DATA ANALYSIS AND VISUALIZATION

About the dataset

→ Titanic dataset basically contains the information on the passengers aboard the RMS Titanic.

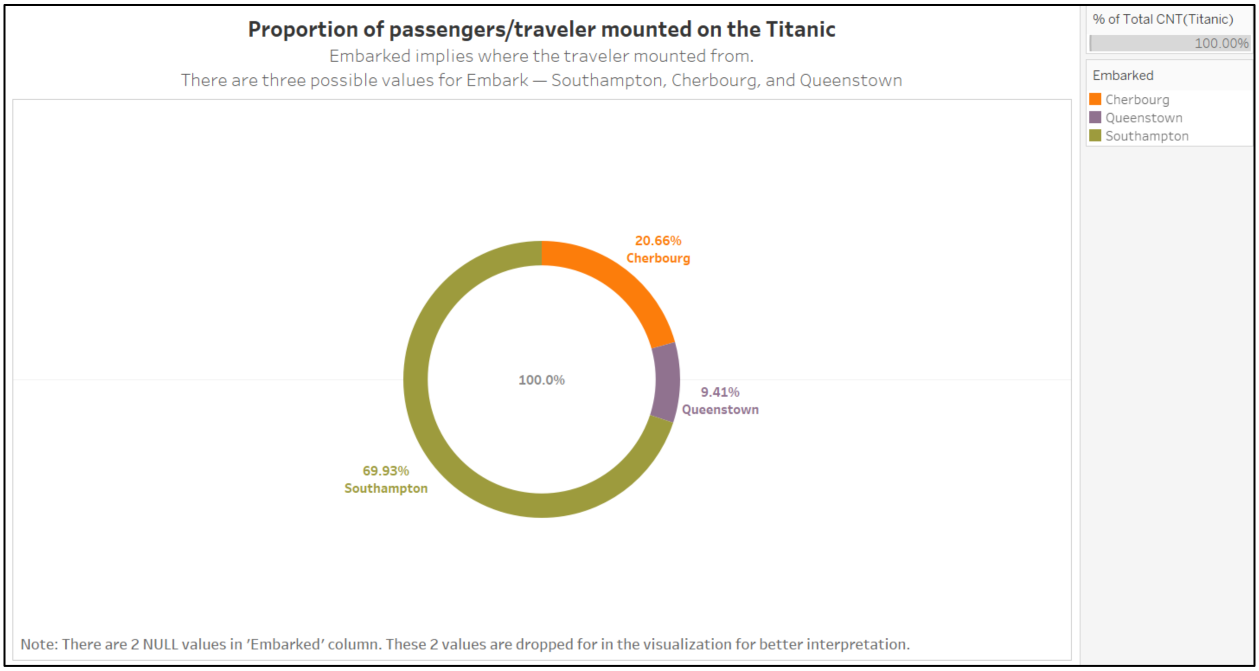
→ This dataset includes various attributes of passengers such as their age, gender, fair paid, passenger class, survival status and others.

→ Data observations:

- Embarked column has 2 NULL values.
- Age has 263 NULL values.
- Fare has only 1 NULL values.
- Cabin has 1014 NULL values.

Problem – 1

→ Proportion Chart


Proportion of passengers/traveler mounted on the Titanic

Embarked implies where the traveler mounted from.

There are three possible values for Embark — Southampton, Cherbourg, and Queenstown

Note: There are 2 NULL values in 'Embarked' column. These 2 values are dropped for in the visualization for better interpretation.

