Enhancing Food Delivery Services: A Predictive Model for Accurate Delivery Time Estimation

Sicheng (Kevin) Lu

Personal Projects # 2

Bachelor of Economics Honours

Minor in Statistics

Introduction

- Rapid growth of food delivery industry have more demand with efficient and accurate delivery time predictions
- Delivery time estimation is a crucial factor with customer satisfaction and loyalty
- Factors that affect delivery times like restaurant preparation time, traffic conditions, weather, and the distance of delivery can cause the accurate prediction more complex
- Will use data analytics and predictive modeling with R and Python offers a solution to this challenge.
- This study is based on creating reliable predictive model for food delivery time estimation to enhance operations and boost user experience.

Where did I get a Dataset??

kaggle

- Received a dataset from Kaggle's website:
  - Kaggle is a data science and artificial intelligence platform
  - Contains with many kind of datasets
  - Published by large companies and organizations
- Datasets: Food_Delivery_Times
- Data link: https://www.kaggle.com/datasets/denkuznetz/food-delivery-time-prediction?resource=download

Data Cleaning with MySQL + Excel

- Initial Data Cleaning can be done in MySQL
- Drop Order_ID as this will not be used for analyzing
- Rename from Delivery_Time_min to Time_Delivery
- Some of missing data, it has replaced as “Unknown”
- After Clean Data saved as CSV file
  - Showing just the text in only one column
  - Can be done in Excel by Text to Column
  - After that, now all the data are in proper column

Regression with R – Part One

Regression Equation:

$$\text{Time\_Delivery} = \beta_0 + \beta_1(\text{Distance\_km}) + \beta_2(\text{Weather}) + \beta_3(\text{Traffic\_Level}) + \beta_4(\text{Time\_of\_Day}) + \beta_5(\text{Vehicle\_Type}) + \beta_6(\text{Preparation\_Time\_min}) + \beta_7(\text{Courier\_Experience\_yrs})$$

Dependent Variable = Time_Delivery

Independent Variable = Distance_km, Weather, Traffic_Level, Time_of_Day, Vehicle_Type, Preparation_Time_min, Courier_Experience_yrs

Regression with R – Part One (Continued)

R Code:

```
model_test <- lm(Time_Delivery ~ Distance_km + Weather + Traffic_Level + Time_of_Day + Vehicle_Type + Preparation_Time_min + Courier_Experience_yrs, data = data)
summary(model_test)
```

```
Call:
lm(formula = Time_Delivery ~ Distance_km + Weather + Traffic_Level +
    Time_of_Day + Vehicle_Type + Preparation_Time_min + Courier_Experience_yrs,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-26.691  -6.220  -0.908   4.424  65.567

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               8.529031   1.997364   4.270 2.15e-05 ***
Distance_km               0.076485   0.001623  47.133  < 2e-16 ***
Weather                   1.189342   0.218161   5.452 6.34e-08 ***
Traffic_Level            -1.721697   0.447934  -3.844 0.000129 ***
Time_of_Day               0.049122   0.339761   0.145 0.885074
Vehicle_Type             -0.290195   0.412998  -0.703 0.482441
Preparation_Time_min      0.958519   0.049947  19.191  < 2e-16 ***
Courier_Experience_yrs   -0.597579   0.123694  -4.831 1.58e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.18 on 962 degrees of freedom
Multiple R-squared:  0.7352,    Adjusted R-squared:  0.7333
F-statistic: 381.6 on 7 and 962 DF,  p-value: < 2.2e-16
```

Figure 1: Regression Output from R

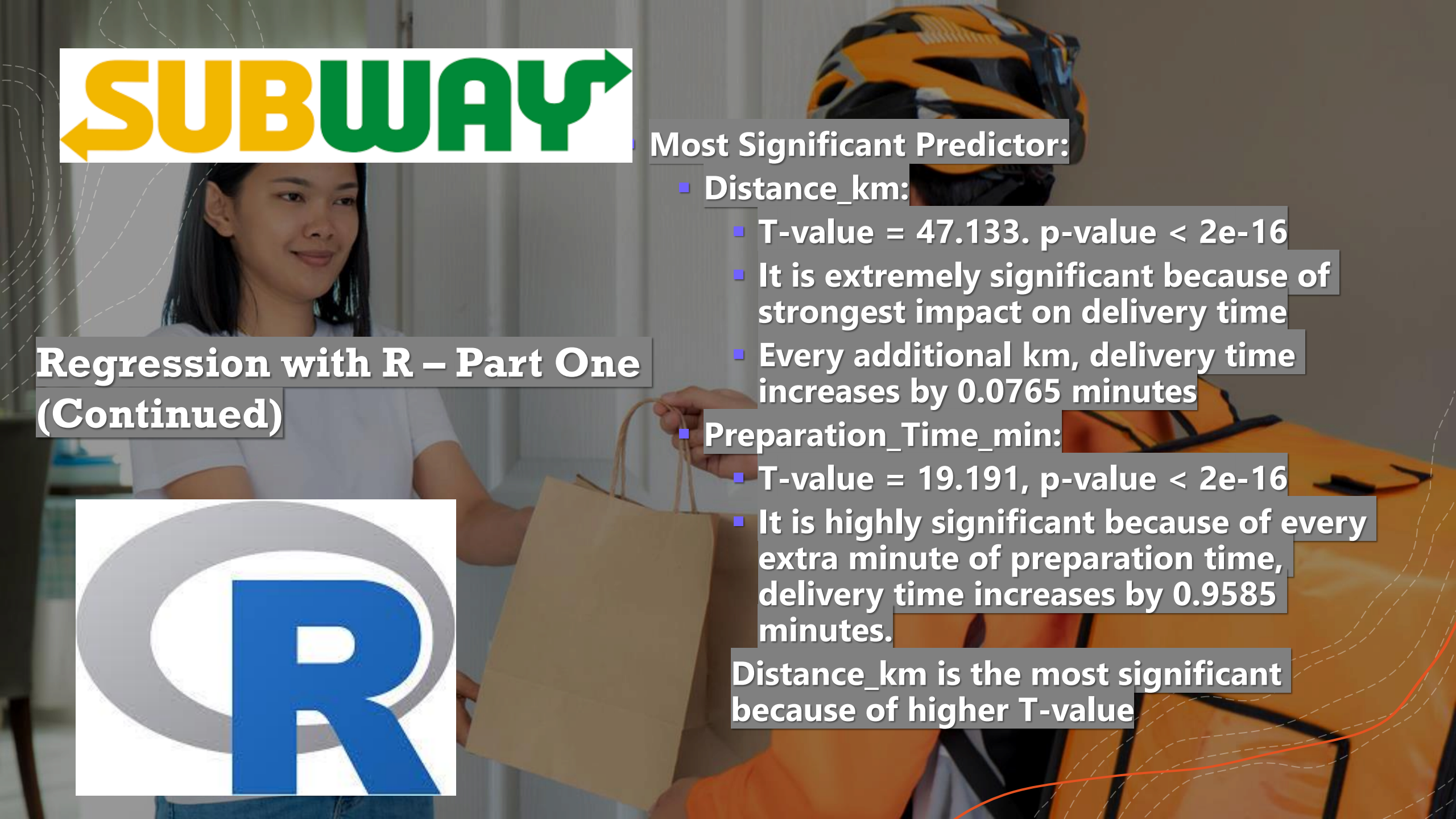
- $RSE = 11.18$ (Better Model fit)
- $R^2 = 0.7352$
- Adjusted R^2 = 0,7333

