DAV HOMEWORK 1

Titanic Dataset Analysis

This analysis is regarding the tragic incident of Titanic, Where many people were stuck in the ship and perished. The analysis will show some specific details using the Titanic dataset.

Problem 1. Proportions and Nested Proportions

Data observations and cleaning:

- Firstly, the survived column and Pclass column has been changed.
  1. Survived column – **0** was replaced with **No** and **1** was replaced with **Yes**
  2. Pclass column – **1**, **2** and **3** were replaced with **One**, **Two** and **Three**

  These changes are done for the better understanding of the data. The majority of the passengers were in third class, followed by first class and second class. More females survived than males.
- The Embarked column has C, S and Q, these are the station names initials. It might be confusing but it's basically Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton) The majority of the passengers embarked from Southampton, followed by Cherbourg and Queenstown.
- There are two more columns named sibsp - Number of Siblings/Spouses Aboard parch - Number of Parents/Children Aboard. These column states all the family travelling together, and it can be related to the age of the people, which helps to determine the age.

Proportions (Donut Chart) –

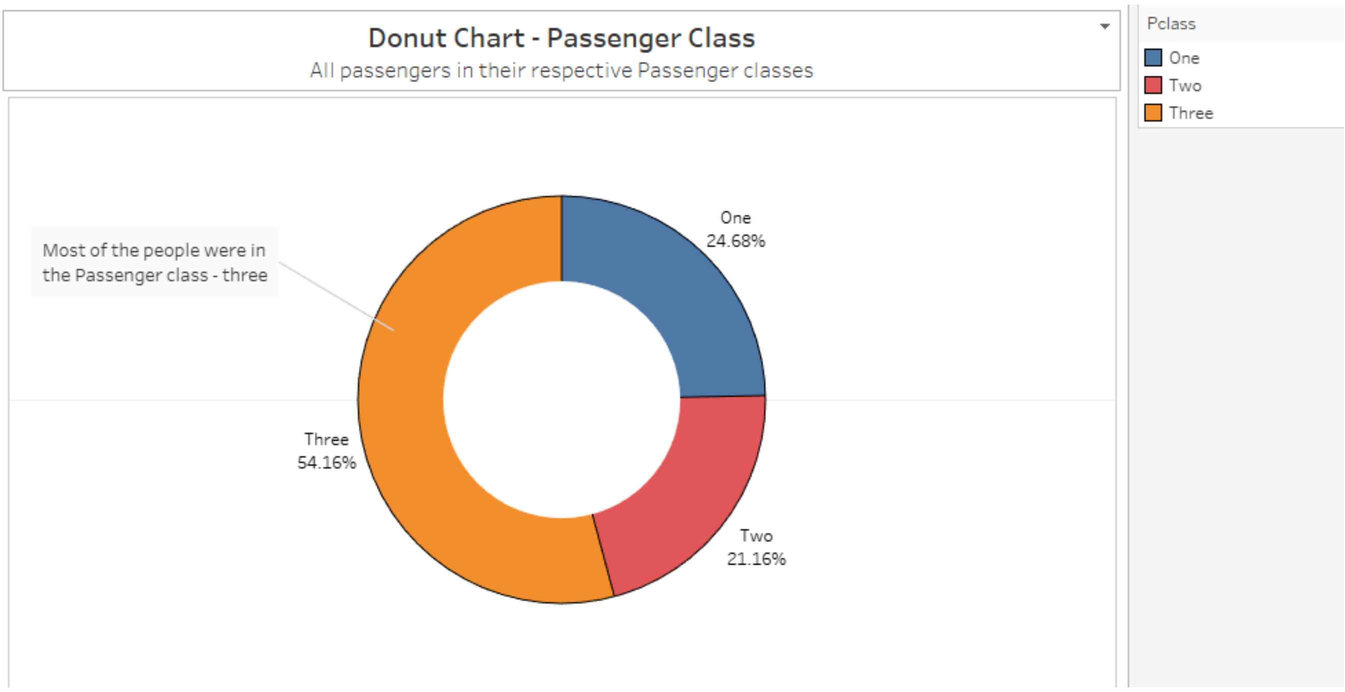

Figure 1: Passenger Class

- The Donut Chart shows all the passengers in their respective passenger class
- Majority of the passengers which is 54.16% were in Third class, whereas the second most passengers were in First class (24.68%), and lastly 21.16% passengers were in the Second class.
- The color are according to the class that is blue is one, pink is two and orange is three Pclass.

