**Introduction**

Over time the increase in gas prices and number of traffic cameras have become more prevalent in the city of Austin, Texas. In addition to these factors, we want to take a look at alcohol consumption and the role that it may play. Vehicle accidents continue to become an increasing issue as the city expands and grows, taking lives and causing havoc on transportation. We are looking for an opportunity to find factors that may have an effect on the frequency of vehicular accidents in order to help mitigate said accidents and improve the quality of road safety in Austin, Texas. One would assume that as gas prices increase, the amount of vehicles on the road would decrease. Alongside this, with the implementation of more traffic cameras, one would hope that they would act as a deterrent to running lights and make people drive with more caution. Alcohol has long been considered a problem when it comes to transportation accidents. We want to see if there has been an uptick in consumption amongst the population and determine the extent of the role that may play in the amount of accidents that occur.

**Data Domain Discussion**

* One of our team members, Bryan Loeffler, works as a Senior Legal Analyst based out of Austin, Texas. In his position, he collaborates closely with various law enforcement agencies providing valuable insights to facilitate part of the criminal legal investigation and exploration process. Due to his background and close involvement in law enforcement investigations and reporting, our team has chosen to use the dataset pertaining to "Traffic Report Data." This dataset in conjunction with the "Traffic Cameras Data" and "Gas Prices Data" will serve as crucial components in our mission to investigate and report on the "Effects of Risk and Mitigation Factors on Frequency of Transportation Accidents."
* The gas price dataset is from the U.S. Energy Information Administration and is a weekly depiction of the state of Texas’ average gas price. The primary data point of interest is the weekly pricing and we intend to map this to accidents to determine if there is a relationship as the expectation is fewer motorists causing fewer accidents as prices increase.
* The traffic camera data is provided by the city of Austin and is an overview of the traffic cameras that were installed in intersections throughout the city. The data of interest for our use is primarily the ‘turn on date’ as we are looking to see if there is a relationship between cameras that were turned on and subsequent motor vehicle accidents.
* The traffic incident data is provided by the city of Austin and is a list of traffic incidents that were reported primarily in Travis County. The primary data of interest is the type of accident as well as the location. We hope to utilize the location based data in order to determine any relation between camera installation and traffic reports.
* The alcohol data is provided by the city of Austin, it contains month and year data for time along with numerical data types for all sales, tax, and per capita columns - spirits, wine, beer, and ale. Our primary datapoint of interest was the sales columns of each type of alcohol, which we then plan to total by our timeframe interval to identify a correlation between accidents and the grouped sales.
* The transportation industry could veer in the direction of more preventative measures, such as cameras, based on quantifiable success of the measures that have been applied. This could result in simply an increase of cameras or that City Officials look into various other methods of safety monitoring to reduce incidents that occur.

**Analysis of Data**

* The traffic data contains a traffic report ID, publish date, issue report, location (latitude and longitudinal coordinates), address, status, and status date. The issue report data is of type object and categorical with the accident types. Location, longitude, and latitude are all numerical data types, status and address are objects. We have 373,755 observations of traffic incidents from 2018 to 2023\* (through September 2023). This dataset was provided by the City of Austin government database. Our team identified the status date, location coordinates, and incident report columns as valuable in our analysis - this allowed us to group the data by date and location with the traffic camera installation and the type of accident allowed us to differentiate reports that would or would not logically be impacted by camera installation.
* The weekly gas price data only contains two columns, a date field and a numerical data type for gas prices. There are 1218 observations from 2000 to September of 2023, all data in this set being relevant as we needed to utilize the weekly grouping with traffic and camera installation datasets. This data was sourced from the U.S. Energy Information Administration website.
* The camera installation data contains camera data like type, ID, enabled or turn on date, status, and location. There is also city data such as district, jurisdiction, location type, and location ID that were mostly unusable for this analysis. The dataset contains primarily objects and keys with regards to location, camera type, and addresses. Dates used for turn on date and numerical data for location coordinates. Though a lot of data is provided, the primary points of interest for our team included the turn on date, the location, the status, and the calculated count of cameras turned on by date. This allowed us to coordinate the linking by week with gas and traffic incident reports. There are 868 observations in this dataset and the data was also obtained through the City of Austin government database.
* The alcohol data set contains primarily numeric data including population, sales and tax data, as well as month and year data. Our main focus was to extract the sales data for grouping in Month/Year keys, the data already being broken out in these columns made the merge piece easier. Because a lot of the numeric data was unnecessary we were able to efficiently strip the necessary columns into a separate data frame making the necessary calculations easier.
* The only issue we encountered with consistency in the data came in the form of how the location coordinates were presented between traffic incidents and camera installation datasets. The formatting prevented usable data in our merge which downstream did not allow us to successfully align or group our data by precise location.

