Business/Data Analyst Take Home Exercise

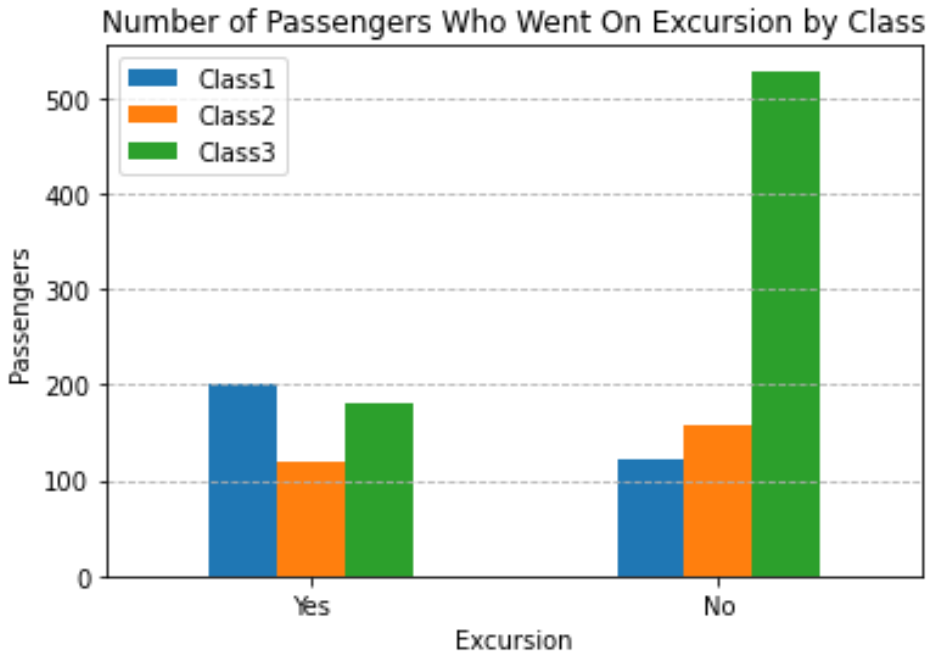
Part A - Visualization and Analysis

Section 1 – Relationships:

Describe three trends or relationships you noticed after exploring the data.

Answer:

First Relationship



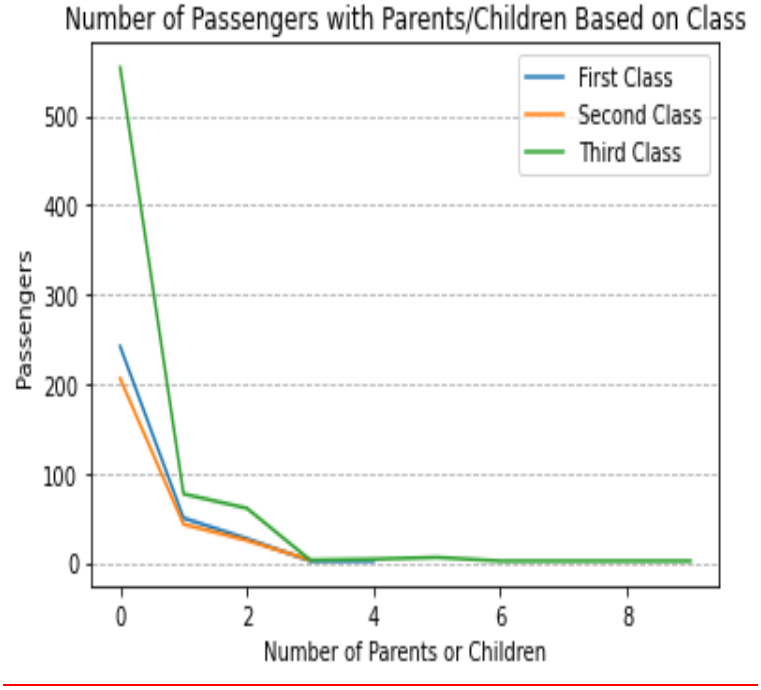
Passenger went to Excursion?		
Class	**Yes**	**No**
1	200	123
2	119	158
3	181	528
Total	**500**	**809**

The first relationship I explored was the number of passengers who went on excursions by class.

Based on the data and graph above, most passengers did not go on an excursion. A total of 809 passengers did not go an excursion vs. 500 who did. Additionally, I found that of the group who did not go, the majority was comprised of third class passengers (~65%). The first and second class passengers were similar in amount of passengers who did not go.

For the group of passengers who did go to an excursion, the amount of first and third class passengers was actually similar (200 vs 181, respectively which is a ~ 9.5% difference).

Second Relationship



Number of Passengers with Parents/Children by Class			
Parent/Children	First Class	Second Class	Third class
0	242	206	554
1	50	43	77
2	27	25	61
3	2	3	3
4	2	0	4
5	0	0	6
6	0	0	2
7	0	0	0
8	0	0	0
9	0	0	2
Total	**323**	**277**	**709**

The second relationship I explored was the number of passengers with parents/children based on class. I was curious to see if class had any relationship on family size. Were third class passengers likely to have more children than first class passengers? Or perhaps less because being in third class might infer that they are of lower socioeconomic status and hence would have a harder time in affording children.

