UDACITY

PROJECT

Investigate a Dataset

A part of the Data Analyst Nanodegree Program

PROJECT REVIEW

CODE REVIEW

NOTES

SHARE YOUR ACCOMPLISHMENT!

Requires Changes

6 SPECIFICATIONS REQUIRE CHANGES

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.

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.

This is somewhat of a complicated issue. The idea of this project is for students to demonstrate basic investigatory skills, documentations skills, and practicing reproducible research. So, although we appreciate the prediction modeling, Investigate a Dataset requires the majority of the project to be descriptive data, not modeling data. Also, it is uncertain from the presentation how the statistics in the Conclusion section were derived. We need this to be apparent in the report and readers should not need to look at the code in order to figure it out. Through the rest of the sections, I will show you what is required, at minimum, to pass the section. You do not have to take the prediction model out of the project, it just can't be the center-point of the project.

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, however, there is no explicit discussion on the missing values and how you have handled (or not handled) them. Please make a section for this.

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

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.

Typical Exploration Phase Steps

For these projects, we are typically looking at a dependent variable and then trying to analyze the relationship they have with the independent variables in the dataset. The **single-variable** analysis involves looking at the variables **by themselves** not in relation to another variable. Raw counts, histograms, box plots, assertions about distributions are made during this phase and is an important part in finding anomalies such as outliers, skewed distributions, missing values, etc. These anomalies can have profound effects on testing and, therefore, guide us to the correct method for testing.

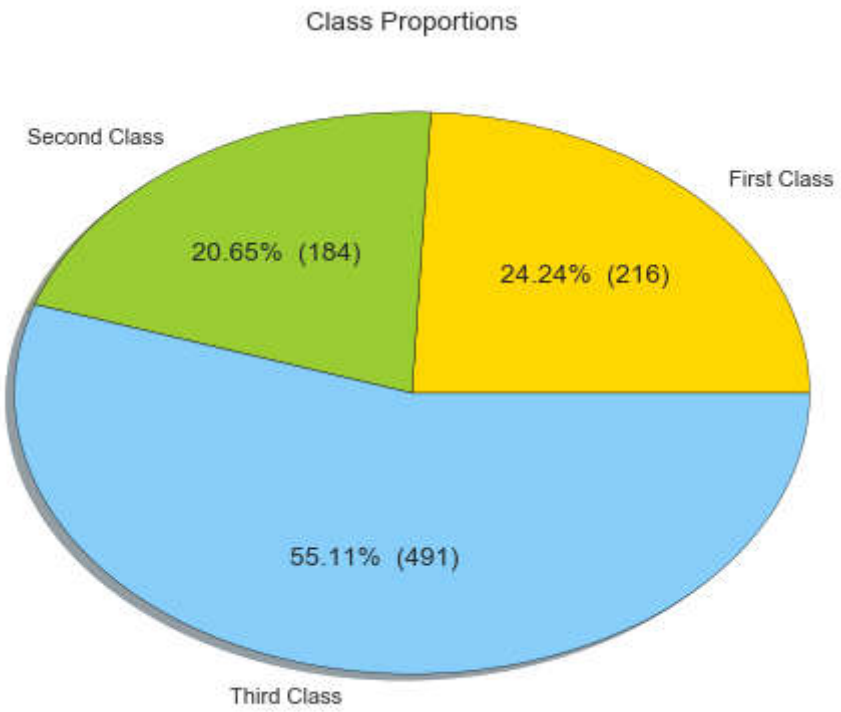
During the **multiple-variable** exploration phase, we start examining each independent variable's relationship to the dependent variable – "How does X affect Y?" In this part, you will hopefully find the independent variables that show close relationship to the dependent variable. The next phase of **multiple-variable** exploration is combining the significant variables together and looking at the relationship to the dependent variable. This fulfills the rubric requirement of investigating **"...stated question(s) from multiple angles."**

