Chapter Goals

- Diagnosing input issues.

General Guidelines

- Tinkering with examples by tuning their code, Gaining more familiarity and intuition.
- Visualizing input representations.

Modules & Datasets Setup

```
# @title
!apt-get install default-jdk
!apt install libgraphviz-dev

Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following additional packages will be installed:
  default-jdk-headless default-jre default-jre-headless fonts-dejavu-
core
  fonts-dejavu-extra libatk-wrapper-java libatk-wrapper-java-jni
libfontenc1
  libice-dev libsm-dev libxkbfile1 libxt-dev libxtst6 libxxf86dga1
  openjdk-11-jdk openjdk-11-jre x11-utils
Suggested packages:
  libice-doc libsm-doc libxt-doc openjdk-11-demo openjdk-11-source
visualvm
  mesa-utils
The following NEW packages will be installed:
  default-jdk default-jdk-headless default-jre default-jre-headless
  fonts-dejavu-core fonts-dejavu-extra libatk-wrapper-java
  libatk-wrapper-java-jni libfontenc1 libice-dev libsm-dev libxkbfile1
  libxt-dev libxtst6 libxxf86dga1 openjdk-11-jdk openjdk-11-jre x11-
utils
0 upgraded, 18 newly installed, 0 to remove and 18 not upgraded.
Need to get 5,518 kB of archives.
After this operation, 15.8 MB of additional disk space will be used.
Get:1 http://archive.ubuntu.com/ubuntu jammy/main amd64 default-jre-
headless amd64 2:1.11-72build2 [3,042 B]
Get:2 http://archive.ubuntu.com/ubuntu jammy/main amd64 libxtst6 amd64
2:1.2.3-1build4 [13.4 kB]
Get:3 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64
openjdk-11-jre amd64 11.0.20.1+1-0ubuntu1~22.04 [213 kB]
```

```
Get:4 http://archive.ubuntu.com/ubuntu jammy/main amd64 default-jre
amd64 2:1.11-72build2 [896 B]
Get:5 http://archive.ubuntu.com/ubuntu jammy/main amd64 default-jdk-
headless amd64 2:1.11-72build2 [942 B]
Get:6 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64
openjdk-11-jdk amd64 11.0.20.1+1-0ubuntu1~22.04 [1,331 kB]
Get:7 http://archive.ubuntu.com/ubuntu jammy/main amd64 default-jdk
amd64 2:1.11-72build2 [908 B]
Get:8 http://archive.ubuntu.com/ubuntu jammy/main amd64 fonts-dejavu-
core all 2.37-2build1 [1,041 kB]
Get:9 http://archive.ubuntu.com/ubuntu jammy/main amd64 fonts-dejavu-
extra all 2.37-2build1 [2,041 kB]
Get:10 http://archive.ubuntu.com/ubuntu jammy/main amd64 libfontenc1
amd64 1:1.1.4-1build3 [14.7 kB]
Get:11 http://archive.ubuntu.com/ubuntu jammy/main amd64 libxkbfile1
amd64 1:1.1.0-1build3 [71.8 kB]
Get:12 http://archive.ubuntu.com/ubuntu jammy/main amd64 libxxf86dga1
amd64 2:1.1.5-0ubuntu3 [12.6 kB]
Get:13 http://archive.ubuntu.com/ubuntu jammy/main amd64 x11-utils
amd64 7.7+5build2 [206 kB]
Get:14 http://archive.ubuntu.com/ubuntu jammy/main amd64 libatk-
wrapper-java all 0.38.0-5build1 [53.1 kB]
Get:15 http://archive.ubuntu.com/ubuntu jammy/main amd64 libatk-
wrapper-java-jni amd64 0.38.0-5build1 [49.0 kB]
Get:16 http://archive.ubuntu.com/ubuntu jammy/main amd64 libice-dev
amd64 2:1.0.10-1build2 [51.4 kB]
Get:17 http://archive.ubuntu.com/ubuntu jammy/main amd64 libsm-dev
amd64 2:1.2.3-1build2 [18.1 kB]
Get:18 http://archive.ubuntu.com/ubuntu jammy/main amd64 libxt-dev
amd64 1:1.2.1-1 [396 kB]
Fetched 5,518 kB in 1s (5,313 kB/s)
Selecting previously unselected package default-jre-headless.
(Reading database ... 120895 files and directories currently
installed.)
Preparing to unpack .../00-default-jre-headless_2%3a1.11-
72build2_amd64.deb ...
Unpacking default-jre-headless (2:1.11-72build2) ...
Selecting previously unselected package libxtst6:amd64.
Preparing to unpack .../01-libxtst6_2%3a1.2.3-1build4_amd64.deb ...
Unpacking libxtst6:amd64 (2:1.2.3-1build4) ...
Selecting previously unselected package openjdk-11-jre:amd64.
Preparing to unpack .../02-openjdk-11-jre_11.0.20.1+1-
0ubuntu1~22.04_amd64.deb ...
Unpacking openjdk-11-jre:amd64 (11.0.20.1+1-0ubuntu1~22.04) ...
Selecting previously unselected package default-jre.
Preparing to unpack .../03-default-jre_2%3a1.11-72build2_amd64.deb ...
Unpacking default-jre (2:1.11-72build2) ...
Selecting previously unselected package default-jdk-headless.
Preparing to unpack .../04-default-jdk-headless_2%3a1.11-
```

```
72build2_amd64.deb ...
Unpacking default-jdk-headless (2:1.11-72build2) ...
Selecting previously unselected package openjdk-11-jdk:amd64.
Preparing to unpack .../05-openjdk-11-jdk_11.0.20.1+1-
0ubuntu1~22.04_amd64.deb ...
Unpacking openjdk-11-jdk:amd64 (11.0.20.1+1-0ubuntu1~22.04) ...
Selecting previously unselected package default-jdk.
Preparing to unpack .../06-default-jdk_2%3a1.11-72build2_amd64.deb ...
Unpacking default-jdk (2:1.11-72build2) ...
Selecting previously unselected package fonts-dejavu-core.
Preparing to unpack .../07-fonts-dejavu-core_2.37-2build1_all.deb ...
Unpacking fonts-dejavu-core (2.37-2build1) ...
Selecting previously unselected package fonts-dejavu-extra.
Preparing to unpack .../08-fonts-dejavu-extra_2.37-2build1_all.deb ...
Unpacking fonts-dejavu-extra (2.37-2build1) ...
Selecting previously unselected package libfontenc1:amd64.
Preparing to unpack .../09-libfontenc1_1%3a1.1.4-1build3_amd64.deb ...
Unpacking libfontenc1:amd64 (1:1.1.4-1build3) ...
Selecting previously unselected package libxkbfile1:amd64.
Preparing to unpack .../10-libxkbfile1_1%3a1.1.0-1build3_amd64.deb ...
Unpacking libxkbfile1:amd64 (1:1.1.0-1build3) ...
Selecting previously unselected package libxxf86dga1:amd64.
Preparing to unpack .../11-libxxf86dga1_2%3a1.1.5-
0ubuntu3_amd64.deb ...
Unpacking libxxf86dga1:amd64 (2:1.1.5-0ubuntu3) ...
Selecting previously unselected package x11-utils.
Preparing to unpack .../12-x11-utils_7.7+5build2_amd64.deb ...
Unpacking x11-utils (7.7+5build2) ...
Selecting previously unselected package libatk-wrapper-java.
Preparing to unpack .../13-libatk-wrapper-java_0.38.0-
5build1_all.deb ...
Unpacking libatk-wrapper-java (0.38.0-5build1) ...
Selecting previously unselected package libatk-wrapper-java-jni:amd64.
Preparing to unpack .../14-libatk-wrapper-java-jni_0.38.0-
5build1_amd64.deb ...
Unpacking libatk-wrapper-java-jni:amd64 (0.38.0-5build1) ...
Selecting previously unselected package libice-dev:amd64.
Preparing to unpack .../15-libice-dev_2%3a1.0.10-1build2_amd64.deb ...
Unpacking libice-dev:amd64 (2:1.0.10-1build2) ...
Selecting previously unselected package libsm-dev:amd64.
Preparing to unpack .../16-libsm-dev_2%3a1.2.3-1build2_amd64.deb ...
Unpacking libsm-dev:amd64 (2:1.2.3-1build2) ...
Selecting previously unselected package libxt-dev:amd64.
Preparing to unpack .../17-libxt-dev_1%3a1.2.1-1_amd64.deb ...
Unpacking libxt-dev:amd64 (1:1.2.1-1) ...
Setting up default-jre-headless (2:1.11-72build2) ...
Setting up libice-dev:amd64 (2:1.0.10-1build2) ...
Setting up libsm-dev:amd64 (2:1.2.3-1build2) ...
Setting up libxtst6:amd64 (2:1.2.3-1build4) ...
```

```
Setting up libxxf86dga1:amd64 (2:1.1.5-0ubuntu3) ...
Setting up openjdk-11-jre:amd64 (11.0.20.1+1-0ubuntu1~22.04) ...
Setting up default-jre (2:1.11-72build2) ...
Setting up libfontenc1:amd64 (1:1.1.4-1build3) ...
Setting up default-jdk-headless (2:1.11-72build2) ...
Setting up libxt-dev:amd64 (1:1.2.1-1) ...
Setting up fonts-dejavu-core (2.37-2build1) ...
Setting up fonts-dejavu-extra (2.37-2build1) ...
Setting up openjdk-11-jdk:amd64 (11.0.20.1+1-0ubuntu1~22.04) ...
update-alternatives: using
/usr/lib/jvm/java-11-openjdk-amd64/bin/jconsole to provide
/usr/bin/jconsole (jconsole) in auto mode
Setting up libxkbfile1:amd64 (1:1.1.0-1build3) ...
Setting up default-jdk (2:1.11-72build2) ...
Setting up x11-utils (7.7+5build2) ...
Setting up libatk-wrapper-java (0.38.0-5build1) ...
Setting up libatk-wrapper-java-jni:amd64 (0.38.0-5build1) ...
Processing triggers for hicolor-icon-theme (0.17-2) ...
Processing triggers for libc-bin (2.35-0ubuntu3.1) ...
/sbin/ldconfig.real: /usr/local/lib/libtbbbind_2_0.so.3 is not a
symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbb.so.12 is not a symbolic
link

/sbin/ldconfig.real: /usr/local/lib/libtbbmalloc.so.2 is not a
symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbbmalloc_proxy.so.2 is not a
symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbbbind_2_5.so.3 is not a
symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbbbind.so.3 is not a symbolic
link

Processing triggers for man-db (2.10.2-1) ...
Processing triggers for fontconfig (2.13.1-4.2ubuntu5) ...
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following additional packages will be installed:
  libgail-common libgail18 libgtk2.0-0 libgtk2.0-bin libgtk2.0-common
  libgvc6-plugins-gtk librsvg2-common libxdot4
Suggested packages:
  gvfs
The following NEW packages will be installed:
  libgail-common libgail18 libgraphviz-dev libgtk2.0-0 libgtk2.0-bin
  libgtk2.0-common libgvc6-plugins-gtk librsvg2-common libxdot4
```

```
0 upgraded, 9 newly installed, 0 to remove and 18 not upgraded.
Need to get 2,433 kB of archives.
After this operation, 7,694 kB of additional disk space will be used.
Get:1 http://archive.ubuntu.com/ubuntu jammy/main amd64 libgtk2.0-
common all 2.24.33-2ubuntu2 [125 kB]
Get:2 http://archive.ubuntu.com/ubuntu jammy/main amd64 libgtk2.0-0
amd64 2.24.33-2ubuntu2 [2,037 kB]
Get:3 http://archive.ubuntu.com/ubuntu jammy/main amd64 libgail18
amd64 2.24.33-2ubuntu2 [15.9 kB]
Get:4 http://archive.ubuntu.com/ubuntu jammy/main amd64 libgail-common
amd64 2.24.33-2ubuntu2 [132 kB]
Get:5 http://archive.ubuntu.com/ubuntu jammy/universe amd64 libxdot4
amd64 2.42.2-6 [16.4 kB]
Get:6 http://archive.ubuntu.com/ubuntu jammy/universe amd64 libgvc6-
plugins-gtk amd64 2.42.2-6 [22.6 kB]
Get:7 http://archive.ubuntu.com/ubuntu jammy/universe amd64
libgraphviz-dev amd64 2.42.2-6 [58.5 kB]
Get:8 http://archive.ubuntu.com/ubuntu jammy/main amd64 libgtk2.0-bin
amd64 2.24.33-2ubuntu2 [7,932 B]
Get:9 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64
librsvg2-common amd64 2.52.5+dfsg-3ubuntu0.2 [17.7 kB]
Fetched 2,433 kB in 1s (2,602 kB/s)
Selecting previously unselected package libgtk2.0-common.
(Reading database ... 121406 files and directories currently
installed.)
Preparing to unpack .../0-libgtk2.0-common_2.24.33-
2ubuntu2_all.deb ...
Unpacking libgtk2.0-common (2.24.33-2ubuntu2) ...
Selecting previously unselected package libgtk2.0-0:amd64.
Preparing to unpack .../1-libgtk2.0-0_2.24.33-2ubuntu2_amd64.deb ...
Unpacking libgtk2.0-0:amd64 (2.24.33-2ubuntu2) ...
Selecting previously unselected package libgail18:amd64.
Preparing to unpack .../2-libgail18_2.24.33-2ubuntu2_amd64.deb ...
Unpacking libgail18:amd64 (2.24.33-2ubuntu2) ...
Selecting previously unselected package libgail-common:amd64.
Preparing to unpack .../3-libgail-common_2.24.33-
2ubuntu2_amd64.deb ...
Unpacking libgail-common:amd64 (2.24.33-2ubuntu2) ...
Selecting previously unselected package libxdot4:amd64.
Preparing to unpack .../4-libxdot4_2.42.2-6_amd64.deb ...
Unpacking libxdot4:amd64 (2.42.2-6) ...
Selecting previously unselected package libgvc6-plugins-gtk.
Preparing to unpack .../5-libgvc6-plugins-gtk_2.42.2-6_amd64.deb ...
Unpacking libgvc6-plugins-gtk (2.42.2-6) ...
Selecting previously unselected package libgraphviz-dev:amd64.
Preparing to unpack .../6-libgraphviz-dev_2.42.2-6_amd64.deb ...
Unpacking libgraphviz-dev:amd64 (2.42.2-6) ...
Selecting previously unselected package libgtk2.0-bin.
Preparing to unpack .../7-libgtk2.0-bin_2.24.33-2ubuntu2_amd64.deb ...
```

```
Unpacking libgtk2.0-bin (2.24.33-2ubuntu2) ...
Selecting previously unselected package librsvg2-common:amd64.
Preparing to unpack .../8-librsvg2-common_2.52.5+dfsg-
3ubuntu0.2_amd64.deb ...
Unpacking librsvg2-common:amd64 (2.52.5+dfsg-3ubuntu0.2) ...
Setting up libxdot4:amd64 (2.42.2-6) ...
Setting up librsvg2-common:amd64 (2.52.5+dfsg-3ubuntu0.2) ...
Setting up libgtk2.0-common (2.24.33-2ubuntu2) ...
Setting up libgtk2.0-0:amd64 (2.24.33-2ubuntu2) ...
Setting up libgvc6-plugins-gtk (2.42.2-6) ...
Setting up libgail18:amd64 (2.24.33-2ubuntu2) ...
Setting up libgtk2.0-bin (2.24.33-2ubuntu2) ...
Setting up libgail-common:amd64 (2.24.33-2ubuntu2) ...
Setting up libgraphviz-dev:amd64 (2.42.2-6) ...
Processing triggers for libc-bin (2.35-0ubuntu3.1) ...
/sbin/ldconfig.real: /usr/local/lib/libtbbbind_2_0.so.3 is not a
symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbb.so.12 is not a symbolic
link

/sbin/ldconfig.real: /usr/local/lib/libtbbmalloc.so.2 is not a
symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbbmalloc_proxy.so.2 is not a
symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbbbind_2_5.so.3 is not a
symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbbbind.so.3 is not a symbolic
link

Processing triggers for man-db (2.10.2-1) ...
Processing triggers for libgdk-pixbuf-2.0-0:amd64 (2.42.8+dfsg-
1ubuntu0.2) ...

# @title
!pip install pygraphviz
!pip install python-javabridge
!pip install python-weka-wrapper3
!pip install sklearn-weka-plugin

Collecting pygraphviz
  Downloading pygraphviz-1.11.zip (120 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 0.0/120.8 kB ? eta -:--:--
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ ━━ 112.6/120.8 kB 4.1 MB/s eta
0:00:01 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 120.8/120.8 kB 3.1
MB/s eta 0:00:00
etadata (setup.py) ... e=pygraphviz-1.11-cp310-cp310-linux_x86_64.whl
```

```
size=175923
sha256=e9e590f2cd5a049d9f9aba02b52073b0366984b2b96dd235953d534c5b5147d
e
  Stored in directory:
/root/.cache/pip/wheels/5b/ee/36/f47a0d35664fbe1a2b5a433ae33c6ad636b00
bb231f68a9aaa
Successfully built pygraphviz
Installing collected packages: pygraphviz
Successfully installed pygraphviz-1.11
Collecting python-javabridge
  Downloading python-javabridge-4.0.3.tar.gz (1.3 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.3/1.3 MB 11.4 MB/s eta
0:00:00
etadata (setup.py) ... ent already satisfied: numpy>=1.20.1 in
/usr/local/lib/python3.10/dist-packages (from python-javabridge)
(1.23.5)
Building wheels for collected packages: python-javabridge
  Building wheel for python-javabridge (setup.py) ...
e=python_javabridge-4.0.3-cp310-cp310-linux_x86_64.whl size=1743154
sha256=c5c5a87058412c3257b3d323435e41d6c062576ded08b0c0f2c1e491c6425ad
5
  Stored in directory:
/root/.cache/pip/wheels/35/58/be/c5d71b71a9dd6585f897fa5b2d021e03962eb
30d6b20797396
Successfully built python-javabridge
Installing collected packages: python-javabridge
Successfully installed python-javabridge-4.0.3
Collecting python-weka-wrapper3
  Downloading python-weka-wrapper3-0.2.14.tar.gz (15.9 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 15.9/15.9 MB 13.7 MB/s eta
0:00:00
etadata (setup.py) ... ent already satisfied: python-javabridge>=4.0.0
in /usr/local/lib/python3.10/dist-packages (from python-weka-wrapper3)
(4.0.3)
Requirement already satisfied: numpy in
/usr/local/lib/python3.10/dist-packages (from python-weka-wrapper3)
(1.23.5)
Requirement already satisfied: packaging in
/usr/local/lib/python3.10/dist-packages (from python-weka-wrapper3)
(23.1)
Collecting configurable-objects (from python-weka-wrapper3)
  Downloading configurable-objects-0.0.1.tar.gz (4.4 kB)
  Preparing metadata (setup.py) ... ple-data-flow (from python-weka-
wrapper3)
  Downloading simple-data-flow-0.0.1.tar.gz (16 kB)
  Preparing metadata (setup.py) ... ple-data-flow
  Building wheel for python-weka-wrapper3 (setup.py) ...
e=python_weka_wrapper3-0.2.14-py3-none-any.whl size=14496261
sha256=0e9abe14d08466e568df4631849bb20233e7a192240d2c1290334612d5808c6
```

```
9
  Stored in directory:
/root/.cache/pip/wheels/80/c5/f2/412fa8d3b181151e11b68d46daa52f96e9b83
2a2eca4bc6c88
  Building wheel for configurable-objects (setup.py) ...
e=configurable_objects-0.0.1-py3-none-any.whl size=4695
sha256=11b5e5fa27fd6933c519301e1336ec4f88d63c8065bdc321ac131c4b355f0ae
e
  Stored in directory:
/root/.cache/pip/wheels/ef/11/bc/75ac8b0592c38dc42412942c37d3947faf0b2
22bad150132a1
  Building wheel for simple-data-flow (setup.py) ... ple-data-flow:
filename=simple_data_flow-0.0.1-py3-none-any.whl size=19063
sha256=e225d77d5a198d841cf5e217d47ba090036b7b3b4c855ae6674aaab1f783157
9
  Stored in directory:
/root/.cache/pip/wheels/b3/02/23/4aec0db3dae7152dd268d6de385905116af55
229c1a8e81303
Successfully built python-weka-wrapper3 configurable-objects simple-
data-flow
Installing collected packages: configurable-objects, simple-data-flow,
python-weka-wrapper3