Based on the data and graph above I see that the third class (green line) had the largest spread of parent/children: 0 - 9. But the majority of third class passengers did not have any parents or children with them on the ship (554/709 which is ~78%).

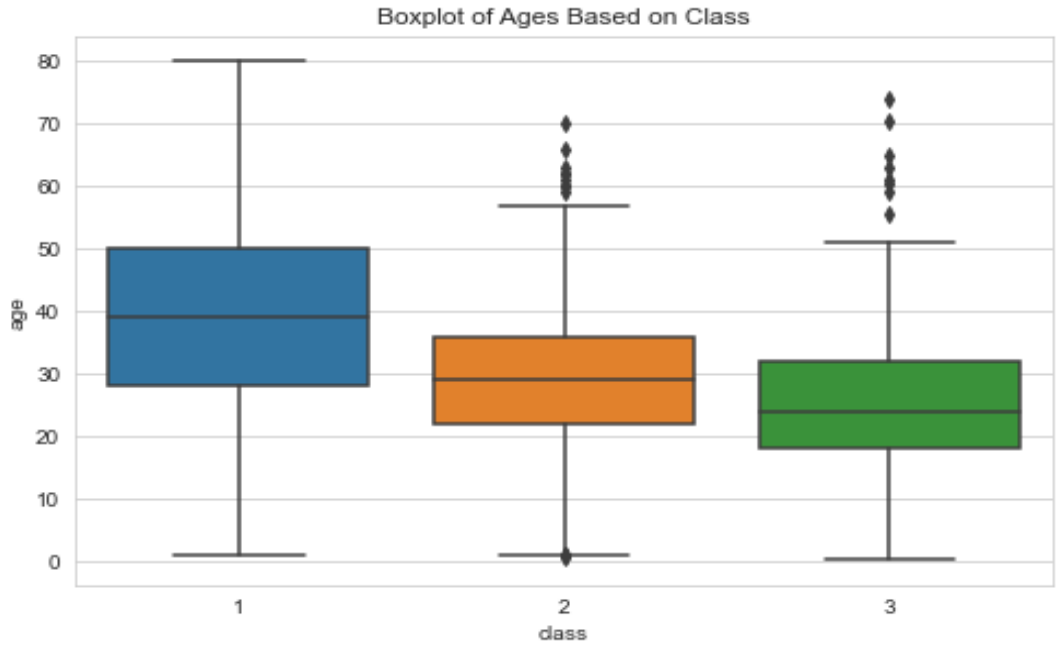
For the first and second class passengers, their trend looks nearly identical. The spread of parent/children looks similar: first class is from 0 – 4 while second class is from 0 – 3. The number of passengers are also similar, with the first class having a few more passengers than the second class for each parent/child size.

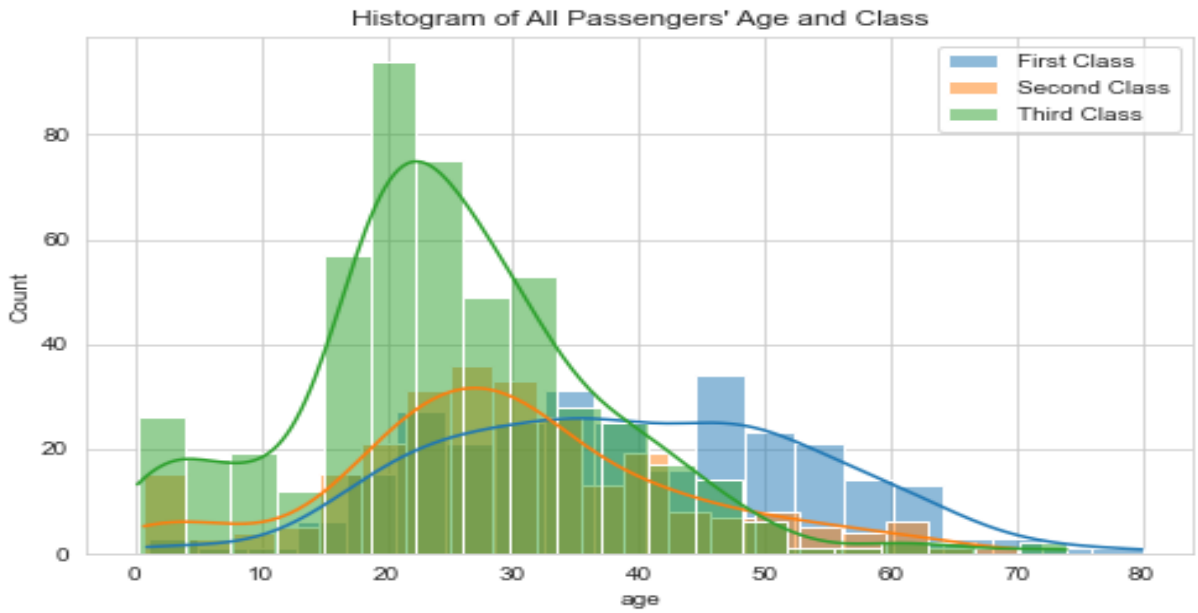
Ex: 242 first class passengers had 0 parents/children vs 206 for second class