Nested Proportions-

- The nested proportions below will show the survival rates of the passengers according to their gender and passenger class.

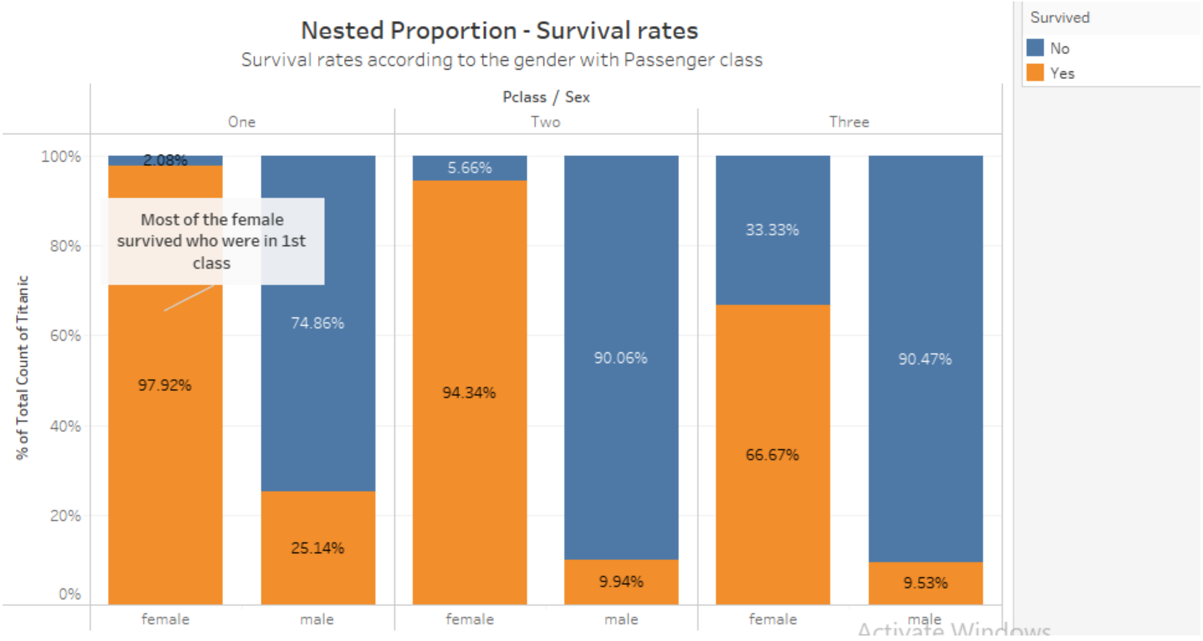

Figure 2: Survival rates of passengers

- From the above chart it is cleared that most of the female passengers survived
- Majority of the female who survived were in the First class which is 97.92%, also the female passengers in second class had a survival rate of 94.34%.
- This chart also gives information that most of the male who perished were in Third class (90.47%) followed by second class (90.06%).

