REPORT ON STATIC CODE

Programming II final assignment

Maryam Jalali

Email: m.jalali@st.hanze.nl

Student ID: 443224

Teacher: Bart Barnard

Inhoud

1- Introduction:

The goal of this report is to give refactoring ideas about a specific codebase. The report will analyze the current state and look at several aspects that contribute to the maintainability, readability, flexibility, and transferability of the code. The aspects include the SOLID principles and design patterns. The codebase that will be analyzed contains only data from students following Mathematics and Portuguese lessons. Therefore, it is only a small sample out of a varying population. Thus, this is only representative for these students. Source Information P. Cortez and A. Silva used Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th Future Business Technology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7. What will be analyzed is to see if there is a correlation between Weekly/Daily Alcohol Consumption, Parents' Education and 3 semesters grades of these two courses.

To compare them, variety of classes and functions have been generated. We could call the classes/functions anytime with other datasets as well. Varity of plots are called and functioned to project a better view and understanding of the dataset. Plus, to answer the questions some data analyses and calculations are done.

Regarding the SOLID technique the codebase changes in the new version and functions which were in script way are sorted under main classes and all classes are divided into integrated functions which are following one purpose (A class can (and usually does) have multiple methods as long as they correspond to the same logic that that class is in charge of handling) which helps maintain readability and transferability as well.

All the plots are divided nicely into classes and subclasses as well, and derived classed are substitutable for their base which is the Liskov Substitution Principle.

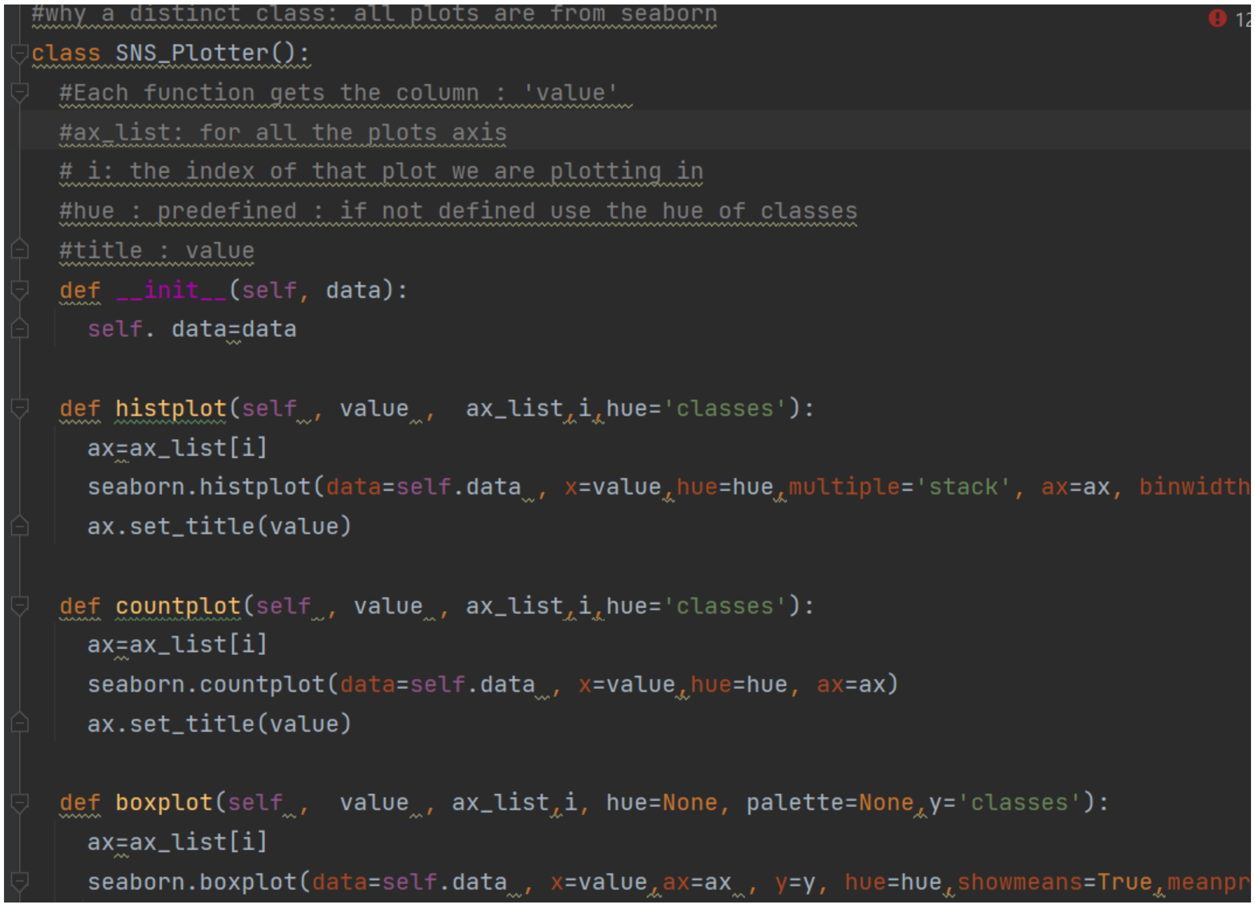
2 – Defining The plotters (Classes)

