Data Exploration

{codenation}®

Learning Objectives

- To generate exploratory reports
- To be able to interpret reports

Reporting

Reporting

We explored our data using methods like .describe()

This helped us work out the data types we were working with, if there were any missing values, any outliers, and gave us our descriptive statistics.

We can use these to get a good overall picture.

Reporting

As a result of your exploring, you might need to change your workflow.

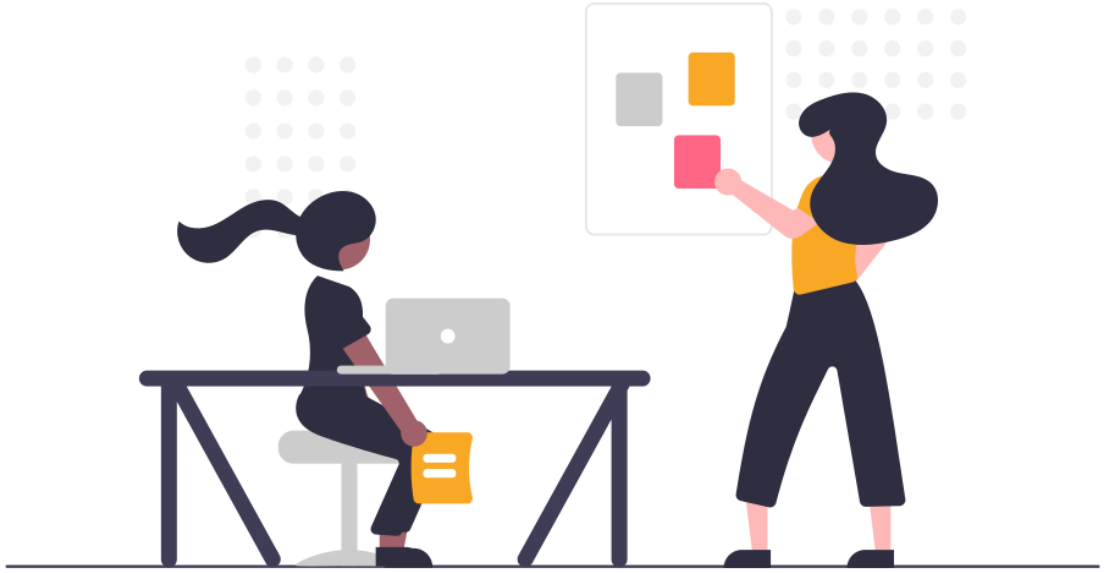
You might need to request more data, more time, or more resources.

You might need to explain why!

Reporting

A **report** is a more readable and accessible way of presenting your findings.

Most people aren't used to reading figures in a dark terminal!



Reporting

Exploring your data and then writing the report can be time consuming.

There are libraries we can use to streamline the process!

As long as we have the data in a dataframe, the **sweetviz** library can report on it for us!

Sweetviz

Sweetviz

sweetviz is a tool that creates a thorough, visually accessible summary of a data set with a few lines of code.

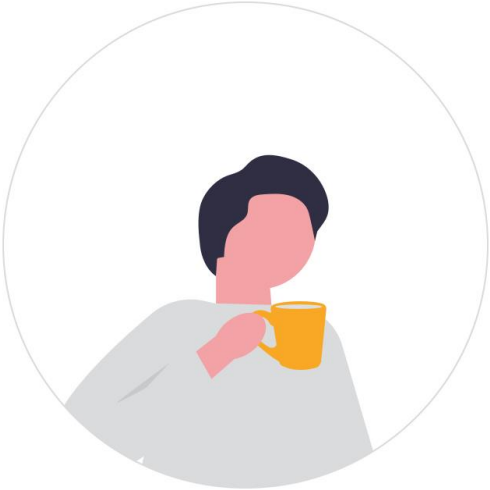
Using a library means your data reports will be consistent, and high quality.

Sweetviz

```
pip install sweetviz
pip install --upgrade setuptools
```

sweetviz is a big library!

