Auto Insurance Expenditures Data:

The data for Auto Insurance Expenditures was obtained from the Insurance Information Institute and is available under the Public Domain. The dataset can be found here, under the 2006-2010 table.

This data was available in an Excel file; however, the file was formatted such that it could not be converted automatically into a CSV. However, aside from this, the table was mostly in a usable form, there were only a few changes that needed to be made.

These changes were as follows:

1. Deleted the leading and trailing rows of the spreadsheet, including the headers, as these were specially formatted and were incompatible with CSV format. Elsewhere, the headers were recorded separately to add back later.
2. Copied the remaining data into a new spreadsheet, and applied the default format to the entire sheet, to eliminate any other artifacts of the spreadsheets special formatting.
3. Using the recorded headers from step 1, added a new row with the headers for the data, properly formatted this time.
4. Deleted data not in the year 2010
5. The data contained a column for rank, however this was left blank for the United States row. To avoid problems arising from a null entry, the United States was given a rank of 0, ensuring that it would display the national data first if sorting by rank.
6. Changed the state name for Washington D.C. from “D.C.” to “District of Columbia”, to ensure that it was consistent with other data sources, for easy joining.
7. Finally, there were a few places where the data contained markers to reference the footnotes of the table. These markers were removed
8. Converted the resulting spreadsheet into a CSV file.

Insurance Information Institute. “Facts + Statistics: Auto Insurance | III.” *Iii.org*, 2016, www.iii.org/fact-statistic/facts-statistics-auto-insurance.

Urban and Rural Classification Data:

The data for Urban and Rural Classification was obtained from the United States Census Bureau and is available under the Public Domain. The dataset can be found here, under the table Percent Urban and Rural in 2010 by State.

This data was available in Excel spreadsheet format, and was already very clean and well formatted. There were only a few changes that needed to be made.

These changes were as follows:

1. Removed columns not relevant to the analysis. For this analysis, the focus was on the population and area percentages for the broad urban and rural classifications, as well as the population density data for each category. All other columns were deleted.
2. The only missing data that needed to be handled was for the District of Columbia: 100% of Washington D.C. is classified as an urban area, and so the population density for rural areas was missing. The missing data was replaced with a 0.
3. Converted the resulting spreadsheet into a CSV file.

US Census Bureau. “2010 Census Urban and Rural Classification and Urban Area Criteria.” *The United States Census Bureau*, 2 Dec. 2019, www.census.gov/programs-surveys/geography/guidance/geo-areas/urban-rural/2010-urban-rural.html.

Population, Age and Gender Data:

The data for Population, Age and Gender was obtained from the United States Census Bureau and is available under the Public Domain. The datasets can be found here, under Annual Estimates of the Resident Population for Selected Age Groups by Sex.

This data was available in spreadsheet format, but was only available in the form of individual spreadsheets by state contained in separate files. In order to have a single master table with data from all the states, a new table needed to be created from scratch.

The steps to do this were as follows:

1. Created a new spreadsheet, and added column headers for my intended columns. The data variables that were analyzed were state names, population over 16, the breakdown of that population by gender, population for various age groups, and median age. Each of these variables were separated in different columns
2. Individually downloaded the data for each state, and copied the data from that spreadsheet into the appropriate column of the new spreadsheet.
3. Once this was complete, removed all formatting to ensure consistency.
4. Converted the resulting spreadsheet into a CSV file.

Bureau, US Census. “State Population by Characteristics: 2010-2019.” *The United States Census Bureau*, www.census.gov/data/tables/time-series/demo/popest/2010s-state-detail.html.

Median Household Income Data

The data for Median Household Income was obtained from the United States Census Bureau and is available under the Public Domain. The dataset can be found here, under table H-8.

This data was available in an Excel file; however, the file was formatted such that it could not be converted automatically into a CSV.

To fix this, the following steps were created:

1. Manually created a new table, and copied over the data from the original Excel file. The final table had columns for State Name, Median Income, and Standard Deviation for the year 2010.
2. Converted the resulting spreadsheet into a CSV file.