Magnitude Chart -

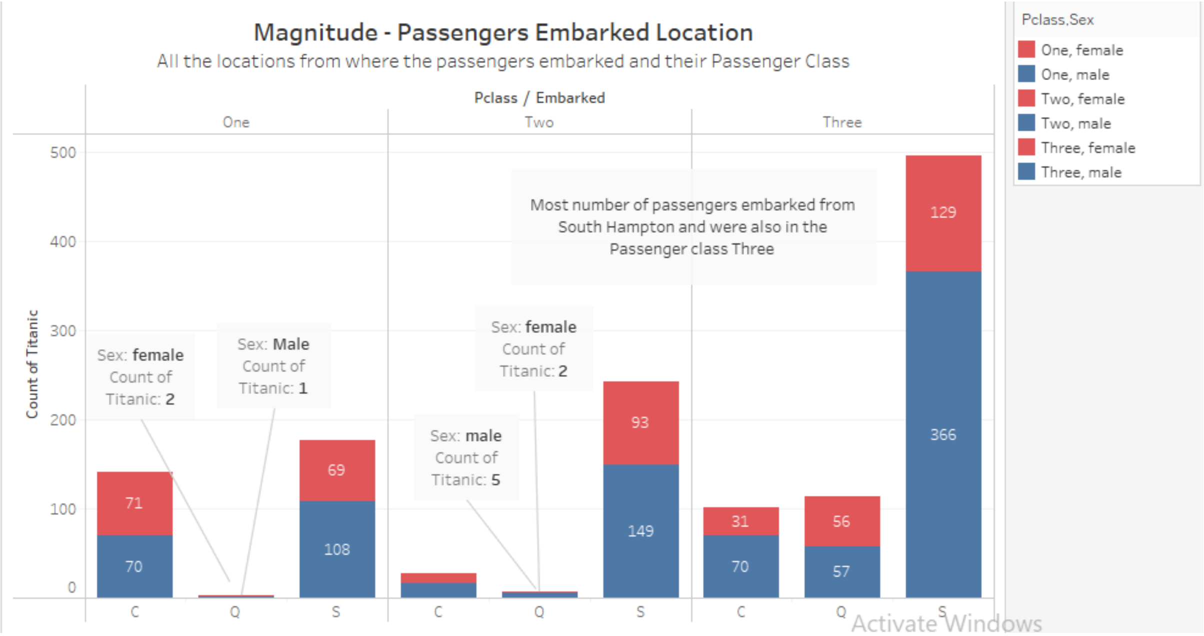

Figure 3: Passengers Embarked Locations and Passenger Class

- From the above chart, it's confirmed that majority of the passengers were in third class as well as most of the passengers embarked from South Hampton.
- There were only 3 people (2 female and 1 male) who embarked from Queenstown and were a first class passenger.
- Similarly, there were 7 people (5 male and 2 female) who embarked from Queenstown and were second class passengers.
- The main focus of the chart which shows a huge number of people who were in third class and embarked from South Hampton.

Problem 2. Distribution

Option 1. Use the titanic dataset you used in problem one and create
1. The distribution of age using a histogram (one chart)

