Canadian Gas Station Purchasing Paradigm: Analysis and Insights

Group Members: Anatole Colevas - Lingfeng Pan - Runze Li - Shiyi Ji

Introduction: The following analysis investigates inventory practices for seven gas station locations in Hamilton, Canada. The stations provided us with invoices detailing their purchase history, location data, gas tank capacity/type data, and fuel level data tracking the inflow and outflow of fuel from each tank. Our datasets include information from 2017-2019.

After an initial investigation of the data, our team established three guiding questions for our final analysis and recommendations.

1.) How well did each station manage its fuel inventory as well as fuel delivery practices?
2.) What inventory management practices did certain locations use to gain a competitive advantage and reduce costs?
3.) Which locations failed to maximize profit due to poor inventory management and what improvements can they make in the future?

The following outline details all data cleaning and analysis tasks followed by recommendations for each station moving forward.

Data Cleaning:

After an initial exploratory analysis of the provided datasets, our team chose to make certain adjustments in an attempt to make the datasets more workable. Adjustments for each of the original datasets provided are detailed below.

Steps:

Fuel_1 / Fuel_2 Data Sets:
- Column headings were adjusted to match one another
- Fuel_1 and Fuel_2 were concatenated to create one cohesive data frame
- Replace ‘T12’ with ‘T 12’ so match other rows tank ID
- Drop rows with null values (2 out of 1.8 million)

Tank Data Cleaning:
- Check for duplicates and null values
- Drop the Premium gas tank data as there is only one in the dataset and therefore not enough to analyze variations between stations
- Convert all U (premium gas) to G (regular gas) as mentioned by Professor Nikandish

Merge Tank and Fuel Dataframes (Combined Dataframe = fuel_level)
- Convert ‘Time Stamp’ column to Datetime datatype

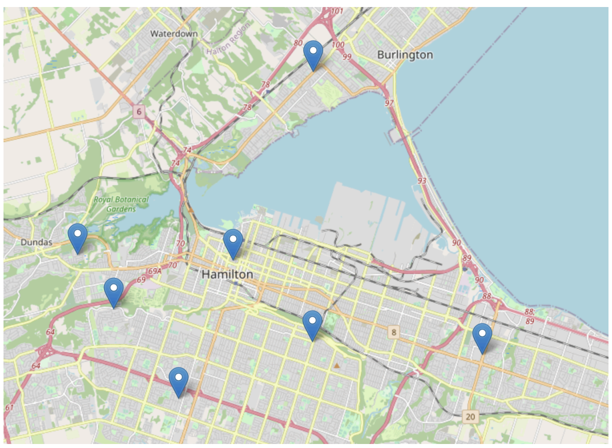
Location Data:
- Check for duplicates and null values

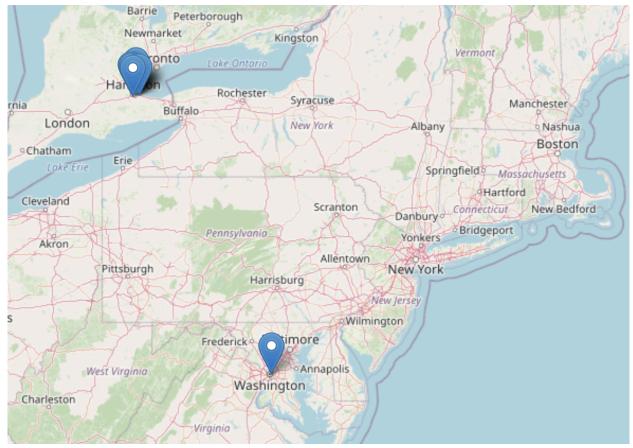
Invoice Data:
- Change “Invoice Date” column to datetime datatype
- Drop 42 rows each with 4 missing values
- Add dollar/liter column to the data frame that includes the station's purchase cost for fuel
- Check for duplicated rows

Exploratory Data Analysis:

Location:

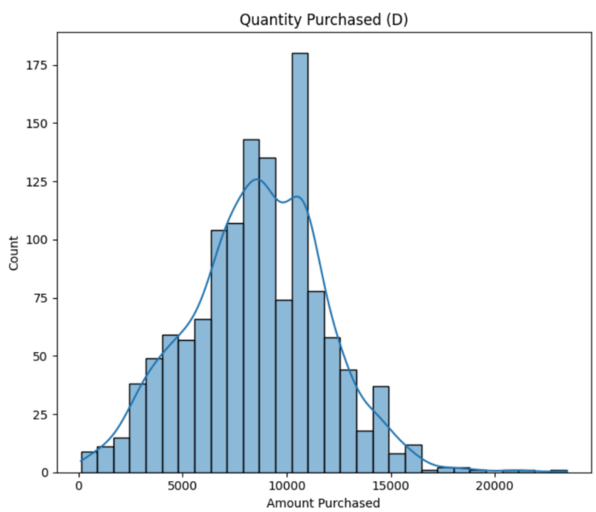
We began our analysis by mapping each of the gas stations based on their coordinates. Seven of the eight stations being evaluated are located in Hamilton, Ontario Canada while one is located in Washington D.C. We chose to drop the D.C. location from our analysis as Hamilton and D.C. are too far away from one another to compare inventory and pricing to one another. Each country also uses a different currency and different units of measurement (liters/gallons).

