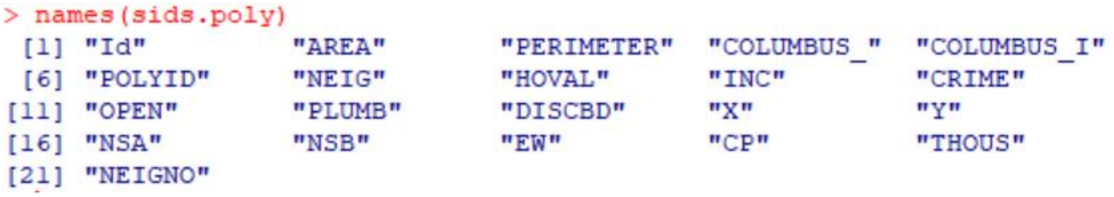
Lab1 Report

JIANG Yuhan 18106651x

Task1: Read *Columbus.shp*

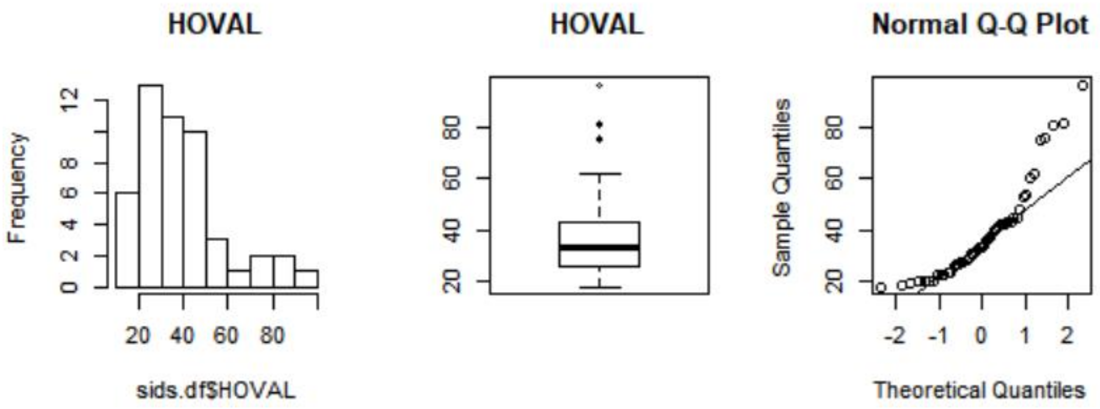


```
> names(sids.poly)
 [1] "Id"         "AREA"       "PERIMETER"  "COLUMBUS_"  "COLUMBUS_I"
 [6] "POLYID"     "NEIG"       "HOVAL"      "INC"        "CRIME"
[11] "OPEN"       "PLUMB"      "DISCBD"     "X"          "Y"
[16] "NSA"        "NSB"        "EW"         "CP"         "THOUS"
[21] "NEIGNO"
```

The positions corresponding to HOVAL, INC, and CRIME are 8, 9 and 10. This message will be used to draw the scatterplot matrix.

Task2: Draw histogram, boxplot and QQ plot of the three variables.

1. HOVAL

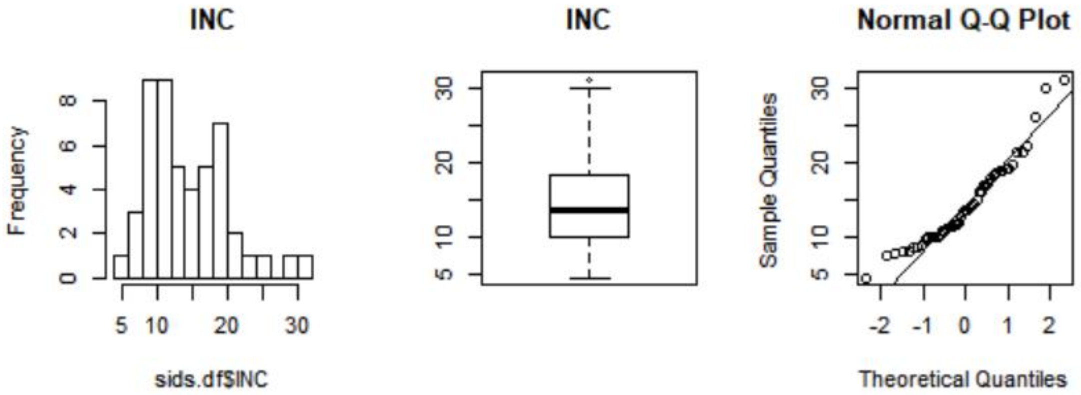


Histogram: The trend is similar to the normal distribution. However, some differences are between the shape of the histogram and that of the normal distribution. The peak is near 20 instead of the middle of this set of data and there is a sharp drop on the right part of the data. Additionally, the shape is not asymmetrical.

Boxplot: There are few suspected outliers above the third quartile and the maximum value and minimum value are not symmetric about the median. What's more, the length of the box is short, which means the dispersion of this data is small.

