Assignment 1 CS4070

Nirmal Roy, 4724429

November 27, 2017

1 Question 1

1.1 Population

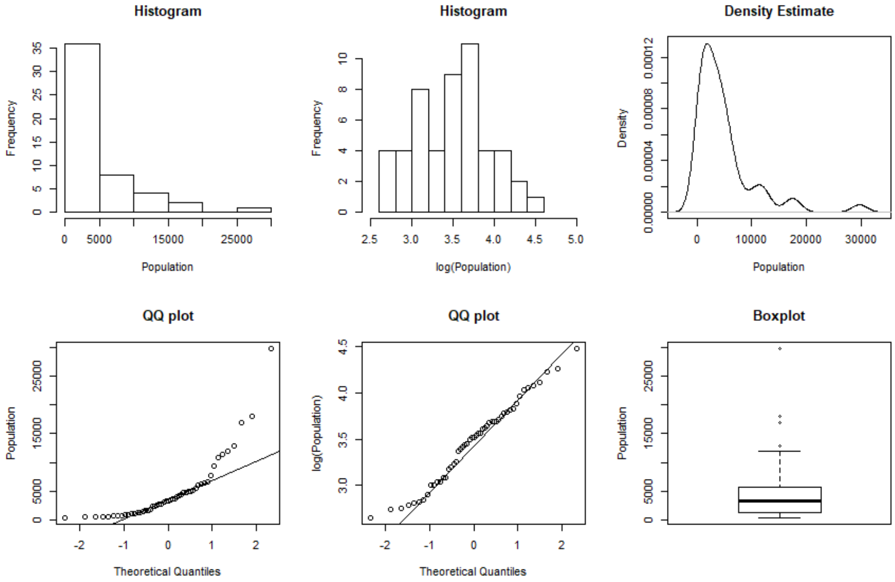

Figure 1: Various plots of the Population variable

- The first image shows a histogram plot of the Population variable where it can be seen that most of the values lies between 5000 and 10000 and hence to properly observe the data, a log_{10} transformation of the variable has been plotted in the next figure. The transformation 'corrects' the histogram plot and detailed observation on the variable can be done. The main peak is somewhere around a population of $10^{3.5} \sim 10^{3.7}$.
- The density estimate shows the presence of a mode in the population variable and it is very high compared to the other peaks which appear.
- The QQ plots are compared with a standard normal distribution. In the first QQ plot we see that the variable is positively skewed because

at both ends the data lies above the line, and we can also see presence of some outliers towards the extreme right. The second QQ plot is log transformed and it appears more linear and normal. But the presence of outliers is hidden in this representation.

- With the boxplot we can easily see the median, first and third quartile. Also the positive skewness of the variable is visible as the top fence is longer than the bottom fence. Moreover, it shows the presence of three outlier values

1.2 Verbal SAT score of students

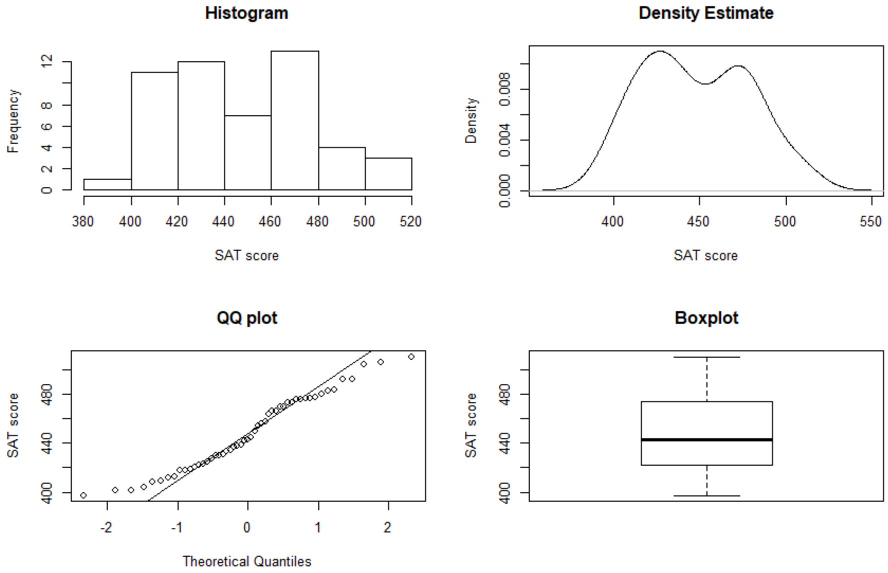

Figure 2: Various plots of the Verbal SAT score variable

- The histogram gives a rough estimate of the presence of two modes and the symmetric data
- The density plot clearly identifies the presence of two modes(one near 425 and the other near 475.
- From the QQ plot we can see that the data doesn't deviate much from the standard normal distribution. Also, near the lower values the data tends to be positively skewed as the points lie above the line whereas towards higher value the data tends to be negatively skewed where the points lie below the line. Hence there's absence of tails in the data. The data roughly follows a normal distribution
- The boxplot also shows the symmetry of the data and the absence of any outlier values. The median is close to 440