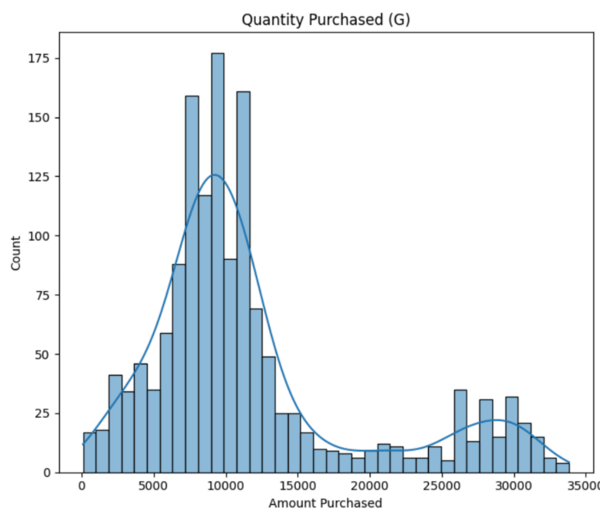




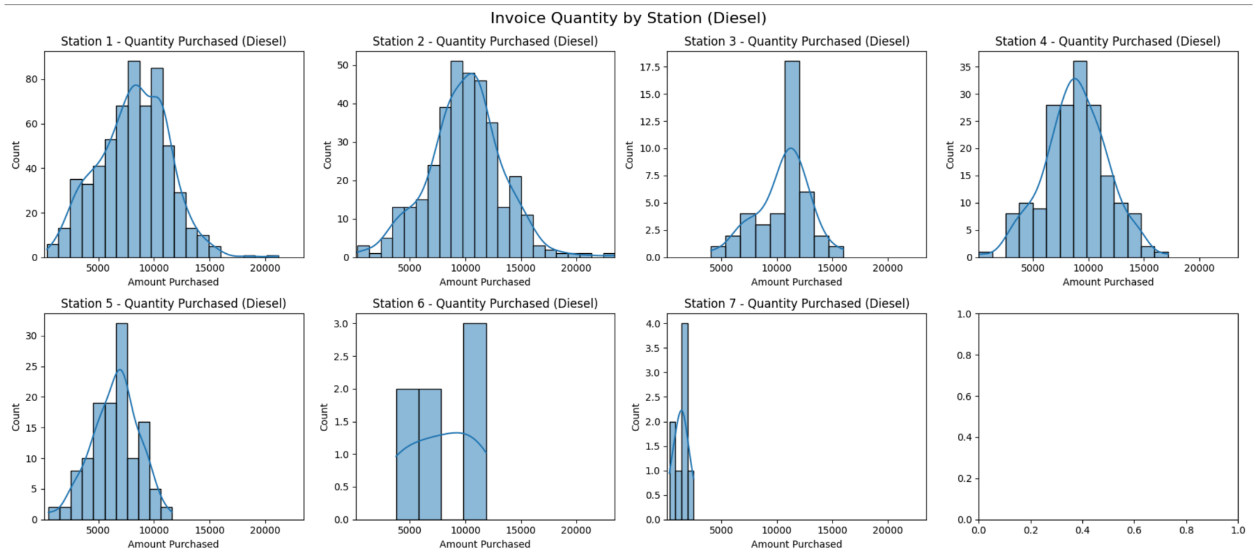
Diesel Analysis:

Before analyzing individual stations' data we split the invoice data into separate data frames with only diesel invoices or regular gas invoices (invoice_d and invoice_g respectively). This allowed us to individually assess fluctuations in the price and purchase amount of each fuel type independently and provided general information useful for later analysis.

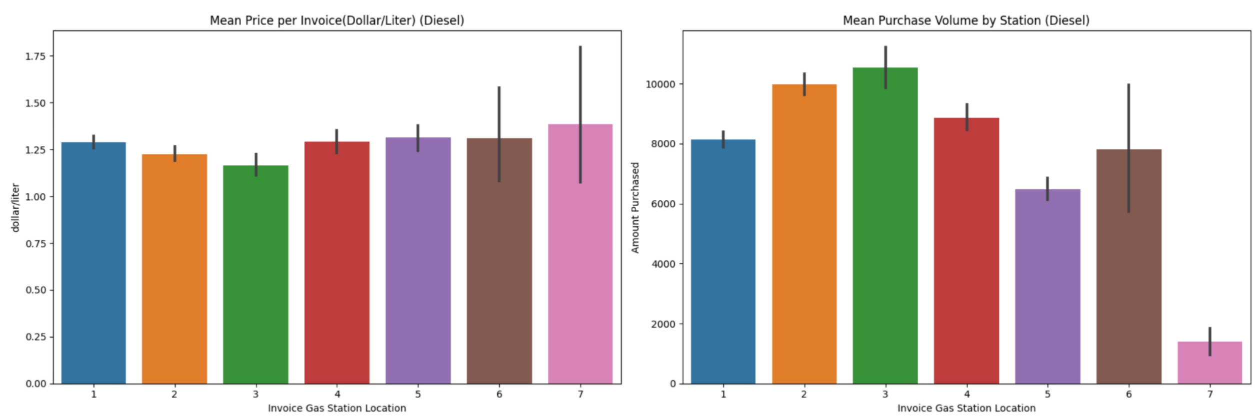




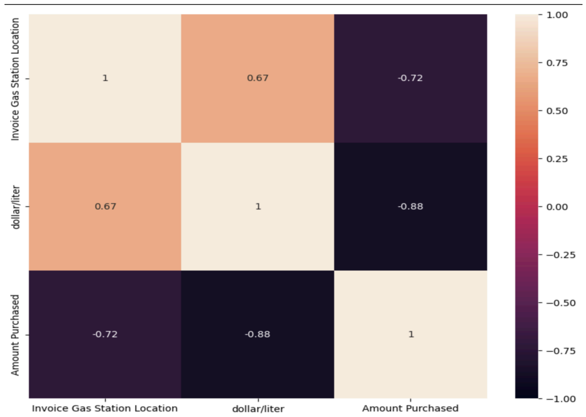
Here we can see the distribution of fuel purchases for both diesel and gas at all stations. Diesel fuel purchases have a close to normal distribution with most purchases falling between 7000-12000 liters. Most gas purchases also fall in a similar range but there is also a small spike in purchases over 25,000 liters.



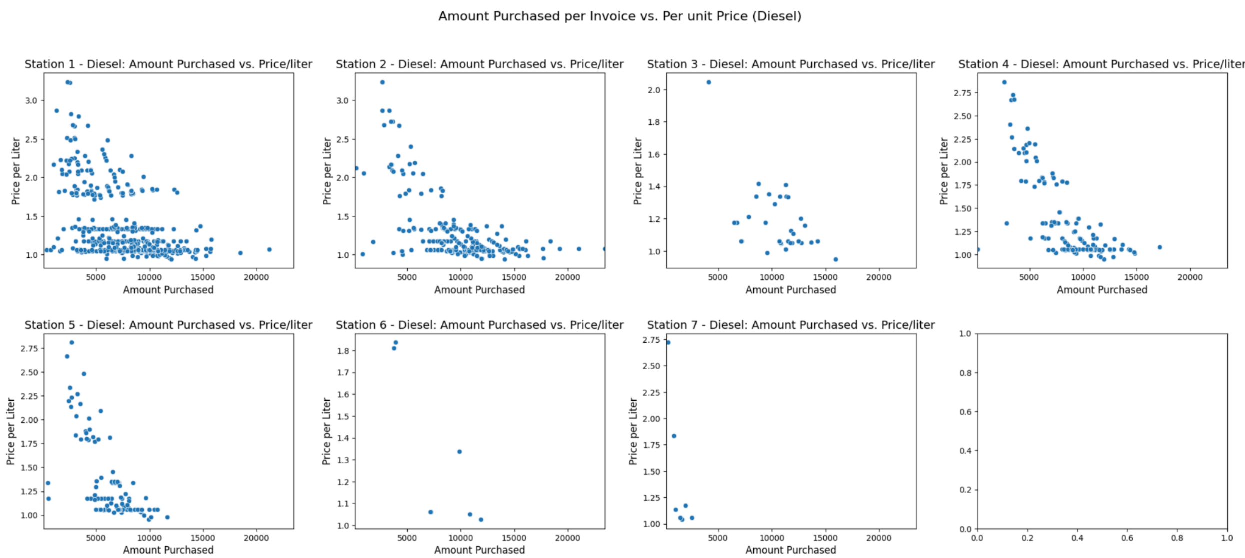
Here we can see the same visual but data is broken into separate charts for each of the seven stations analyzed. By looking at the y-axis ranges we can see that stations 1, 2, 4, and 5 are buying larger quantities of diesel fuel compared with other stations. We can also see slight variations in buying habits, with station 1 buying at a lower volume (<10,000 liters) more often than station 2, 3 whose typical buying volume is slightly higher. Lastly, we see that stations 6 and 7 both have very few diesel purchases in the dataset, indicating they are not making as many sales. Station 7 is particularly interesting as it has very few total fuel purchases and all purchases are low volume (>5000 liters). This will likely lead to higher average purchasing costs. Let's investigate further.



