ANALYSIS OF FOOD DELIVERY TIME DATA

Farhan Nugraha Pratama

Project Vision

01. Analyze the factors that affect food Delivery Time

02. Build a delivery time prediction model based on features such as distance, weather, traffic, and courier experience

DATA PREPROCESSING

01

Collected from Kaggle

02

1000 Data with 9 Column

#	Column	Non-Null Count	Dtype
---	------	--------------	-----
0	Order_ID	1000 non-null	int64
1	Distance_km	1000 non-null	float64
2	Weather	1000 non-null	object
3	Traffic_Level	1000 non-null	object
4	Time_of_Day	1000 non-null	object
5	Vehicle_Type	1000 non-null	object
6	Preparation_Time_min	1000 non-null	int64
7	Courier_Experience_yrs	1000 non-null	float64
8	Delivery_Time_min	1000 non-null	int64

03

Check Missing Value and 4 column have 30 missing values (3%)

Missing Values Info for df:

	Missing Values	Percentage
Order_ID	0	0.0
Distance_km	0	0.0
Weather	30	3.0
Traffic_Level	30	3.0
Time_of_Day	30	3.0
Vehicle_Type	0	0.0
Preparation_Time_min	0	0.0
Courier_Experience_yrs	30	3.0
Delivery_Time_min	0	0.0

04

Data cleaning is done with fill, the result is no missing values.

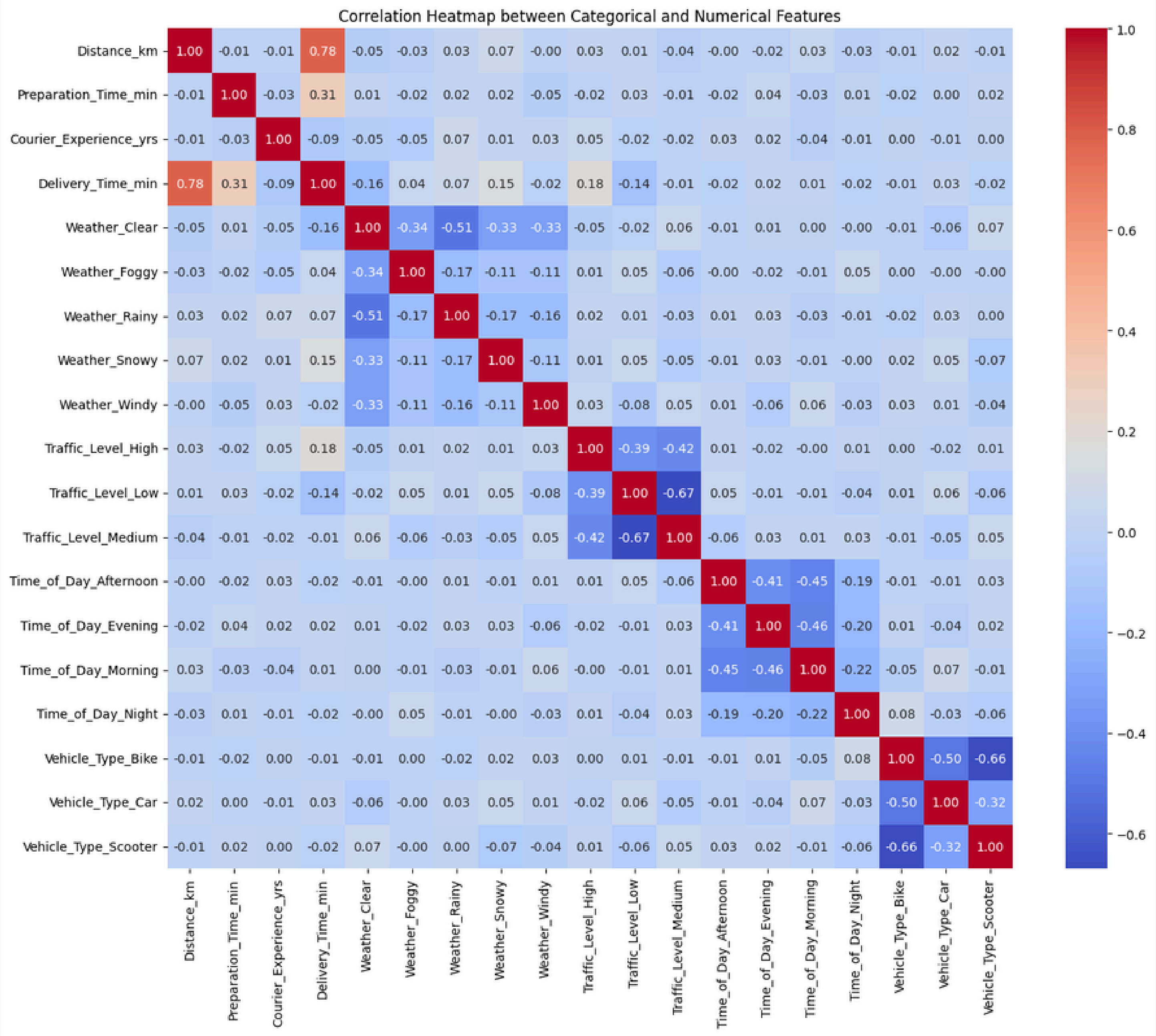
Missing Values Info for df1:

