Data Analyst Nanodegree Program

Udacity Nanodegree (In Collaboration with Kaggle)
Course-1: Introduction to Data Analysis
Project-1: Investigate a Dataset

Meets Specifications.
Congratulations! You have made it! You have put great effort into this project. Keep up the good work. Good luck with your future Projects! Happy Learning!

Code Functionality :
All code is functional and produces no errors when run. The code given is sufficient to reproduce the results described. Well done! All code is functional and runs without any errors.

Learning Notes:

- Jupyter Notebook Tips and Tricks
- Markdown Cheatsheet

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. Awesome! You have used Pandas dataframe and series, and built-in functions for your analysis e.g. head(), dtypes, shape, info(), drop(), describe(), nunique(), isnull(), etc.

Learning Notes: For more details about the Pandas' functions and vectorized operations. You may refer to the below links:

- Important Pandas' Functions
- Essential Basic Functionality of Pandas
- Pandas Tutorials
- Vectorization And Parallelization In Python With Numpy And Pandas
- A Beginner's Guide to Optimizing Pandas Code for Speed

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.

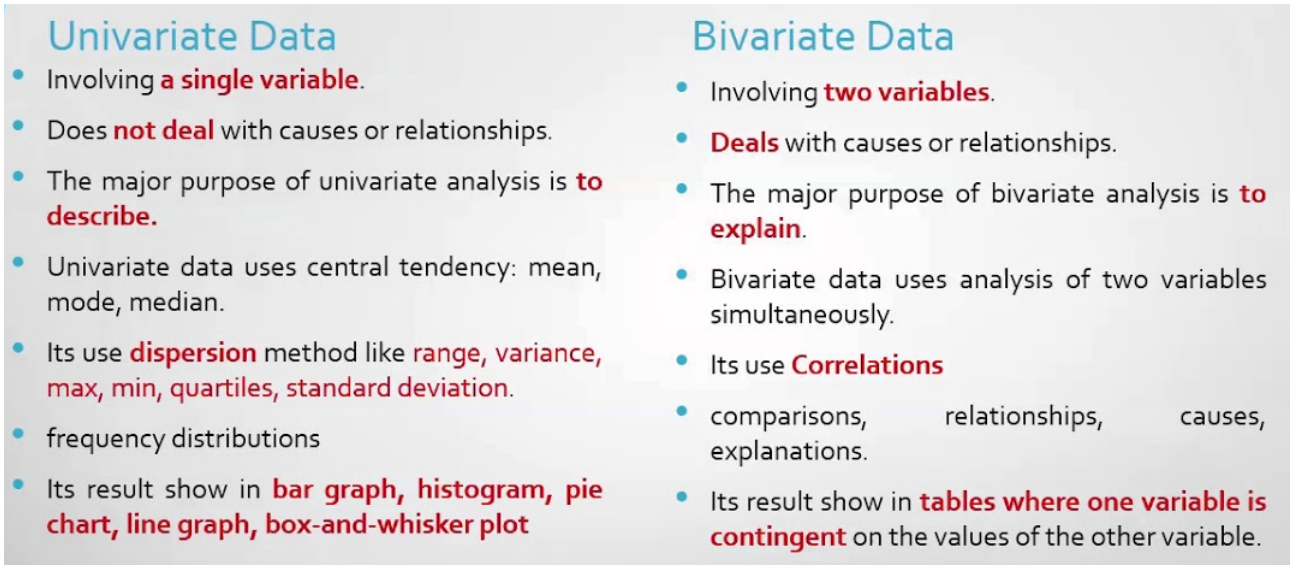
Data Wrangling Phase:
The project documents any changes that were made to clean the data, such as merging multiple files, handling missing values, etc. Well done! You have documented the data cleaning done on the dataset such as handling missing values, removing extraneous columns, modifying data types, checking of duplicate values, etc.

Learning Notes: For more details about handling missing values and merging multiple files. You may refer to the below links:

- Handling Missing Data
- How to handle missing values
- Merging files in Pandas

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.
Good job! You have investigated the stated questions from multiple angles using both single-variable (1d) and multiple-variable (2d) explorations.