Above we have visuals looking at the mean price (dollar/liter) of diesel invoices for each gas station (left) and the mean purchase volume of diesel by station (right). We can see that station 3 has the lowest mean purchase price by volume and the highest mean purchase volume per invoice. Of all stations, stations 2 and 3 typically purchase diesel in higher volume when compared to other stations. There also appears to be a negative correlation between increasing fuel volume per purchase and decreasing dollar/liter prices. Let's investigate.

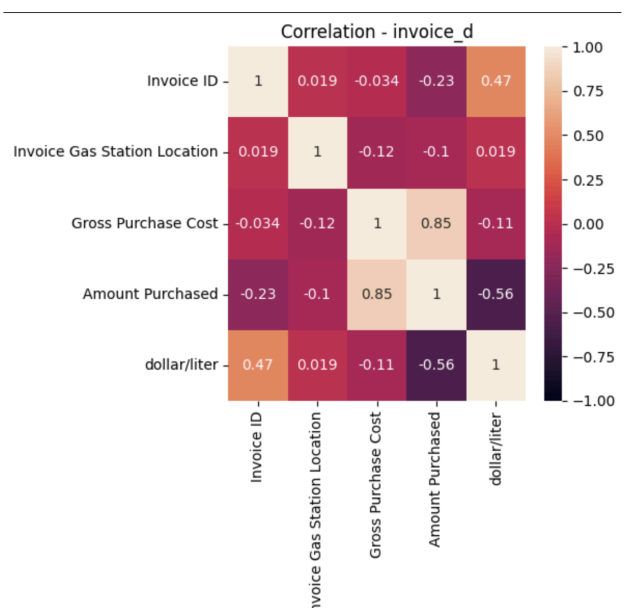


- As expected, there is a negative relationship between the mean dollar/liter price and the mean amount of fuel purchased. Stations should attempt to buy fuel in larger volumes to reduce their dollar/liter cost.

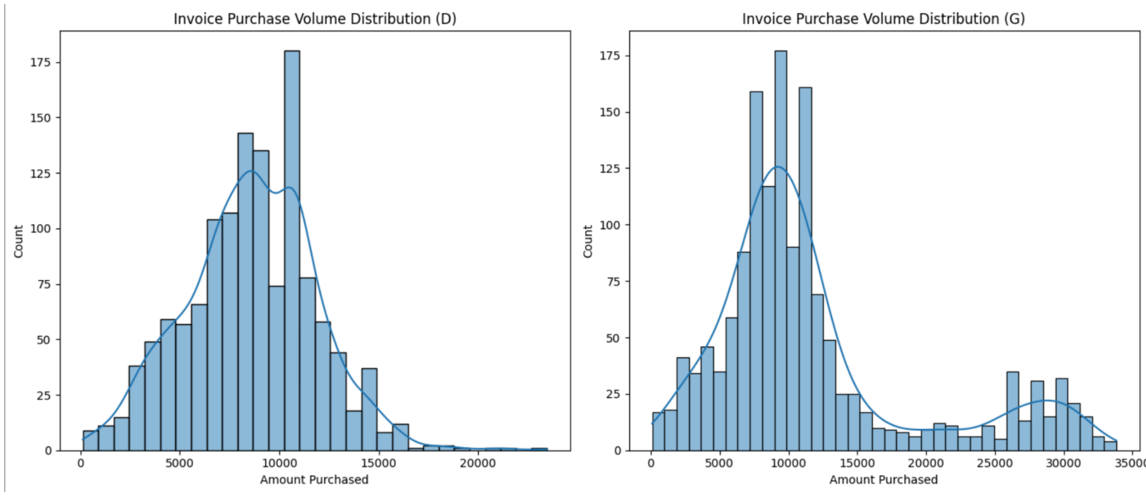


Above we are looking at scatterplots comparing the dollar/liter cost of each invoice to the volume of fuel (Diesel) purchased for each station individually. There seems to be a relationship between the increasing volume of purchases and the lower dollar/liter price of fuel (negative correlation) for the majority of stations. Some stations, such as 6 and 7 have so few purchases in the dataset that they are not particularly meaningful for analysis. Let's investigate the correlation.

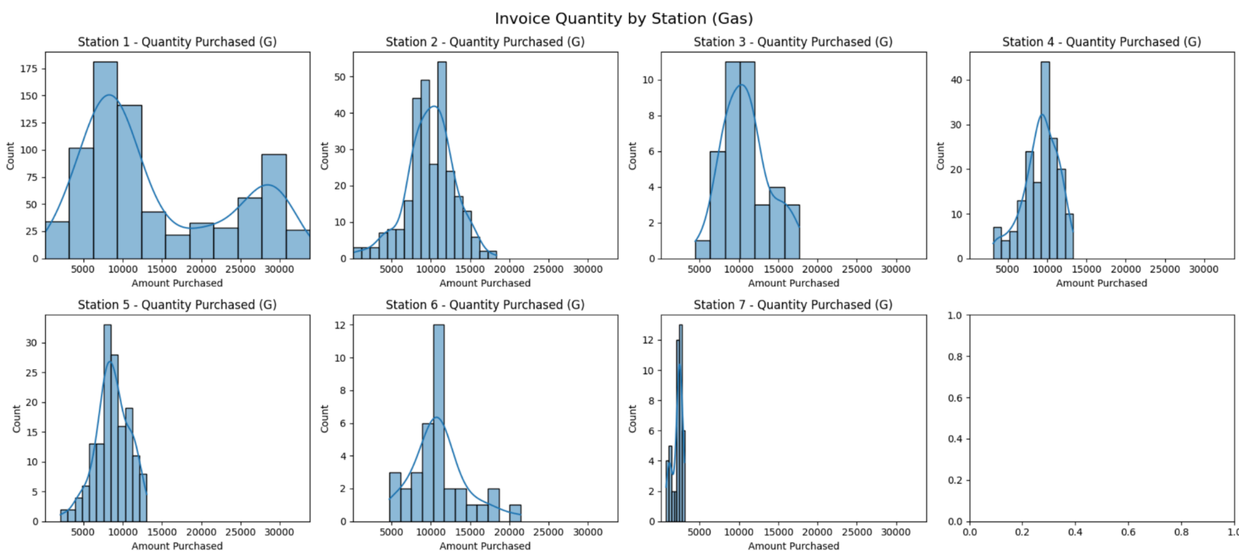
- We can see here that there is a -0.56 correlation between the amount of fuel purchased and dollar/liter (diesel).