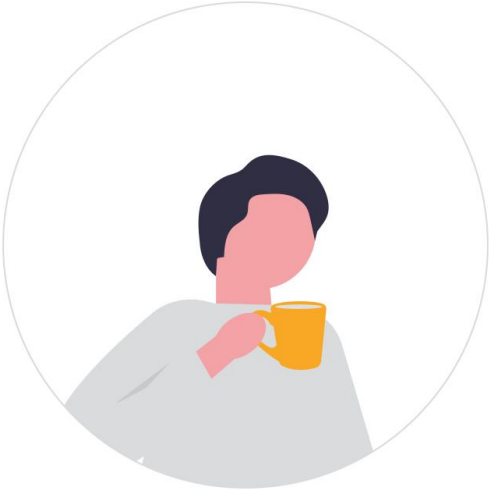
It could take a while to install.



Sweetviz

```
pip install sweetviz
pip install --upgrade setuptools
```

setuptools will help pip manage the packages sweetviz needs.



Sweetviz

We need to take three steps with **sweetviz**:

1. Import the data as a frame
2. Turn the frame to profile
3. Turn the profile into a web page

To turn our terminal stats into this…

SweetViz 2.3.1

Get updates, docs & report issues here

Created & maintained by Francois Bertrand
Graphic design by Jean-Francois Hains

DataFrame

NO COMPARISON TARGET

45315	ROWS
0	DUPLICATES
16.9 MB	RAM
9	FEATURES
1	CATEGORICAL
2	NUMERICAL
6	TEXT

ASSOCIATIONS

DataFrame

1 date

VALUES:	45,315 (100%)
MISSING:	---
DISTINCT:	15,649 (35%)

66	<1%	2012-02-29
64	<1%	2016-03-29
60	<1%	2008-03-26
59	<1%	2014-03-05
56	<1%	2023-11-21
56	<1%	2012-11-14
55	<1%	2022-03-29
44,899	>99%	(Other)

2 home_team

VALUES:	45,315 (100%)
MISSING:	---
DISTINCT:	313 (<1%)

600	1%	Brazil
580	1%	Argentina
567	1%	Mexico
533	1%	Germany
530	1%	England
514	1%	Sweden
510	1%	France
41,481	92%	(Other)

3 away_team

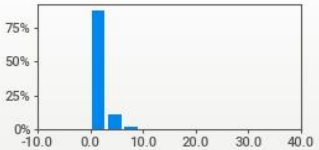
VALUES:	45,315 (100%)
MISSING:	---
DISTINCT:	308 (<1%)

565	1%	Uruguay
551	1%	Sweden
529	1%	England
495	1%	Hungary
478	1%	Paraguay
464	1%	Germany
452	<1%	Argentina
41,781	92%	(Other)

4 home_score

VALUES:	45,315 (100%)
MISSING:	---
DISTINCT:	26 (<1%)
ZEROES:	10,959 (24%)

MAX	31.0	RANGE	31.0
95%	5.0	IQR	1.00
Q3	2.0	STD	1.75
AVG	1.7	VAR	3.05
MEDIAN	1.0		
Q1	1.0	KURT.	11.5
5%	0.0	SKEW	2.19
MIN	0.0	SUM	78,817



5 away_score

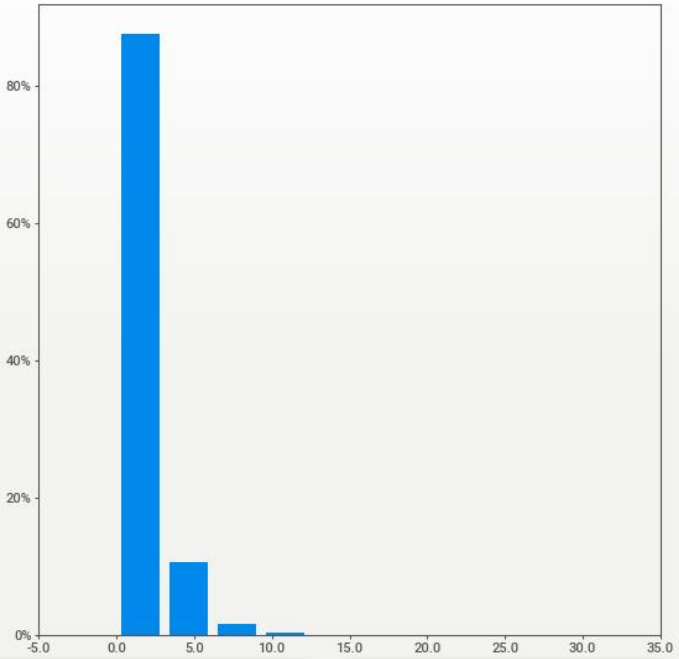
VALUES:	45,315 (100%)
MISSING:	---
DISTINCT:	22 (<1%)

