Data Science Career Track
Capstone Two: Exploratory Data Analysis

Overview

Now that you've obtained, cleaned, and wrangled your dataset into a form that's ready for analysis, it's time to perform exploratory data analysis (EDA). Use the outline below as a reminder of what steps to follow. Keep in mind that the goal of the EDA work is to get familiar with the features in your dataset, investigate the relationships between features, and generally understand the core characteristics of your dataset. Be creative and think about interesting figures and plots you can create to help deepen your understanding of the data.

You can also review the EDA work you did for the guided capstone, the DSM Medium article. And the EDA cheatsheet pdf for reference.

Project Steps

Estimated Time: 6-10 Hours

All of the following steps should be completed in a Jupyter notebook. Please provide adequate notation and structure so that your mentor can better understand the work you've done.

Goal: Explore the data relationships of all your features and understand how the features compare to the response variable.

- Build data profiles and tables
    - Inferential statistics
- Explore data relationships
    - Data visualization
- Feature Selection and Engineering

Inferential Statistics Hint: do any particular results in your data strike you as abnormal? *What hypotheses can you form (Null and Alternative hypotheses) which you could go on to test? Take some time to recall your statistical concepts, such as that the* p-value *of an observation is the probability of seeing data at least as extreme as that observation, on the assumption of the null hypothesis.*

Data Visualization Hint: recall your matplotlib *and* seaborn *functions for data visualization:* matplotlib:

- plt.plot()
- plt.xlabel()
- plt.show()
- plt.hist(),

Seaborn:

- sns.relplot()
- sns.lmplot()
- sns.catplot().

Remember to always start with an idea of what you want to achieve*, and use these libraries and their functions as your toolkit to make that idea a reality.*

Feature Selection and Engineering Hint: feature selection is where data storytelling starts: we tell a story as soon as we include certain features and omit others. But how we manipulate - or engineer - *those fields is just as important. Recall the crucial elements to feature engineering:*

- *If you have categorical features, you might need to* one-hot encode *them*
- *You may need to* binarize *your columns and* bin *your values.*
- *To handle missing data, think about how appropriate the methods of* listwise deletion, data imputation, replacing missing values with constants *or simply attempting to* find the missing values *are for your data.*

- *Think about whether you need to* standardize, log-transform *or* normalize *your data, as well as statistically valid ways to remove outliers.*

Consider the following questions and use your understanding of your dataset to answer them:

- Are there variables that are particularly significant in terms of explaining the answer to your project question?
- Are there significant differences between subgroups in your data that may be relevant to your project aim?
- Are there strong correlations between pairs of independent variables or between an independent and a dependent variable?
- What are the most appropriate tests to use to analyze these relationships?

Student Examples

Get some inspiration from these student examples:

Example 1: School Shooter Investigation - Tyler Schmalz

Example 2: Sports team performance - Rob Chudzik