

[← Back to Data Analyst Nanodegree](#)

# Investigate a Dataset

## REVIEW

## HISTORY

### Meets Specifications

Greetings Student,

This was a good implementation and I congratulate you for passing all rubric items with this submission. It was delightful reviewing your work as it was well thought-out. I encourage you to keep up the good work as it will make you a great Data Analyst. Way to go! 😊

### Code Functionality



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

The code ran perfectly and did not produce any errors when run. Good.



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 project has used NumPy, Pandas, and DataFrames wherever appropriate. The good thing about python is that there are many built-in functions which we can directly use for the analysis.  
You have done a great job in this part.



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

Your code is properly commented and contain good variable names which is making your code easy to read.

## Quality of Analysis



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

Good work.

The project has clearly stated the questions and then they are addressed in the rest of the analysis. The questions are brief and precise.

## 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.**

The stated questions were studied and investigated from various angles. You have included multiple variables for 1d and 2d explorations. Great work.



**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.**

Visualizing data requires a lot of patience and determination because it's not easy selecting the best visualization to match with a given data type. Well enough, the project rightly builds descriptive visualizations using a variety of plots.

## 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 results are clearly presented and analysis does not state that one change causes another based solely on a correlation.

## Communication



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

You have provided the reasoning after each plot, making it easy to understand. Good work.

Your analysis is written in explanatory terms allowing your audience to fully understand the work done and results. 🙌. In this direction, I like [this post](#) very much, it explains the importance of telling a history using Data.

As a suggestion, this project is a great opportunity for you to create a new repository in Github that becomes part of your online portfolio and allow potential employers to review your work, in case you are not familiar with Github, this is a [great post](#) and an [Udacity Course](#) for deeper understanding. This report defines your credentials, so it is important that you put special attention not just to the technical side of the project but also the communications side since this is a critical characteristic for any data scientist. For your reference, check this [Kaggle post](#) for further reference, as you can see this is really a hot topic in the data science world! 😊



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

The visualizations are clear and easily interpreted. The plots contain labels, titles, and legends wherever appropriate.

Excellent work. 👍

Whenever we make visualizations, we must ask ourselves, "Can the reader understand what the plot is conveying without looking at code or comments?", visualizations included are properly formatted and help to better understand the different data exploration conclusions. 🙌