(Result: Indicate a good model fit)

- F-Statistic = 381.6
- P-value < 2.2e-16

This model is highly significant (Good explanation of the variation in delivery time using the given predictors)

- Degree of Freedom: 7 predictors + intercept



Regression with R – Part One (Continued)

Most Significant Predictor:

- **Distance_km:**
  - T-value = 47.133. p-value < 2e-16
  - It is extremely significant because of strongest impact on delivery time
  - Every additional km, delivery time increases by 0.0765 minutes
- **Preparation_Time_min:**
  - T-value = 19.191, p-value < 2e-16
  - It is highly significant because of every extra minute of preparation time, delivery time increases by 0.9585 minutes.

Distance_km is the most significant because of higher T-value

95% Confidence Interval from Regression Part One

```
R code:
conf_intervals <- confint(model_test, level = 0.95)
conf_intervals
```

A matrix: 8 × 2 of type dbl		
	2.5 %	97.5 %
(Intercept)	4.60933790	12.44872413
Distance_km	0.07330084	0.07966993
Weather	0.76121530	1.61746957
Traffic_Level	-2.60073670	-0.84265670
Time_of_Day	-0.61763666	0.71588084
Vehicle_Type	-1.10067592	0.52028595
Preparation_Time_min	0.86050231	1.05653621
Courier_Experience_yrs	-0.84031965	-0.35483770

Figure 2: Output of 95% CI based on regression from Part One

- Intercept:
  - CI = (4.61. 12.45), Significant
- Distance_km:
  - CI = (0.0733, 0.0797), Significant
- Weather:
  - CI = (0.761, 1.617), Significant
- Traffic_Level:
  - CI = (-2.601, -0.843), Significant
- Time_of_Day:
  - CI = (-0.618, 0.716), Not Significant
- Vehicle_Type:
  - CI = (-1.101, 0.520), Not Significant
- Preparation_Time_min:
  - CI = (0.861, 1.057), Significant)

Courier_Experience_yrs:

- CI = (-0.840, -0.355), Significant

Residuals from Regression – Part One

- Most residuals cluster between -20 and 60, with most of them around 0
- It looks like uneven spread which is called heteroscedasticity
- Have more extreme positive outliers
- No clear trend across index values, looks like random distribution
- Looks like have more extreme positive residuals (up to 60) than negative ones (down to -20)
- Showing dense concentration of points around zero, showing significant that have good model fit