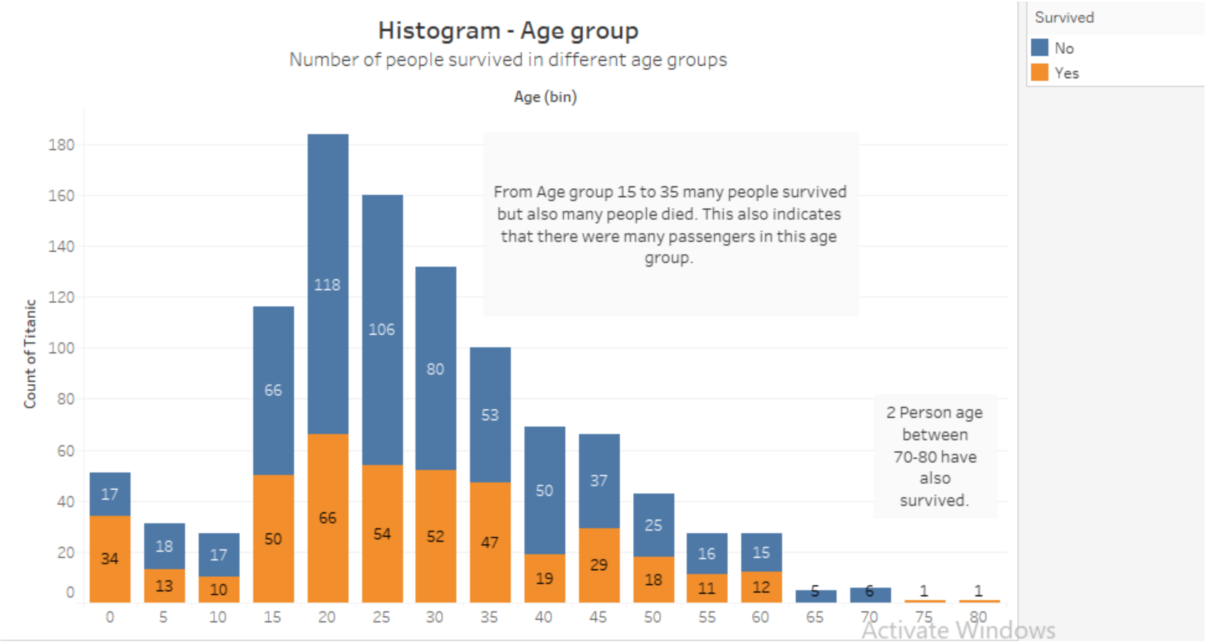

Figure 4: Survived passengers according to the age group

- The main focus of the histogram is that most of the passengers travelling in Titanic were young people.
- The passengers from the age group of 15 to 45 were more likely to be young people, where 184 (66 + 118) passengers are in the age group of 20-24.
- Most of the Passengers who died were young, and in the age group of 15 to 45.
- One of the most noticeable thing is that there were 2 people who were in the age group of 70-80 and they both survived.

2. Cumulative histograms of the age variable (one chart)

- The Cumulative Histogram below shows the sum of age.

