

[Return to Classroom](#)

# Investigate a Dataset

## REVIEW

## HISTORY

### Meets Specifications


Greetings Student,

I applaud you for having adopted all the guidance on this project. This was a good execution. It was a pleasure to evaluate your work, because it was carefully thought through. I appreciated you've designed the report structure.

I encourage you to try various different types of plots. [Here](#), you can find lots of different types of visualizations with code for each visualization.

You should also learn Why use of [Functions to Avoid Code Repetition](#) is important and keep up the good work as it will make you a great Data Analyst. Way to go!

You can inquire on the [Knowledge Portal](#) for any queries. At this Pportal mentors are open 24/7 to resolve your concerns.

Stay ! Stay Safe!

### Code Functionality



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

All codes are **functional and error free**.

---

## Tips

- As you have used Jupyter notebook, this [shortcuts with Jupyter Notebook](#) will make your learning experience smooth.
- You should also try [how to use Python to write programs that do in minutes what would take you hours to do by hand-no prior programming experience required](#).



- 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.

Excellent effort to simplify the work using the Pandas libraries.

Built-in methods in Pandas which are extremely beneficial to explore variables in this project:

- [Boolean-Indexing](#)
- [Group-by](#)
- [Value-Counts](#)
- [Series.map](#)
- [Working-with-text-data:](#)

I want to suggest how you can optimise Pandas code for performance based on your code and **How to Vectorize Data Aggregation.**:

- [A Beginner's Guide to Optimizing Pandas Code for Speed](#)
- [How to Vectorize Data Aggregation](#)

Also, [Python Pandas: Tricks & Features You May Not Know](#) and [10 Python Pandas tricks that make your work more efficient](#)



- The code makes use of at least 1 function to avoid repetitive code.
- The code contains good comments and meaningful variable names, making it easy to read.

Excellent effort to prevent repeat code and to include good comments and variable names over the whole project.

---

## Suggestions

## Suggestions

- You can include [docstring](#) for user-defined functions.
- Check out this link for more info on [DRY principals](#)
- [10 Tips for Writing Cleaner & Better Code](#)
- [Use functions to avoid code repetition](#)
- [Importance of Commenting Python Code](#)

## 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.

Good job mentioning **data wrangling/preparation steps** for the target variable especially.

Do checkout: [Data Wrangling with pandas, Cheat Sheet](#). This cheat sheet is on data exploration operation in Python using Pandas is your go-to resource to know each step involved in data exploration. You will find cheat codes for reading & writing data, a preview of data frames, rename columns of data frame, aggregate the data, etc.

## Exploration Phase



- The project investigates the stated question(s) from multiple angles.
- The project explores at least three variables in relation to the primary question. This can be an exploratory relationship between three variables of interest, or looking at how two independent variables relate to a single dependent variable of interest.
- The project performs both single-variable (1d) and multiple-variable (2d) explorations.

The analysis makes use of **both single and multiple variable explorations** to investigate different features and the relations between these features in the dataset.

- You can try various different types of plots. [Here](#), for each visualisation you may discover several

different types of visualisations and code.

## Tips

- [A complete tutorial on data exploration](#)
- [What is Exploratory Data Analysis?](#)
- [Exploratory Data Analysis \(EDA\) and Data Visualization with Python](#)
- [Quick and Dirty Data Analysis with Pandas](#)

