



Doordash Data Science Project



By Pranav Kumarsubha



01

Intro and Background

Why I chose Doordash?



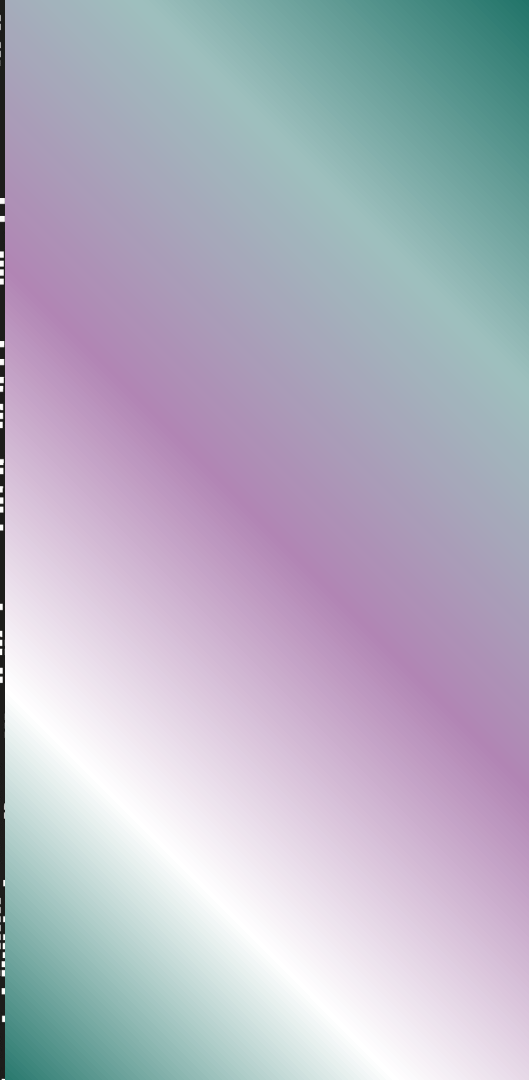


Introduction

Having just completed the “Python for Machine Learning & Data Science Masterclass” course on Udemy, where I improved on and learned topics such as pandas, data visualization, machine learning models, and NLP, I knew the next step to becoming a complete data scientist was to immerse myself in developing projects and my models. I scoured the internet for ideas and concepts I could immediately start working on. Eventually, I decided to begin a project about something I had experience in, but from a different angle.

Why Doordash?

I began driving for Doordash in April of 2023 during college. I was using it to make income to help support myself outside of class. It was a stressful job at times, driving so many hours and keeping focused throughout, but I started to use it as my primary income source over the summer season when I went back home. That is why when I came across the Doordash take-home case study project online, I knew it would be a great way to start diving into the project atmosphere because of the background information I possessed.



Purpose

My first initial goal was to mainly test and analyze the different variables of Doordash and how they are correlated from both the consumer side and the driver side. I wanted to analyze how potentially the total delivery time affected the tip total or for example, how the time the order was placed affected how fast the order arrived.

Questions like these were what I focused on while creating the various charts and graphs in my project. After the analysis, my final idea was to build my own machine learning model to predict the delivery time. I believed utilizing my new machine learning knowledge would help me make a successful predictor model.