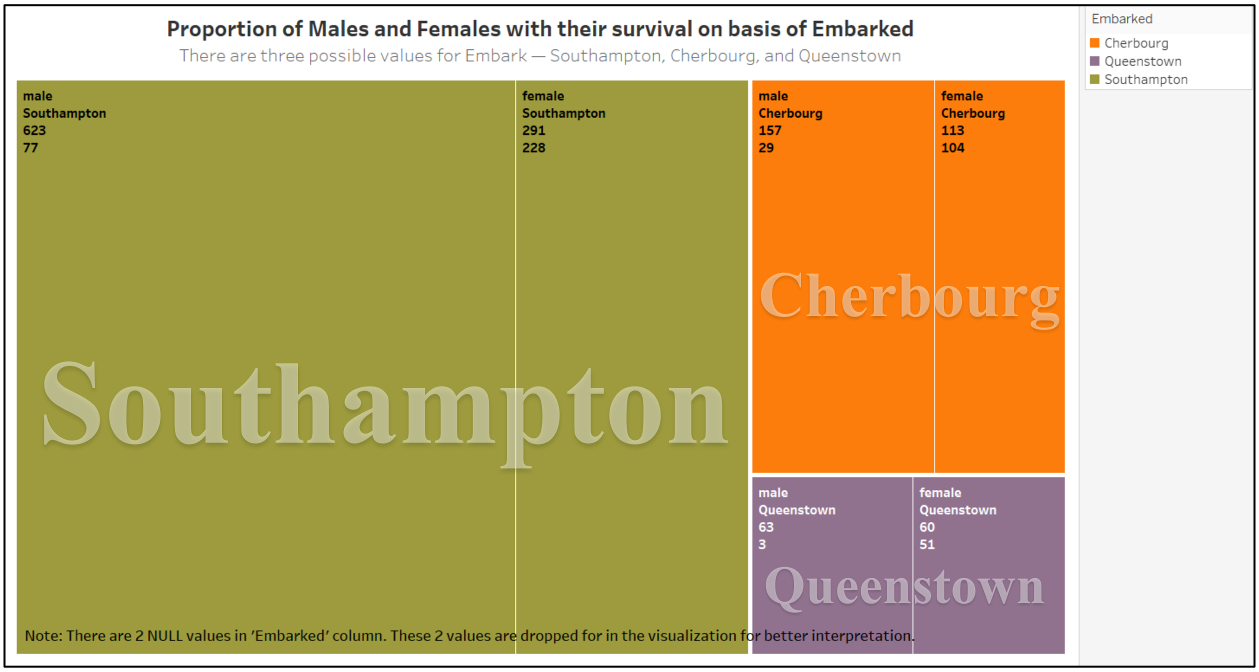
→ This chart shows the proportion of passengers/traveler mounted on the Titanic.

→ From this donut chart we can infer that, 69.93% of passengers boarded the Titanic from Southampton, 20.66% from Cherbourg, and 9.41% from Queenstown.

→ There were 2 NULL values in the Embarked that has been dropped in the above visualization for better interpretation.

→ We can say using the above chart that, most people i.e., almost 70% of passengers were mounted on Titanic from Southampton region.

→ <u>Nested Proportion Chart</u>



→ The above chart is the nested proportion chart.

→ It shows the proportion of males and females with their survival on basis of Embarked.

→ We can observe that, from all the three embarked, males were more in number, and females were less.

→ But surprisingly, we can see that, number of females survived were more than the males for all the three regions.

→ For Southampton, number of males were 623 and number of females were 291, while the survival of male was only 77 out of 623, but for female it was 228 out of 291.

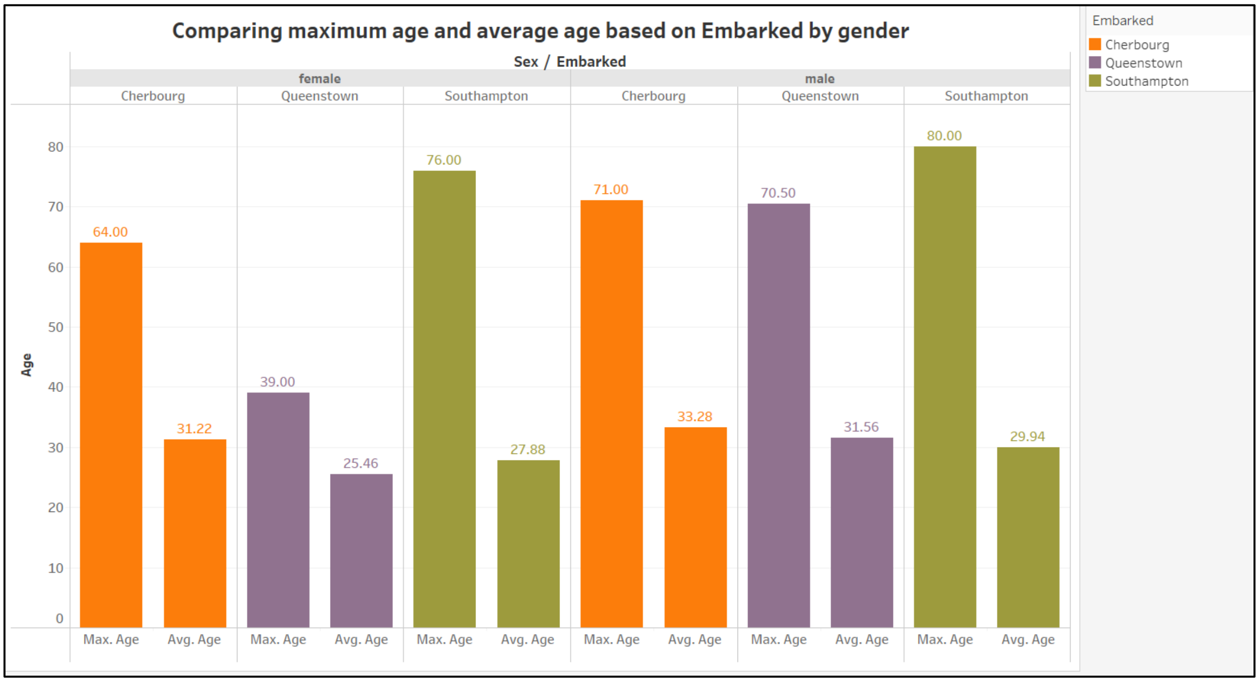
→ For Cherbourg, number of males were 157 and number of females were 113, while the survival of male was only 29 out of 157, but for female it stood to 104 out of 113.

