DAND Investigate a Dataset Final Project

Titanic Data

Numpy, Pandas, DataFrames, Python

Jan 24, 2017

Final Review by Udacity Team

Meets Specifications

Next, Data Wrangling and MongoDB and SQL

It is very clear you have put in a lot of effort into this project and your should feel proud. Congratulations on completing this difficult course. Next, we will be doing some data wrangling with python - cleaning up datasets so they can be analyzed meaningfully. This is an incredible skill considering the bulk of Big Data is not standardized. You may have even noticed this about the Subway Dataset. Another incredible skill which will become even more important as we acquire more and more data is the ability to work with NoSQL and SQL databases. We will go over all this next. Good luck on the next course. It was a pleasure to look over your project. Keep up the great work!

Code Functionality

All code is functional and produces no errors when run. The code given is sufficient to reproduce the results described.

Code reflects the work done in the analysis and produces no errors.

The project uses NumPy arrays and Pandas Series and DataFrames where appropriate rather than Python lists and dictionaries. Where possible, vectorized operations and built-in functions are used instead of loops.

Great work applying the `numpy` and `pandas` libraries appropriately throughout the report!

The code makes use of functions to avoid repetitive code. The code contains good comments and variable names, making it easy to read.

Quality of Analysis

The project clearly states one or more questions, then addresses those questions in the rest of the analysis.

You have postulated several points of exploration and have provided a thoughtful investigation in turn. Asking the right questions is, arguably, the most important part in data analysis. Well done!

Data Wrangling Phase

The project documents any changes that were made to clean the data, such as merging multiple files, handling missing values, etc.

You have done a great job documenting your work and talking about the missing values for the variables you have explored. I would highly recommend just giving a brief overview of all the missing values in the dataset. This gives the reader a better idea of the quality of the dataset as a whole and adds to the transparency of the report.

- What variables had missing values?
- How many values were missing?
- How have you decided to handle this?

Hint
A really simple way to document missing values is the following...

```
# df stands for whatever the name is of your pandas dataframe
>>> df.info()
```

Exploration Phase

The project investigates the stated question(s) from multiple angles. At least three variables are investigated using both single-variable (1d) and multiple-variable (2d) explorations.

The use of univariate and multivariate plots/stats to investigate the answers to your questions from several perspectives is thorough and compelling.

The project's visualizations are varied and show multiple comparisons and trends. Relevant statistics are computed throughout the analysis when an inference is made about the data.

At least two kinds of plots should be created as part of the explorations.

To Exceed Expectations

The following are some statistical tests you can perform on the dataset. Pay attention to the assumptions of the test before applying. I have attached documentation for the tests but it is up to you to do the research and find the appropriate one.

I will assume you have an understanding of Z-tests, T-tests, etc. The following tests are for some specific situations you will encounter i the dataset that we may not have gone into deeply. This part is OPTIONAL, as the rubric indicates, but will be valuable to your career.

Dependent Variable is Continuous

Independent T-test

- Dependent variable is continuous
- **Samples are Independent**, meaning we are comparing the means of different samples. If this seems confusing, look at the converse – Dependent T-test definitions. An example might be if we were to test the significance of height between Males and Females. Obviously, the samples are different since one group are Males and the other are Females – they are independent.
- Assumes normal distribution of the dependent variable
- Population parameters are unavailable (usually less than 30 samples) https://docs.scipy.org/doc/scipy-0.15.1/reference/generated/scipy.stats.ttest_ind.html

Dependent T-test

- Dependent variable is continuous
- **Samples are Dependent**, meaning we are comparing the means of the same samples. This is sometimes called a *Paired test* or *before and after test* because we are taking samples, measuring a result, then applying an effect, and measuring them after the effect to see if there are significant effects of the affect. There is one group and it is being tested twice.
- Assumes normal distribution of the dependent variable
- Population parameters are unavailable (usually less than 30 samples) https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_rel.html

Mann Whitney U-test

- Dependent variable is continuous
- Does not assume normal distribution of the dependent variable
- Often used in place of the **Independent T-test** when samples are not normal and as such, assumes the samples are **Independent**. http://docs.scipy.org/doc/scipy-0.17.0/reference/generated/scipy.stats.mannwhitneyu.html

Wilcoxon Signed-Rank Test

- Dependent variable is continuous
- Does not assume normal distribution of the dependent variable
- Often used in place of the **Dependent T-test** when samples are not normal and as such, assumes the samples are **Dependent**. https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.wilcoxon.html

One-Way ANOVA Test

- Dependent variable is continuous
- Dependent variable is Normally Distributed (this needs to be determined before testing.