R Code:

```
(plot(model_test$residuals))
```

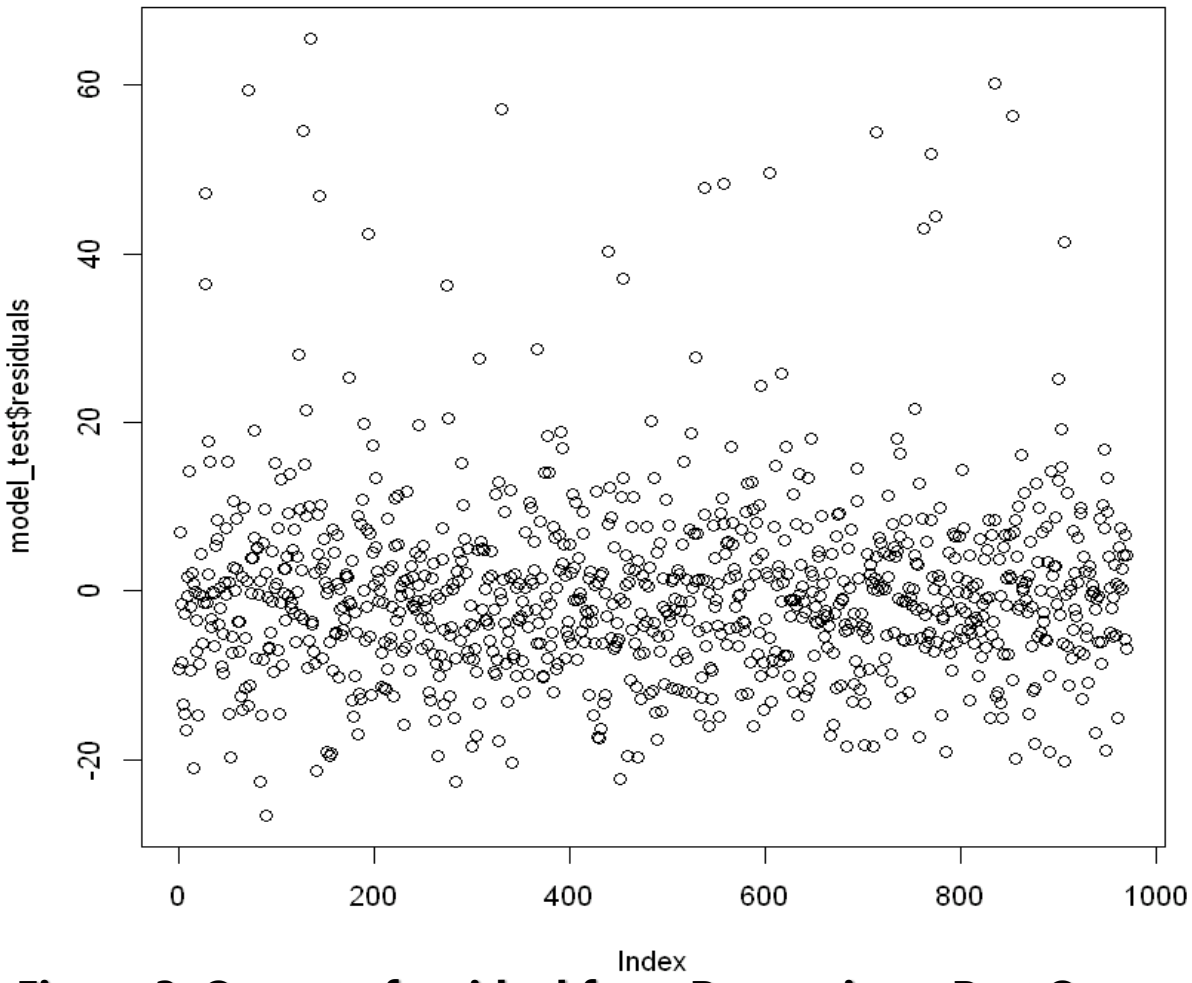
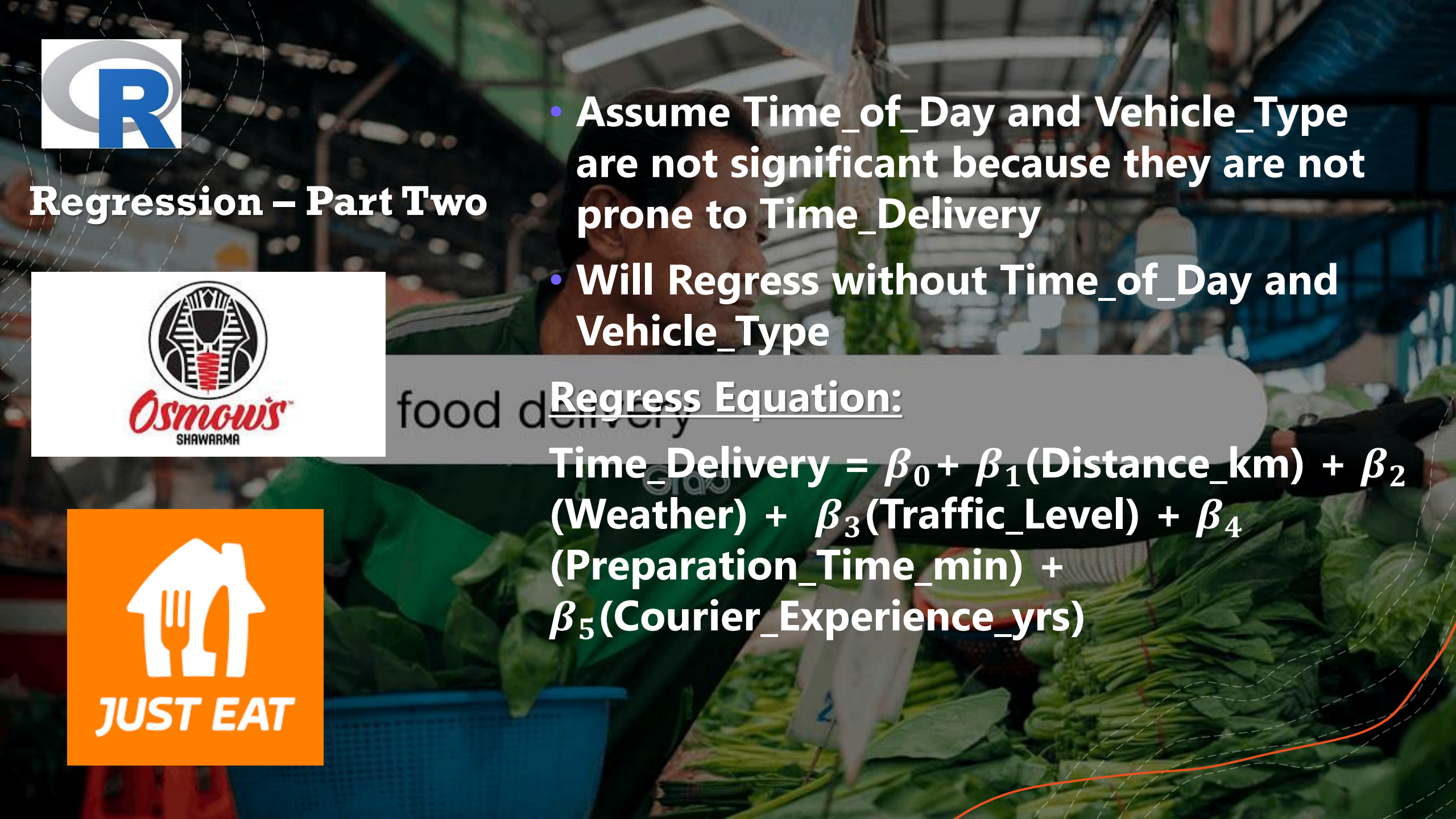


Figure 3: Output of residual from Regression – Part One



Regression – Part Two

- Assume Time_of_Day and Vehicle_Type are not significant because they are not prone to Time_Delivery
- Will Regress without Time_of_Day and Vehicle_Type

Regress Equation:

$$\text{Time\_Delivery} = \beta_0 + \beta_1(\text{Distance\_km}) + \beta_2(\text{Weather}) + \beta_3(\text{Traffic\_Level}) + \beta_4(\text{Preparation\_Time\_min}) + \beta_5(\text{Courier\_Experience\_yrs})$$

Regression – Part Two (Continued)

R Code:

```
model_test2 <- lm(Time_Delivery ~ Distance_km +
Weather + Traffic_Level + Preparation_Time_min +
Courier_Experience_yrs, data = data)
summary(model_test2)
```

```
Call:
lm(formula = Time_Delivery ~ Distance_km + Weather + Traffic_Level +
    Preparation_Time_min + Courier_Experience_yrs, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-26.720  -6.250  -0.792   4.284  65.769

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           8.148859   1.681848   4.845 1.47e-06 ***
Distance_km           0.076484   0.001621  47.171  < 2e-16 ***
Weather               1.194333   0.217886   5.481 5.39e-08 ***
Traffic_Level        -1.732633   0.447088  -3.875 0.000114 ***
Preparation_Time_min  0.957783   0.049894  19.196  < 2e-16 ***
Courier_Experience_yrs -0.598697   0.123395  -4.852 1.43e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.17 on 964 degrees of freedom
Multiple R-squared:  0.7351,    Adjusted R-squared:  0.7337
F-statistic:   535 on 5 and 964 DF,  p-value: < 2.2e-16
```

