Lab 1 - DAVE3625-1 23H Introduksjon til Kunstig Intelligens

Most of the time spent working on AI, is actually time spent preparing data. You need to figure out what datapoints to use, and if you can combine datapoints to get a better model.

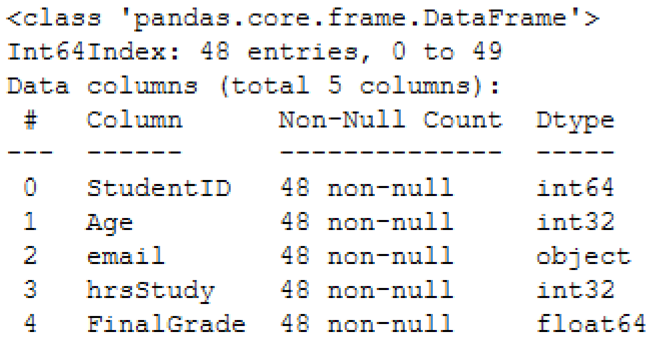
The first task when working with a new dataset is to clean the data and solve data errors. In the file stud.csv, we have 50 entries with:

StudentID, Age, email, hrsStudy, FinalGrade

In this lab, you will import the csv file into pandas:

```
Hint:
df = pd.read_csv(url, sep=',')
df.head()
```

You will then clean the data set so df.info() produce



```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 48 entries, 0 to 49
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   StudentID   48 non-null     int64  
 1   Age         48 non-null     int32  
 2   email       48 non-null     object 
 3   hrsStudy    48 non-null     int32  
 4   FinalGrade  48 non-null     float64
```

```
Hint:
df.isna().sum() #show missing values
df=df.replace(r'^\s*$', np.nan, regex=True) #Replace blank values with
np.nan values

df['Column'] = df['Column'].astype(str).astype(int) #Convert from obj to
int
```

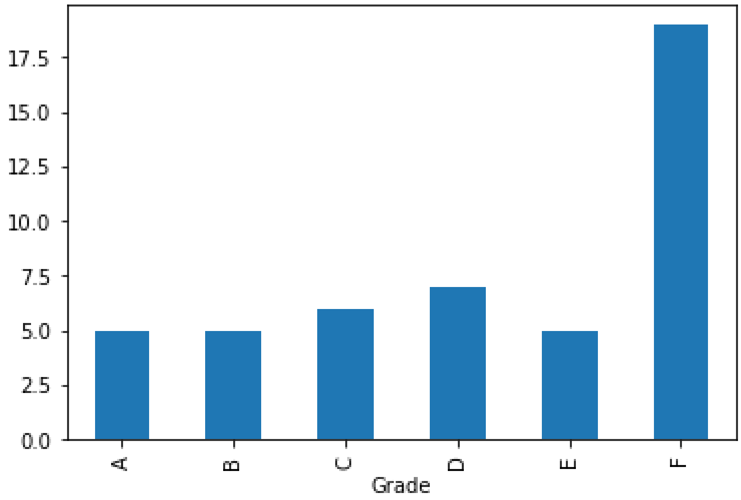
Then idenify and remove the outliers in the «FinalGrade» column

```
Hint : df["FinalGrade"].plot.box()
```

Finally add a column "Grade" where you transform the grade from float to a char:

91 - 100 = A
81 - 90 = B
71 - 80 = C
61 – 70 = D
51 – 60 = E
> 50 = F

And produce this plot:


A
B
C
D
E
F
Grade