Utilizing Data Visualization Techniques to Improve a Landscaping Company

Author

Prabin Kumar Shrestha

COMP 6934 - Introduction to Data Visualization

MASc. In Software Engineering
Memorial University of Newfoundland

June 2023

1 Introduction

The purpose of this project report is to showcase how visualizations derived from given datasets can contribute to the improvement of a company. In addition to discussing approaches, data processing, and data visualization techniques, this report also incorporates my personal experiences. It is important to note that this is not an exact submitted report; rather, it is a compilation of selected portions created for my personal use in presenting my work in written form.

While I have accumulated four years of experience in software development, data science has not been a direct part of my background. However, driven by curiosity, I ventured into this field and found it to be an immensely enjoyable and educational experience. I owe a debt of gratitude to my professor, Terrence Tricco, for his teaching techniques, which went beyond theoretical concepts and emphasized practical learning through hands-on exercises. The majority of the content in this report is derived from the professor's notes [1], although explicit references are not made. Additionally, resources such as Stack Overflow for coding assistance, as well as the documentation for Matplotlib and Seaborn, were invaluable in supporting my efforts. I followed the professor's notes for aspects such as data processing, cleaning, exploration, and final touches, including angles, focus, frames, annotations, and colors. Furthermore, I conducted additional research on relevant websites.

The professor provided three datasets from the company, which served as the foundation for the project's overarching goal of improving the company. In order to achieve this goal, it was necessary to identify sub-goals and justify them through visualizations. Two sub-goals were identified, focusing on job types and customer attributes, both contributing to the overall improvement of the company. The datasets provided included $employees.csv$, $landscape.csv$, and $calendar.csv$. These datasets encompassed information on wages, completed job shifts, as well as job details and various customer-related information. The data covered the period from spring to fall in 2022.

Data processing techniques were applied to all three datasets, which involved writing multiple iterations of code, beginning with simple implementations and gradually progressing to more complex ones. The final codes were thoroughly refined. Various techniques and assumptions were made during this process, such as estimating hours per shift to calculate total employee wage expenses. Data exploration played a significant role in the project, with an emphasis on quick visualizations to gain initial insights.

A total of four data visualizations were created, with the final one being interactive in nature. These visualizations effectively demonstrated how exploring and leveraging the concepts can contribute to the company's improvement, while also justifying the sub-goals. The entire coding process was conducted in Jupyter notebooks, and the report includes the code necessary for generating the four data visualizations.

2 Requirements

The project had specific requirements, with the main objective being the improvement of the company as a broad goal. Our task was to identify two specific sub-goals and develop exactly four data visualizations to support both the broad and sub-goals. In compliance with the project requirements, it is necessary to include four specific types of data visualizations: exploratory, explanatory, interactive, and derived. Each of these categories must be represented by at least one visualization in order to effectively support the goals of improving company performance. Moreover, the project required us to incorporate at least three standard types of data visualization. Here are some examples of standard types of data visualization:

- Trends: Line plot
- Correlations: Scatter plot, Bubble plot
- Comparisons: Bar plot, Heat map, Radar plot, Polar plot
- Distributions: Histogram, Box plot, Violin plot, Strip plot, Swarm plot,
- Rug plot, Contour plot
- Part-to-Whole: Pie chart, Treemap, Sunburst
- Hierarchies: Treemap, Sunburst chart, Dendrogram, Venn diagram
- Connections: Sankey diagram, Chord diagram
- Spatial: Choropleth map, Isarithmic map

3 Objective

We were given the broad goal of assisting the landscaping company in improving. We identified two sub-goals within the data to support this broad goal, which are as follows:

1. Identify potential opportunities for growth and development of the company by conducting an analysis of the various service types.
2. Deeper understanding of the customers by analyzing their attributes and using the information to develop more targeted and effective strategies.

4 Dataset

Three datasets were provided for the project, covering the period from Spring 2022 to Fall 2022: $landscaping.csv$, $employees.csv$, and $calendar.csv$. Each dataset offers valuable information that contributes to our analysis and supports the goals of improving the landscaping company.

