**Experiment No.4**

**Title:** Applying and interpreting different plots

**Batch: A4 Roll No.:1914078 Experiment No.: 4**

**Aim:** Identifying suitable dataset for applying different plots and interpretation. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Resources needed:** Any programming language/ Rapid Miner, any data source (RDBMS/Excel/CSV)

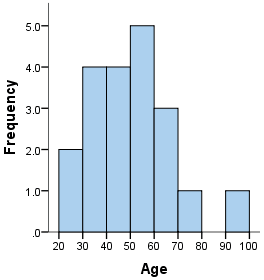
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Theory:**

For data preprocessing to be successful, it is essential to have an overall picture of your data. Basic statistical descriptions alone cannot be used to identify properties of the data and highlight which data values should be treated as noise or outliers. The plots such as Box Plot, Q-Q Plot, Histogram and Scatterplots provide various information to the data analyst. Data visualization is very much needed because a visual summary of information makes it easier to identify patterns and trends than looking through thousands of rows. Before applying plots suitability of the attribute should be checked.

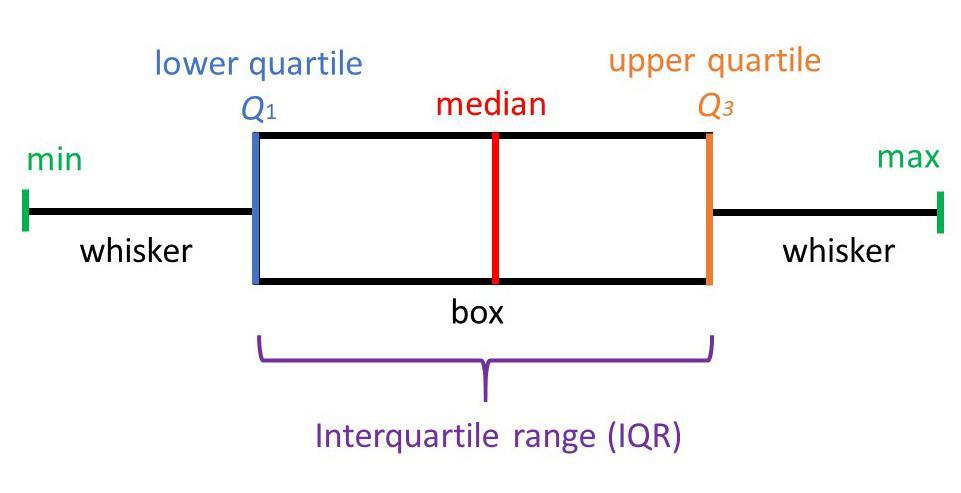
**Histogram**

Histogram gives accurate representation of the distribution of numeric data. A histogram is a chart that shows frequencies for intervals of values of a continuous variable. It summarize a Univariate Data set​. In histogram of a continuous frequency table, x-axis marks class intervals on a suitable scale and y-axis marks frequency of each class interval. The interval of value is known as bin and they all have the same widths.​ The upper and lower class limits of the new exclusive type classes are known as class boundaries. Histograms also give us much more complete information about our data. ​



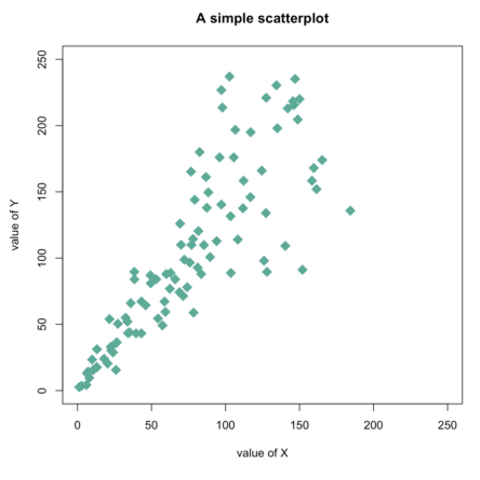
**Box plot**

Boxplot also known as box-and-whisker plot is a way to show the distribution of values based on the five-number summary: minimum, first quartile, median, third quartile, and maximum. The minimum and the maximum are just the min and max values from the data set. The median is the value that separates the higher half of a data from the lower half. The first quartile is the median of the data values to the left of the median in our ordered values. The third quartile is the median of the data values to the right of the median in our ordered values. Boxplot can also show outliers and IQR(Inter Quartile range) .



