

# Home work 11

Matti Kuikka 24.11.2023

NOTE: This Home work contains EXTRA tasks at the end where you are able to earn additional 5 points!

- Other Data Management skills you learned (where you can earn max 3 extra points)
- Data Management skills not covered in this course, but should have covered (for Introduction to Data Engineering AND/OR Introduction to AI, where you can earn extra 2 points)

## 1. Reading data with Pandas, data exploration and cleaning

Reading data with Pandas

- How to use Pandas to read data from different sources and formats, such as CSV, Excel, JSON, and HTML files.

Data exploration

- How to inspect and explore the data using various Pandas methods and attributes.

Data cleaning

- How to clean and modify the data using various Pandas methods and attributes.

```
In [ ]: # Add Python libraries you need for data reading and exploration  
# - use import statements
```

```
import pandas as pd  
import numpy as np  
import matplotlib.pyplot as plt  
import seaborn as sb  
import requests  
import json  
import csv
```

```
In [ ]:
```

```
In [ ]: # Add in here example how you read csv data and explore the contents
```

```
In [ ]: csv_data = pd.read_csv('my_spotify.csv')  
csv_data.head(2)
```

	Track name	Artist name	Album	Playlist name	Type	ISRC
0	MISA MISA!	CORPSE	MISA MISA!	MISA MISA! - CORPSE, Scarlxrd, Kordhell	Playlist	QZNWR2293683
1	Famous	Phonk Mane	Famous	MISA MISA! - CORPSE, Scarlxrd, Kordhell	Playlist	CAHQJ2235882

```
In [ ]:
```

```
In [ ]: # Add in here example how you read JSON data and explore the contents  
# - cover both files containing JSON and if pure JSON objects are read  
  
json_data = pd.read_json('patients.json')  
json_data.head(2)
```

```
Out[ ]:   Name  Gender  Nationality  Age  
0    John     Male        UK      10  
1    Nick     Male    French      25
```

```
In [ ]:
```

```
In [ ]: # Add in here examples how you clean data  
# Data can be the CSV you have already read or it could be also Excel or HTML format data
```

```
In [ ]: csv_data.dropna(inplace=True)  
csv_data.drop_duplicates(inplace=True)  
csv_data.head(2)
```

```
Out[ ]:   Track name  Artist name  Album  Playlist name  Type  ISRC  
0  MISA MISA!  CORPSE  MISA MISA!  MISA MISA! - CORPSE, Scarlxrd, Kordhell  Playlist  QZNWR2293683  
1    Famous  Phonk Mane  Famous  MISA MISA! - CORPSE, Scarlxrd, Kordhell  Playlist  CAHQJ2235882
```

## 2. Data formats

Converting data from one format to another using Pandas.

- you write the dataframes into different formats, such as CSV, Excel, JSON, and HTML files.

```
In [ ]: # Add in here example how you convert data from JSON to CSV
```

```
In [ ]: csv_file_path = 'my_spotify.csv'  
json_file_path = 'my_spotify.json'  
  
csv_data = []  
with open(csv_file_path, 'r') as csv_file:  
    csv_reader = csv.DictReader(csv_file)  
    for row in csv_reader:  
        csv_data.append(row)  
  
with open(json_file_path, 'w') as json_file:  
    json.dump(csv_data, json_file, indent=2)
```

```
In [ ]: # json_data.to_csv('patients.csv')  
# csv_data.to_json('patients.json')  
  
converted_json_data = pd.read_csv('patients.csv')  
converted_json_data.head(2)
```

```
out[ ]:   Unnamed: 0  Name  Gender  Nationality  Age
```

