LINEAR MODEL

# Assignment 4

**GROUP MEMBERS**

**PROJECT GUIDE :**

**Prof . Priya Deshpande**

**NAMES PRN**

**SUBHADIP GHOSH 23060641089**

**DEVASHRI JOSHI 23060641061**

# Introduction :

In the rapidly growing online food delivery industry, predicting delivery times accurately is crucial for ensuring customer satisfaction and business success . In today's fast- paced world, online food delivery services have become increasingly popular, offering convenience and variety to customers. Timely delivery is essential for customer satisfaction and retention. Therefore, this project focuses on developing a predictive model for estimating delivery times in the online food delivery industry. By analyzing factors such as delivery boy age, delivery person ratings, distance between the restaurant and delivery location, type of order, and type of vehicle, this model aims to provide accurate delivery time predictions. This predictive model will not only improve operational efficiency but also enhance customer experience, ultimately contributing to the success of online food delivery businesses. "In the ever-evolving landscape of the online food delivery industry, efficient delivery operations play a pivotal role in ensuring customer satisfaction and loyalty. With the objective of enhancing delivery efficiency and providing customers with accurate delivery estimates, our project focuses on predicting delivery times using a multiple linear regression model. To achieve this, we have gathered a comprehensive dataset containing various factors that influence delivery times. These factors include delivery person age, ratings, distance between the restaurant and delivery location, type of order, and type of vehicle used for delivery. By analyzing these factors, our goal is to develop a model that can accurately predict delivery times, taking into account the diverse variables involved in the delivery process.

Through our analysis, we aim to identify the significant factors affecting delivery times and understand how delivery person characteristics, distance, and mode of transportation impact the overall delivery process. By leveraging data-driven insights, we seek to provide online food delivery platforms with valuable information to optimize their delivery operations. Our model will enable platforms to better estimate delivery times, leading to improved service efficiency and enhanced customer satisfaction. By streamlining the

delivery process and providing more accurate delivery estimates, online food delivery platforms can strengthen their competitive position in the market.

The insights gained from our analysis will not only benefit online food delivery platforms but also contribute to the broader understanding of delivery logistics in the e-commerce sector. Through this project, we aim to address the challenges associated with predicting delivery times accurately in the online food delivery industry. By combining data analysis with machine learning techniques, we are confident that our model will provide valuable insights to stakeholders in the online food delivery ecosystem. Ultimately, our goal is to improve the overall customer experience by ensuring timely and efficient food delivery services."

# Problem Statement :

The online food delivery industry faces challenges in accurately estimating delivery times, leading to customer dissatisfaction and operational inefficiencies. Inaccurate predictions can result in delayed deliveries, impacting customer experience and loyalty. Therefore, there is a need to develop a predictive model to estimate delivery times accurately.

# Aim Of the Project :

The aim of this project is to develop a multiple linear regression model that can predict delivery times for online food orders with precision. By analyzing factors such as delivery boy age, delivery person ratings, distance between the restaurant and delivery location, type of order, and type of vehicle, the model will provide accurate estimations of delivery times. This predictive model aims to enhance operational efficiency, optimize resource allocation, and improve overall customer satisfaction in the online food delivery industry**.**

# Data Set :

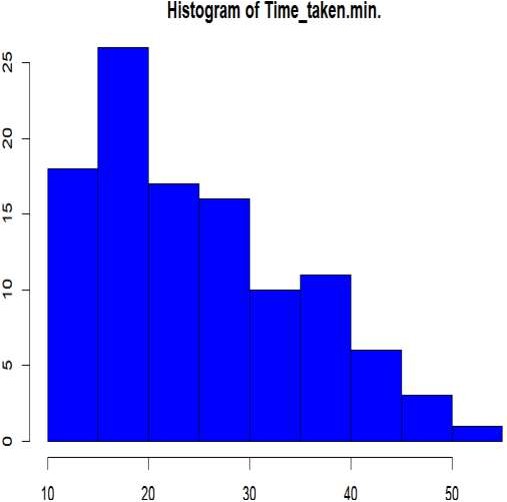
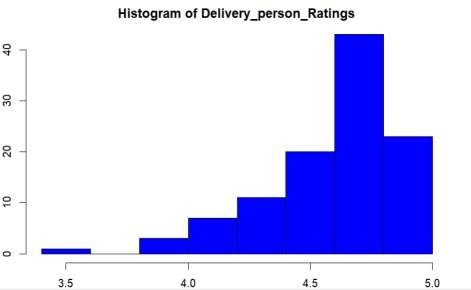
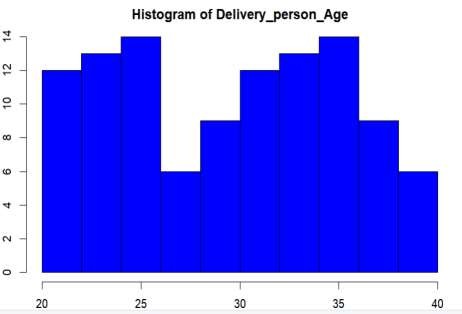
* Name : Food delivery Time Case Study
* Total number of observations : 546
* Data Source : Kaggle
* Data Link : <https://www.kaggle.com/devashripravinjoshi>