Figure 4: Output of Regression Part Two

- RSE = 11.17
- R-Squared = 0.7351, Adjusted R-squared = 0.7337
  - Suggest marginally better at capturing variance in the data
- F-statistic = 535, p-value < 2.2e -16
  - Higher F-Statistic which prove to have stronger significance
- Degrees of Freedom: 5 predictors + Intercept

Comparison Regression Part One and Two

- RSE:
  - Part One = 11.16, Part two = 11.17
  - Slightly higher RSE in part two shows marginally lower predictive accuracy
- R-Squared:
  - R-squared:
    - Part One = 0.7357, Part Two = 0.7351
  - Adjusted R-squared:
    - Part One = 0,7340, Part Two = 0.7337
    - Part One model shows more variability (0.06% higher R-squared)
- F-statistic:
  - Part One: F = 427.2, p-value < 2.2e-16, Part Two: F = 535, p-value < 2.2e-16
  - Part Two have higher F-Statistic, shows stronger overall significance
- Degrees of Freedom:
  - Part One: 962 residual degrees (7 predictors + intercept)
  - Part Two: 964 residual degrees (5 predictors + intercept)
  - Fewers predictors in Part Two have more degrees of freedom show more efficient

95% Confidence Interval from Regression Part Two

R Code:

```
conf_intervals2 <- confint(model_test2, level = 0.95)
conf_intervals2
```

A matrix: 6 × 2 of type dbl

	2.5 %	97.5 %
(Intercept)	4.84835453	11.44936443
Distance_km	0.07330252	0.07966641
Weather	0.76674847	1.62191774
Traffic_Level	-2.61001103	-0.85525414
Preparation_Time_min	0.85986934	1.05569750
Courier_Experience_yrs	-0.84085102	-0.35654291

Figure 5: Output of 95% CI from Regression Part Two

- Intercept: expected delivery time between 4.85 and 11.45 minutes
- Distance_km: delivery time increases by 0.073 to 0.080 minutes
- Weather: Adverse weather conditions increase delivery time by 0.77 to 1.62 minutes
- Traffic_Level: Low traffic levels reduce delivery time by 0.86 t0 2.61 minutes (Likely complex relationship)
- Preparation_Time_min: Every extra minute, delivery time increases by 0.86 to 1.06 minutes
- Courier_Experience_yrs: More experiences reduces delivery time by 0.26 to 0.84 minutes per year

- As the result, all of the predictors are significant and more prone to Time_Delivery.

Residual from Regression – Part Two

R Code:

```
(plot(model_test2$residuals))
```

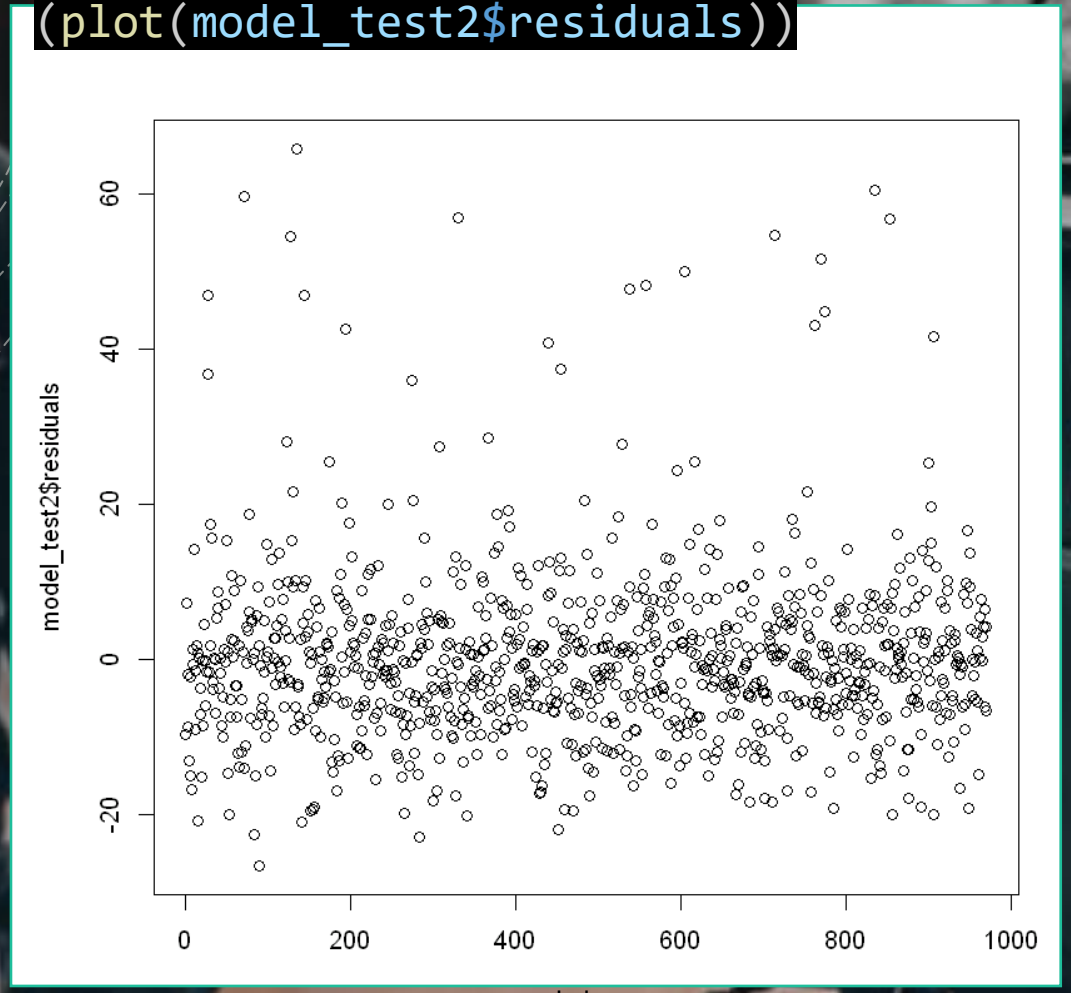


Figure 6: Output of Residual Graph from Regression Part Two

Show exactly the same pattern and distribution of residuals as Part One's residual
Same range (-20 to +60)
Clustering around zero
Outlier pattern with more extreme positive values
No visible differences between two plots

