Python for Data Analysis

Project Report

Team Five

Group Members:

Avril Zhang

Kaiying Zhu

Rongyi Lai

Tsega Tsewameskel

1 Introduction

1.1 Brief Introduction

This project aims to analyze the fuel price and purchasing order data to evaluate how well the gas stations manage their fuel tanks' inventory. The goal is to suggest a better inventory policy that can save the gas stations money while maintaining an excellent customer service level.

To achieve this goal, our team visualizes the inventory evolution trajectory and gain insights into the inventory management practices. In order to quantify performance, we compare the amount of money saved to the maximum potential savings possible if the purchasing strategy is optimized. It is important to consider inflation in the calculations, as the purchasing power of money changes over time. Therefore, our team found Canada's monthly inflation rates, and created a new dataset with these rates, and join it with the existing data to ensure that the calculations are accurate.

Overall, our team aims to achieve a thorough understanding of the inventory management practices, cost structure, and overall efficiency of the gas stations.

1.2 Data Description

The datasets described are related to gas stations and fuel tanks.

1.2.1 Locations

The "Locations.csv" dataset contains information about gas station locations, including a unique ID for each gas station, the gas station name, address, latitude and longitude.

1.2.2 Tank

The "Tank.csv" dataset contains information about fuel tanks, including a unique ID for each tank, the gas station it belongs to, a tank number, the fuel type (G for regular gas, D for Diesel, and P for premium) and the tank capacity in liters.

1.2.3 Invoices

The "Invoices.csv" dataset contains information about purchases of different fuel types by gas stations from their suppliers, including the date of purchase, a unique ID for each invoice, the gas station location, the total cost in Canadian Dollars, the total amount of fuel purchased in liters, and the fuel type purchased.

1.2.4 Fuel Level

The "Fuel_Level_Part_1.csv" and "Fuel_Level_Part_2.csv" datasets contain fuel level information in each tank at frequent and mostly regular time stamps, including

the tank ID, the amount of remaining fuel in liters, and the time of inventory reporting.

1.2.5 Canada Inflation Rate

The "canada_infl.csv" dataset contains information about Canada's monthly inflation rate from January 2017 to December 2019. The data is obtained from the website "https://www.statcan.gc.ca/".

1.3 Overview

The organization of the paper is as follows. Section 2 describes the data processing. Section 3 describes the data exploration. Section 4 presents the business questions analysis with results. Section 5 briefly concludes this report.

2 Data Processing

2.1 Data Cleaning

Firstly, we import all CSV files, and we extract all columns to see what data we have, then we rename all columns to make it valid. Since the tank type G and U are the same thing, so we use replace operation to replace U with G. Secondly, we check for duplicates and null values then drop them. Lastly, we explore the data with info operation.

2.2 Data Merging

We use concat function to merge fuel_1 and fuel_2 file, then we merge invoices and tanks file to create dataframe inventory. The last step is using to_datetime operation to transfer 'Invoice_date' to date format.

3 Data Exploration

In this part, our code is aiming to do some prior preparation for further analysis.

Firstly, we calculate the total volume of fuel purchased for each fuel type at each gas station. we perform a groupby operation on the "invoices" dataframe, grouping by the columns "Station_location" and "Fuel_type". We then calculate the sum of the Amount_purchased column for each group and create a new dataframe called fuel_purchased. The .reset_index() method is used to reset the index of the resulting dataframe to a default index, and the ".head()" method is used to display the first few rows of the resulting dataframe.

Secondly, in order to check for outliers, we use box plot operations, and it shows many outliers above the 75% IQR. Then we create a histogram and heat map to see

the fuel price distribution and correlation between purchase amount and cost. It shows that the purchase price is concentrated in 5000-15000 and the amount purchased has high relation with the total cost.