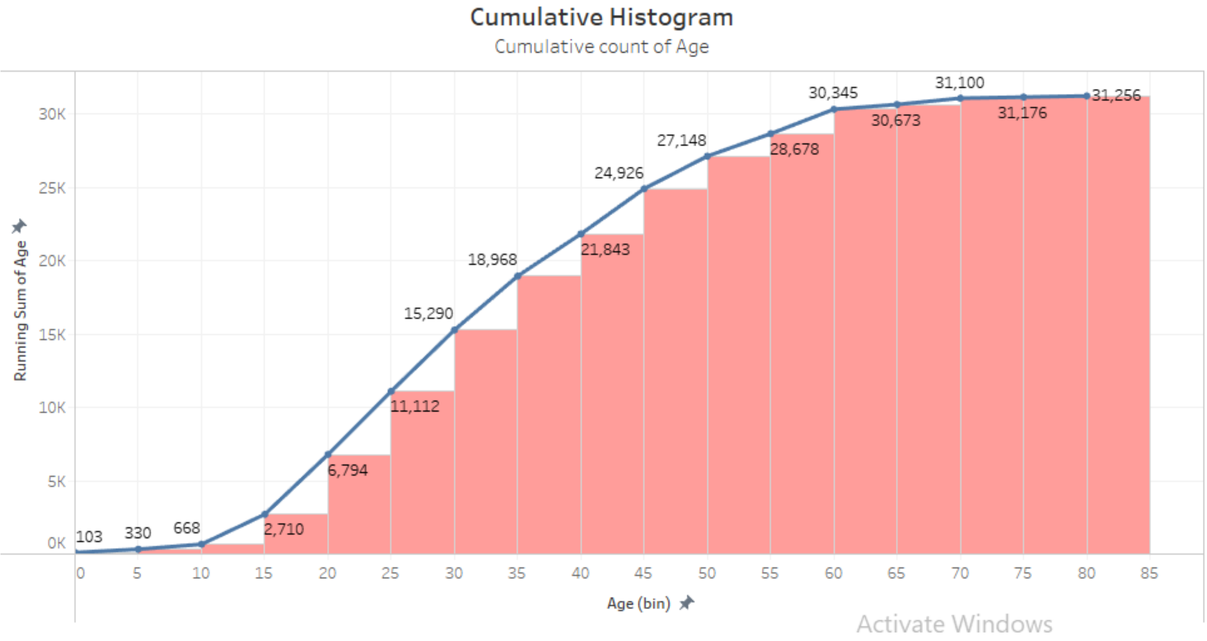

Figure 5: Cumulative Histogram

- The chart shows the age bins and the sum of the age making a total cumulative of 31,256.
- The cumulation started from 2710 and went up till 31,256.
- From observing the cumulative histogram with the age variable in the Titanic dataset, we can see there is a gradual increase in the number of passengers from 30 years old, followed by a steep increase in the number of passengers from 40 years old.
- Overall, the cumulative histogram with the age variable provides valuable insight into the age distribution of passengers on the Titanic and can be useful in understanding the demographics of the passengers who were on board the ship.

3. Compare the distribution of male and female passengers (one chart)

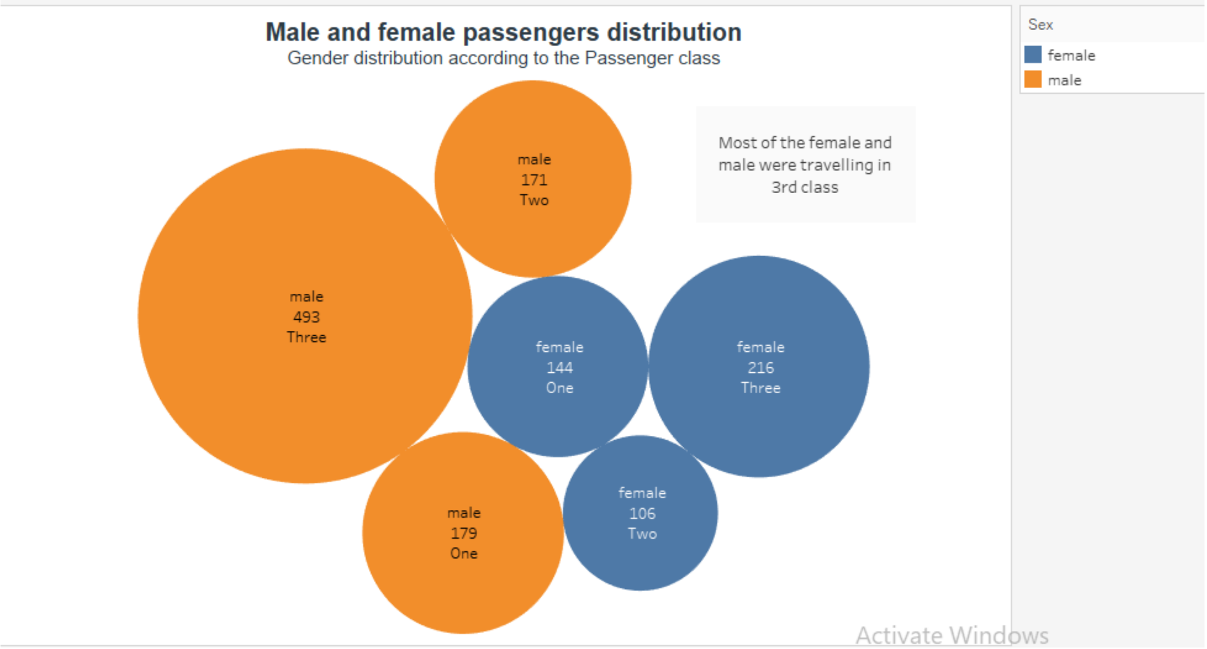

Figure 6: Male and Female Passenger distribution

- In the above chart, orange represents male and blue represents female. Also, the numbers are the count of those passengers and the chart also shows the Passenger class according to the male and female distribution.
- Majority of the passengers were male and also were in passenger class Three.
- Similarly, majority of the female were in passenger class Three.

