*A new analytics program we developed as a change agent for DoorDash using the CRISP-DM process.*

**Assignment**

**6**

A6

ALY6120 Leadership In Analytics

Assignment 6 – Signature Assessment

**PREPERATION:**

By: John DiSessa

For: Professor Dooley

On: August 20th, 2022

Introduction

Based on my previous analyses, I was selected to lead DoorDash’s new Corporate Strategy Analytics Department. DoorDash’s leadership was impressed with my thorough analysis despite limited data and decided my skills as a leader will guide their new team to the forefront of analytical thinking. As part of my interview process, we discussed my leadership strengths and weaknesses. First and foremost, I am a data-driven decision maker. I believe that decisions made without data are subject to potential biases and can often be no better than simple guessing. Data-driven decision making allows for many areas of interest to be scaled, quantified, and projected which can give corporate leadership the tools they need to budget and allocate resources. As someone who values everyone’s opinions and perspectives, my inclusivity is a skill that DoorDash was looking for since their goal was to build a competent and diverse Corporate Strategy Analytics Department. I find value in everyone and my personality can bring out the best in all of my team members. One condition of my hire was that I needed to improve my patience. Patience is a weakness of mine both personally and as a leader. Because I lack patience, I would struggle teaching my teammates skills or processes if they are not quick learners. DoorDash was willing to hire top talent in order to reduce the amount of teaching I would need to do, but there will still be many team members that need to learn from me. DoorDash has given my department the time it needs to work properly as long as I improve my patience. DoorDash executives and I have agreed to a timeline and work pattern, via Gantt Chart, in order to set work quality and work timing expectations. We also agreed to use the CRISP-DM process to guide our work due to its industry-wide acceptance, flexibility, and thoroughness. Our new team will consist of top talent from other analytics departments within DoorDash as well as outside hires with extensive leadership and analytical expertise. We will have data engineers working on the database and data analysts processing the data to gain valuable insights.

Understanding the Business Problem

**Organizational Objectives**

The first step in the CRISP-DM process is to establish a clear understanding of the business problem. DoorDash faces intense competition in the mobile food-delivery business with multi-million-dollar competitors in Grubhub, Uber Eats, Go Puff, and Postmates. Our DASH stock is down 58% from the IPO in December of 2020, down 60% over the last year, and down 33% over the last 6 months. Our net income is down 51% over the last year despite a 35% increase in revenue. Profitability is our main concern right now.

In order to clearly and unambiguously identify objectives that can improve their business problem, the objectives need to be SMART (specific, measurable, attainable, realistic, and timely). The SMART acronym is usually associated with goal setting, but can apply to the determination of DoorDash’s objectives in order to ensure that we clearly state what we need to analyze and how we can be sure our analysis accurately applies to our objectives.

Our 3 objectives over the next year for the DoorDash CRISP-DM process are:

* Increase the average customer rating
* Decrease the average delivery time
* Identify if DashPass increases customer monthly order frequency

Customers rate each order they receive on a scale from 1 to 5 with 5 being the best. This feature allows us to clearly see if the customer experience improves or not. Delivery time is also an essential function to the customer experience. The DoorDash app tracks the time from order-placed to order-delivered. Again, we will be able to clearly see if the average delivery time decreases or not. Lastly, the DashPass is an essential offering to entice customers to stay. Since there are essentially no switching costs for customers to use competitors’ services, the DashPass incentivizes customers to use DoorDash by offering free delivery and various discounts. If DoorDash can clearly find a correlation between DashPass usage and an increase in customer order frequency, we can be confident customers use DoorDash over competitors’ services.

**Leadership Requirements**

In order to effectively utilize and implement an analytics solution, my leadership will have these 3 requirements in place:

* Empowered Employees
* Team Diversity
* Patience

Empowering employees by giving them the freedom to use resources, explore creative solutions, and make decisions is essential to motivating them. Employees are more likely to view their role as a career and not just as a job. They will have more positive attitudes and thus be more productive. Diversity is also essential to getting the most out of a team. A purely analytical approach will not suffice since DoorDash is a people business at its core. We need a team with very different perspectives since DoorDash is used by a diverse customer base and for a slew of various reasons. People use DoorDash for business lunches, personal meals, snacks, late-night, or other reasons. Lastly, I must display patience. I need patience to teach employees the technical and analytical skills, I need patience to try various theories and wait for concrete results, and I need to have patience to allow time for my employees to implement their own ideas. In such a competitive landscape, DoorDash cannot rush their decisions and risk million-dollar mistakes.