Thirdly, we calculate the frequency of replenishment and fuel purchase for each tank location per day. We perform aggregation operations on the "invoices" dataset and calculate various statistics for each gas station and fuel type. The statistics are then renamed with more readable column names. We also convert the earliest and latest transaction dates to date-time format then calculate the number of days between them (Transaction_Date_diff) for each gas station and fuel type. Additionally, we calculate the average fuel purchased per day (Fuel_Purchased_per-day), which is the total fuel purchased divided by the number of days between the earliest and latest transaction dates. Finally, we reset the index with reset_index() function for further analysis and visualization.

Fourthly, we sort the data by tank ID and timestamp with sort_values operation to check the change of each tank in a period of time.

Lastly, we create a data frame from a csv with monthly inflation rates of Canada from 2017-2019 found from Statistics Canada and a new column Month_year containing the month of and year of the invoices. Then we merge the inflation dataframe with invoices to adjust the Gross Purchase Cost column for each invoice based on the inflation rate of the month it was purchased. After dropping null values, we draw boxplots to make an overview of fuel prices in a week for each type in each location.

4 Data Analysis

This project focuses mainly on the following four business questions to achieve a thorough understanding of the inventory management practices, cost structure, and overall efficiency of the gas stations.

4.1 Business Question One

4.1.1 Key Question

How to order to minimize the cost and maximize the discount but not run out of fuel for each location?

4.1.2 Question Solving

In order to find the optimal replenishment frequency and potential improvement for each gas station and fuel type, our group followed the following steps:

1. Define two functions: "optimal_replenishment" that takes the quantity of fuel purchased and returns the discount rate based on the purchased quantity, and "days_to_highest_discount" that takes the quantity of fuel purchased and returns the number of days needed to reach the highest discount rate.
2. Calculate the average number of days between replenishment for each gas station and fuel type using the "diff" method of the "Invoice_date" column, the average replenishment quantity for each gas station and fuel type, and merge the frequency and quantity data into a new dataset called "replenishment_summary".
3. Apply the "optimal_replenishment" function to the average replenishment quantity to get the discount rate for each gas station and fuel type.
4. Calculate the total cost, the number of days needed to reach the highest discount rate, the optimal replenishment frequency and the potential improvement for each gas station and fuel type based on the optimal replenishment quantity.
5. Print out the optimal replenishment frequency and potential improvement for each gas station and fuel type.

4.1.3 Result

Optimal replenishment frequency and potential improvement:

For station 1 and fuel type D, the optimal replenishment frequency is 0.92 days with a potential improvement of 56.69%.

For station 1 and fuel type G, the optimal replenishment frequency is 1.45 days with a potential improvement of 18.91%.

For station 2 and fuel type D, the optimal replenishment frequency is 2.6 days with a potential improvement of 38.75%.

For station 2 and fuel type G, the optimal replenishment frequency is 4.0 days with a potential improvement of 31.16%.

For station 3 and fuel type D, the optimal replenishment frequency is 8.26 days with a potential improvement of 57.87%.

For station 3 and fuel type G, the optimal replenishment frequency is 10.16 days with a potential improvement of 56.56%.

For station 4 and fuel type D, the optimal replenishment frequency is 1.87 days with a potential improvement of 64.32%.

For station 4 and fuel type G, the optimal replenishment frequency is 2.04 days with a potential improvement of 63.29%.

For station 5 and fuel type D, the optimal replenishment frequency is 1.99 days with a

potential improvement of 74.02%.

For station 5 and fuel type G, the optimal replenishment frequency is 2.17 days with a potential improvement of 65.02%.

For station 6 and fuel type D, the optimal replenishment frequency is 43.96 days with a potential improvement of 68.78%.

For station 6 and fuel type G, the optimal replenishment frequency is 16.3 days with a potential improvement of 48.44%.

For station 7 and fuel type D, the optimal replenishment frequency is 5.24 days with a potential improvement of 94.46%.

For station 7 and fuel type G, the optimal replenishment frequency is 1.82 days with a potential improvement of 91.49%.