# Variable Used :

The data set, named "Online Food Delivery Time," contains the following factors:

* Delivery boy age
* Delivery person ratings
* Restaurant latitude and longitude
* Delivery location latitude and longitude
* Distance between the restaurant and delivery location
* Type of order (snacks, drinks, buffet, meal)
* Type of vehicle (scooter, electrical scooter, motorcycle)
* Time taken for delivery
* **EDA (Exploratory Data Analysis) :**

# Histogram :

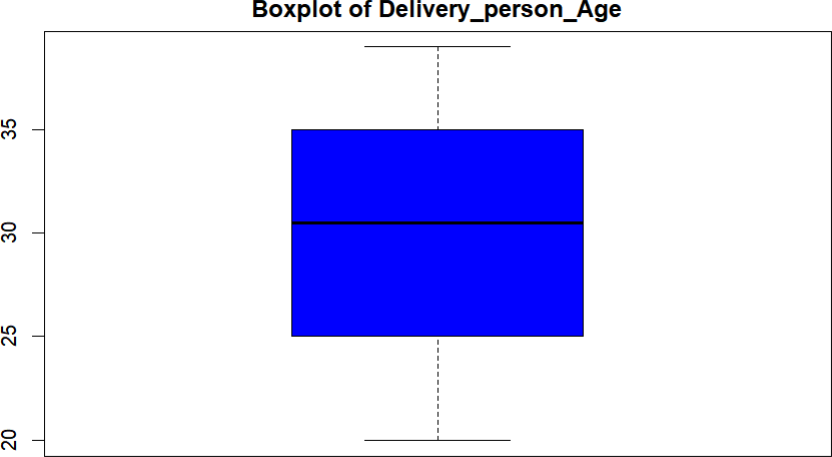


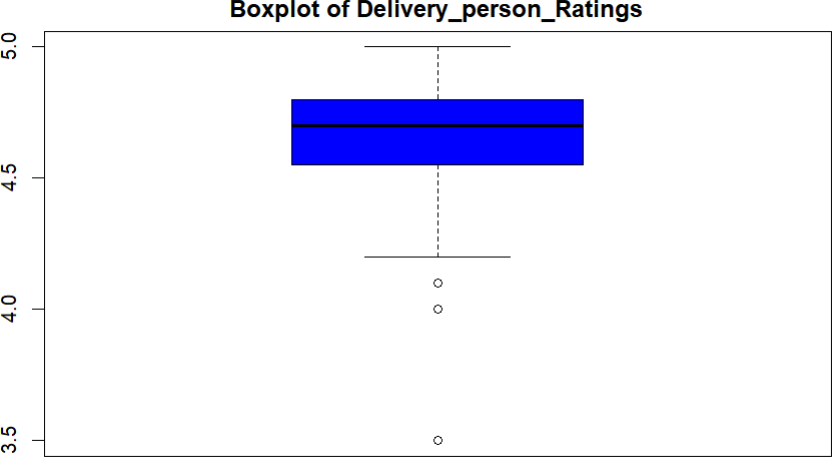
**Interpretation :**

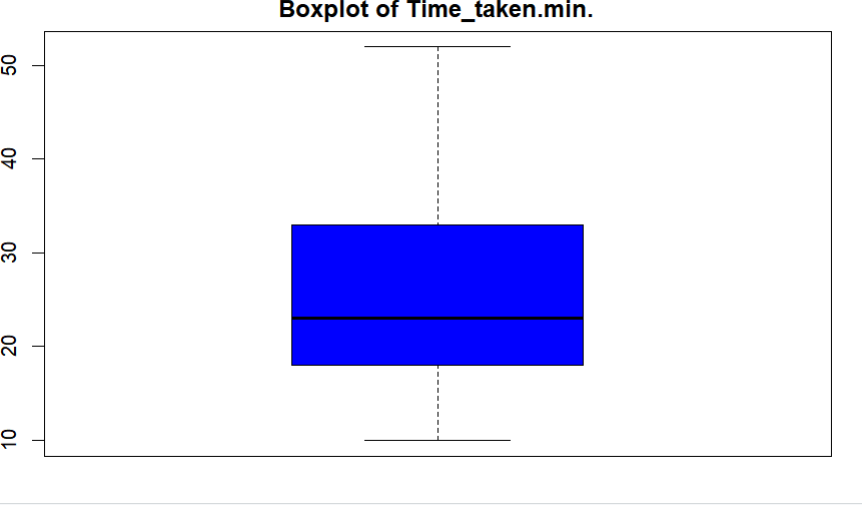
Histogram is used to represent the distribution of the numerical data.

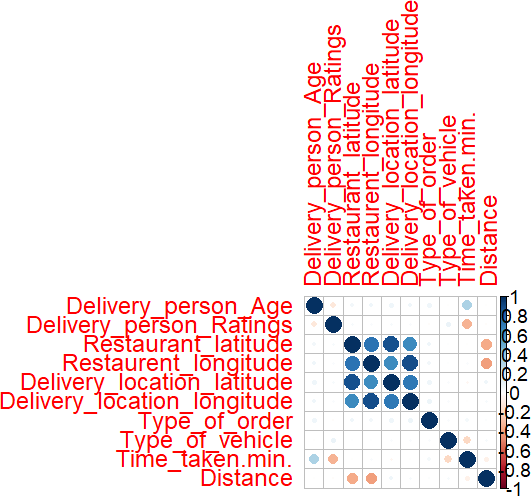
* It does show that the distribution of delivery person ages is skewed to the left, meaning there are more younger delivery people than older delivery people.
* The skewness suggests that there might be a tendency for customers to be more likely to leave a rating when they are satisfied with the service than when they are unsatisfied.
* It reveals that most tasks took around 20 minutes to finish, with a spread of times in both shorter and longer duration.
* If histogram is not bell shaped then various techniques like transformation , alternative measures of center and spread can be used to normalize the distribution.

# Box plot :









* **Interpretation :**

A box plot, also called a box-and-whisker plot and it is a graphical tool used to depict the distribution of a dataset .

* + It is a versatile and informative way to explore and visualize the distribution of data, especially when dealing with potential outliers.
  + The boxplot suggests that most delivery people are between 25 and 30 years old and there is one outlier who is 42 years old.
  + The median rating most delivery people received is 4.5 and there is no significant skew in the distribution.
  + The box is slightly skewed to the right, meaning there might be a few more tasks that took longer than 25 minutes compared to the number of tasks that took less than 15 minutes.
  + The whiskers extend outward from the box. The whiskers indicate the range of data points that fall outside the central IQR but are not considered outliers.

# Heat map :

* **Interpretation :**
  + A heatmap is a visualization tool used to represent data as color-coded squares. Imagine a grid where each square corresponds to a data point. The colour intensity of the square reflects the value of the data point.
  + Weak positive correlation between delivery person age and ratings (possibly meaning slightly higher ratings for older workers).
  + Weak positive correlation between ratings and time taken (deliveries that take longer might receive slightly higher ratings).
  + Moderate negative correlation between type of vehicle and distance (certain vehicles tend to be used for shorter distances).
  + Weak positive correlation between distance and time taken (longer distances tend to take more time).