Successfully installed configurable-objects-0.0.1 python-weka-
wrapper3-0.2.14 simple-data-flow-0.0.1
Collecting sklearn-weka-plugin
  Downloading sklearn-weka-plugin-0.0.7.tar.gz (69 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 69.8/69.8 kB 1.4 MB/s eta
0:00:00
etadata (setup.py) ... ent already satisfied: numpy in
/usr/local/lib/python3.10/dist-packages (from sklearn-weka-plugin)
(1.23.5)
Requirement already satisfied: python-weka-wrapper3>=0.2.5 in
/usr/local/lib/python3.10/dist-packages (from sklearn-weka-plugin)
(0.2.14)
Collecting sklearn (from sklearn-weka-plugin)
  Downloading sklearn-0.0.post9.tar.gz (3.6 kB)
  Preparing metadata (setup.py) ... ent already satisfied: python-
javabridge>=4.0.0 in /usr/local/lib/python3.10/dist-packages (from
python-weka-wrapper3>=0.2.5->sklearn-weka-plugin) (4.0.3)
Requirement already satisfied: packaging in
/usr/local/lib/python3.10/dist-packages (from python-weka-
wrapper3>=0.2.5->sklearn-weka-plugin) (23.1)
Requirement already satisfied: configurable-objects in
/usr/local/lib/python3.10/dist-packages (from python-weka-
wrapper3>=0.2.5->sklearn-weka-plugin) (0.0.1)
Requirement already satisfied: simple-data-flow in
/usr/local/lib/python3.10/dist-packages (from python-weka-
wrapper3>=0.2.5->sklearn-weka-plugin) (0.0.1)
Building wheels for collected packages: sklearn-weka-plugin, sklearn
```

```
  Building wheel for sklearn-weka-plugin (setup.py) ...
e=sklearn_weka_plugin-0.0.7-py3-none-any.whl size=27346
sha256=16da304e5f99621b04488ff61c56ee93f0d262b81c243fd21b2d55b3bd8a8e1
8
  Stored in directory:
/root/.cache/pip/wheels/51/6d/e5/458ea9a1be729f39ed4cf14aab2f87eb51470
47b690402605b
  Building wheel for sklearn (setup.py) ... e=sklearn-0.0.post9-py3-
none-any.whl size=2952
sha256=149c7e5b99d540c7bbb02dc296cb7af9fa5284158c9f2a0156f8ce1372d7601
8
  Stored in directory:
/root/.cache/pip/wheels/33/a3/d2/092b519e9522b4c91608b7dcec0dd9051fa1b
ff4c45f4502d1
Successfully built sklearn-weka-plugin sklearn
Installing collected packages: sklearn, sklearn-weka-plugin
Successfully installed sklearn-0.0.post9 sklearn-weka-plugin-0.0.7

# @title
#Restart runtime after installing the dependencies

# @title
import os
import glob
import numpy as np
import pandas as pd
import weka.core.jvm as jvm
from weka.core import converters
import matplotlib.pyplot as plt

# @title
data_dir = 'data'

# @title
#!rm -r weka
#!rm -r data

# @title
#jvm.stop()
jvm.start(packages=True)

DEBUG:weka.core.jvm:Adding bundled jars
DEBUG:weka.core.jvm:Classpath=['/usr/local/lib/python3.10/dist-
packages/javabridge/jars/rhino-1.7R4.jar',
'/usr/local/lib/python3.10/dist-packages/javabridge/jars/runnablequeue
.jar',
'/usr/local/lib/python3.10/dist-packages/javabridge/jars/cpython.jar',
'/usr/local/lib/python3.10/dist-packages/weka/lib/mtj.jar',
'/usr/local/lib/python3.10/dist-packages/weka/lib/core.jar',
'/usr/local/lib/python3.10/dist-packages/weka/lib/weka.jar',
'/usr/local/lib/python3.10/dist-packages/weka/lib/arpack_combined.jar'
```

```
, '/usr/local/lib/python3.10/dist-packages/weka/lib/python-weka-
wrapper.jar']
DEBUG:weka.core.jvm:MaxHeapSize=default
DEBUG:weka.core.jvm:Package support enabled

# @title
# Preparing Datasets
if not os.path.exists(data_dir):
    !mkdir $data_dir
    for file in ['airline.arff', 'breast-cancer.arff', 'contact-
lenses.arff', 'cpu.arff', 'cpu.with.vendor.arff', 'credit-g.arff',
'diabetes.arff', 'glass.arff', 'hypothyroid.arff', 'ionosphere.arff',
'iris.2D.arff', 'iris.arff', 'labor.arff', 'segment-challenge.arff',
'segment-test.arff', 'soybean.arff', 'supermarket.arff',
'unbalanced.arff', 'vote.arff', 'weather.nominal.arff',
'weather.numeric.arff',]:
        url =
'https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/' + file
        !wget -P $data_dir $url
    loader =
converters.Loader(classname="weka.core.converters.ArffLoader")
    saver =
converters.Saver(classname="weka.core.converters.CSVSaver")
    for file in glob.glob(os.path.join(data_dir, '*.arff')):
        dataset = loader.load_file(file)
        filename, file_extension = os.path.splitext(file)
        saver.save_file(dataset, filename + '.csv')
    !wget -P $data_dir https://raw.githubusercontent.com/Rytuo/ITMO-
CT/master/Others/AdvancedML/data/OpenML/data/1438.arff
    !rm -r weka
```

```
--2023-09-29 11:53:24--
https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/airline.arff
Resolving git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)...
130.217.218.43
Connecting to git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)|
130.217.218.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 2357 (2.3K) [text/plain]
Saving to: 'data/airline.arff'

airline.arff        100%[===================>]   2.30K  --.-KB/s    in
0s

2023-09-29 11:53:26 (1.18 GB/s) - 'data/airline.arff' saved
[2357/2357]

--2023-09-29 11:53:26--
```

```
https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/breast-cancer.arff
Resolving git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)...
130.217.218.43
Connecting to git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)|
130.217.218.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 29418 (29K) [text/plain]
Saving to: 'data/breast-cancer.arff'

breast-cancer.arff  100%[===================>]  28.73K  94.1KB/s    in
0.3s

2023-09-29 11:53:27 (94.1 KB/s) - 'data/breast-cancer.arff' saved
[29418/29418]

--2023-09-29 11:53:28--
https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/contact-lenses.arff
Resolving git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)...
130.217.218.43
Connecting to git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)|
130.217.218.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 2890 (2.8K) [text/plain]
Saving to: 'data/contact-lenses.arff'

contact-lenses.arff 100%[===================>]   2.82K  --.-KB/s    in
0s

2023-09-29 11:53:29 (1.11 GB/s) - 'data/contact-lenses.arff' saved
[2890/2890]

--2023-09-29 11:53:29--
https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/cpu.arff
Resolving git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)...
130.217.218.43
Connecting to git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)|
130.217.218.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 5561 (5.4K) [text/plain]
Saving to: 'data/cpu.arff'

cpu.arff            100%[===================>]   5.43K  --.-KB/s    in
0s

2023-09-29 11:53:30 (107 MB/s) - 'data/cpu.arff' saved [5561/5561]

--2023-09-29 11:53:30--
```

```
https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/cpu.with.vendor.arff
Resolving git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)...
130.217.218.43
Connecting to git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)|
130.217.218.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 6960 (6.8K) [text/plain]
Saving to: 'data/cpu.with.vendor.arff'

cpu.with.vendor.arf 100%[====================>]   6.80K  --.-KB/s    in
0s

2023-09-29 11:53:31 (114 MB/s) - 'data/cpu.with.vendor.arff' saved
[6960/6960]

--2023-09-29 11:53:31--
https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/credit-g.arff
Resolving git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)...
130.217.218.43
Connecting to git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)|
130.217.218.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 162270 (158K) [text/plain]
Saving to: 'data/credit-g.arff'

credit-g.arff        100%[====================>] 158.47K    173KB/s    in
0.9s

2023-09-29 11:53:34 (173 KB/s) - 'data/credit-g.arff' saved
[162270/162270]

--2023-09-29 11:53:34--
https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/diabetes.arff
Resolving git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)...
130.217.218.43
Connecting to git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)|
130.217.218.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 37443 (37K) [text/plain]
Saving to: 'data/diabetes.arff'

diabetes.arff        100%[====================>]  36.57K    120KB/s    in
0.3s

2023-09-29 11:53:35 (120 KB/s) - 'data/diabetes.arff' saved
[37443/37443]
```

```
--2023-09-29 11:53:35--
https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/glass.arff
Resolving git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)...
130.217.218.43
Connecting to git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)|
130.217.218.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 17850 (17K) [text/plain]
Saving to: 'data/glass.arff'

glass.arff          100%[===================>]  17.43K  55.9KB/s    in
0.3s

2023-09-29 11:53:37 (55.9 KB/s) - 'data/glass.arff' saved
[17850/17850]

--2023-09-29 11:53:37--
https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/hypothyroid.arff
Resolving git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)...
130.217.218.43
Connecting to git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)|
130.217.218.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 310897 (304K) [text/plain]
Saving to: 'data/hypothyroid.arff'

hypothyroid.arff    100%[===================>] 303.61K   249KB/s    in
1.2s

2023-09-29 11:53:39 (249 KB/s) - 'data/hypothyroid.arff' saved
[310897/310897]

--2023-09-29 11:53:40--
https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/ionosphere.arff
Resolving git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)...
130.217.218.43
Connecting to git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)|
130.217.218.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 80487 (79K) [text/plain]
Saving to: 'data/ionosphere.arff'

ionosphere.arff     100%[===================>]  78.60K  85.7KB/s    in
0.9s

2023-09-29 11:53:42 (85.7 KB/s) - 'data/ionosphere.arff' saved
[80487/80487]
```

```
--2023-09-29 11:53:42--
https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/iris.2D.arff
Resolving git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)...
130.217.218.43
Connecting to git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)|
130.217.218.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 3492 (3.4K) [text/plain]
Saving to: 'data/iris.2D.arff'

iris.2D.arff          100%[===================>]   3.41K  --.-KB/s    in
0s

2023-09-29 11:53:43 (41.9 MB/s) - 'data/iris.2D.arff' saved
[3492/3492]

--2023-09-29 11:53:43--
https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/iris.arff
Resolving git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)...
130.217.218.43
Connecting to git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)|
130.217.218.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 7486 (7.3K) [text/plain]
Saving to: 'data/iris.arff'

iris.arff             100%[===================>]   7.31K  --.-KB/s    in
0s

2023-09-29 11:53:44 (63.8 MB/s) - 'data/iris.arff' saved [7486/7486]

--2023-09-29 11:53:44--
https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/labor.arff
Resolving git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)...
130.217.218.43
Connecting to git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)|
130.217.218.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 8279 (8.1K) [text/plain]
Saving to: 'data/labor.arff'

labor.arff            100%[===================>]   8.08K  --.-KB/s    in
0s

2023-09-29 11:53:46 (113 MB/s) - 'data/labor.arff' saved [8279/8279]
```

```
--2023-09-29 11:53:46--
https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/segment-challenge.arff
Resolving git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)...
130.217.218.43
Connecting to git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)|
130.217.218.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 200410 (196K) [text/plain]
Saving to: 'data/segment-challenge.arff'

segment-challenge.a 100%[===================>] 195.71K   160KB/s    in
1.2s

2023-09-29 11:53:48 (160 KB/s) - 'data/segment-challenge.arff' saved
[200410/200410]

--2023-09-29 11:53:48--
https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/segment-test.arff
Resolving git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)...
130.217.218.43
Connecting to git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)|
130.217.218.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 109984 (107K) [text/plain]
Saving to: 'data/segment-test.arff'

segment-test.arff   100%[===================>] 107.41K   117KB/s    in
0.9s

2023-09-29 11:53:50 (117 KB/s) - 'data/segment-test.arff' saved
[109984/109984]

--2023-09-29 11:53:51--
https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/soybean.arff
Resolving git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)...
130.217.218.43
Connecting to git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)|
130.217.218.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 202935 (198K) [text/plain]
Saving to: 'data/soybean.arff'

soybean.arff        100%[===================>] 198.18K   162KB/s    in
1.2s

2023-09-29 11:53:53 (162 KB/s) - 'data/soybean.arff' saved
[202935/202935]
```

```
--2023-09-29 11:53:53--
https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/supermarket.arff
Resolving git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)...
130.217.218.43
Connecting to git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)|
130.217.218.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 2025871 (1.9M) [text/plain]
Saving to: 'data/supermarket.arff'

supermarket.arff    100%[===================>]   1.93M   926KB/s    in
2.1s

2023-09-29 11:53:57 (926 KB/s) - 'data/supermarket.arff' saved
[2025871/2025871]

--2023-09-29 11:53:57--
https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/unbalanced.arff
Resolving git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)...
130.217.218.43
Connecting to git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)|
130.217.218.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 186360 (182K) [text/plain]
Saving to: 'data/unbalanced.arff'

unbalanced.arff     100%[===================>] 181.99K   199KB/s    in
0.9s

2023-09-29 11:54:00 (199 KB/s) - 'data/unbalanced.arff' saved
[186360/186360]

--2023-09-29 11:54:00--
https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/vote.arff
Resolving git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)...
130.217.218.43
Connecting to git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)|
130.217.218.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 40261 (39K) [text/plain]
Saving to: 'data/vote.arff'

vote.arff           100%[===================>]  39.32K   129KB/s    in
0.3s

2023-09-29 11:54:01 (129 KB/s) - 'data/vote.arff' saved [40261/40261]
```

```
--2023-09-29 11:54:01--
https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/weather.nominal.arff
Resolving git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)...
130.217.218.43
Connecting to git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)|
130.217.218.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 587 [text/plain]
Saving to: 'data/weather.nominal.arff'

weather.nominal.arf 100%[===================>]     587  --.-KB/s    in
0s

2023-09-29 11:54:03 (355 MB/s) - 'data/weather.nominal.arff' saved
[587/587]

--2023-09-29 11:54:03--
https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/weather.numeric.arff
Resolving git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)...
130.217.218.43
Connecting to git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)|
130.217.218.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 495 [text/plain]
Saving to: 'data/weather.numeric.arff'

weather.numeric.arf 100%[===================>]     495  --.-KB/s    in
0s

2023-09-29 11:54:04 (133 MB/s) - 'data/weather.numeric.arff' saved
[495/495]

--2023-09-29 11:54:08--  https://raw.githubusercontent.com/Rytuo/ITMO-
CT/master/Others/AdvancedML/data/OpenML/data/1438.arff
Resolving raw.githubusercontent.com (raw.githubusercontent.com)...
185.199.108.133, 185.199.109.133, 185.199.110.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|
185.199.108.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 25381 (25K) [text/plain]
Saving to: 'data/1438.arff'

1438.arff            100%[===================>]  24.79K  --.-KB/s    in
0.001s

2023-09-29 11:54:08 (33.9 MB/s) - 'data/1438.arff' saved [25381/25381]
```

```
rm: cannot remove 'weka': No such file or directory
```