0	0	John	Male	UK	10
1	1	Nick	Male	French	25

```
In [ ]: converted_csv_data = pd.read_json('my_spotify.json')
converted_csv_data.head(2)
```

```
out[ ]:   Track name  Artist name  Album  Playlist name  Type  ISRC
0  MISA MISA!  CORPSE  MISA MISA!  MISA MISA! - CORPSE, Scarlxrd, Kordhell  Playlist  QZNWR2293683
1  Famous  Phonk Mane  Famous  MISA MISA! - CORPSE, Scarlxrd, Kordhell  Playlist  CAHQJ2235882
```

```
In [ ]: # Add in here example how you can convert Pandas DataFrame column to other format
# - to numbers
# - to strings

csv_data = pd.read_csv('patients.csv')

csv_data['Age'] = csv_data['Age'].astype('int64') # Convert to int64
csv_data['Age'].head(2)
```

```
out[ ]: 0    10
1    25
Name: Age, dtype: int64
```

```
In [ ]: # How you convert data incluing ',' as decimal sign to '.' as decimal sign? (Example: 3,2)

csv_data['Age'] = csv_data['Age'].astype('str') # Convert to string
csv_data['Age'] = csv_data['Age'].str.replace(',', '.') # Replace ',' with '.'
csv_data['Age'].head(2)
```

```
out[ ]: 0    10
1    25
Name: Age, dtype: object
```

```
In [ ]:
```

### 3. Data visualization

You can use Pandas, Matplotlib, and Seaborn to create various types of data visualizations, such as line plots, bar plots, scatter plots, histograms, box plots, and heatmaps.

You can also use various parameters and functions to modify the appearance and style of the plots, such as the labels, the titles, the legends, the colors and the grids.

```
In [ ]: sb.get_dataset_names()
```

```
out[ ]: ['anagrams',
'anscombe',
'attention',
'brain_networks',
'car_crashes',
'diamonds',
'dots',
```

```
'exercise',
'flights',
'fmri',
'geyser',
'glue',
'healthexp',
'iris',
'mpg',
'penguins',
'planets',
'seaice',
'taxis',
'tips',
'titanic']
```

```
In [ ]: # Add Python libraries you need for data visualization
# - use import statements

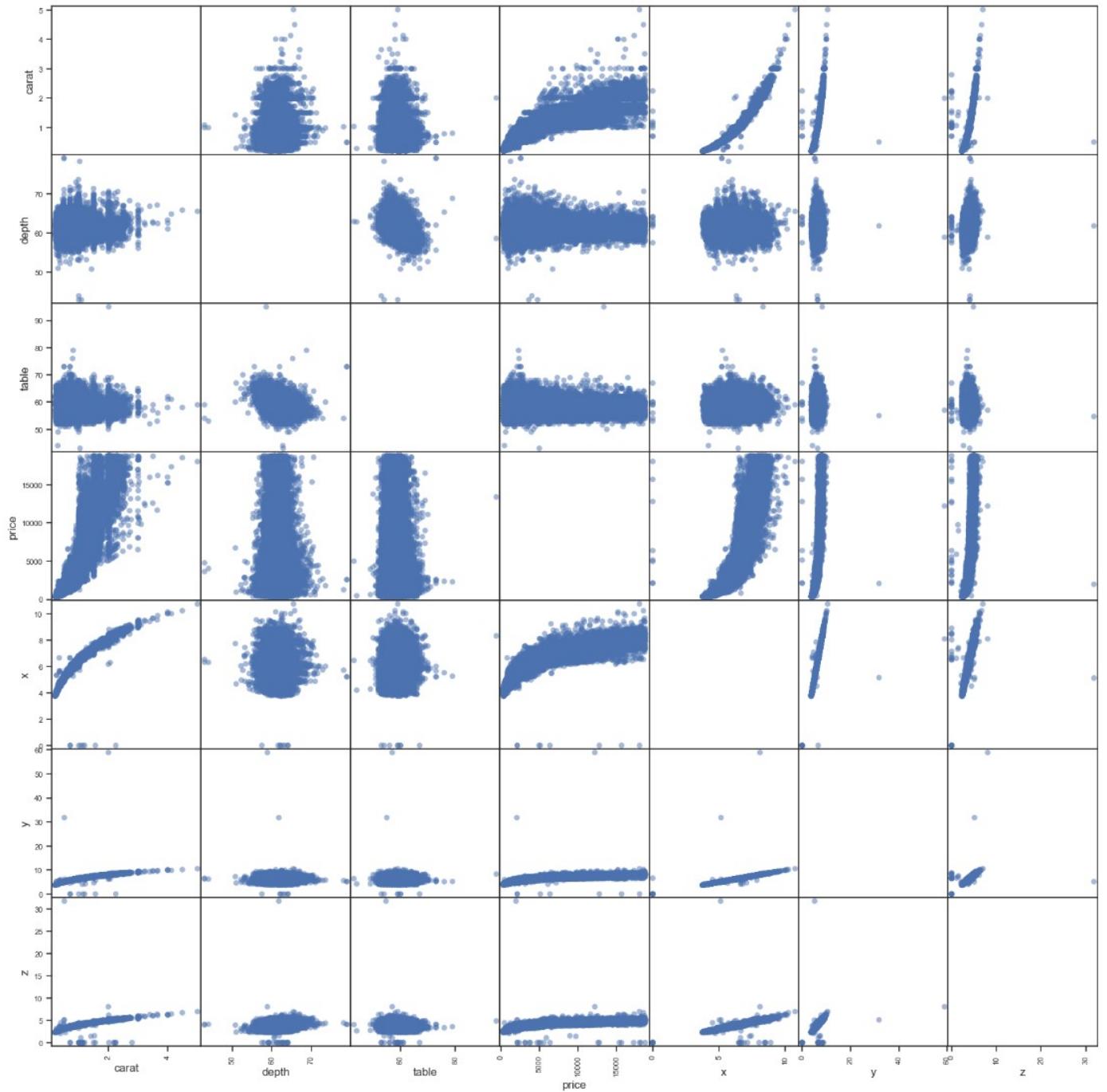
visual_data = sb.load_dataset('diamonds')
visual_data.head(2)
```

```
out[ ]:   carat      cut  color  clarity  depth  table  price     x     y     z
          0    0.23    Ideal      E     SI2    61.5    55.0    326  3.95  3.98  2.43
          1    0.21  Premium      E     SI1    59.8    61.0    326  3.89  3.84  2.31
```

```
In [ ]:
```

```
In [ ]: # Add in here example how you create graph (e.g. bar graph, boxplot or histogram) using Pandas library

pd.plotting.scatter_matrix(visual_data, figsize=(20,20), diagonal='kde', alpha=0.5, marker='o')
plt.show()
```



