FINAL PROJECT REPORT

BU.510.615.81.SP24

Python for Data Analysis

Team 12

Team Member: Zhang, Beiming/ Sun, Shiming/ Wang, Anna/ Kang, Yuanyuan

Data Processing and Data Cleaning

First, all datasets are read from CSV files and stored in data frames.

To consider inflation in our calculations, Canadian Consumer Price Index (CPI) data[1] is imported. For each month, we obtain the current month's (1 + month-on-month inflation rate), and multiply the previous month's cumulative inflation rate by the current month's factors to get the current month's cumulative inflation rate. As shown below：

	Month	*month-on-month*	*Cumulative Inflation Rate*
0	2017-01	0.009	1.0
1	2017-02	0.002	1.002
2	2017-03	0.002	1.0040
3	2017-04	0.004	1.0080
4	2017-05	0.001	1.0090

For Fuel_level datasets, we clean the NaN of the fuel_level data, where only two values are missing. Specifically, for rows with indexes 3 and 111, we fill in the missing values from their "Tank ID" and "Fuel Level" columns with the corresponding values for adjacent rows.

Then, the two datasets of fuel level data are merged, and the Time stamp is converted to date type. We divide the data into groups based on different values in the "Tank ID" column and store each group in a dictionary.

For Locations dataset, we group Tank ids according to Tank Location and Tank Type to know which tanks are in the same area and of the same Tank Type. As shown below：

	Tank ID	*Fuel Level*	*Time stamp*
2	T 12	26934.0	1/1/2017 0:40
3	T 12	26934.0	1/1/2017 0:45
110	T 12	21566.0	1/2/2017 3:25
111	T 12	21566.0	1/2/2017 3:35

[1] Source：Statistics Canada

To clean the invoice data, we remove NA, and change G to U to conform to other datasets.

Here we noticed that in the original data set, Amount Purchased and Gross Purchase Cost may have incorrect row names. We have corrected this and will explain the reasons for our correction later.

Cumulative Inflation Price Calculating

To calculate the unit price and discount rate for each transaction, we calculate the unit price of each product and give different discounts according to the quantity purchased. After that, we extract the months of the invoice date and match each month in the invoice data with the corresponding cumulative inflation rate, and we divide the total purchase cost by the cumulative inflation rate to get the adjusted total purchase cost. Finally, we add this column Adjusted Gross Purchase Cost to the data frame:

	Invoice Date	*Invoice ID*	*Invoice Gas Station Location*	*Gross Purchase Cost*	*Amount Purchased*	*Fuel Type*	*Unit Price*	*Unit Price Before Discount*	*Cumulative Inflation Rate*	*Adjusted Gross Purchase Cost*
0	2017-01-02	10000	1	7570.82	6609.6	U	1.15	1.15	1.0	7570.82
1	2017-01-02	10001	1	12491.85	9338.74	D	1.34	1.34	1.0	12491.85
3	2017-01-02	10002	2	17034.34	13377.82	D	1.27	1.27	1.0	17034.34
5	2017-01-02	10003	2	12616.77	9432.11	D	1.34	1.34	1.0	12616.77
6	2017-01-02	10004	4	11363.8	9139.2	D	1.24	1.24	1.0	11363.8

Given that we don't know the exact price of oil daily, and the CPI figures are monthly, we directly use the average price of each oil product per month to represent the monthly oil price of that oil product.

