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

The project is already very well done and only requires some small tweaks required by the project guidelines. I have made suggestions in the appropriate evaluation field. Once these areas are addressed, the project should be complete and ready for portfolio submission. Keep up the great work, you are almost there!

> "Genius is one percent inspiration and ninety-nine percent perspiration."

-- Thomas Alva Edison

Code Functionality

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

Please note The project instructions require the analysis be presented in either a PDF or HTML document. I have reviewed your project through the submitted `.ipynb` file in order to save unnecessary submission attempts, however, `.ipynb` is not a universal file format for documentation and cannot be opened on all

machines. It is fortunate that I am able to open, however, this may not be the case for other reviewers. If this was a GitHub repo link, please note that even though GitHub has native support for `.ipynb` files, reviewers do not get access to the repo. Instead, Udacity downloads the contents of the repo and sends us an archive. This is for the privacy of your other repos. For future submissions, please submit scripts **as well** as the analysis in PDF or HTML formats.

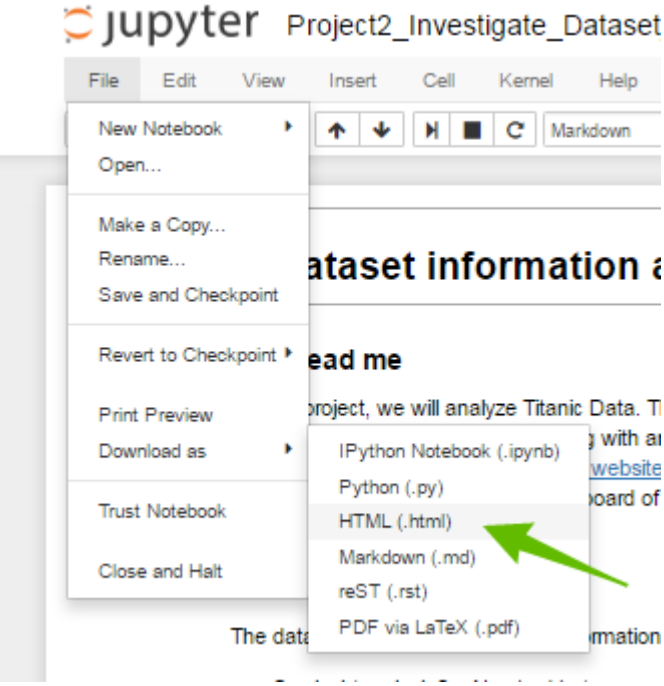
Submission

What to include in your submission

1. A PDF or HTML file containing your analysis. This file should include:
   - A note specifying which dataset you analyzed
   - A statement of the question(s) you posed
   - A description of what you did to investigate those questions
   - Documentation of any data wrangling you did
   - Summary statistics and plots communicating your final results
2. If the code you used to perform your analysis is not included in the above, you should submit the code separately in `.py` file(s).
3. A list of Web sites, books, forums, blog posts, github repositories, etc. that you referred to or used in creating your submission (add N/A if you did not use any such resources).

IPython notebook instructions

If you used IPython notebook to create your analysis, you can include your code directly in the notebook and do not need to submit it separately. To download your notebook as an HTML file, click on File -> Download.As -> HTML (.html) within the notebook. If you get an error about "No module name", then open a terminal and try installing the missing module using `pip install <module_name>` (don't include the "<" or ">" or any words following a period in the module name).



https://www.udacity.com/course/viewer#!/c-nd002/l-3176718735/m-5464560482

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.

```
# Number of people embarked
```

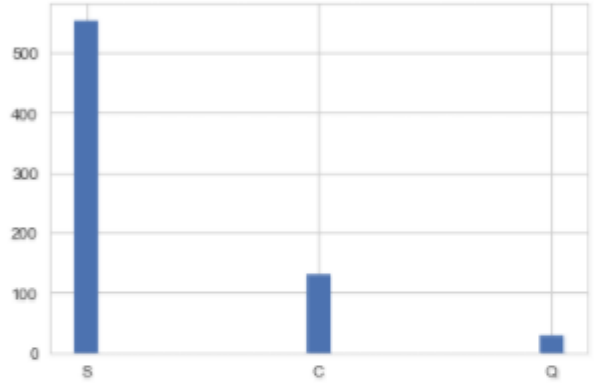
```
# on S, C and Q.
S_list = []
C_list = []
Q_list = []

for value in new_data['Embarked']:
    if value == 'S':
        S_list.append(value)
    elif value == 'C':
        C_list.append(value)
    elif value == 'Q':
        Q_list.append(value)

x = np.arange(3)
plt.bar(x, height=[len(S_list), len(C_list), len(Q_list)], width=0.1)
plt.xticks(x, ['S', 'C', 'Q'])

print "Passengers from S : ", len(S_list)
print "Passengers from C : ", len(C_list)
print "Passengers from Q : ", len(Q_list)
```

```
Passengers from S :  554
Passengers from C :  130
Passengers from Q :  28
```