Understanding the Data

I got hired by DoorDash based on my independent analysis on their profitability. They found my analysis interesting and believe I can achieve even greater results when given access to their internal database. One dataset I first used focused on restaurant reviews and the other dataset focused on delivery reviews. I considered combining datasets to simplify the data understanding, however the datapoints were too different. I felt like I needed both types of data for a comprehensive analysis. Both of the data sources were externally acquired on Kaggle, however, it is unclear where my sources on Kaggle originally acquired the data. I assumed their acquisition to be external as well to avoid assuming that the datasets are completely reliable.

**Restaurant Dataset**

The first dataset I chose contains restaurant data for over 3,200 of Canada’s DoorDash restaurant partners in Canada’s 8 largest cities. First, each restaurant contains an average customer rating as well as the total number of reviews each restaurant has. It is possible that more reviews are byproducts of more orders. More orders could be byproducts of desirable restaurants. Second, each restaurant contains an average delivery distance for all their customer’s orders. Delivery distance is likely correlated with delivery time so that could give us some initial insights. It can also indicate how well DoorDash manages their driver fleet. If delivery distance is low, that could indicate a high geographical customer concentration, it could indicate DoorDash uses many drivers to increase the likelihood that drivers are available nearby restaurants at the time of orders, or it could indicate that DoorDash can better predict customer demand and relocate their drivers more efficiently. Lastly, it is possible assess customers’ attitudes on price with a price rating category or even identify city popularity by comparing the 8 Canadian cities to each other.

**Delivery Dataset**

The second dataset I chose contains data on 480 individual deliveries. This dataset contained crucial variables needed to assess delivery times and customer ratings. This dataset primarily focused on DoorDash’s driver utilization and customer experience by showing how delivery and wait times can affect a customer’s rating. By assuming that a customer’s rating is indicative of their experience, we can pinpoint which steps in DoorDash’s service the customers value the most. We have time logs for dasher to restaurant time, dasher wait time, and restaurant to customer time as well as cumulative times, on-time ratings, and delivery methods.

Data Preparation

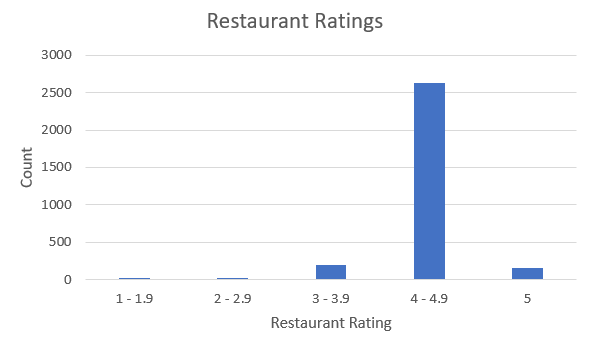
I assessed the quality of each data set by fixing typos, consolidating data types, removing blank entries, and finding outliers. In each dataset, there were several typos such as restaurant names with special characters and ratings spelled out alphabetically (5 = five). I removed the incorrect restaurant names and updated all the ratings to be on a 1-5 scale. The price rating was listed as $, $$, $$$, or $$$$ so I also converted these to a 1-4 scale. I then removed any entry with a blank rating since there were thousands of valid data points and did not need to corrupt our data for the sake of sample size. Lastly, I tried to identify any outliers. Since most of the ratings were on a scale from 1-5, these results were expected and not removed. The delivery distance category was not on a consistent scale so I created a boxplot to see if any deliveries were extremely far and misrepresented the average delivery distance. As you can see from the table below, all of the restaurant’s average delivery distances ranged from .01 miles to 30 miles. Because 75% of the deliveries were 1.6 miles or less, longer delivery distances could have been considered outliers. We also calculated a standard deviation of just over 9 miles. With a mean delivery distance of 5 miles, all of the restaurants are within 3 standard deviations of the mean. Since the mean is significantly greater than the median, delivery distance is right-skewed. The longer delivery distances skew the mean, considering half of the deliveries are 1 mile or less. However, contrary to these statistics, I decided not to remove any far deliveries from the dataset. 15% of the entries were 20 miles or further. That is too significant of a sample to remove. Also, these are valid entries and far deliveries will always be an obstacle that DoorDash will have to deal with.

|  |  |
| --- | --- |
| **BoxPlot Delivery Distance** | |
| Min | 0.01 |
| 1Q | 0.5 |
| Median | 1 |
| 3Q | 1.6 |
| Max | 30 |
| IQR | 1.1 |
| Upper Fence | 3.25 |
| Lower Fence | 0 |