→ Similarly, for Queenstown, the number of males were 63 and out of it only 3 survived while number of females were 60 and 51 of them survived for this region.

→ In a nutshell, we can say that a greater number of females survived in terms of number than males for all the three regions.

→ Just for note, there were 2 NULL values in Embarked, and both of them were dropped for better interpretation.

→ Magnitude Chart



→ The above chart is the magnitude chart.

→ This chart shows the comparison of maximum age and average age based on Embarked by gender.

→ We can view the maximum age and average age of males and females for the three specified regions, mainly (Southampton, Cherbourg, and Queenstown).

→ For females, the maximum age for Cherbourg, Queenstown, and Southampton is 64, 39, and 76, while their average stood at 31.22, 25.46, and 27.88, respectively.

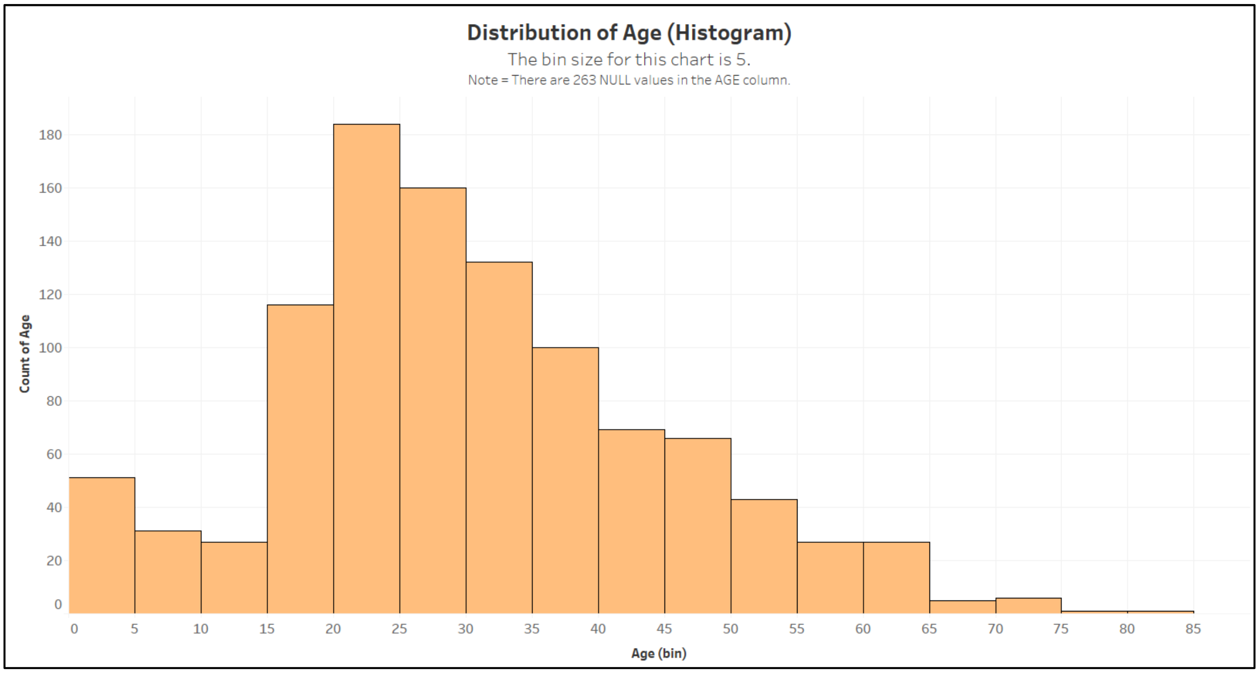
→ Similar to it, male's maximum age for Cherbourg, Queenstown, and Southampton is 71, 70.50, and 80, while their average were 33.28, 31.56, and 29.94, respectively.