- Typically used for 3+ Independent Variables (If 2 or less, the T-test should be performed since it is easier computationally and performs just as well). http://docs.scipy.org/doc/scipy-0.16.0/reference/generated/scipy.stats.f_oneway.html

Kruskal–Wallis Test

- Dependent variable is continuous
- Similar to an **ANOVA Test** but for non-normal distributions (distribution still should be determined just in case a regular ANOVA test can be performed.)
- Typically used for 3+ Independent Variables (If 2 or less, the Mann-Whitney U-test should be performed since it is easier computationally and performs just as well). http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mstats.kruskalwallis.html

Dependent Variable is Categorical

Difference in Proportions Test

- Dependent variable is categorical
- Typically used to compare the Dependent variable of 2 Independent-Variable Conditions
- Usually involved creating a Contingency Table http://statsmodels.sourceforge.net/devel/generated/statsmodels.stats.proportion.proportions_ztest.html#statsmodels.stats.proportion.proportions_ztest

Chi-Squared Test

- Dependent variable is categorical
- Typically used for 3+ Independent-Variable Conditions
- Should be equivalent to the **Difference in Proportions Test** for less than 3 Conditions
- Usually involved creating a Contingency Table http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chi2_contingency.html

Choosing the Correct Test
http://www.ats.ucla.edu/stat/stata/whatstat/

Conclusions Phase

The results of the analysis are presented such that any limitations are clear. The analysis does not state or imply that one change causes another based solely on a correlation.

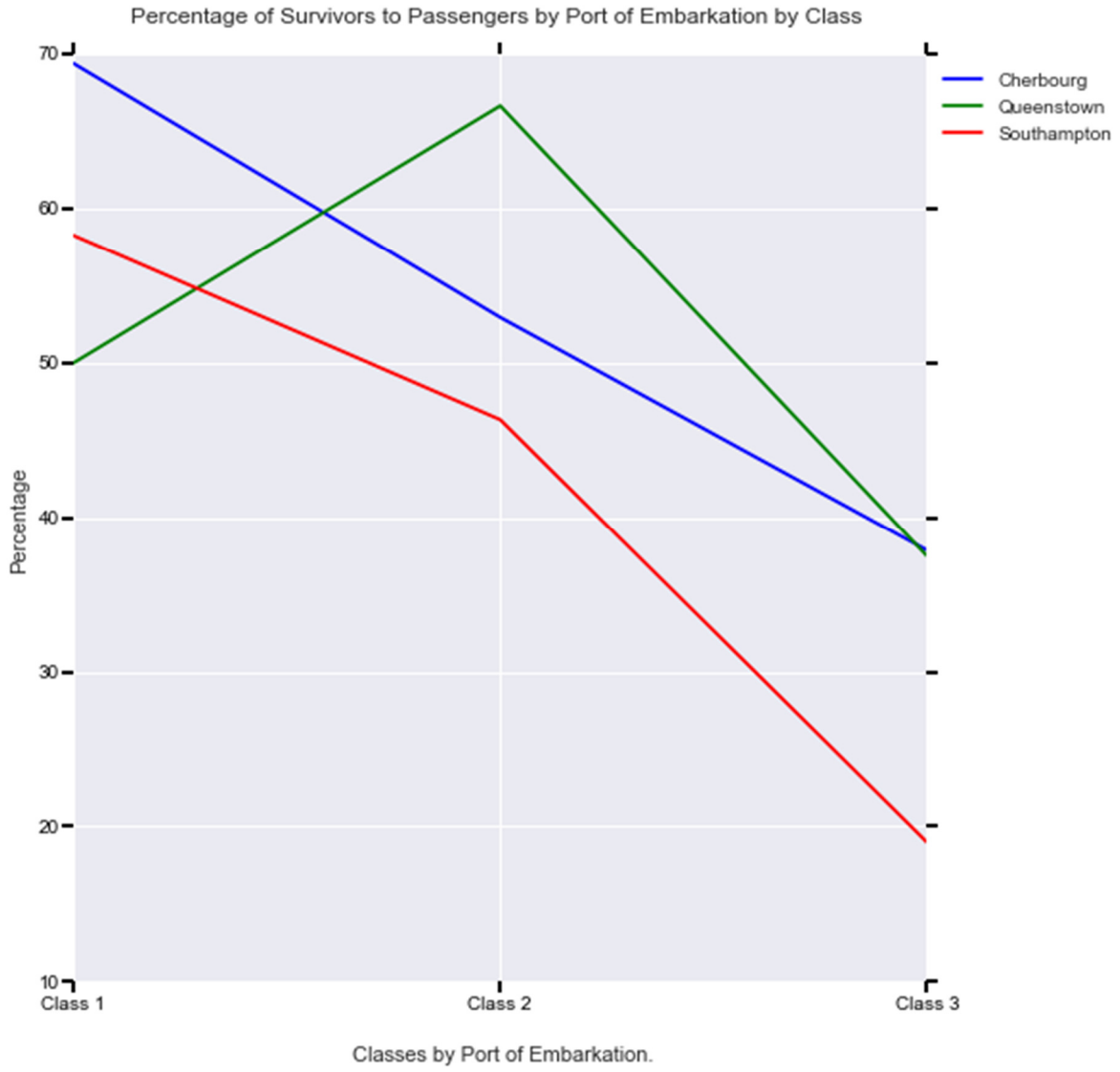
Outstanding work on presenting the investigation in a way that makes the limitations clear.