US Census Bureau. “Historical Income Tables: Households.” *Census.gov*, 28 Aug. 2018, www.census.gov/data/tables/time-series/demo/income-poverty/historical-income-households.html.

Vehicle Type Data

The data for the breakdown of types of vehicles that were involved in fatal crashes were found from the National Highway Traffic Safety Administration data portal. The data set can be found here under the year 2010 for the entire United States under the tables: Vehicles Involved in Fatal Crashes by Vehicle Type - State: USA, Year: 2010 and Vehicles Involved in Fatal Crashes by Body Type - State: USA, Year: 2010

To prepare the data, the following steps were completed:

1. Downloaded the files as 2 separate Excel files (one with the overview of each vehicle class and one with detailed breakdown of each type of vehicle in each class) and then converted it to a csvs.
2. The csvs were imported using Pandas with the appropriate separator and line terminator parameters and stored as a dataframes
3. Empty rows and columns were removed
4. The extra "\n"s were deleted from the beginning of the first columns

National Highway Traffic Safety Administration. (2010c). FARS Encyclopedia: Vehicles - All Vehicles [Dataset]. https://www-fars.nhtsa.dot.gov/Vehicles/VehiclesAllVehicles.aspx

Fatal Accident Rates for Each Age Group

The data for the breakdown of the ages of motor occupants was found in the National Safety Council data portal. The data set can be found here under the year 2018 for the entire United States under the section: Age of Driver. The year 2018 was chosen since it was the year closest to 2010 for which there was available data. This data included the Rate of Fatal Accidents per 100,000 licensed drivers, which was used in this study

To prepare the data, the following steps were completed:

1. Downloaded the Excel file.
2. Created a new Excel file with Age Group Bins that were decided previously in our group
3. Calculated the Average Rate of Fatal Accidents Per 100,000 licensed by averaging the dates in the dataset for each of the predetermined age groups
4. The csv was imported using Pandas and stored as a dataframe.

Age of Driver. (2018, February 4). [Dataset]. National Safety Council.
https://injuryfacts.nsc.org/motor-vehicle/overview/age-of-driver/

Traffic Fatalities by State Data

The data for the breakdown of fatal crashes for each state was found in the National Highway Traffic Safety Administration data portal. The data set can be found here under the year 2010 for the entire United States under the table: 2010 Traffic Fatalities by STATE and Percent Change from 2009 - State: USA

To prepare the data, the following steps were completed:

1. Downloaded the Excel file and then converted it to a csv.
2. The csv was imported using Pandas with the appropriate separator and line terminator parameters and stored as a dataframes
3. Empty rows and columns were removed
4. The extra “\n”s were deleted from the beginning of the first columns

National Highway Traffic Safety Administration. (2010b). FARS Encyclopedia: States - Crashes and All Victims [Dataset]. https://www-fars.nhtsa.dot.gov/States/StatesCrashesAndAllVictims.aspx

Weather Conditions of Fatal Crashes Data

The data for the breakdown of weather conditions of fatal crashes can be found in the National Highway Traffic Safety Administration data portal. The data set can be found here under the year 2010 for the entire United States under the table: Fatal Crashes by Weather Condition and Light Condition - State: USA, Year: 2010

To prepare the data, the following steps were created:

1. Downloaded the Excel file and then converted it to a csv.
2. The csv was imported using Pandas with the appropriate separator and line terminator parameters and stored as a dataframes
3. Empty rows and columns were removed
4. This also had to be rearranged because pandas imported some of the data as index labels
5. The extra “\n”s were deleted from the beginning of the first columns

National Highway Traffic Safety Administration. (2010). FARS Encyclopedia: Crashes - Time [Dataset]. https://www-fars.nhtsa.dot.gov/Crashes/CrashesTime.aspx

Rural vs. Urban Fatal Crash Data

The data for the breakdown of motor vehicle deaths per 100 million miles traveled was found on the Insurance Institute for Highway Safety Data Institute. The data set can be found here under the table: Motor Vehicle crash deaths per 100 million miles traveled by land use, 1977-2019.