## R code for EDA :

#EDA

library(dplyr) library(ggplot2) library(corrplot) par(mfrow=c(1,1)) par(mar=c(2, 2, 2, 2))

for (i in 3:11) {

hist(data2[,i], main = paste("Histogram of", names(data12)[i]))

}

par(mfrow=c(1,1)) for (i in 3:11) {

boxplot(data2[,i], main = paste("Boxplot of", names(data12)[i]))

}

numeric\_data <- data2[, sapply(data2, is.numeric)]

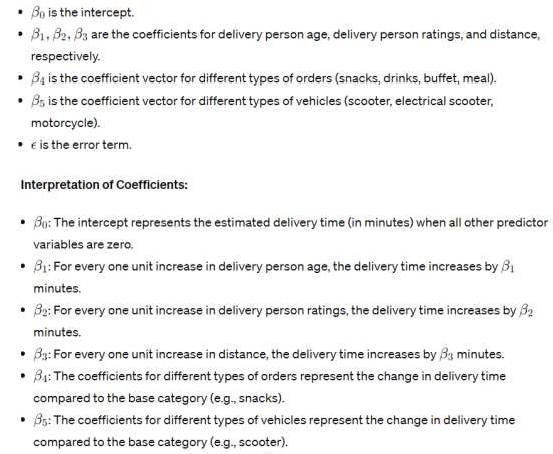
correlation\_matrix <- cor(data2[, sapply(data2, is.numeric)]) print(correlation\_matrix)

corrplot(correlation\_matrix, method = "circle")

# Fitting The Model :

### Multiple linear regression formula

The formula for a multiple linear regression is:



# Out put :

### Interpretation of the Model:

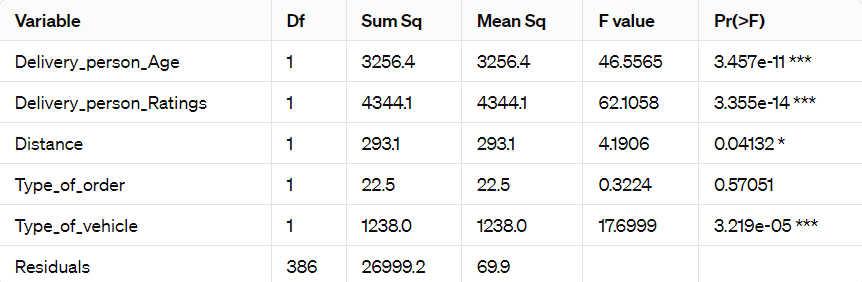
* **Intercept (β₀):** The intercept is -16.88. It represents the estimated delivery time (in minutes) when all other predictor variables are zero. However, since having zero delivery person age, zero delivery person ratings, or zero distance is not meaningful, the interpretation of the intercept is limited.
* **Delivery\_person\_Age (β₁)**: For every one unit increase in delivery person age, the delivery time increases by 0.083 minutes.
* **Delivery\_person\_Ratings (β₂):** For every one unit increase in delivery person ratings, the delivery time increases by 0.487 minutes.
* **Distance (β₃):** For every one unit increase in distance, the delivery time increases by

0.317 minutes.

* **Type\_of\_order (β₄, β₅, β₆)**: The coefficients for different types of orders (snacks, drinks, buffet) are not significant (p > 0.05), indicating that the type of order does not have a significant effect on delivery time.
* **Type\_of\_vehicle (β₇, β₈):** The coefficients for different types of vehicles (electrical scooter, motorcycle) are not significant (p > 0.05), indicating that the type of vehicle does not have a significant effect on delivery time.

Delivery person age, delivery person ratings, and distance significantly influence delivery time. However, the type of order and type of vehicle do not have a significant effect on delivery time. This information can be valuable for optimizing delivery operations and improving service efficiency in the online food delivery industry.

# Anova Table :



## Interpretation:

* **Delivery\_person\_Age:** Delivery person age significantly affects the delivery time **(F = 46.56, p < 0.001).**
* **Delivery\_person\_Ratings**: Delivery person ratings significantly affect the delivery time

### (F = 62.11, p < 0.001).

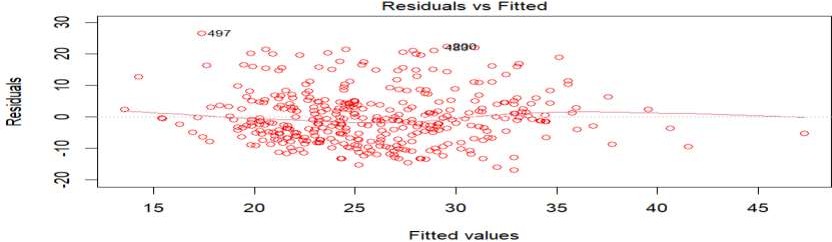
* **Distance:** The distance between the restaurant and delivery location has a marginally significant effect on delivery time **(F = 4.19, p = 0.041).**
* **Type\_of\_order**: The type of order does not have a significant effect on delivery time **(F**

### = 0.32, p = 0.570).

* **Type\_of\_vehicle:** The type of vehicle significantly affects delivery time **(F = 17.70, p < 0.001).**

# Conclusion:

Delivery person age, delivery person ratings, distance, and type of vehicle significantly influence delivery time. However, the type of order does not have a significant effect on delivery time. This information can be valuable for optimizing delivery operations and improving service efficiency in the online food delivery industry.