In this part, Plotters are defined as our code classes.

1. SNS Plotter with all subcategories of which we can find Correlation plotters
   - Hist plot
   - Count plot
   - Box plot
   - Correlation Map

Plotters above have two arguments Ax list and i which are the location and a value which defines the column.

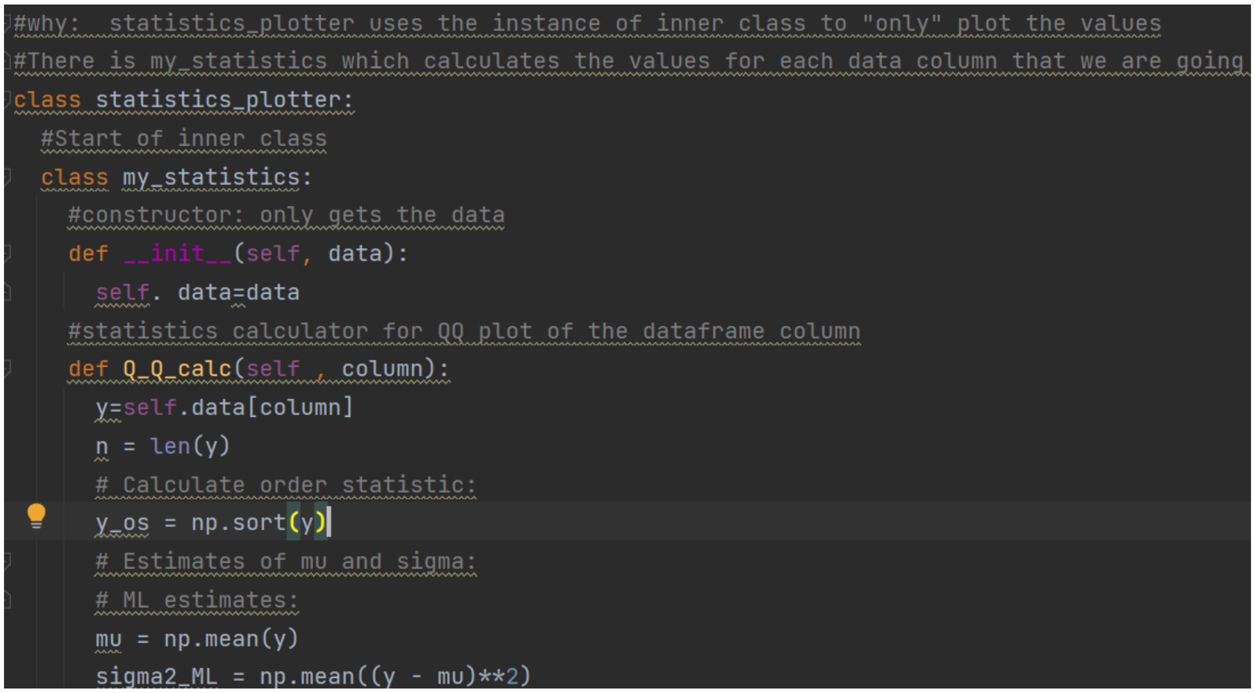
The only Dunder used in this case is __init__ which is called automatically by python, and all the functions of different plotters are defined under the same plotter class which projects a good practice of defining classes and initializing the functions related. All the functions are related to SNS plotters and they are specifically doing one particular action.



```
#why a distinct class: all plots are from seaborn
class SNS_Plotter():
  #Each function gets the column : 'value'
  #ax_list: for all the plots axis
  # i: the index of that plot we are plotting in
  #hue : predefined : if not defined use the hue of classes
  #title : value
  def __init__(self, data):
    self. data=data

  def histplot(self , value ,  ax_list,i,hue='classes'):
    ax=ax_list[i]
    seaborn.histplot(data=self.data , x=value,hue=hue,multiple='stack', ax=ax, binwidth
    ax.set_title(value)

  def countplot(self , value , ax_list,i,hue='classes'):
    ax=ax_list[i]
    seaborn.countplot(data=self.data , x=value,hue=hue, ax=ax)
    ax.set_title(value)

  def boxplot(self ,  value , ax_list,i, hue=None, palette=None,y='classes'):
    ax=ax_list[i]
    seaborn.boxplot(data=self.data , x=value,ax=ax , y=y, hue=hue,showmeans=True,meanpr
```