50 first class passengers had 1 parents/children vs 43 for second class

27 first class passengers had 2 parents/children vs 25 for the second class

Third Relationship





The third relationship I explored was the age of passengers on class. Were first class passengers more likely to be older than second and third class? If someone is older then they may have more wealth than someone younger, therefore the first class group may be older than second and third.

The charts above show that the first class passengers were, on average, older than second and third class. The interquartile range for each boxplot shows this. The blue boxplot has a larger

vertical range and higher median. This is similarly shown in the histogram wherein the blue line is flatter than the others and spread out more evenly compared to the orange and green lines.

Additionally, the second and third class boxplots have smaller vertical ranges and lower medians. We can see this to reflect in the histogram wherein the middle of the orange and green lines are skewed right. This means that the median ages are lower for both second and third class.

The third class's lower median age and lower interquartile range is probably caused by a plethora of passengers in their lower 20's. We can see this by the sharp green jump in the histogram.

In summary, the third class has the most amount of passengers. A large proportion of them are made up of young adults. The second class is older and has less passengers overall. Lastly, the first class has the oldest median and average age the age range is more evenly spread for adults when compared to the second and third class.

Section 2 – Communication:

Pick one of the trends or relationships you described in Section 1.

Discuss why you think the trend/relationship might exist and what significance these trends/relationships may have.

Please also address:
- Possible limitations or caveats to your analysis
- Additional things you would want or need to know to evaluate or validate the trend or relationship.

Answer: I think the trend for the Third Relationship: Passenger's Age on Class exists due to socioeconomic status. Younger people are less likely to have wealth when compared to older people. If they have less wealth they probably cannot afford to purchase a first class ticket. Also, the majority of the passengers are in the third class. This might also reflect the phenomenon of wealth wherein most people in society are not wealthy. The significance of this phenomenon is that the passengers may not receive equal treatment. One group may have better living conditions, access to higher quality food, and might enjoy advantages that the other group does not.

One possible limitation to this analysis is that we do not have income data of the passengers. There may be many exceptions to our analysis. For example, perhaps some wealthier passengers purchased third class tickets because they are frugal and that was one of the qualities that allowed them to save their wealth. Or perhaps some less-wealthy passengers purchased first class tickets as a means to splurge/celebrate.

Section 3 – Visualization:

Create two visualizations that highlight and describe different aspects of the data.

Answer: Please see Part A Section 1.

Please see jupyter notebook (saved as HTML) titled “Business Data Analyst Part A - Visualization and Analysis 23May2022” to see the code.

Part B – Data Analytics

● What is the median age of the women who went on excursions?

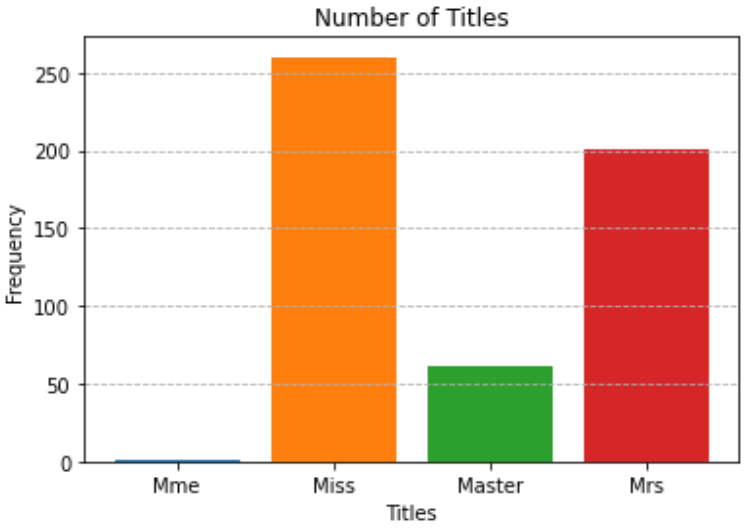
Answer: The median age of women who went on excursions = 28.5 years old.

● Add a column which denotes whether the names have a double letter in them. Possible values for this column can be “true” and “false”.

Answer: Please see column H in the attached excel file titled “CX BA I Take Home - Ship Data - Take Home - Ship Data Found Duplicates”.

● What is the breakdown of titles (For simplicity, just Mme, Miss, Master, Mrs)?

Answer:



Title	Frequency
Mme	1
Miss	260
Master	61
Mrs	201

Please see jupyter notebook (saved as HTML) titled “Business Data Analyst Part B - Data Analytics 23May2022” to see the code.