Since, I removed two predictor who are not significant to the Time delivery, it has no meaningful impact on this residual structure, So this residual plots appear to be identical to Part One's residual plot

Average Time of Delivery based on Weather

```
Python Code:
Weather_avg =
df.groupby('Weather')['Time_Delivery'].mean().reset_index()
plt.figure(figsize=(8,4))
sns.barplot(Weather_avg, x = 'Weather', y =
'Time_Delivery')
plt.title('Average Time of Delivery based on Weather')
plt.xlabel('Weather')
plt.ylabel('Average of Time Delivery')
plt.show()
```

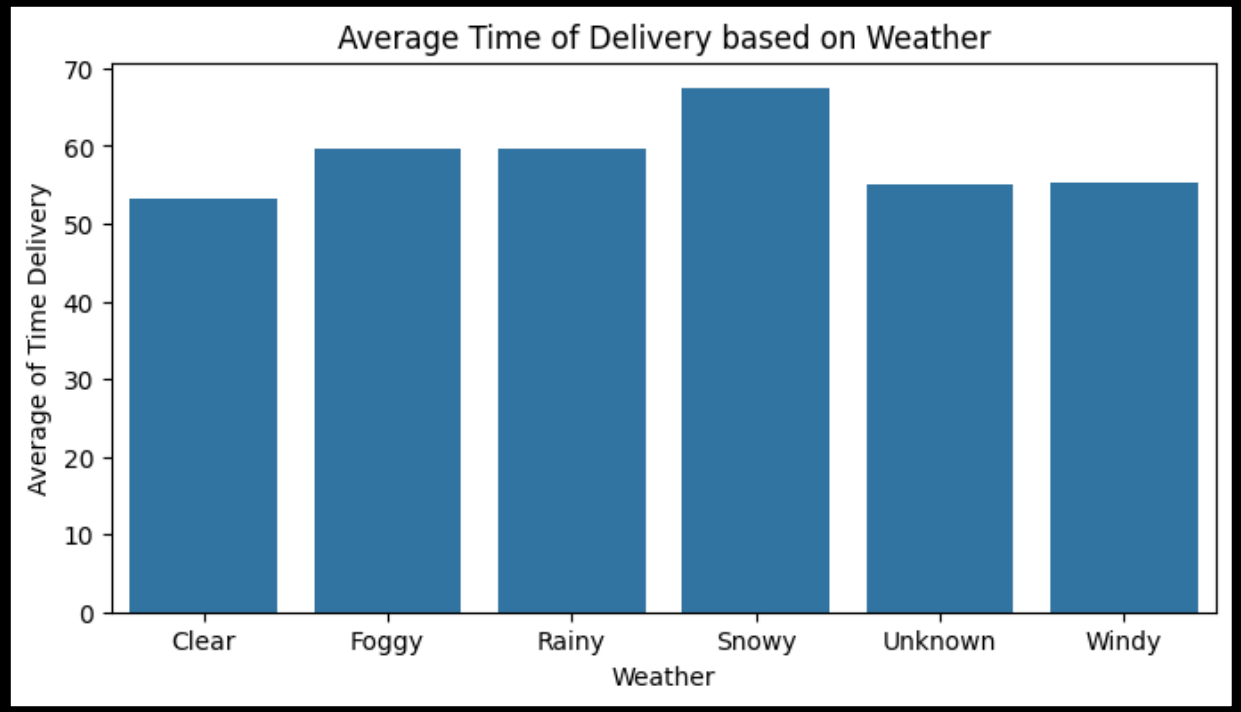
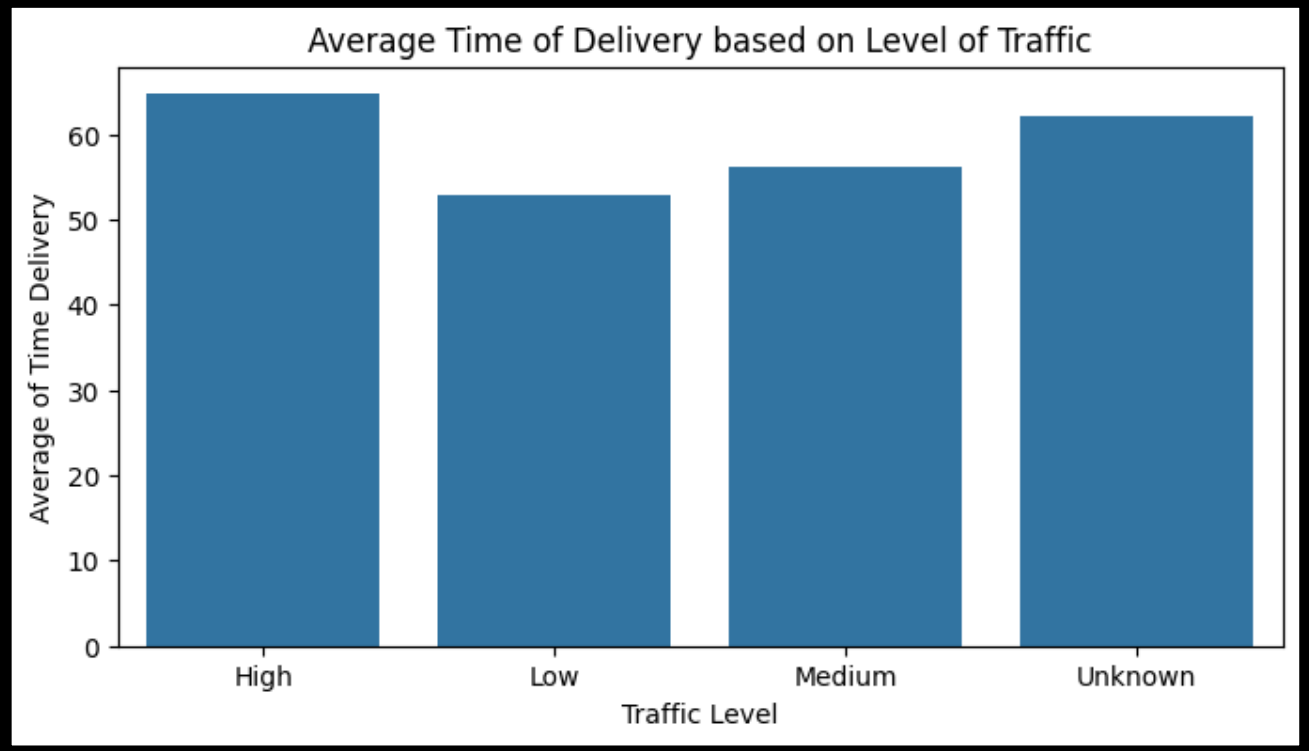


