VIJAY PULAVARTHI

FBI CRIME DATA ANALYSIS

Here is a relatively short explanation of the work completed thus far:

- Data cleaning: Removed columns with high missing values, handled duplicates, and checked for data types.
- Exploratory data analysis: Examined the distribution of categorical and numeric variables, identified the top 10 percentages of unsolved cases, and visualized the number of reported cases by state and year.
- Graphing relationships: Created a bar chart showing cases reported by state, a line chart showing aggregated reported cases by year, a line plot showing the number of solved cases each year, and a Catplot showing the distribution of case counts for each region.
- Rankings: Identified the top 30 states with the overall highest occurrences of human trafficking for all of the past 8 years.

Overall, the data analysis revealed a number of key findings, including:

- The average actual count of human trafficking offenses has increased over the years.
- The number of solved cases each year has generally decreased.
- There is significant regional variation in case volumes.
- The top 30 states with the highest occurrences of human trafficking are relatively consistent over time.

Here are some issues I ran into while analyzing the FBI human trafficking dataset:

- Missing values: The PUB_AGENCY_UNIT and UNFOUNDED_COUNT columns had a high number of missing values. I mitigated this by removing these columns from the dataset, as they were not essential for my analysis.
- Duplicate rows: There were a small number of duplicate rows in the dataset. I mitigated this by using the drop_duplicates() method to remove the duplicate rows.
- Data types: Some of the data types were not consistent throughout the dataset. For example, the DATA_YEAR column was initially stored as a string, but I needed it to be a numeric value in order to perform certain calculations. I mitigated this by using the pd.to_numeric() function to convert the DATA_YEAR column to a numeric value.

In the future, I plan to mitigate these issues by:

- Handling missing values more carefully: If I encounter a dataset with missing values in important variables, I will explore different options for handling the missing values, such as imputing them with reasonable values or removing the rows with missing values.
- Using a data validation pipeline: I will develop a data validation pipeline to systematically check for duplicate rows, inconsistent data types, and other data quality issues. This will help me to identify and address these issues early on in the data analysis process.
- Documenting my data cleaning steps: I will carefully document the steps I take to clean the data, so that I can reproduce the data cleaning process in the future and so that others can understand how the data was prepared.

To finish my project, I still need to:

- Analyze the data in more depth: I have explored the data at a high level, but I need to delve deeper into the data to identify specific trends and patterns. For example, I could look at the relationship between different variables, such as the type of trafficking, the age and gender of victims, and the location of trafficking incidents.
- Develop data visualization tools: I have created some basic data visualizations, but I would like to develop more sophisticated data visualization tools that can help to communicate the key findings of my analysis to a wider audience. For example, I could create interactive dashboards or maps that allow users to explore the data and filter it by different criteria.