2.1 What's a Concept?

```python
from sklearn.utils import all_estimators
for typeFilter in ['classifier', 'regressor', 'cluster',
'transformer']:
    print(typeFilter)
    for estimatorName, estimatorClass in
all_estimators(type_filter=typeFilter):
        print('\t', estimatorName)

for i in
pkgutil.iter_modules(path=weka.__path__,prefix="weka.classifiers."):
 print(i.name)
```

```
weka.classifiers.associations
weka.classifiers.attribute_selection
weka.classifiers.classifiers
weka.classifiers.clusterers
weka.classifiers.core
weka.classifiers.datagenerators
weka.classifiers.experiments
weka.classifiers.filters
weka.classifiers.flow
weka.classifiers.plot
weka.classifiers.timeseries
```

```python
import weka.core.packages as packages
for module in ['classifiers', 'clusterers', 'associations']:
    print(module)
    for suggestion in packages.suggest_package('weka.' + module):
        print('\t', suggestion)
```

```
classifiers
        alternatingDecisionTrees
        alternatingModelTrees
        averagedOneDependenceEstimators
        bayesianLogisticRegression
        bestFirstTree
        citationKNN
        classAssociationRules
        classificationViaClustering
        clojureClassifier
        complementNaiveBayes
        conjunctiveRule
        dagging
        decorate
```