Communication

Reasoning is provided for each analysis decision, plot, and statistical summary.

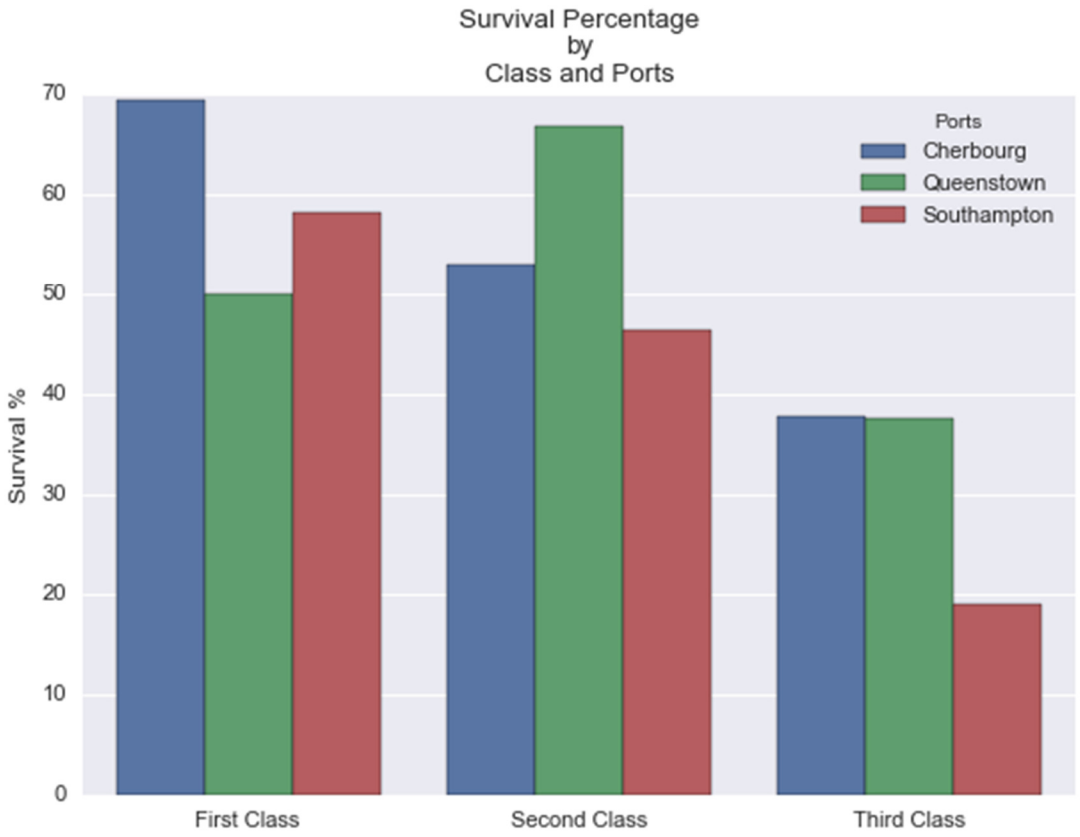
I really appreciate the level of depth with your communication. You do a really great job guiding the reader.

