1. **I have analyzed the titanic_data**

Quality of Analysis

2. **Questions:**

a. What factors lead to more survival
b. How many passengers embarked from each station
c. Which city has more survival
d. What correlation between passengers belonging to a particular class and survival
e. Correlation between class and survivors

3. **Data Wrangling:**

a. Removed the missing values from the data by using dropna() function in Pandas

b. Removed '()' and the data within the parenthesis from the Names column of the data as that data was not required during analysis

c. Presented the names in the format Salutation, First Name, Last name to make it more readable.

4**. Exploration Phase**

a. Explored the data and found out the total number of survivors and deaths. This helped me verify with the number every time i calculated survivals according to the pclass, age, etc. Filtered the fields whose 'Survival' values were 1 and then counted the total number of such fields.

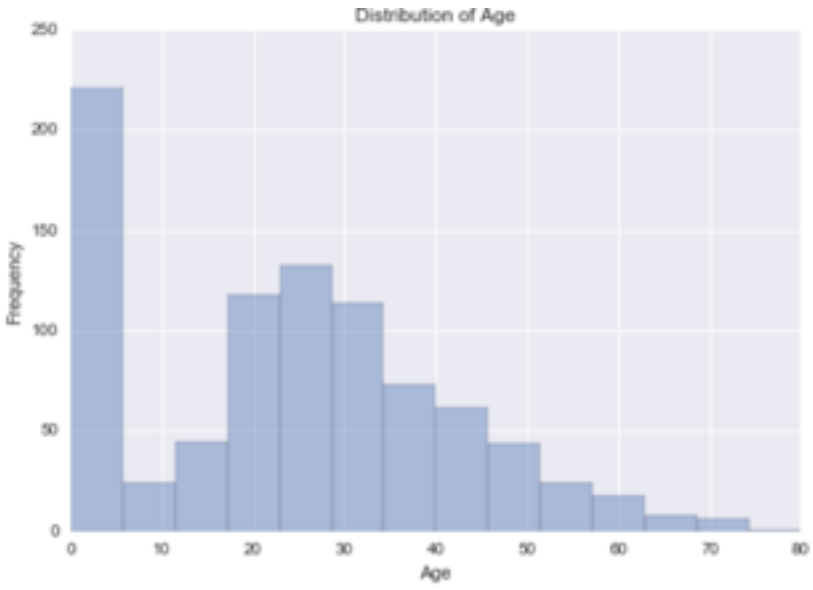
To check out the factors that lead to more survival:

i. Filtered survivors according to different **passenger classes**, 1,2,3, and counted the survivors in each class to check out which class has the maximum survivors
ii. Filtered the survivors according to **age** and counted the survivors that were below 30 years and above 30 years to check whether young or the old people were the maximum survivors.
iii. Filtered survivors according to the **siblings and parents** they had on board to check whether the passenger with the maximum number of family on board survived.

For all these, I considered the 'Survived' as the dependent variable and 'Pclass','Age','SibSp' and 'Parch' as the independent variables.

Single variable analysis:

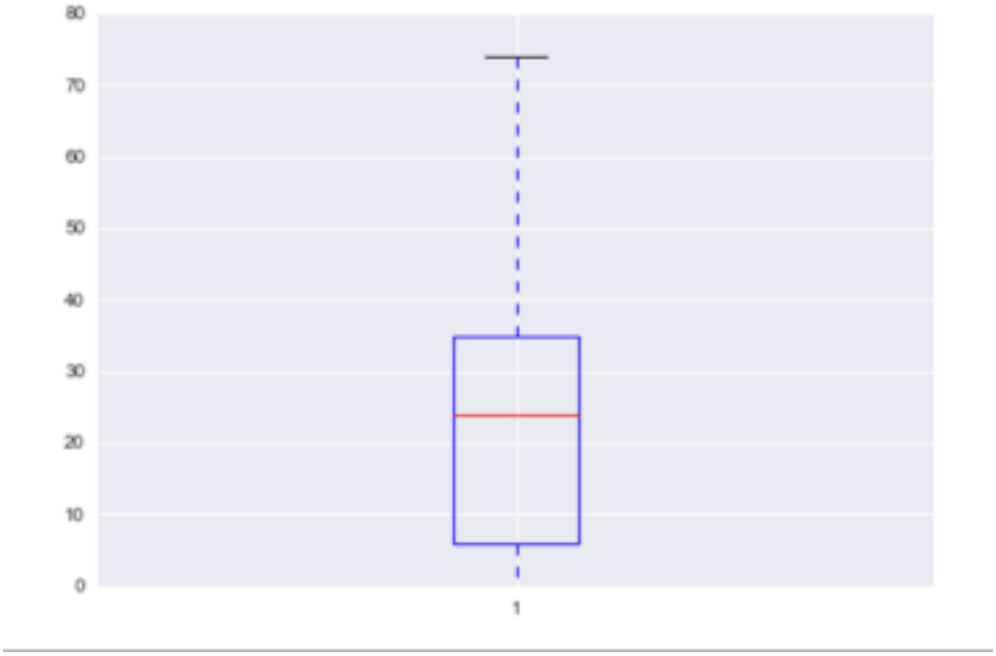
Analyzed all the variables without considering their relationship with other variables.