discriminantAnalysis
distributedWekaBase
distributedWekaSparkDev
DMNBtext
DTNB
dualPerturbAndCombine
elasticNet
ensembleLibrary
ensemblesOfNestedDichotomies
extraTrees
functionalTrees
fuzzyLaticeReasoning
fuzzyUnorderedRuleInduction
grading
gridSearch
hiddenNaiveBayes
hyperPipes
isolationForest
isotonicRegression
iterativeAbsoluteErrorRegression
J48graft
kerasZoo
kernelLogisticRegression
largeScaleKernelLearning
lazyBayesianRules
leastMedSquared
LibLINEAR
LibSVM
localOutlierFactor
logarithmicErrorMetrics
massiveOnlineAnalysis
metaCost
multiBoostAB
multiInstanceLearning
multilayerPerceptronCS
multiLayerPerceptrons
naiveBayesTree
NNge
oneClassClassifier
ordinalClassClassifier
ordinalLearningMethod
ordinalStochasticDominance
paceRegression
partialLeastSquares
percentageErrorMetrics
probabilityCalibrationTrees
racedIncrementalLogitBoost
RBFNetwork
realAdaBoost

```
        ridor
        rotationForest
        RPlugin
        scriptingClassifiers
        simpleCART
        simpleEducationalLearningSchemes
        SPegasos
        stackingC
        supervisedAttributeScaling
        thresholdSelector
        timeseriesForecasting
        userClassifier
        votingFeatureIntervals
        wekaPython
        winnow
clusterers
        cascadeKMeans
        CLOPE
        distributedWekaBase
        optics_dbScan
        sequentialInformationalBottleneckClusterer
        wekaPython
        XMeans
associations
        classAssociationRules
        generalizedSequentialPatterns
        hotSpot
        predictiveApriori
        tertius
```

2.2 What's in an Example?

```
df = pd.read_csv("data/weather.numeric.csv")
df
```

```
      outlook  temperature  humidity  windy play
0       sunny           85        85  False   no
1       sunny           80        90   True   no
2    overcast           83        86  False  yes
3       rainy           70        96  False  yes
4       rainy           68        80  False  yes
5       rainy           65        70   True   no
6    overcast           64        65   True  yes
7       sunny           72        95  False   no
8       sunny           69        70  False  yes
9       rainy           75        80  False  yes
10      sunny           75        70   True  yes
11   overcast           72        90   True  yes
```

```
12   overcast            81        75  False  yes
13      rainy            71        91   True   no
```

Task 2.2.1 Show the instance with highest temperature.

Relations

```
df = pd.DataFrame(
    [
      ["Peter","Male","?","?"],
      ["Peggy","Female","?","?"],
      ["Steven","Male","Peter","Peggy"],
      ["Graham","Male","Peter","Peggy"],
      ["Pam","Female","Peter","Peggy"],
      ["Ian","Male","Grace","Ray"]
    ],

    columns = ["Name", "Gender", "Parent1", "Parent2"]
)

df
```

```
     Name  Gender Parent1 Parent2
0   Peter    Male       ?       ?
1   Peggy  Female       ?       ?
2  Steven    Male   Peter   Peggy
3  Graham    Male   Peter   Peggy
4     Pam  Female   Peter   Peggy
5     Ian    Male   Grace     Ray
```

Task 2.2.2 Who is the parent of *Steven*?

Task 2.2.3 Modify the dataframe so that *Peter* has children more than *Peggy*.

Other Example Types

```
loader =
converters.Loader(classname="weka.core.converters.ArffLoader")
dataset = loader.load_file("data/1438.arff")
print(dataset)

@relation accelerometer

@attribute id {'?'}
@attribute bag relational
@attribute y numeric
@attribute x numeric
@attribute z numeric
@end bag
@attribute class {walking,running,standing,'?'}
```

@data
?,'0.392035,-0.145497,9.850266\n0.489183,-0.286804,9.984836\n0.427212,-0.273332,9.95924\n0.676144,-0.286055,10.198592\n0.796494,-0.182021,9.773476\n0.727787,-0.189955,9.761351\n0.87538,-0.468226,9.888437\n0.571062,-0.402214,10.228979\n0.817899,-0.422571,9.79099\n0.752785,-0.343685,9.985884\n0.714764,-0.302072,9.854158\n0.793201,-0.300276,9.914333\n0.78392,-0.404758,10.169552\n0.768352,-0.248932,10.239158\n1.783092,-0.515977,9.588311\n2.439327,-0.407453,9.943672\n3.572471,-0.169298,9.000483\n4.120481,-0.438139,9.193432\n4.266428,-0.424817,9.113049\n4.453689,-0.412542,8.93956\n4.224665,-0.250729,8.827294\n4.205654,-0.234862,9.073531\n4.17392,-0.268841,8.940159\n4.040997,-0.444276,9.023835\n3.716172,-0.164508,9.269025',standing
?,'9.597592,0.281564,2.32766\n9.59355,0.172292,2.181563\n9.676029,0.239951,2.131418\n9.662257,0.201182,2.196383\n9.660012,0.140109,1.982926\n9.688453,0.171094,1.9476\n9.713301,0.229173,1.958228\n9.720935,0.177381,2.055226\n9.676927,0.140558,1.955833\n9.662108,0.151635,1.994752\n9.724079,0.070054,1.918111\n9.678274,0.097447,1.888772\n9.767339,0.088316,1.807192\n9.676927,0.073048,1.831142\n9.69923,0.127235,1.898652\n9.666748,0.113614,1.802252\n9.683363,0.115859,1.800306\n9.664203,0.133373,1.768572\n9.663305,0.165705,1.794319\n9.665401,0.182021,1.808389\n9.737401,0.165107,1.760938\n9.649085,0.188159,2.059268',standing
?,'0.116608,-9.580527,1.314566\n-0.008981,-9.49266,1.2743\n-0.010778,-9.953102,1.46171\n-0.222288,-9.608968,1.396895\n-0.310903,-9.734407,1.103805\n-0.371078,-9.748328,1.232687\n-0.396825,-9.72827,1.107397\n-0.39907,-9.684711,0.951272\n-0.339045,-9.54086,1.18763\n-0.315244,-9.704469,1.327889\n-0.170645,-9.694291,1.078507\n-0.142953,-9.762399,1.058898\n-0.26435,-9.68501,1.157842\n-0.192949,-9.718989,0.945733\n-0.311053,-9.710307,1.07237\n-0.287253,-9.70881,1.097817\n-0.339344,-9.729168,1.10126\n-0.392484,-9.670041,1.026865\n-0.321232,-9.736802,0.907413\n-0.228275,-9.886341,0.996029\n-0.294438,-9.695937,1.047821\n-0.226629,-9.776919,0.829874\n-0.176633,-9.657916,0.835413\n-0.174836,-9.670191,0.956062',standing
?,'6.181246,-0.448318,7.75642\n6.306685,-0.366139,7.661368\n6.355933,-0.315095,7.656279\n6.418054,-0.345631,7.396269\n6.423592,-0.36075,7.329808\n6.479426,-0.381856,7.592212\n6.468648,-0.310155,7.554939\n6.486312,-0.402663,7.666907\n6.411318,-0.383952,7.434141\n6.66983,-0.446821,7.195686\n6.427634,-0.470771,7.413933\n6.652017,-0.422721,7.59296\n6.488856,-0.451461,7.442224\n6.456524,-0.453108,7.640412\n6.212531,-0.425415,7.654482\n6.162535,-0.436792,7.982151\n6.164331,-0.268542,7.611971\n6.216573,-0.36554,7.610324',standing
?,'3.272345,-1.473835,9.213341\n2.650239,-1.40917,9.568851\n3.037034,-1.598676,9.41557\n3.241958,-1.816323,8.402327\n4.609365,-

1.350342,8.784483\n4.667893,-1.621578,8.372839\n4.771328,-
1.773662,8.657547\n4.887037,-1.826502,8.308622\n5.002747,-
1.857787,8.242909\n5.111121,-1.978586,8.097561\n5.205276,-
2.184557,8.182584\n5.128186,-2.04864,8.250992\n5.169201,-
2.105222,8.332273\n5.248386,-2.237248,7.980954\n5.043911,-
2.114203,8.210127\n5.125192,-2.209705,8.106093\n5.2177,-
2.179019,8.123457\n5.220843,-2.211351,8.102201\n5.325326,-
2.220033,8.19426\n5.182074,-2.249971,7.970924\n5.262906,-
2.150129,8.065228\n5.162465,-2.081721,8.239167\n4.432582,-
1.535956,8.317753\n4.499044,-1.981729,8.622669',standing
?,'-6.72297,-3.135979,6.406528\n-6.402037,-3.009492,6.572233\n-
6.739136,-2.821782,6.465954\n-6.70366,-2.752027,6.666537\n-6.46775,-
2.952161,6.889723\n-6.639892,-2.819537,6.590345\n-6.678362,-
2.827171,6.651119\n-6.661897,-2.803221,6.608757\n-6.567742,-
2.989583,6.659951\n-6.614296,-2.96234,6.742279\n-6.635252,-
2.955304,6.664591\n-6.525979,-3.066373,6.692583\n-6.662645,-
2.81774,6.740633\n-6.712342,-2.821333,6.724017\n-6.76608,-
2.718796,6.609805\n-6.547834,-2.895279,6.637198\n-6.524333,-
2.956502,6.825357\n-6.580616,-2.942281,6.665339\n-6.521788,-
2.92357,6.712192\n-6.616092,-2.847828,6.676416\n-6.622079,-
2.854265,6.553223\n-6.670279,-3.02476,6.738088\n-6.556516,-
3.406466,6.586453\n-6.704258,-3.102897,6.694529',standing
?,'-1.641187,-7.671547,7.094946\n-1.637595,-7.620353,5.497767\n-
1.587299,-7.790849,5.920638\n-1.616488,-7.667505,5.686076\n-1.610351,-
7.745343,5.99668\n-1.609603,-7.775132,5.663173\n-1.727258,-
7.661368,5.812862\n-1.653013,-7.790999,5.776488\n-1.664838,-
7.795639,5.785319\n-1.701961,-7.683821,5.650749\n-1.635649,-
7.754474,5.601951\n-1.490002,-7.89144,5.609136\n-1.647175,-
7.802824,5.677993\n-1.772614,-7.726782,5.639972\n-1.619782,-
7.931556,5.517975\n-1.872307,-7.828121,5.827981\n-1.887275,-
7.749684,5.704338\n-1.742227,-7.839947,5.724246\n-1.863774,-
7.728878,5.700446',standing
?,'-1.641187,-7.671547,7.094946\n-1.637595,-7.620353,5.497767\n-
1.587299,-7.790849,5.920638\n-1.616488,-7.667505,5.686076\n-1.610351,-
7.745343,5.99668\n-1.609603,-7.775132,5.663173\n-1.727258,-
7.661368,5.812862\n-1.653013,-7.790999,5.776488\n-1.664838,-
7.795639,5.785319\n-1.701961,-7.683821,5.650749\n-1.635649,-
7.754474,5.601951\n-1.490002,-7.89144,5.609136\n-1.647175,-
7.802824,5.677993\n-1.772614,-7.726782,5.639972\n-1.619782,-
7.931556,5.517975\n-1.872307,-7.828121,5.827981\n-1.887275,-
7.749684,5.704338\n-1.742227,-7.839947,5.724246\n-1.863774,-
7.728878,5.700446',standing
?,'-9.548643,-0.945584,1.500779\n-9.787996,-0.868344,1.71693\n-
9.845027,-0.888702,1.419498\n-9.611962,-1.120121,1.174607\n-9.647288,-
0.936453,1.297202\n-9.62723,-0.953068,1.176254\n-9.716445,-
1.01459,1.481469\n-9.530831,-1.172512,1.338217\n-9.764944,-
0.913401,1.30783\n-9.762998,-0.897384,1.335672\n-9.716744,-
0.99558,1.466051\n-9.737551,-0.972528,1.306783\n-9.653875,-
1.048869,1.173709\n-9.748179,-0.898282,1.255888\n-9.695638,-