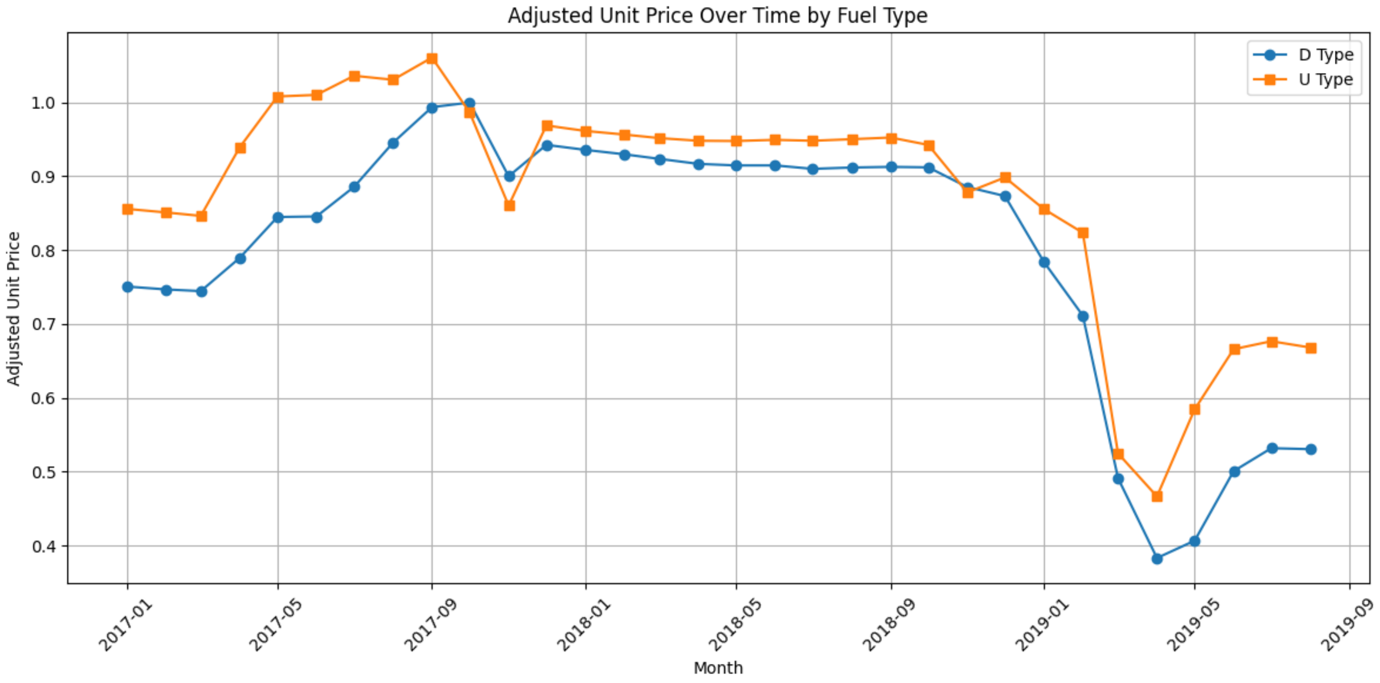
Month	*Fuel Type*	*count*	*mean*	*std*	*min*	*25%*	*50%*	*75%*	*max*
2017-01	D	45.0	1.33	0.02	1.24	1.34	1.34	1.34	1.37
2017-01	U	51.0	1.18	0.04	1.15	1.17	1.17	1.17	1.27
2017-02	D	40.0	1.34	0.01	1.34	1.34	1.34	1.34	1.37
2017-02	U	46.0	1.19	0.04	1.03	1.17	1.17	1.17	1.27
2017-03	D	49.0	1.34	0.01	1.34	1.34	1.34	1.34	1.39

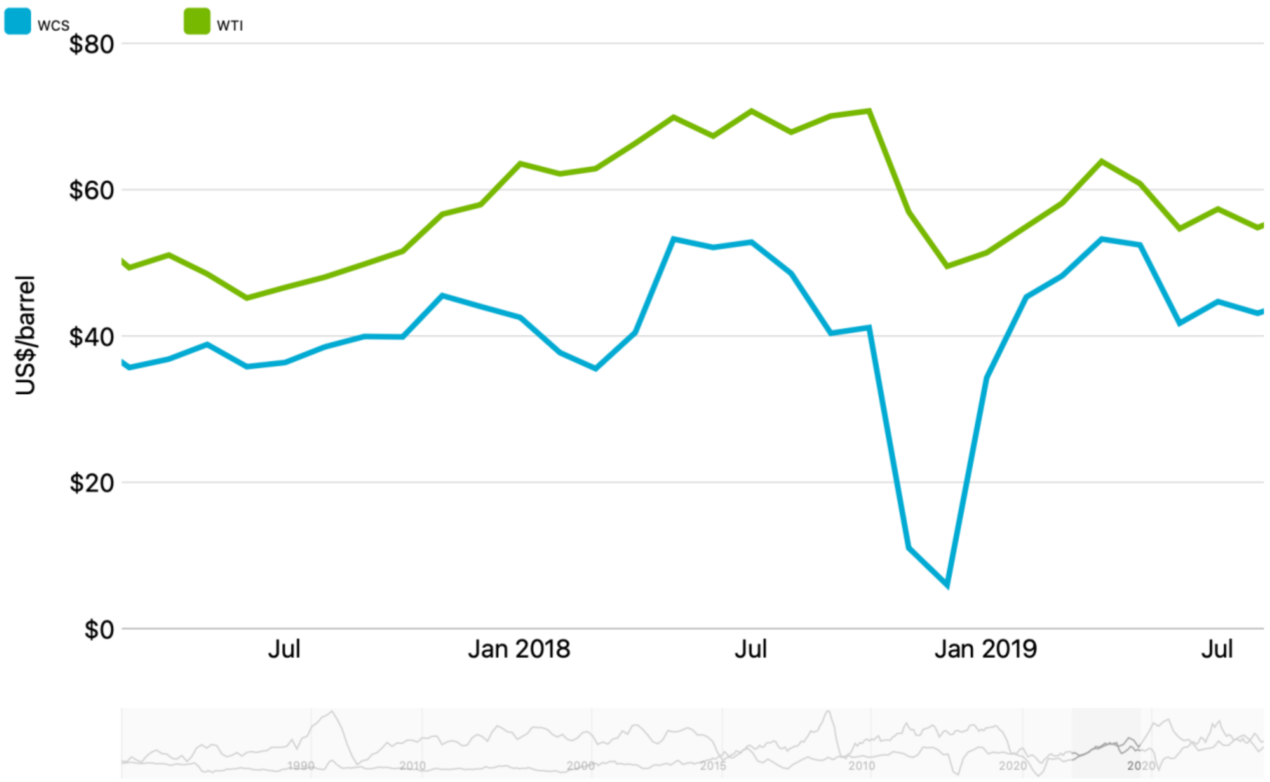
To summarize the monthly adjusted price of each type of oil, we group the data by invoice date to calculate the average unit price for each month and fuel type. The corresponding cumulative inflation rate for each month is then extracted from the Canadian Consumer Price Index (CPI) data set and connected with the invoice data.