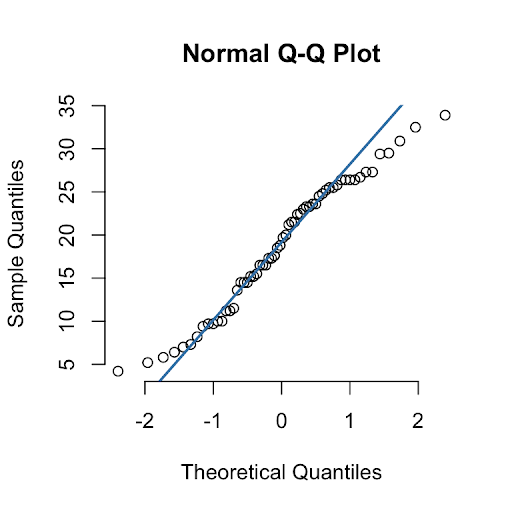
**Scatterplots**

A scatter plot is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data.​ A scatter plot can be used either when one continuous variable that is under the control of the experimenter and the other depends on it or when both continuous variables are independent. A scatter plot can suggest various kinds of correlations between variables with a certain confidence interval.



**Quantile-Quantile Plot**

A Q–Q (quantile-quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Many distributional aspects can be obtained from a q-q plot like, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. It  helps to assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential.  ​



