PROJECT: TITANIC SURVIVORS DATA ANALYSIS

QUESTION FOR INVESTIGATION:

Q1: What factors made people more likely to survive?

1. To answer the factors responsible for survival, following relationships are analysed in detail from given data:

    - Relationship between survived people & passenger class (Pclass)
    - Relationship between survived people & gender
    - Relationship between survived people & age
    - Relationship between survived people & marital status (single/family)
    - Relationship between survived people, passenger class & fare
    - Relationship between survived people & embark
    - Relationship between survived people, age & gender
    - Relationship between survived people, age & passenger class (Pclass)
    - Statistical chi squared test between survived & gender

A. Importing required libraries:

1. 'Numpy' and 'Pandas' are loaded for converting data from csv file to dataframe and analyse it further.
2. For plotting purposes, 'matplotlib' is imported and 'seaborn' library is utilized for enhancing the representation of the figures.

B. Loading the data and cleaning it:

1. Using the csv package, the given data is converted and saved into pandas dataframe variable 'titanic_data'. We can see that it consists of 12 variables & 891 observations.

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

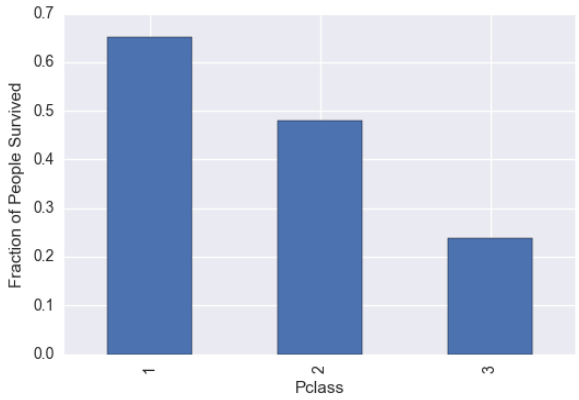
2. A modified data frame named 'tdn' is created by removing the columns 'Cabin' & 'Ticket' as they are not required in our analysis.

