QUESTION 1: What categories of passengers were most likely to survive the Titanic disaster? ->4 points (the mark includes 2 points for the answer to the main questions and 2 points for the detailed explanation of the logic of the analysis)

After doing crosstab between different groups, I have concluded the following categories of passengers have much higher than average overall survival rate of 0.384:

- Females at 0.742
    - Male has a survival rate of 0.189
    - This makes sense since women were offered a space in the limited lifeboats before the men
- Social class of 1 at 0.63
    - Has the highest survival rate compared to other classes
    - Very likely that higher class was able to secure a spot on the lifeboat
- With one sibling or spouse at 0.536
    - Has the highest survival rate
    - This could mean this passenger is likely a child with another sibling or likely a woman with a husband. Both groups have higher survival rate
- With at least one parent/children at 0.51
    - I agglomerated the survival rate to one or more children or parents since Parch>2 has low counts.
    - Either having at least one children or parent has the highest survival rate compared to not having any. Parents with children and children were more likely to have a seat on the lifeboat.
- With a Cabin letter(cabin number) at 0.667
    - Only 204 passengers have a cabin letter/ cabin number. Survival rate is high because 0.86 of theses passengers with letters have a social class of 1.
    - Website has shown that cabin letters and numbers were hard to obtained. Most of these data were based on a first class passenger list from steward Herbert Cave
        - https://www.encyclopedia-titanica.org/cabins.html
    - It has been shown that social class of 1 has higher survival rate, thus passengers with cabin letters by association has higher survival rate.
- Embarked from Cherbourg at 0.554
    - Highest survival compared to others.
    - 0.506 of passengers from Cherbourg are from social class one
    - Average fare from Cherbourg is also at $60 compared to average of $32
    - Thus, the higher survival for passengers from Cherbourg is likely from class 1 and higher ticket fare leading to higher survival. (More likely to get on the lifeboat)
- <=12 years old (Age_group=”Minor) at 0.575
    - Highest survival out of all the other age groups.
    - Lifeboats were offered to children first.
- Ticket fare >$31 at 0.581
    - Highest survival out of all the fare groups
    - 0.973 of people in >$31 fare group is in social class 1, which leads to the same explanation highest social class was offered a seat on the lifeboat first.

QUESTION 2: What other attributes did you use for the analysis? Explain how you used them. Provide a complete list of all attributes used. -> 2 points

```
Survived              891 non-null category
Pclass                891 non-null category
Title                 891 non-null object
First_Middle Name     891 non-null object
Last_Name             891 non-null object
Sex                   891 non-null category
Age                   891 non-null float64
Age_Group             891 non-null category
SibSp                 891 non-null category
Parch                 891 non-null category
Ticket_Letter_Code    226 non-null object
Ticket_Number         891 non-null object
Fare                  891 non-null float64
Fare_Group            891 non-null category
Cabin_Letter          204 non-null category
Cabin_Number          204 non-null object
Embarked              891 non-null category
```

This is a complete list of attributes I have in the column. I used every single one that is labelled category for my analysis except Ticket_Letter_Code and Ticket_Number. I use the title column for treating missing ages, and embarked column to treat its own missing value (See question 4).

QUESTION 3: Did you engineer any attributes? If yes, explain the rationale and how the new attributes were used in the analysis? If you have excluded any attributes from the analysis, provide an explanation why you believe they have to be excluded ->2 points

I engineered the following attributes:

- Title, First_Middle Name, Last_name
    - Using regular expression on the original Name column, I was able to split it into three columns: Title, First_Middle Name, and Last_name
    - I used the title column to treat missing values in the age column
- Age_Group
    - I grouped the ages into minors (<=12), teenager (<=17), young adults (<=35), middle-age (<=65), and seniors (>65), and put each passenger into the one of the groups.
    - I used this age group to find if any particular group have a higher survival rate.
- Ticket_Letter_Code and Ticket_Number
    - I parsed the original ticket column into Ticket_Letter_Code (letters) and Ticket_Number (number)
    - I didn't use these two engineered attributes
- Fare_Group
    - I grouped the fare based on quartiles of fare into 4 groups: <=$7.9, ($7.9,$14.5), ($14.5, $31), and >$31
    - I used this attribute to find if any group has higher survival rate.
- Cabin_Letter and Cabin_Number

- I parsed the cabin into Cabin_Letter and Cabin_Number using regular expression.
- I used this attribute to find if any cabin letter or having a cabin letter has higher survival rate.

No attribute was excluded.

QUESTION 4: How did you treat missing values? Provide a detailed explanation in the comments. ->2 points

There are 177 ages missing. Trying to extrapolate using the following strategy:

- Define the median age of boys with title master. Fill in the median age for any missing age in this group
- For men with a Mister title (adult men), use the median age to fill in the missing ages
- For women, based on their title/martial status (Miss vs Mrs), fill in with the median age of Miss or Mrs.
- For Dr. replacing the missing doctor's age with the median of Drs' ages.

There are two embarked entries missing. I filled in the empty embarked fields with the mode of whole data frame's embarked.