The $employees.csv$ dataset provides details about employee wages and their corresponding employee ids. This data allows us to evaluate the company's wage expenses, analyze variations in wages among employees, and explore potential correlations between performance and wages.

With a total of 1276 entries, the *landscaping.csv* dataset contains comprehensive information about each job completed during the specified timeframe. It includes essential details such as job id, type, invoice and material costs, request date, start date, completion date, as well as customer information like customer id, type, postal code, and customer rating. This dataset serves as the primary source of information for our project due to its wealth of data.

The *calendar.csv* dataset offers insights into the jobs performed by each employee on specific dates. It follows a format where columns represent employee ids and rows represent dates. By analyzing this dataset, we can evaluate trends in work rates, assess the overall work done by employees, and derive meaningful insights.

In subsequent sections, I will delve into the data processing employed to extract valuable insights from these datasets and achieve our project objectives.

5 Approach

I utilized Python as the primary programming language for data visualization, with the option of using JavaScript (JS) as well. To handle data processing tasks, I relied on the powerful pandas library. While exploring different visualization libraries like matplotlib, seaborn, and plotly, I chose to primarily work with seaborn due to its convenience as a library built on top of matplotlib. I opted for the object-oriented approach in seaborn, as it provides greater flexibility when working with smaller visualization components.

In practical data analysis, the workflow of planning, processing data, creating visualizations, and finalizing them is rarely a linear process. It often involves frequent iterations and back-and-forth. Questions may arise that cannot be answered solely with the available data, and challenges can be encountered in finding suitable visualizations to convey desired insights. Flexibility and adaptability are crucial in navigating these iterative and non-linear aspects of practical data analysis.

Data processing plays a crucial role in data visualization workflows, and having a strong understanding of tools like NumPy and pandas greatly facilitates this process. My background in software development and databases enabled me to adapt to these tools more easily. Visualizing data in the mind, understanding relationships, and determining the best visual representations are key skills in manipulating and transforming data to extract meaningful insights. Before diving into data visualization, I followed the professor's suggestions and focused on thoroughly understanding the data. This involved exploring datasets and examining data from various angles to gain comprehensive insights.

After conducting data exploration, my next step is to extract meaningful insights and craft a compelling narrative based on the data. I begin by asking ”why” questions to uncover relationships and patterns, such as why certain types of jobs are less frequent or why job ratings tend to be lower. While these questions can be addressed through data analysis, I find that quick data visualizations like bar plots, line graphs, or pie charts help to visualize and understand patterns more easily. Visualizations provide concise and intuitive representations, enabling efficient identification of trends and connections.

The data analysis process involves iterative exploration, data processing, visualization, questioning, and revisiting the data. This iterative approach allows for deeper insights and a better understanding of underlying patterns and relationships. Sometimes, questions cannot be fully answered with the available data, leading to a temporary halt in the analysis. On the other hand, effective visualizations can serve as a basis for further analysis and storytelling if they communicate the desired message and insights.

Merging different visualizations, either in their entirety or selectively, is a valuable technique. By combining visualizations, a comprehensive view of the data is obtained, facilitating the discovery of complex relationships and patterns that may not be apparent when examining individual visualizations. It's important to note that the focus, until this stage, is primarily on the data analysis process itself, with final touches such as angles, focus, framing, colors, text, annotations, and other visual elements being considered in later stages when refining and presenting the analysis.

6 Data Preparation and Processing

Data preparation and processing result from frequent data exploration, involving data visualization to enhance its quality. Throughout the visualization finalization process, I had to frequently modify the data preparation and processing code, such as categorizing the data for color. Ultimately, it is crucial to optimize and clean the code as well.

The files *landscaping.csv*, *calendar.csv*, and *employees.csv* are converted into dataframes using pandas' *read_csv* function. They are assigned to the variables *df_landscaping*, *df_calendar*, and *df_employee*, respectively. The request_date and completion_date columns of *df_landscaping* are parsed since we require them for future date processing during data visualization.