Problem 3. Design and Redesign –

Scatter Plot-

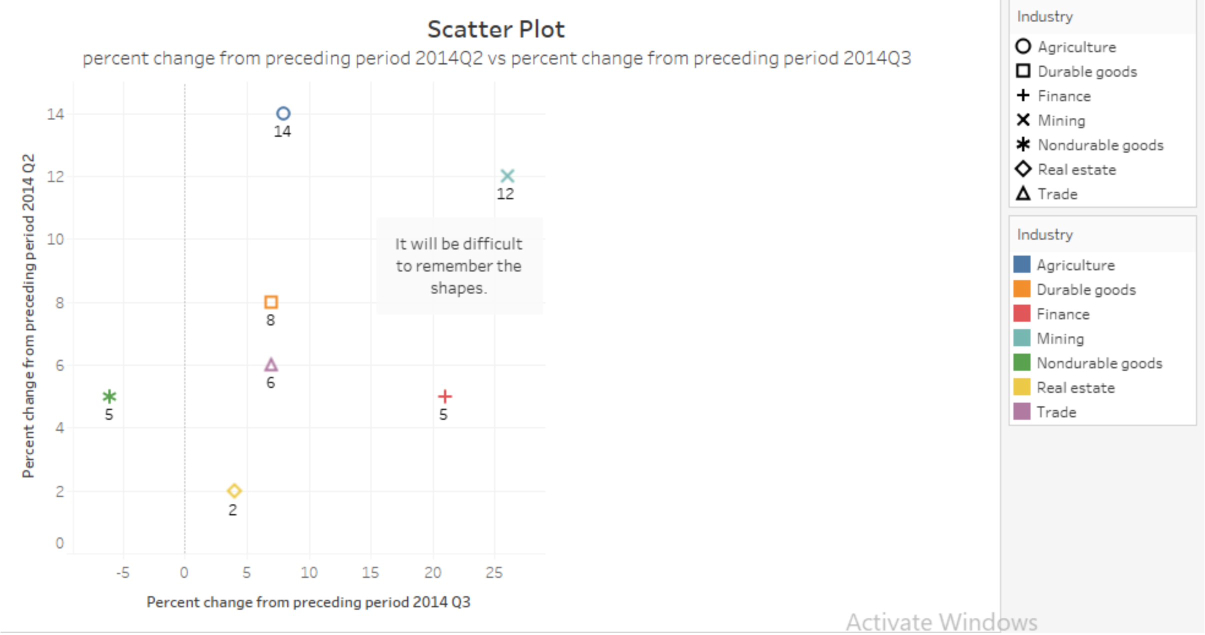

Figure 7: Scatter Plot

- The above scatter plot does not clearly depict the Q2 and Q3 of the year 2014 according to the industry.
- It will be difficult to understand this scatter plot as the viewer has to remember the colour and the shape of everything.
- One has to study the above graph to detail in order to figure out the trends.