3. Next, we see that ‘Age’ and ‘Embarked’ columns have ‘NA’ values present. These observations cannot be used for constructing relationships. So, using ‘dropna’ function these observations are removed from data and finally the data is saved in ‘titanic_data_final’. As we can see that only 712 observations are valid for analysis as they have all the variable values available.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 712 entries, 1 to 891
Data columns (total 9 columns):
Survived    712 non-null int64
Pclass      712 non-null int64
Name        712 non-null object
Sex         712 non-null object
Age         712 non-null float64
SibSp       712 non-null int64
Parch       712 non-null int64
Fare        712 non-null float64
Embarked    712 non-null object
dtypes: float64(2), int64(4), object(3)
memory usage: 55.6+ KB
```

C. 1st Relationship – Survived people and passenger class:

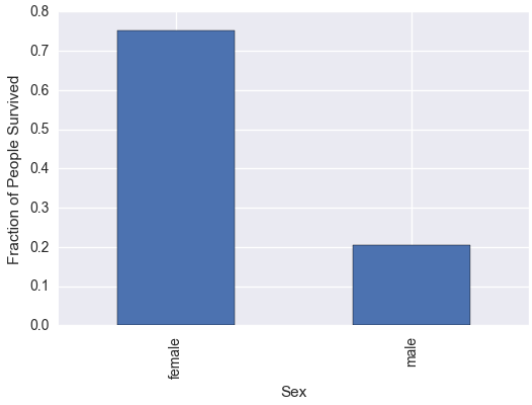
1. There are 3 different passenger class on the ship. (1, 2 & 3) For analysing this relation, the people who were from each class are grouped together respectively. In those classes percent of people who survived is plotted in the form of bar graph.



2. Class 1 people had 65.2% survival rate, while the Class 2 and Class 3 had 48% and 24% of people alive respectively.
3. This relation shows that Class 1 must be made up of rich people as they were given priority while evacuating the ship. Similarly, Class 3 people would mostly consist of poor people.
4. Further, we can see that number of survivors from Class 1,2 & 3 were 120, 83 & 85 respectively. This shows that number of rich and middle-class people on board were almost same. Most number of people were from poor class.

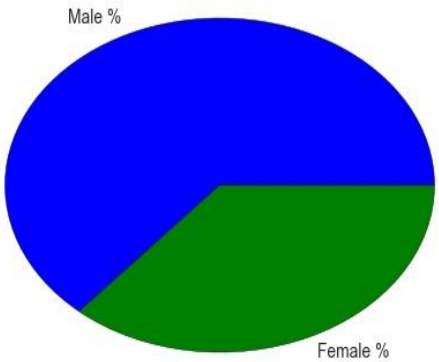
D. 2nd Relationship – Survived people and Gender:

1. Plotting the graph like the earlier relation, we see that there is huge difference in survival rate based on gender. (male/female)



2. Females had 75.2% survival rate, while unbelievably 79.5 % of Males did not make it. This is in spite the fact that % of Males was more than Females on board. (See Pie Chart)
3. This statistic confirms the general trend in any culture where ladies are prioritized over gents. Children with the females would also had been given more importance. This is analysed in detail in the age factor.
4. These are the stats for actual number of people survived grouped my gender:

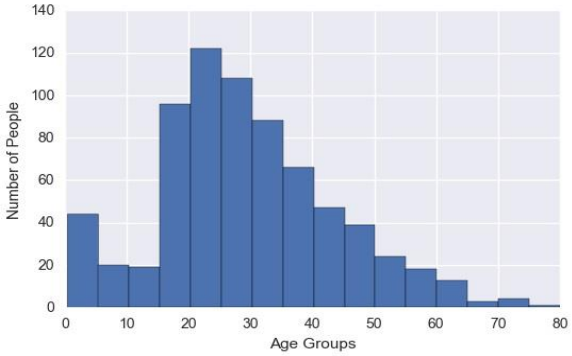
```
Sex
female    0.752896
male      0.205298
Name: Survived, dtype: float64
Male Survivors Number: 93
Female Survivors Number: 195
Total Survivors Number 288
Percent of Male Survivors: 20
Percent of Female Survivors: 75
```



PIE CHART FOR % PEOPLE BY GENDER

E. 3rd Relationship – Survived people and Age:

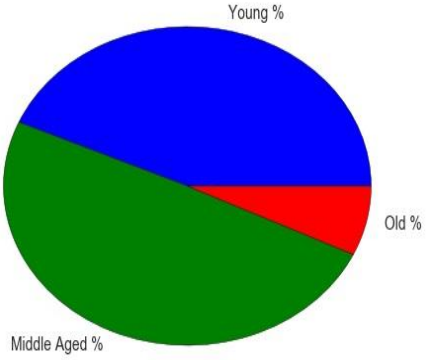
1. Firstly, people in the dataset are grouped together into 3 different age groups as follows:
    - Young (0-25 years)
    - Middle Aged (26-50 years)
    - Old (51-80 years)