In *df_landscaping*, I have added two new fields: completion_date_week_number and request_date_week_numb which represent the respective week numbers. Additionally, I calculated the net_profit by subtracting the material_costs from the invoice_amount. I performed these calculations at the beginning because I will be using the net_profit multiple times in subsequent data visualizations.

The *calendar.csv* data was not in a tidy format, so I applied the *melt* method in pandas to transform it into a tidy structure. This conversion allowed me to create a dataframe *df_job_employee_link* that consisted of columns for date, employee id, and job id. This tidy format was easier to work with and provided a clearer representation of the data. It was also important to handle any missing values or duplicates in the dataset by using methods such as *dropna* or *drop_duplicates*.

In *df_employee*, the wage for each employee is specified. Assuming an average shift hour of 4 hours, I calculated the total wage per shift for each employee. Since this information was not provided, I made this assumption. I then merged this dataframe with the *df_job_employee_link* so that each job shift includes the employee wages. This allows me to consider the employee wages when calculating the profit, in addition to the invoice_amount and material_costs.

7 Data Visualizations

In this section, I have presented four data visualizations along with their corresponding explanations. The code to develop these visualizations is attached separately.

7.0.1 Trends in Invoice Amount and Requested Jobs through Spring to Fall 2022

The data visualization, Figure 1, depicts the trends and patterns in the total invoice amount and the number of jobs requested on a weekly basis from March to November, covering week numbers 9 to 44. The visualization utilizes a bar graph to show the total number of jobs requested per week and a line graph to represent the total invoice amounts per week.

The graph displays a peak in the number of job requests during week 10, followed by a pattern that starts from the bottom and gradually increases from week 12 to week 27, before declining until week 44. Surprisingly, the total invoice amount appears to remain relatively constant throughout the entire period, averaging around $35,000 per week, despite the fluctuating frequency of requested jobs. The consistency in the total invoice amount can be attributed to the varying invoice amount of different job types.

A deeper analysis of the reasons for the fluctuation in job requests can provide valuable insights into the company's operations. To increase profitability, the company can conduct a thorough analysis of the underlying causes for the decline in job requests during the period of low job requests and focus on increasing the job count by implementing effective strategies. By closely monitoring and analyzing the trends and patterns in job requests and invoice amounts on a weekly basis, businesses can improve their revenue and overall performance.

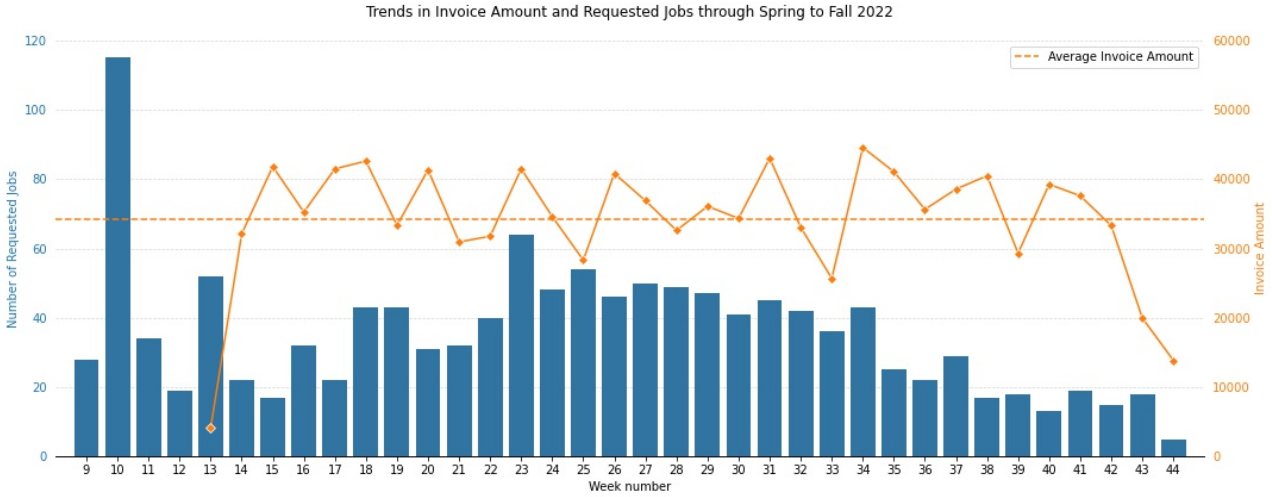


Figure 1: Trends in Invoice Amount and Requested Jobs through Spring to Fall 2022