Horizontal Bar Chart –

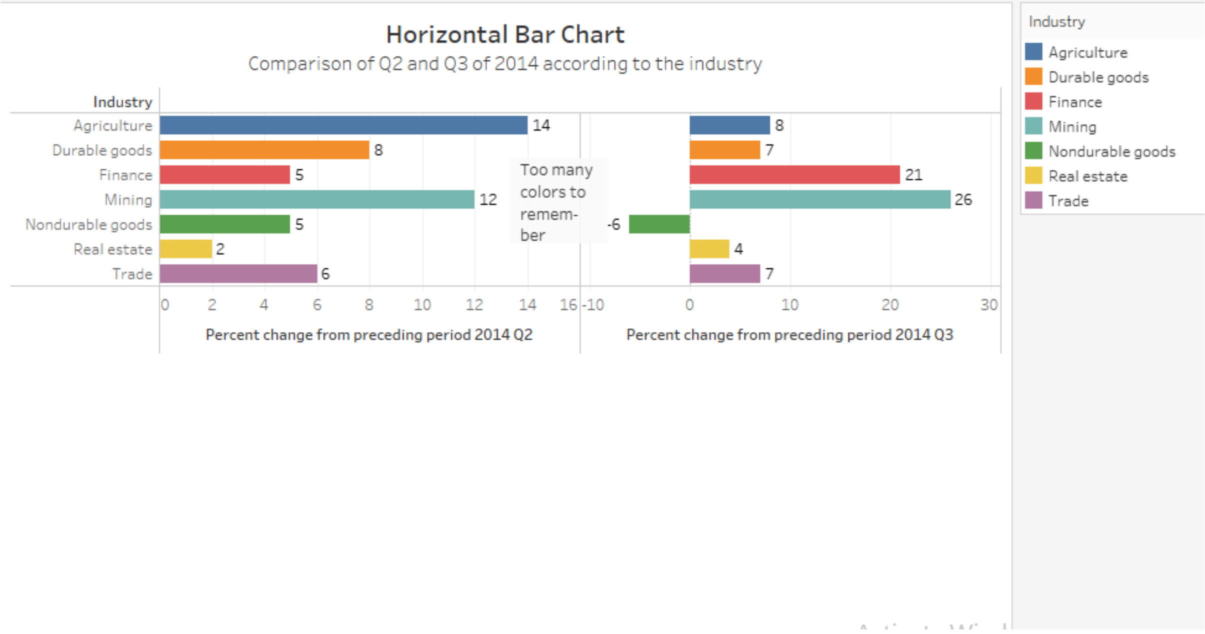

Figure 8: Horizontal Bar Chart

- The Horizontal Bar chart is comparatively easy to understand than the above scatter plot.
- This chart is easy to understand all the values but there are too many colors in this chart
- The viewer has to remember the color to make it easy for them to understand the difference in Quarter 2 and 3 of the year 2014 according to the industry.