The Dataset

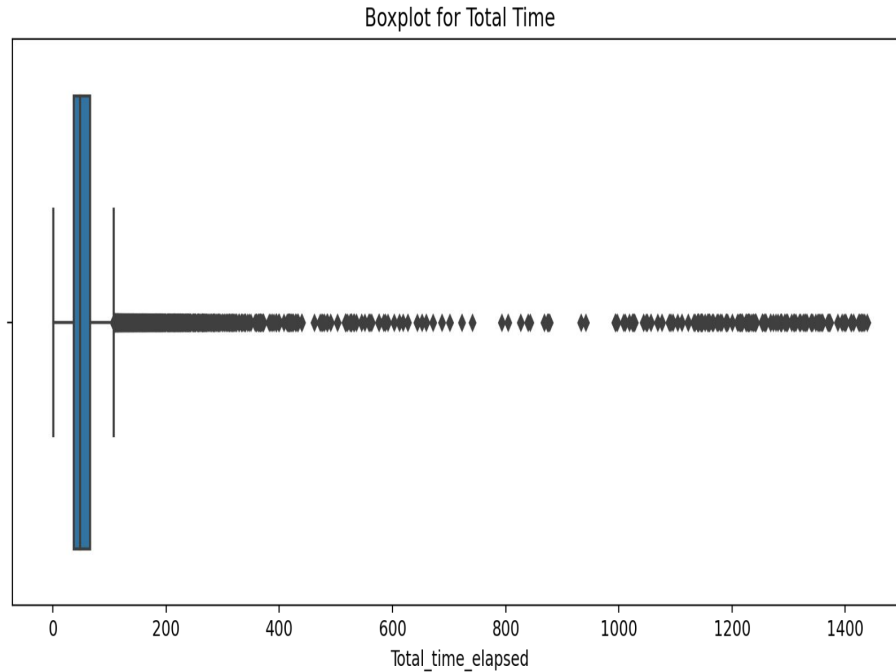
I found this Doordash dataset while doing my research on their case study project. The original dataset came with 16 columns and 18078 entries.

Unnamed: 0

	Customer_placed_order_datetime	Placed_order_with_restaurant_datetime	Driver_at_restaurant_datetime	Delivered_to_consumer_datetime	Driver_ID
0	14 20:27:45	14 20:29:41	14 20:39:32	14 20:52:03	86
1	07 20:16:28	07 20:17:32	07 20:36:00	07 20:49:02	325
2	13 19:35:09	13 19:39:26	13 20:28:16	13 20:52:44	200
3	22 19:47:53	22 19:56:08	22 20:01:20	22 20:18:01	154
4	03 19:01:52	03 19:09:08	03 19:36:20	03 19:45:26	332

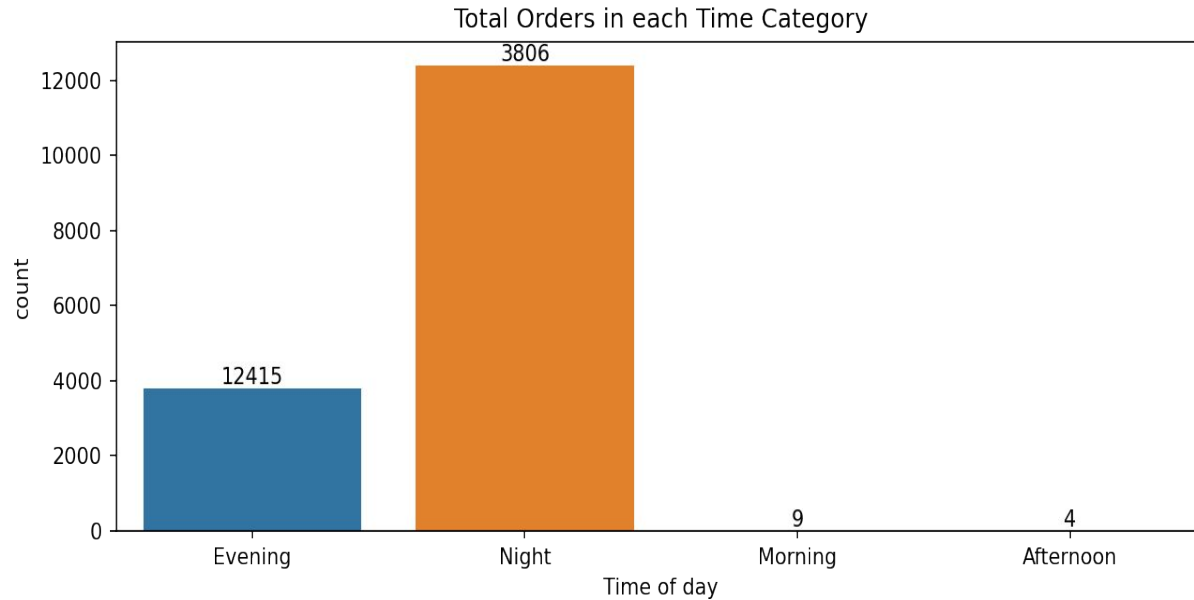
Restaurant_ID	Consumer_ID	Is_New	Delivery_Region	Is_ASAP	Order_total	Amount_of_discount	Amount_of_tip	Refunded_amount	Total_time_elapsed
12	5	False	Palo Alto	True	20.45	20.45	3.07	0.0	0 days 00:24:18.000000000
66	5	False	Palo Alto	True	40.62	40.62	3.73	0.0	0 days 00:32:34.000000000
124	5	False	Palo Alto	True	37.78	37.78	1.89	0.0	0 days 01:17:35.000000000
5	14	False	Palo Alto	True	39.66	0.00	1.98	0.0	0 days 00:30:08.000000000
9	14	False	Palo Alto	True	39.66	0.00	5.95	0.0	0 days