0.909209,1.27445\n-9.738149,-0.869392,1.119971\n-9.591305,-0.99049,0.943039\n-9.722881,-1.063838,1.167871\n-9.662407,-1.423839,2.040108\n-9.56481,-1.214125,0.534538',standing
?,'0.169298,7.512578,6.562503\n0.12978,7.550149,6.28498\n-0.025746,7.501201,6.79976\n0.078138,7.494315,6.310427\n0.08173,7.572603,5.97258\n-0.000898,7.533833,6.581813\n-0.040566,7.528445,6.460565\n-0.013023,7.593858,6.446494\n0.040865,7.515422,6.141878\n-0.093106,7.666907,6.100863\n-0.124691,7.687713,6.440507\n-0.210013,7.855814,6.272856\n-0.061522,7.622898,6.154751\n0.001946,7.526199,6.207891\n0.104782,7.392527,6.217022\n0.194595,7.805518,5.876629\n0.375868,7.864047,5.745203\n0.549956,7.77603,5.639972\n0.450114,7.863298,5.720804\n0.310604,7.797735,5.713319',standing
?,'-8.595575,-2.442621,4.119733\n-8.193661,-3.277435,4.284241\n-8.149054,-3.51065,4.372707\n-8.27105,-3.444337,4.256548\n-8.228389,-3.303181,4.221671\n-8.248298,-3.218008,4.412823\n-8.16058,-3.30348,4.071833\n-8.242011,-3.454965,4.181554\n-8.289612,-3.229384,4.222868\n-8.191566,-3.430117,4.089047\n-8.243657,-3.342699,4.111201\n-8.235873,-3.381468,4.148922\n-8.270002,-3.368296,4.210594\n-8.159083,-3.41919,4.192182\n-8.233778,-3.225792,4.42869\n-8.417596,-3.022365,4.527485\n-8.292755,-3.014881,4.405788\n-8.27584,-2.942281,4.50578',standing
?,'1.135988,0.125589,9.514964\n1.114732,-0.37018,10.830129\n1.140329,-0.856818,9.295969\n0.628992,-0.478705,9.02623\n0.629591,-0.985401,8.362061\n0.403112,0.492775,8.889564\n-0.996478,-1.189427,13.289215\n0.66836,-0.48589,11.817176\n1.01459,-0.622705,9.537117\n0.592917,-1.010399,8.561895\n-0.240101,-0.409099,9.561966\n0.051044,-1.478925,8.219557\n1.065484,-0.418081,10.208471\n1.333727,-0.708177,9.219029\n0.419278,-1.238225,9.097332\n-0.645907,-0.268242,10.20443\n1.14946,-1.047522,9.900262\n0.854124,-1.309028,9.857002',walking
?,'1.191822,-1.396596,8.293953\n0.583636,0.08652,9.125623\n0.844843,-1.68789,11.181747\n1.227897,-0.916544,11.172018\n0.782573,-0.278421,9.693692\n1.190923,-1.041983,8.420889\n0.508792,-0.898432,10.275981\n0.849184,-1.433868,8.267009\n1.197659,-1.556912,12.141552\n1.271306,-1.427731,10.879825\n1.326691,-1.438658,10.191557\n0.91834,-0.639171,8.719518\n0.222138,-0.909958,9.700727\n0.693957,-0.727487,8.647218\n0.377066,-1.0098,10.317744',walking
?,'-8.283774,-3.196004,4.305796\n-8.095465,-2.773433,2.874323\n-9.725276,-0.517474,1.347348\n-8.032446,-1.534609,3.955225\n-10.943742,-3.129243,3.700006\n-9.418564,-2.664459,2.704276\n-7.700587,-3.323239,1.92814\n-7.387737,-1.303489,2.295477\n-8.134534,-2.419568,0.738864\n-10.518327,-2.121838,0.893492\n-8.471184,-1.708248,1.466051\n-13.235776,-2.92866,1.696273\n-10.041269,-1.968706,2.817142\n-8.366402,-2.875071,1.699566\n-8.599916,-2.109114,2.520159\n-7.631879,-2.482737,2.338288\n-9.035211,-2.944377,3.6066\n-9.316776,-2.025887,3.073708',walking
?,'-8.394693,-3.435057,2.848876\n-7.234456,-3.709286,1.868714\n-

11.722273,-4.349805,3.82784\n-9.538165,-3.921096,3.236719\n-7.596403,-4.385131,2.303111\n-7.9585,-2.378853,2.987487\n-8.587792,-2.399959,3.596421\n-8.587492,-3.830834,2.532134\n-10.674752,-3.598816,3.449726\n-9.211694,-5.398673,2.391128\n-8.694071,-3.528313,1.945504\n-7.560478,-3.964356,1.474135\n-8.481812,-2.919379,2.591261\n-8.380323,-3.503614,0.731529\n-12.866793,-5.218149,2.46223\n-9.121133,-3.866909,3.492837\n-8.381371,-3.248694,3.113376\n-9.458381,-2.534529,3.915408',walking
?,'-2.651137,-9.937535,4.842729\n-2.112856,-9.189391,3.588787\n-1.700464,-8.069719,2.237397\n-2.787653,-8.875344,2.668351\n-1.203198,-8.678054,4.115991\n-1.64478,-10.232422,3.764971\n-1.658401,-8.961564,2.531386\n-0.88466,-8.094418,2.032623\n-2.098486,-7.722291,2.891687\n-1.370251,-9.344019,2.965334\n-0.629591,-8.758886,5.503605\n-2.622397,-11.522888,3.221002\n-0.704435,-8.966354,2.262545\n-2.20222,-8.654403,2.29877\n-0.987796,-8.746162,1.619632\n-1.098266,-8.803194,3.14496\n-0.548909,-10.519824,2.500849\n-0.383802,-8.555908,2.352957\n-1.722767,-8.99719,4.21733\n-0.713566,-8.71832,2.826871\n-2.077081,-11.128159,2.994074\n-0.92328,-8.760383,3.891009\n-1.237477,-8.261171,3.413951\n-1.736689,-9.064101,3.057392\n-1.224454,-9.118588,3.667523\n-0.614772,-9.051078,3.604355\n-1.822011,-8.440648,4.227209',walking
?,'11.312875,1.594035,4.046685\n9.991423,0.01916,3.987408\n8.07421,1.685794,2.542014\n9.035361,-0.436642,3.13523\n9.960886,1.373843,1.88039\n12.453203,1.856589,3.060685\n10.628348,0.739313,2.729424\n8.540789,1.898801,1.801354\n9.155711,1.066532,2.202819\n9.040001,1.535058,2.068997\n9.376052,2.80352,1.653013\n11.382181,0.917442,3.561843\n8.020022,2.442471,3.010839\n9.303154,0.485441,3.506159\n9.087453,1.706451,4.262386',walking
?,'1.387914,10.01672,0.85113\n0.915646,11.413166,0.959056\n0.107926,9.316027,1.879941\n1.082699,9.260942,0.937351\n1.836082,10.072554,2.728825\n1.188678,10.193503,2.809657\n-0.166454,10.419682,2.172432\n0.917143,8.919801,2.286795\n0.473914,8.610694,1.078058\n0.93301,10.642269,2.63063\n2.104025,10.650802,1.721121\n-0.356858,9.988579,1.663191\n1.929488,9.043444,2.212848\n1.349144,8.972342,0.187859\n1.322051,9.358988,2.184857\n1.478925,10.40172,3.018024\n-1.325793,9.147777,2.299668\n0.530048,8.990604,3.796256\n-0.451611,9.486074,1.602717\n0.37048,10.179731,1.852697',walking
?,'0.621508,8.966953,3.425028\n-0.416284,8.609946,1.124162\n-0.541723,10.007739,0.786165\n0.101489,9.063053,2.15447\n0.343685,8.854836,0.291294\n0.863854,10.322384,2.187102\n1.172212,9.075777,0.938548\n1.772614,12.991933,1.533711\n0.611628,10.315948,1.76019\n-0.629591,9.094338,0.593067\n0.693059,8.50202,1.426683\n0.608335,9.675879,1.681304\n-0.530048,11.163785,3.237468\n-0.515977,10.032587,2.727628\n-0.324226,9.157207,1.078507\n0.515827,9.012159,1.776655',walking
?,'-8.677305,-1.64942,3.346591\n-9.98304,-2.328258,2.656526\n-