## Summary: Differences between univariate and bivariate data.

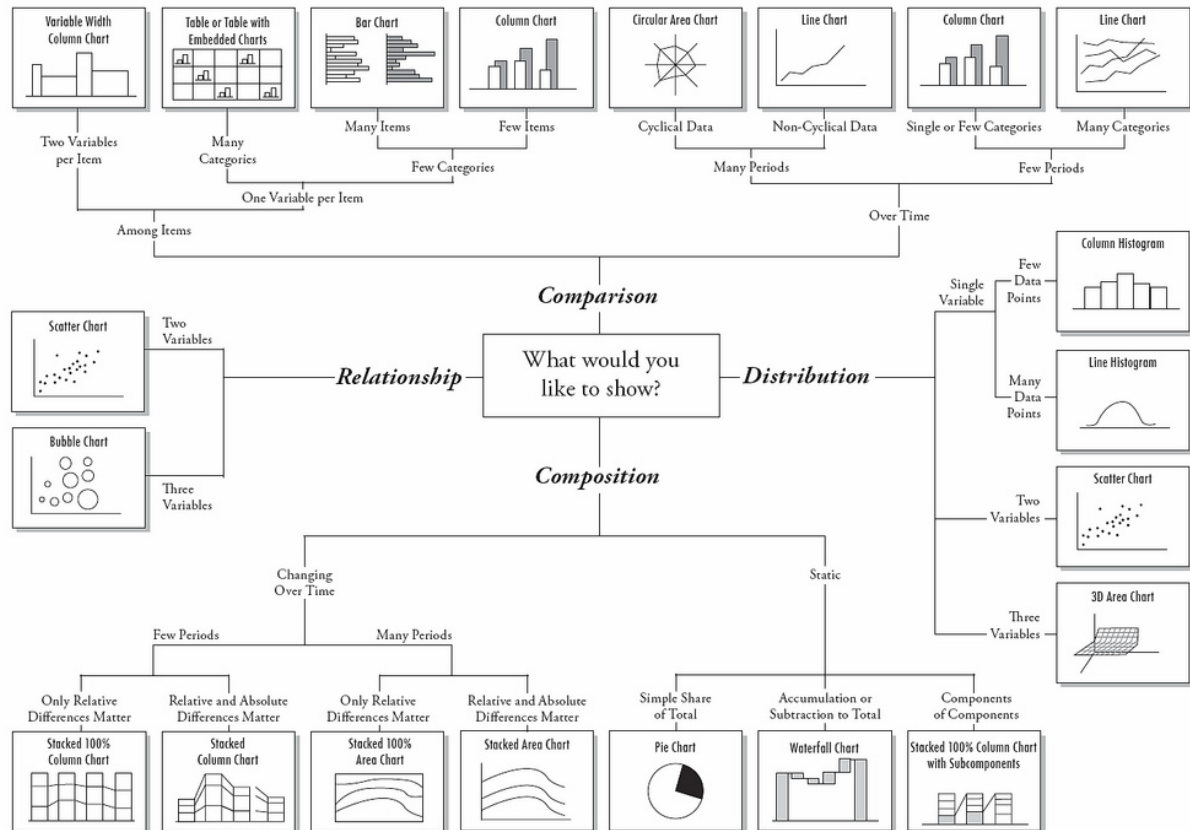
Univariate Data	Bivariate Data
<ul style="list-style-type: none"> <li>• involving a <b>single variable</b></li> </ul>	<ul style="list-style-type: none"> <li>• involving <b>two variables</b></li> </ul>
<ul style="list-style-type: none"> <li>• does not deal with causes or relationships</li> </ul>	<ul style="list-style-type: none"> <li>• deals with causes or relationships</li> </ul>
<ul style="list-style-type: none"> <li>• the major purpose of univariate analysis is to describe</li> </ul>	<ul style="list-style-type: none"> <li>• the major purpose of bivariate analysis is to explain</li> </ul>
<ul style="list-style-type: none"> <li>• central tendency - mean, mode, median</li> <li>• dispersion - range, variance, max, min, quartiles, standard deviation.</li> <li>• frequency distributions</li> <li>• bar graph, histogram, pie chart, line graph, box-and-whisker plot</li> </ul>	<ul style="list-style-type: none"> <li>• analysis of two variables simultaneously</li> <li>• correlations</li> <li>• comparisons, relationships, causes, explanations</li> <li>• tables where one variable is contingent on the values of the other variable.</li> <li>• independent and dependent variables</li> </ul>
<b>Sample question:</b> How many of the students in the freshman class are female?	<b>Sample question:</b> Is there a relationship between the number of females in Computer Programming and their scores in Mathematics?



- The project's visualizations are varied and show multiple comparisons and trends.
- At least two kinds of plots should be created as part of the explorations.
- Relevant statistics are computed throughout the analysis when an inference is made about the data.

The report uses **different charts** to look at the analysis and show the insights and outcomes. 🙌

## Chart Suggestions—A Thought-Starter



© 2006 A. Abela — a.abelag@gmail.com

## Suggestions

Do checkout more ways of [Plotting with categorical data](#)

## Conclusions Phase



- The Conclusions have reflected on the steps taken during the data exploration.
- The Conclusions have summarized the main findings in relation to the question(s) provided at the beginning of the analysis accurately.
- The project has pointed out where additional research can be done or where additional information could be useful.
- The conclusion should have at least 1 limitation explained clearly.
- The analysis does not state or imply that one change causes another based solely on a correlation.

Excellent work displaying the analytical findings and **limitations**.

## Tips

- [What are the limitations of a study and how to write them?](#)

## Communication



- The code should have ideally the following sections: Introduction; Questions; Data Wrangling; Exploratory Data Analysis; Conclusions, Limitation.
- Reasoning is provided for each analysis decision, plot, and statistical summary.
- Interpretation of plots and application of statistical tests should be correct and without error.
- Comments are used within the code cells.
- Documented the flow of analysis in the mark-down cells.

It is vital to effectively convey the findings as a future data analyst; nevertheless, description of every action, analysis or graph is also quite significant. This will enable your viewers to see what you do and how you do it. In addition, your work **becomes structured, formal and sophisticated** via reasoning you added.



Visualizations made in the project depict the data in an appropriate manner (i.e., has appropriate labels, scale, legends, and plot type) that allows plots to be readily interpreted.

Awesome! The graphs are easy to read and nicely labelled.

In the end, I would like you to ask yourself some questions:

1. Can you determine what is being plotted?
2. Is there another plot type that would be more appropriate for this data?
3. Is the scale chosen appropriately so that the data is visible?

I feel this is going to have a big effect on your analysis and lead you to a better data analyst.

 [DOWNLOAD PROJECT](#)

[RETURN TO PATH](#)

Rate this review

START