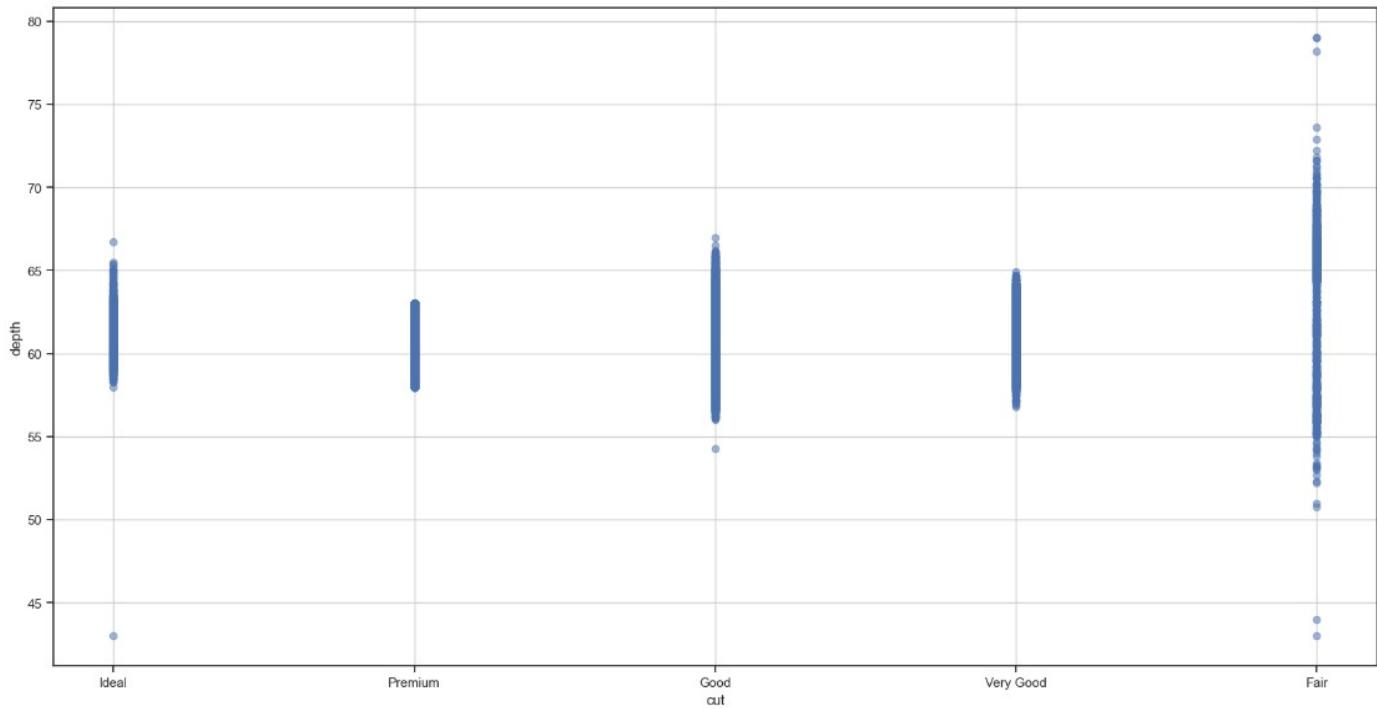
In [ ]:

In [ ]:

```
# Add in here example how you create graph (e.g. bar graph, boxplot or histogram) using Matplotlib
```

In [ ]:

```
plt.figure(figsize=(20,10))
plt.scatter(visual_data['cut'], visual_data['depth'], alpha=0.5, marker='o')
plt.xlabel('cut')
plt.ylabel('depth')
plt.grid(True)
plt.show()
```



```
In [ ]: # Add in here example how you create graph (e.g. scatter diagram, bar graph, boxplot or histogram)
```

```
In [ ]:
visual_data = sb.load_dataset('car_crashes')
visual_data.head(2)

sb.set(style='ticks')
sb.pairplot(visual_data, hue='abbrev', palette='husl')

plt.show()
```

c:\Python311\Lib\site-packages\seaborn\\_oldcore.py:1498: FutureWarning: is\_categorical\_dtype is deprecated and will be removed in a future version. Use isinstance(dtype, CategoricalDtype) instead  
if pd.api.types.is\_categorical\_dtype(vector):  
c:\Python311\Lib\site-packages\seaborn\\_oldcore.py:1498: FutureWarning: is\_categorical\_dtype is deprecated and will be removed in a future version. Use isinstance(dtype, CategoricalDtype) instead  
if pd.api.types.is\_categorical\_dtype(vector):  
c:\Python311\Lib\site-packages\seaborn\\_oldcore.py:1498: FutureWarning: is\_categorical\_dtype is deprecated and will be removed in a future version. Use isinstance(dtype, CategoricalDtype) instead  
if pd.api.types.is\_categorical\_dtype(vector):  
c:\Python311\Lib\site-packages\seaborn\\_oldcore.py:1498: FutureWarning: is\_categorical\_dtype is deprecated and will be removed in a future version. Use isinstance(dtype, CategoricalDtype) instead  
if pd.api.types.is\_categorical\_dtype(vector):  
c:\Python311\Lib\site-packages\seaborn\\_oldcore.py:1498: FutureWarning: is\_categorical\_dtype is deprecated and will be removed in a future version. Use isinstance(dtype, CategoricalDtype) instead  
if pd.api.types.is\_categorical\_dtype(vector):  
c:\Python311\Lib\site-packages\seaborn\\_oldcore.py:1498: FutureWarning: is\_categorical\_dtype is deprecated and will be removed in a future version. Use isinstance(dtype, CategoricalDtype) instead  
if pd.api.types.is\_categorical\_dtype(vector):  
c:\Python311\Lib\site-packages\seaborn\\_oldcore.py:1498: FutureWarning: is\_categorical\_dtype is deprecated and will be removed in a future version. Use isinstance(dtype, CategoricalDtype) instead  
if pd.api.types.is\_categorical\_dtype(vector):  
c:\Python311\Lib\site-packages\seaborn\\_oldcore.py:1498: FutureWarning: is\_categorical\_dtype is deprecated and will be removed in a future version. Use isinstance(dtype, CategoricalDtype) instead



