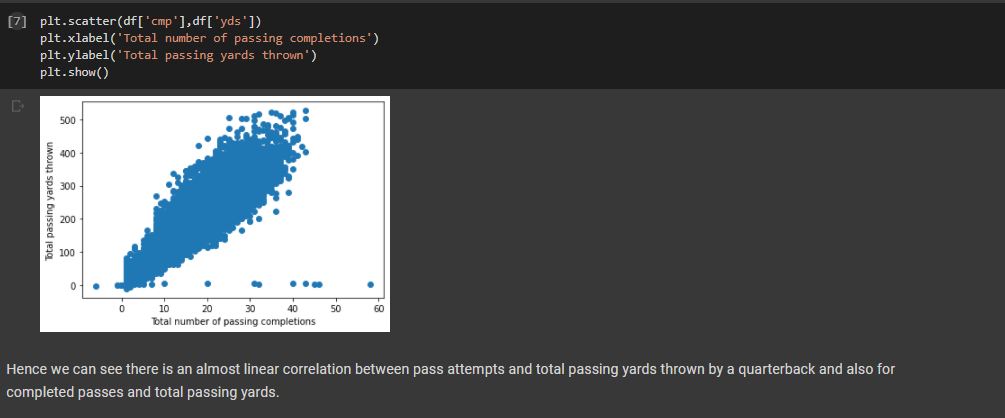
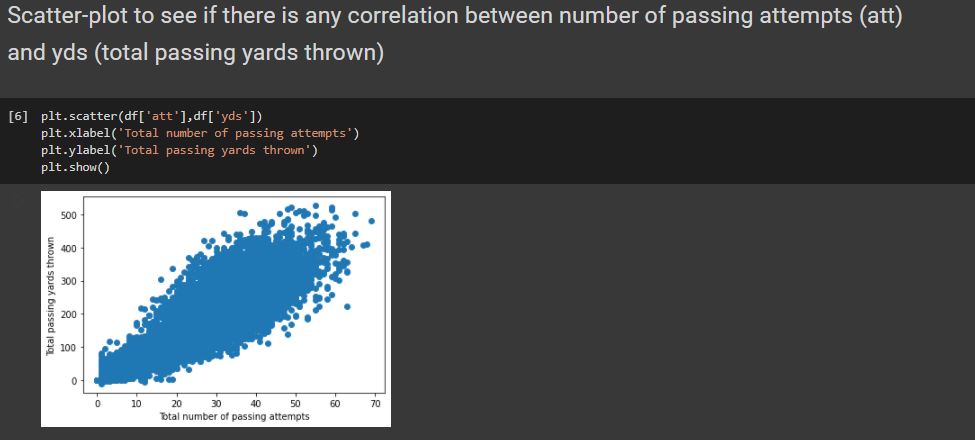
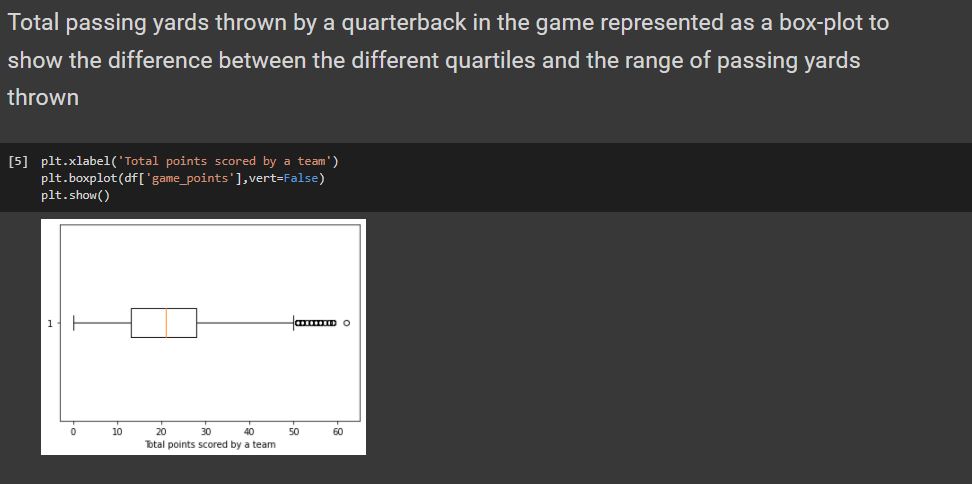
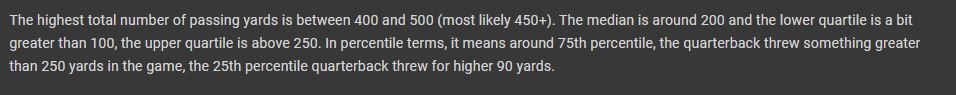
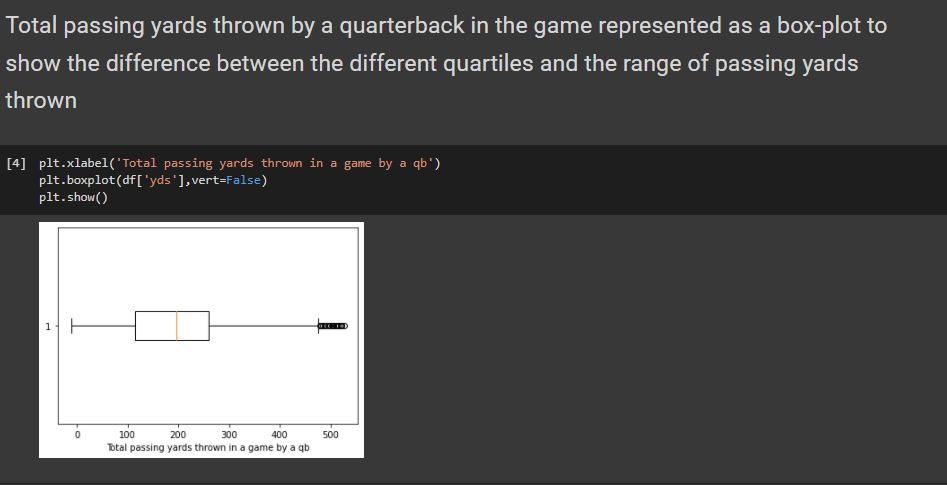
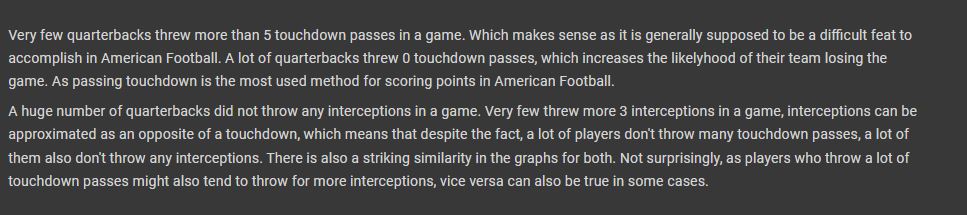
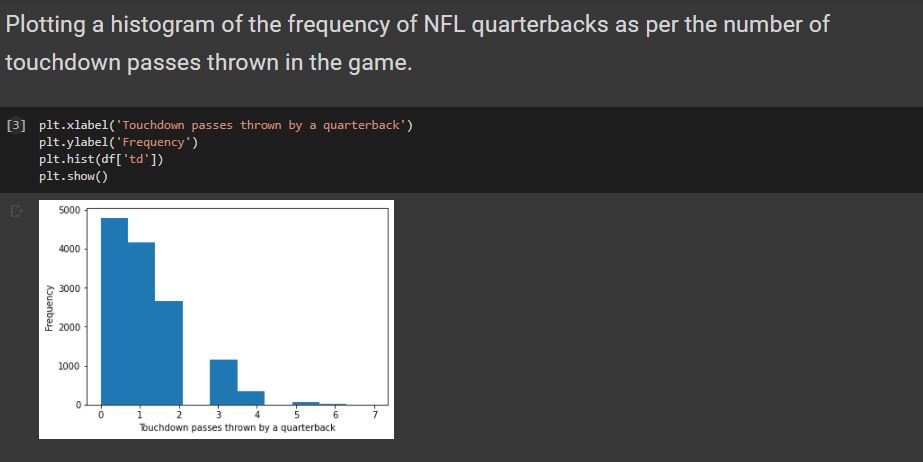
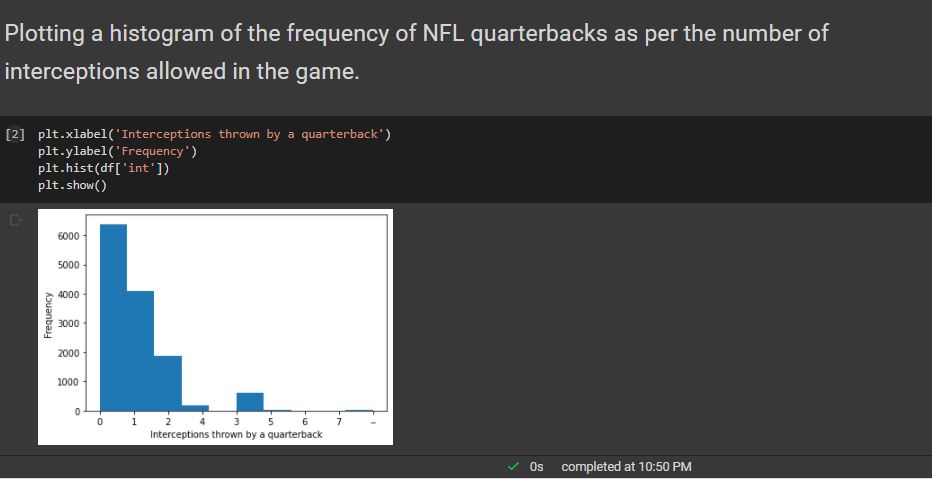
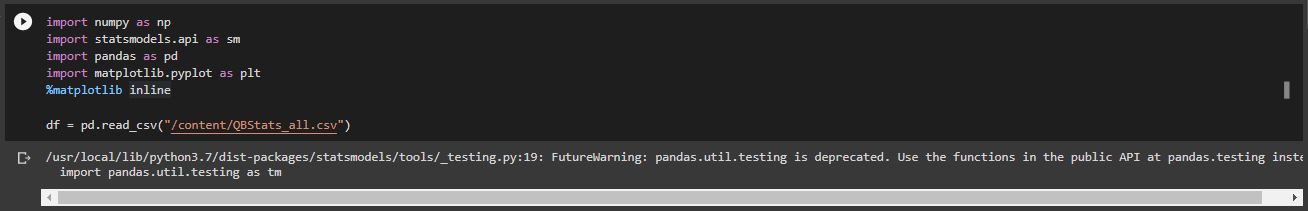
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