→ This clearly shows that, for all three regions average age of males boarded on Titanic were higher than females. Same for maximum age, the oldest person for these regions were male.

→ This chart also had 2 NULL values in Embarked, and both of them were dropped for better interpretation.

Problem – 2

→ Distribution of Age (Histogram)



→ The above chart shows the distribution of age using the histogram.

→ We had set the bin size for this histogram to be 5.

→ There are 263 NULL values in the age variable.

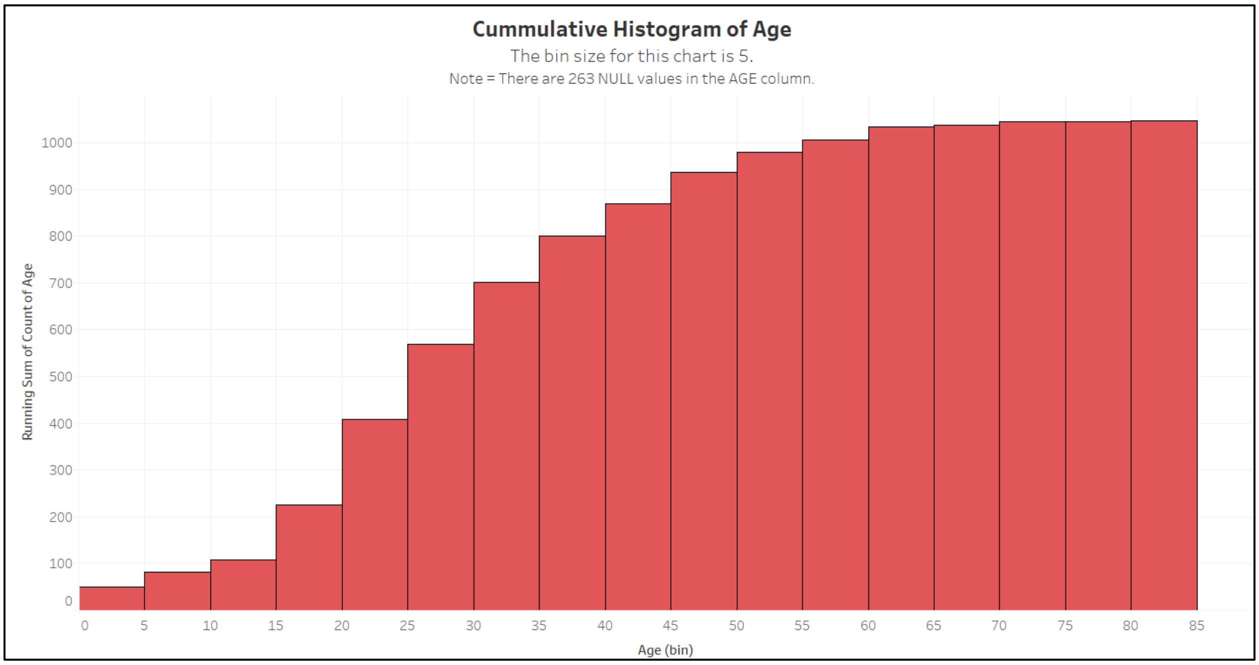
→ If we compare our distribution with normal distribution, it seems that we got a right-skewed distribution.

→ As we know that, for right-skewed histogram, mean > median > mode. Therefore in our case, it is also true because mean of age is 29.88 which is greater than median of age, 28 which is greater than mode of age, 24.

→ Our peak of the distribution of age is at the Age (bin) 20.