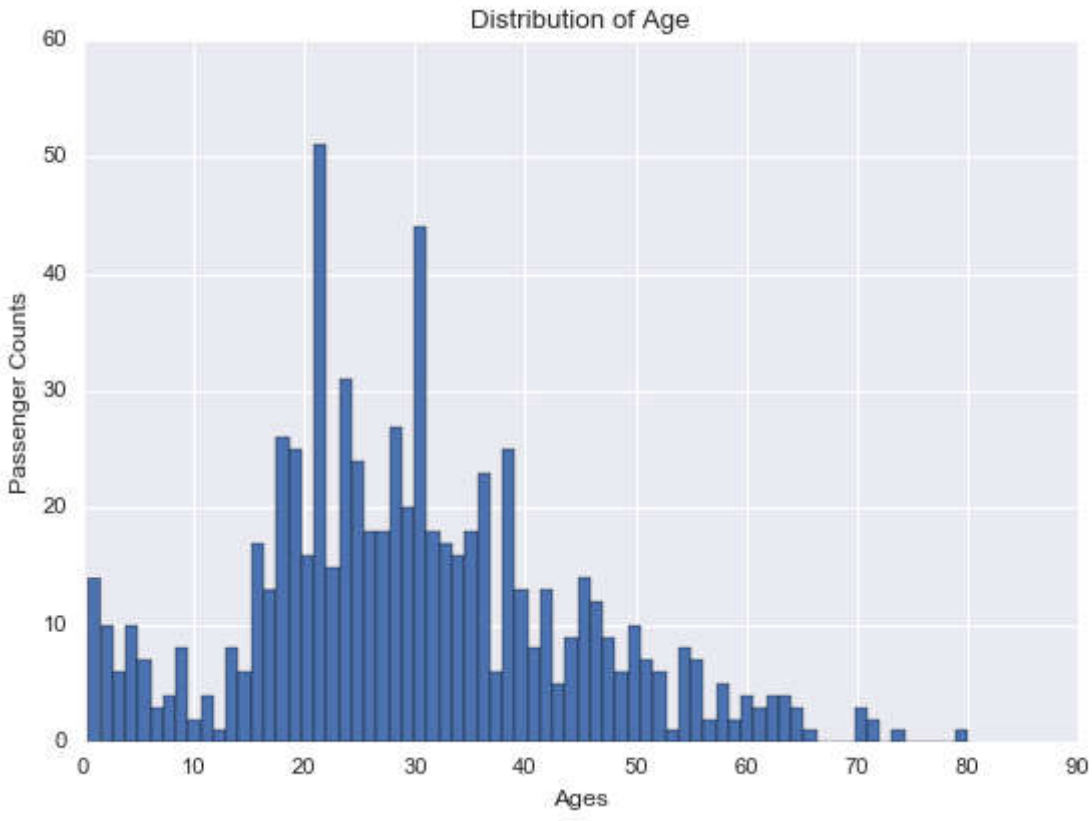
Pertaining to your Analysis

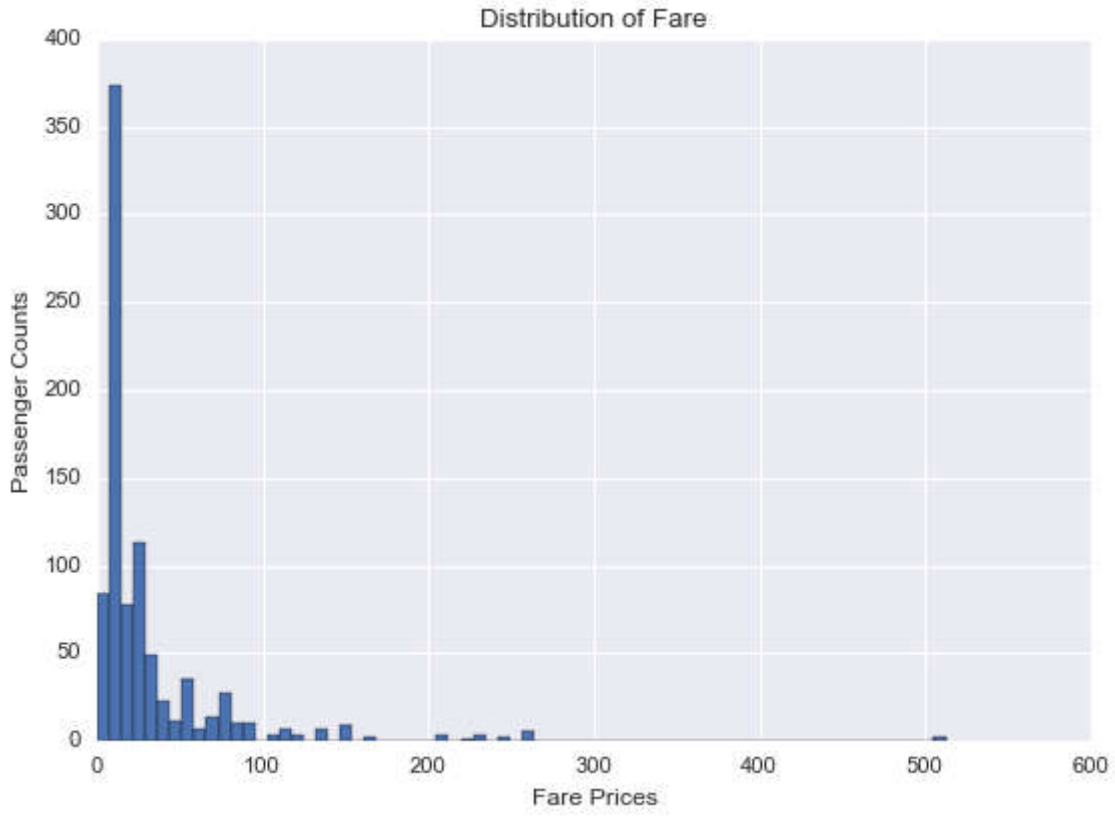
Alright, visualizations are a big part of the Explorations Phase. Here are some ideas to get your rolling.

Remember to comment on the insights. Currently, the statistics provided in the conclusion section doesn't have any visualizations to back them up.

