**Assignment 1**

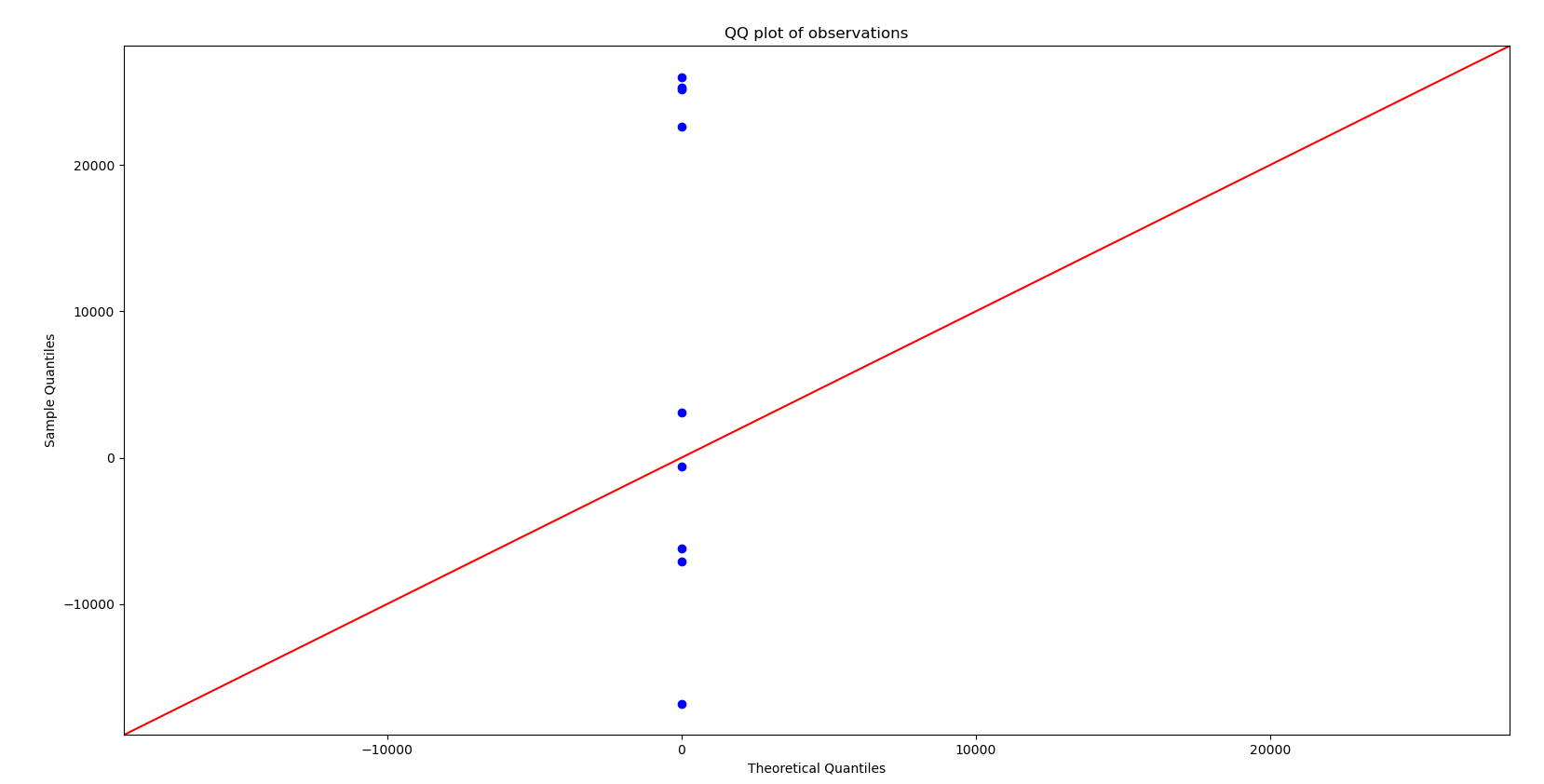
**Ido Levkovitch – 205619919**

**Eilon Toledano – 206903445**

**Question 1:**

We took the 9 observations of the annual rates of tax return of different companies and doubles it by thousands.

1. We constructed a Q-Q plot of the observations and got the following result:

****

As we know the Q-Q plot compares our observations to Normal distribution, based on its shape. If our data would have been normally distributed, we would have seen our values approximately around the x=y line.

Since our results are far from being close to the line, it indicates that our observations distribution is not a normal one.

1. The Shapiro-Wilk test is another test to see whether our data distributes like a normal distribution. Its results are:
   1. P value - Indicates that if the value is less than the alpha level, the null hypothesis is rejected, and if it is higher than it can’t be rejected.
   2. Statistics - The statistic will take a value between 0 and 1 with 1 being a perfect match.