MAX	21.0	RANGE	21.0
95%	4.0	IQR	2.00
Q3	2.0	STD	1.39
AVG	1.2	VAR	1.94
MEDIAN	1.0		

home_score

MISSING: ---

Auto 5 15 30



NUMERICAL ASSOCIATIONS

(PEARSON, -1 to 1)

away_score	-0.14

CATEGORICAL ASSOCIATIONS

(CORRELATION RATIO, 0 to 1)

neutral	0.04

MOST FREQUENT VALUES

1	13,429	29.6%
0	10,959	24.2%
2	9,941	21.9%
3	5,305	11.7%
4	2,758	6.1%
5	1,311	2.9%
6	711	1.6%
7	383	0.8%
8	207	0.5%
9	126	0.3%
10	66	0.1%
11	36	<0.1%
12	27	<0.1%
13	16	<0.1%
14	12	<0.1%

SMALLEST VALUES

0	10,959	24.2%
1	13,429	29.6%
2	9,941	21.9%
3	5,305	11.7%
4	2,758	6.1%
5	1,311	2.9%
6	711	1.6%
7	383	0.8%
8	207	0.5%
9	126	0.3%
10	66	0.1%
11	36	<0.1%
12	27	<0.1%
13	16	<0.1%
14	12	<0.1%

LARGEST VALUES

31	1	<0.1%
30	1	<0.1%
24	1	<0.1%
22	1	<0.1%
21	2	<0.1%
20	1	<0.1%
19	3	<0.1%
18	1	<0.1%
17	3	<0.1%
16	6	<0.1%
15	8	<0.1%
14	12	<0.1%
13	16	<0.1%
12	27	<0.1%
11	36	<0.1%

Sweetviz

Create a new file called **reporting.py**

Import the necessary libraries and classes.

```
import pandas as pd
import sweetviz as sv
```

Sweetviz

Create a dataframe from the **results.csv** file.

```
df = pd.read_csv('results.csv')
my_report = sv.analyze(df)
my_report.show_html()
```

Sweetviz

Create a dataframe from the **results.csv** file.

```
df = pd.read_csv('results.csv')
my_report = sv.analyze(df)
my_report.show_html()
```

Sweetviz

Use Sweetviz's **analyze()** function to analyse the dataframe, and save those results to **my_report**

```
df = pd.read_csv('results.csv')
my_report = sv.analyze(df)
my_report.show_html()
```

Sweetviz

The show_html() method will export the my_report object to an HTML file called: "SWEETVIZ_REPORT.html"

```
df = pd.read_csv('results.csv')
my_report = sv.analyze(df)
my_report.show_html()
```

Navigating the Report

Navigating the report

The overview section gives us information similar to df.info()

It tells us the shape of the frame, how many duplicates there are, and the data types of the columns.

Navigating the report

Each column (variable) becomes a tab.

Each tab overviews the contents – how many values are there? How many are missing? What are the distinct percentages?

Navigating the report

The tabs show a snapshot of the data, typically sorted from most frequent – least frequent.

Clicking the tabs gives a much larger list!

Navigating the report

For numerical data types, sweetviz will construct visualisations to help us do quick comparisons.

It will also show lots of the data from df.describe() like the five-figure summary.

Navigating the report

The associations section shows the strength of the relationship between the variables.

It measures how much the difference between the values can be explained by each other.

There's no relationship between results – a team scoring a goal doesn't suggest the opponent will score a goal.

It also doesn't suggest if a team scores a goal, the other team will score -1 goal.

Review

Review

Which method did you prefer?

Why?

Which method do you think would be more useful to a non-analyst?

Learning Objectives

- To generate exploratory reports
- To be able to interpret reports

Activity 1

Produce exploratory reports for the other two data sets given.

What information can you see from them?

We will explore visualisations and what they mean more next week!