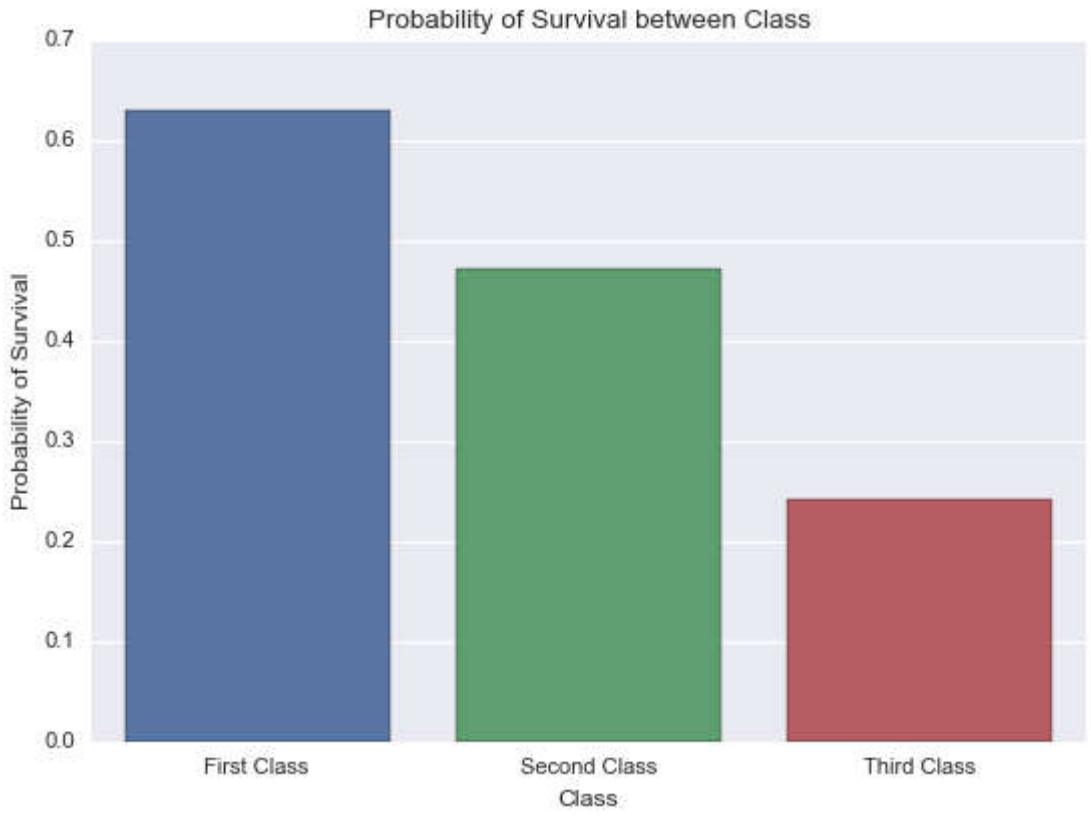
Single

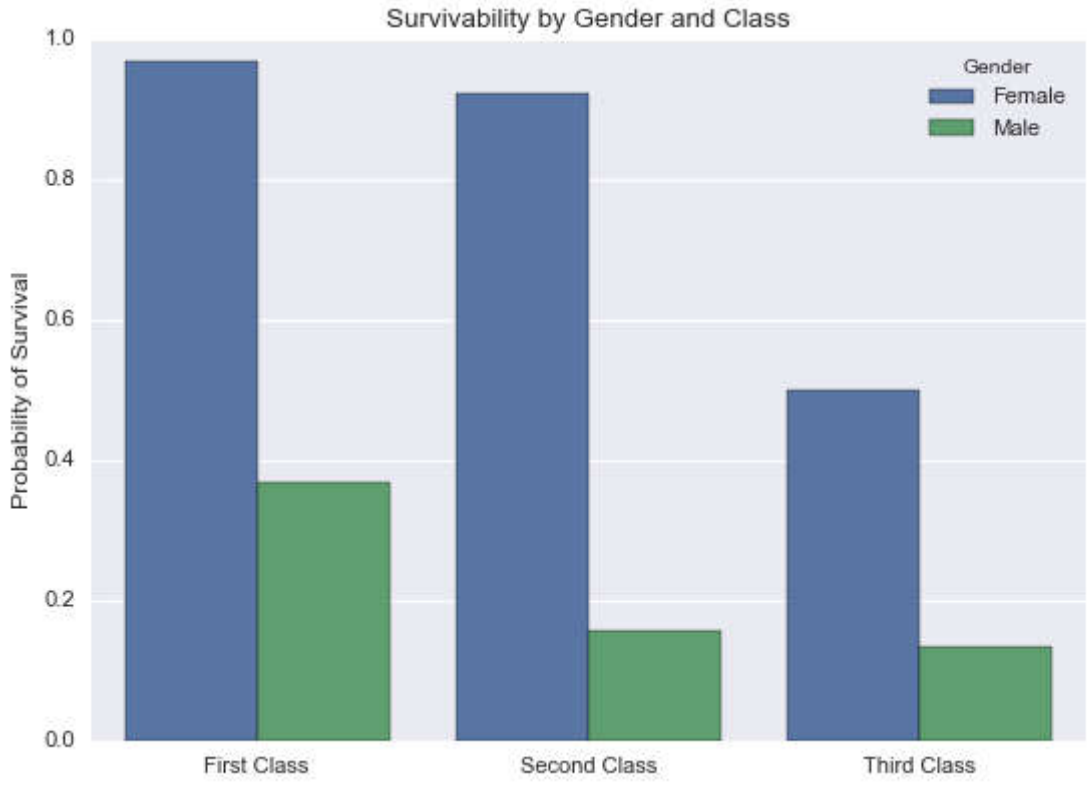





Distribution of Fare
Passenger Counts
Fare Prices
0
50
100
150
200
250
300
350
400
0
100
200
300
400
500
600

Multiple


Probability of Survival between Class
Probability of Survival
Class
First Class
Second Class
Third Class
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7



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.

