UDACITY

‹ Return to "Data Analyst Nanodegree" in the classroom

DISCUSS ON STUDENT HUB

Investigate a Dataset

REVIEW

HISTORY

Requires Changes

1 SPECIFICATION REQUIRES CHANGES

Very impressive work ! your project reflects your hard work and I have to congratulate you for that 😄 Your code is very solid as well, you only need some modifications order to continue. Good luck in your next submission !

Don't hesitate to reach your Student Hub mentor or use the slack channel in order to get help, we are here to help you succeed 🏆

Code Functionality

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

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.

Good work
Suggestion

Here are a few Pandas built-in methods that are very useful for exploring variables in this project:

- Boolean-Indexing
- Group-by
- Value-Counts
- Series.map
- Working-with-text-data

The code makes use of functions to avoid repetitive code. The code contains good comments and variable names, making it easy to read.

Excellent job! solid code and well documented 👍

Quality of Analysis

The project clearly states one or more questions, then addresses those questions in the rest of the analysis.

Very deep and important questions 🏆

Data Wrangling Phase

The project documents any changes that were made to clean the data, such as merging multiple files, handling missing values, etc.

Good work in implementing a Data Wrangling Phase
Suggestion

The most important aspect of Data Wrangling is to clean or transform the data preparing it for analysis.

One main issue is having missing data while conducting analysis, which can provide skew/bias results. Luckily there are a few methods that Pandas provide to deal with these issues:

- The first thing to do is to always Identify the missing values within the dataset. The few steps after this explain how to deal with the missing data
- If there are columns with a few rows of missing data the Dropna method could be used to drop the missing rows.
- If there are rows with missing data the Fillna-method can be used instead of dropping them completely (This method can vary with the data and the project)
- The final option is if there are way too many missing values within a column it is best to drop the column completely using the Drop-column-method

Data Wrangling does not only involve Identifying and dealing with missing values but also involves in transforming the data to a more effective state to target the analysis. Here are other wrangling methods:

- Binning or Cutting Groups continuous or numerical values into smaller groups or 'bins'
- Pandas-Dummies Transforms categorical data into dummy/indicator variables

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!

This link summarises the difference between bivariate and univariate data.

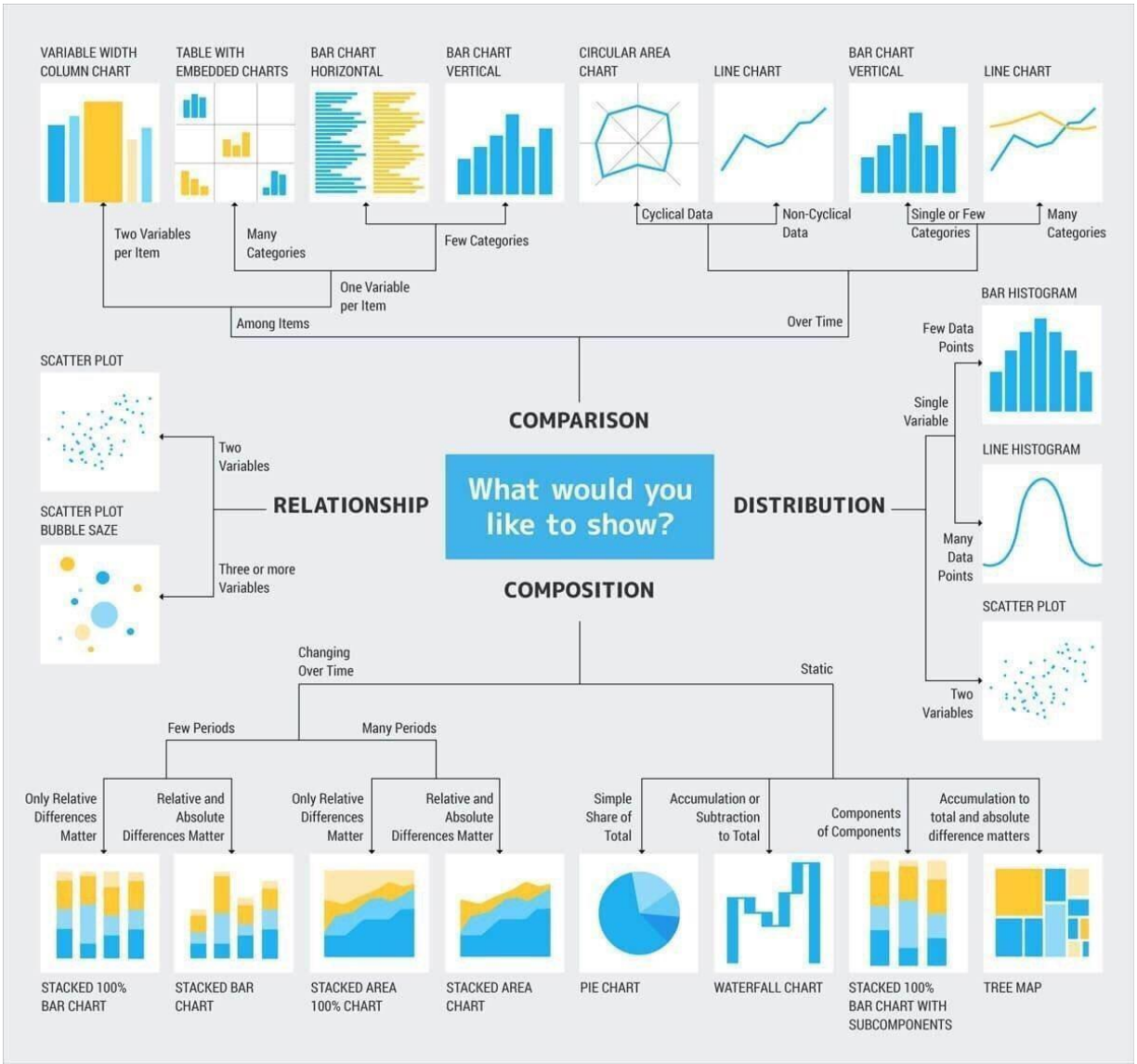
Univariate Data	Bivariate Data
• involving a **single variable**	• involving **two variables**
• does not deal with causes or relationships	• deals with causes or relationships
• the major purpose of univariate analysis is to describe	• the major purpose of bivariate analysis is to explain
• central tendency - mean, mode, median • dispersion - range, variance, max, min, quartiles, standard deviation. • frequency distributions • bar graph, histogram, pie chart, line graph, box-and-whisker plot	• analysis of two variables simultaneously • correlations • comparisons, relationships, causes, explanations • tables where one variable is contingent on the values of the other variable. • independent and dependent variables
Sample question: How many of the students in the freshman class are female?	**Sample question:** Is there a relationship between the number of females in Computer Programming and their scores in Mathematics?