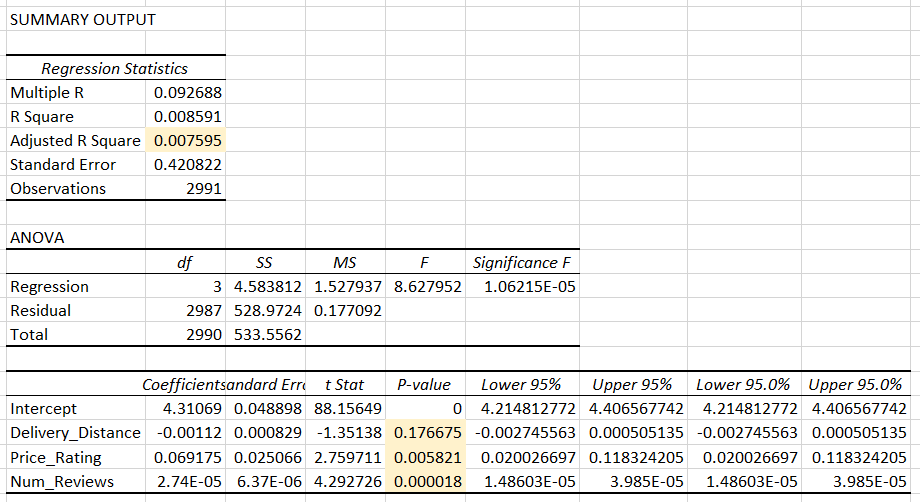
Modeling & Evaluation

**Restaurant Analysis**

I started by creating a bar graph of all the restaurant ratings in our dataset. Since the ratings were on a scale of 1-5, I created 5 bins grouping all the ratings within those bins. 88% of the restaurants in our dataset were rated between 4 and 4.9. There were hardly any restaurants below 3 because, if restaurants continued to rate that poorly, DoorDash would have removed them as a partner. There were also very few restaurants with perfect 5-star ratings, but future analysis could look into the characteristics of these 5-star restaurants to see any overlapping similarities. Any commonalities could be used as teaching material for other restaurants to improve their ratings. The main focus would consist of improving restaurants in the low 4’s to get them into the high 4’s. Future analysis could break the 4’s group into 4 – 4.5 and 4.5 – 5 to get a better idea of the smaller differences in restaurant quality.

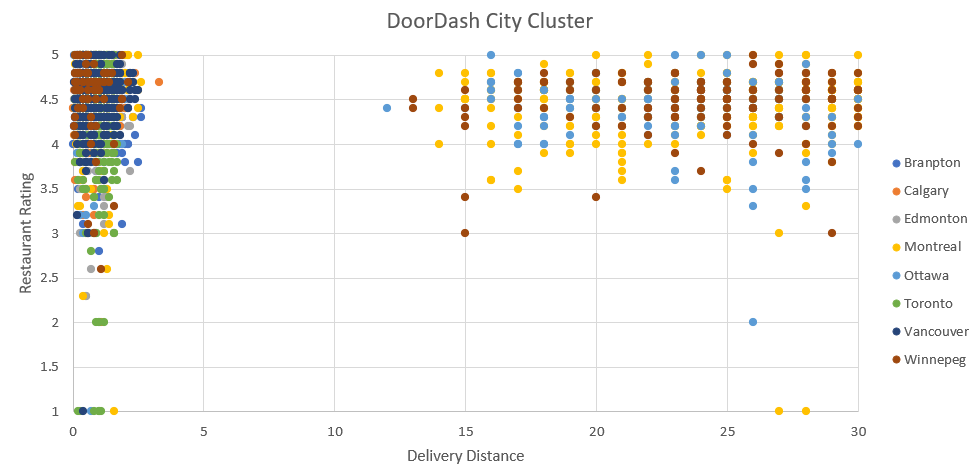


I then looked at delivery distance, number of reviews, and price rating to see if any of those variables were good predictors of restaurant rating. I conducted 3 different regressions for each variable to try and predict restaurant rating. As seen below, all 3 variables are actually insignificant predictors of how people rate the restaurants.