For station 8 and fuel type D, the optimal replenishment frequency is 32.79 days with a potential improvement of 66.2%.

For station 8 and fuel type G, the optimal replenishment frequency is 21.55 days with a potential improvement of 56.26%.

4.2 Business Question Two

4.2.1 Key Question

What day to order? Which day of a week is the cheapest one for each location?

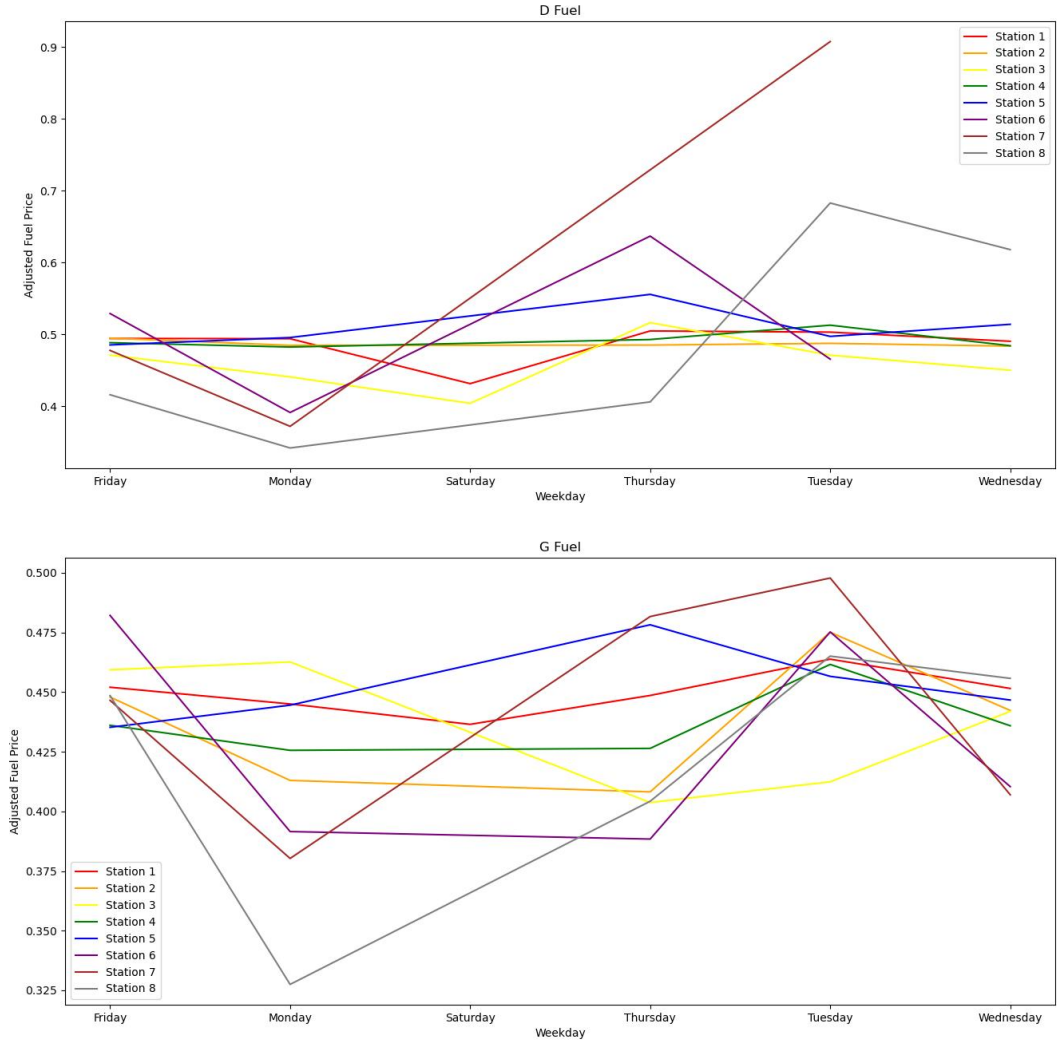
4.2.2 Question Solving

In this part, we analyzed historical data to identify the cheapest day of the week to order fuel for each location, which can help save money and optimize fuel purchase. Firstly, we import seaborn and matplotlib.pyplot to create lineplot for the mean adjusted fuel price for each weekday.