And here's some more inspirations for you: https://seaborn.pydata.org/tutorial/categorical.html

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.

Very good ! for future projects let me recommend you these tools to choose your visualizations



Conclusions Phase

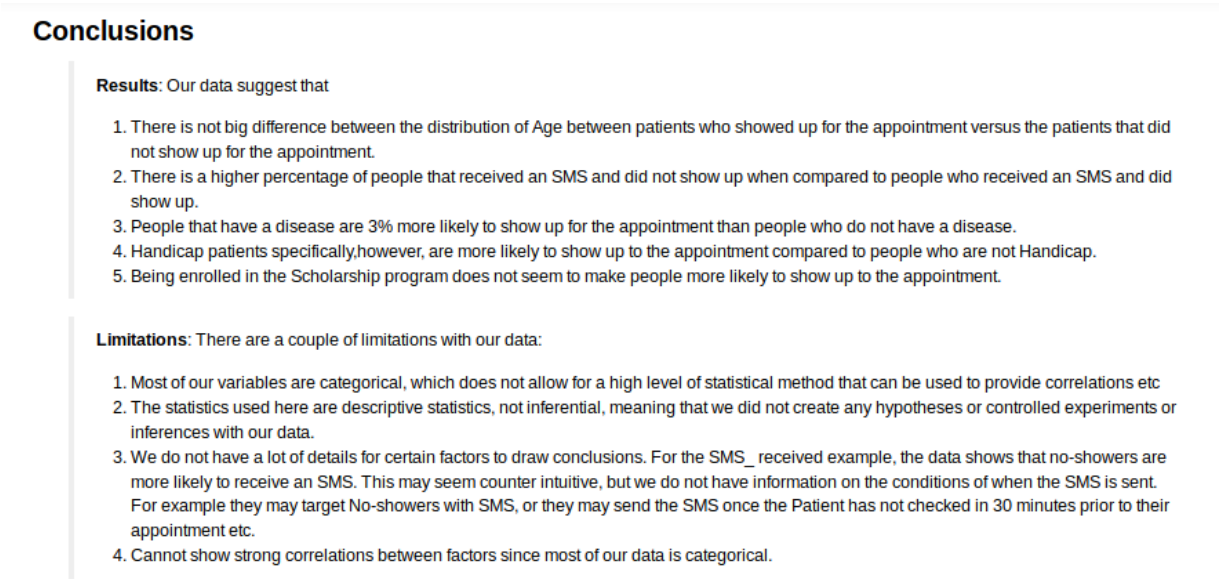
The results of the analysis are presented such that any limitations are clear. The analysis does not state or imply that one change causes another based solely on a correlation.

Conclusion section is not included in your analysis. You need to include conclusion section in the end. The conclusion is intended to help the reader understand why your EDA should matter to them after they have finished reading the analysis. A conclusion is not merely a summary of the main topics covered or a re-statement of your research problem but a synthesis of key points .

There should be a separate subsection inside the conclusion section called 'Limitations' where you would have to discuss the limitations of this dataset which might have adversely affected your analysis. Examples would be null or missing values, whether this samples are an effective representation of the population or

not or maybe that you could dive deeper in your analysis with additional specific information.

The conclusions and limitations have the following structure:



Conclusions

Results: Our data suggest that

1. There is not big difference between the distribution of Age between patients who showed up for the appointment versus the patients that did not show up for the appointment.
2. There is a higher percentage of people that received an SMS and did not show up when compared to people who received an SMS and did show up.
3. People that have a disease are 3% more likely to show up for the appointment than people who do not have a disease.
4. Handicap patients specifically,however, are more likely to show up to the appointment compared to people who are not Handicap.
5. Being enrolled in the Scholarship program does not seem to make people more likely to show up to the appointment.

Limitations: There are a couple of limitations with our data:

1. Most of our variables are categorical, which does not allow for a high level of statistical method that can be used to provide correlations etc
2. The statistics used here are descriptive statistics, not inferential, meaning that we did not create any hypotheses or controlled experiments or inferences with our data.
3. We do not have a lot of details for certain factors to draw conclusions. For the SMS_ received example, the data shows that no-showers are more likely to receive an SMS. This may seem counter intuitive, but we do not have information on the conditions of when the SMS is sent. For example they may target No-showers with SMS, or they may send the SMS once the Patient has not checked in 30 minutes prior to their appointment etc.
4. Cannot show strong correlations between factors since most of our data is categorical.

Communication

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

Fantastic 👏

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

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

Rate this review