2. Statistic Plots
    - Q-Q-Calc
    - Normal Binomial
    - Q-Q-Plot

Basically, the same method is used here to define these classes and their functions and what makes it different is every line for the calculation has a comment and if a viewer or another programmer wants to review the code, the comments assist them to have a better understanding of the code. So, readability has been taken to account in this particular area.

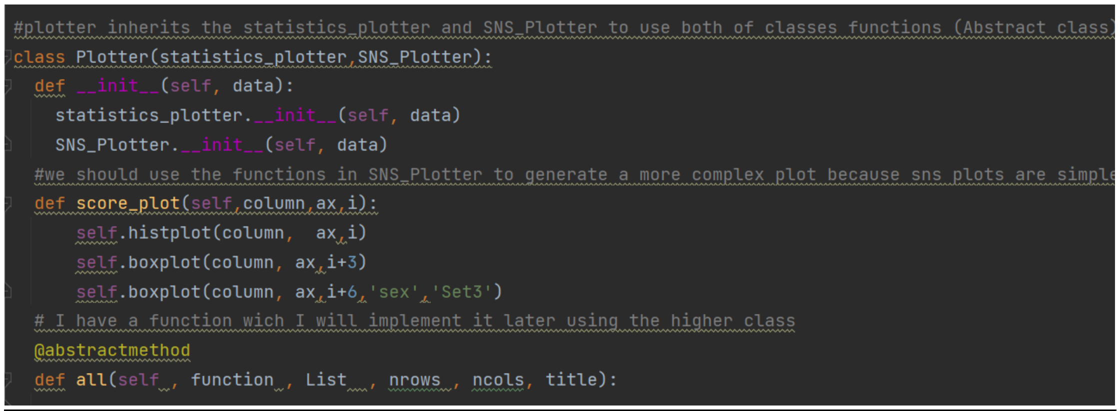


```
#why:  statistics_plotter uses the instance of inner class to "only" plot the values
#There is my_statistics which calculates the values for each data column that we are going
class statistics_plotter:
  #Start of inner class
  class my_statistics:
    #constructor: only gets the data
    def __init__(self, data):
      self. data=data
    #statistics calculator for QQ plot of the dataframe column
    def Q_Q_calc(self , column):
      y=self.data[column]
      n = len(y)
      # Calculate order statistic:
      y_os = np.sort(y)
      # Estimates of mu and sigma:
      # ML estimates:
      mu = np.mean(y)
      sigma2_ML = np.mean((y - mu)**2)
```

3 - Inheritance with classes (SNS plotter and Statistics)

In this part of the code two main plots are combined with all the characteristics of and score plot has been added which its functions are to draw another column and it has an abstract method named “all” which will be called later in the code.

