UDACITY

PROJECT

Investigate a Dataset

A part of the Data Analyst Nanodegree Program

PROJECT REVIEW

CODE REVIEW

NOTES

SHARE YOUR ACCOMPLISHMENT!

Meets Specifications

Code Functionality

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

Please avoid from including very long tables in the report, it distracts your readers from the main objective of the report. Instead consider using summary statistics or .info() or .describe().

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 analysis makes use of the NumPy and Pandas libraries, vector operators are employed instead of loops and lists.

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.

The report states clear and relevant questions that are being addressed by the following 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.

Well Done for identifying the missing values in the dataset and documenting the changes for the data set. This is important because others will be able to repeat your analysis if needed.

Please consider to include a simple histogram (continues features) or barplot (categorical features) for each feature that is included in the analysis. That will allow you to describe the distribution, identify outliers and make changes to the data set before starting the analysis.

Please note that in most cases more data will benefit the analysis and the results, so if it is possible please prefer to work with all the data.

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.

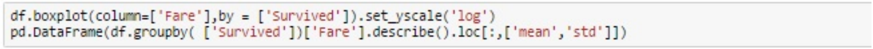
The analysis makes use of both single-variable (1d) and multiple-variable (2d) explorations to investigate the age survived, gender, pclass and other features.

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.

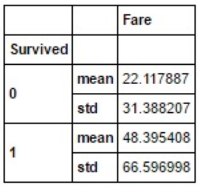
At least two kinds of plots should be created as part of the explorations.

The analysis uses different chart type to explore the dataset and depict the results and insights. Other chart types that can be used here,

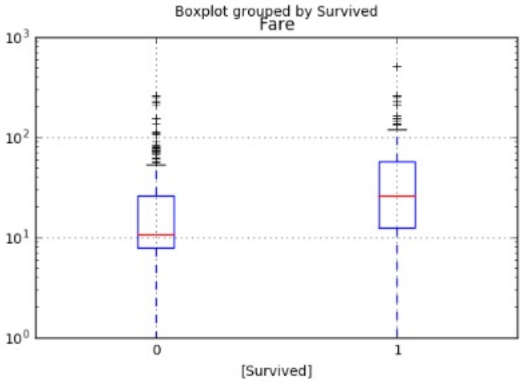
A box plot depict the distribution of a continues feature for different categories,



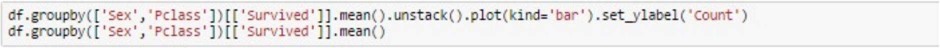
```
df.boxplot(column=['Fare'],by = ['Survived']).set_yscale('log')
pd.DataFrame(df.groupby( ['Survived'])['Fare'].describe().loc[:,['mean','std']])
```



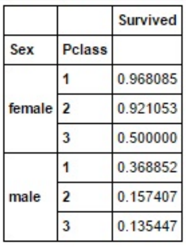
		Fare
Survived		
0	mean	22.117887
	std	31.388207
1	mean	48.395408
	std	66.596998



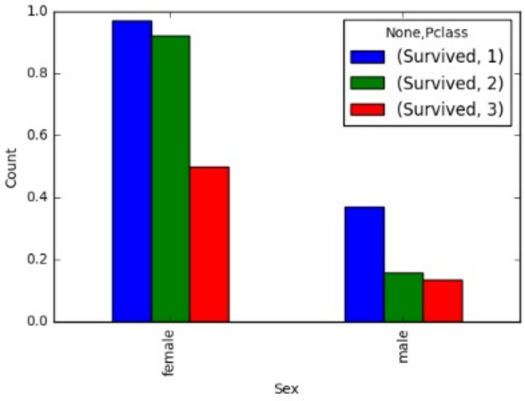
Bar plot can depict the relation between more than 2 features,



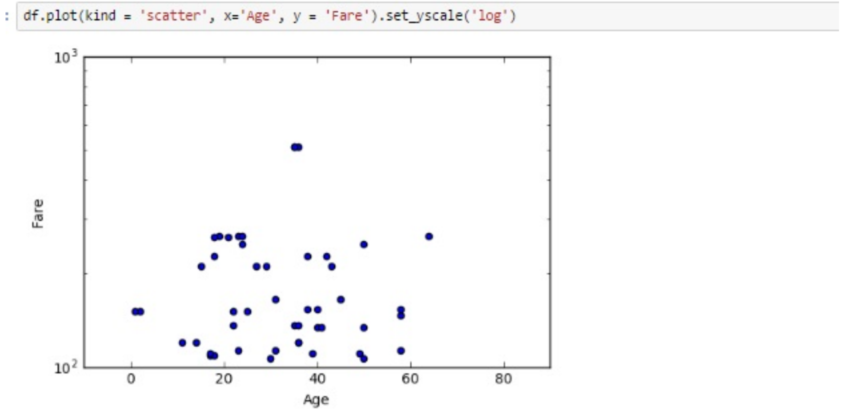
```
df.groupby(['Sex','Pclass'])[['Survived']].mean().unstack().plot(kind='bar').set_ylabel('Count')
df.groupby(['Sex','Pclass'])[['Survived']].mean()
```



		Survived
Sex	Pclass	
female	1	0.968085
	2	0.921053
	3	0.500000
male	1	0.368852
	2	0.157407
	3	0.135447



Scatter plot to depict the relation between 2 continues features,



```
: df.plot(kind = 'scatter', x='Age', y = 'Fare').set_yscale('log')
```

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.

The report discusses the limitations and shortcomings of the dataset and the analysis.
Optionally you can calculate the significance with the appropriate statistical test
To calculate the significance of survival rate for different categories you can use the chi-square test
http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chisquare.html

Another statistical test that is useful to appreciate if the distribution for different categories is different is the Mann-Whitney rank test
http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html

Communication

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

The analysis follows a logical flow, the discussion includes reasonings, explanations about the analysis and relevant statistics to quantify the results and insights.

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

DOWNLOAD PROJECT

RETURN TO PATH

Student FAQ