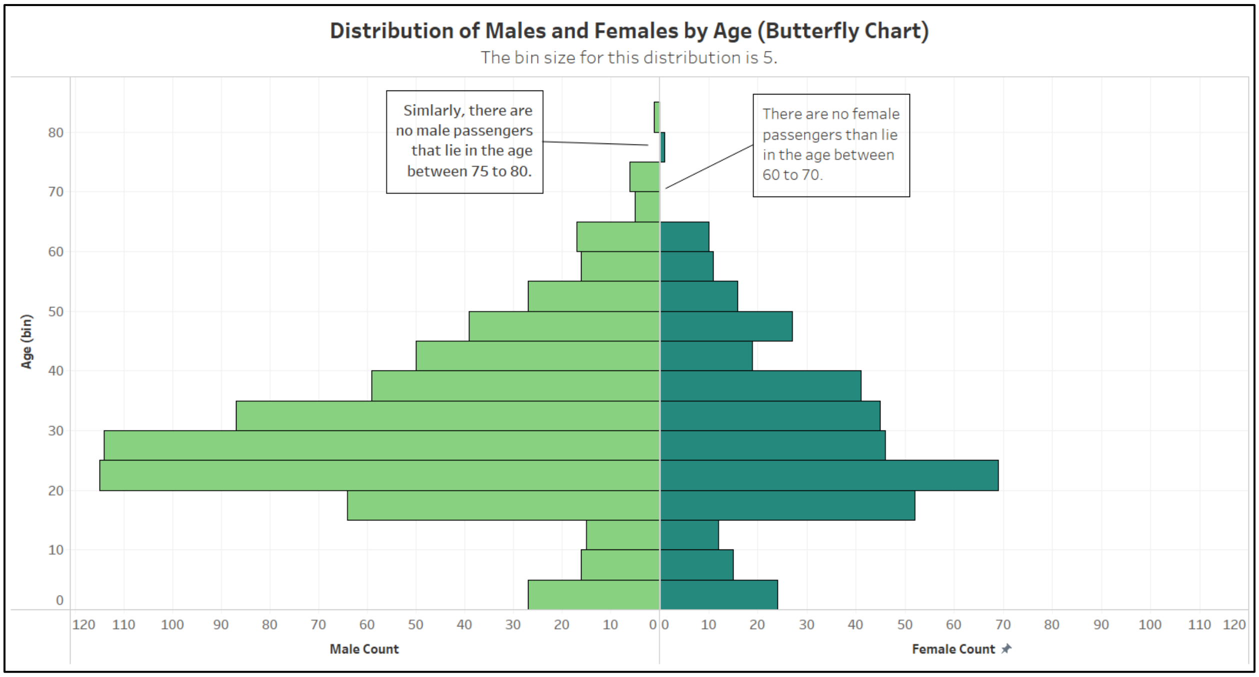
→ We can conclude from this chart that, most of the people aboard on Titanic were between the age 15 to 40.

→ Cumulative histogram of Age



→ The above chart shows the cumulative histogram of age.
→ The bin size for this chart is set to be 5.
→ There are 263 NULL values in the age column.
→ From this chart, we can see that the cumulative curve is concave up.
→ Initially the cumulative curve is concave up, later it straightens outs and appears linear in the middle and finally, it turns concave down, which indicates that the number of people of age of 65 and above, are less.

→ <u>Comparison of distribution of Male and Female Passengers</u>



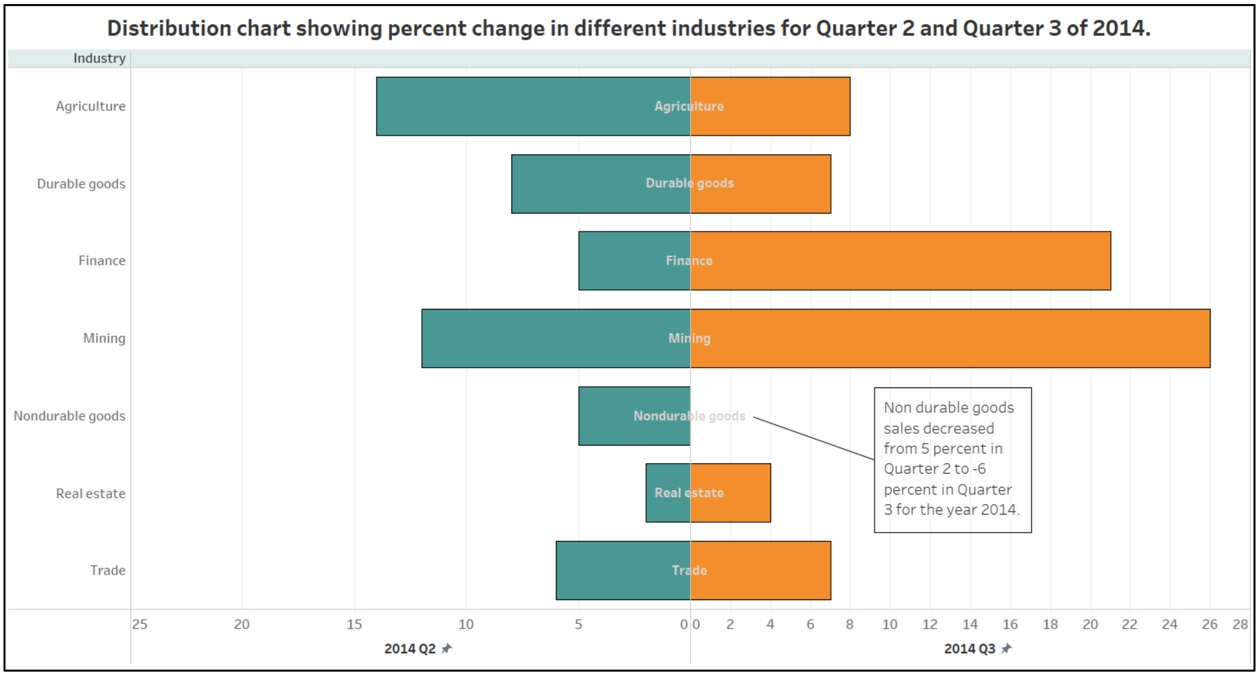
Simlarly, there are no male passengers that lie in the age between 75 to 80.