Figure 7: Bar plot of Average Time Delivery based on Weather

- Made this Bar plot graph in Python
- Snowy weather has the highest average delivery time
  - Possibly due to reduced visibility, road slippery, traffic congestion, difficult to reach location, or need extra precautions could affect Delivery time.
- Foggy and Rainy Weather shows more time on delivery compared to clear conditions
  - Could be road slippery or reduce visibility
- Clear Weather shows lowest average delivery time
- Windy conditions looks slightly higher than clear weather

Average of Delivery Time Based on Level of Traffic

```
Python Code:
Traffic_avg =
df.groupby('Traffic_Level')['Time_Delivery'].mean().reset_index()
plt.figure(figsize=(8,4))
sns.barplot(Traffic_avg, x = 'Traffic_Level', y = 'Time_Delivery')
plt.title('Average Time of Delivery based on Level of Traffic')
plt.xlabel('Traffic Level')
plt.ylabel('Average of Time Delivery')
plt.show()
```

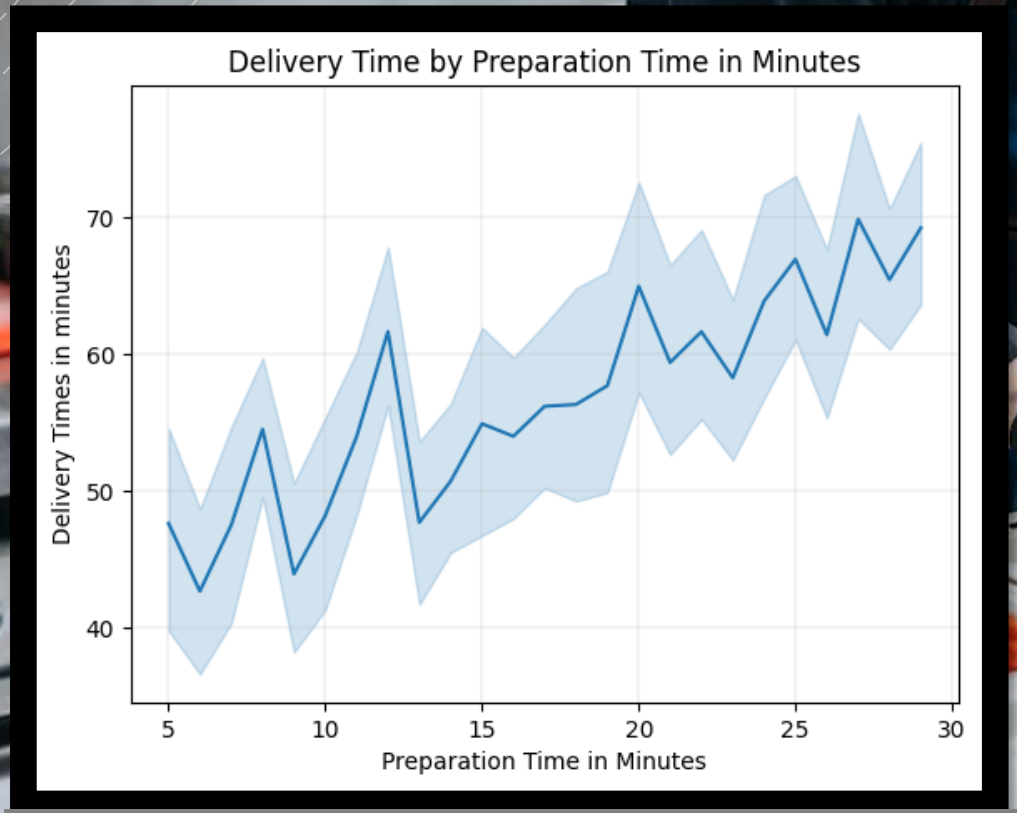


- High Traffic shows increase the average of Time of Delivery
  - Possibly due to busy roads, accidents, road closure, construction, or events that blocking the road
- Low Traffic shows less Time of Delivery
  - Showing no traffic, make smooth delivery without disruptions
- Medium Traffic shows slightly higher than Low Delivery
- There are unknown which they could not determine the on the traffic for Time Delivery

Figure 8: Box Plot of Average Time of Delivery based on Level of Traffic

Delivery Time by Preparation Time in Minutes

```
Python Code:
sns.lineplot(df, x = 'Preparation_Time_min', y =
'Time_Delivery')
plt.title('Delivery Time by Preparation Time in
Minutes')
plt.xlabel('Preparation Time in Minutes')
plt.ylabel('Delivery Times in minutes')
plt.grid(linewidth = 0.2)
plt.show()
```