Gas Analysis:



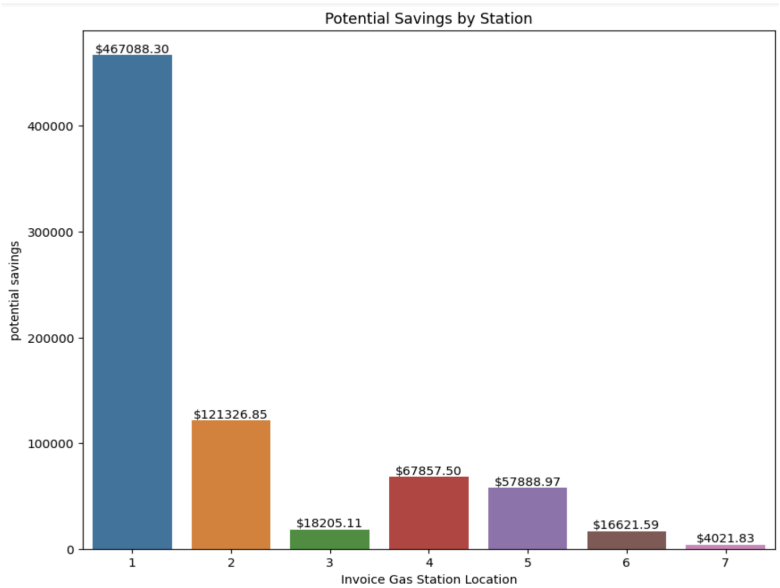
Here we are again looking at the side-by-side comparison of diesel and gas invoice purchase volumes. When compared to the approximately normal distribution of diesel purchases, there is an interesting spike in gas purchases >25,000 liters.



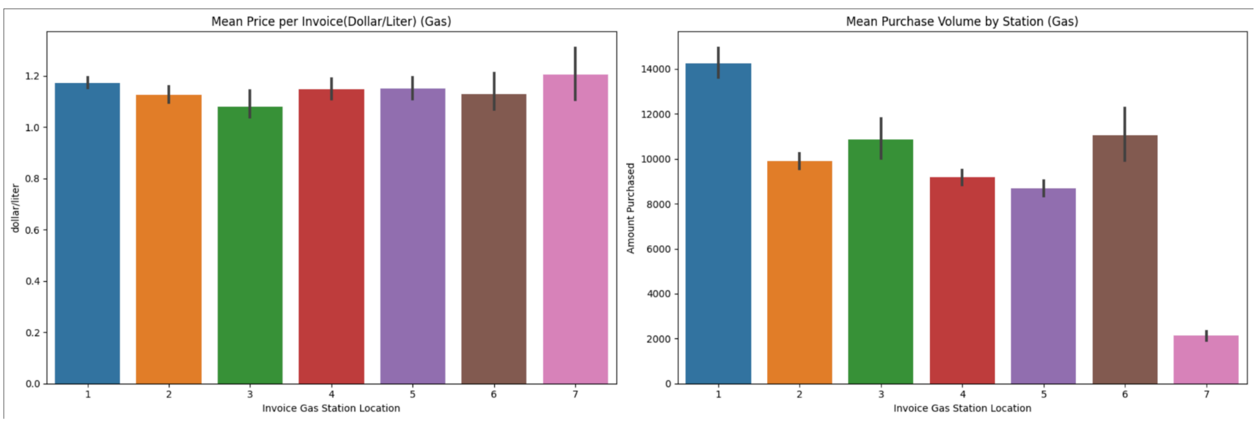
While stations 2-7 purchasing (Gas) follows a relatively normal distribution, station 1 has several purchases of 20,000 liters or more. Almost all of the purchases of gas above 20,000 liters were by station 1. We then subset those invoices with more than 20000 liters and less than 20000 liters to compare their relative costs.

```
#Look at dollar/liter price difference of buying 20000+ liters at a time versus the
df_1=invoice_g[invoice_g['Invoice Gas Station Location']==1]
df_1 = df_1[df_1['Amount Purchased']>= 20000]
df_2=invoice_g[invoice_g['dollar/liter']<20000]

low_price = df_1["dollar/liter"].mean()
high_price = df_2["dollar/liter"].mean()

save = high_price-low_price
print(f'Stations could save {round(save, 2)}$ per liter if they bought in incraments of greater than 20.000 liters')

Stations could save 0.04$ per liter if they bought in incraments of greater than 20.000 liters
```

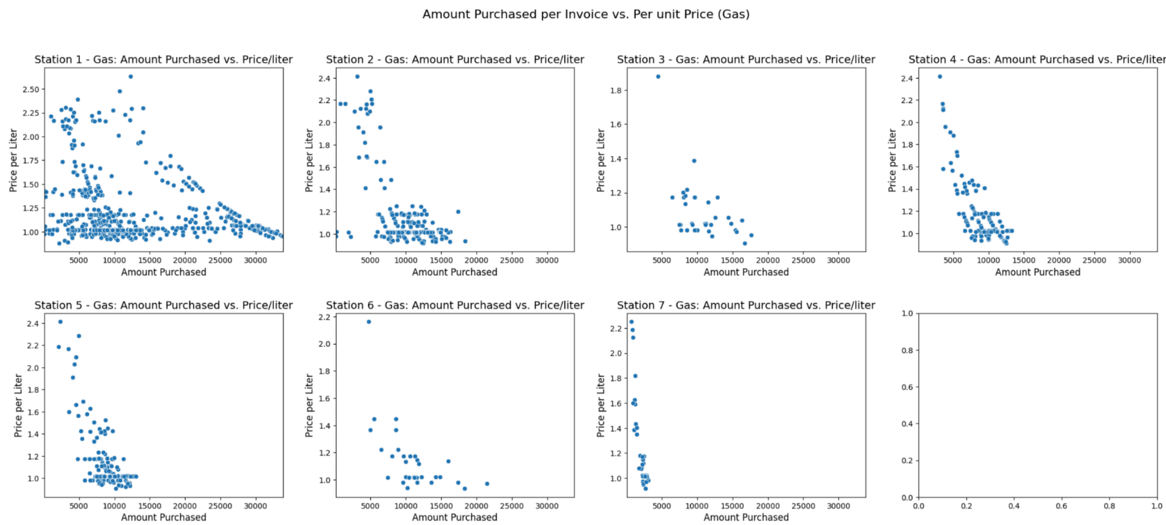
Based on our data, we can see that stations could save approximately 4 cents per liter if they buy gas in increments of for than 20,000 liters at a time. Let's look at the potential savings by station



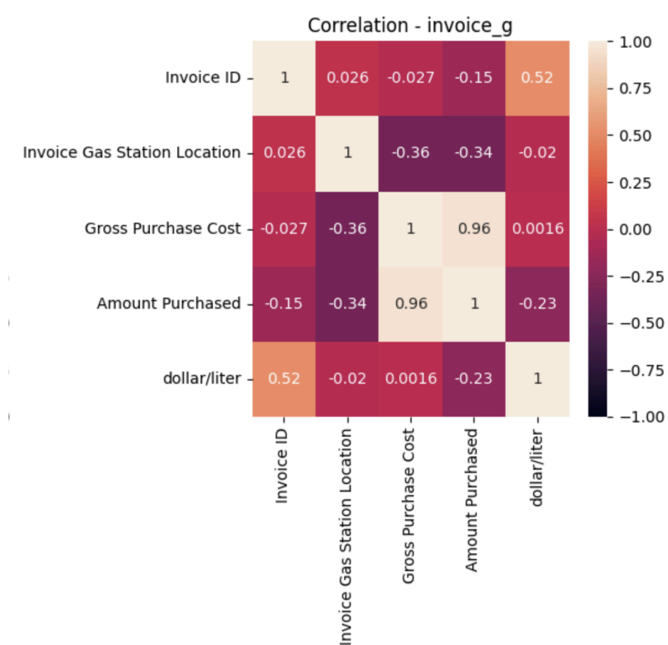
- We can see that gas stations stand to save significant amounts of money by buying large volumes (<20,000 liters) of fuel with each invoice.
- The large gap in potential savings between stations is largely a function of the volume of gas they buy and sell. Station one buys and sells much more fuel than other stations and therefore stands to save more.