Visualizations made in the project depict the data in an appropriate manner that allows plots to be readily interpreted.



- I noticed `seaborn` being imported but not really utilized in the visualizations.
- The `seaborn` package is extremely useful in many instances to reduce the amount of code you have and in some cases create better visualizations.
- For example, the plot above should not be represented with lines.
- These are discrete variables and there are no values in between the classes or port. A line is used to represent continuous variables that do have values in between, like Age or Height.

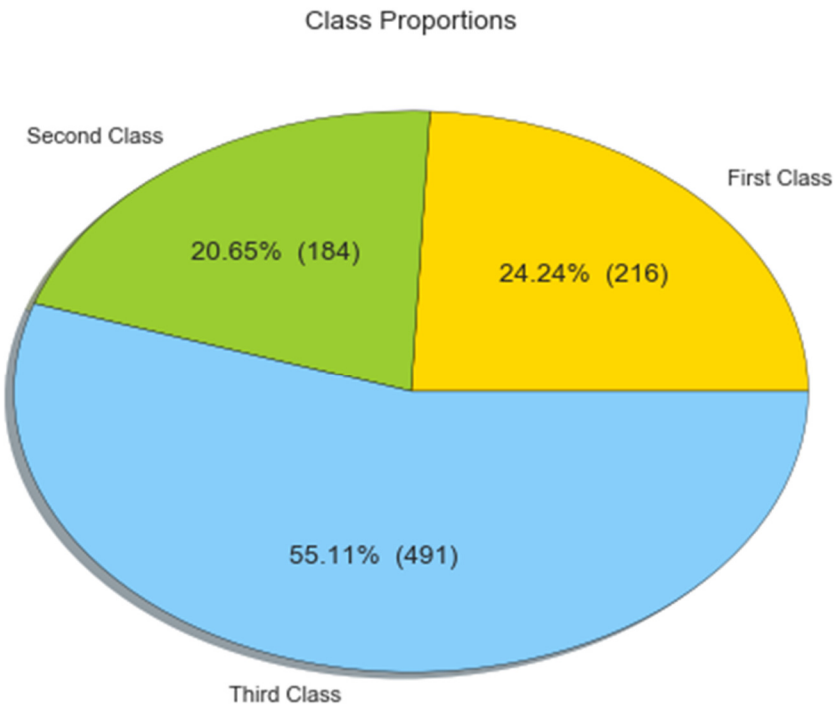
- A bar plot is more appropriate and it is very easy to do in `seaborn`.