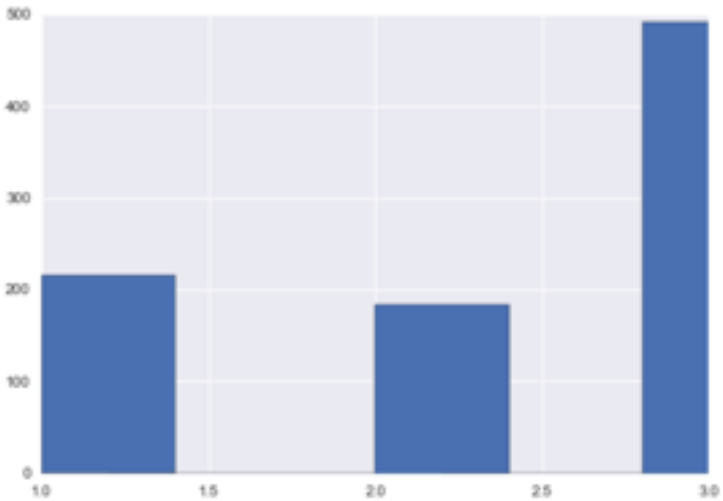
Analysis steps include:
i.Calculated how many standard deviations each value is away from the mean
ii.**Age**:

Descriptive statistics:

a. The shape of the distribution is positively skewed.
b. The mode is in the interval 0-10. Most passengers were between the age 0-10
c. Mean = 23.799293
d. Median = 24.0000
e. Standard deviation = 17.596074
f. Outliers are present in the data which I visualized using box plots

