ASSIGNMENT 3

The third assignment is again only Python. There are three files: **Assignment3.ipynb**, **1.csv**, and **2.csv** (each corresponding to its question number). Please follow all the instructions, which are repeated here. **When you are finished, please submit the Python notebook file ONLY. Please do not rename the file, but do enter your Student ID (numerical), in the notebook**.

(1) Do not delete any cells
(2) Please put code only in cells where it has

```
# YOUR CODE HERE
```

(3) Make sure you give your answers in the cells below the question.
(4) Make sure you follow the naming of the variables according to the instructions.
(5) Remember that unless specified each plot should have a title and axis labels (and a legend there are multiple lines/plots on the same axis).
(6) Make sure you delete *raise notImplemented()/*

Question 1 - Linear models and residuals

Run the first cell to load the dataset.

a.) Use *numpy.polyfit()* to fit a linear model to the data, storing the parameters into variables called m and b respectively for the slope and y-intercept.
b.) Compute the residual of the data with respect to the model and store them in a variable called *res*
c.) Plot an unnormalised histogram of the residuals with 40 bins.
d.) Is the linear model a good fit? (one or two sentences)
e.) Complete the function below to compute hinges of the input x, which should be set at the 30th and 70th quantile.
f.) Use the above function to compute the inner fences of the residuals. Store the results in the variables *lower_fence* and *upper_fence* respectively.
g.) Create a new Dataframe called Y with all the values corresponding to residuals which are beyond the inner fences removed.
h.) Fit a new linear model and plot the outliers as red points, the remainder of the points (non-outliers) as blue stars (*), the original model as a black line and the new model shown as a green line. Store the new model parameters in the variables *m_new* and *b_new*.
j.) Is what you did above different from computing (and filtering) the outliers on the original data? If so, describe how, otherwise explain why they are the same. (a few sentences)

Question 2

The next dataset is on the sales of video games. Use Pandas to read in the file **games.csv** into a dataframe called *games*.

a.) Find answers to the following questions about the dataframe.
- How many genres are there? Store the answer in a variable called *num_genres*.
- How many publishers are there? Store the answer in a variable *num_publishers*.
- How many platforms are there? Store the answer in a variable *num_platforms*.

b.) Create a box-and-whisker plot of Global Sales by genre.

c.) Find the upper outer fence thresholds for each genre (assuming hinges at the 25th and 75th percentile). Store them in a dictionary {genre:upper outer fence} called *ufence*.

d.) Plot the unnormalised histogram with 20 bins for all sports games whose Global sales which are beyond the outer fence.

e.) these are successful games - what can you say about the maximum versus the typical successful game? What are the problems with using the average (overall or just the successful games).

f.) Create a line plot of Global Sales as the dependent variable and year as the independent variable, with a line for each of the following platforms: 2600, NES, SNES, GEN, N64, GC, WII, PS, PS2, PS3, PS4, X360, XB, XOne.