Total Time Boxplot



At first glance I was shocked to see how many outliers there were in this dataset. Only then did I go back and noticed how some rows for example said the order was delivered a day later which is practically impossible. I knew this data would later ruin much of my analysis and especially my model so I decided to eliminate the rows part of the 10% highest delivery times to eradicate many of the outliers existent.

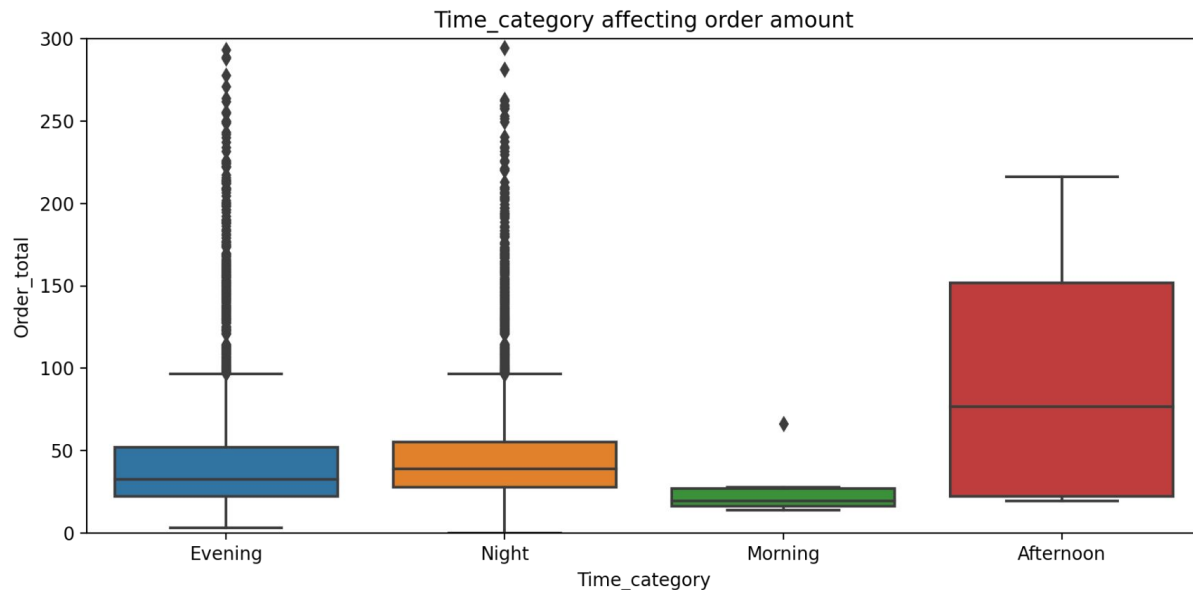
Orders by Time Category



This barplot was made just so I could quickly analyze which parts of the day where most of the orders made and it's clear that the Night orders far exceed any other category while morning and afternoon orders in these regions were scarce surprisingly.