This shows that there were most passengers from S followed by C and Q.

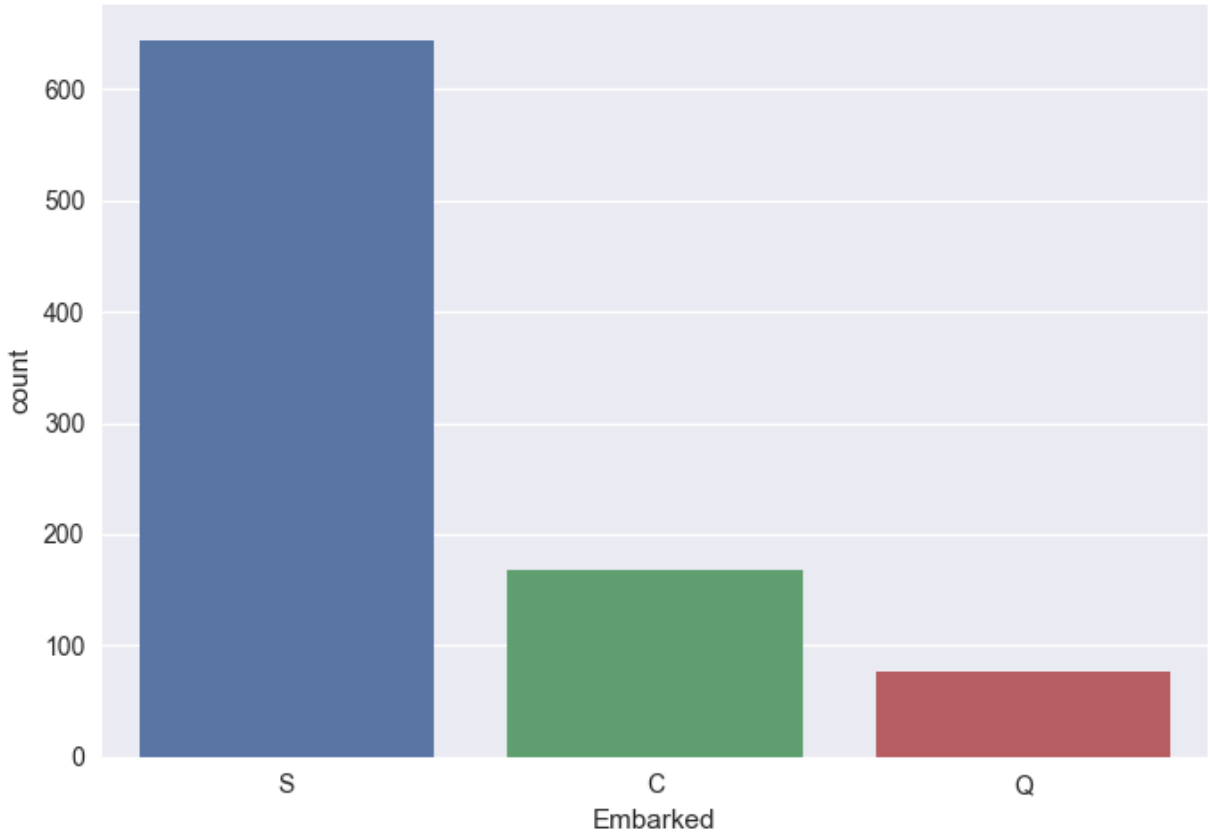
```
# Finding the total female and male passengers
import seaborn as sns
female = 0
male = 0
total_passengers = len(new_data)
for people in new_data['Sex']:
    if people == 'female':
        female = female + 1
    elif people == 'male':
        male = male + 1

sns.countplot(x="Sex", data=new_data)
print "Total male passengers: ", male
print "Total female Passengers: ", female
```

```
Total male passengers:  453
Total female Passengers:  261
```

- **Required** Loops are not needed for this. You can directly use sns.countplot like this.

```
sns.countplot(data = titanic, x = 'Embarked')
```



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.