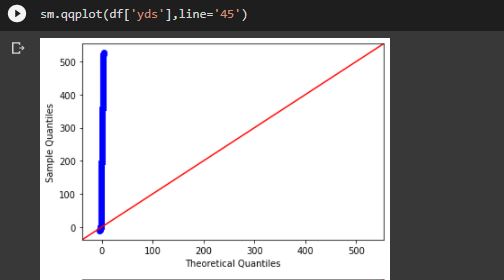
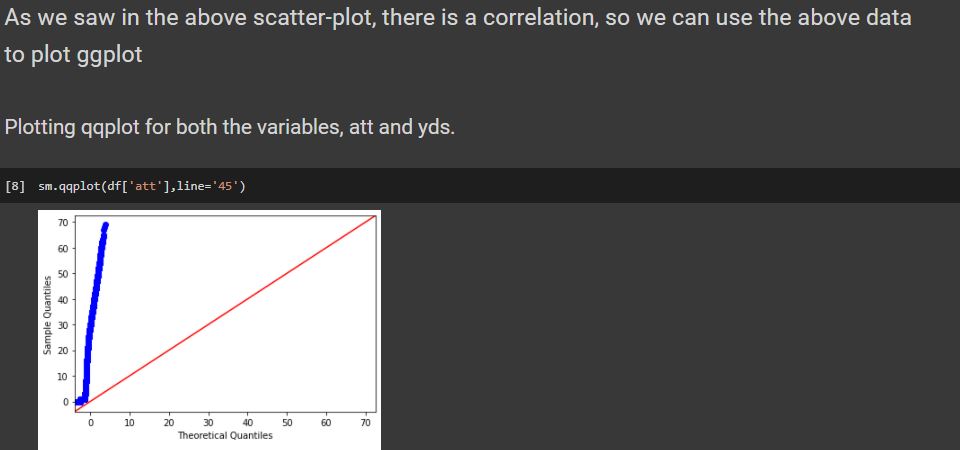
**Procedure / Approach /Algorithm / Activity Diagram:**

