De video waar ik mij op het gebaseerd om het project aan te vatten is: *Introduction to data visualization with Python.* Een cursus voorzien op PluralSight.

Om aan data visualisatie te doen wordt aangeraden om het *Anaconda* pakket te downloaden via https://www.anaconda.com/distribution/

In dit pakket zitten alle benodigdheden om met Python aan te slag te gaan om data visueel voor te stellen. Libraries die in dit pakket voorzien zijn:

* Matplotlib: wordt gebruikt om de data effectief te visualiseren
* Pandas: om data te importeren en te verwerken
* Jupyter: een interactieve Python notebook

Ik heb er voor gekozen om de 64-bit grafische interface voor Windows te installeren.

Eens geïnstalleerd kan je via *Anaconda Navigator,* ***Jupyter*** *Notebook* openen. Hiermee kan je code schrijven, organiseren en testen.

Om **Matplotlib** te gebruiken, dien je eerst de pyplot module te importeren. In de Jupyter Notebook geef je volgend commando in: *“from matplotlib import pyplot as plt”* Dit commando geeft aan dat je de pyplot module moet geïmporteerd worden vanuit matplotlib en plt moet genoemd worden.

Om vervolgens een eerste grafiek te tonen, kan je in Jupyter het volgende commando ingeven:

plt.plot([1,2,3],[1,4,9])

Wil je meerdere lijnen op een grafiek tonen:

plt.plot([1,2,3],[1,4,9])

plt.plot([1,2,3],[10,20,30])

Dit zegt dat je op de X-as punten 1, 2 ,3 moet gebruiken en op de Y-as punten 1, 4, 9. Om dit dan te tonen:

plt.show()

Labels toevoegen doe je op de volgende manier:

plt.xlabel('x-as')

plt.ylabel('y-as')

plt.title('test')

Een legend toevoegen:

plt.legend([‘set1’,’set2’])

Exporteren van de grafiek (afbeelding):

* Download afbeelding vanuit de browser
* Met het commando *plt.savefig(‘naamVanDeFile’)* het bestand wordt dan opgeslagen in dezelfde map als deze waar het project in zit.

**Pandas**

Eerst opnieuw imports doen:

from matplotlib import pyplot as plt

import pandas as pd

Eerst gaan we dummy data ingeven:

data = {'Year':[2008, 2012, 2016],

'attendees':[112,321,729],

'average age':[24,43,31]}

Vervolgens steken we deze data in een variabele om er een dataframe van te maken:

df = pd.DataFrame(data)

Wanneer je deze data dan toont met *df* krijg je de data set te zien in tabelvorm.

Om slecht één enkele kolom te selecteren, kan je dan df[‘Year’] (of data.Year) ingeven. Het resultaat is dan een *Panda series* datatype en is vergelijkbaar met een list in Python.

Hier kan je dan vergelijkingsoperatoren op toepassen: df['Year'] < 2013 die dan een reeks van Booleans zullen terug geven.

Wanneer je deze vergelijking in een nieuwe variabele steekt, dan kan je met df[‘nieuweVariabele’] de data van deze vergelijking laten zien. Wat Boolean Indexing wordt genoemd.

Om de waardes dan tov elkaar af te zetten en te tonen in een grafiek, kunnen we ons beroepen op het voorgaande:

plt.plot(df['Year'], df['attendees'])

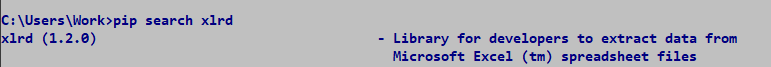
plt.plot(df['Year'], df['average age'])

plt.legend(['attendees', 'average age'])

plt.show()

Data importeren:

* Importeren vanuit een excel file

1. Installeer de xlrd library voor Python indien je deze niet hebt.  
     
     
   Installeren kan via de command line met commando: pip install xlrd
2. xlrd importeren via:  
   import xlrd
3. Om de data in te lezen kies ik eerst het pad waar ik het databestand opgeslagen heb. Ik ga naar het pad en stel de current working directory in. Ik lees de data in met de functie read\_excel()  
   Onderstaand een code voorbeeld:  
   path='/Users/Work/Documents/TIN2/Data Advanced/PE/voetbal'

os.chdir(path)

cwd = os.getcwd()

data = pd.read\_excel("voetbal.xlsx", sheetname="gegevens")

1. zie csv voor verdere uitleg

* Importeren vanuit een csv file.

1. In het csv-bestand de ‘,’ vervangen door het ‘.’. Anders zal de data van gewicht geïnterpreteerd worden als een object en niet als een float.
2. Pandas en pyplot importeren

import pandas as pd

from matplotlib import pyplot as plt

1. Zorg ervoor dat het csv bestand in dezelfde map zit dan het project.
2. De dataset inladen en toekennen aan een variabele: data = pd.read\_csv('voetbal.csv',sep=’;’)
   1. data geeft dan de ganse tabel weer
   2. data.head() geeft de eerste 5 rijen weer.
3. Waarde van de geboortedatum aanpassen + de kolom inzet vullen:

start = pd.to\_datetime('2011-01-01')

end = pd.to\_datetime('2011-12-31')

start\_u = start.value//10\*\*9

end\_u = end.value//10\*\*9

data['inzet'] = ''

for x, row in data.iterrows():

y = (pd.to\_datetime(np.random.randint(start\_u, end\_u), unit='s'))

data.at[x,'geboortedatum'] = (pd.to\_datetime(np.random.randint(start\_u, end\_u), unit='s'))

if y < pd.to\_datetime('2011-04-01'):

data.at[x, 'inzet'] = 'zeer goed'

elif y < pd.to\_datetime('2011-10-01'):

data.at[x, 'inzet'] = 'goed'

else:

data.at[x, 'inzet'] = 'matig'

1. Gewicht sorteren

gewicht\_gesorteerd = sorted(set(data.gewicht))

1. Grafiek weergeven met de correlatie tussen lengte en gewicht

for x in gewicht\_gesorteerd:

y = data[data.gewicht == x]

plt.scatter(y.gewicht, y.lengte)

plt.xlim(19,31)

plt.ylim(110,140)

plt.xlabel('gewicht')

plt.ylabel('lengte')

plt.show()