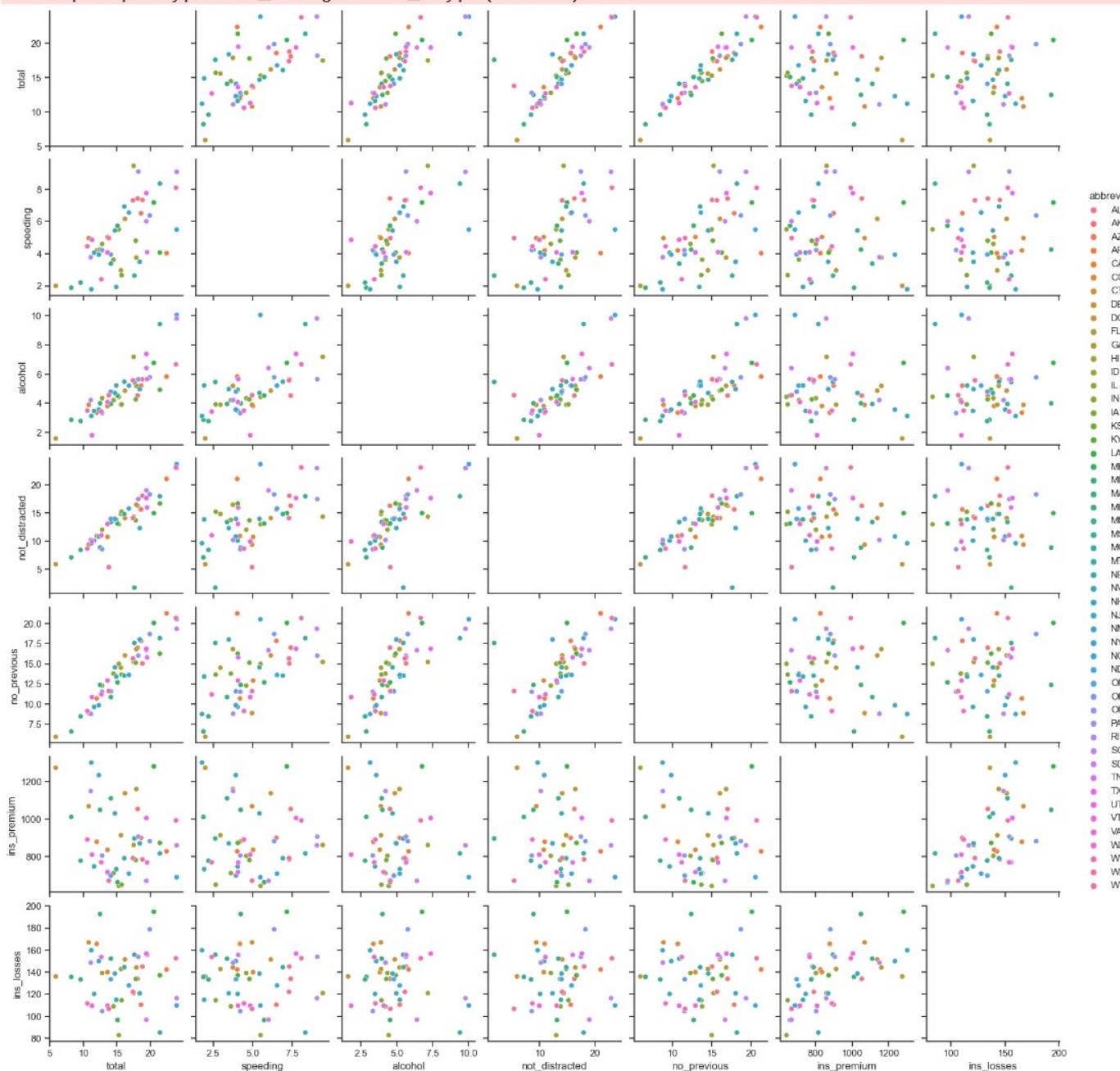




```

pe is deprecated and will be removed in a future version. Use isinstance(dtype, categorica
l Dtype) instead
    if pd.api.types.is_categorical_dtype(vector):
c:\Python311\Lib\site-packages\seaborn\_oldcore.py:1498: FutureWarning: is_categorical_dty
pe is deprecated and will be removed in a future version. Use isinstance(dtype, categorica
l Dtype) instead
    if pd.api.types.is_categorical_dtype(vector):
c:\Python311\Lib\site-packages\seaborn\_oldcore.py:1498: FutureWarning: is_categorical_dty
pe is deprecated and will be removed in a future version. Use isinstance(dtype, categorica