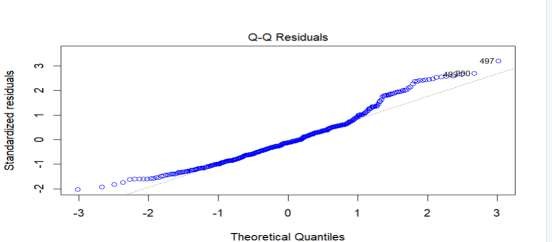
# Plots :

## Residuals vs Fitted

**Interpretation**

The points should be randomly scattered around the horizontal line at zero. This indicates that the residuals have constant variance and are not dependent on the predicted values. The residuals are randomly scattered around the horizontal line at zero. some points that deviate significantly from the horizontal line at zero that indicate potential outliers or influential observations.

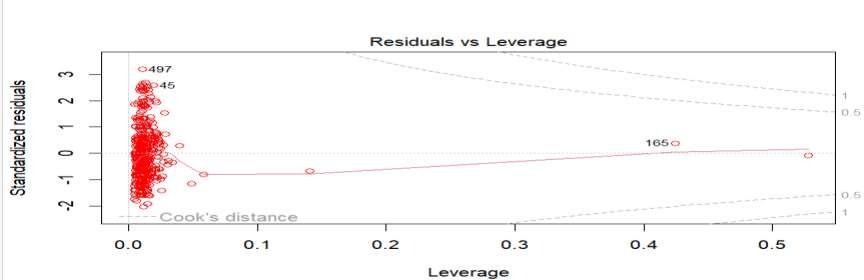
## Normal Probability Plot :



**Interpretation**

If the points on the Q-Q plot approximately follow a straight line, it suggests that the residuals are normally distributed. This indicates that the assumption of normality for the residuals is reasonable, which is important for making valid inferences and predictions in a multiple linear regression model. Points deviating upwards indicate positive skewness .the tails of the Q-Q plot deviate from the straight line, it indicate that heavier or lighter tails than the normal distribution.

## Residuals Vs Leverage Plot :



**Interpretation :**

If a data point has a high residual, it means that the model did a poor job of predicting the value of that point. This could be due to an outlier in the data, or it could be a sign that the model is not appropriate for the data.If a data point has a high residual, it means that the model did a poor job of predicting the value of that point. This could be due to an outlier in the data, or it could be a sign that the model is not appropriate for the data.

# Summary of The Data :

Minimum**:** -**16.877**

1st Quartile (Q1): -**6.018** Median: -**1.049**

3rd Quartile (Q3): **4.341**

Maximum: **26.597**

Residual Standard Error: **8.363 on 386** degrees of freedom Multiple R-squared: **0.2532**

This indicates that approximately **25.32%** of the variability in the response variable (delivery time) can be explained by the independent variables in the model.

Adjusted R-squared: **0.2435**

This is the R-squared value adjusted for the number of predictors in the model. It provides a more accurate indication of the model's goodness-of-fit.

F-statistic: **26.18 on 5 and 386** degrees of freedom

This tests the overall significance of the regression model.

The low p-value **(< 2.2e-16)** suggests that the overall model is statistically significant.

## Conclusion:

The model explains approximately **25.32%** of the variability in the delivery time. The model is statistically significant **(p < 2.2e-16).**

However, the adjusted R-squared value is relatively low, indicating that the model may not fit the data well. Further investigation may be needed to improve the model's performance.