We performed a python simulation and got the following results:



Therefore, according to *α* = 0*.*1, we reject the theory that the data is distributed like a normal distribution, and it corresponds to the results of the Q-Q plot.

1. In order to find the regression equation Y = aX + b (Y = distance, X = Age) we need to calculate a and b variables according to the formulas:

From the data we received the mean and variance of X and Y and the correlation.

The correlation formula is

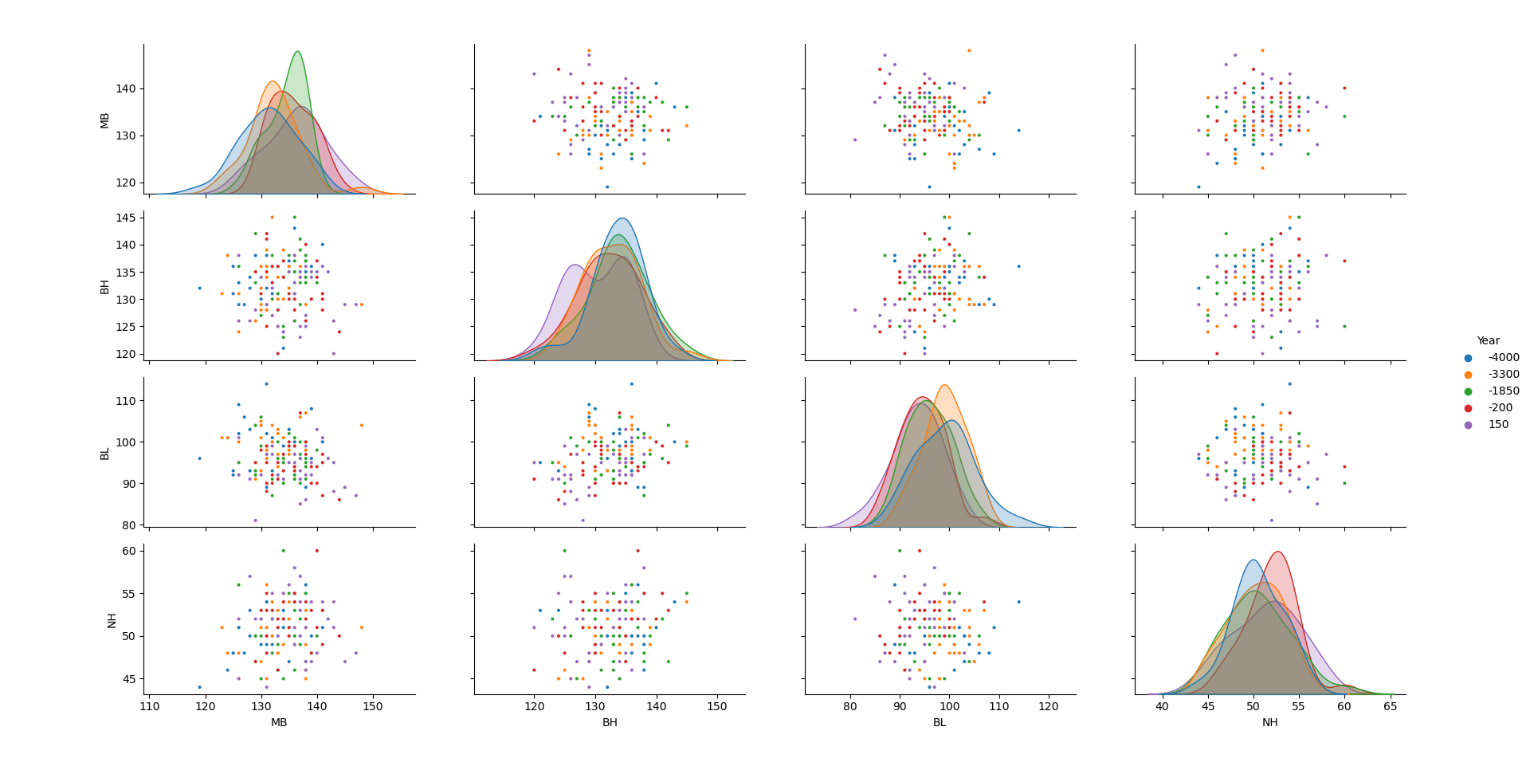
From it we extracted the value of a:

We ran a python script and got the following results:



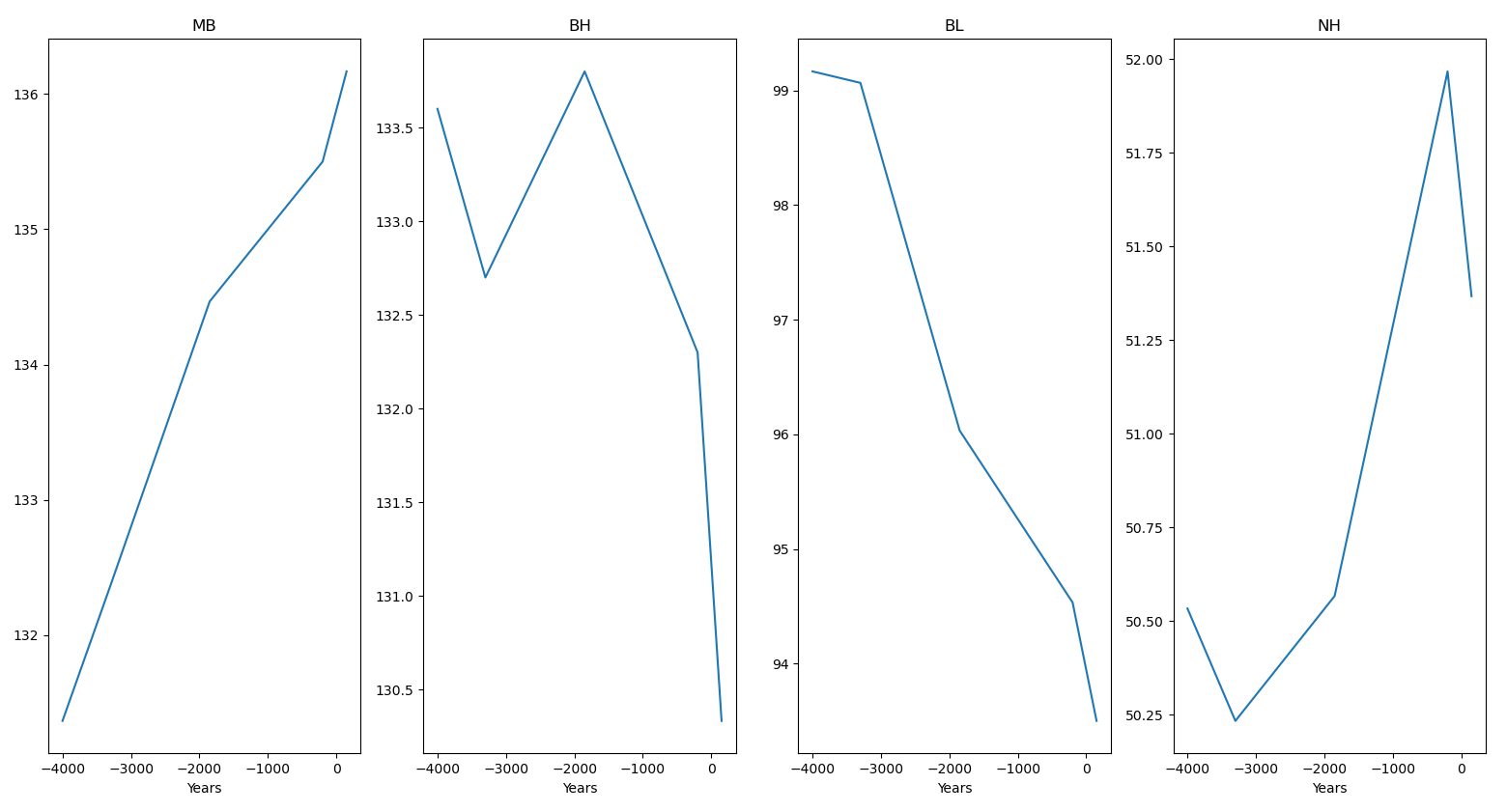
**Question 2:**

We used three different visual EDA methods that show how the skull size has changed, if at all, over time:

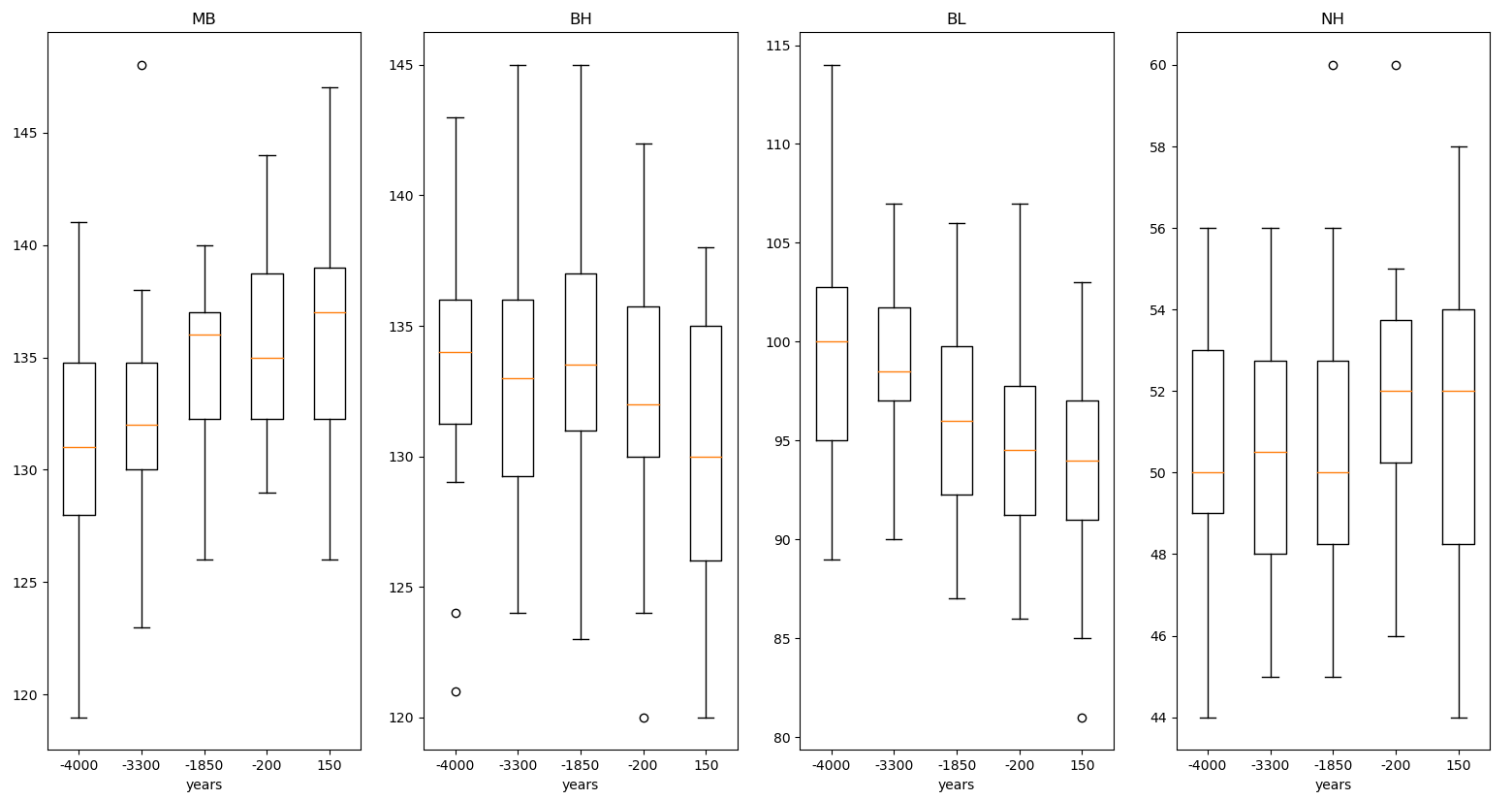


Pair plot of two attributes in each year (differentiate in color)

Plot of the mean each attribute change over the years



Box plot of attributes for each year (differentiate in color)



The first is a pair-plot graph according to each year, a graph that plots the values of the attributes according to each year. Each year value is marked with a different color. Based on it, if we check two parameters, we can’t infer to which year is the skull with high probability. The second graph is the mean for each feature over the years, and the third graph is a box plot graph for the features in each year. Based on both of them we can conclude that maximum breadth of skull and the nasal height of skull increased in the over all by a small percentage of about 5%, and the basibregmatic height of skull and basialveolar length of skull decreased in a similar percentage, but all of them have high variance.

**Question 3:**

1. Aa
   1. output[68] show all association rules with more than 1 item on the left side

with a hitmap such that higher confidence in the role is colored brighter. For example we can see that if a book is catalog as childBks and artBks it will be GeogBks in 62% of the times.

* 1. leverage(A, B) = Support(A|B) - Support(A)\*Support(B)

lift uses ratio, while leverage uses difference. Because of the difference in the formula, leverage favors item sets with higher support, while lift can find strong associations among less frequent item set.

Conviction(A, B) = (1 - Support(B))/(1 - Confidence(A, B))

give us the frequency that role would be incorrect in indipendate data, compare to our data

Unlike lift, conviction is sensitive to rule direction so if we got that the Conviction((YouthBks), (ChildBks)) is 1.73 that means that the rule would be incorrect 73% more often if the association between (YouthBks) and (ChildBks) was purely random chance.