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. This is required regardless if the variables are being explored or not. It gives the reader an idea of the dataset's *health*. Please make a section for this.

- 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 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.

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?
- Does the investigation distinguish between **Correlation** and **Causation**?
  Correlation does not Imply Causation
  Funny Spurious Correlations

These are just some common ideas students have talked about. You may or may not find these limitations are applicable to your project but it is important to look for them.

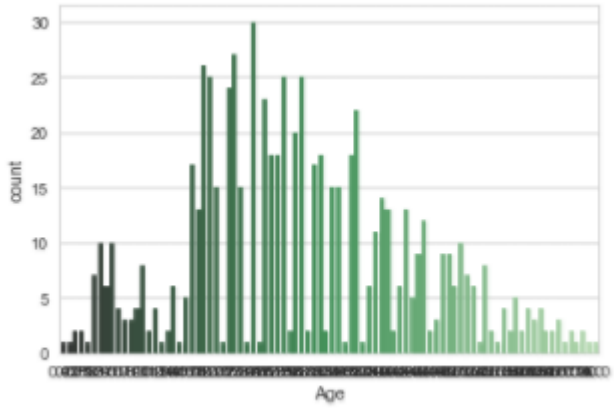
Communication

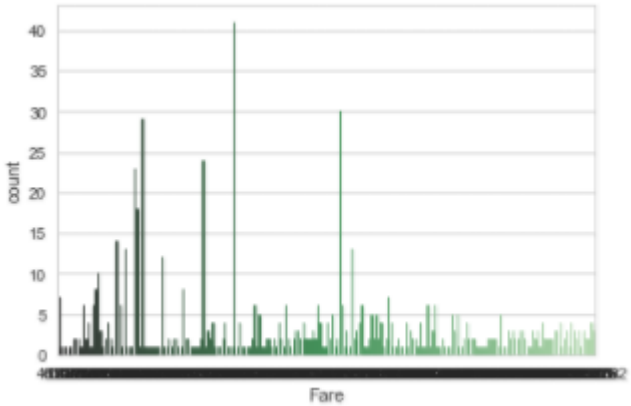
Reasoning is provided for each analysis decision, plot, and statistical summary.

There are a lot of stats and visuals but very little commenting in each of these sections. Remember to guide your reader. These projects will be submitted to your Job-Ready portfolio and the content, reasoning, and decisions should be clear to the reader and presented in a polished manner. Please talk about every section and answer the following.....

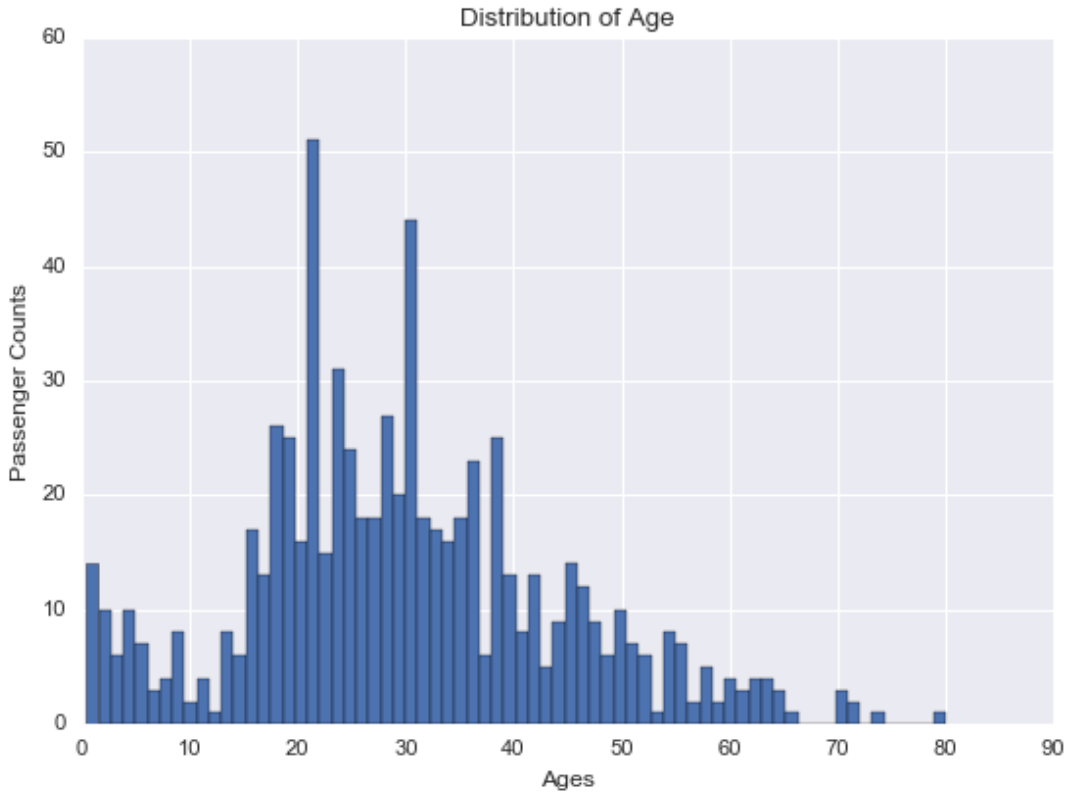
1. Why have you chosen/rejected certain variables?
2. Why did you perform this step? *Because it was required is not sufficient :)*
3. What insight have you derived from it?
4. Are there limitations?