7.0.2 The Relationship between Invoice Amount and Profit Percentage

The scatter plot, Figure 2, illustrates the correlation between the invoice amount and profit percentage of jobs, classified by customer type (commercial or residential). The profit percentage is obtained by dividing the difference between the invoice amount and material cost by the material cost.

Upon analyzing the plot, it is apparent that the density of data points is the highest in the top left corner of the graph. This area indicates a cluster of jobs with low invoice amounts but high profit percentages. In contrast, a gap is evident in the middle of the plot, with a lighter concentration of data points towards the bottom right side, indicating a smaller number of jobs with higher invoice amounts and lower profit percentages.

These observations suggest that the company has a larger number of jobs with lower invoice amounts and a smaller number of jobs with higher invoice amounts. Moreover, jobs with lower invoice amounts tend to have higher profit percentages, whereas those with higher invoice amounts tend to have lower profit percentages.

Furthermore, the density of commercial customer types is slightly higher than that of residential customer types, implying that commercial customers generate higher profit percentages. However, the number of data points in the commercial customer category is lower than that of the residential customer category.

Therefore, the company can benefit from investigating the reasons for the lower number of jobs in the higher invoice amount group and increase the frequency of that group. Additionally, focusing on increasing the number of commercial customers may prove advantageous.

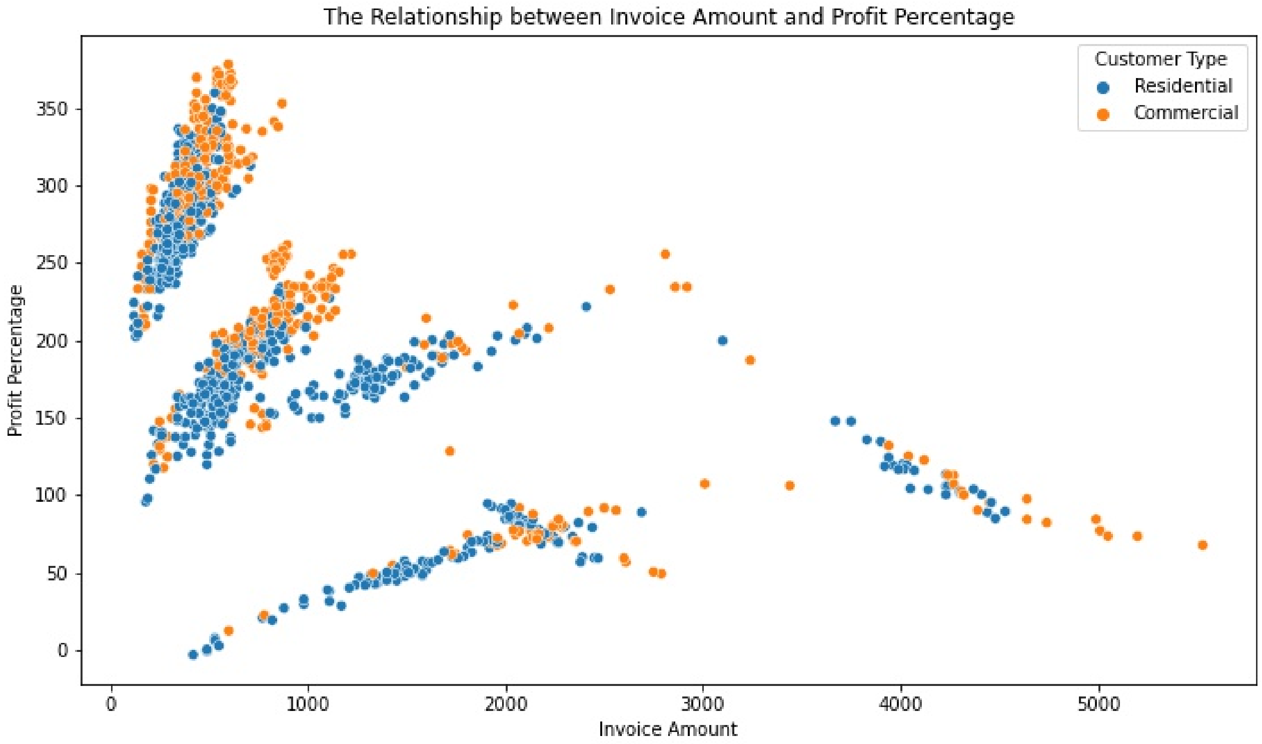
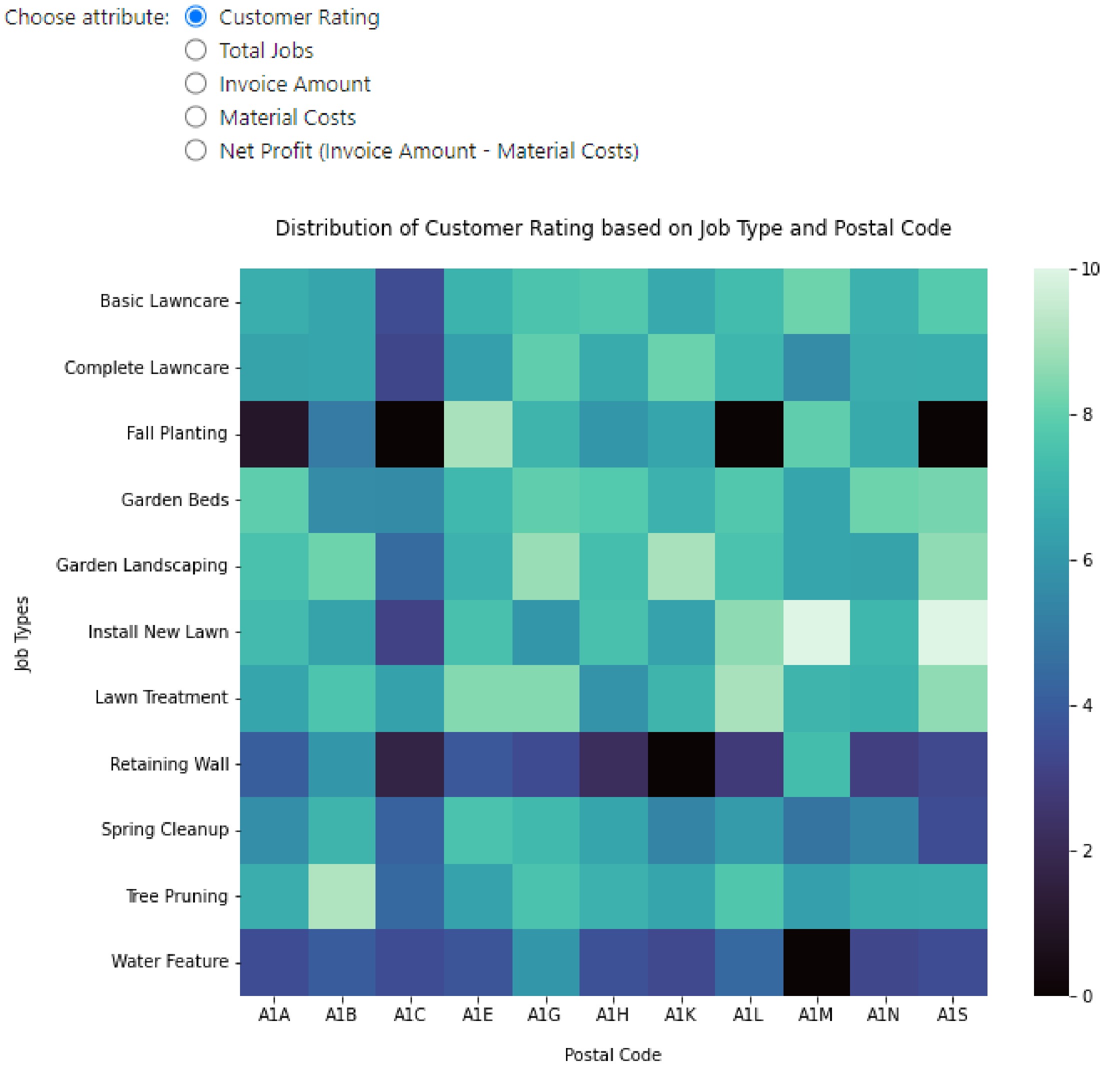