1. Identify the attributes where it will be sensible to apply the below given plots.  
   1. Box Plot
   2. Q Q Plot
   3. Histogram
   4. Scatter Plot

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Results: (Program printout with output / Document printout as per the format)**

****

****

**Questions:**

1. **Why is it important to measure the dispersion in the dataset?**

While measures of central tendency are used to estimate "normal" values of a dataset, measures of dispersion are important for describing the spread of the data, or its variation around a central value.

Two distinct samples may have the same mean or median, but completely different levels of variability, or vice versa.

A proper description of a set of data should include both of these characteristics.

There are various methods that can be used to measure the dispersion of a dataset, each with its own set of advantages and disadvantages.

1. **Discuss the other purposes/advantages of the plots used in this experiment.**

1. **Box plot**   
     
   A box plot is a way of summarizing a set of data measured on an interval scale. It is often used in explanatory data analysis. This type of graph is used to show the shape of the distribution, its central value, and its variability.  
     
   Advantages:

* Graphically display a variable's location and spread at a glance.
* Provide some indication of the data's symmetry and skewness.
* Unlike many other methods of data display, boxplots show outliers.
* By using a boxplot for each categorical variable side-by-side on the same graph, one quickly can compare data sets.

**Q-Q plot**  
  
The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential.

Advantages:

* The sample sizes do not need to be equal.
* Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.

**Histogram**

A histogram is used to summarize discrete or continuous data. In other words, it provides a visual interpretation. This requires focusing on the main points, facts of numerical data by showing the number of data points that fall within a specified range of values (called “bins”). It is similar to a vertical bar graph.

The main advantages of a histogram are its simplicity and versatility. It can be used in many different situations to offer an insightful look at frequency distribution. For example, it can be used in sales and marketing to develop the most effective pricing plans and marketing campaigns.  
  
**Scatter plot**

Scatter plots’ primary uses are to observe and show relationships between two numeric variables. The dots in a scatter plot not only report the values of individual data points, but also patterns when the data are taken as a whole.

Advantages:

* + - It shows the relationship between two variables.
    - It is the best method to show you a non-linear pattern.
    - The range of data flow, i.e. maximum and minimum value, can be determined.
    - Observation and reading are straightforward.
    - Plotting the diagram is easy.

Overall advantages:

- Summarizes variation in large datasets visually. Shows outliers. Compares multiple distributions. Indicates symmetry and skewness to a degree

- Used in many different situations to offer an insightful look at frequency distribution.

- The sample sizes do not need to be equal. Many distributional aspects can be simultaneously tested.

**Outcomes:**  **CO2** - Comprehend descriptive and proximity measures of data

**Conclusion: (Conclusion to be based on the objectives and outcomes achieved)**

Applied data visualization and formed inferences and conclusions from various patterns obtained in each graph .Understood the significance of various plots like box plots, qq plot and histograms etc.

**Grade: AA / AB / BB / BC / CC / CD /DD**

Signature of faculty in-charge with date

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**References:**

Books/ Journals/ Websites:

1. Han, Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann 3nd Edition