To fix this, the following steps were completed

1. Saved the data on an Excel file and saved it as a csv
2. The csv was imported using Pandas with the appropriate separator and line terminator parameters and stored as a dataframes .
3. Empty rows and columns were removed

Fatality Facts 2019: Urban/rural comparison. (1977–2019). [Dataset]. Insurance Institute of Highway Safety. https://www.iihs.org/topics/fatality-statistics/detail/urban-rural-comparison#:%7E:text=In%202019%2C%20the%20rate%20of%20crash%20deaths%20per,percent%20in%20urban%20areas%20%28from%202.35%20to%200.86%29.

Insurance Claims Data

This dataset comes from the Emcien Patterns Knowledge database and is a sample machine learning dataset. It can be found here. The dataset includes locational information, as well as policy information and claim amounts. Due to this being a sample machine learning dataset, the data was already quite clean, and prepared for machine-learning for the most part. Some step we took however included the following:

1. Removed unnecessary columns, such as Customer ID, as well as country due to all the entries being within the United States.
2. Ensured categorical data worked with dummy variables in Python.
3. Cleaned blanks and nulls.
4. Used Correlation Matrix to confirm most significant relationships between columns.

Sample Data Sets: EmcienPatterns. Leading Prescriptive Analytics. (n.d.). https://emcien.com/sample-data-sets-2/.

Average Annual Miles Traveled by Each Vehicle Category

The data for the average miles traveled by each vehicle category was found on the US Department of Energy data portal. The data set can be found here for the year 2018 for the entire United States under the table: Average Annual Vehicle Miles Traveled by Major Vehicle Category.

To fix this, the following steps were created:

1. Downloaded the Excel file.
2. Created a new Excel file with the Vehicle Categories listed in Vehicle Type Accident Data (Passenger Vehicles, Light Trucks, Large Trucks, Motorcycles, Bus, and Other/Unknown)
3. Aggregated the categories provided in this set to fit the previously determined categories
4. The csv was imported using Pandas and stored as a dataframe.

Alternative Fuels Data Center: Maps and Data - Average Annual Vehicle Miles Traveled by Major Vehicle Category. (2018). [Dataset]. US Department of Energy. https://afdc.energy.gov/data/10309

Aggregated Master State Data

Once all of the data that was state based was obtained and cleaned, we wanted to be able to combine all of that data into a single table, so that the various columns could be examined together for correlation matrices and visualizations.

To create this master table, the following steps were completed:

1. Created a new Excel spreadsheet.
2. Loaded all of the individual categories of the state-based data that we had, using the cleaned CSV files as the data sources.
3. Once each of these files were loaded as separate queries, a new query was created in Power Query. Within this query, each of the datasets were merged, using state name as the key.
4. Once the datasets had been merged into a single query, each column was expanded out, except for the state, so that the resulting table contained all of the columns from each of the datasets, without duplicate columns.
5. With these steps, a master_state_data file was created and stored as a CSV file. However, I wanted to clean out columns that would not be used in the final analysis, as well as add a couple of new calculated columns.
6. First, Power Query was used to remove the unneeded columns. This included the accident data from 2009, the percent change in accidents, the liability, collision and comprehensive insurance totals, the standard deviation of the median income, and the age breakdowns for population, as this had proven to not be fruitful in our analysis.
7. Calculated the accident rate column by dividing the number of accidents column by the population over 16 column and multiplied the result by 1,000 to obtain the accident rate.
8. To handle the gender breakdown, a new column was created containing the percent of the population that is male, by dividing the male population column by the total 16+ calculation column.
9. Once this new column was created, the total gender population columns were no longer needed, so the columns for male and female population counts were deleted.
10. Finally, loaded the query and saved the resulting file as a CSV.