We then obtain the inflation-adjusted unit price by dividing the original unit price for each group by the corresponding cumulative inflation rate. The result is a new data frame that contains the average unit price for each month and fuel type, the corresponding cumulative inflation rate, and the adjusted unit price.

	Month	*Fuel Type*	*Unit Price Before Discount*	*Cumulative Inflation Rate*	*Adjusted Unit Price*
0	2017-01	D	1.33	1.0	1.33
1	2017-01	U	1.18	1.0	1.18
2	2017-02	D	1.34	1.0	1.34
3	2017-02	U	1.19	1.0	1.18
4	2017-03	D	1.34	1.0	1.34

To visualize the previously calculated average unit price data for each month and fuel type, we divide the data into two data frames based on fuel type, one containing the data of fuel type "D" and the other containing the data of fuel type "U". The "Month" column in each data frame is then converted into a date-time format to ensure that the time series is properly represented in the chart.





[2]

We have obtained the prices of WTI and WCS in Canada for this period, and the trend of the data we calculated is very close to the actual oil price trend. Adding to this, our extensive sampling tests also indicate that the actual fuel consumption values are quite close to the "Gross Purchase Cost" column rather than the "Amount Purchased" column.

Therefore, we have sufficient reason to believe that there is a problem with the column names in the original invoice data. The analysis section below will proceed based on the modified invoice dataset.

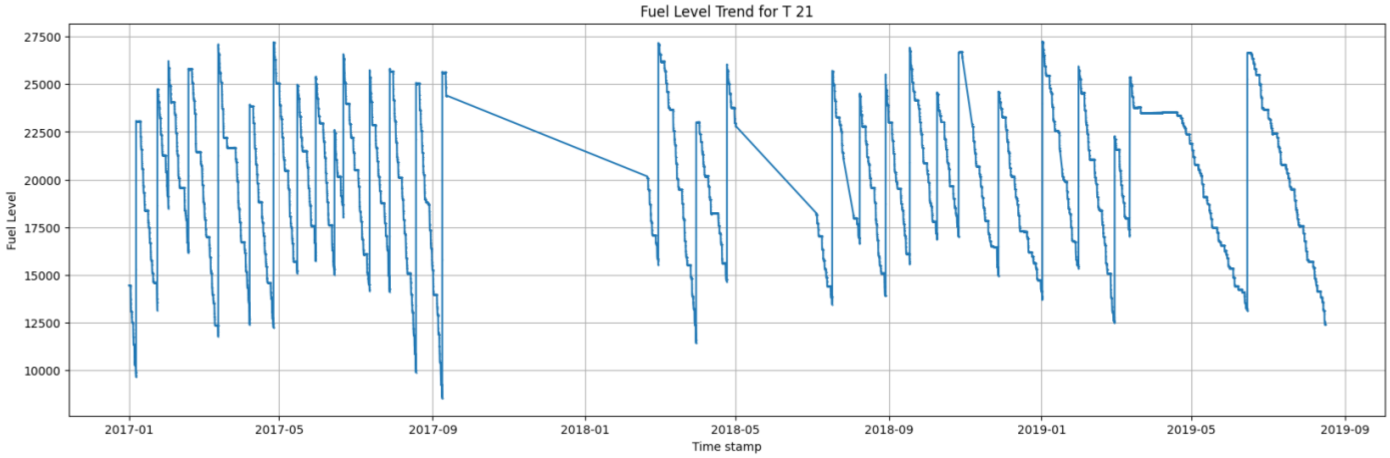
[2] Source: Alberta Economic Dashboard

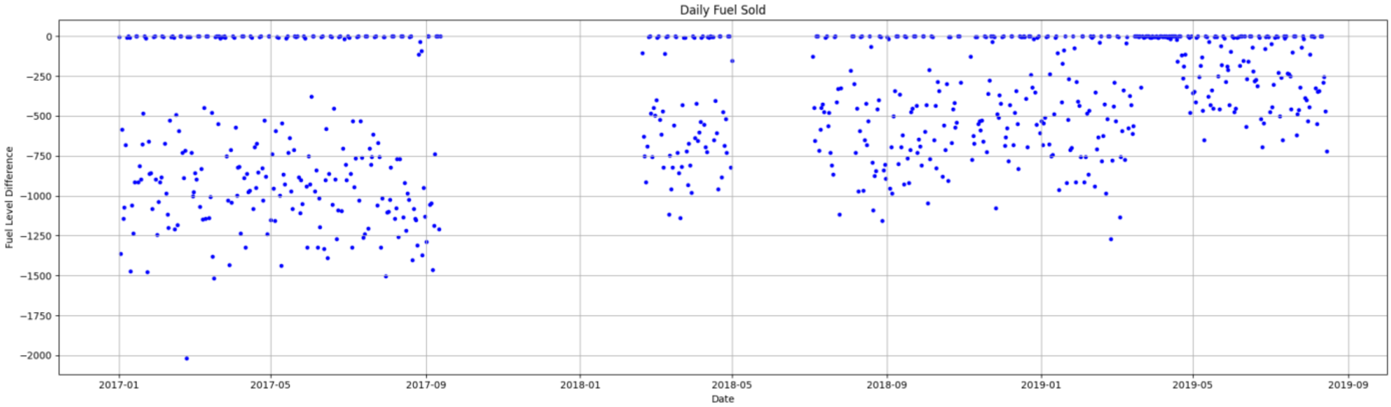
Analysis example

To plot the data, we define some functions, such as the function that plots the process of Fuel Level change over time, that plots the Fuel Level in a specific time, that calculates the daily fuel sales volume, and that plots the daily sales volume scatter plot. Now, let's take the T21 tank as an example for analysis.