Secondly, to calculate the exact mean price, we used the following steps.

1. Use For Loop through each station location and fuel type combination.
2. Create a new column 'Weekday' which maps the day of the week from the 'Invoice_date' column to the corresponding weekday name.
3. Calculate the mean price of the fuel type for each weekday using the groupby function and sort the resulting dataframe in ascending order based on the adjusted fuel price.
4. Extract the weekday with the lowest adjusted fuel price from the sorted dataframe and store it in the 'best_day_value' variable along with its corresponding price.
5. Give the result for type D fuel and type G fuel along with the station location and weekday where the lowest price occurred.

4.2.3 Result



We can see that for both D Fuel and G Fuel, station 8 has the lowest price on Monday.

The second part of the code shows the following result:

The best date price of Type D fuel in the station 1 is Saturday with a price of 0.43

The best date price of Type G fuel in the station 1 is Saturday with a price of 0.44

The best date price of Type D fuel in the station 2 is Wednesday with a price of 0.48

The best date price of Type G fuel in the station 2 is Thursday with a price of 0.41

The best date price of Type D fuel in the station 3 is Saturday with a price of 0.40

The best date price of Type G fuel in the station 3 is Thursday with a price of 0.40

The best date price of Type D fuel in the station 4 is Monday with a price of 0.48

The best date price of Type G fuel in the station 4 is Monday with a price of 0.43

The best date price of Type D fuel in the station 5 is Friday with a price of 0.49

The best date price of Type G fuel in the station 5 is Friday with a price of 0.44

The best date price of Type D fuel in the station 6 is Monday with a price of 0.39
The best date price of Type G fuel in the station 6 is Thursday with a price of 0.39
The best date price of Type D fuel in the station 7 is Monday with a price of 0.37
The best date price of Type G fuel in the station 7 is Monday with a price of 0.38
The best date price of Type D fuel in the station 8 is Monday with a price of 0.34
The best date price of Type G fuel in the station 8 is Monday with a price of 0.33
The lowest price for type D fuel is 0.34 in station 8 on Monday
The lowest price for type G fuel is 0.33 in station 8 on Monday

4.3 Business Question Three

What is the optimal tank size? Should we increase the capacity of tank size? If we increase the capacity of tank size, how much would we save per year?

4.3.2 Question Solving

Firstly, we calculate the tank capacity and average fuel volume for each gas station and tank and compute the utilization rate of each tank. The concrete steps are here as follows:

1. We group the "inventory" dataset by gas station and tank ID, calculating the maximum tank capacity for each tank.
2. We group the "invoices" dataset by gas station, calculating the average fuel volume for each gas station.
3. Then we merge the tank capacity and average fuel volume data by gas station to calculate the utilization rate of each tank, which is the amount of fuel purchased divided by the tank capacity.
4. We use the Seaborn library's barplot function to group the tank utilization rate by tank ID and plot it as a bar chart for each gas station, for better comparison and visualization.