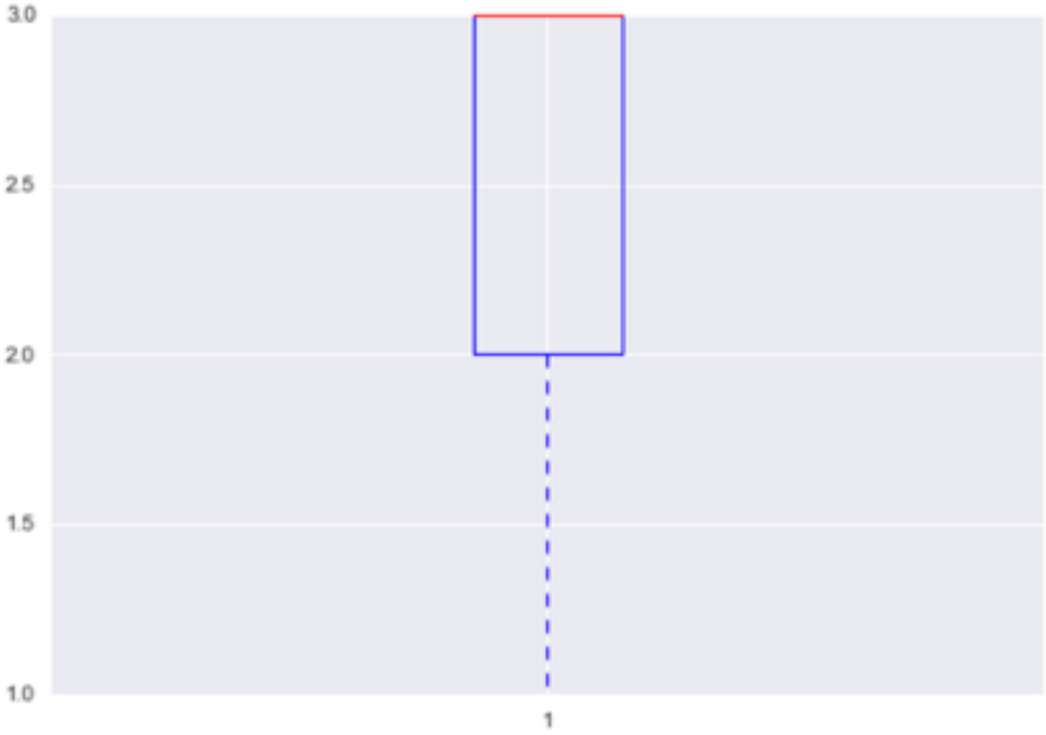


iii. **Class**:

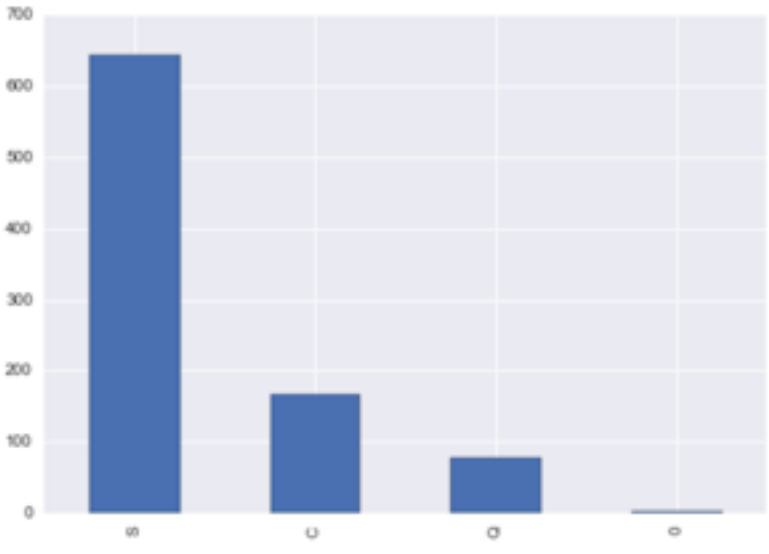


Descriptive statistics:

a. The shape of the distribution is negatively skewed.
b. The mode is in the interval 3. Most passengers were in passenger class 3
c. Mean = 446.000000
d. Median = 3.0000
e. Standard deviation = 257.353842
f. Visualized using box plots



iv. Embarked city details



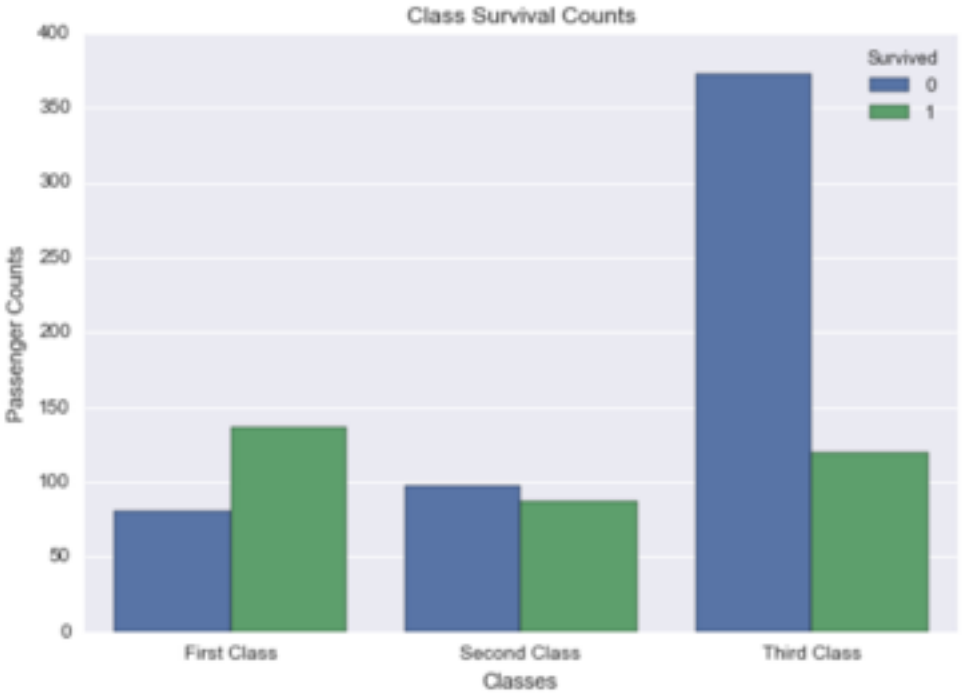
Descriptive statistics:

a. The shape of the distribution is positively skewed.
b. Most passengers embarked from the city of Southampton