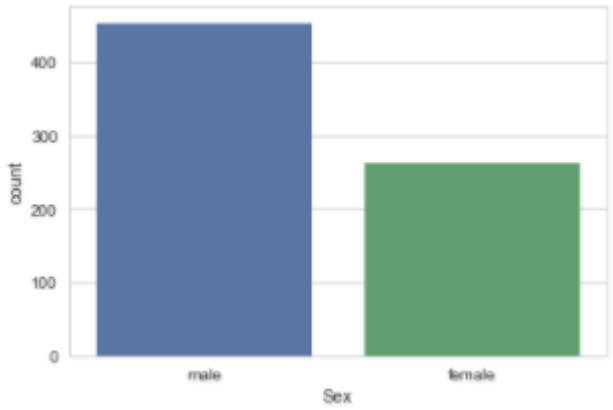
Visualizations made in the project depict the data in an appropriate manner that allows plots to be readily interpreted.



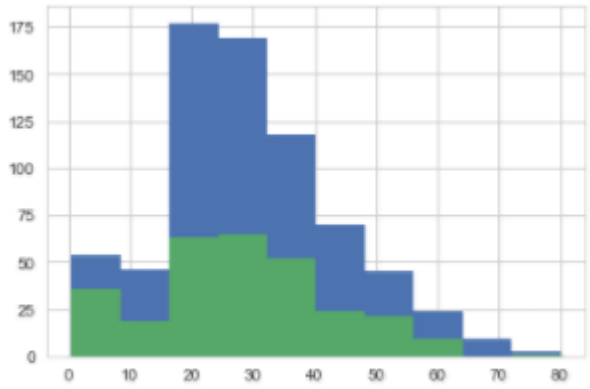


- **Required** A countplot is not appropriate for this since the x axis values are completely unreadable. A histogram would be ideal.
  Example

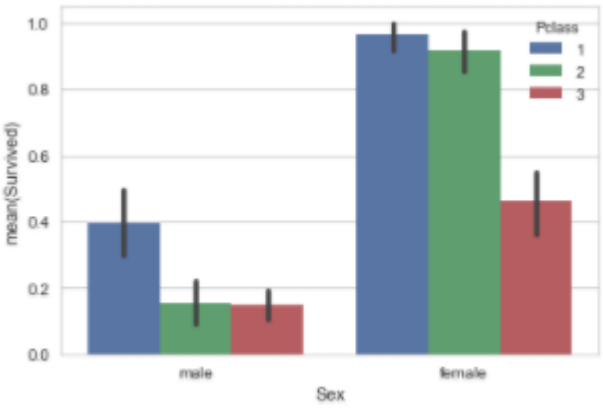




- **Required** Is this a subset of gender, is this gender distribution of the entire dataset. We don't know because the plot does not have a title for context. All axes should be labeled and all plots should have titles.



- **Required** All plots should have titles
- **Required** All axes should be labeled



- **Required** mean(survived) should be renamed Survival Rate or Survival Proportion since that is what that axes really is.

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)

RETURN TO PATH

Student FAQ