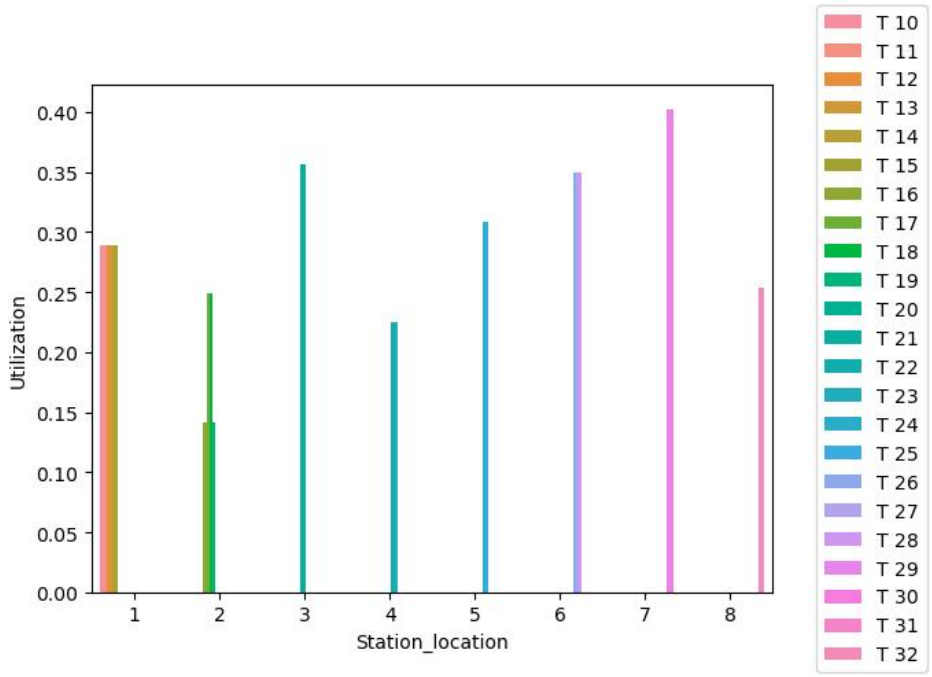
Secondly, we explore the best tanks capacity for each tank. So, we define two functions ‘calculate_new_tank_size’ and ‘generate_recommendations’ to generate recommendations for stations to switch to a larger tank size based on their utilization and current tank size.

1. The calculate_new_tank_size function takes in a station's utilization, current tank size, and the number of days between deliveries and returns the optimal tank size.
2. The generate_recommendations function takes in a DataFrame with information about stations' locations, fuel types, current tank sizes, utilization, and potential cost savings associated with switching to a larger tank size. It iterates over each

row of the DataFrame and constructs a recommendation string based on the station's information, potential cost savings, and the recommended new tank size.

3. Then we merge data on station performance and tank capacities, calculate the optimal tank size for each station using the calculate_new_tank_size function, sort the recommendations by potential cost savings, and the code returns the sorted DataFrame.
4. In the end, we generate a list of recommendations based on the input DataFrame recommendations_df. The list of recommendations is then printed to the console using a for loop.

4.3.3 Result



We can see that station 7 has the largest utilization, with station 3 following closely behind.

The result of recommended tanks capacity for each tank is shown below:

For station 1 and fuel type G, switching to a 11564.0479 gallon tank can save up to $1624421.7 in gross purchase costs per year. This would result in a 15.0% savings.

For station 1 and fuel type D, switching to a 11564.0479 gallon tank can save up to $1026449.34 in gross purchase costs per year. This would result in a 21.0% savings.

For station 2 and fuel type D, switching to a 11564.0479 gallon tank can save up to $561409.42 in gross purchase costs per year. This would result in a 16.0% savings.

For station 4 and fuel type D, switching to a 9942.1101 gallon tank can save up to $335436.39 in gross purchase costs per year. This would result in a 21.0% savings.