Multiple variable analysis:

Examined the relationship between the dependent and independent variables:

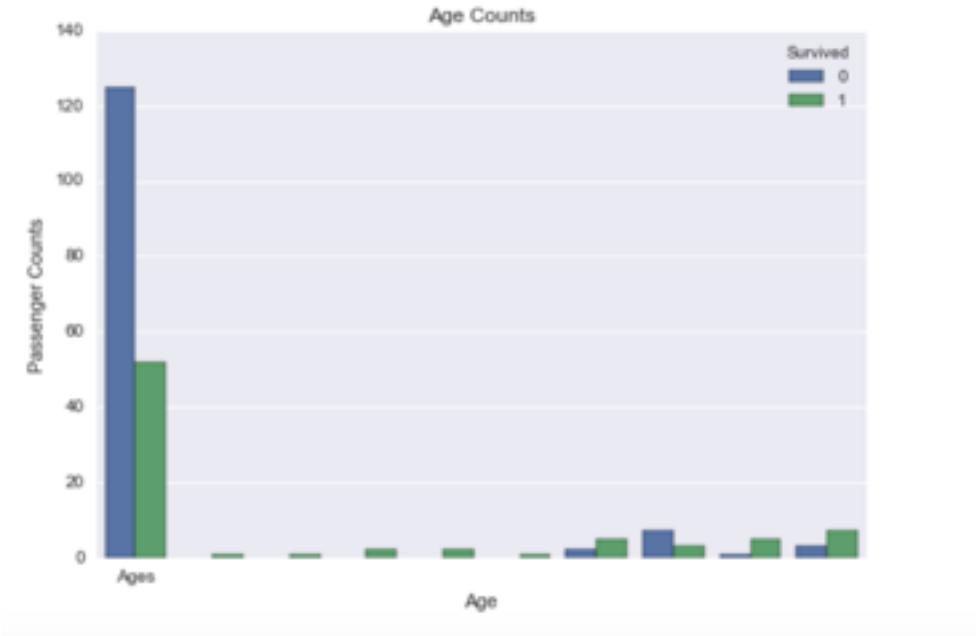
i. Relationship between class and survival:



Observations:

The people of third class survived the least and the maximum number of survivors belonged to first class.

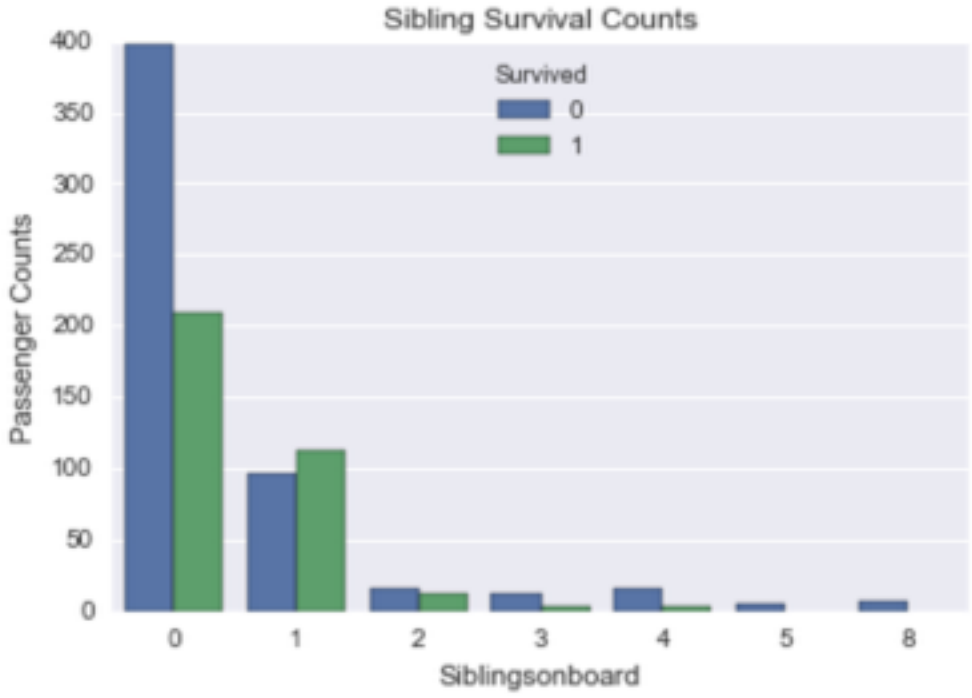
ii. Relationship between age and survivals:



Observations:

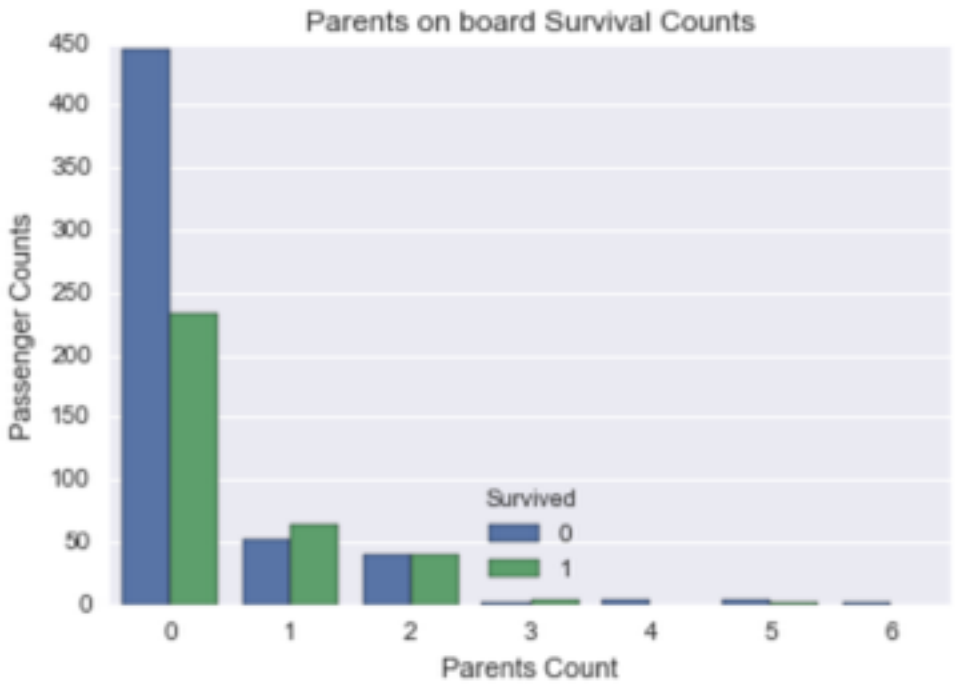
Maximum number of passengers were of age between 0 - 30. And the maximum number of deaths occurred for passengers between age 0-10

iii.Relationship between siblings and survivals:



Observations:
People having the maximum number of siblings did not survive at all

iv. Relationship between parents on board and survivals:

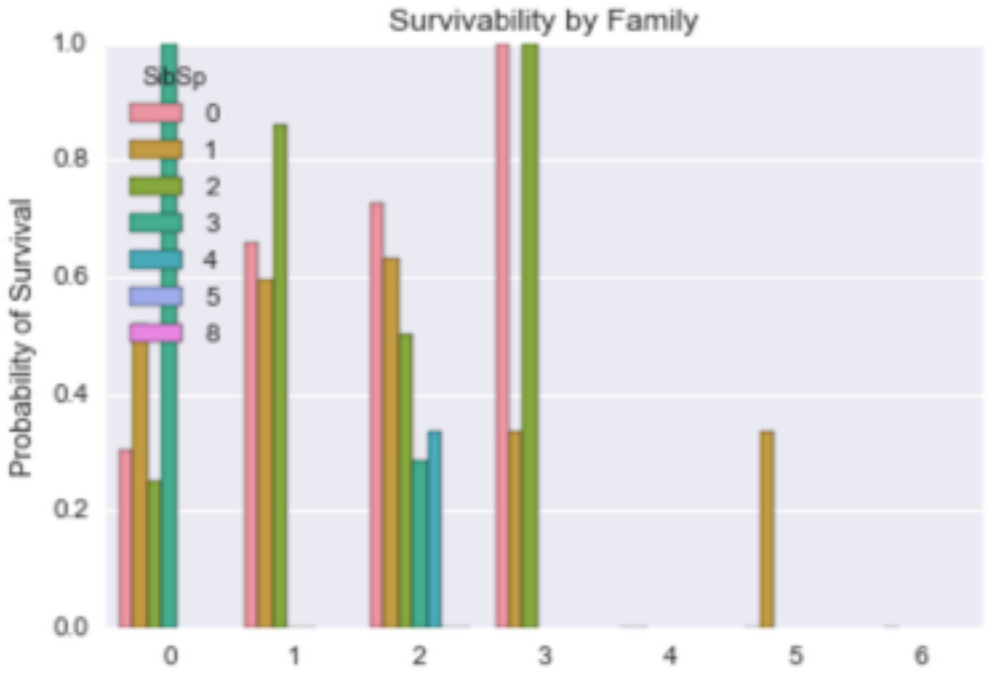


---

v. Relationship between gender, class and Survival:

Observations:
In each class, the female survivors were more, highest being in the first class.

vi. Relationship between gender, class and Survival:



Observations:
The shape of the distribution is positively skewed. Lower the family members, higher is the survival rate.

5. Limitations:

a. The **age** column consists of a lot of missing values, which makes it difficult to calculate statistics relative to the entire set of people. On dropping the nan values, using axis or index, the entire row or column with missing values get dropped.
b. The **ticket** column was unnecessary as it does not provide any basic insight into the analysis
c. The **cabin** column also had a lot of missing values. If cabin values had been properly provided, and which class had which cabins, another filter using survivals by cabin could have been easily calculated.

d. Most of the data are in the string format or have only two values, due to which calculation of means and comparing between other groups is limited, which rules out the use of statistical testing. Therefore, have to rely on the percentage, mean and standard deviation calculation which does not provide any confidence interval of our assumptions. One does not know if

adverse impact truly exists in some defined population.Therefore, we make the best decision we can based on the results of statistical testing.

6. **Conclusions about my findings:**

The main concern about analyzing the titanic data, is to draw a pattern between the factors that lead to people's survival and deaths. Based on the fields provided, those fields had to be extracted that can relate to the survival or death patterns. Those fields were,

- Class
- Age
- SibSp
- Parch
- Embarkment
- Gender

On exploring the data in excel itself, I found out that there was some relation of the survival or death rate with these fields. For example, I filtered the data according to Pclass = 1, and then I found most of the survival values were 1 for that class.
Similarly, I filtered the values according to various SibSp and Parch values and I found higher the number, lower were the survivals. Therefore, these fields made the basis of my analysis and I wrote code using Pandas ad Numpy and visualized the data as well and came to the following conclusions:

a. According to the analysis performed, people of first class survived more as those class people were the people paying the highest fare.
b. Young people whose age was less than 30 survived more
c. People who embarked from Southampton survived more
d. Females survived more as compared to males
e. People with huge family(SibSp,Parch) survived less
f. Survival and passenger class are negatively correlated

One limitation is the correlation between Survival and age are not accurate. On calculating the correlation, it shows positive correlation between ages and survival which is not the case because according to calculations, lower the age higher was the survival.

List of websites I referred:

1. http://onlinestatbook.com/2/graphing_distributions/boxplots.html
2. http://flowingdata.com/2008/02/15/how-to-read-and-use-a-box-and-whisker-plot/
3. http://stackoverflow.com/questions/17950374/converting-a-column-within-pandas-dataframe-from-int-to-string
4. http://whatis.techtarget.com/definition/positive-correlation
5. https://stanford.edu/~mwaskom/software/seaborn/
6. http://pandas.pydata.org/
7. http://docs.scipy.org/doc/numpy-1.10.0/reference/generated/numpy.array.html
8. http://adverseimpact.org/CalculatingAdverseImpact/StatisticalTesting.htm
9. http://cogmaster-stats.github.io/python-cogstats/auto_examples/plot_pandas.html

10. http://onlinestatbook.com/2/graphing_distributions/boxplots.html
11. http://stackoverflow.com/questions/31029560/plotting-categorical-data-with-pandas-and-matplotlib
12. http://stackoverflow.com/questions/17950374/converting-a-column-within-pandas-dataframe-from-int-to-string