While we can see that stations 2 and 3 had the lowest mean purchase price/liter, they do not have the highest mean purchase volumes. This does not reflect the same negative relationship apparent in diesel data.



Visual analysis seems to show a negative correlation between purchase volume and price per liter. Let's investigate further by calculating correlation coefficient.



- When investigating gas, the correlation between dollar/liter and amount purchased is not very strong(-0.23)
- This correlation could become stronger with a larger dataset as visual analysis does seem to indicate a negative correlation.

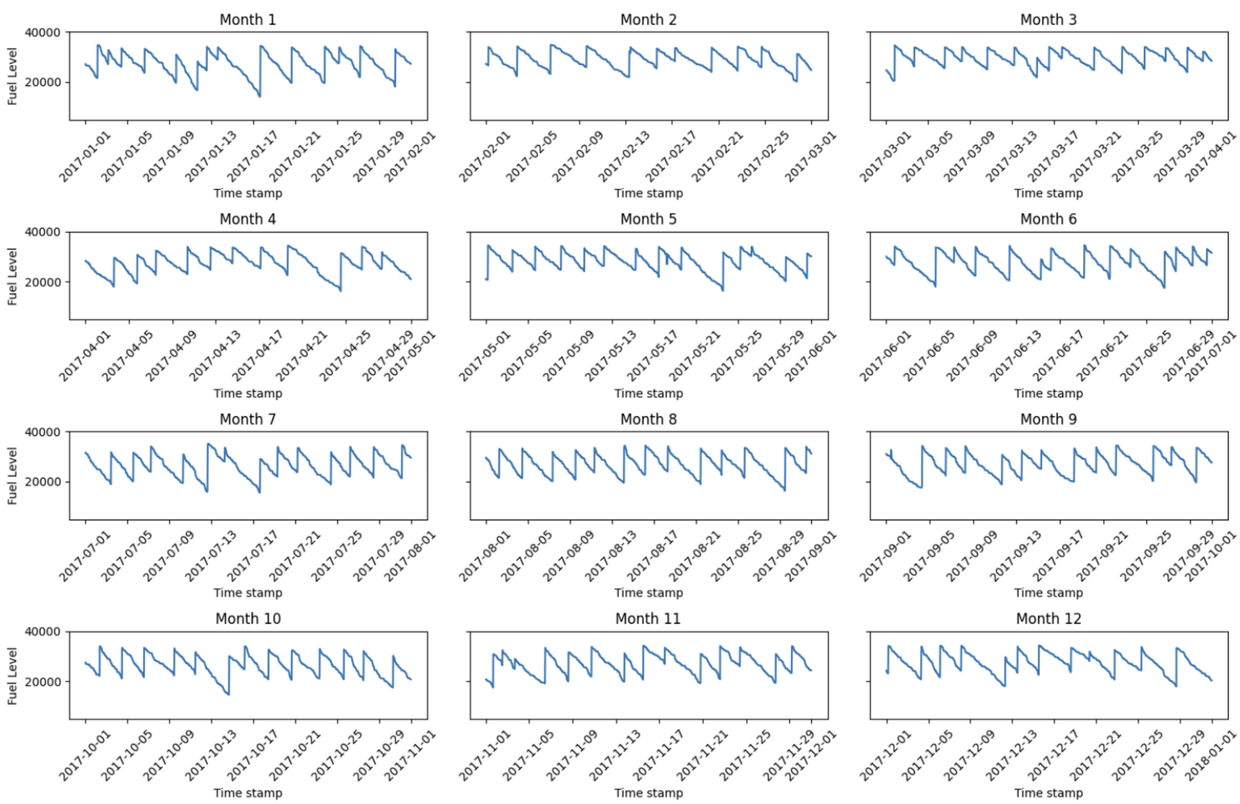
Fuel Level:

Diesel:

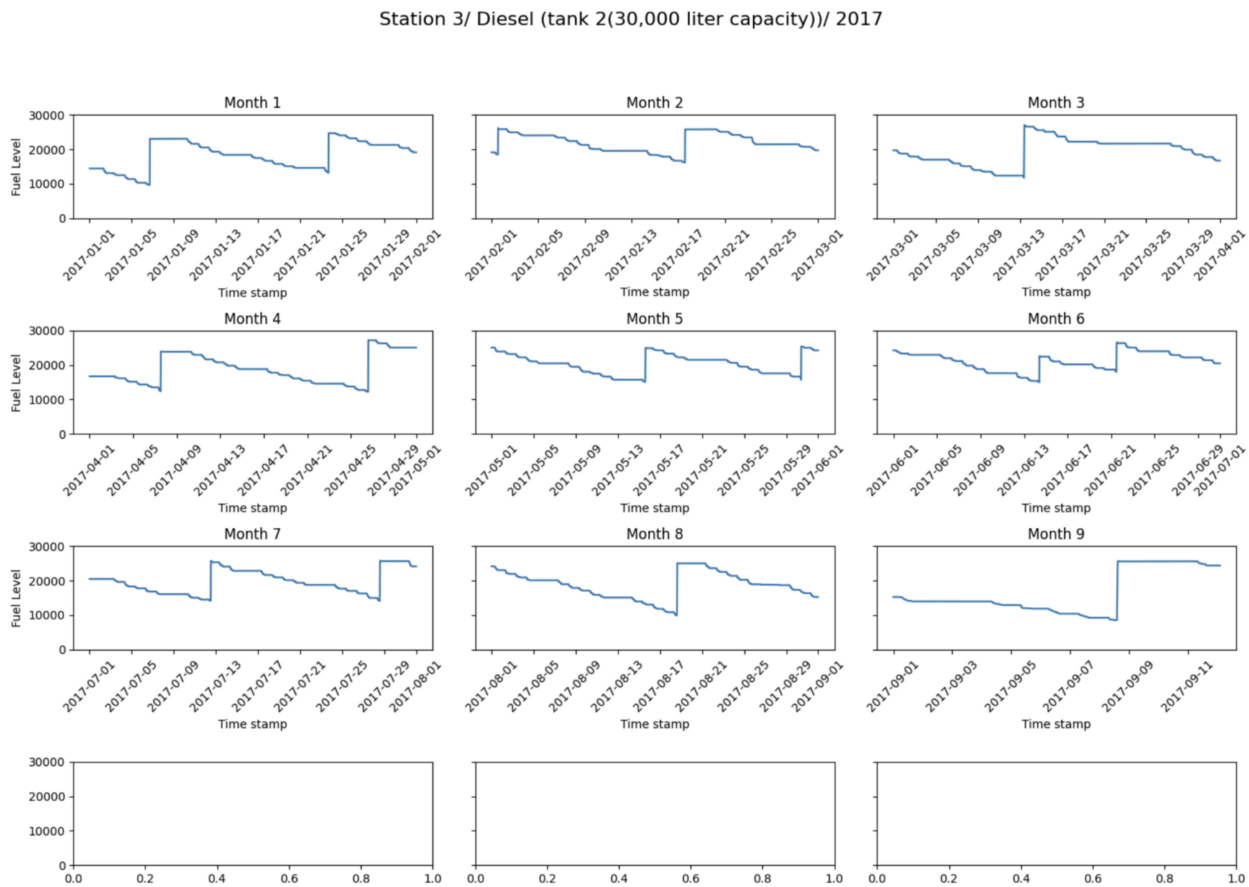
While our analysis above indicates that stations should attempt to buy both gas and diesel in greater bulk, other factors affect their ability to do so. Stations have limited tank space and can only purchase as much gas as they can store. We must investigate the fuel_level dataset, measuring fluctuations in fuel level between various stations, tanks, and fuel types. Changing fuel levels allows us to gain greater insight into fuel purchases, storage efficiency, and fuel sales.