Order Amount by Time Category

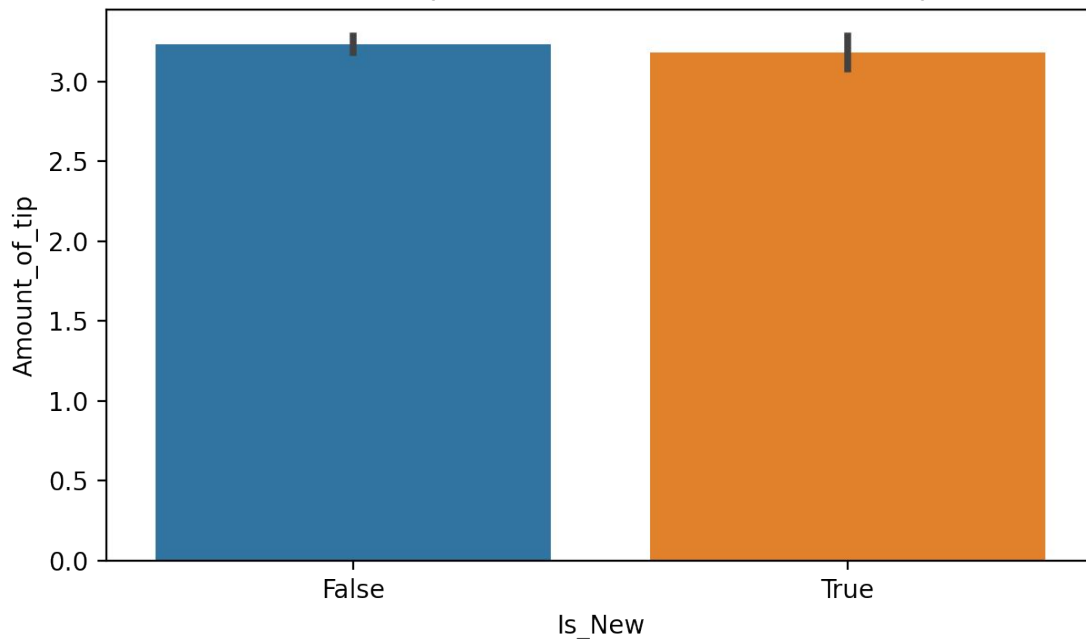


The most interesting part I discovered from this boxplot was that there were so many outliers in this data set. It also was a little hard to analyze the different time categories just because of how few data there was in the morning and afternoon categories respectively.