11.258089,-0.774939,3.138972\n-11.240276,-0.308209,2.675985\n-9.026978,-2.127376,2.545157\n-7.912097,-0.577649,1.242865\n-8.074359,-1.496438,0.12499\n-10.112671,-1.478775,-0.179327\n-10.453811,0.021256,1.617237\n-9.156459,-2.941982,3.197351\n-9.495504,-0.757425,2.563868\n-9.007369,-1.09138,1.965413\n-9.556876,-2.105821,1.548829\n-8.701555,-1.897155,3.098706',walking
?,'-11.33907,0.445623,5.004693\n-9.690099,-2.100133,3.734135\n-8.907527,-0.975671,2.37062\n-7.864196,-2.009421,1.494342\n-8.553064,-0.797991,2.339335\n-10.113868,-1.863325,3.293152\n-9.26663,-1.398093,2.0196\n-12.569661,-2.945874,2.999762\n-9.099428,-2.836302,2.646497\n-8.500224,-0.724044,4.370761\n-8.17974,-1.262774,2.498754\n-9.480685,-0.659379,3.186424\n-7.858209,-2.649191,0.779429\n-11.710747,-2.65952,1.922752',walking
?,'0.819097,-0.867596,7.60284\n1.593736,-1.878144,14.999259\n2.576143,-2.109713,4.984634\n2.730023,-3.523972,8.327333\n1.462159,-2.139501,11.195519\n1.518592,-0.546663,3.756888\n1.866319,-0.594564,7.544312',running
?,'0.839903,-1.443448,11.49774\n1.689088,-1.084794,6.030659\n1.156196,-2.020049,13.881084\n1.118923,-0.786315,6.488258\n0.999472,-0.188009,6.549031\n3.801046,-2.648592,10.250983\n1.150957,-1.16877,9.350156',running
?,'0.969534,-1.392704,7.70373\n2.539469,-1.211581,11.471844\n1.390159,-0.185764,9.551039\n0.632734,-2.836601,4.455934\n-1.670077,-4.38573,22.006487\n0.958457,-2.400259,4.435277\n1.137634,-5.023254,5.643115\n0.716261,-3.446732,16.221617',running
?,'-1.093626,-4.398004,3.321593\n-4.70322,-9.031319,12.117601\n-0.412392,-4.157604,7.385791\n-1.709445,-4.080365,5.728587\n-2.844086,-6.625822,13.747562\n-0.266296,-5.048252,9.053473\n-0.76461,-4.095633,4.809499\n-1.479523,-4.027525,8.725805\n-0.984503,-3.993845,11.415112\n-0.635429,-5.122049,6.569838\n-2.198478,-4.273164,11.770773\n-1.328188,-4.957391,10.150991',running
?,'-0.016615,-4.68376,5.92932\n-2.411785,-3.591781,7.133266\n-5.140909,-6.875502,9.678723\n0.299976,-5.111121,5.543722\n-0.184566,-4.890929,3.056494\n-2.687661,-6.440806,8.81457\n-1.990561,-7.389234,12.007131\n-0.87523,-3.675008,5.424869\n-0.8655,-3.599864,9.749825',running
?,'-1.64942,-3.17385,7.155121\n-0.880918,-1.799558,6.681955\n-5.63099,-7.811207,11.674523\n-2.589914,-4.445605,3.752696\n-0.524659,-2.77448,5.946384\n-7.924371,-7.21799,14.31563\n-1.29855,-3.159779,5.45331\n-1.599873,-1.355132,5.556745\n-3.054997,-7.369325,11.310929\n-2.688858,-1.971251,10.566078',running
?,'-0.611029,-0.618963,6.067034\n-4.035907,-6.65082,7.111412\n-5.667814,-3.76991,6.119125\n-0.239951,-2.107617,4.054319\n-10.647658,-7.689809,8.527317\n-5.752687,-5.105882,8.617729\n0.113464,-1.906585,5.423671\n-9.039552,-8.466993,14.222823\n-0.775388,-5.131329,5.172643',running
?,'-1.573079,-2.984194,3.696862\n-5.79954,-5.98231,6.300548\n-2.449506,-6.926995,4.664899\n-7.239396,-9.995315,9.124426\n-5.249134,-5.523514,9.071286\n-1.229543,-3.21232,5.329667',running

```
?,'-4.186644,-8.220306,10.40651\n-1.940116,-4.969366,5.418133\n-
3.801345,-4.309538,6.562503\n-6.435867,-9.353599,7.478149\n-5.114864,-
5.572612,3.916007\n-3.029251,-4.378395,9.493708\n-5.37637,-
8.113278,8.99749\n-3.045118,-5.364993,7.144792',running
?,'7.48743,4.69828,7.051237\n3.575914,2.674039,4.144881\
n12.489578,5.536986,12.904515\n3.924838,2.505639,6.148464\
n4.305347,3.102299,7.979008\n6.595734,3.666625,11.361823\n-
0.241149,1.691033,3.72186\n7.857161,5.656138,12.665911\
n4.654122,1.377585,8.64617',running
?,'2.832859,0.442031,9.29522\n4.666845,1.804048,11.62752\
n2.057172,0.942889,4.695735\n6.869814,2.089355,6.885681\
n8.81472,5.42906,8.615783\n1.419049,2.776576,2.516417\
n4.951254,3.776796,9.490115\n3.147804,3.792364,9.379196\
n2.315984,1.315764,4.214486',running
?,'1.055755,1.62547,5.677843\n4.124523,3.886368,7.078929\
n8.183033,1.760639,12.550052\n0.622106,1.992656,4.603976\
n6.514902,4.588558,8.197403\n6.098768,2.373764,11.048375\
n3.092868,3.284021,5.205575\n3.716921,1.306483,6.505771',running
?,'1.183888,1.703009,4.262686\n10.038874,4.267176,12.061618\
n3.763474,2.785108,5.585784\n4.054768,3.211422,4.718338\
n8.289163,4.915927,12.990287\n4.034859,3.45257,8.098459\
n1.495241,1.683699,2.222428',running
?,'0.89424,3.04886,4.672533\n8.546927,4.307592,12.346775\
n4.201164,1.428479,6.825806\n4.593198,2.695145,7.550299\
n4.473597,3.733835,5.774542\n9.590407,6.032306,3.708089\
n12.535083,6.576125,8.747659\n1.515748,3.849245,3.331472\
n7.422615,3.475173,8.825348\n6.301147,3.38521,9.357192',running
?,'4.184099,2.930456,6.067483\n1.970203,3.261568,6.074219\
n6.863078,6.515351,7.589368\n1.905537,2.61102,2.658621\
n14.496155,5.665868,8.681347\n1.920806,2.502346,3.171605\
n7.500004,4.136797,9.133557\n7.11246,1.705104,11.961476\
n4.277954,2.06765,8.298593',running
?,'4.646039,1.477877,15.343094\n3.551814,1.637295,5.666167\
n2.755919,3.082839,1.911824\n13.585898,4.316274,10.211914\
n1.647773,1.769919,3.641927\n7.700437,6.022426,9.590407\
n5.172194,2.511627,11.731105\n1.56769,1.447041,6.964717',running
?,'0.883014,2.852019,4.448749\n5.920937,-0.21136,14.627133\
n1.025966,1.594185,8.54528\n0.869242,2.79873,6.136938\
n12.073593,4.971911,12.56562\n-0.058079,1.485511,3.700455\
n10.174642,5.331463,8.277936\n8.599318,5.241051,9.186247\
n1.091081,2.158661,5.273384\n3.239264,2.58228,7.644004',running
```

#2.3 What's in an Attribute?

```
df = pd.read_csv("data/weather.numeric.csv")
df

      outlook  temperature  humidity  windy play
0       sunny           85        85  False   no
```

```
1       sunny               80          90   True    no
2    overcast               83          86   False   yes
3       rainy               70          96   False   yes
4       rainy               68          80   False   yes
5       rainy               65          70   True    no
6    overcast               64          65   True    yes
7       sunny               72          95   False   no
8       sunny               69          70   False   yes
9       rainy               75          80   False   yes
10      sunny               75          70   True    yes
11   overcast               72          90   True    yes
12   overcast               81          75   False   yes
13      rainy               71          91   True    no
```

```python
# All unique values of some feature
df["outlook"].unique()
```

```
<StringArray>
['sunny', 'overcast', 'rainy']
Length: 3, dtype: string
```

```python
# Counts of all unique values of some feature
df[ ["outlook"] ].value_counts()
```

```
outlook
rainy        5
sunny        5
overcast     4
dtype: int64
```

```python
# Visualizing counts
df[ ["outlook"] ].value_counts().plot.bar()
```

```
<Axes: xlabel='outlook'>
```
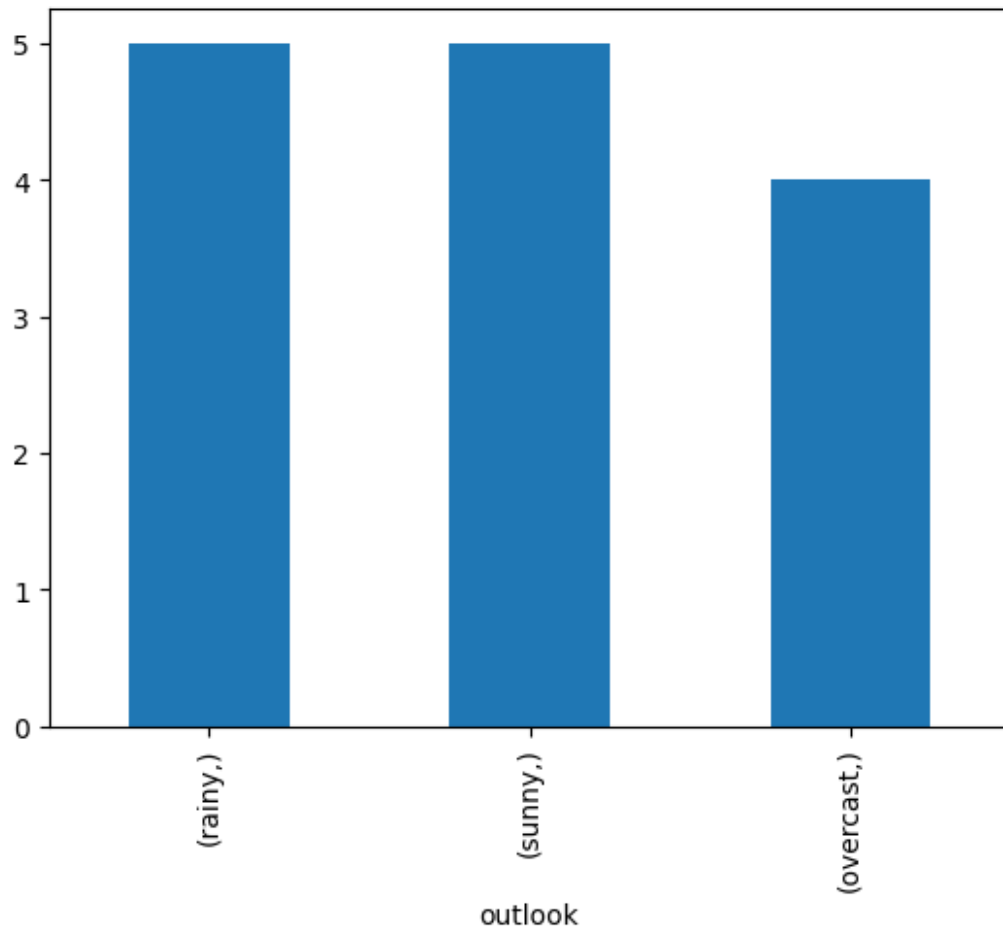
```
# Histogram
df[ ["temperature", "humidity"] ].plot.hist(alpha=0.75)

<Axes: ylabel='Frequency'>
```
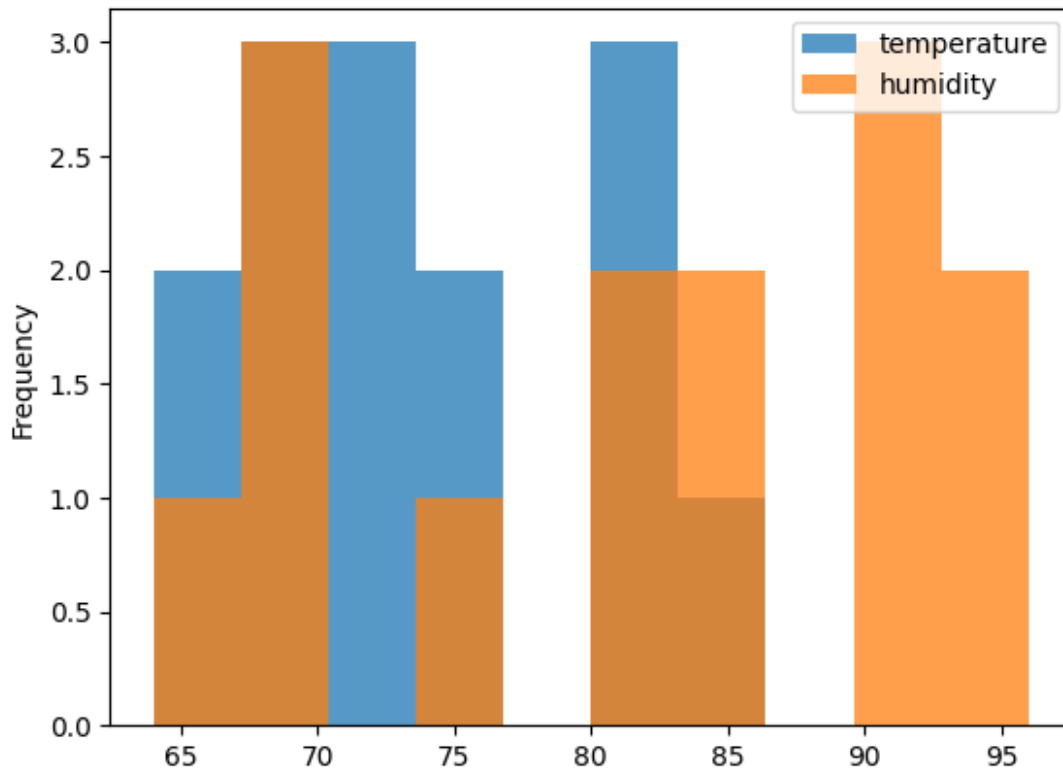
Task 2.3.1 Visualize other features using the same methods

Challenge 2.3.2 How can we visualize the correlation between more than 2 variables?