Below we can see the fluctuating fuel level in station one's third fuel tank (diesel). This visual only looks at changes in fuel levels through the year 2017, month by month.

Station 1/ Diesel (tank 3 (40,000 liter capacity))/ 2017



One of the details immediately visible is the frequency with which the station refills its tanks. We can assume that each vertical spike in tank level corresponds to a delivery of fuel, and therefore, an invoice. We can also see that the total fuel level in the tank rarely drops below 20,000 liters throughout the year, meaning the station is refueling more frequently than necessary. While the station certainly wants to maintain a small amount of fuel in tanks at all times to avoid running out of stock, keeping 20,000 of supply on hand is not only expensive, it also prevents the stations from reducing cost with larger fuel purchase volumes. While refilling practices vary slightly by tank the general trend of frequent diesel refills with large remaining volumes is common among all stations (station specific charts included in python file) with the exception of station 3.

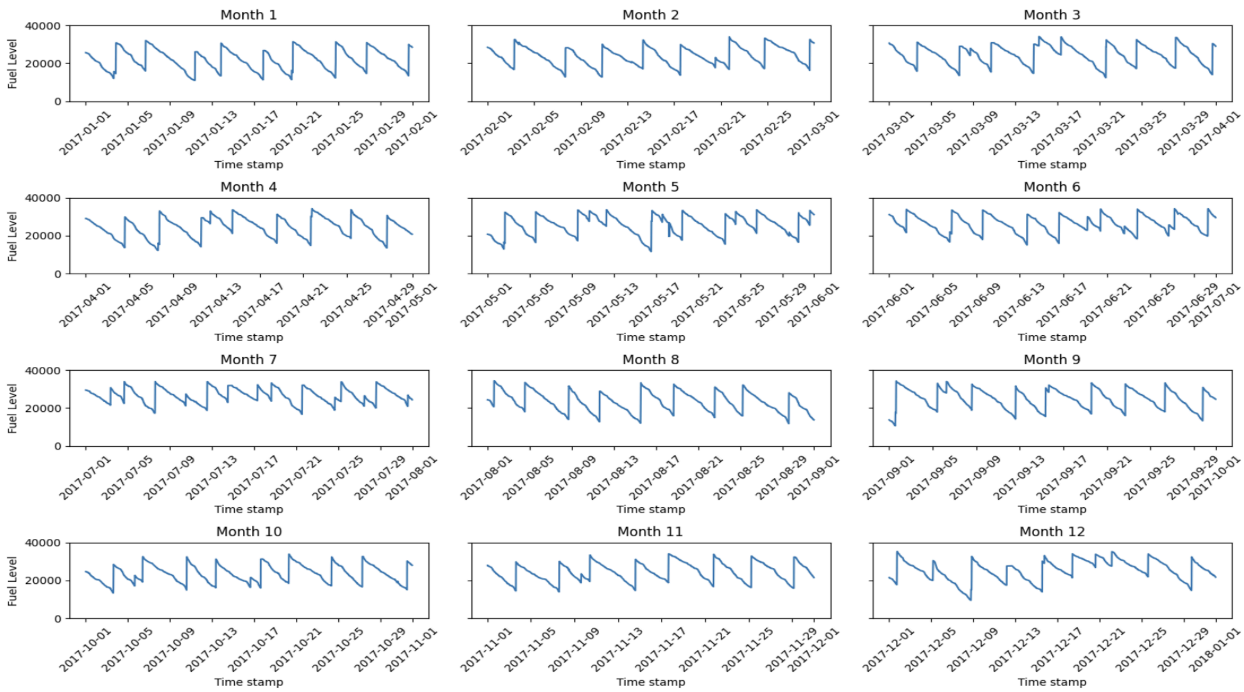


Station 3’s diesel fuel level chart indicates a higher average number of liters per refill (invoice) which is also reflected in the station 3 invoice analysis above. This is likely one of the reasons station 3 has the lowest average dollar/liter purchase cost for its diesel invoices. Taht beig said, station 3’s fuel level rarely drops below 10,000 liters and therefore the station still has the ability to reduce the frequency of fuel purchases, increase their average purchase volume, and increase savings. Charts for each of the 7 stations analyzed are included in the attached Python file.

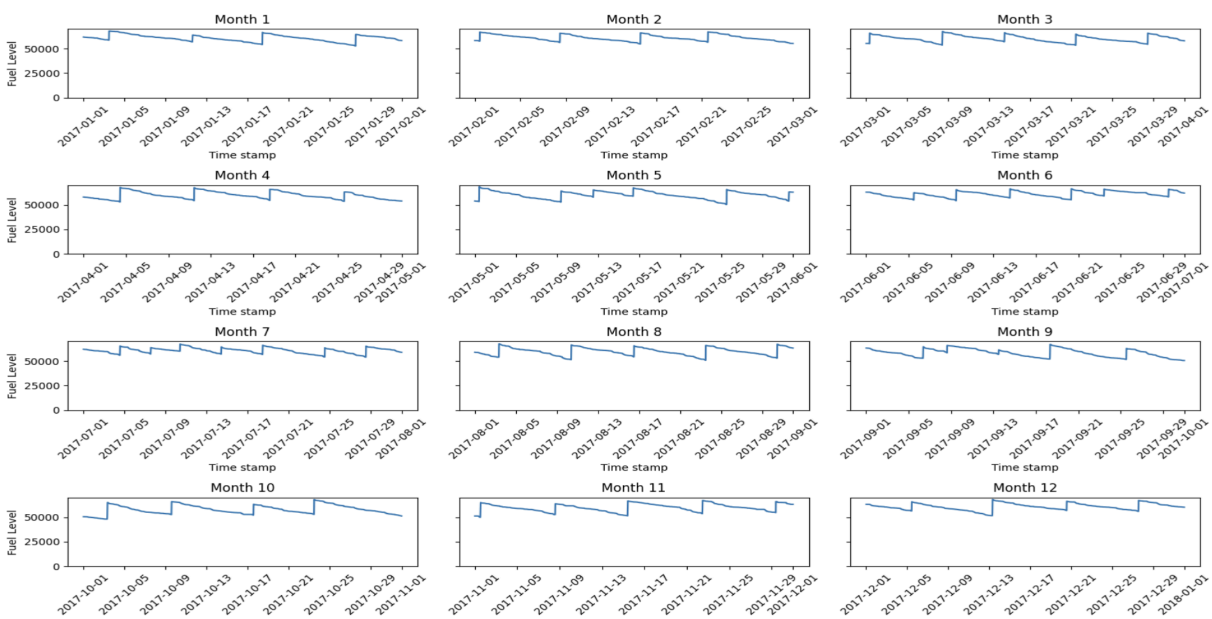
Gas:

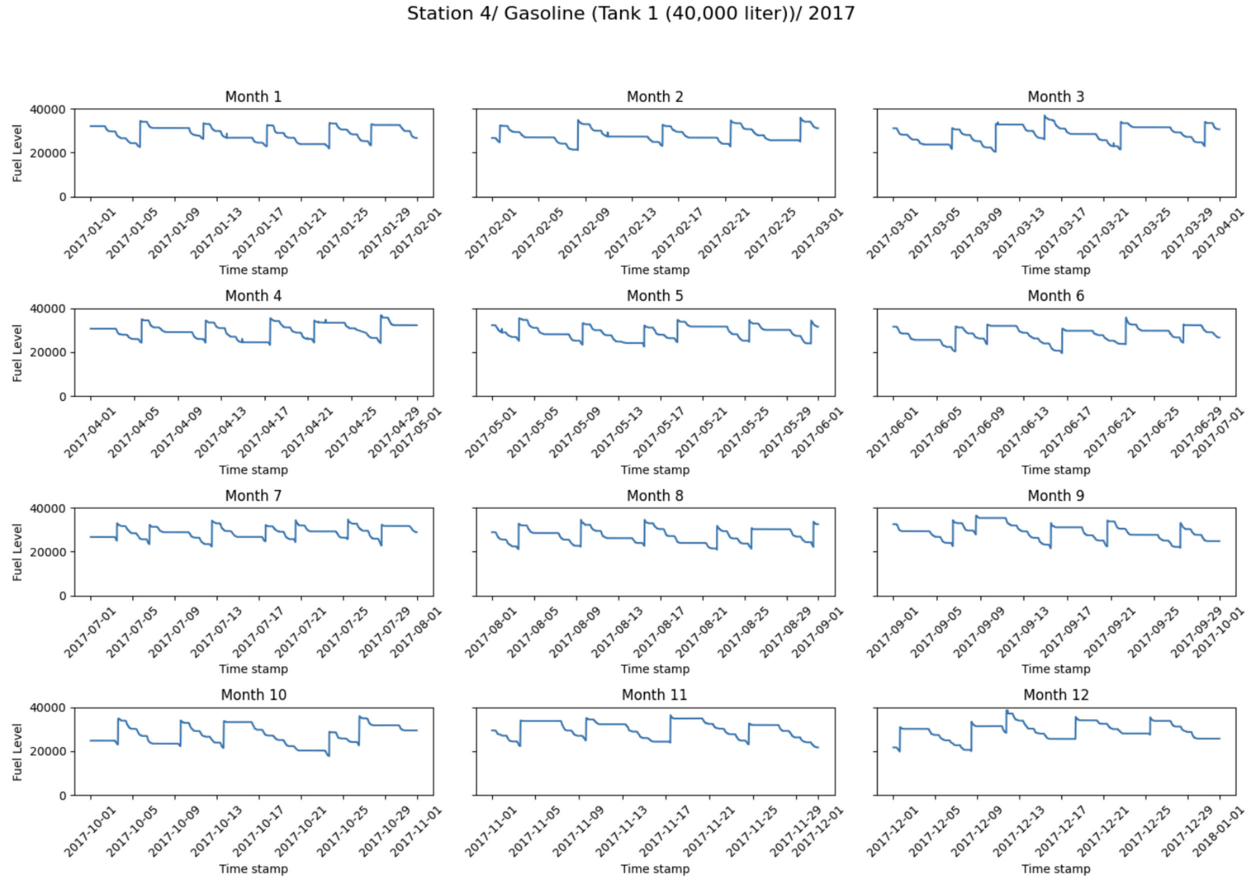
Below we look into gasoline fuel levels in a variety of stations for insights as to how stations can improve their inventory strategy.

Station 1/ Gasoline (Tank 2 (40,000 liter))/ 2017



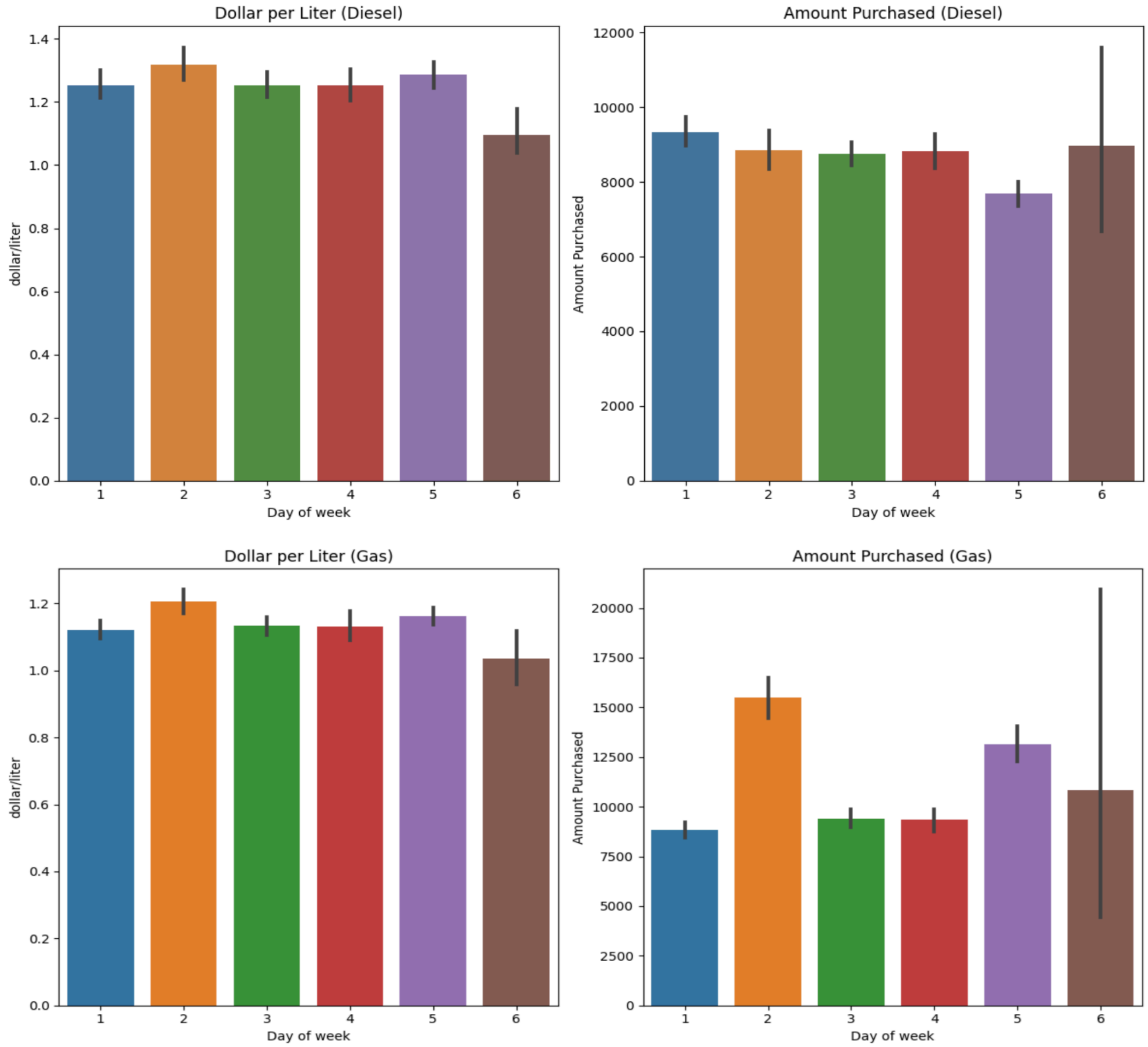
Station 2/ Gasoline (Tank 3 (70,000 liter))/ 2017