There are no female passengers than lie in the age between 60 to 70.

- → The above chart shows the distribution count of male and female passengers on the Titanic by age.
- → This chart lies in the category of distribution chart and is known as butterfly chart.
- → The bin size is set to 5 for this distribution.
- → The x-axis shows the male and female counts and y-axis shows the age with a bin of 5.
- → We can get various insights from this chart, and one of the basic inference is that, there were more number of males on the Titanic than females.
- → For the age 20 to 45, the number of males were more than the number of females, and it can be clearly seen using the chart.
- → For other ages, we can also see that the males were higher in number than females.
- → There are no female passengers that lie in the age between 60 to 70, similarly there are no male passengers that lie in the age between 75 to 80.

Problem – 3

→ Distribution Chart



Non durable goods sales decreased from 5 percent in Quarter 2 to -6 percent in Quarter 3 for the year 2014.

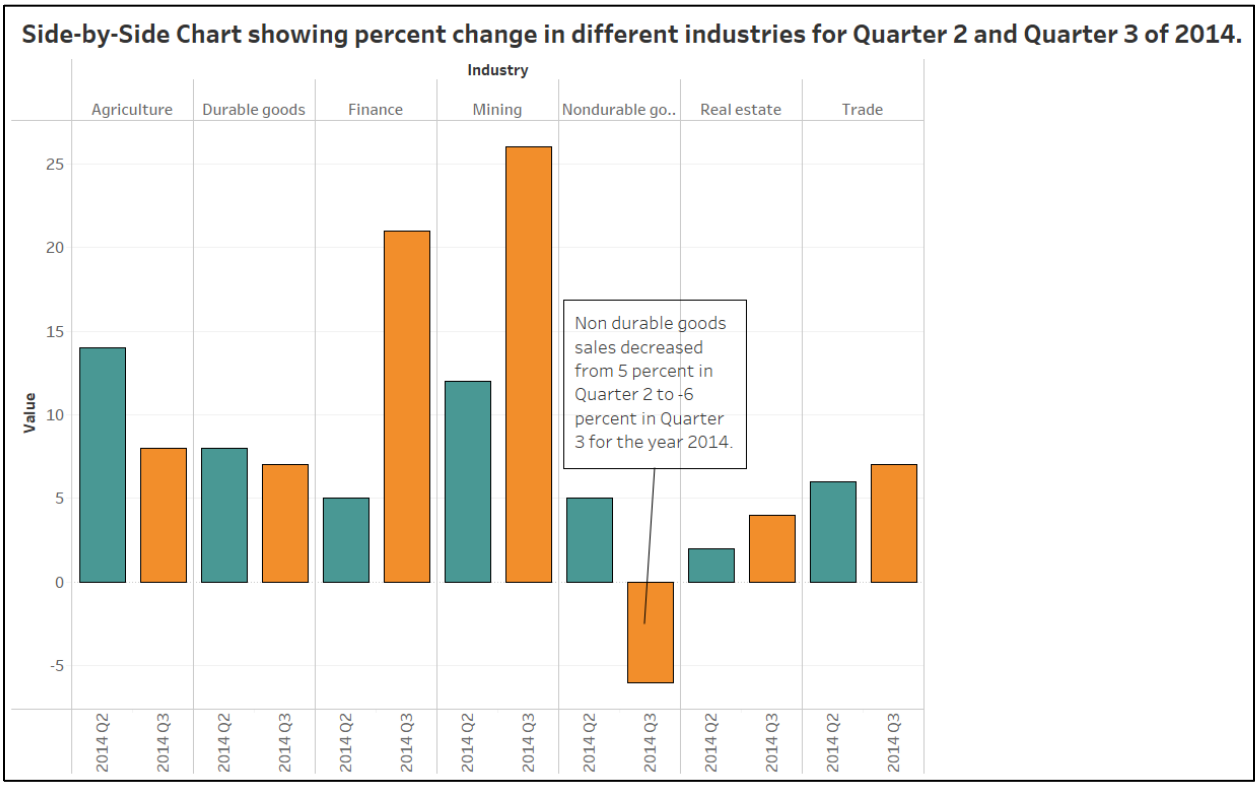
❖ Pros:

- → This chart helps to find the difference in the result of the Quarter 2 and Quarter 3 of 2014.
- → It helps us to compare the pattern side by side.
- → Example: We can easily see the difference of Mining industry for Quarter 2 and Quarter 3 of 2014.
- → It is easy to read and understand, even for those who are not familiar with statistics.
- → This chart helps us to find the comparative analysis, which can help us to find similarities and differences.

❖ Cons:

- → We can see that Nondurable goods has negative value, and we can’t see this value in this chart.
- → We can only use two parameters at a time, like in the above chart we were only able to use Industry and Quarter results at one time.

→ Side-by-Side Chart



❖ Pros:

→Side-by-Side chart is used to compare two different values using bar chart.

→For example, here we are comparing the different industry for their percent change for quarter 2 and quarter 3 of 2014.

→This chart is used to show multiple data sets in a single chart.

→Here, we can also see the negative value easily and read its value from y-axis.

❖ Cons:

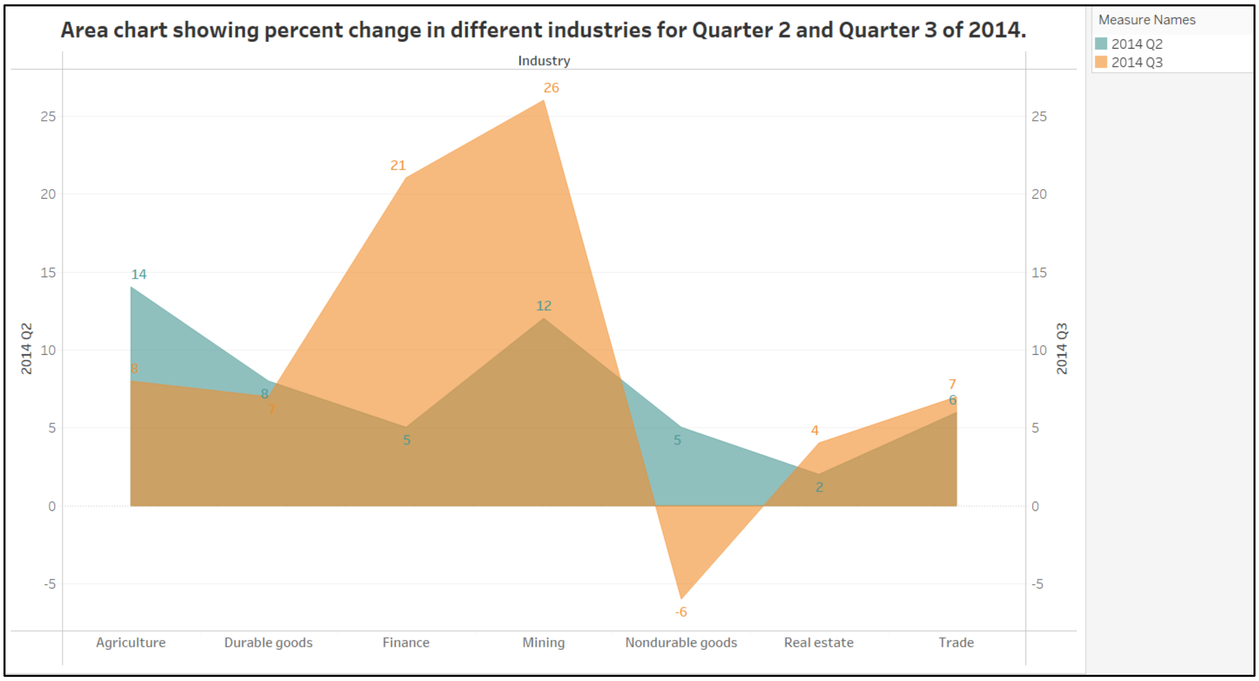
→It is difficult to compare two different sectors, if they are far from each other.

→For example, if we want to compare agriculture sector and real estate sector for quarter 3 of 2014, it is bit difficult to read it's percent change from y-axis.

→Since, we have only 7 industry in our dataset, it is easy for us to read the graph. But, if we would have more than 10 industries, it would become cluttered and difficult to read.