Learning Notes: For further reading on this topic, you may refer to this blogpost:

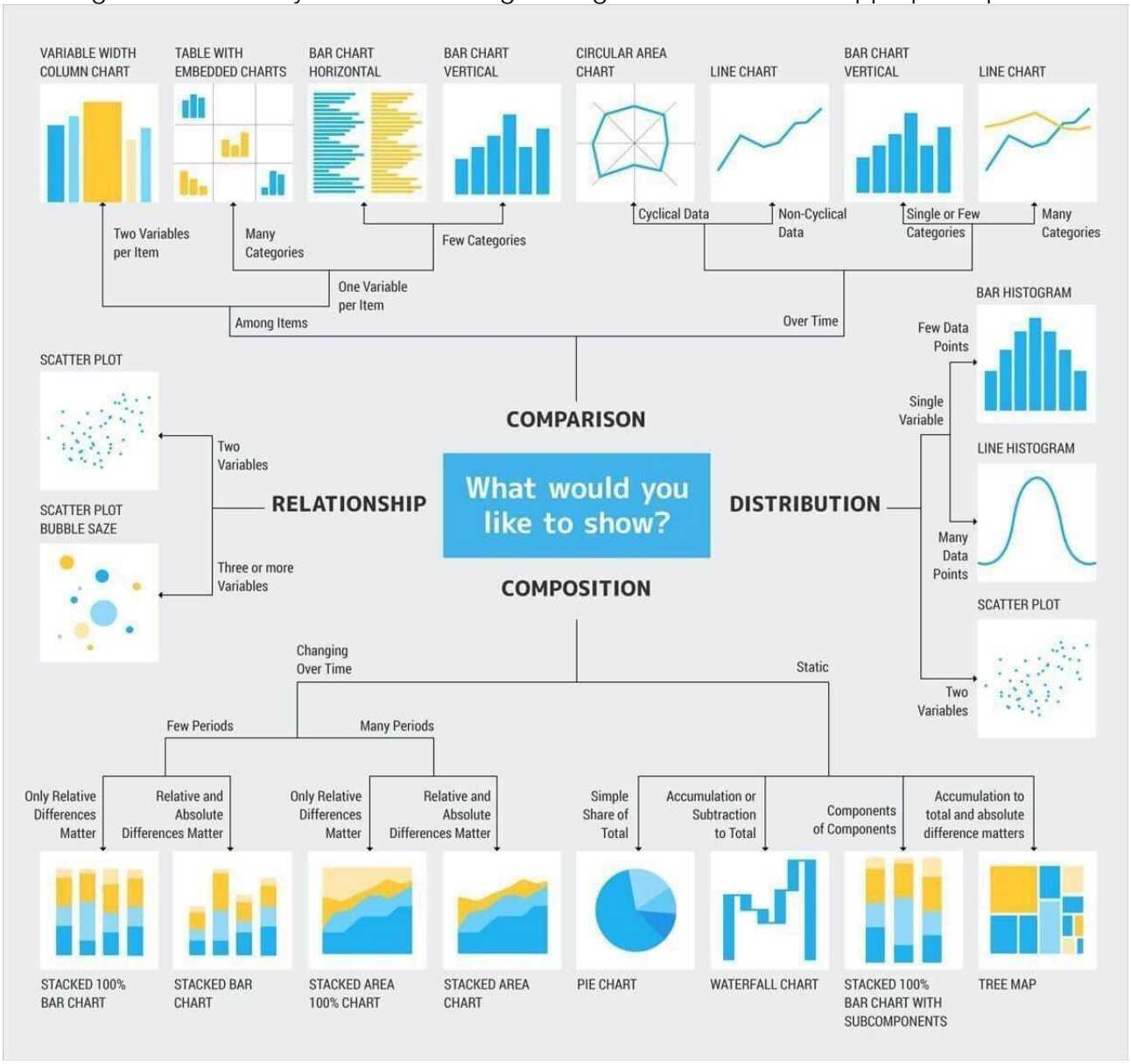


- A Comprehensive Guide to Data Exploration

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.

Good job! You have done a great analysis. Using various visualizations techniques(bar chart, histogram, and scatter plot), multiple comparisons and trends are shown. You have calculated the relevant statistics throughout the analysis.

Learning Notes: You may use the following chart guide to choose the appropriate plots:



For more details, you may refer this
- Visualization using Pandas

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. Well done! You have included the results of the analysis in the conclusion section, also mentioning its limitations.

Suggestions:
Mentioning the limitations of the result of the analysis is important in the data analysis process. Limitations that we face during data analysis could be as below:
- Incomplete data/Any duplicates or outliers in the data/Missing values in the data.
- Any inconsistencies or errors present in the data.
- Constraints of the method of analysis etc.

Also, solely based on correlation we should not conclude causation. You may refer to the below links to understand this in details:
- Correlation vs. Causation
- Correlation does not imply causation

Communication:
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. Awesome! You have included appropriate visualizations with appropriate plot title, axes title in your analysis. This makes plots more readable and could be easily interpreted.

Learning Notes: For basic details about the Matplotlib below blogpost is a good read along with documentation.

- Data Visualization in Python using Matplotlib