l Dtype) instead
    if pd.api.types.is_categorical_dtype(vector):
c:\Python311\Lib\site-packages\seaborn\_oldcore.py:1498: FutureWarning: is_categorical_dty
pe is deprecated and will be removed in a future version. Use isinstance(dtype, categorica
l Dtype) instead
    if pd.api.types.is_categorical_dtype(vector):
c:\Python311\Lib\site-packages\seaborn\_oldcore.py:1498: FutureWarning: is_categorical_dty
pe is deprecated and will be removed in a future version. Use isinstance(dtype, categorica
l Dtype) instead
    if pd.api.types.is_categorical_dtype(vector):

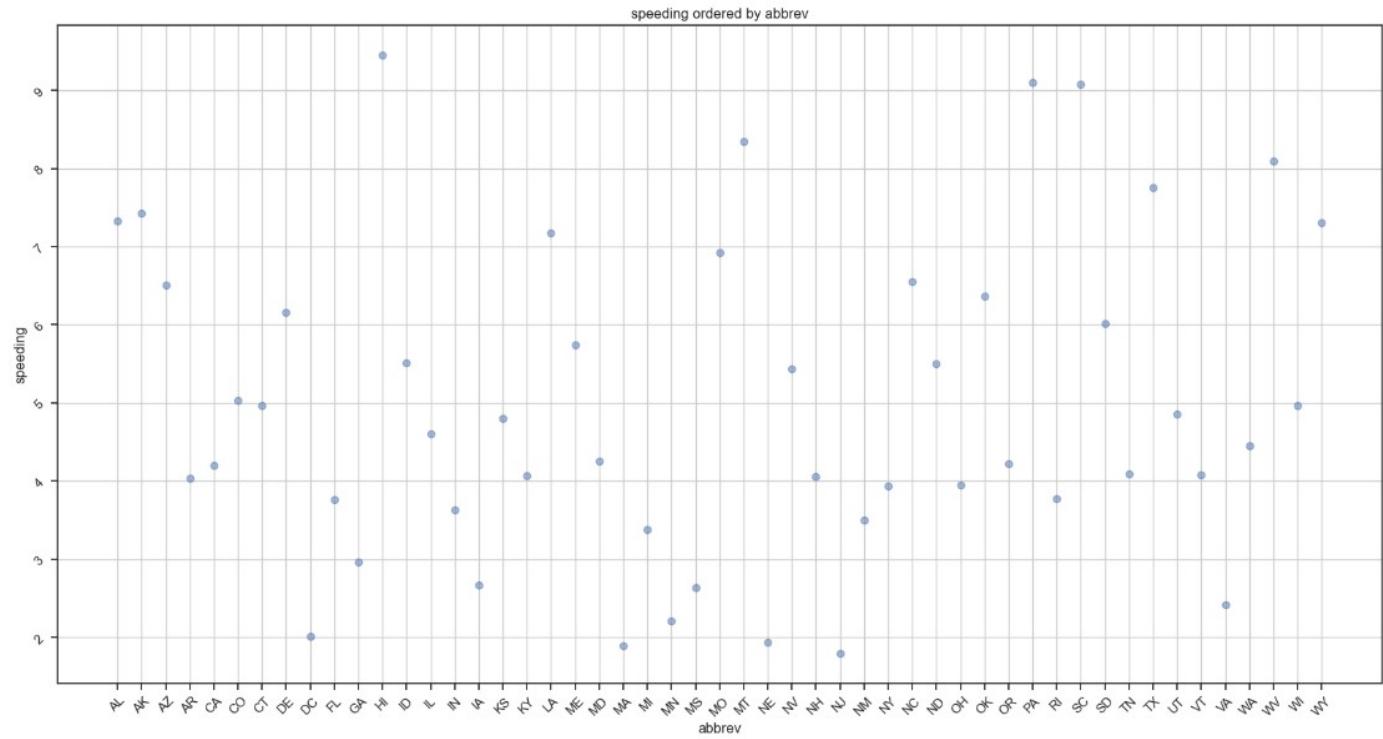
```