  Survival Rate by Pclass and Embarked
- `titanic['Class'] = titanic.Pclass.map({1 : 'First Class', 2 : 'Second Class', 3 : 'Third Class'})`
- `titanic['Ports'] = titanic.Embarked.map({'C' : 'Cherbourg', 'Q' : 'Queenstown', 'S' : 'Southampton'})`
- `df = (titanic.groupby(['Class', 'Ports'])['Survived'].mean()*100).reset_index()`
- `df`
- ```
          Class        Ports   Survived
  0   First Class    Cherbourg  69.411765
  1   First Class   Queenstown  50.000000
  2   First Class  Southampton  58.267717
  3  Second Class    Cherbourg  52.941176
  4  Second Class   Queenstown  66.666667
  5  Second Class  Southampton  46.341463
  6   Third Class    Cherbourg  37.878788
  7   Third Class   Queenstown  37.500000
  8   Third Class  Southampton  18.980170
  ```
- `sns.barplot(data=df,x='Class',y='Survived',hue='Ports')`
- `plt.title('Survival Percentage\nby\nClass and Ports')`
- `plt.ylabel('Survival %')`
- `plt.xlabel('')`


  ```

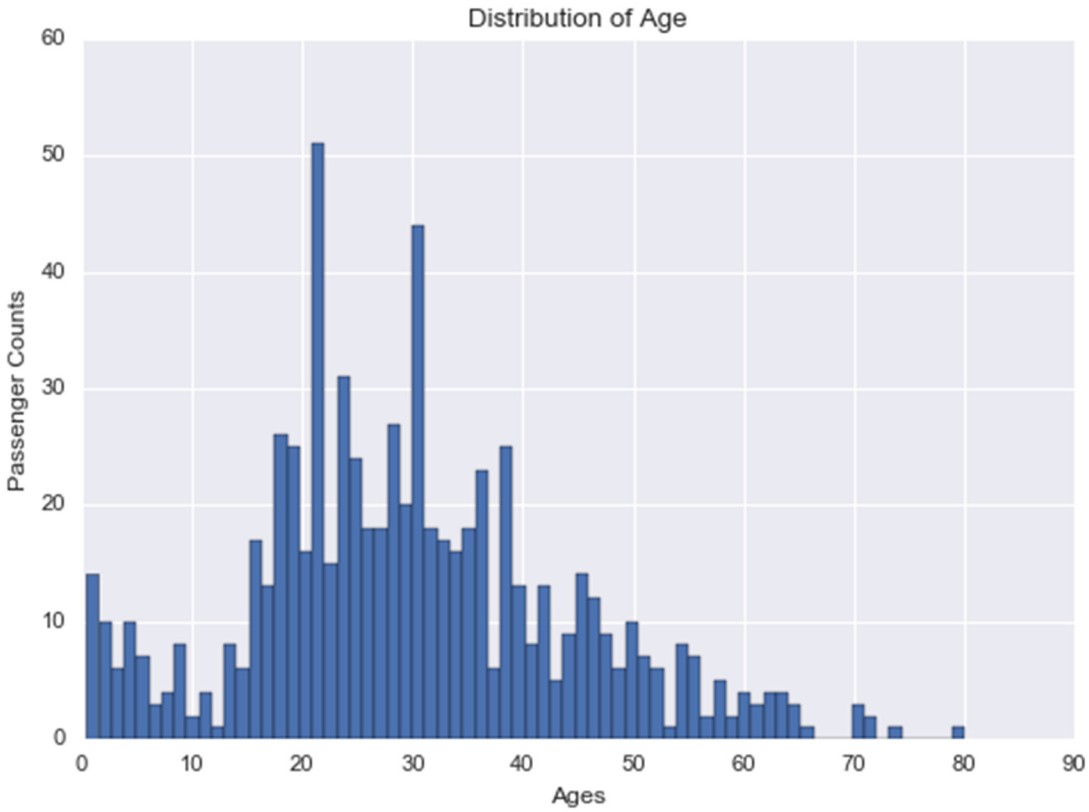
Here are some ideas and demos of popular plotting packages I personally use.

Matplotlib Demos
Seaborn Demos

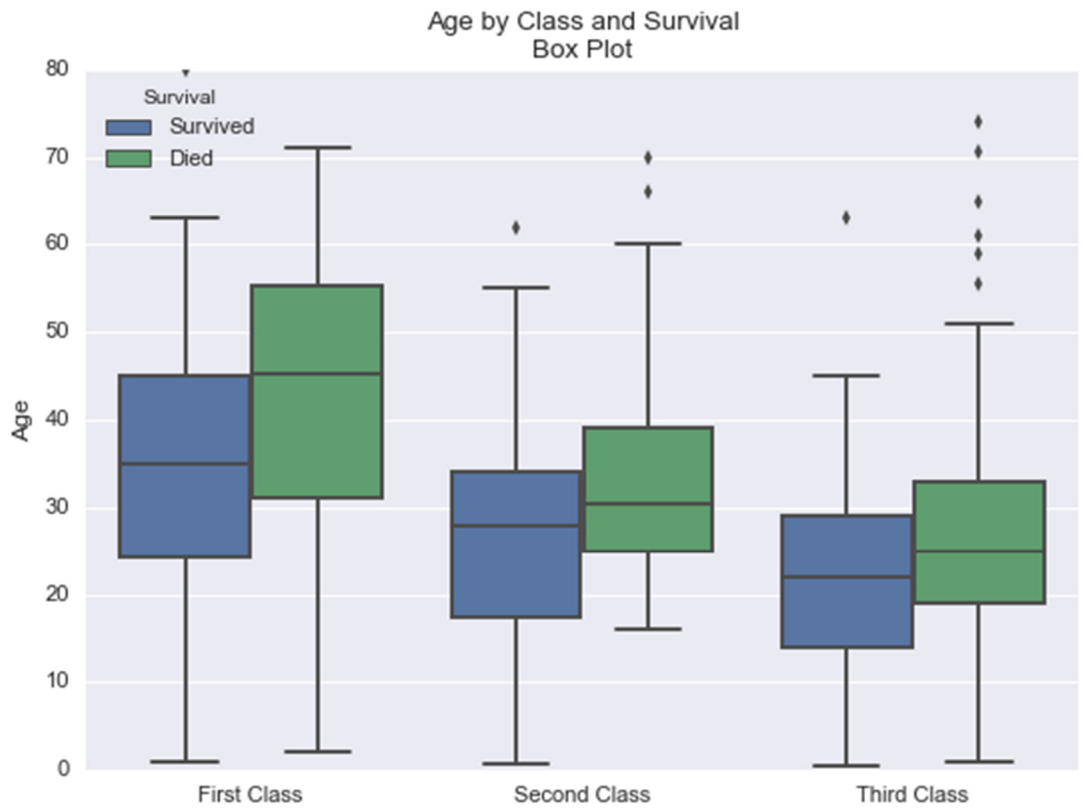
Here are some ideas....



`pandas.plot.pie`

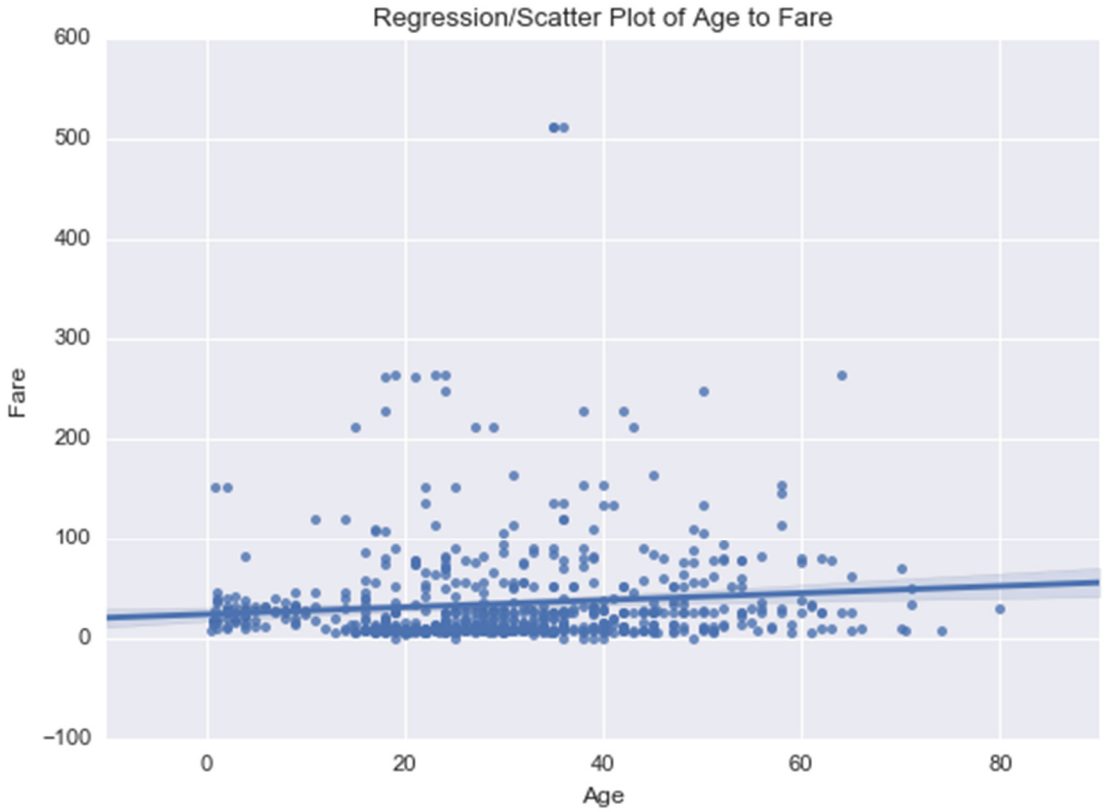



```
seaborn.distplot
hist
```



`seaborn.boxplot`
`pandas.plot.boxplot`

bar
seaborn.barplot



seaborn.regplot
pandas.plot.scatter