**Data Cleaning**

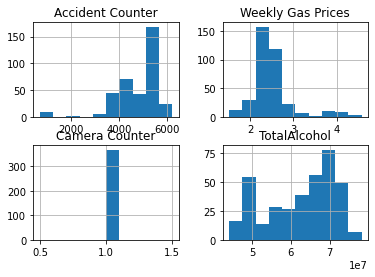
* The overall quality of our datasets could be classified as noisy but not overly complicated when it came to the variables of value. All datasets contained relevant date/time frames that allowed us to group and key off the date ranges and easily filter the categorical data once the ambiguous and irrelevant data was accounted for.
* The camera installation dataset contained the highest number of irrelevant columns or variables. The primary column we were concerned with, Turn On Date, had both missing values and unnecessary categories as we only needed a count of cameras that were turned on in a given timeframe. Additionally we needed to strip the date data into separate Month and Year columns to allow for time frame grouping which then provided the necessary keys to create a calculated column of camera enablement counts.
* The gas price dataset was the most straightforward of the two given there were only two columns, week and average gas price. Aside from removing irrelevant data at the top of the file the primary action was to strip special characters from the price column to allow for mathematical operations once the data was split into Month and Year columns.
* For traffic accident data, while also stripping the data into month and year columns for grouping our concern was first to remove any type of report that was unlikely to result in a traffic camera installation - for example, boating accidents. Once the irrelevant incidents were removed and the timeframes grouped that allowed to perform another count on accidents to tie into the other datasets.
* The alcohol data already being broken out into Month and Year columns by default removed a step in the process that was consistent for the other sets. With that we were able to strip the relevant columns, alcohol sales by type, and create a separate dataframe to remove noisy columns altogether instead of programmatically stripping them. From there totaling all sales to create a single calculated column grouped by data allowed us to format the dataset in a way that fit into the merge plan seamlessly.
* The biggest challenge we faced was finding a way to merge the time frames across the datasets. The gas data provided weekly averages, the camera installation, alcohol sales, and traffic reports were isolated by date but not necessarily at a consistent level. The month/year grouping that resolved the timeframe problem also removed a level of granularity that you might hope to see to create a more vast dataset. The second of our problems was the programmatic determination of addresses and zip codes given coordinates in the traffic incident and camera datasets. While we found a way to accurately obtain the zip code that we aspired to also group on, the limits the library allowed for presented problems created the relevant data in our output as we were unable to resolve the throttling issue for the large datasets ranging between 800 and 373k rows.

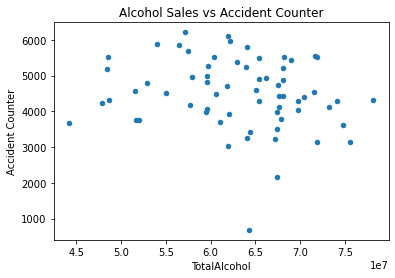
**Data Merging**

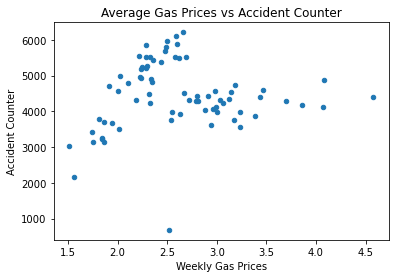
* Two separate merges were performed, one large merge that created a single dataset based on an inner join of the Month/Year key. The second merge performed was a merge of a set of two cleansed datasets for a more granular analysis, all also on the Month/Year key grouping. The issue of grouping by granularity was resolved by breaking the data out into Month/Year in the cleaning process. While this removed some of the precision that may be desired it offered a clean way for all four datasets to merge effectively.