In [ ]:

```
# Add in here example how you change the layout of figures using Matplotlib
```

```
In [ ]:
plt.figure(figsize=(20,10))
plt.scatter(visual_data['abbrev'], visual_data['speeding'], alpha=0.5, marker='o')
plt.xlabel('abbrev')
plt.ylabel('speeding')
plt.grid(True)
plt.title('speeding ordered by abbrev')
plt.xticks(rotation=45)
plt.yticks(rotation=45)
plt.show()
```



## 4 Database reading and writing with Python

You use Pandas to read and write data from and to a relational database.

You use the SQLite database engine, which is a lightweight and self-contained database system that does not require any installation or configuration.

```
In [ ]:
# Add Python libraries you need for management of databases
# - use import statements
import sqlite3
```

```
In [ ]:
data = pd.read_csv('patients.csv')
data.head(2)
```

```
In [ ]:
conn = sqlite3.connect('test.sqlite3')
csv_data.to_sql('data', conn, if_exists='replace', index=False)
```

```
In [ ]:
# Add in here example how store data to SQLite database
```

```
In [ ]:
sql_data = pd.read_sql('SELECT * FROM data', conn)
sql_data.head(2)
```

```
In [ ]: # Add in here example how read data from SQLite database and present it
```

```
In [ ]:
```

```
In [ ]: # Add in here example how modify data in SQLite database
```

```
In [ ]: conn = sqlite3.connect('test.sqlite3')
cursor = conn.cursor()

cursor.execute('UPDATE data SET Age = 30 WHERE Age > 12')

conn.commit()
conn.close()
```

```
In [ ]: # Add in here example how read data using Pandas from SQL database
```

```
In [ ]: conn = sqlite3.connect('test.sqlite3')
sql_data = pd.read_sql('SELECT * FROM data', conn)
sql_data.head(3)
```

## 5 Linear algebra with Python

You will use Python and Numpy to perform basic linear algebra operations, such as creating and manipulating vectors and matrices, performing arithmetic operations, and solving linear systems. You can use the np.array function to create Numpy arrays, which are the main data structures for storing and manipulating numerical data in Numpy.