- X-Axis = Preparation Times in Minutes
- Y-Axis = Delivery Times in Minutes

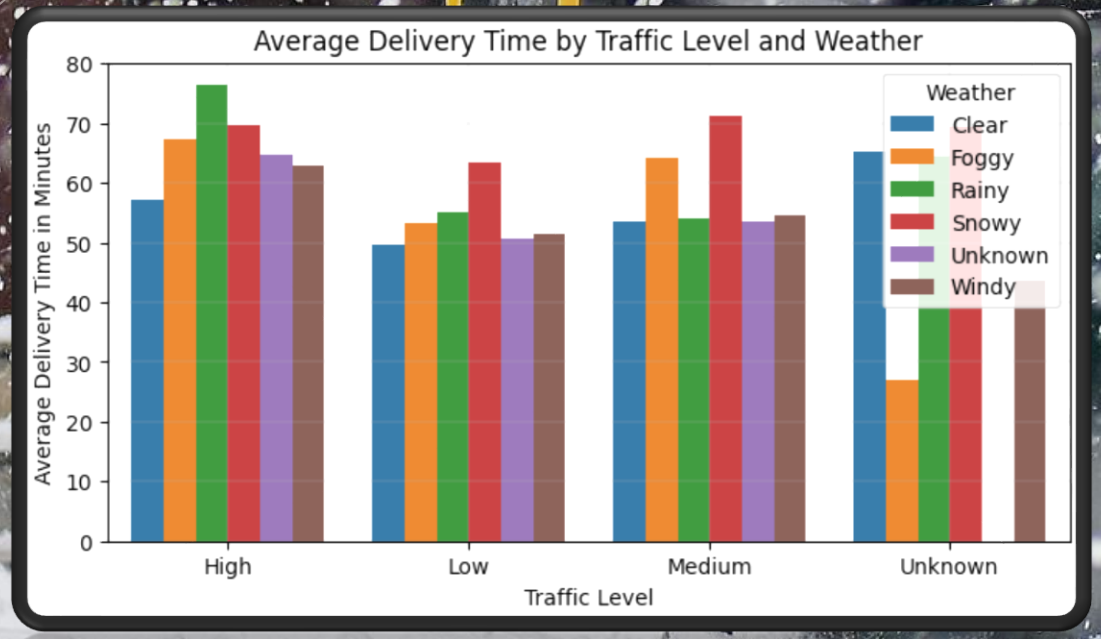
- Looks like if restaurants need more time to prepare for food for delivery, it may increase the delivery times.
  - Ex: If they prepared the meal for 30 minutes, it mostly likely to complete delivery up to roughly 65 minutes

- However, it does not mean they prepared 10 to 15 minutes and complete sooner.
  - Usually Advisory Weather, Road conditions, or traffic could affect the delivery times if they prepared less than 15 minutes.

Figure 9: Line plot for Delivery Time by Preparation Time in Minutes

Average Delivery Time by Traffic Level and Weather

```
Python Code:
pivot = df.groupby(['Traffic_Level', 'Weather'])
['Time_Delivery'].mean().reset_index()
plt.figure(figsize=(8, 4))
sns.barplot(pivot, x = 'Traffic_Level', y = 'Time_Delivery',
hue = 'Weather')
plt.title('Average Delivery Time by Traffic Level and
Weather')
plt.xlabel('Traffic Level')
plt.ylabel('Average Delivery Time in Minutes')
plt.grid(linewidth = 0.2)
plt.show()
```



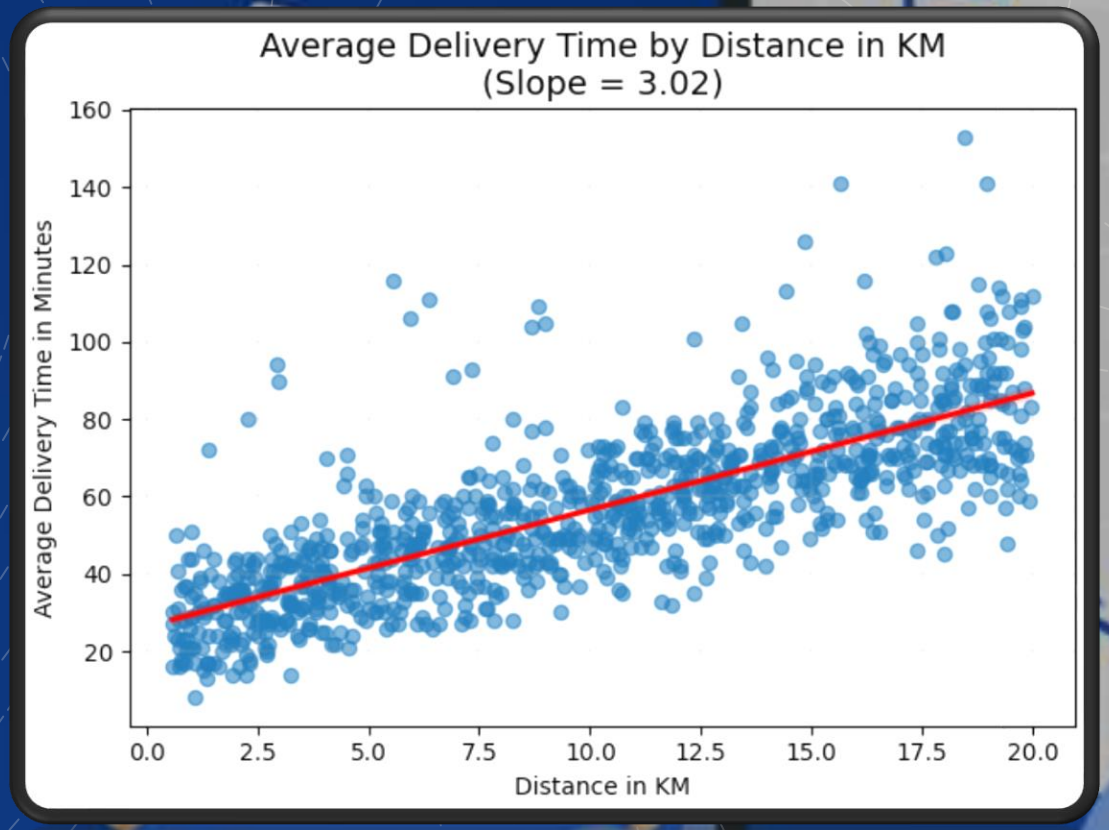
- High Traffic: Rainy and snow weather takes more delays. Rainy is the highest.
- Low Traffic: Snowy weather increase delivery time more than other weather type
- Medium Traffic: Rainy and snow takes longer delivery times
- Unknown Traffic: Shows likely represents missing or incomplete data based on traffic conditions

- Weather is more prone to traffic that can affect time delivery because of adverse conditions which can lead to risk of accidents, difficult to navigate on road, and cause congestion.