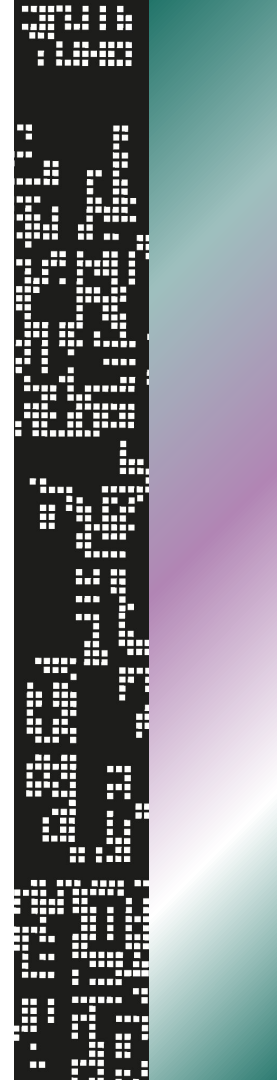


Tip and Driver Experience

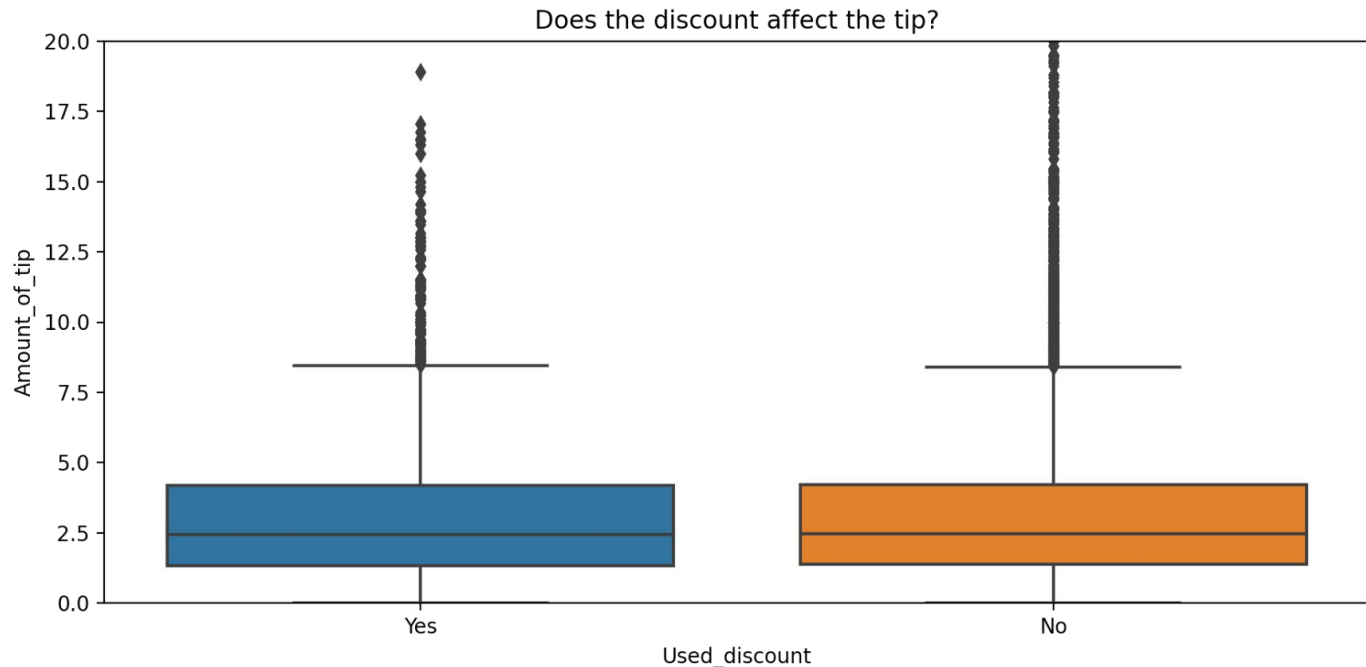
Does the experience of a Driver affect the tip?



When I initially thought to measure the experience of a driver with the tip amount I had expected there to be a big difference because as a driver myself I knew that those who do more orders get higher paying ones with higher tips as well but I was very surprised to find that there was little to no difference in tip versus experience.