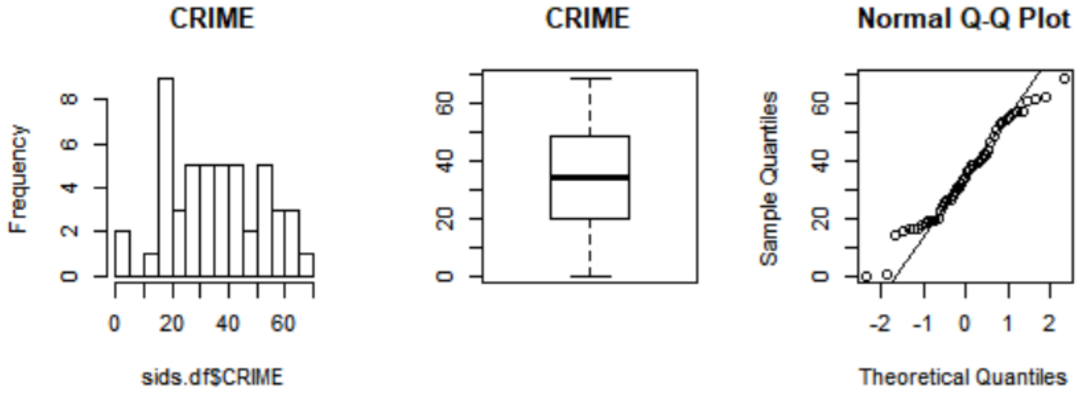
QQ plot: Some points at the tail are above the line which means they are larger than the expected value. And they go further and further from the expected value. Those part of the data maybe corresponds to the outliers on the boxplot. The rest part of the data is just on the line. So, this set of data is more likely to be a Student's-t (kurtotic) data set.

2. INC



Histogram: There are two peaks in this histogram while there is only one peak in the histogram of normal distribution. Additionally, the higher peak is not at the middle of this set of data and the shape is not asymmetrical. However, the trend is similar to the normal distribution. The frequency becomes small at both ends.
Boxplot: There are still suspected outliers above the third quartile. But they are less than the outliers in HOVAL. Furthermore, the maximum value and minimum value are not symmetric about the median. And the length of the box is longer than that of HOVAL, which means the dispersion of this data is larger than that of HOVAL.
QQ plot: Some points at both ends are obviously above the line which means they are larger than the expected value. The rest part of the data is just on the line. So, this set of data is more likely to be chi-squared (skewed) data set.

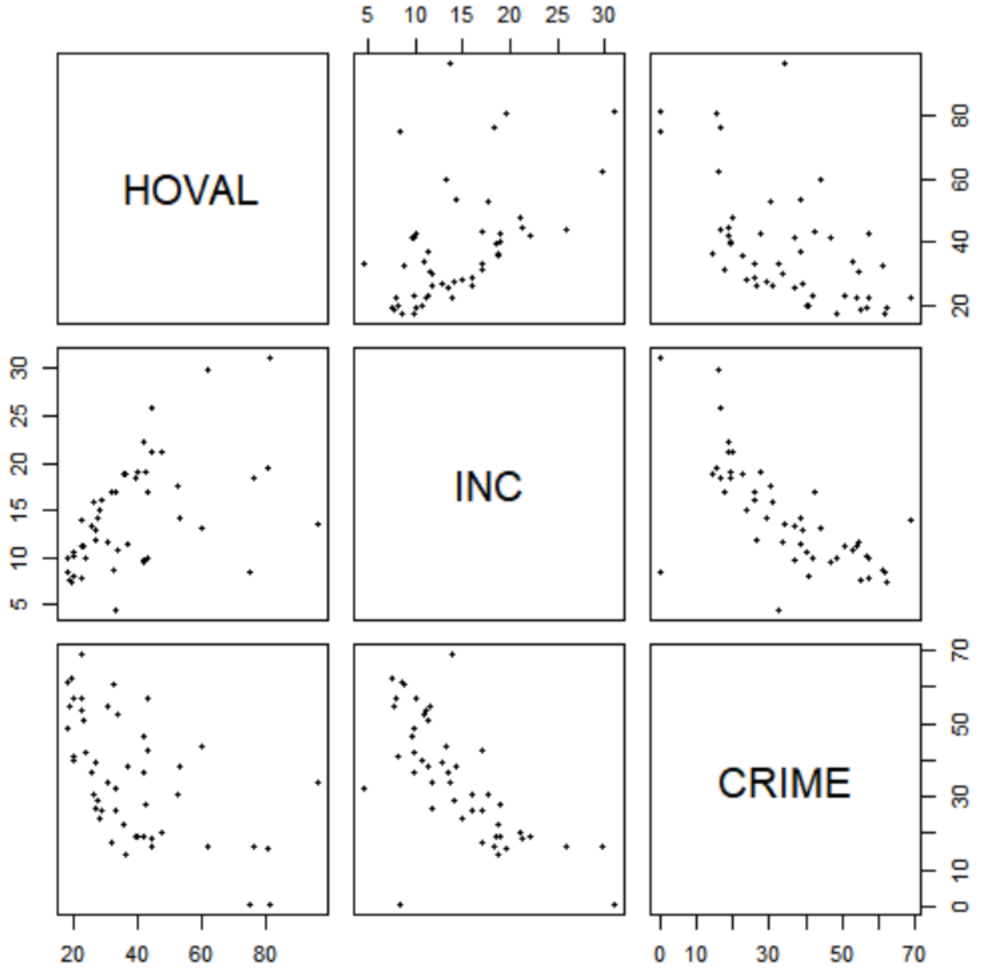
3. CRIME



Histogram: There is a peak around 20 instead of the mean of this set of data which is around 35. And there is a sharp drop besides 20. The data is not particularly concentrated. However, except for the highest bar, the trend of the rest data is like normal distribution and approximately asymmetrical.
Boxplot: There is no suspected outliers above the third quartile. Furthermore, the maximum value and minimum value are symmetric about the median. However, the length of the box is the longest, which means the dispersion of this data is the largest.
QQ plot: Some points are above the line while some of them are below the line. This

also shows this set of data is asymmetric. The rest part of the data is just on the line. So, this set of data is approximately normally distributed.

TASK3: Scatterplot matrix of the three variables



This picture shows the correlation between the two. CRIME and INC have the strongest correlation and they are negatively correlated. HOVAL and INC have weak correlation and they are positive correlated. HOVAL and CRIME also have weak correlation and they are negative correlated.