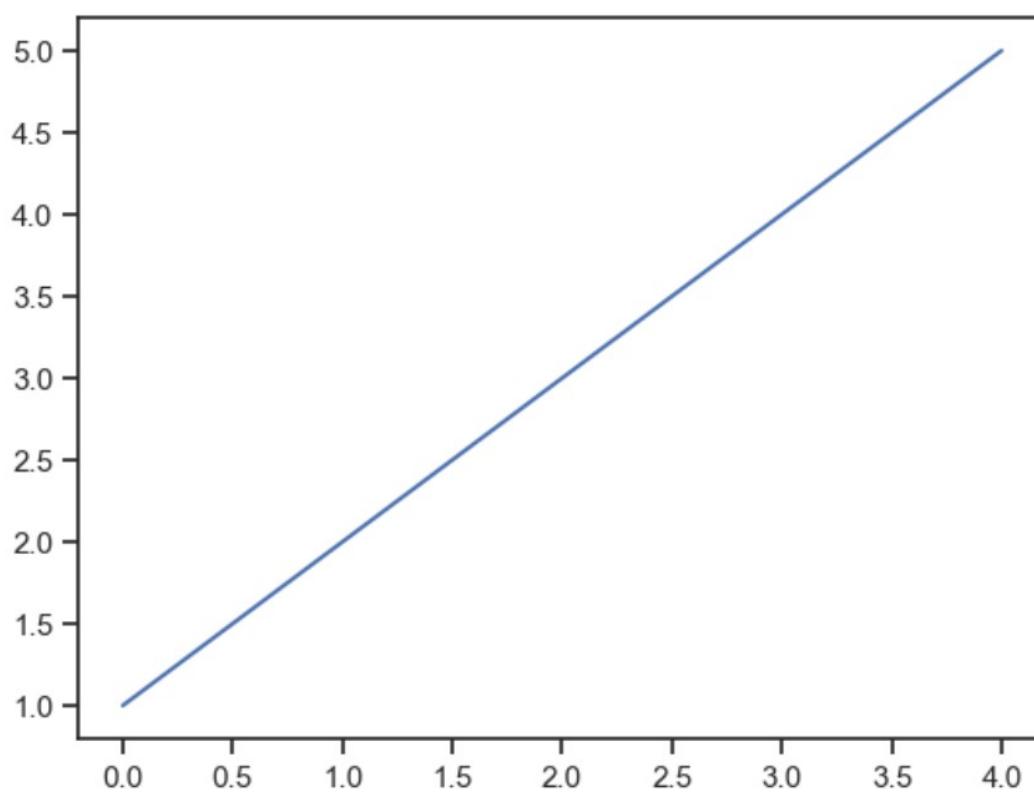
```
In [ ]: # Add Python libraries you need for linear algebra (matrices, vectors and linear models)
# - use import statements

from sklearn import linear_model
from sklearn.metrics import mean_squared_error, r2_score
```

```
In [ ]:
```

```
In [ ]: # Add in here examples how create vectors
# include examples where you create vector directly using numpy and from Pandas DataFrame
```

```
In [ ]: vector = np.array([1,2,3,4,5])
# display vector in a form of a graph
plt.plot(vector)
plt.figure(figsize=(5,5))
plt.show()
```



<Figure size 500x500 with 0 Axes>

In [ ]:

```
# Add in here examples how create matrices  
# include examples where you create matrix directly using numpy and from Pandas DataFrame
```

In [ ]:

```
matrix1 = np.array([[1,2,3,4,5],[6,7,8,9,10]])  
matrix2 = pd.DataFrame([[1,2,3,4,5],[6,7,8,9,10]])  
# display matrix in a form of a graph  
print("Matrix from numpy:",'\n', matrix1, '\n')  
print("Matrix from pd dataframe:",'\n', matrix2)
```

Matrix from numpy:  
[[ 1 2 3 4 5]  
 [ 6 7 8 9 10]]

Matrix from pd dataframe:  
0 1 2 3 4  
0 1 2 3 4 5  
1 6 7 8 9 10

In [ ]:

```
# Add in here example how you transpose matrice  
Transposed_matrix = matrix1.transpose()  
  
print("Transposed matrix1:",'\n', Transposed_matrix)
```

Transposed matrix1:  
[[ 1 6]  
 [ 2 7]  
 [ 3 8]  
 [ 4 9]  
 [ 5 10]]

In [ ]:

In [ ]:

In [ ]:

## EXTRA: Other Data Management skills you learned

You may earn 1 - 3 additional points from this final task if you:

- provide examples of other Data Management skills you learned during this course or with self-study
- provide short summary like in below for 5 tasks and then for each of them similarly examples

In [ ]:

### EXTRA 1

In [ ]:

In [ ]:

### EXTRA 2

In [ ]:

In [ ]:

### EXTRA 3

In [ ]:

## EXTRA: Data Management skills not covered in this course

You may earn 1 - 2 additional points from this final task if you:

- provide information what was missing in this course but what skills you should have for other introduction courses

Additional data management skills needed for Introduction to Data Engineering?

# Specify in here if any?

Additional data management skills needed for Introduction to AI?

# Specify in here if any?