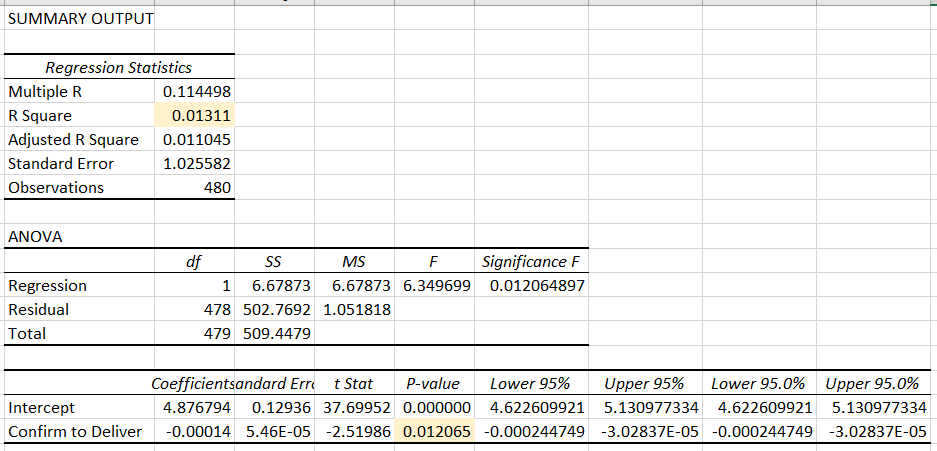
Using all 3 variables, the model was only able to explain less than 1% of the variation in restaurant rating. In fact, delivery distance was statistically insignificant since it’s p-value is greater than our alpha of .05. I also ran regression for all 3 variables independently and achieved similar results. While these regression findings do not warrant further analysis, I still learned that these variables are not good predictors of customer restaurant rating and ignoring them, or monitoring them very loosely, is a better strategy than wasting resources trying to monitor them closely.

Lastly, I used clustering to identify any noticeable differences in restaurant rating. I graphed all restaurant ratings by city and delivery distance since that variable had the most variance. As you can see from the chart below, the data can be broken down into two clusters: deliveries over 11 miles and deliveries under 5 miles. Some cities, like Toronto and Branpton, had no restaurants that consistently delivered that far. For restaurants with average deliveries over 11 miles, there were very few ratings less than 3.5. However, for deliveries under 5 miles, there were significantly more ratings under 3.5. It is clear that future analysis needs to analyze these restaurants in two groups and assess their qualities independently. Only after identifying any differences can city differences be accurately assessed.

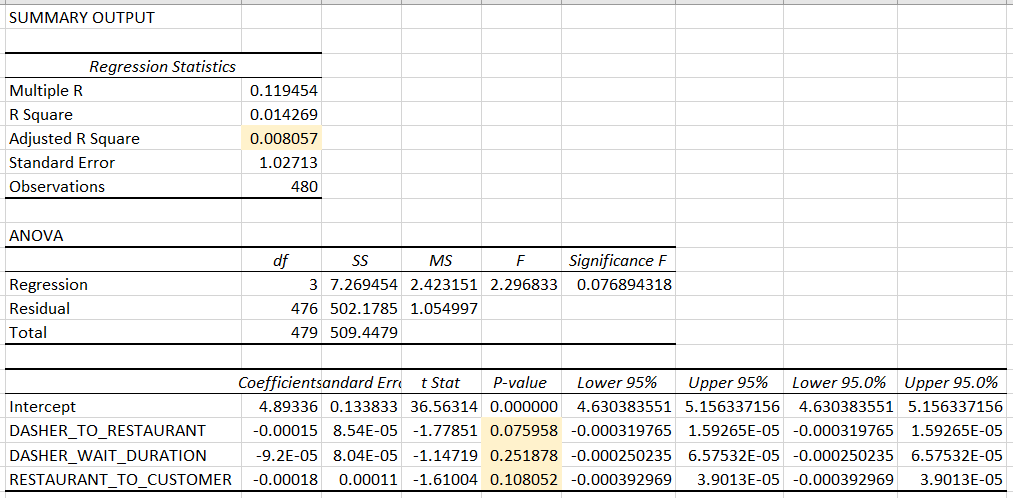


**Delivery Analysis**

I conducted two regression analyses to analyze actual delivery times and their effects on delivery rating. I figured customers care more about the delivery time than they do about the delivery distance. My first regression looked at the cumulative confirm to order delivery time. Surprisingly, this variable only accounted 1.3% of the variation in delivery rating. While it was statistically significant, confirm to deliver time by itself, is not a good predictor of delivery rating.



Rather than looking at the cumulative delivery time, I analyzed each stage of the delivery independently to see if any one stage in particular affected delivery rating. The following regression used dasher to restaurant time, dasher wait time, and restaurant to customer (final delivery) time. All 3 variables are insignificant with p-values over .05 and the model as a whole is a bad predictor of delivery rating.