	Missing Values	Percentage
Order_ID	0	0.0
Distance_km	0	0.0
Weather	0	0.0
Traffic_Level	0	0.0
Time_of_Day	0	0.0
Vehicle_Type	0	0.0
Preparation_Time_min	0	0.0
Courier_Experience_yrs	0	0.0
Delivery_Time_min	0	0.0

05

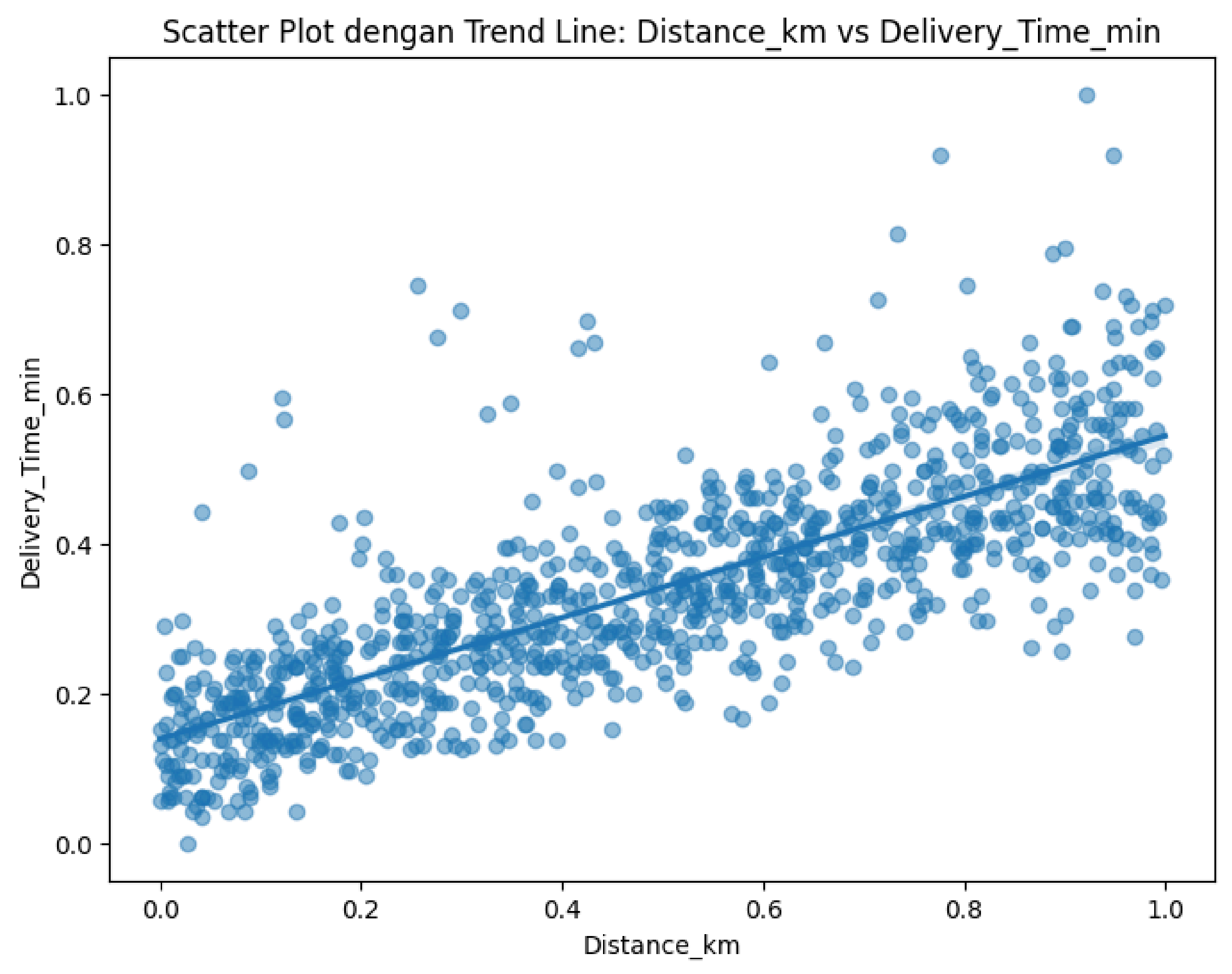
Duplicates Data 1000 Complete with 0 Duplicates Data