Step one, we calculate the daily fuel sales of the T21 by function daily_fuel_sold().

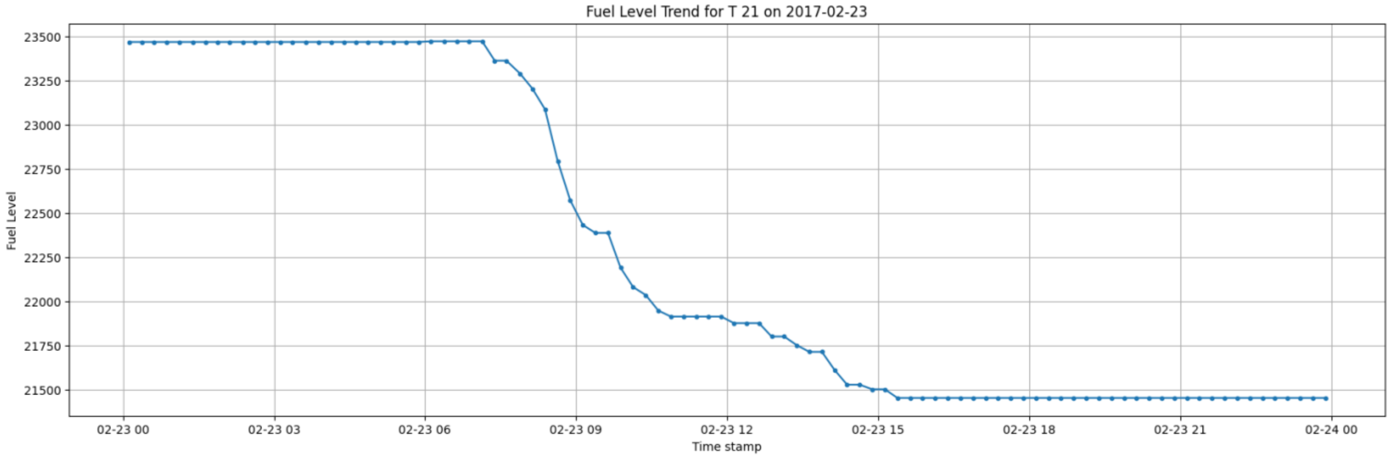
Step two, we plot the trend of T21's Fuel Level over time and plot a scatter plot of T21's daily sales.



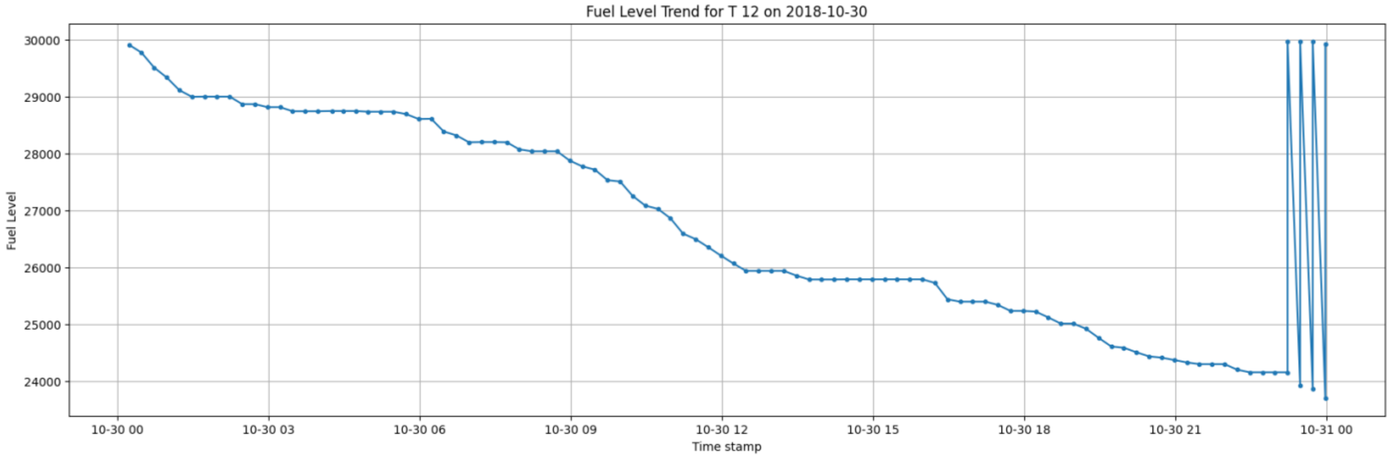


Step three, we use the image to find possible extreme points. In this example, we can easily find an extreme point at about the point where the fuel level difference is less than -2,000, and we filter out the fuel level difference of less than -2,000. In T21, the day is 2017-02-23.