For station 2 and fuel type G, switching to a 11564.0479 gallon tank can save up to $221023.31 in gross purchase costs per year. This would result in a 8.0% savings.
For station 5 and fuel type D, switching to a 9942.1101 gallon tank can save up to $182946.86 in gross purchase costs per year. This would result in a 23.0% savings.
For station 4 and fuel type G, switching to a 9942.1101 gallon tank can save up to $152154.99 in gross purchase costs per year. This would result in a 10.0% savings.
For station 5 and fuel type G, switching to a 9942.1101 gallon tank can save up to $140976.04 in gross purchase costs per year. This would result in a 10.0% savings.
For station 3 and fuel type D, switching to a 11564.0479 gallon tank can save up to $61482.05 in gross purchase costs per year. This would result in a 14.0% savings.
For station 6 and fuel type G, switching to a 10691.6722 gallon tank can save up to $34538.34 in gross purchase costs per year. This would result in a 9.0% savings.
For station 8 and fuel type G, switching to a 7708.7385 gallon tank can save up to $25702.09 in gross purchase costs per year. This would result in a 12.0% savings.
For station 3 and fuel type G, switching to a 11564.0479 gallon tank can save up to $25446.43 in gross purchase costs per year. This would result in a 6.0% savings.
For station 8 and fuel type D, switching to a 7708.7385 gallon tank can save up to $11922.86 in gross purchase costs per year. This would result in a 16.0% savings.
For station 6 and fuel type D, switching to a 10691.6722 gallon tank can save up to $11387.6 in gross purchase costs per year. This would result in a 21.0% savings.
For station 7 and fuel type G, switching to a 9019.9051 gallon tank can save up to $10596.53 in gross purchase costs per year. This would result in a 11.0% savings.
For station 7 and fuel type D, switching to a 9019.9051 gallon tank can save up to $2113.68 in gross purchase costs per year. This would result in a 19.0% savings.

4.4 Business Question Four

What is the optimal frequency to refill, and what's the maximum discount in this strategy compared with the current discount? Identify which fuel stations would benefit most.

4.4.2 Question Solving

We used the following steps to solve this question:

1. Perform an inventory analysis on a DataFrame invoices that contains information about fuel purchases for different stations on different dates. Then we group the data by station and date using the groupby method and calculate the total fuel purchased and total cost for each group. Finally we calculate the average fuel

price for each group, the inventory change for each group, and combine all the results into a single DataFrame inventory_analysis.

2. Compute various statistics related to fuel consumption using the invoices DataFrame. We group the DataFrame by station, fuel type, and invoice date using the groupby method, and calculate the total fuel purchased for each group. We then add a new column Day_of_Week to the DataFrame that contains the day of the week for each invoice date. Next, the code calculates the average daily consumption for each gas station and fuel type by grouping by station, fuel type, and invoice date, and applying the mean method to the Amount_purchased column using the transform method. The result is stored in a new column Avg_Daily_Consumption. Finally, we calculate the daily change in consumption for each gas station and fuel type by grouping by station and fuel type, and applying the diff method to the Avg_Daily_Consumption column. The result is stored in a new column Daily_Change.
3. Calculate the potential cost savings for each fuel purchase using an optimal replenishment function, and then group the data by gas station and fuel type to calculate the actual cost savings and savings percentage. The resulting data is stored in a DataFrame called performance_df, which includes columns for the gas station location, fuel type, potential cost savings, actual cost savings, savings potential, and savings percentage.
4. The last step is to print the maximum possible savings that could be achieved by a gas station for each fuel type, assuming that the station always purchased fuel at the highest discount rate. We use data from a performance_df DataFrame which contains the actual and potential cost savings for each gas station and fuel type.

Moreover, In order to find the fuel station and tank type that benefit most, we perform data analysis on a dataset related to fuel stations, fuel types, and potential cost savings. We sort the dataset by the Savings_Percentage column in descending order, groups the data by Station_location column, calculate the sum of Savings_Potential and Potential_Cost columns for each group, and calculate the Savings_Percentage column.