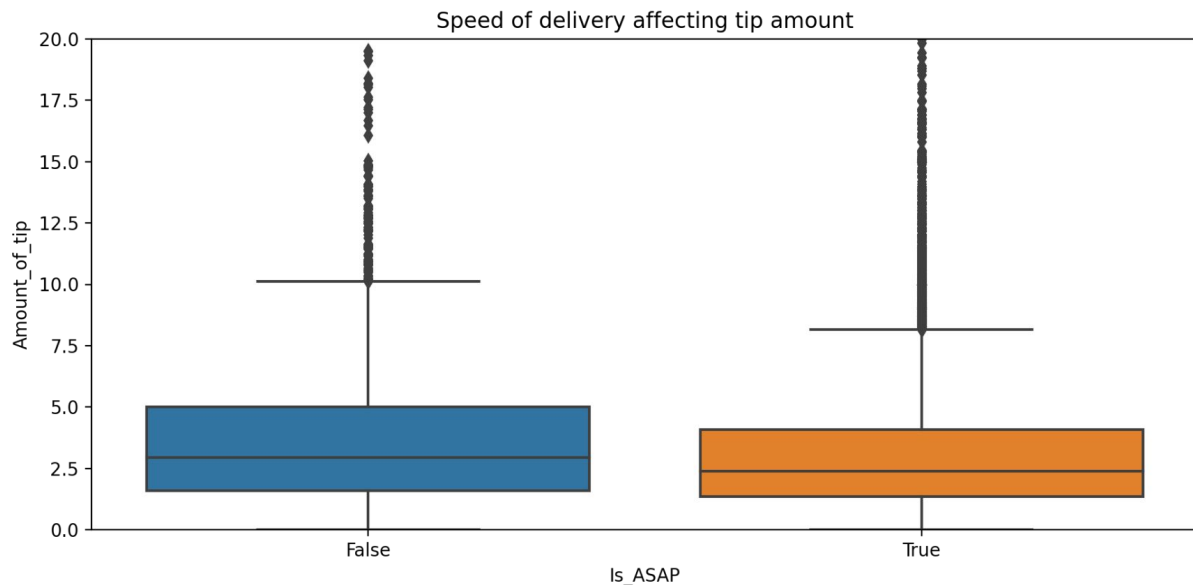


Tip and Discount



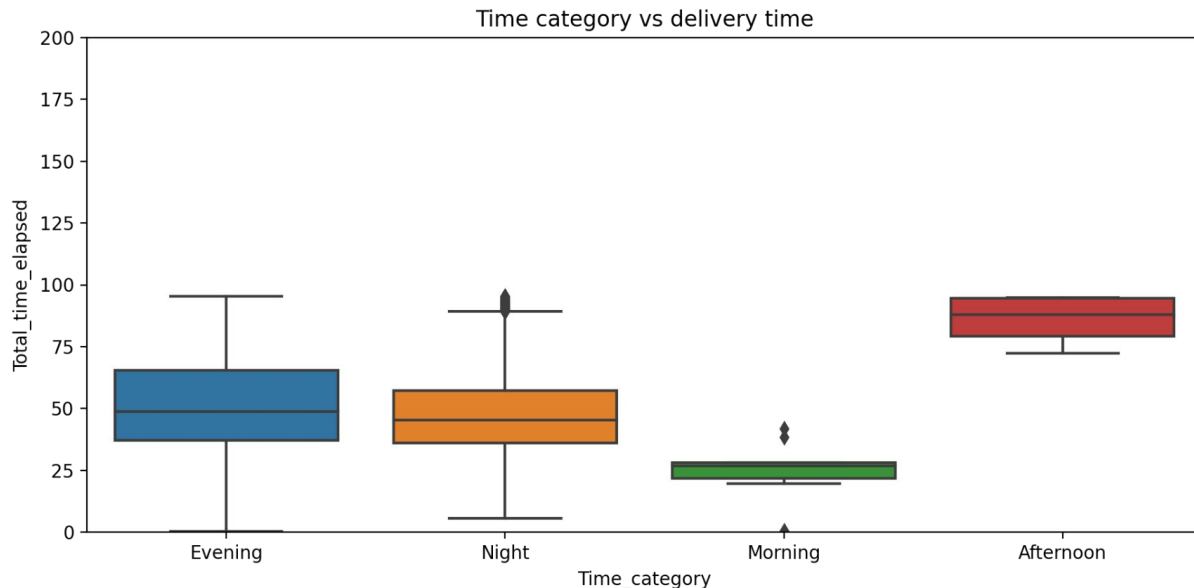
I had believed initially that the orders with a discount would naturally tip more just because of the fact that they would no longer be spending as much but there was little to no correlation between whether a discount was used and the tip amount.

Speed of Delivery vs Tip



This discovery was most shocking to me because I could only naturally assume that orders that came as soon as possible would be tipped more than orders that weren't so the fact that the boxplot shows a higher median and quartiles for tip amount when the order is not as soon as possible is very surprising.

Time of day vs delivery time



This specific graph was the one that most aligned with my predictions because as a driver I have seen myself getting stuck in traffic for what felt like an eternity during the afternoon times but that the evening and night deliveries were usually smaller and faster orders. Those deliveries are also more likely to be fast food so it was a much easier delivery time.