Despite insignificant regression results, I still identified dasher\_wait\_duration as the variable DoorDash should be primarily concerned about. Dasher\_wait\_duration had the largest coefficient of variation, which looked at the standard deviations of each variable relative to the mean of each variable. A high dasher\_wait\_duration indicates an inefficiency in DoorDash’s algorithm for assigning drivers to orders. Their algorithm miscalculated how long it would take a restaurant to prepare an order and how long it would take the driver to arrive. Drivers waiting for orders also makes the driver experience terrible because drivers do not get paid for the time they wait. As well as increasing the time it takes for a customer to receive their order, this variable affects the driver experience as well and can affect DoorDash’s overall driver availability.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Dasher\_To\_Restaurant** | **Dasher\_Wait\_Duration** | **Restaurant\_To\_Customer** |
| Mean | 768.36 | 606.17 | 835.14 |
| Standard Deviation | 563.51 | 597.26 | 432.46 |
| Coefficient of Variation | .73 | .99 | .52 |

Deployment

My restaurant and delivery analysis yielded disappointing results. All of my regressions were insignificant which surprised me at first. However, my analysis did not include any food quality ratings. Customers enjoying their food is perhaps the biggest factor in ratings so my results should not have been surprising. While clustering provided me some concrete insights for future analysis, it is clear I need to dig deeper with the data I have as well as find some new datapoints. Future analysis could combine restaurant quality, driver quality, delivery time, app functionality, or restaurant offerings. DoorDash could add some customer surveys or have a quick push notification after each customer rating to ask them why they rated the way they did. Despite not being ready to deploy any models, I laid the foundation for my new analytics team at DoorDash to follow once I get my hands on their internal dataset. Even though I followed the CRISP-DM process, it is clear I need to restart the process a few more times with new datasets and new analysis. As the new analytical leader for the Corporate Strategy Analytics Department, my team will conduct rigorous testing at each stage of the cycle. We will have a 3-point sign-off process for each stage. Before moving onto the next stage in the CRISP-DM cycle, we need approval from me, the lead data analyst, and the lead data engineer. All 3 of us will also have to agree whether to move on to the next step or restart the cycle if given new information or insights. My team will be open to this because of my constant communication with DoorDash executives to plan for any corporate-directive changes. By maintaining a strong level of communication, flexibility, and patience, my team will have DoorDash operating optimally in no time.

Summary

Even though I was not able to gain many concrete insights as an independent researcher, I feel confident in following the CRISP-DM process. Restarting the cycle with new information is part of the process and necessary for delivering thorough and insightful results. My research proved to DoorDash executives that I am the best analytical leader to run their Corporate Strategy Analytics Department. I am data-driven, highly motivated, detail orientated, and inclusive all as evidenced by my research above. DoorDash values my skills and is willing to work with me on my patience and delivery. I have proven my ability to learn, adapt, and improve over time and I will improve my weaknesses in order to become a well-rounded leader. If I follow sound leadership and analytical principles, I will set the gold-standard for how analytics departments should be operated.

Citations

“Adjusted R Squared.” *IBM*, https://www.ibm.com/docs/en/cognos-analytics/11.1.0?topic=terms-adjusted-r-squared.

*Calculating and Displaying Regression Statistics in Excel*. https://www.rwu.edu/sites/default/files/downloads/fcas/mns/calculating\_and\_displaying\_regression\_statistics\_in\_excel.pdf.

“Excel Regression Analysis Output Explained.” *Statistics How To*, 3 June 2021, https://www.statisticshowto.com/probability-and-statistics/excel-statistics/excel-regression-analysis-output-explained/.

Hotz, Nick. “What Is CRISP DM?” *Data Science Process Alliance*, 16 Apr. 2022, https://www.datascience-pm.com/crisp-dm-2/.

Bhandari, Pritha. “How to Find and Remove Outliers.” *Scribbr*, 20 May 2022, https://www.scribbr.com/statistics/outliers/.

Dalgleish, Debra. “How to Build an Excel Box Plot Chart.” *How to Make an Excel Box Plot Chart*, https://www.contextures.com/excelboxplotchart.html.

Ihewitt. “Doordash.” *Kaggle*, 9 May 2022, https://www.kaggle.com/datasets/isabellahewitt/doordash.

Satoshi\_S. “Food Delivery across Canada (Door Dash).” *Kaggle*, 3 June 2022, https://www.kaggle.com/datasets/satoshiss/food-delivery-in-canada-door-dash?resource=download.