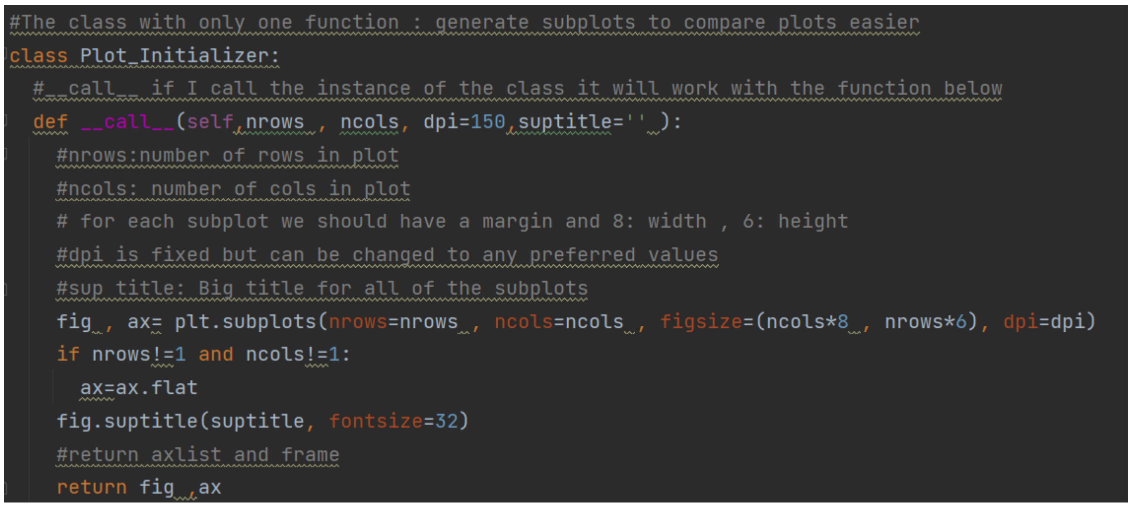
The reason we use this is to call the other function under a class with their subfunctions without defining them every time for all the functions. It avoids repetition and prevents from occupying too many memory units.



```python
#plotter inherits the statistics_plotter and SNS_Plotter to use both of classes functions (Abstract class
class Plotter(statistics_plotter,SNS_Plotter):
  def __init__(self, data):
    statistics_plotter.__init__(self, data)
    SNS_Plotter.__init__(self, data)
  #we should use the functions in SNS_Plotter to generate a more complex plot because sns plots are simple
  def score_plot(self,column,ax,i):
      self.histplot(column,  ax,i)
      self.boxplot(column, ax,i+3)
      self.boxplot(column, ax,i+6,'sex','Set3')
  # I have a function wich I will implement it later using the higher class
  @abstractmethod
  def all(self , function , List , nrows , ncols, title):
```

Plot initializer

By calling its Dunder function (“call”) we can generate the subplots we have defend in the program.



```python
#The class with only one function : generate subplots to compare plots easier
class Plot_Initializer:
  #__call__ if I call the instance of the class it will work with the function below
  def __call__(self,nrows , ncols, dpi=150,suptitle='' ):
    #nrows:number of rows in plot
    #ncols: number of cols in plot
    # for each subplot we should have a margin and 8: width , 6: height
    #dpi is fixed but can be changed to any preferred values
    #sup title: Big title for all of the subplots
    fig , ax= plt.subplots(nrows=nrows , ncols=ncols , figsize=(ncols*8 , nrows*6), dpi=dpi)
    if nrows!=1 and ncols!=1:
      ax=ax.flat
    fig.suptitle(suptitle, fontsize=32)
    #return axlist and frame
    return fig ,ax
```

<u>Prof Plotter Class</u>

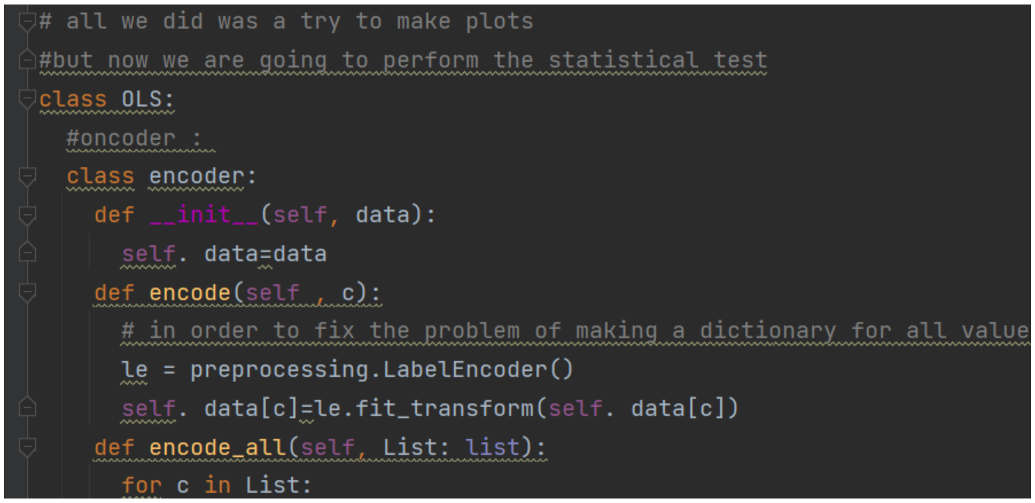
Having all the codes provided, we initialized the Prof-Plotter to draw the subplots needed.