Machine Learning Model

ML model Findings

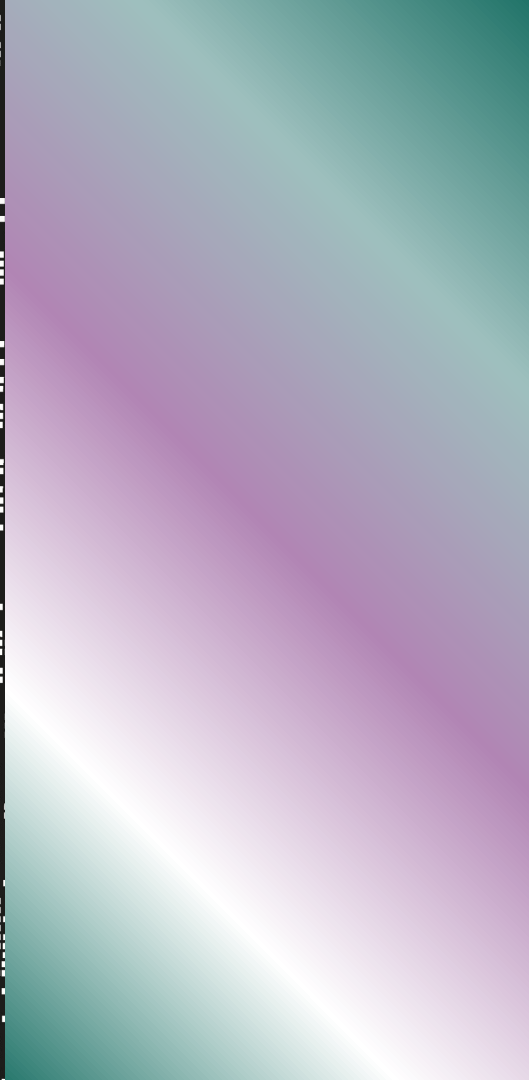
```
In [99]: mean_absolute_error(y_test,y_pred)
         #testing error methods
```

```
Out[99]: 4.25572172734861
```

```
In [100]: np.sqrt(mean_squared_error(y_test,y_pred))
          #testing error methods
```


```
Out[100]: 6.1464162318275815
```

I used a Linear Regression Model with Elastic Net to help predict the delivery times. With a RMSE value of 6.146 and comparing it to the mean delivery time of the data which was approximately 48.72 minutes I would conclude that the model was semi-accurate in its predictions. It definitely was not perfect, but it was not very far off from the time it actually took for the delivery to be made.





Reflection and Difficulties

A man with a beard, wearing a red shirt, is shown from the side, looking at a tablet. The tablet displays a dashboard with various data visualizations. At the top, there are four circular progress indicators with percentages: 25%, 50%, 65%, and 75%. Below these are several bar charts and a line graph. The background is a blurred office setting with windows.

**A picture is worth a
thousand words**



Difficulties



By far, the biggest difficulty I had during this entire project was cleaning the data. From my course experience and research, I had heard and seen that cleaning data was always the part of the process that was most time consuming, but I never really understood it till I got into doing my first project on my own with this dataset. The various methods I tried to use to convert the time columns into timedelta or datetime format was taking much longer than expected and at times it had felt like I had no solution. The process of filling in the missing data for the driver arrival time at the restaurant also was a long process with much trial and error. Besides the cleaning data step, I felt the rest of the project wasn't nearly as time consuming because it was all very straightforward and just an implementation of what I had learned previously.

Reflection and Changes

I believe there definitely were a few changes I could have made to potentially help carry out this project better and make a more accurate machine learning model to predict the delivery time.

- A better way for me to handle the outliers could have been used that may have greatly affected my model performance.
- Another way to convert the date and time columns that also involved the date the order was delivered to add in another variable.
- There is maybe a potential use to include the restaurant ID and the customer ID that could have helped the predictions
- I think with more time and experience, I could find better ways to map the data and analyze the graphs to produce more findings.