Figure 10: Bar plot of Average Delivery Time by Traffic Level and Weather

Average Delivery Time by Distance in KM

Figure 11: Scatter with line of fitting for Average Delivery Time by Distance in KM




```
Python Code:
from scipy.stats import linregress
sns.regplot(df, x = 'Distance_km', y = 'Time_Delivery',
scatter_kws={'alpha': 0.6}, line_kws={'color':'red'})
slope, intercept, r_value, p_value, std_err =
linregress(df['Distance_km'], df['Time_Delivery'])
plt.title(f'Average Delivery Time by Distance in KM\n(Slope
= {slope:.2f})', fontsize=14)
plt.xlabel('Distance in KM')
plt.ylabel('Average Delivery Time in Minutes')
plt.grid(linewidth = 0.2)
plt.tight_layout()
plt.show()
```

- Slope = 3.02: additional KM adds about 3 minutes to delivery
- Longer distance could increase delivery times
- Looks like it spreads out:
  - Traffic conditions, weather, or using detour route could affect the delivery times
  - Courier with more experience could spread for longer distance due to better at navigating complex routes and finding shortcuts while less experience stick to standards routes or follow the GPS

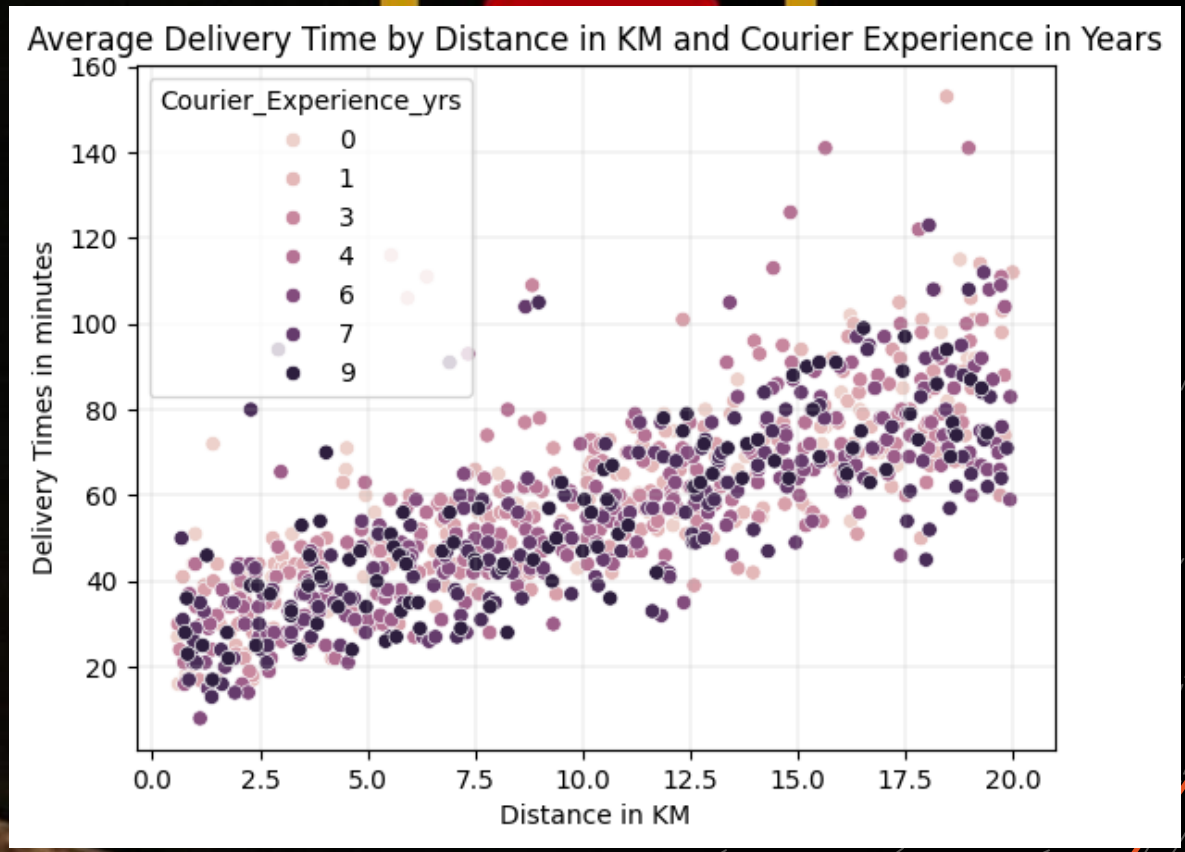
Average Delivery Time by Distance in KM and Courier Experience in Years

Python Code :

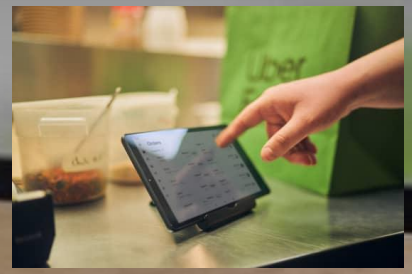
```
courier_experience =
df.groupby(['Courier_Experience_yrs',
'Distance_km'])
['Time_Delivery'].mean().reset_index()
sns.scatterplot(courier_experience, x =
'Distance_km', y = 'Time_Delivery', hue =
'Courier_Experience_yrs')
plt.title('Average Delivery Time by Distance in
KM and Courier Experience in Years')
plt.xlabel('Distance in KM')
plt.ylabel('Delivery Times in minutes')
plt.grid(linewidth = 0.2)
plt.show()
```

- As mentioned in previous slides that Courier with experience could spread out in longer distance
- It may handle different types of routes (Ex: City, town, or countryside) lead to different delivery times
- Individual couriers may vary due to efficiency, speed, or familiarity with specific routes
- Longer distance likely reduce the effect of experience because of physical constraints like limits of speed.

Figure 12: Average Delivery Time by Distance in KM and Courier Experience in Years



Conclusion



- Model Performance:
  - R-Squared = 0.735 (good predictive power)
  - Removing Time_of_Day and Vehicle_Type since they are not significant to Time_Delivery
- Distance in KM are the most significant predictor and more prone to time delivery with each of kilometer adding by roughly 3 minutes to delivery
- Weather affect delivery times:
  - Snow/Rain/Foggy can affect On-Time delivery as this can cause most delays
  - Clear Weather shows fastest delivery times due to smooth traffic flows
- Traffic:
  - High Traffic can increase delivery time (Rush Hours or any thing that can cause congestion)
  - Weather combined with High traffic can cause more delays
- Preparation Time:
  - Preparation Time affects delivery times:
    - Restaurants may be slower or might be busier, depending on the volume of number of customer and number of orders
- Suggestion for Next Step:
  - Model Improvements with Non-linear Relationships
    - Due to clear non-linear patterns (spreading out at longer distances)