2.4 Preparing the Input

Gathering the Data Together

Create dataset manually

```python
from weka.core.dataset import Attribute, Instance, Instances

# create attributes
num_att = Attribute.create_numeric("num")
date_att = Attribute.create_date("dat", "yyyy-MM-dd")
nom_att = Attribute.create_nominal("nom", ["label1", "label2"])

# create dataset
dataset = Instances.create_instances("helloworld", [num_att, date_att,
nom_att], 0)

# add rows
values = [3.1415926, date_att.parse_date("2014-04-10"), 1.0]
inst = Instance.create_instance(values)
```

```
dataset.add_instance(inst)

values = [2.71828, date_att.parse_date("2014-08-09"),
Instance.missing_value()]
inst = Instance.create_instance(values)
dataset.add_instance(inst)

print(dataset)

@relation helloworld

@attribute num numeric
@attribute dat date yyyy-MM-dd
@attribute nom {label1,label2}

@data
3.141593,2014-04-10,label2
2.71828,2014-08-09,?
```

Create dataset from lists

```
from weka.core.dataset import create_instances_from_lists
from random import randint

# pure numeric
x = [[randint(1, 10) for _ in range(5)] for _ in range(10)]
y = [randint(0, 1) for _ in range(10)]
dataset = create_instances_from_lists(x, y, name="generated from
lists")
print(dataset)

@relation 'generated from lists'

@attribute x1 numeric
@attribute x2 numeric
@attribute x3 numeric
@attribute x4 numeric
@attribute x5 numeric
@attribute y numeric

@data
10,4,9,6,7,1
4,1,5,5,4,1
5,5,9,3,1,0
1,10,10,8,4,1
8,10,7,4,6,1
8,2,8,5,8,1
7,6,3,10,2,1
10,10,9,3,1,0
```

```
2,4,6,3,3,1
4,4,10,3,2,0

dataset = create_instances_from_lists(x, name="generated from lists
(no y)")
print(dataset)

@relation 'generated from lists (no y)'

@attribute x1 numeric
@attribute x2 numeric
@attribute x3 numeric
@attribute x4 numeric
@attribute x5 numeric

@data
10,4,9,6,7
4,1,5,5,4
5,5,9,3,1
1,10,10,8,4
8,10,7,4,6
8,2,8,5,8
7,6,3,10,2
10,10,9,3,1
2,4,6,3,3
4,4,10,3,2

# mixed data types
x = [["TEXT", 1, 1.1], ["XXX", 2, 2.2]]
y = ["A", "B"]
dataset = create_instances_from_lists(x, y, name="generated from mixed
lists", cols_x=["text", "integer", "float"], col_y="class")
print(dataset)

@relation 'generated from mixed lists'

@attribute text string
@attribute integer numeric
@attribute float numeric
@attribute class string

@data
TEXT,1,1.1,A
XXX,2,2.2,B
```

Create dataset from matrices

```
from weka.core.dataset import create_instances_from_matrices
import numpy as np
```

```python
# pure numeric
x = np.random.randn(10, 5)
y = np.random.randn(10)
dataset = create_instances_from_matrices(x, y, name="generated from matrices")
print(dataset)

@relation 'generated from matrices'

@attribute x1 numeric
@attribute x2 numeric
@attribute x3 numeric
@attribute x4 numeric
@attribute x5 numeric
@attribute y numeric

@data
0.629392,-0.09832,-0.422484,-0.169946,0.05149,-0.692909
-0.541271,1.009759,1.631953,3.368145,1.303312,0.776066
-0.179628,0.003113,1.915605,-2.529908,-0.943525,-0.352515
-1.013789,-0.943359,-0.129905,0.696045,0.975933,0.538208
0.818717,0.894472,1.063312,1.547528,1.483321,-1.277839
0.751582,0.009669,-0.596318,1.299259,0.524294,2.876909
-0.286807,-1.020932,-0.351447,-1.774938,-1.131307,0.553186
-2.112942,-0.15905,0.647894,-2.249435,-0.046986,0.596745
-0.308123,-1.472949,0.364102,-1.161712,0.585897,1.148289
-1.156549,0.491881,-0.832956,-1.214837,0.791667,-1.662772

dataset = create_instances_from_matrices(x, name="generated from matrix (no y)")
print(dataset)

@relation 'generated from matrix (no y)'

@attribute x1 numeric
@attribute x2 numeric
@attribute x3 numeric
@attribute x4 numeric
@attribute x5 numeric

@data
0.629392,-0.09832,-0.422484,-0.169946,0.05149
-0.541271,1.009759,1.631953,3.368145,1.303312
-0.179628,0.003113,1.915605,-2.529908,-0.943525
-1.013789,-0.943359,-0.129905,0.696045,0.975933
0.818717,0.894472,1.063312,1.547528,1.483321
0.751582,0.009669,-0.596318,1.299259,0.524294
-0.286807,-1.020932,-0.351447,-1.774938,-1.131307
-2.112942,-0.15905,0.647894,-2.249435,-0.046986
```

```
-0.308123,-1.472949,0.364102,-1.161712,0.585897
-1.156549,0.491881,-0.832956,-1.214837,0.791667

# mixed data types
x = np.array([("TEXT", 1, 1.1), ("XXX", 2, 2.2)], dtype='S20, i4, f8')
y = np.array(["A", "B"], dtype='S20')
dataset = create_instances_from_matrices(x, y, name="generated from
mixed matrices", cols_x=["text", "integer", "float"], col_y="class")
print(dataset)

@relation 'generated from mixed matrices'

@attribute text string
@attribute integer numeric
@attribute float numeric
@attribute class string

@data
TEXT,1,1.1,A
XXX,2,2.2,B
```

Task 2.4.1 Create new datasets with different attribute types and rows entries, and display the dataset schema.

Attribute Types

```
df = pd.read_csv("data/weather.numeric.csv")
df

      outlook  temperature  humidity  windy play
0       sunny           85        85  False   no
1       sunny           80        90   True   no
2    overcast           83        86  False  yes
3       rainy           70        96  False  yes
4       rainy           68        80  False  yes
5       rainy           65        70   True   no
6    overcast           64        65   True  yes
7       sunny           72        95  False   no
8       sunny           69        70  False  yes
9       rainy           75        80  False  yes
10      sunny           75        70   True  yes
11   overcast           72        90   True  yes
12   overcast           81        75  False  yes
13      rainy           71        91   True   no

# Show data types
df.dtypes

outlook        object
temperature     int64
```

```
humidity          int64
windy              bool
play             object
dtype: object

# Modify outlook column to be a string data type
df["outlook"] = df["outlook"].astype("string")
```

Task 2.3.1 Modify *play* column data type to be boolean.

```
dataset = pd.read_csv(os.path.join(data_dir, 'labor.csv'))
display(dataset)

    duration wage-increase-first-year wage-increase-second-year  \
0          1                        5                         ?
1          2                      4.5                       5.8
2          ?                        ?                         ?
3          3                      3.7                         4
4          3                      4.5                       4.5
5          2                        2                       2.5
6          3                        4                         5
7          3                      6.9                       4.8
8          2                        3                         7
9          1                      5.7                         ?
10         3                      3.5                         4
11         2                      6.4                       6.4
12         2                      3.5                         4
13         3                      3.5                         4
14         1                        3                         ?
15         2                      4.5                         4
16         1                      2.8                         ?
17         1                      2.1                         ?
18         1                        2                         ?
19         2                        4                         5
20         2                      4.3                       4.4
21         2                      2.5                         3
22         3                      3.5                         4
23         2                      4.5                         4
24         1                        6                         ?
25         3                        2                         2
26         2                      4.5                       4.5
27         2                        3                         3
28         2                        5                         4
29         3                        2                       2.5
30         3                      4.5                       4.5
31         3                        3                         2
32         2                      2.5                       2.5
33         2                        4                         5
34         3                        2                       2.5
```

35	2	2	2
36	1	2	?
37	1	2.8	?
38	3	2	2.5
39	2	4.5	4
40	1	4	?
41	2	2	3
42	2	2.5	2.5
43	2	2.5	3
44	2	4	4
45	2	4.5	4
46	2	4.5	4
47	2	4.6	4.6
48	2	5	4.5
49	2	5.7	4.5
50	2	7	5.3
51	3	2	3
52	3	3.5	4
53	3	4	3.5
54	3	5	4.4
55	3	5	5
56	3	6	6

	wage-increase-third-year	cost-of-living-adjustment	working-hours \
0	?	?	40
1	?	?	35
2	?	?	38
3	5	tc	?
4	5	?	40
5	?	?	35
6	5	tc	?
7	2.3	?	40
8	?	?	38
9	?	none	40
10	4.6	none	36
11	?	?	38
12	?	none	40
13	5.1	tcf	37
14	?	none	36
15	?	none	37
16	?	?	35
17	?	tc	40
18	?	none	38
19	?	tcf	35
20	?	?	38
21	?	?	40
22	4.6	tcf	27
23	?	?	40
24	?	?	38

25	2	none	40
26	?	tcf	?
27	?	none	33
28	?	none	37
29	?	?	35
30	5	none	40
31	2.5	tc	40
32	?	?	38
33	?	none	40
34	2.1	tc	40
35	?	none	40
36	?	tc	40
37	?	none	38
38	2	?	37
39	?	none	40
40	?	none	?
41	?	none	38
42	?	tc	39
43	?	tcf	40
44	?	none	40
45	?	?	40
46	?	none	40
47	?	tcf	38
48	?	none	38
49	?	none	40
50	?	?	?
51	?	tcf	?
52	4.5	tcf	35
53	?	none	40
54	?	none	38
55	5	?	40
56	4	?	35

	pension	standby-pay	shift-differential	education-allowance	\
0	?	?	2	?	
1	ret_allw	?	?	yes	
2	empl_contr	?	5	?	
3	?	?	?	yes	
4	?	?	?	?	
5	?	?	6	yes	
6	empl_contr	?	?	?	
7	?	?	3	?	
8	?	12	25	yes	
9	empl_contr	?	4	?	
10	?	?	3	?	
11	?	?	4	?	
12	?	?	2	no	
13	?	?	4	?	
14	?	?	10	no	

15	empl_contr	?	?	?
16	?	?	2	?
17	ret_allw	2	3	no
18	none	?	?	yes
19	?	13	5	?
20	?	?	4	?
21	none	?	?	?
22	?	?	?	?
23	?	?	4	?
24	?	8	3	?
25	none	?	?	?
26	?	?	?	yes
27	?	?	?	yes
28	?	?	5	no
29	none	?	?	?
30	?	?	?	no
31	none	?	5	no
32	empl_contr	?	?	?
33	none	?	3	no
34	none	2	1	no
35	none	?	?	no
36	ret_allw	4	0	no
37	empl_contr	2	3	no
38	empl_contr	?	?	?
39	?	?	4	?
40	none	?	?	yes
41	empl_contr	?	?	yes
42	empl_contr	?	?	?
43	none	?	?	?
44	none	?	3	?
45	?	?	2	no
46	?	?	5	?
47	?	?	?	?
48	?	14	5	?
49	ret_allw	?	?	?
50	?	?	?	?
51	empl_contr	?	?	yes
52	?	?	?	?
53	empl_contr	?	6	?
54	empl_contr	10	6	?
55	?	?	?	?
56	?	?	14	?

	statutory-holidays	vacation	longterm-disability-assistance	\
0	11	average	?	
1	11	below_average	?	
2	11	generous	yes	
3	?	?	?	
4	12	average	?	