Figure 2: The Relationship between Invoice Amount and Profit Percentage

7.0.3 Distribution of Different Attributes based on Job Type and Postal Code

The main purpose of the visualization, Figure 3, is to offer valuable insights into the complex relationship between postal codes and job types across various parameters. The visualization is interactive and comprises a set of radio buttons that allow users to select different parameters, such as customer satisfaction, total jobs, invoice amount, and material costs. Upon selecting any of these parameters, a corresponding heatmap is generated, which showcases the relationship between postal codes and job types based on the chosen parameter.



Choose attribute:
● Customer Rating
○ Total Jobs
○ Invoice Amount
○ Material Costs
○ Net Profit (Invoice Amount - Material Costs)

Distribution of Customer Rating based on Job Type and Postal Code

Figure 3: Distribution of Different Attributes based on Job Type and Postal Code

The interactive features of this visualization provide viewers with the opportunity to explore further. Two inquiries that I anticipate are as following.

1. The distribution patterns across job types and postal codes for different attributes, such as customer satisfaction. Viewers can explore the heatmap to identify areas where specific job types

or postal codes consistently exhibit higher or lower levels of satisfaction. Further exploration can be done for other parameters as well.

2. The possible absence of services in a specific area can be determined by examining the visualization, which displays black squares indicating the absence of certain services in that location. By selecting the ”total jobs” parameter, it is possible to identify areas where there is a high or low demand for a specific job type.

In this particular viusalization, the distribution of customer ratings is shown. However, when executed in a Jupyter notebook, this is an interactive diagram where different attributes can be selected. For example, if the net profit attribute is chosen, it will display the distribution of net profit across postal codes.