1. Three variables in particular drives the probability of survival in the Titanic disaster.
    a. Sex
        i. The sex variable increases the probability of survival of a passenger by almost 45% in the 3rd class (7.09% vs 52.7%) and 52% in a first class passenger (88.16% vs 36.45%)
    b. Passenger Class
        i. A first class passenger is passenger is almost thrice as likely to survive the disaster as compared to a third class passenger (50.03% vs 13.04%)
    c. More counter intuitively, port of embarkation in particular, Queenstown
        i. An average man who is of the mean age, mean siblings, mean parent child count, and paid the mean fare is almost twice as likely to survive if his port of embarkation is Queenstown as compared to boarding inn Chebourg (13.04% vs 7.91%)

It is unclear how these statistics were derived. I want to show you a demonstration of how to convey statistics more transparently.

Probability of Survival by Gender

```
titanic.groupby('Gender')['Survived'].mean()
Gender
Female    0.742038
Male      0.188908
```

Probability of Survival by Class

```
titanic.groupby('Class')['Survived'].mean()
Class
First Class     0.629630
Second Class    0.472826
Third Class     0.242363
```

...Both

```
titanic.groupby(['Class', 'Gender'])['Survived'].mean()
Class         Gender
First Class   Female    0.968085
              Male      0.368852
Second Class  Female    0.921053
              Male      0.157407
Third Class   Female    0.500000
              Male      0.135447
```

This is quite significant! But wait! Proportions without the raw counts don't mean anything. 60% could mean 60/100 or 6/10, the former being more significant by a factor of 10.

```
titanic.groupby(['Class', 'Gender', 'Survival']).apply(len)
Class         Gender  Survival
First Class   Female  Died          3
                      Survived     91
              Male    Died         77
                      Survived     45
Second Class  Female  Died          6
                      Survived     70
              Male    Died         91
                      Survived     17
Third Class   Female  Died         72
                      Survived     72
              Male    Died        300
                      Survived     47
```

- This is a typical path of exploration.
- We must include the raw counts since proportions only give one side of the story.
- This should be done for the variables you have chosen.
- Remember to be thorough.

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.

I do not see an explicit discussion dealing with the limitations of the analysis. There are always limitations when analyzing a limited dataset. It is important for analysts to scrutinize their own analysis for integrity. Please include a section explicitly dedicated to discussing the limitations of the analysis. Here are some ideas to talk about....

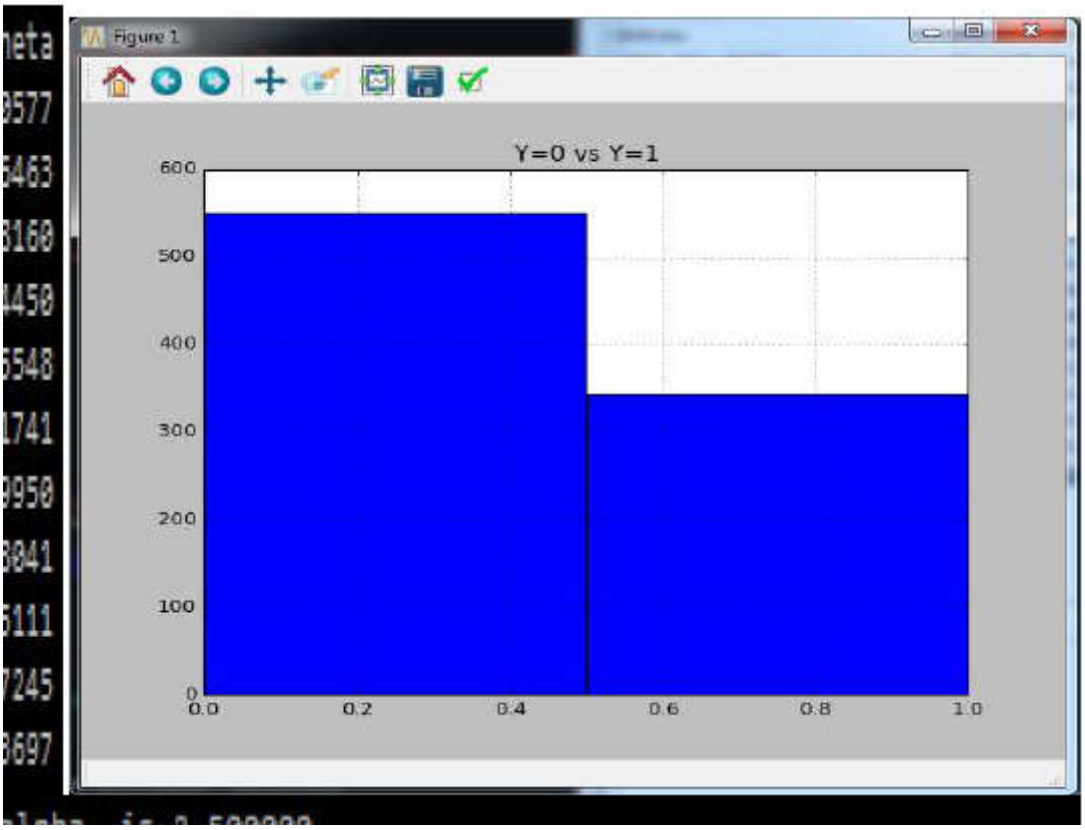
- The dataset is filled with missing values. Whatever way we choose to handle these missing values (omitting, imputing, etc.) presents its own pros and cons.
- What are the limitations of making assumptions without statistical testing (t-test, z-test, etc.)?
- Could there by other variables not included in the dataset that could have been useful in the analysis?