5	12	average	?
6	12	generous	yes
7	12	below_average	?
8	11	below_average	yes
9	11	generous	yes
10	13	generous	?
11	15	?	?
12	10	below_average	no
13	13	generous	?
14	11	generous	?
15	11	average	?
16	12	below_average	?
17	9	below_average	yes
18	11	average	no
19	15	generous	?
20	12	generous	?
21	11	below_average	?
22	?	?	?
23	10	generous	?
24	9	generous	?
25	10	below_average	?
26	10	below_average	yes
27	12	generous	?
28	11	below_average	yes
29	10	average	?
30	11	average	?
31	10	below_average	yes
32	10	average	?
33	10	below_average	no
34	10	below_average	no
35	11	average	yes
36	11	generous	no
37	9	below_average	yes
38	10	average	?
39	12	average	yes
40	11	average	no
41	12	generous	yes
42	12	average	?
43	11	below_average	?
44	10	below_average	no
45	10	below_average	no
46	11	average	?
47	?	?	yes
48	11	below_average	yes
49	11	average	yes
50	11	?	yes
51	?	?	yes
52	13	generous	?
53	11	average	yes

		generous		yes

54	11	generous	yes
55	12	average	?
56	9	generous	yes

	contribution-to-dental-plan	bereavement-assistance	\
0	?	yes	
1	full	?	
2	half	yes	
3	?	yes	
4	half	yes	
5	?	?	
6	none	yes	
7	?	?	
8	half	yes	
9	full	?	
10	?	yes	
11	full	?	
12	half	?	
13	full	yes	
14	?	?	
15	full	yes	
16	?	?	
17	half	?	
18	none	no	
19	?	?	
20	full	?	
21	?	?	
22	?	?	
23	half	?	
24	?	?	
25	half	yes	
26	none	?	
27	?	yes	
28	full	yes	
29	?	yes	
30	half	?	
31	half	yes	
32	?	?	
33	none	?	
34	half	yes	
35	none	yes	
36	none	no	
37	half	?	
38	?	yes	
39	full	yes	
40	none	no	
41	none	yes	
42	?	yes	
43	?	yes	

44	none	?
45	half	?
46	full	yes
47	half	?
48	?	?
49	full	yes
50	full	?
51	half	yes
52	?	yes
53	full	?
54	?	?
55	half	yes
56	full	yes

	contribution-to-health-plan	class
0	?	good
1	full	good
2	half	good
3	?	good
4	half	good
5	?	good
6	half	good
7	?	good
8	?	good
9	?	good
10	full	good
11	?	good
12	half	bad
13	full	good
14	?	good
15	?	good
16	?	good
17	none	bad
18	none	bad
19	?	good
20	full	good
21	?	bad
22	?	good
23	full	good
24	?	good
25	full	bad
26	half	good
27	full	good
28	full	good
29	full	bad
30	?	good
31	full	bad
32	?	bad
33	none	bad

```
34                         full    bad
35                         full    bad
36                         none    bad
37                         none    bad
38                         none    bad
39                         half   good
40                         none    bad
41                         full    bad
42                            ?    bad
43                            ?    bad
44                         none    bad
45                         half    bad
46                         full   good
47                         half   good
48                         full   good
49                         full   good
50                            ?   good
51                            ?   good
52                         full   good
53                         full   good
54                         full   good
55                         half   good
56                         full   good
```

Sparse Dataframe & Missing Values

```python
df = pd.read_csv("data/supermarket.csv")

# replace each "?" entry with Null
df = df.applymap(lambda x: np.NaN if x == "?" else x)

df
```

```
     department1 department2 department3 department4 department5
department6  \
0            NaN         NaN         NaN         NaN         NaN
NaN
1              t         NaN         NaN         NaN         NaN
NaN
2            NaN         NaN         NaN         NaN         NaN
NaN
3              t         NaN         NaN         NaN         NaN
NaN
4            NaN         NaN         NaN         NaN         NaN
NaN
...          ...         ...         ...         ...         ...
...
4622         NaN         NaN         NaN         NaN         NaN
NaN
```

```
4623        NaN          NaN          NaN            t          NaN
NaN
4624        NaN          NaN          NaN          NaN          NaN
NaN
4625        NaN          NaN          NaN          NaN          NaN
NaN
4626          t          NaN          NaN          NaN          NaN
NaN

      department7  department8 department9 'grocery misc'  ...
department208  \
0             NaN          NaN          NaN            NaN  ...
NaN
1             NaN          NaN          NaN            NaN  ...
NaN
2             NaN          NaN          NaN            NaN  ...
NaN
3             NaN          NaN          NaN            NaN  ...
NaN
4             NaN          NaN          NaN            NaN  ...
NaN
...           ...          ...          ...            ...  ...
...
4622          NaN          NaN          NaN            NaN  ...
NaN
4623          NaN          NaN          NaN            NaN  ...
NaN
4624          NaN          NaN          NaN            NaN  ...
NaN
4625          NaN          NaN          NaN            NaN  ...
NaN
4626          NaN          NaN          NaN            NaN  ...
NaN

      department209 department210 department211   department212
department213  \
0             NaN          NaN          NaN           NaN
NaN
1             NaN          NaN          NaN           NaN
NaN
2             NaN          NaN          NaN           NaN
NaN
3             NaN          NaN          NaN           NaN
NaN
4             NaN          NaN          NaN           NaN
NaN
...           ...          ...          ...           ...
...
4622          NaN          NaN          NaN           NaN
```

```
NaN
4623                NaN            NaN            NaN            NaN
NaN
4624                NaN            NaN            NaN            NaN
NaN
4625                NaN            NaN            NaN            NaN
NaN
4626                NaN            NaN            NaN            NaN
NaN

      department214 department215 department216 total
0                NaN            NaN            NaN  high
1                NaN            NaN            NaN   low
2                NaN            NaN            NaN   low
3                NaN            NaN            NaN   low
4                NaN            NaN            NaN   low
...              ...            ...            ...   ...
4622             NaN            NaN            NaN   low
4623             NaN            NaN            NaN  high
4624             NaN            NaN            NaN   low
4625             NaN            NaN            NaN   low
4626             NaN            NaN            NaN  high

[4627 rows x 217 columns]
```

Task 2.4.2 Give two special observations in the dataframe above.

Answer

```python
# Map each column to a new column of total True and False counts
df.isnull().apply(
    lambda col: col.value_counts()
)
```
```
        department1  department2  department3  department4  department5
\
False          1047          131           90           84          175

True           3580         4496         4537         4543         4452


        department6  department7  department8  department9  'grocery
misc'  \
False             2           67          NaN           82
178
True           4625         4560       4627.0         4545
4449

        ...  department208  department209  department210  department211
\
```

```
False    ...             NaN             NaN             191             207

True     ...          4627.0          4627.0            4436            4420


        department212  department213  department214  department215  \
False              38             22            NaN            NaN
True             4589           4605         4627.0         4627.0

        department216    total
False             NaN   4627.0
True           4627.0      NaN

[2 rows x 217 columns]
```

Task 2.4.3 Compute a new column counting the total sum along all *department* columns.

Task 2.4.4 Visualize the total sum to confirm your intuitive guess of in *task 2.4.2*.

Challenge 2.4.5 Explore methods of dimesionality reduction and its role in learning from data.
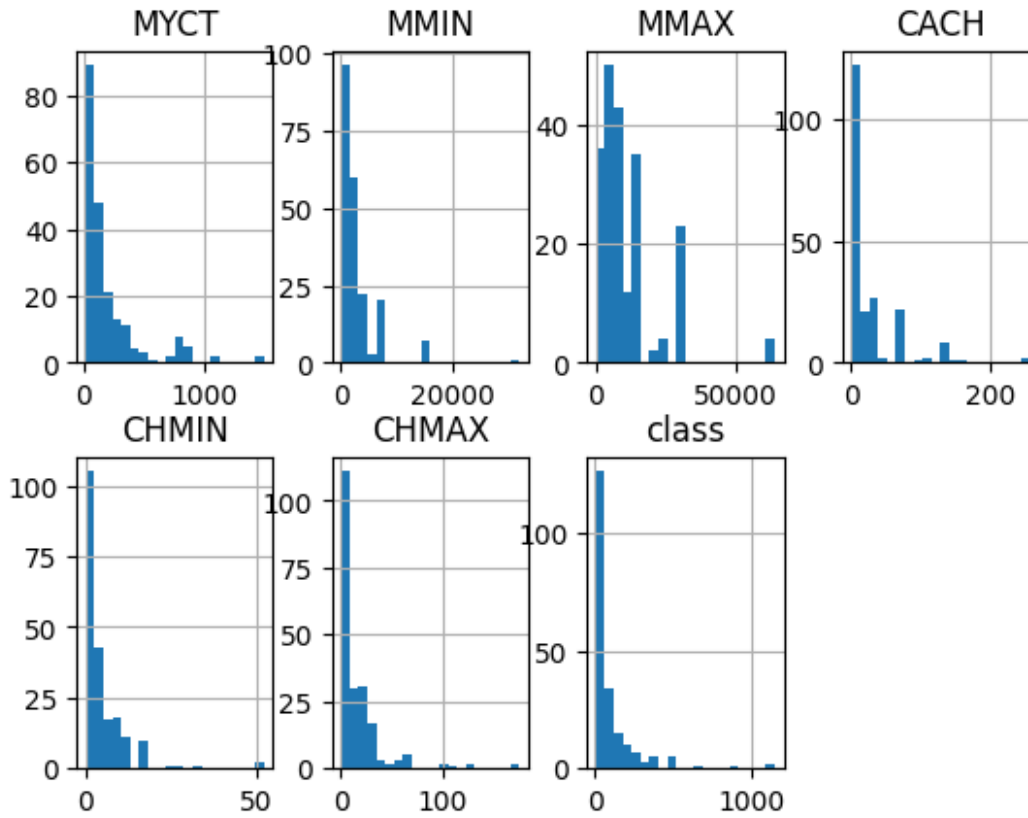
##Inaccurate Values

```
dataset = pd.read_csv(os.path.join(data_dir, 'cpu.csv'))
dataset.hist(layout=(2,4), bins=20)

array([[<Axes: title={'center': 'MYCT'}>,
        <Axes: title={'center': 'MMIN'}>,
        <Axes: title={'center': 'MMAX'}>,
        <Axes: title={'center': 'CACH'}>],
       [<Axes: title={'center': 'CHMIN'}>,
        <Axes: title={'center': 'CHMAX'}>,
        <Axes: title={'center': 'class'}>, <Axes: >]], dtype=object)
```
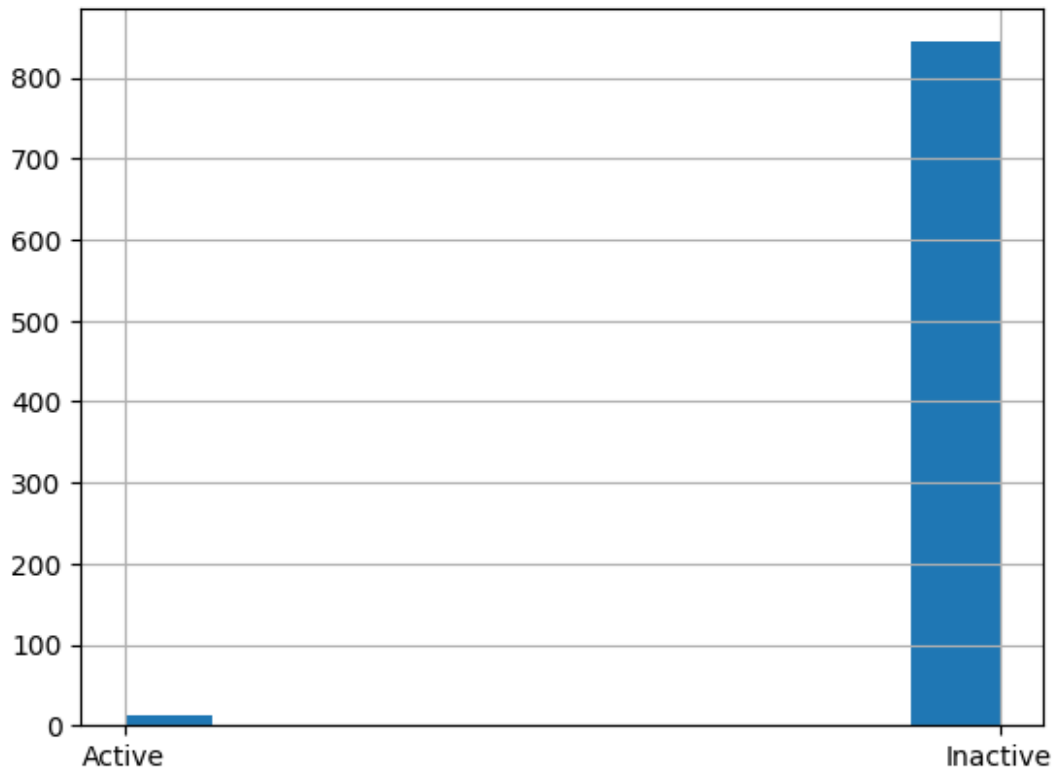
Unbalanced Data

```
df = pd.read_csv("data/unbalanced.csv")
df.iloc[:, -1].hist()
<Axes: >
```

Challenge 2.4.6 Explore why is imbalanced data an issue, and how to overcome it by imposing a penalty.

Challenge 2.4.7 The imbalance was shown on *active-inactive* class. What if a similar graph was found on a feature. Do we consider the data imbalanced in that case? Why?

Project. Phase 2

- Data exploration
- Data cleansing
- Data engineering