




PROJECT

Investigate a Dataset
A part of the Data Analyst Nanodegree Program

PROJECT REVIEW	CODE REVIEW	NOTES
<div>Meets Specifications<div>SHARE YOUR ACCOMPLISHMENT</div><div></div><p>Dear student,</p><p>well done addressing the last remaining issue and completing your submission. Aside from a Pro Tip regarding missing values and an advanced one regarding hypothesis testing, there's nothing relevant I can further add to your good work.</p><p>Congratulations on passing your exam!</p></div>		
<div>Code Functionality</div> <div><div>All code is functional and produces no errors when run. The code given is sufficient to reproduce the results described.</div><div>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.</div><div>The code makes use of functions to avoid repetitive code. The code contains good comments and variable names, making it easy to read.</div></div>		
<div>Quality of Analysis</div> <div><div>The project clearly states one or more questions, then addresses those questions in the rest of the analysis.</div><div>Well done proposing some interesting questions, asking the right question is really the most important thing when analysing data, you might find interesting the following article: http://www.datapine.com/blog/data-analysis-questions/</div></div>		
<div>Data Wrangling Phase</div> <div><div>The project documents any changes that were made to clean the data, such as merging multiple files, handling missing values, etc.</div><div>Pro Tip: There are several options to deal with missing values like: a. Replacing the values with means or medians. b. Remove the features that have an exceeding number of missing values. c. More complex approaches rely on analysing the distribution of missing values: https://en.wikipedia.org/wiki/Missing_data http://scikit-learn.org/stable/modules/preprocessing.html</div></div>		
<div>Exploration Phase</div> <div><div>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.</div><div>You might want to enrich your report by taking advantage of the great variety of plots available in Python, here are some useful sources: https://matplotlib.org/ examples: https://matplotlib.org/gallery.html https://seaborn.pydata.org examples: https://seaborn.pydata.org/examples/index.html</div><div>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.</div><div>At least two kinds of plots should be created as part of the explorations.</div><div>Advanced Pro Tip: statistically testing survival in groups This is beyond what is required here though you might be interested in testing whether some groups had a statistically significant higher survival rates than others. This might be quite tricky as survival is a categorical variable. T-tests, Z-tests, ANOVA, F-test, and other parametric tests assume normality of the dependent variables being tested (survival). A categorical variable cannot be distributed normally and doesn't have the properties of a normal variable. (For instance we can't really average survived with not survived as the result would have no meaning, the same applies to standard deviations and other statistics). If you want to perform a test regarding this you should use a contingency table and a Chi-Square test, which is used to compare categorical variables. http://www.statisticshowto.com/what-is-a-contingency-table/ https://en.wikipedia.org/wiki/Contingency_table</div></div>		
<div>Conclusions Phase</div> <div><div>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.</div></div>		
<div>Communication</div> <div><div>Reasoning is provided for each analysis decision, plot, and statistical summary.</div><div>Visualizations made in the project depict the data in an appropriate manner that allows plots to be readily interpreted.</div></div>		
<div> DOWNLOAD PROJECT</div>		