Tree Map –

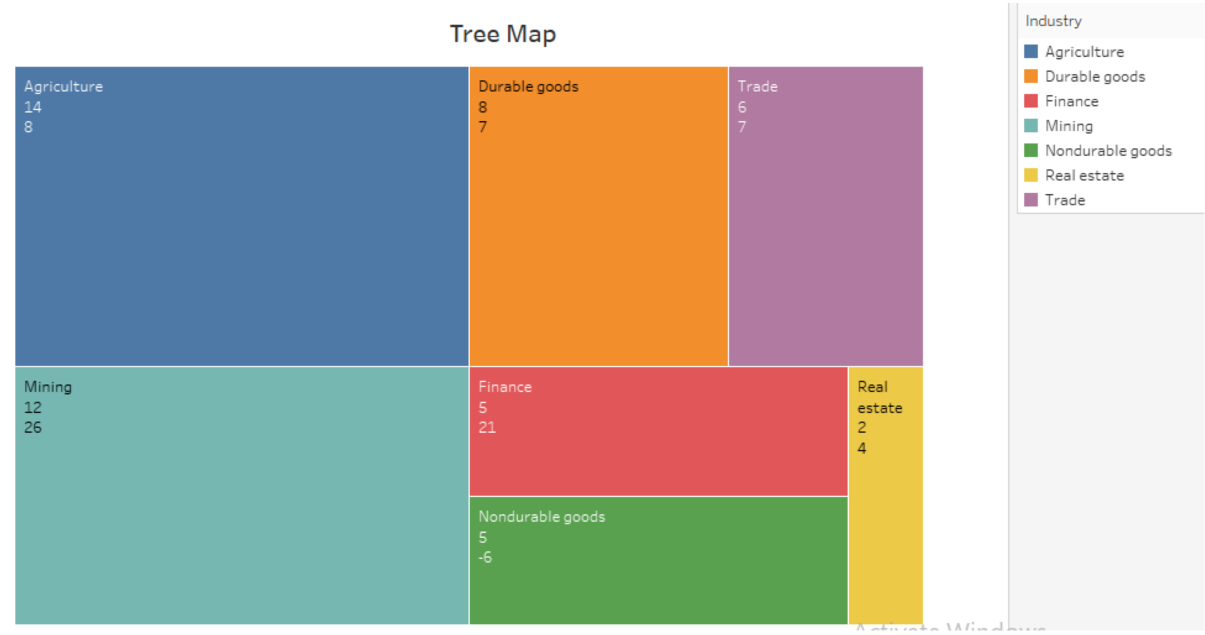

Figure 9: Tree Map

- In the above Tree Map, all the industry attributes whose values are close have similar size which makes it difficult to interpret the data.
- The values are close to one another which makes it difficult to find the trends.

- Comparing with the above Horizontal Bar Chart and Scatter plot, Tree map is a better option.

Scatter Plot (Circle) –

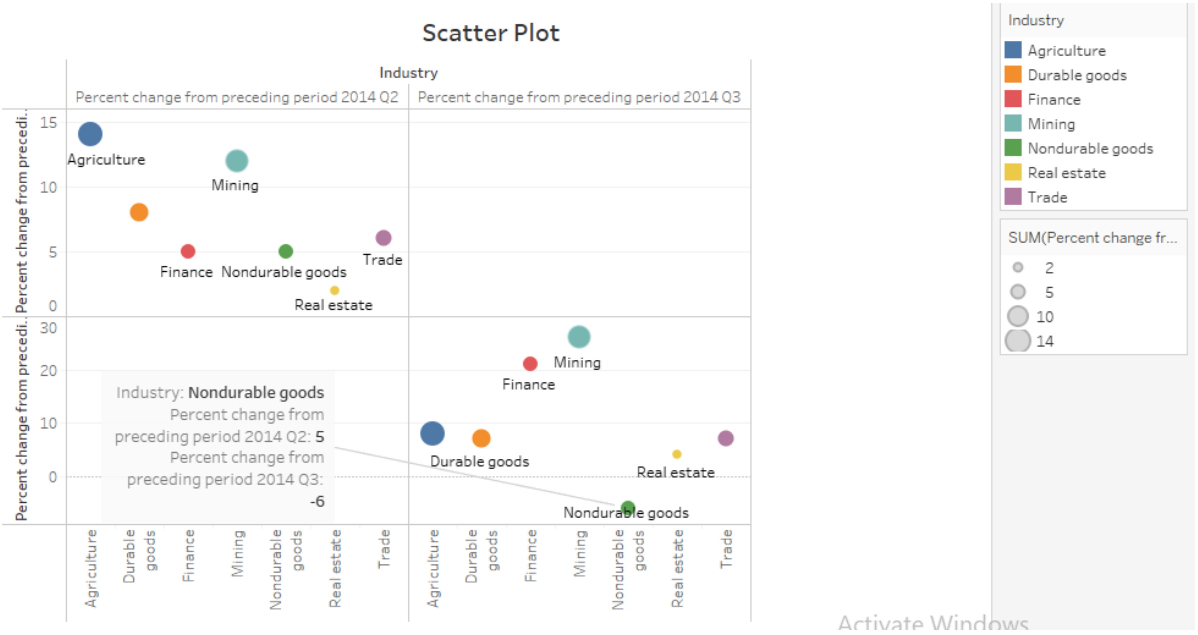

Figure 10: Scatter Plot

- In the above plot, everything is properly displayed compared to the other charts. It is easier to identify trends and figure out the maximum and minimum values displayed.
- As a viewer, the nondurable goods value declines in Q3 of the year 2014.
- It is also easier to figure out quarterly growths industry wise.