7.0.4 Water Feature and Retaining Wall: Less Frequent and Rated Job Types, Yet Highly Profitable

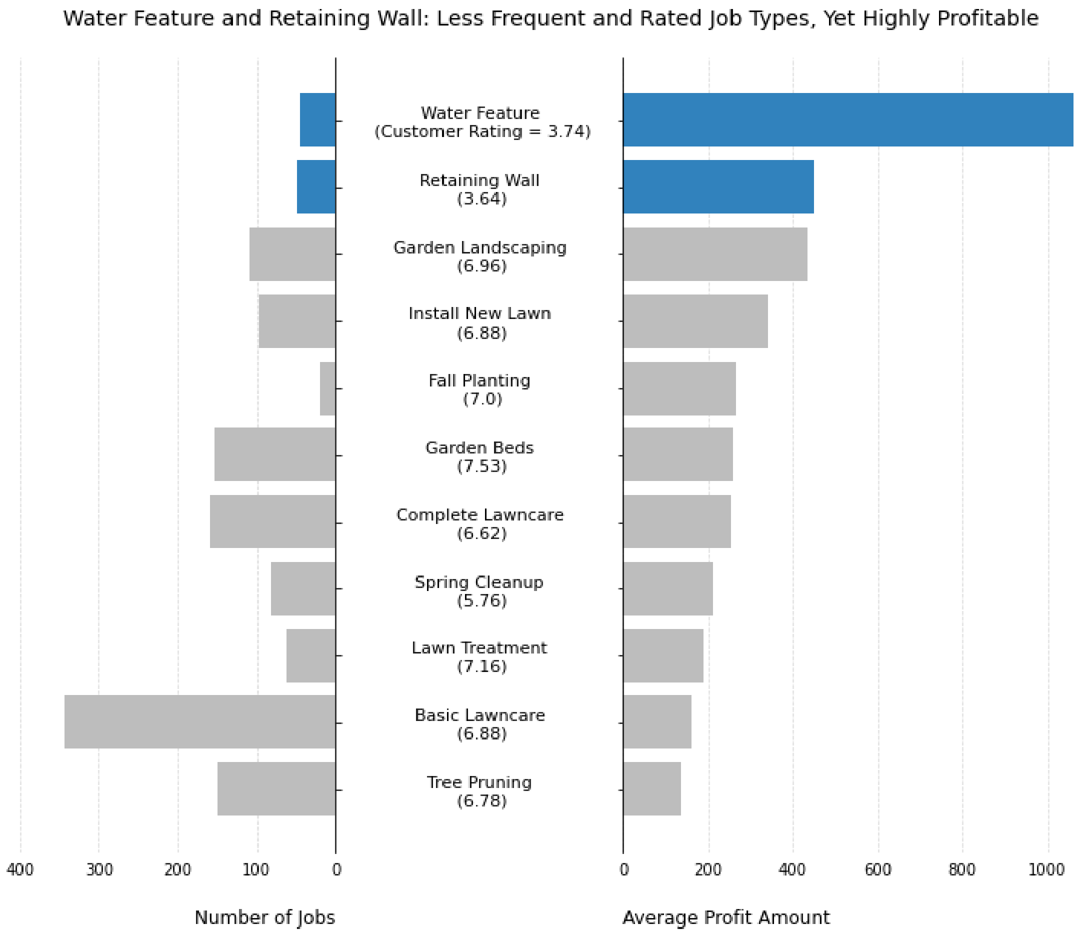


Figure 4: Water Feature and Retaining Wall: Less Frequent and Rated Job Types, Yet Highly Profitable

The presented data visualization, Figure 4, offers a clear and concise representation of the profitability and frequency of various job types. By utilizing a bar chart, viewers can easily compare and contrast the different job types based on their profitability and frequency, with profitability being calculated based on invoice amount, material costs, and employee wages, assuming 4 hours per shift.

Interestingly, the two most profitable job types, Water Feature and Retaining Wall, are also the least rated in terms of customer satisfaction, and among the least frequently requested job types. This leads us to question whether the lower frequency of these job types is due to customer dissatisfaction and impacting the overall potential profit.

One potential solution to this issue could be to focus on improving customer satisfaction and increasing the number of jobs in these less popular but profitable job types. This has the potential to significantly increase profits for the company.

8 Conclusion

In conclusion, this project report has demonstrated the potential for utilizing data visualizations to drive improvements within a company. Through the application of data processing techniques and the creation of insightful visual representations, I have successfully addressed the main goal of enhancing the company. By identifying and justifying sub-goals related to job types and customer attributes, I have showcased how visualizations can contribute to overall organizational growth.

The project journey, despite my initial lack of direct experience in data science, proved to be a valuable learning experience. Thanks to the practical and hands-on teaching approach of my professor, supplemented by resources like Stack Overflow and the documentation of visualization libraries, I was able to navigate through the data processing and visualization stages successfully. The creation of four visualizations, including an interactive one, highlighted the importance of data exploration and visualization in supporting decision-making processes and achieving the company's objectives.

Overall, this project not only allowed me to gain valuable insights into the field of data science but also provided an opportunity to apply these skills in a real-world context. Moving forward, I believe that the integration of data visualization techniques will continue to play a pivotal role in empowering businesses to make data-driven decisions and drive sustainable growth.

References

[1] T. Tricco, "Introduction to data visualization," Winter 2023.