<u>OLS</u>

It contains an inner class called encode to do all the encodings.

Here we need to define a dunder __all__ to set and generate the formulas and print the tables needed.

Also, subclasses are defined with indentation and their own functions like a good practice which helps to have more readability of the code.



```
# all we did was a try to make plots
#but now we are going to perform the statistical test
class OLS:
  #oncoder :
  class encoder:
    def __init__(self, data):
      self. data=data
    def encode(self , c):
      # in order to fix the problem of making a dictionary for all value
      le = preprocessing.LabelEncoder()
      self. data[c]=le.fit_transform(self. data[c])
    def encode_all(self, List: list):
      for c in List:
```

4- Comparison

As I checked the first version of my code there was no cohesion or coherence between the script and plots and regarding readability it was very difficult to understand and distinguish the code with a glimpse or even in-depth consideration. There was also lack of changeability which by organizing every single part of the code we can easily have access to different classes and also classes can be easily called in other classes and they inherit from their main source. What I changed was to put the lessons learnt into practice and organize all the classes, their subclasses and functions under one purpose categories and draw them as figure below.

First Version code

- Importing too many libraries
- Data Observation and cleaning
- Scripts
    - Functions
    - A Class
- Functions
- Projecting Plots
- Scripts
- Functions
- Projecting Plots
- Scripts
- Conclusion

New Version Code

- Importing a few needed libraries
- Defining Plotters as classes
    - Class A
        - Function x
        - Function y
    - Class B (with Functions)
    - Class C (with Functions)
    - Class D (with Functions)
- Defining Plot initializers
    - Class E
        - Function z
- Generate Plots
        - Function i
- Call all the plotters and compare
- Conclusion

5- Conclusion:

To answer the 3 questions asked

1. Is there a significant difference between alcohol consumption and the parents’ education?
2. Is there a significant difference between alcohol consumption and the parents job?
3. does alcohol consumption during the weekend influence the students final grade?

Some Information about Columns name:

Medu = Mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary and

Fedu = Father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary

Mjob = Mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g., administrative or police), 'at home'

Fjob = Father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g., administrative or police), 'at home'

Walc = Weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)

G3 = Final Grade in exam (numeric: from 0 to 20, output target)

Health = Current health status (numeric: from 1 - very bad to 5 - very good)

- from a first visual inspection we can state the following: We can state that: Subjects with low Walc (1,2) are getting marks higher than 6 to 19 except one 0
- Subjects with medium Walc (3) are getting marks from 8 to 18 and females are in a good status
- Subjects with high Walc (4,5) contains the highest number of failed ones, however there is an exception which has the highest number among all Based on the scatter plot above, there is no influence of alcohol consumption on final grade, for the Portuguese course students. Here as a same data type with different group of student as Portuguese lesson We can state that Subjects with low Walc(1) are mostly getting higher marks than the others
- Subjects with medium Walc (2-3) females are getting the highest and the lowest marks 4,5,6 and 17,18

- Subjects woth high Walc (5) are getting average G3 from 7 to 13 Based on the heatmap for both data frame, we can draw the following conclusions regarding the research questions:
- There is positive correlation between student grades and Alcohol consumption. There is positive correlation between parents' education and Alcohol consumption. The job of parents has effect on Alcohol consumption in teenager. from the scatter plot it can be observed that in the weekend the alcohol consumption in the male students is more than female students. but it seems that in general alcohol consumption in the weekend does not influence the final grade (G3). The box plot above also confirms the result of the previous scatter plot. Walc does not influence the G3. and male students consume more alcohol during the weekend. Boxplot shows that the health condition of the students that consume less alcohol better than the others (category 4 and 5). this is more obvious in the data of male students. Form this data we can conclude that the students who have raised by the higher educated parents drink less alcohol and at the same time parents' job could influence student alcohol consumption. Furthermore, Alcohol consumption did not show any effect on final grade of the students also male students drink more alcohol than female students during the weekend. There is a significant difference between the alcohol consumption in comparison with the jobs of the parents ($p=0.000209 < 0.05$).