Step four, we plot the trend of fuel level on this unusual day.



The data here is normal. However, we still have some abnormal values that need to be manually processed, such as the data for T10 on 2018-10-30. The specific methods of handling vary, and we have included all the processing procedures in the source code.



Step five, we organize and group the data. To simplify the problem, the model here is defined as that when the safety capacity is triggered at the end of a certain day, the gas station will order oil. The amount of ordered oil is the total capacity minus the current amount, and the oil will arrive before the next day.

Step six, we group the data as shown below:

Weekday	*count*	*mean*	*std*	*min*	*25%*	*50%*	*75%*	*max*
Friday	102.0	-592.57	366.99	-1518.0	-865.0	-594.0	-341.75	0.0
Monday	104.0	-642.25	409.55	-1503.0	-942.75	-679.5	-290.25	0.0
Saturday	101.0	-8.63	41.99	-398.0	-4.0	0.0	0.0	0.0
Sunday	102.0	-4.02	10.31	-72.0	-4.0	0.0	0.0	0.0
Thursday	103.0	-659.1	385.85	-2017.0	-906.5	-670.0	-359.5	0.0
Tuesday	104.0	-617.7	377.23	-1473.0	-892.0	-630.5	-352.5	0.0
Wednesday	102.0	-619.11	351.27	-1464.0	-862.0	-583.5	-421.0	0.0

Is US Holiday	*count*	*mean*	*std*	*min*	*25%*	*50%*	*75%*	*max*
False	700.0	-445.64	424.49	-2017.0	-765.75	-431.0	-4.0	0.0
True	18.0	-658.22	409.35	-1363.0	-933.5	-691.0	-307.75	0.0

We can learn from the tables that the most suitable day for refueling is the weekend, because no one goes to refueling. Followed by Friday, the average is relatively lowest, but the change in working days is not significant, so we think it is better to say that the weekend is more appropriate, because there are more people refueling on Monday. In addition, the American holiday will significantly increase its sales, so it is better to choose to refuel before the holiday.

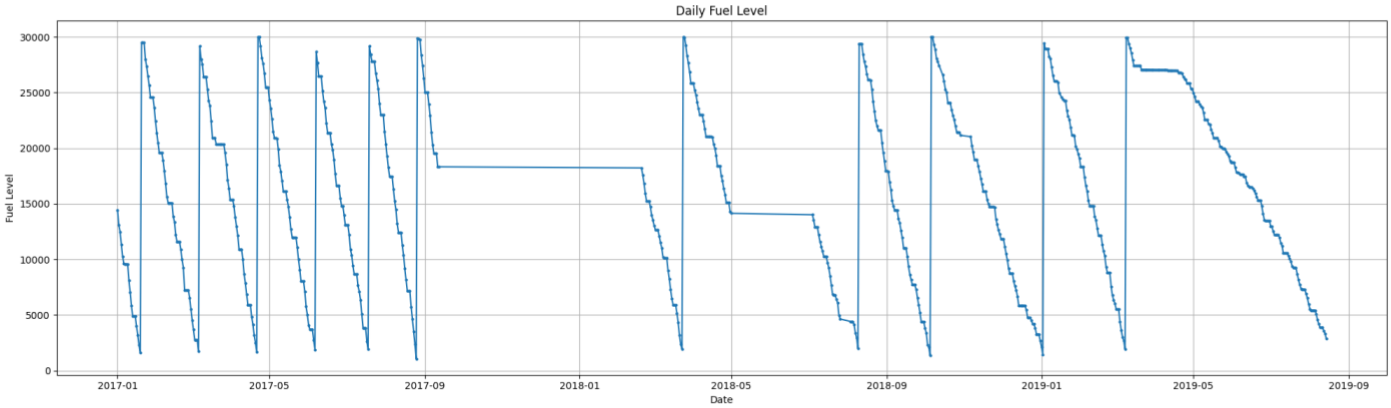
Because the fluctuation of the fuel tank at this position is relatively small, we choose a more aggressive strategy based on the fuel tank capacity of 30,000. Here, we plan to set the safe capacity to a relatively aggressive position, 1,500, and relax the refueling threshold to 2,000 on special dates.

So, the final strategy is as follows: check how much oil is left at the end of the day, if tomorrow is a weekend or holiday, then use a relaxed threshold of 2,000, if not, use a normal threshold of 1500, below which the next day fill up the tank.

Following this strategy, we update the values of the "Purchase", "Fuel Level", and "Fuel Cost" columns in the data frame to reflect fuel sales and costs. We can use the slices to see if the strategy is implemented.

	Fuel Sales	*Weekday*	*Is US Holiday*	*Purchase*	*Fuel Level*	*Fuel Cost*
2017-01-16	-916.0	Monday	True	0	3993.0	0.0
2017-01-17	-814.0	Tuesday	False	0	3179.0	0.0
2017-01-18	-897.0	Wednesday	False	0	2282.0	0.0
2017-01-19	-678.0	Thursday	False	0	1604.0	0.0
2017-01-20	-484.0	Friday	False	28396	29516.0	37052.69
2017-01-21	0.0	Saturday	False	0	29516.0	0.0
2017-01-22	-12.0	Sunday	False	0	29504.0	0.0
2017-01-23	-1476.0	Monday	False	0	28028.0	0.0
2017-01-24	-659.0	Tuesday	False	0	27369.0	0.0
2017-01-25	-863.0	Wednesday	False	0	26506.0	0.0

Step seven, we make a plot to confirm that the strategy is sound.