1.3 Maths SAT score

- The SATM variable has almost similar properties as the SATV variable of the previous section. This data is also bimodal, roughly symmetric and doesn't have any outlier values. This histogram fails to represent the bimodality of the data which is clear in the density estimate. The absence of heavy tails is apparent both from the QQ plot and the density estimate plot. The median SAT score is close to 490

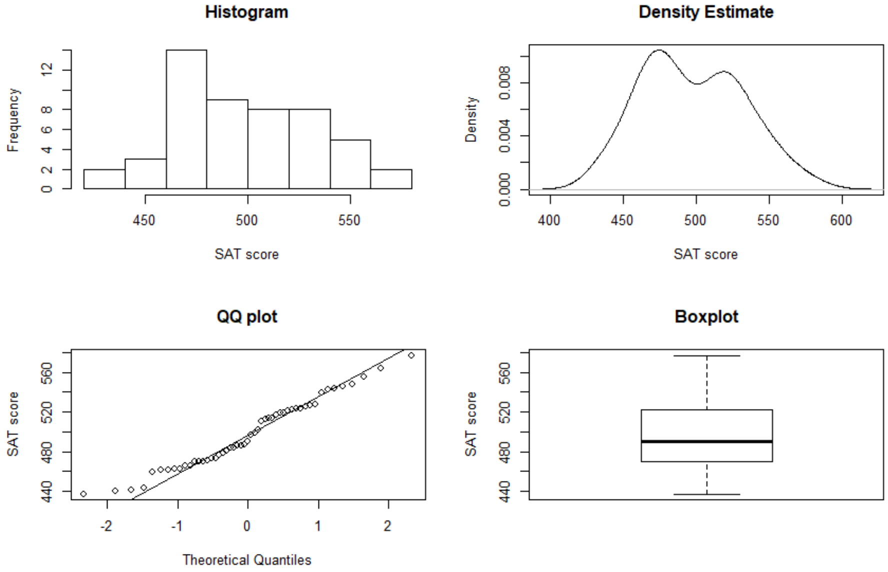

Figure 3: Various plots of the Maths SAT score variable

1.4 Percentage of graduating high school students

- As seen in 4 the histogram identifies the absence of the value 30% and the two peaks are also visible in the data
- The two peaks of the density estimate represents the two modes of the data where as the plot also indicates that the data is spread across the entire range. The data is approximately symmetric around 40. But it lacks normality.
- The QQ plot shows that the data doesn't follow normal behavior. The points above the line near the bottom shows that it starts as a positively skewed manner and it ends as a negatively skewed manner as indicated by the points below the line towards the end
- The box plot shows that the data interquartile range lasts for a considerable period from 10% to about 60% and that there is no unusual value in the data. Inspite of the wide spread the median is at 25% that is near the bottom