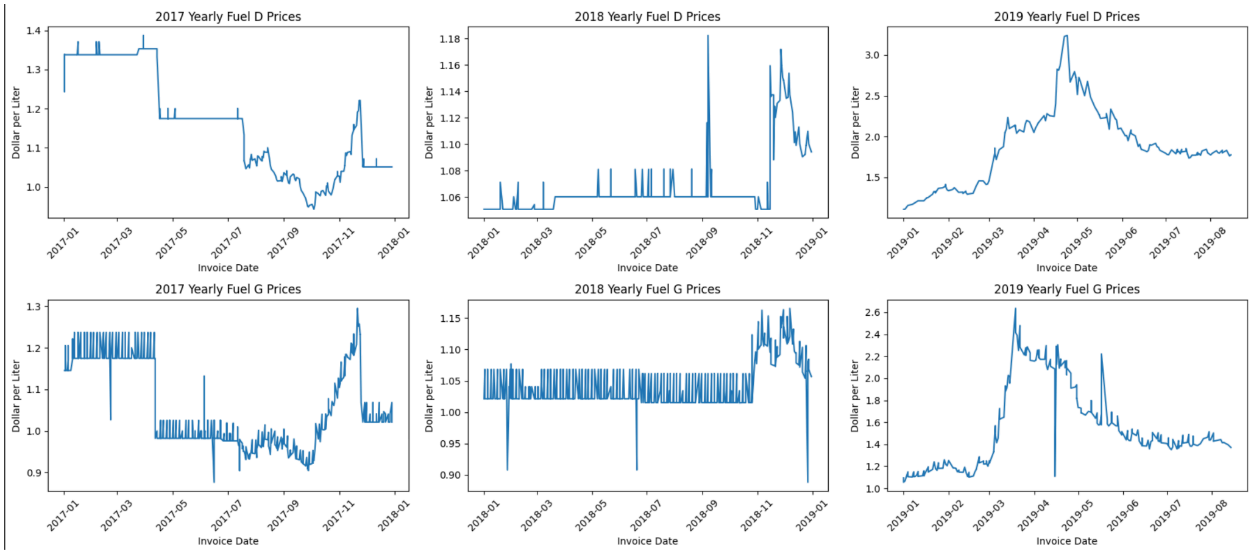


Much like the diesel data, we can see that stations frequently refilled its thanks and often had ample fuel left in the tank before the refill. There were a variety of refill pfrequencies among stations but some locations had particularly poor inventory strategy and failed to take advantage of their tank size. Station 2 for example has a 70,000 liter gas tank that is kept mostly full throughout the year. It would be advantageous for the station to refill the tank less frequently but with greater refill volumes to save on costs. Charts for each of the 7 stations analyzed are included in the attached Python file.

Cheapest Time to Buy Fuel:



Here we analyzed the mean fuel price by day of the week. Numbers on the x-axis correspond to Monday-Sunday (1-7). There is no seven included as fuel is not delivered on Sundays. We can see that Saturday had the cheapest mean fuel price. Stations should attempt to plan their purchases for saturday if possible. However their were only five diesel invoices and 8 gas invoices for saturday in the dataset. This sample size may not be large enough to conclusively say that Saturday is the cheapest day to buy.



We can also visualize the overall change in fuel price from 2017-2019 using invoice prices. Trends in this graph likely are reflective of the overall fuel commodity market and can't easily be used to recommend inventory policy.

Recommendations:

Based on the analysis above, our general recommendations for each station are as follows

- Reduce the frequency or fuel orders
- Order Fuel in larger volumes
- Allow fuel tank levels to reach low levels before ordering more fuel.
- Order fuel on Saturday when possible

These principles lead to lower overall fuel costs and significant savings. They are applicable to both diesel and Gas purchases.

- **Station 1**: Stations 1 typically buys diesel fuel in volumes under 10,000 liters which fails to maximize its potential price savings. Gas purchases have a greater range of invoice volumes but the station could have saved more than 400,000$ if it only purchase gas in volumes greater than 20,000 liters. By reducing the average number of refuels per month from eight to six, the stationwould still have ample tank capacity to handle demand and would reap the savings mentioned previously.
- **Station 2**: Station 2 has better refueling habits than station 1, buying in higher volumes of both diesel and gas. This translate to lower mean costs for both fuel types. Station 2 refills its diesel tanks more frequently than is necessary. Gas tank fuel levels are kept extremely high, rarely dropping below 80% capacity, meaning their is a significant opportunity to scale back the frequency of fuel orders and save.
- **Station 3**: Station 3 manages its fuel more efficiently than other stations, refilling less frequently and buying is larger volumes. For this reason, station 3 has the lowest mean dollar/liter price of all stations analyzed.
- **Station 4**: Station 4 has the same concerns as station 1 with regard to consistently high fuel levels and overly frequent fuel orders. This station should also attempt to order less frequently and in higher volume to same.
- **Station 5**:
- **Stations 6-7**: Stations six and seven have minimal invoice data included in our available dataset. With this limitation there is little we can do to analze and provide recommendations for these locations.

Notes on Chat GPT Usage:

- Chat GPT was used frequently to reformat visuals for readability as well as quickly replicate code for various stations and invoices
- All code was initially written by the team. Only subsequent adjustments and replications were created using GPT
- Prompts were typically written as follows:
    - "Here is my code for visualizing the relationship between … and … I would like to create subplots with axis [...] by [...] that recreates the visualizations above for each station in my dataset. Please create a subplot that fits all axis labels."