This strategy maximizes the amount refueled each time while ensuring there is no shortage in supply. The results meet our requirements, so we can consider our strategy effective.

Furthermore, during the comparison of costs before and after calculations, we realized that there were issues with the original dataset. If we were to operate based on the original dataset and compare costs, our costs would be higher than the previous ones. After randomly selecting several dates, comparing refueling data, and validating our suspicions, we decided to modify the invoice dataset. Clearly, the modified dataset is more reasonable.

Tank Refill Strategy Overview

All the following analyses do not consider T13, as its invoice data is missing, and there also seems to be anomalies in the fuel level data.

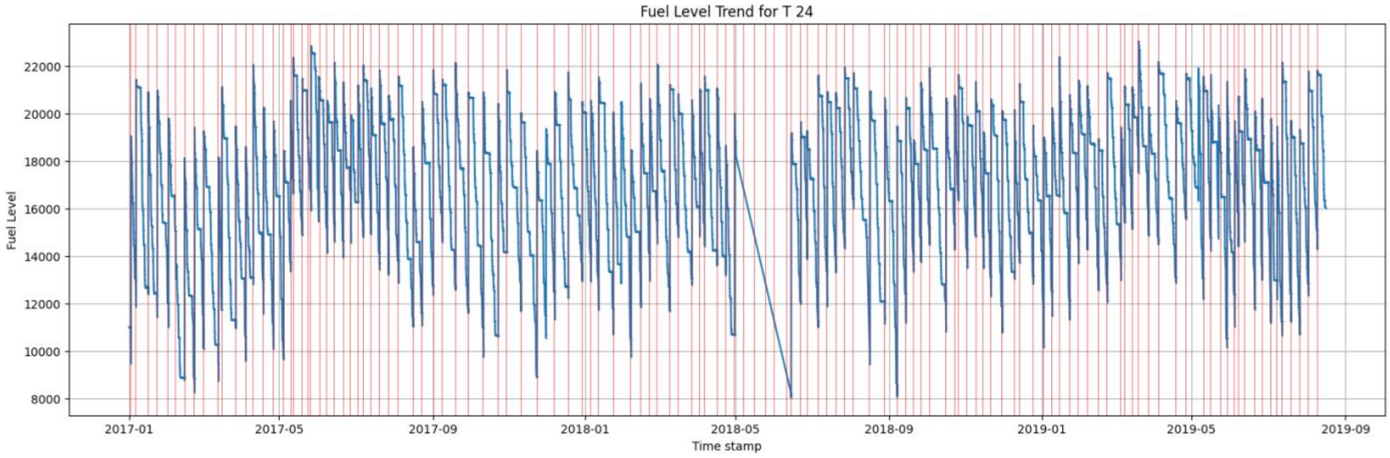
According to the case of T21, we analyze all tanks and summarize the results into the same table：

	Tank ID	*Tank Location*	*Tank Number*	*Tank Type*	*Tank Capacity*	*Best Day*	*Consider Holidays*	*Regular Threshold*	*Special Day Threshold*	*New Total Cost*
0	T 10	1	1	U	40000	Sunday	False	10000	12000	2087698.74
1	T 11	1	2	U	40000	Sunday	False	9000	10000	3420847.22
2	T 12	1	3	D	40000	Sunday	True	10000	11000	3054461.64
3	T 13	1	4	P	40000					
4	T 14	1	5	U	40000	Sunday	False	7500	9000	2337339.58
5	T 15	1	6	D	40000	Sunday	True	10000	11000	1721713.38
6	T 16	2	1	U	70000	Sunday	True	3300	3300	1299439.46
7	T 17	2	2	D	40000	Sunday	True	6000	7000	1802083.7
8	T 18	2	3	U	40000	Sunday	True	3500	4000	1197326.77
9	T 19	2	4	D	70000	Sunday	True	5500	6000	1207883.25
10	T 20	3	1	U	30000	Saturday	True	2300	2500	316108.13
11	T 21	3	2	D	30000	Sunday	True	1500	2000	245116.57
12	T 22	4	1	U	40000	Sunday	True	4000	4500	1388584.49
13	T 23	4	2	D	40000	Sunday	True	6000	6500	1411101.77
14	T 24	5	1	D	25000	Saturday	True	3000	3500	738949.18
15	T 25	5	2	U	25000	Sunday	True	3500	3500	1213768.92
16	T 26	6	1	U	30000	Sunday	False	800	900	70632.82
17	T 27	6	2	U	30000	Sunday	True	1100	1200	259322.03
18	T 28	6	3	D	30000	Sunday	True	750	750	48418.25
19	T 29	7	1	U	5000	Sunday	True	400	400	80232.89
20	T 30	7	2	D	5000	Saturday	False	200	250	11033.82
21	T 31	8	1	D	40000	Sunday	True	800	800	55540.67
22	T 32	8	2	U	40000	Saturday	True	800	900	160707.12

We figure out the sum of the original Adjusted Gross Purchase Cost and the sum of New Total Cost. According to our calculations, according to the original refueling plan, the adjusted total cost is approximately 2,841,2944.96 Canadian dollars, and the adjusted total cost is approximately 24,128,310.40 Canadian dollars, with a difference of 4,284,634.56 Canadian dollars, a saving of about 15%.