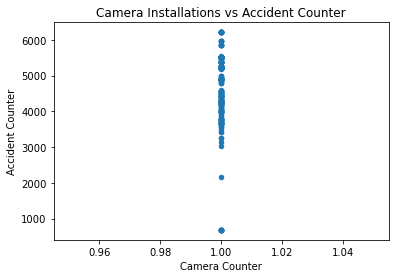
**Visualization of Data**

* Our hypothesis was that increased gas prices and decreased alcohol sales would have a correlation on the number of traffic incidents reported. After visualization and running correlation analysis, the relationships were not as strong as we anticipated them being. Evidenced below:
* When it comes to attempting to visualize our data, we ran into an issue with the sheer differences in values. The only two values on a slightly similar scale are the gas prices and the camera installations in the single/double digit range. For car accidents, the values jump into the thousands and for alcohol, that value skyrockets to the hundreds of thousands and millions. This difference in scale causes normal visualization to always have one data piece dwarfed when compared to accidents.
* We ran simple scatter plots to see if there is any correlation but everything was skewed to one side or the other. The only decent “visualization” obtained was doing a simple correlation, which turned up less than desirable results, and a histogram of the four major variables. The histogram showed that most of our data is fairly skewed either to the left or the right, with TotalAlcohol being the only semi-normal one.









**Additional Documentation**

* Upon analysis of our data, we have determined that none of it was worth attempting to carry forward to future classes. While we did start out with some csv files containing >500 rows, once all merging and totals were done, we ended up with ~54 rows. This was due in part to merging on month and year, so we are limited by the date ranges available for each dataset.
* In addition to the lack of rows disqualifying the dataset, only 4 out of our 9 possible variables had meaningful correlations (corr > .15 or corr < -.15).

**Flow diagram of your project**

* [Link to Flow Diagram](https://lucid.app/lucidchart/a98d963a-5944-4b3d-af50-6fed71339597/edit?invitationId=inv_50d1fdec-d3dd-4587-9663-79155d645b9b)
* Separate image provided as well

**Instructions for code**

* The code is broken up in separate chunks (separated by walls of #), each one dealing with an individual csv file and the final two dealing with merging them and the visualization. To ensure proper execution of code, please run each section at a time as you go down the code. Code files include comments with step by steps instructions on what each statement is doing prior to execution.

1. Run import statements to retrieve all the necessary libraries
2. Set the ‘path’ variable to the directory where your datasets reside as well as where you want any output to be written to
   1. If you do not change the name of the original source files then you do not need to modify the file\_x variables
   2. Modify the output files to the desired names
3. Run the statements to read in the CSV files into pandas dataframes
4. Data Cleaning:
   1. Drop any camera records where Turn on Date is missing
   2. Set list of statuses that are not applicable
   3. Loop through status column and remove any irrelevant statuses
   4. Parse datetime column to month and year for grouping
   5. Create calculated column ‘Camera Counter’ then run a sum based on the datetime grouping done in step d
   6. (Optional: Uncomment and run command to write to CSV to see output)
   7. Rename ambiguous gas column
   8. Parse datetime column to month and year for grouping
   9. Rename verbose column name for later visualization and easier viewing
   10. Replace the $ character in the price column to allow for mathematical operations
   11. Group prices by month/year and calculate the average
   12. (Optional: Uncomment and run command to write to CSV to see output)
   13. Extract relevant columns from the alcohol dataset and create a new dataframe for analysis
   14. Use the average of the ale column to account for 0 values as to not skew the results
   15. Create a total alcohol sales calculated column based on the already grouped month/year columns
   16. (Optional: Uncomment and run command to write to CSV to see output)
   17. Parse datetime column to month and year for grouping accident data
   18. Create list of accidents unlikely to cause camera installation
   19. Loop through undesired accident types for removal
   20. Create calculated column Accident Counter’ then run a sum based on the datetime grouping done in step o
   21. (Optional: Uncomment and run command to write to CSV to see output)
   22. (Optional: On a smaller dataset run the reverse geolocation step to extract the address and strip the zipcode for location grouping. As noted, this is throttled on larger datasets but code is included for review)
5. Data Merging
   1. All datasets were parsed to key on Month/Year
   2. First merge: merge all datasets into one master file on Month/Year with inner joins
   3. Second merge: Create separate files (ex. Accident and camera, accident and alcohol, etc) with inner joins based on Month/Year keys
   4. (Optional: Uncomment and run command to write to CSV to see output)
6. Data Visualization
   1. Run correlation function to observe calculated value
   2. Run correlation variable output to csv to observe
   3. Run large scatter matrix on all data
   4. Isolate desired variables to review and create multiple scatter plots for analysis
   5. Extract calculated columns and create histogram