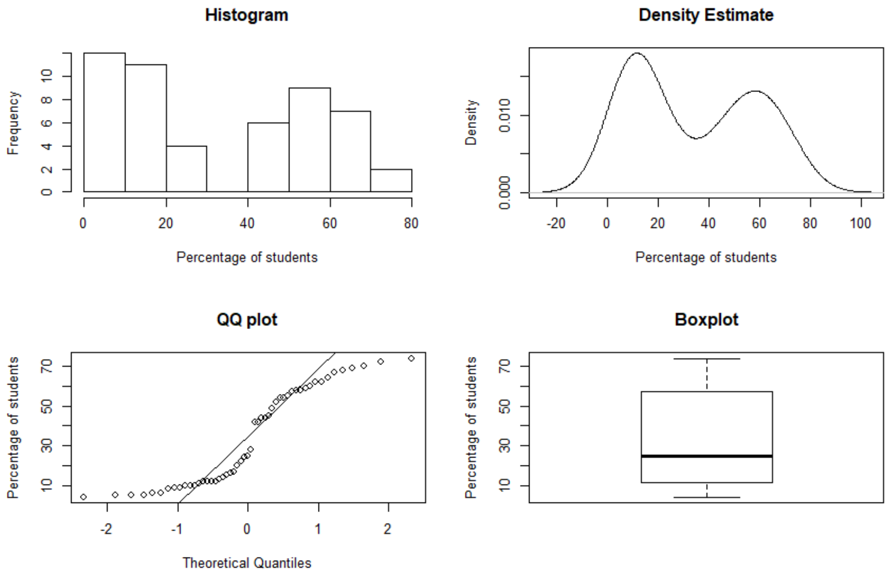

Figure 4: Various plots of the Percentage of graduating high school students variable

1.5 State Spending in Education

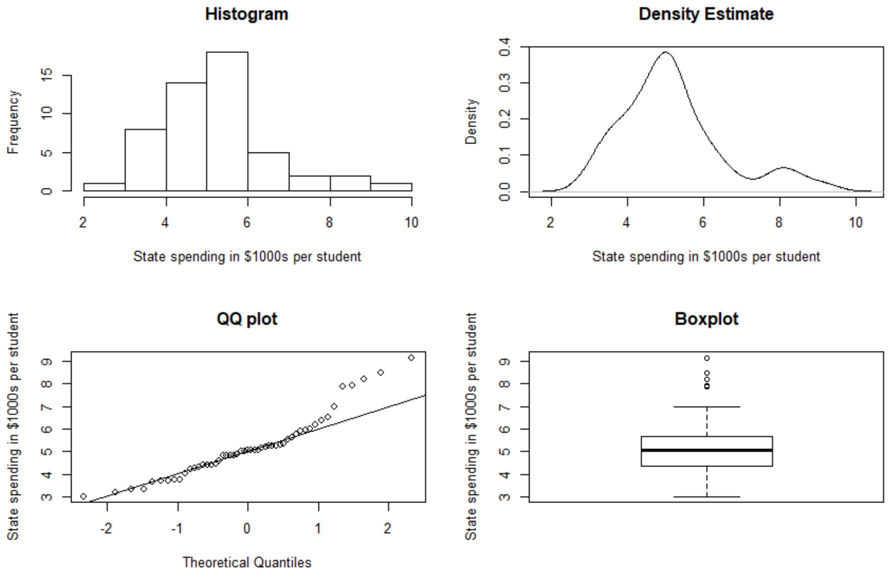

Figure 5: Various plots of the state spending on education in $1000s per student

- The histogram and density estimates shows a peak at the value 5 which indicates that the data has one mode. And as observed from the density estimate plot the data looks fairly symmetric in the beginning but tapers off at the end

- As concluded from the density estimate plot, the QQ plot confirms that the data follows straight line until the value of 6. Whereas the points much above the line indicates the positive skewness of the data at the end and indicates the presence of a heavy tail distribution.
- The boxplot with it's equal fence width and hinge width shows the data is fairly symmetric and in addition confirms the presence of four outlier values which indicates that four states spends much more money in comparison to the other states. The median is at 5