# Predictions :

## R code

# Make predictions using the multiple linear regression model predictions <- predict(model, newdata = test\_data)

# View the predictions predictions

## Output

1 2 9 15 17 18 19 21 28

32

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 34.88332 | 21.65089 | 26.06331 | 24.73728 | 29.36245 | 34.87365 | 27.15628 | 31.11385 | 26.08 |
| 37 | 38 | 43 | 44 | 46 | 47 | 49 | 56 |  |
| 24.94879 | 19.83402 | 22.46669 | 35.54890 | 16.24748 | 29.34195 | 27.38858 | 12.70451 | 25.17 |
| 62 | 63 | 65 | 68 | 71 | 78 | 79 | 82 |  |
| 24.45912 | 24.92120 | 23.58485 | 20.07166 | 24.07147 | 25.83348 | 19.84060 | 27.79472 | 30.01 |
| 95 | 99 | 101 | 103 | 109 | 112 | 124 | 126 |  |
| 21.17004 | 20.12653 | 37.30853 | 23.42780 | 21.87133 | 25.62806 | 18.87722 | 26.85890 | 24.96 |
| 133 | 138 | 140 | 142 | 144 | 149 | 150 | 157 |  |
| 22.97962 | 27.37559 | 22.32620 | 30.69272 | 28.22906 | 27.61270 | 20.79035 | 20.48667 | 16.09 |
| 169 | 171 | 173 | 180 | 182 | 183 | 187 | 189 |  |
| 33.11822 | 22.85833 | 33.32463 | 32.48142 | 21.86569 | 30.39934 | 35.97612 | 25.09387 | 37.69 |
| 193 | 201 | 203 | 206 | 207 | 216 | 219 | 220 |  |
| 19.18179 | 23.78844 | 15.58673 | 21.57411 | 24.72367 | 29.12349 | 27.10367 | 24.93985 | 24.21 |
| 231 | 233 | 239 | 240 | 241 | 246 | 247 | 248 |  |
| 24.45913 | 24.53997 | 24.68109 | 24.96892 | 23.63289 | 20.77677 | 30.49078 | 28.70355 | 30.18 |
| 271 | 275 | 276 | 281 | 283 | 287 | 293 | 296 |  |
| 22.90152 | 20.56900 | 25.19240 | 22.49213 | 24.87746 | 19.17896 | 19.40294 | 31.34781 | 25.80 |
| 306 | 307 | 311 | 313 | 314 | 320 | 323 | 324 |  |
| 30.89370 | 28.50046 | 22.32821 | 30.92096 | 28.23923 | 33.11685 | 30.43848 | 31.57115 | 25.17 |

23.43295

515

35

58

33.79732

285

60

86

27.82775

148

87

131

23.43867

106

132

162

26.48394

179

167

190

30.23778

096

192

221

25.59792

660

227

256

20.50940

873

260

298

25.40781

958

300

327

34.81589

493

333 334 338 341 342 345 354 356 357 358

22.03491

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 30.45316 | 26.70026 | 28.91308 | 26.05711 | 27.36421 | 23.63266 | 22.52351 | 25.62410 26.05 |
| 370 | 376 | 377 | 385 | 390 | 395 | 397 | 402 |
| 23.64824 | 21.66569 | 28.24828 | 18.30346 | 29.36638 | 23.63064 | 29.58730 | 26.50543 22.29 |
| 412 | 413 | 414 | 418 | 422 | 423 | 434 | 438 |
| 22.10325 | 27.53845 | 34.42917 | 25.76461 | 26.92570 | 30.90214 | 18.52773 | 19.46554 22.92 |
| 442 | 450 | 451 | 453 | 458 | 467 | 468 | 472 |
| 23.42387 | 21.16347 | 21.79825 | 31.77122 | 16.45741 | 24.08200 | 23.56936 | 26.70337 23.41 |
| 475 | 479 | 484 | 505 | 506 | 511 | 512 | 513 |
| 26.28209 | 25.15806 | 28.06975 | 24.50808 | 25.17685 | 22.99711 | 27.80862 | 34.42505 23.42 |
| 521 | 524 | 529 | 532 | 535 | 536 | 537 | 539 |
| 24.86888 | 25.30833 | 24.23303 | 20.72097 | 32.45241 | 26.49367 | 26.87265 | 24.54116 23.79 |
| 548 | 549 | 551 | 552 | 553 | 555 | 560 |  |
| 27.31330 | 17.41262 | 22.77098 | 28.47826 | 29.52401 | 22.96205 | 20.78536 |  |

709

359

404

18.08821

832

406

440

27.60276

979

441

473

22.18410

027

474

514

21.15475

015

515

540

25.41279

631

543

26.27897

Lets take a example :

### Code :

new\_data <- data.frame(Delivery\_person\_Age = 25, Delivery\_person\_Ratings = 4.5,

Distance = 10,

Type\_of\_order = 3,

Type\_of\_vehicle = 0)

predicted\_time <- predict(model, newdata = new\_data) print(predicted\_time)

### Output :

Predicted time is **27.18212**

## Code :

sqrt(deviance(model) / df.residual(model)) summary(model)$r.squared

## Output and analysis :

The “residual standard error” of the multiple linear regression model is **8.36338**.