However,   given the missing of some invoice data and the integrity of fuel_level data, we cannot know the exact total price, so only the estimated savings value is here. In fact, the previous cost should be about 10% larger than the value we're using now. We estimate that the final cost reduction will be in the range of 4-6%.

We can identify some reasons for the deviation in the following graph. In the graph below, we have marked the refueling days with red lines based on the fuel level data. It is apparent that from May 2018 to July 2018, there is a period where we only have invoice data without corresponding fuel level data. This issue exists for almost all fuel tanks, and thus, it is the primary reason for the data inflation.

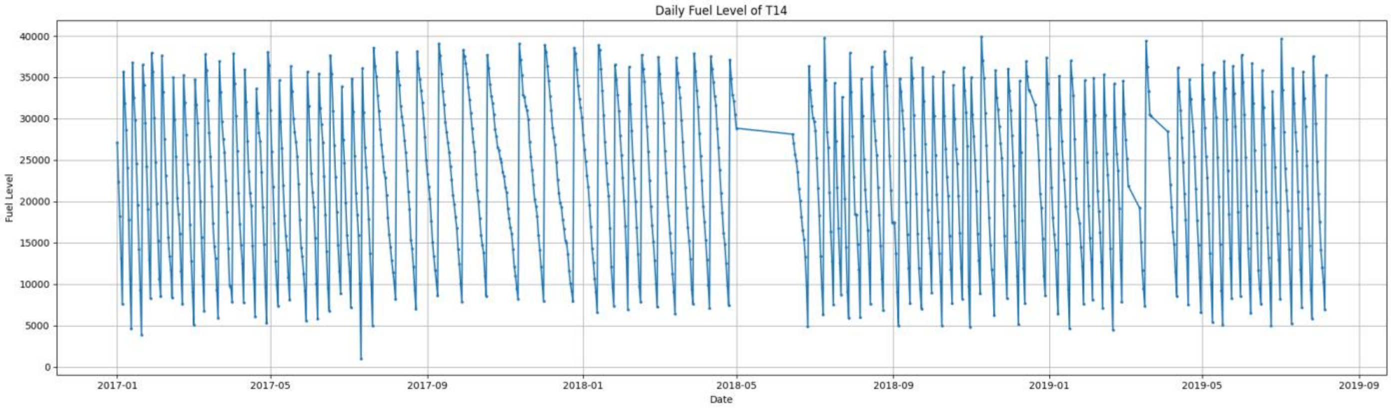


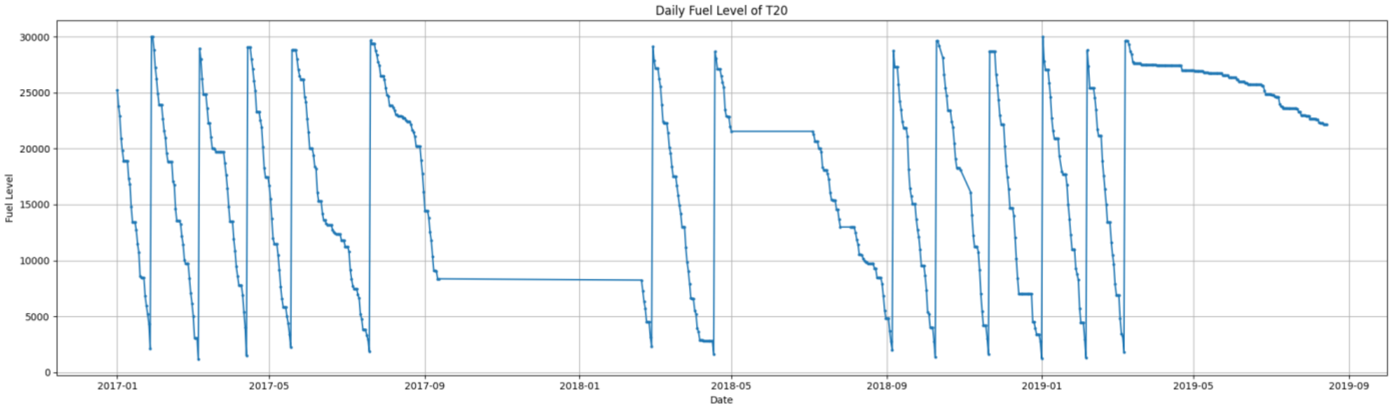
Subsequently, we attempted to correct this clearly inflated ratio by excluding fuel purchases present in the invoice but not reflected in the fuel level data. After doing so, our new raw cost totaled 26,414,367.08 Canadian dollars. Based on this value, we saved approximately 8.65% of the total cost. This value remains slightly inflated (due to the partial discrepancy between fuel metering data and invoices). However, we believe it is very close to our actual value and is sufficient to demonstrate that our approach is correct and effective.