4.4.3 Result

Maximum Possible Savings:

For station 1 and fuel type D, the maximum possible savings that could be achieved if

the gas station always purchased fuel at the highest discount rate is up to $1026449.34 in gross purchase costs per year. This would result in a 21.00% savings.
For station 1 and fuel type G, the maximum possible savings that could be achieved if the gas station always purchased fuel at the highest discount rate is up to $1624421.70 in gross purchase costs per year. This would result in a 15.00% savings.
For station 2 and fuel type D, the maximum possible savings that could be achieved if the gas station always purchased fuel at the highest discount rate is up to $561409.42 in gross purchase costs per year. This would result in a 16.00% savings.
For station 2 and fuel type G, the maximum possible savings that could be achieved if the gas station always purchased fuel at the highest discount rate is up to $221023.31 in gross purchase costs per year. This would result in a 8.00% savings.
For station 3 and fuel type D, the maximum possible savings that could be achieved if the gas station always purchased fuel at the highest discount rate is up to $61482.05 in gross purchase costs per year. This would result in a 14.00% savings.
For station 3 and fuel type G, the maximum possible savings that could be achieved if the gas station always purchased fuel at the highest discount rate is up to $25446.43 in gross purchase costs per year. This would result in a 6.00% savings.
For station 4 and fuel type D, the maximum possible savings that could be achieved if the gas station always purchased fuel at the highest discount rate is up to $335436.39 in gross purchase costs per year. This would result in a 21.00% savings.
For station 4 and fuel type G, the maximum possible savings that could be achieved if the gas station always purchased fuel at the highest discount rate is up to $152154.99 in gross purchase costs per year. This would result in a 10.00% savings.
For station 5 and fuel type D, the maximum possible savings that could be achieved if the gas station always purchased fuel at the highest discount rate is up to $182946.86 in gross purchase costs per year. This would result in a 23.00% savings.
For station 5 and fuel type G, the maximum possible savings that could be achieved if the gas station always purchased fuel at the highest discount rate is up to $140976.04 in gross purchase costs per year. This would result in a 10.00% savings.
For station 6 and fuel type D, the maximum possible savings that could be achieved if the gas station always purchased fuel at the highest discount rate is up to $11387.60 in gross purchase costs per year. This would result in a 21.00% savings.
For station 6 and fuel type G, the maximum possible savings that could be achieved if the gas station always purchased fuel at the highest discount rate is up to $34538.34 in

gross purchase costs per year. This would result in a 9.00% savings.
For station 7 and fuel type D, the maximum possible savings that could be achieved if the gas station always purchased fuel at the highest discount rate is up to $2113.68 in gross purchase costs per year. This would result in a 19.00% savings.
For station 7 and fuel type G, the maximum possible savings that could be achieved if the gas station always purchased fuel at the highest discount rate is up to $10596.53 in gross purchase costs per year. This would result in a 11.00% savings.
For station 8 and fuel type D, the maximum possible savings that could be achieved if the gas station always purchased fuel at the highest discount rate is up to $11922.86 in gross purchase costs per year. This would result in a 16.00% savings.
For station 8 and fuel type G, the maximum possible savings that could be achieved if the gas station always purchased fuel at the highest discount rate is up to $25702.09 in gross purchase costs per year. This would result in a 12.00% savings.

To identify which fuel stations would benefit most, we found that Fuel station 1 would benefit most, with 0.17 percentage savings. Fuel station 5 type D would benefit most, with 0.23 percentage savings.

5 Conclusion

To conclude, we found that for the optimal replenishment frequency and potential improvement, station 7 with fuel type D has the highest potential improvement of 94.46% with the optimal replenishment frequency being approximately 5 (5.24) days. As for the question what day to order, we found from both figure and number results that for both type D and type G Fuel, station 8 has the lowest price on Monday, which is 0.34 and 0.33 respectively. When it comes to tank capacity analysis, we found that station 7 has the largest utilization. When station 5 switches to a 9942.11 gallons tank for fuel type D, it can save the largest percentage with 23.0% savings, which is $182946.86 in gross purchase costs per year. Finally, we found that after applying maximum discount in new strategy compared with the current discount, fuel station 1 would benefit most, with 0.17 percentage savings. For each type of the fuel, type D in station 5 would benefit most, with 0.21 percentage savings.