2. Number of survivors from each of these groups are extracted and analysed. Also, a histogram plot for number of people on board grouped by age is plotted. From this plot, we can see that people aged between 20-25 years were most common, closely followed by age group 25-30 years.



3. More detailed statistics on this relationship is given by 'describe' function. It is given below. Oldest person on board was 80 years old while the youngest one was aged 0.42 years old. (5 months) The mean was 29.6 years.

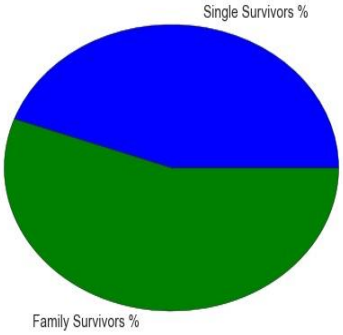
```
count    712.000000
mean      29.642093
std       14.492933
min        0.420000
25%       20.000000
50%       28.000000
75%       38.000000
max       80.000000
Name: Age, dtype: float64
```

4. By grouping the people into 3 categories as mentioned above, we extract that 124 young people, 143 middle aged people & 21 old people survived. The limits for dividing age groups can be varied by using the function 'AgeGroups'.
5. A pie for survived people by these age groups is created. We can see that middle-aged people had maximum survival rate.

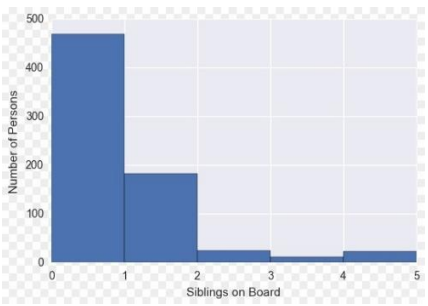


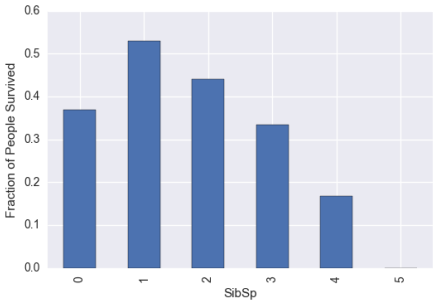
F. 4th Relationship – Survived people and Marital Status:

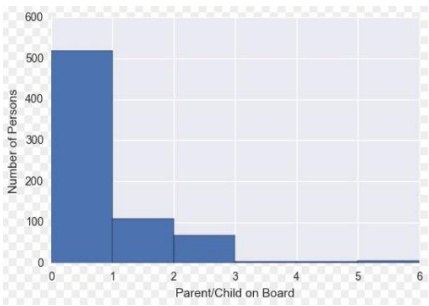
1. For determining whether a person is single or with a family on board, factors named 'SibSp' (Siblings on board) and 'Parch' (Parent/Child on board) are utilized. If a person has no sibling or parent/child on board, then he is considered single. So, the data is saved in two new variables:
    - 'Single_Survivors_Num'
    - 'Family_Survivors_Num'
2. Number of single survivors are 128 while number of family members who were alive are 160. Percent of survivors based on these two categories are plot on pie chart to get further information. We can see that family members that greater survival success.

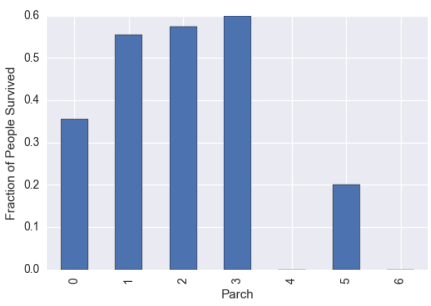


3. Number of people on board grouped by siblings and then by parent/child are plotted. Also, their survival percent are plotted on another set of bar graphs. Single people were most common on ship. People with more parent/child and siblings (4 or 5) had less survival success. It is interesting to note that people with 1 or 2 family members on board had more survival rate than single people. Most likely this could be females carrying their children, who were given priority.



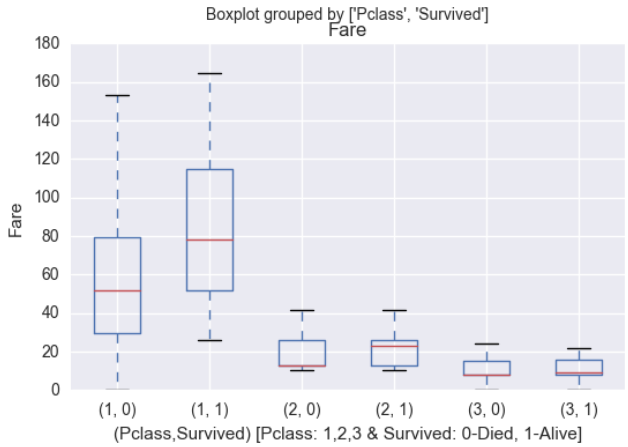






G. 5th Relationship – Survived people, Passenger Class and Fare:

1. People on board are grouped together by passenger class and survival in this relation. A box plot is shown to indicate fare distribution among these groups.



2. We can see that majority proportion of people survived in Pclass 1, which had highest fare. Similarly, other classes had lower proportion of people survived with Pclass 3 being worst. Higher fare indicates people consisting of rich class and lower fare consists of poor people.

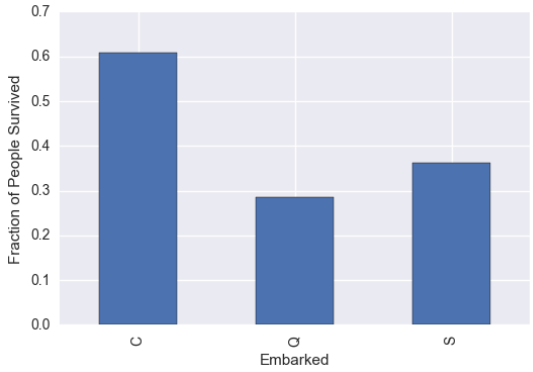
H. 6th Relationship – Survived people and Embarkment:

1. In the given data, the factor 'Embarkment' is divided into 3 classes: C, Q & S. Obtaining stats on this factor, we get following details:

```
Embarked
C         count    130.000000
          mean       0.607692
          std        0.490153
          min        0.000000
          25%        0.000000
          50%        1.000000
          75%        1.000000
          max        1.000000
Q         count     28.000000
          mean       0.285714
          std        0.460044
          min        0.000000
          25%        0.000000
          50%        0.000000
          75%        1.000000
          max        1.000000
S         count    554.000000
          mean       0.362816
          std        0.481247
          min        0.000000
          25%        0.000000
          50%        0.000000
          75%        1.000000
          max        1.000000
Name: Survived, dtype: float64
```

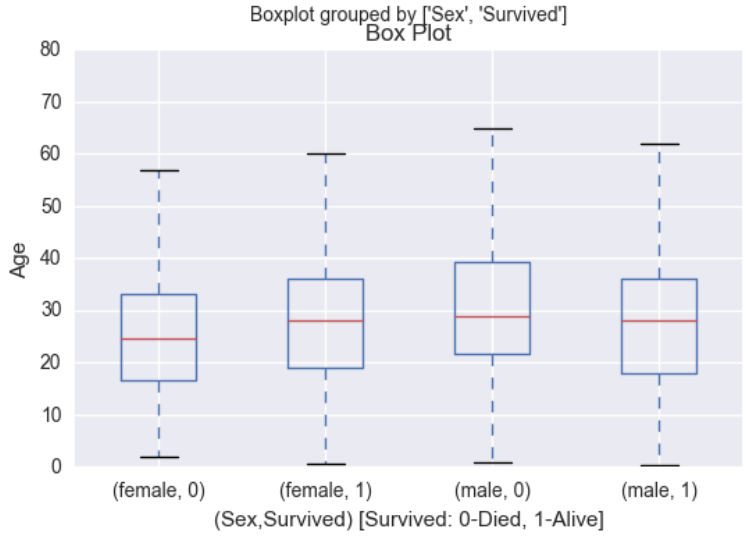
2. Out of the total 712 people we are analysing, 130 attempted to reach the location named 'C', while 28 and 554 people tried to travel the rescue points named 'Q' and 'S' respectively.

3. Location ‘C’ had highest success rate of 60.7%. The next successful location for survival was ‘S’, where 36.2% people arrived safely. The worst location in terms of survival rate was ‘Q’. (28% success) This is also indicated by the following bar plot:



I. 7th Relationship – Survived people, Age & Sex:

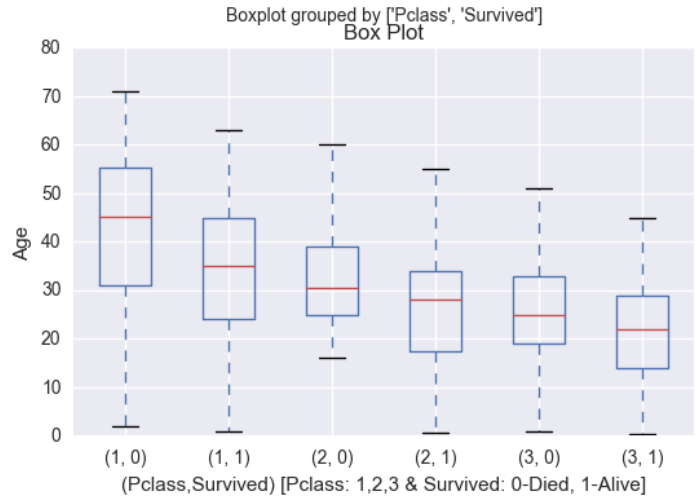
1. This multi variable relation is explored to see whether a person was more successful at surviving based on age and gender. For this, a box plot is created.



2. It is clear that older women and younger male were more successful in surviving. From the plot, it seems that there was no significant advantage for survival based on age and gender combined.

J. 8th Relationship – Survived people, Age & Pclass:

1. Like last relation, survival chances are analysed based on age and passenger class of travelling people using box plot.



2. For people in class 1 (Rich class), survived ones had mean age around 35 years while the ones who did not made it had mean age around 45 years. Similar trend is seen in class 2 and 3 but the mean ages dropping further. Therefore, younger aged people were more successful in survival in all the respective classes.

K. 9th Relationship – Statistical Test Between Survived People and Gender:

1. To answer the question whether females had better chance of survival than males, a statistical test is performed. Chi squared test is preferred as the dependent variable 'Sex" is categorical is nature. (P-test & Z-test are not appropriate here) Set of null and alternative hypothesis statements for the test are given as:

   H_0: Survival chances are independent of gender
   H_A: Survival chances are not independent of gender