These are just some ideas. Please explore for yourself.

Communication Phase

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

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

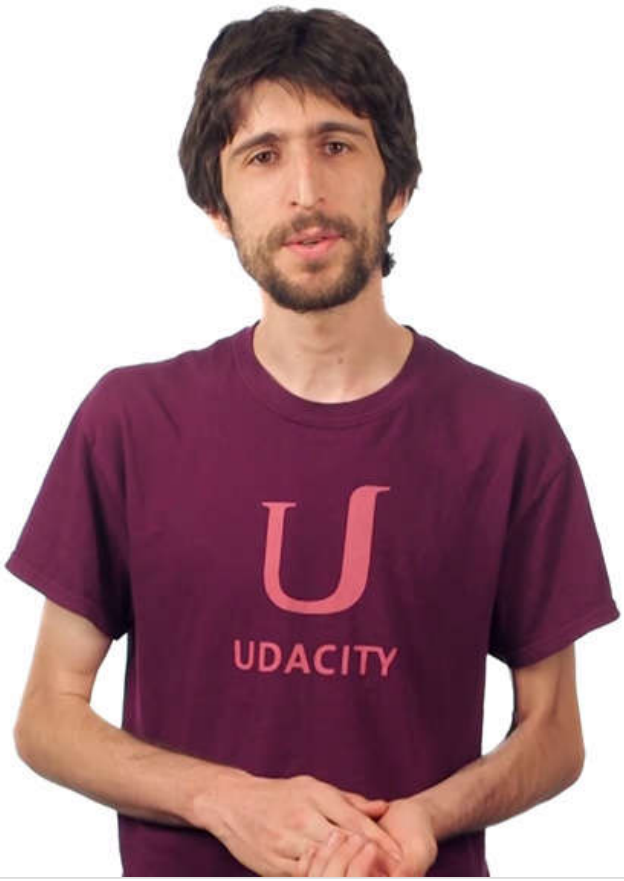


- Readers should be able to know what is happening in the plot.
- Plots need to be self-explanatory and should not rely on displayed code or descriptions for interpretability.

- Please include **descriptive** axis labels and Titles for all visualizations.
- Be sure to be meticulous in detailing your plots making sure everything is readable.

RESUBMIT

DOWNLOAD PROJECT



Best practices for your project resubmission

Ben shares 5 helpful tips to get you through revising and resubmitting your project.

Watch Video (3:01)

Have a question about your review? Email us at review-support@udacity.com and include the link to this review.

RETURN TO PATH

Student FAQ