1.6 Average teacher's salary in the state

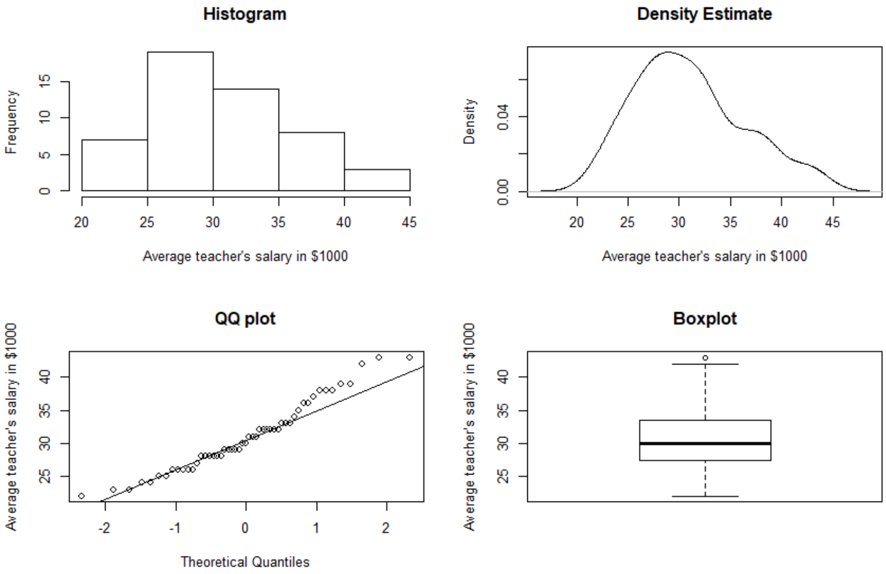

Figure 6: Various plots of the average teacher's salary in the state in $1000s

- The histogram and density estimate plots show that the data is a bit skewed positively and having one peak(mode) at around $25 - 30$
- The positive skewness of the data can be confirmed from the QQ plot where points at both ends lie above the line. But the points follow the line in the middle values which shows that the data has quite a good amount of normality
- The slight skewness is also apparent from the longer upper fence of the boxplot. There is one outlier value and the median is at 30

1.7 Discussion

Histograms

Histogram having the proper bin size gives a fair estimate regarding the number of modes and skewness of the data. But since it's not a continuous plot it provides and approximate estimate of the data properties and might lead to misinterpretations. Also properties like normality and unusual values are hidden in this representation.

Non Parametric Density Estimate Plots

These are perhaps the best representations of the data as it is simple, intuitive and reveals almost all properties of the data. The peaks tell us about the number of nodes, the shape tells us about the skewness/heavy tail of the data or whether the data is symmetric and follows the normal distribution. It will fail to say if a value is 'unusual' but can be guessed if a plot has a very heavy tailed nature.

QQ Plot

These plots gives us an idea about the normality of data. With these plots data can be compared with any 'standard' distribution and we can check how similar the data is with that distribution. Hence, if we compare a data with the normal distribution we can check how far the data deviates from being normal. Moreover we can easily check the skewness of the data if points towards the end lie above or below the line. However, information like unusual values, median, mode are generally hidden in these plots

Boxplots

The main usefulness of boxplots lies in the fact that it gives a clear indication of the number of outlying values, about the median and interquartile range of the data. We can check the range where the data is mostly concentrated. If the fence width and hinge width are observed the boxplots can also give an idea about the skewness/symmetry of the data. However boxplots might not be intuitive for observing skewness and doesn't reveal information about normality or the number of modes in the data.

2 Question 2

2.1 Relationship between Per Capita Income and Infant Mortality Rate

The first plot shows the relationship between per capita income and infant mortality rate. As expected when there is low per capita income rate the infant mortality rate is much higher. The dense lower region indicates that there are a lot of countries with low per capita income and they usually have infant mortality. The tailing off indicates that the number of countries with high per capita income is low and with low infant mortality rate too. There are a few outlier values clearly seen in the plot, which has a very high infant mortality rate

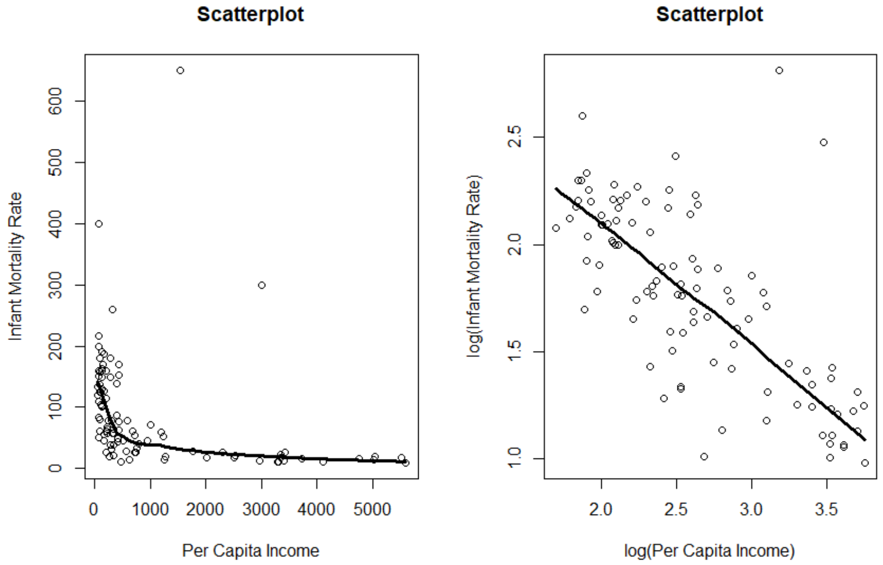

Figure 7: Relationship between per capita income in USdollars and infant mortality rate per 1000 live births with a lowess regression fit having a span of $\frac{2}{3}$

compared to countries having the same per capita income rate. The cause can be attributed to poor healthcare systems in those countries. The second figure is a plot of the log transformation of the both the response and explanatory variable. Sometimes plots like in figure one hides some important data in the cluster so a transformation stretches those parts to reveal the data.