```
# 9th Relationship: Survived & Sex Statistical Test:

# Chi-Squared Tests
from scipy.stats import chi2_contingency

# Sex to Survivability
pivot = pd.pivot_table(data = titanic_data_final[['Survived', 'Sex']], index = 'Survived', columns = ['Sex'], aggfunc = len)
print pivot,"\n"

chi2, p_value, dof, expected = chi2_contingency(pivot)

print "Results of Chi-Squared test on Sex to Survival:"
print "Chi-Squared Score = " + str(chi2)
print "Pvalue = " + str(p_value)
print "\n"
```

```
Sex       female  male
Survived
0             64   360
1            195    93

Results of Chi-Squared test on Sex to Survival:
Chi-Squared Score = 202.869448776
Pvalue = 4.93941668545e-46
```

2. Observing the values for Chi squared score and P-value, the results are significant for all reasonable alphas. So, we must **reject the null hypothesis.**
3. Further, by looking at the pivot table, we clearly see that proportion of females survived are much greater than males. Survivability and gender are dependent on each other with females more likely to survive than males.
4. One important part to note that though the test indicates that females had an advantage over males, it does not reveal complete picture. As the gender is not the only variable on which survival is dependent, further tests must be carried out which considers other variables also to answer this question more deeply.

CONCLUDING REMARKS, LIMITATIONS AND OBSERVATIONS

1. Being in the rich class, it was likelier to survive as priority privileges were given to this class. Also, females were significantly more successful in making it than males. So, if a female person was in rich class, she was most likely to survive.
2. Middle aged people were most successful in surviving, followed by young people. Old people were most likely to die due to obvious reason that they are limited in physical aspect than people from other class.
3. Family people were more likely to make it than single people. (especially people with 1 or 2 family members.) This can be relatable as people with young children must be given priority.
4. Location 'C' had highest survival rate among 3 given. The reasons for this can be derived by looking more into geographical details of the incident place. This information is not provided in the data set.
5. Statistically, it can be seen that older women and young men had more chance surviving. But the difference is not large enough to consider it as important factor in survival.
6. Being younger in all the classes was certainly a factor in survival. They were likelier to survive. This trend is slightly different than one analysed in survival vs age. The reason middle class people had slightly more chance of surviving than young people because of the limits of age selected while defining those age groups. These values can be appropriately selected so that similar trend can be observed in both the factors.
7. Many of the entries in the data set had values 'NA'. Due to this analysis of only 712 people out of 891 ones could be made. Complete data set would have led to better accuracy in results.