→Here we are using two colors to distinguish quarter 2 and quarter 3, but we need to be aware of the color used to make it accessible for color-blind people.

→ <u>Area Chart</u>



Area chart showing percent change in different industries for Quarter 2 and Quarter 3 of 2014.

❖ Pros:

→ Area chart is used to display as an area between the x-axis and the line or curve representing the data.

→ This chart is useful to show the change in data over time.

→ In our case, we can easily represent our data for each quarter of 2014 for different industry.

→ We can also represent and show the negative value, i.e., we can see the negative value of nondurable goods easily.

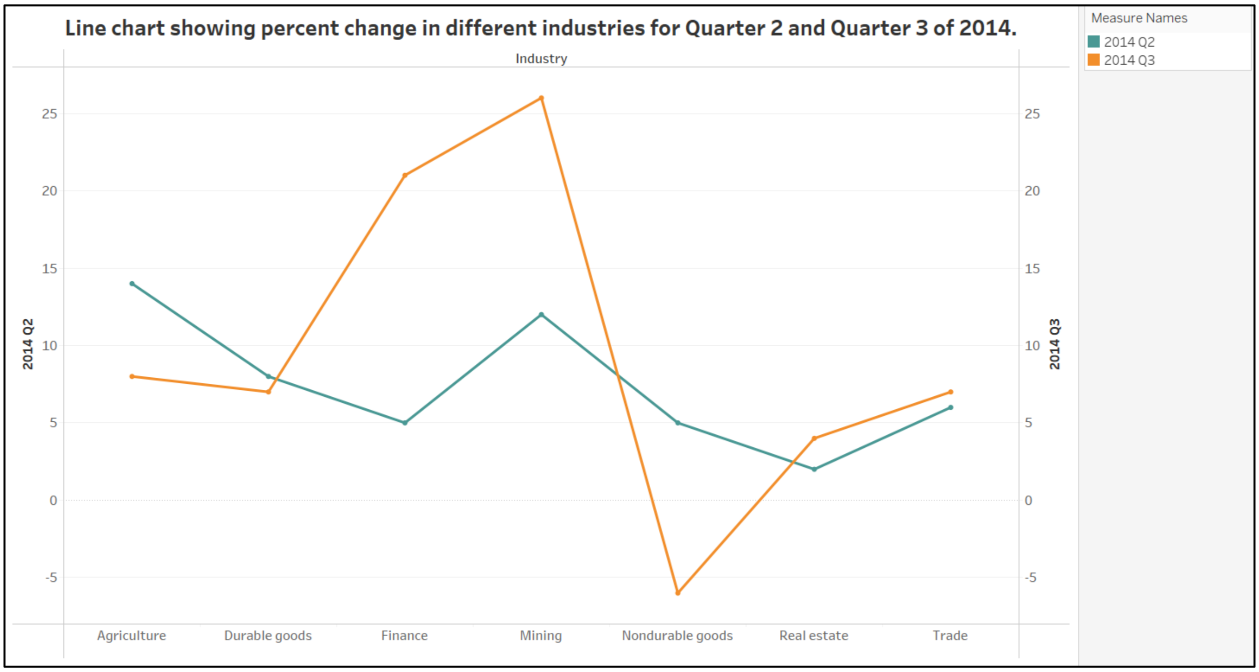
❖ Cons:

→ When we represent two quarters here, we must use different color in order to view both the data properly. Even we need to reduce the transparency so that we can distinguish them easily.

→ Another thing is, if we would have the same values for both the data, then it would be hard to interpret the data, since it will overlap each other.

→ It is less suitable for categorical data, but in our case, we are using numerical data, so it is fine.

→ <u>Line Chart</u>



❖ Pros:

→ Line chart is easy to read and understand, as the data is displayed as a continuous line with data points marked on the line.

→ It shows accurate result because the data is represented as specific data points connected by straight lines.

→ In our case, we can easily see and compare both the quarter results.

→ Even for negative values, i.e., for nondurable goods, we can represent it quite easily.

❖ Cons:

→ It doesn't show all the relevant details, unless annotated.

→ It is difficult to represent proportion using line charts, and for it we need to stick to pie chart or bar chart.

→ It is less suitable for categorical data.

<u>My best visual</u>

The visual which I like and would prefer would be **line chart**. This is because line chart is able to display the results of quarters of 2014 easily. Other than that, we can easily compare quarter 2 and quarter 3 percent change by industry. We are able to distinguish them by using different colors and can get exact value of the percent change using **markers**. Also, nondurable goods had negative value, which was also displayed properly and can be compared with quarter 2 results. These are the reasons why I would choose line chart for this dataset.