This means that the average difference between the observed values and the values predict ed by the model is approximately **8.36338** minutes.

The “coefficient of determination (R-squared)” of the multiple linear regression model is app roximately **0.2532**.This indicates that approximately **25.32%** of the variability in the delivery time is explained by the independent variables included in the model.

# CONCLUSION OF MY PROJECT :

Based on the results of the multiple linear regression model:

The residual standard error of the model is **8.363** minutes, indicating the average difference between observed and predicted delivery times.The R-squared value of the model is approximately  suggesting that approximately  of the variability in delivery time is explained by the independent variables.The adjusted R-squared value, accounting for the number of predictors, is **0.2435.**The F-statistic is **26.18** with a p-value **< 2.2e-16,** indicating the overall model's statistical significance.While the model is statistically significant, it only explains a moderate proportion of the variability in delivery time.

Factors such as delivery person age, ratings, distance, and type of vehicle contribute to the model's predictive power.However, there are likely other variables not included in the model that also influence delivery time.Further research and refinement of the model may be necessary to improve its predictive accuracy.The model provides a foundation for optimizing delivery operations and enhancing service efficiency in the online food delivery industry.Overall, the model's insights can assist online food delivery platforms in providing more accurate delivery time estimates and improving customer satisfaction.

### R code

data <- read.csv("C:/Users/bhama/OneDrive/Desktop/data.csv") View(data2)

calculate\_distance <- function(lat1, lon1, lat2, lon2) {

R <- 6371

lat1 <- lat1 \* pi / 180 lon1 <- lon1 \* pi / 180 lat2 <- lat2 \* pi / 180 lon2 <- lon2 \* pi / 180

dlon <- lon2 - lon1 dlat <- lat2 - lat1

a <- sin(dlat/2)^2 + cos(lat1) \* cos(lat2) \* sin(dlon/2)^2 c <- 2 \* atan2(sqrt(a), sqrt(1-a))

distance <- R \* c

return(distance)

}

# Calculate distance for each sample in the dataset data$Distance <- sapply(1:nrow(data), function(i) {

calculate\_distance(data[i, "restaurant\_latitude"], data[i, "restaurant\_longitude"], data[i, "delivery\_location\_latitude"], data[i, "delivery\_location\_longitude"])

})

str(data2)

summary(data2)

#EDA

library(dplyr) library(ggplot2) library(corrplot) par(mfrow=c(1,1)) par(mar=c(2, 2, 2, 2))

for (i in 3:11) {

hist(data2[,i], main = paste("Histogram of", names(data12)[i]))

}

par(mfrow=c(1,1)) for (i in 3:11) {

boxplot(data2[,i], main = paste("Boxplot of", names(data12)[i]))

}

numeric\_data <- data2[, sapply(data2, is.numeric)]

correlation\_matrix <- cor(data2[, sapply(data2, is.numeric)]) print(correlation\_matrix)

corrplot(correlation\_matrix, method = "circle") set.seed(123) # for reproducibility

train\_index <- sample(1:nrow(data2), 0.7 \* nrow(data2)) train\_data <- data2[train\_index, ]

test\_data <- data2[-train\_index, ]

# Build the multiple linear regression model

model <- lm(Time\_taken.min. ~ Delivery\_person\_Age + Delivery\_person\_Ratings + Distance

+ Type\_of\_order + Type\_of\_vehicle , data = train\_data)

summary(model) plot(model)

sqrt(deviance(model) / df.residual(model)) summary(model)$r.squared

new\_data <- data.frame(Delivery\_person\_Age = 25, Delivery\_person\_Ratings = 4.5,

Distance = 10,

Type\_of\_order = 3,

Type\_of\_vehicle = 0)

predicted\_time <- predict(model, newdata = new\_data) print(predicted\_time)

predictions <- predict(model, newdata = test\_data) predictions