2.2 Relationship between Region and Infant Mortality Rate

The boxplots (see fig 8) clearly reveal Africa has a very high infant mortality rate. This can be attributed to the low per capita income in African countries. Asia and America has similar medians but the interquartile range of Asia is larger indicating that the spread of data is higher i.e the infant mortality rate takes a lot of different values. The data also has a relative positive skewness in case of Asia. Both the median and the spread of infant mortality rate is low in Europe. Needless to say average per capita income is highest in Europe. All regions have one or two outlying data.

2.3 Relationship between whether a country exports oil and Infant Mortality Rate

There is not much difference between the median(fig 9 of the infant mortality rate of an oil exporting country and a non oil exporting country. However the former has higher spread and higher fence which means they have some countries with higher values of infant mortality rates. This makes sense as oil exporting countries are prone to health and pollution hazards.

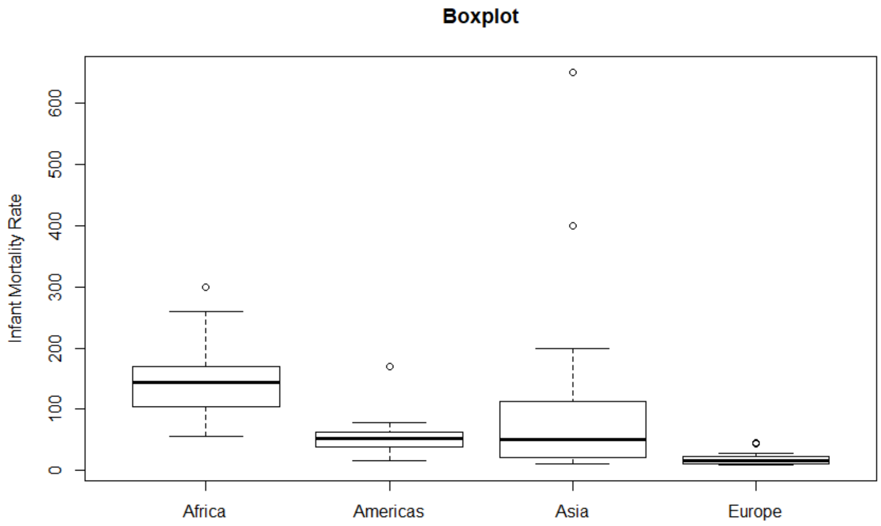

Figure 8: Relationship between Region and infant mortality rate per 1000 live births with boxplots

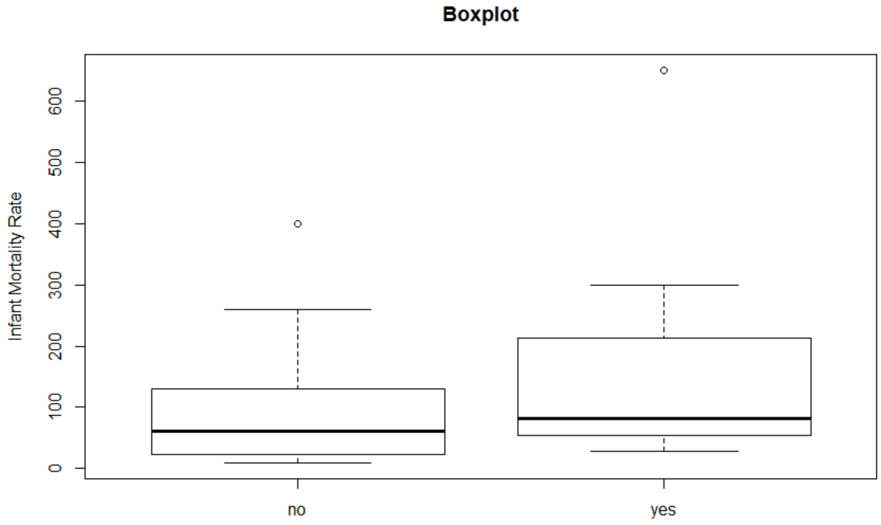

Figure 9: Relationship between oil export and infant mortality rate per 1000 live births with boxplots