Based on the principles and results of the calculation, we can consider this strategy to be significantly effective.

Tank Expansion Strategy Overview

Our decision to expand fuel tank capacity is based on the following process: when the fuel tank capacity approaches the discount threshold + safety threshold, we consider increasing the tank's capacity to this value. The logic behind this is straightforward, as we aim to secure the maximum discount while ensuring our supply. When considering expansion decisions, we also need to consider the Return on Investment (ROI). Unfortunately, we do not have data on the cost of expanding fuel tanks; hence, we will observe the fueling cycle to decide. Generally, more frequent fueling cycles indicate a higher total volume of fueling, and our costs for expanding the fuel tank will be amortized more quickly. Some tanks, like T11 and T12, clearly fit this strategy. Therefore, we will present our tank expansion strategy in a quantitative descriptive manner. However, it is challenging to provide a quantified return on investment (considering the objective conditions, limitations, schedule, and costs related to tank expansion are unclear), so this strategy is presented independently of any purchasing strategy and will not be further integrated with it. We want to clarify that our proposed solutions are based on the medium to long-term ROI (possibly 3-5 years) of the "tank expansion" decision.





We have selected two representative fuel tanks. T14 has a high number of refueling instances throughout the entire period, while its capacity is capped at 40,000, making it suitable for expanding the tank size. On the other hand, T20 has fewer refueling instances and is also further from the next supplier discount threshold (40,000), therefore not suitable for expanding the tank size, as the potential ROI is too low.

After our observation, T10, T12, T11, T14, T15, T17, T18, T22, T23, T24, T25 are the fuel tanks suitable for expansion. Interestingly, their original tank capacities all corresponded to a certain supplier discount threshold, so we made some simplifications in our calculations. Our final fuel tank expansion plan is as follows:

	Tank ID	*Tank Location*	*Tank Number*	*Tank Type*	*Tank Capacity*	*Regular Threshold*	*Special Day Threshold*	*If Expansion*	*Refill Frequency*	*Final Capacity*
0	T 10	1	1	U	40000	10000	12000	True	85.0	52000
1	T 11	1	2	U	40000	9000	10000	True	122.0	50000
2	T 12	1	3	D	40000	10000	11000	True	126.0	51000
3	T 13	1	4	P	40000			False		40000
4	T 14	1	5	U	40000	7500	9000	True	85.0	49000
5	T 15	1	6	D	40000	10000	11000	True	66.0	51000
6	T 16	2	1	U	70000	3300	3300	False	22.0	70000
7	T 17	2	2	D	40000	6000	7000	True	63.0	47000
8	T 18	2	3	U	40000	3500	4000	True	37.0	44000
9	T 19	2	4	D	70000	5500	6000	False	23.0	70000
10	T 20	3	1	U	30000	2300	2500	False	13.0	30000
11	T 21	3	2	D	30000	1500	2000	False	11.0	30000
12	T 22	4	1	U	40000	4000	4500	True	44.0	44500
13	T 23	4	2	D	40000	6000	6500	True	52.0	46500
14	T 24	5	1	D	25000	3000	3500	True	43.0	28500
15	T 25	5	2	U	25000	3500	3500	True	62.0	28500
16	T 26	6	1	U	30000	800	900	False	3.0	30000
17	T 27	6	2	U	30000	1100	1200	False	10.0	30000
18	T 28	6	3	D	30000	750	750	False	2.0	30000
19	T 29	7	1	U	5000	400	400	False	20.0	5000
20	T 30	7	2	D	5000	200	250	False	3.0	5000
21	T 31	8	1	D	40000	800	800	False	2.0	40000
22	T 32	8	2	U	40000	800	900	False	5.0	40000

In our expansion plan, fuel tanks with higher refueling frequencies can bring about higher ROI. For example, T11 and T12 are notably significant in yielding higher returns from expanding the tank capacity, followed by T10 and T15, and so forth. We refrain from providing quantitative rankings here due to the absence of data for some fuel level readings. However, from a qualitative perspective, the ROI ranking we provide is as follows: T11/T12 - T10/T15 - T15/T17/T25 - T23 - T22/T24 - T18. Tanks listed earlier have higher priority for tank expansion. Of course, we have not considered many unknown factors, so this ranking is for reference purposes only.

Conclusion and Summary

In this assignment, we cleaned the original data, corrected necessary errors, and introduced new datasets to reflect the impact of inflation.

Regarding the tanks refill plan, we calculated the daily sales and replenishment volumes for each tank and determined different thresholds based on these data, as well as the most suitable day for refueling within a week. The solution we finally proposed effectively increased the refueling volume per occurrence and significantly reduced the total cost by approximately 8.65%, despite a certain degree of overestimation in the data.

For the tank expansion plan, based on the previous refueling plan, we qualitatively identified suitable tanks for expansion and set expansion targets. We also provided a reference priority ranking. We believe that these strategies can effectively help gas stations increase profits while maintaining safe fuel reserves.

However, there are some limitations to these solutions. For instance, we only selected one day as the recommended refueling day and did not provide further plans to boost sales. Nevertheless, considering the limitations of the dataset and incomplete information, we believe the solutions we proposed are acceptable and have significant effects.