EDA: Matrix Analysis



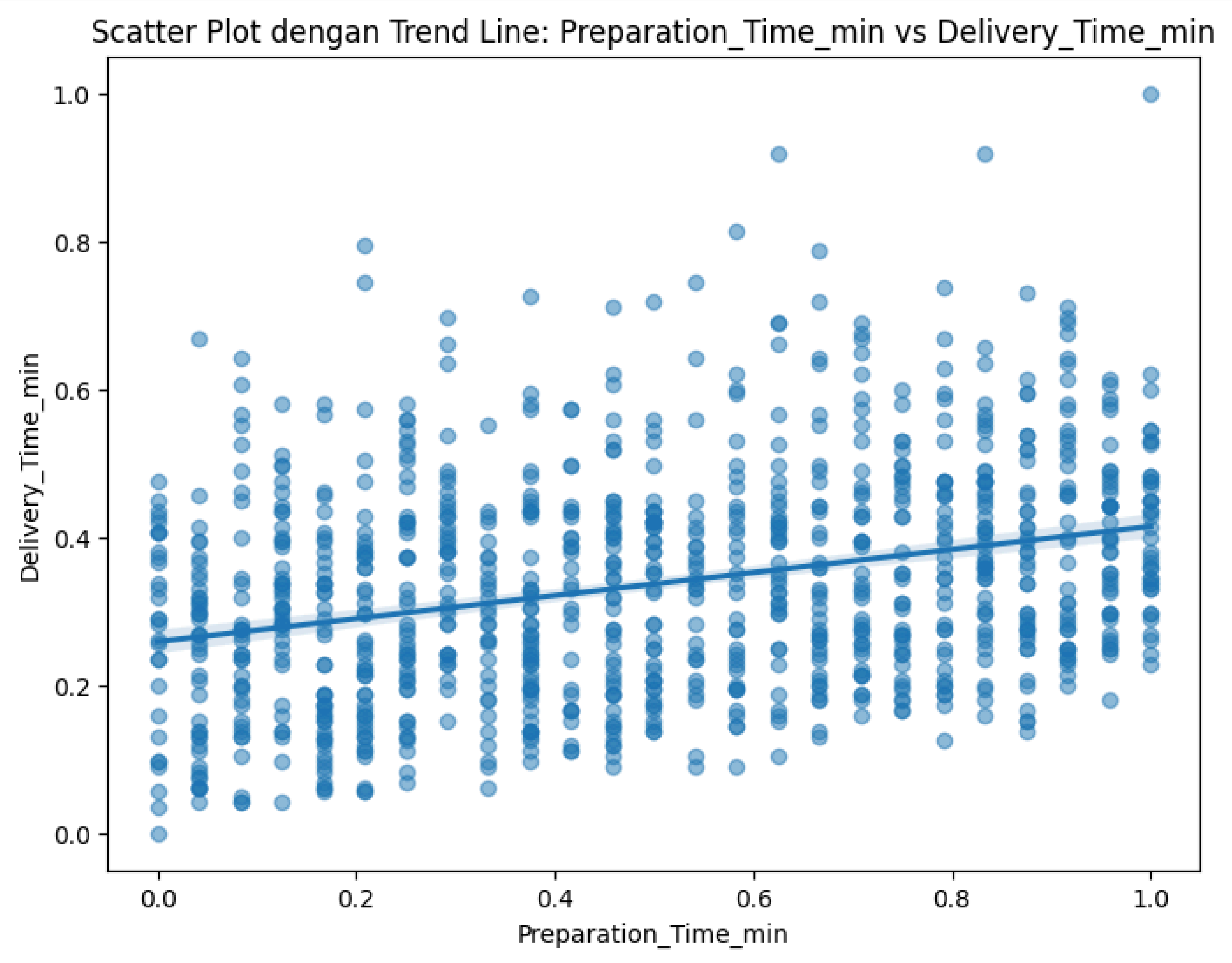
- Correlation values range from -1 (strong negative) to +1 (strong positive)
- Distance and preparation time are the most impactful predictors of delivery time.
- External factors, including traffic and weather, also influence delivery performance, albeit to a lesser extent.

EDA: Distance and Delivery Time



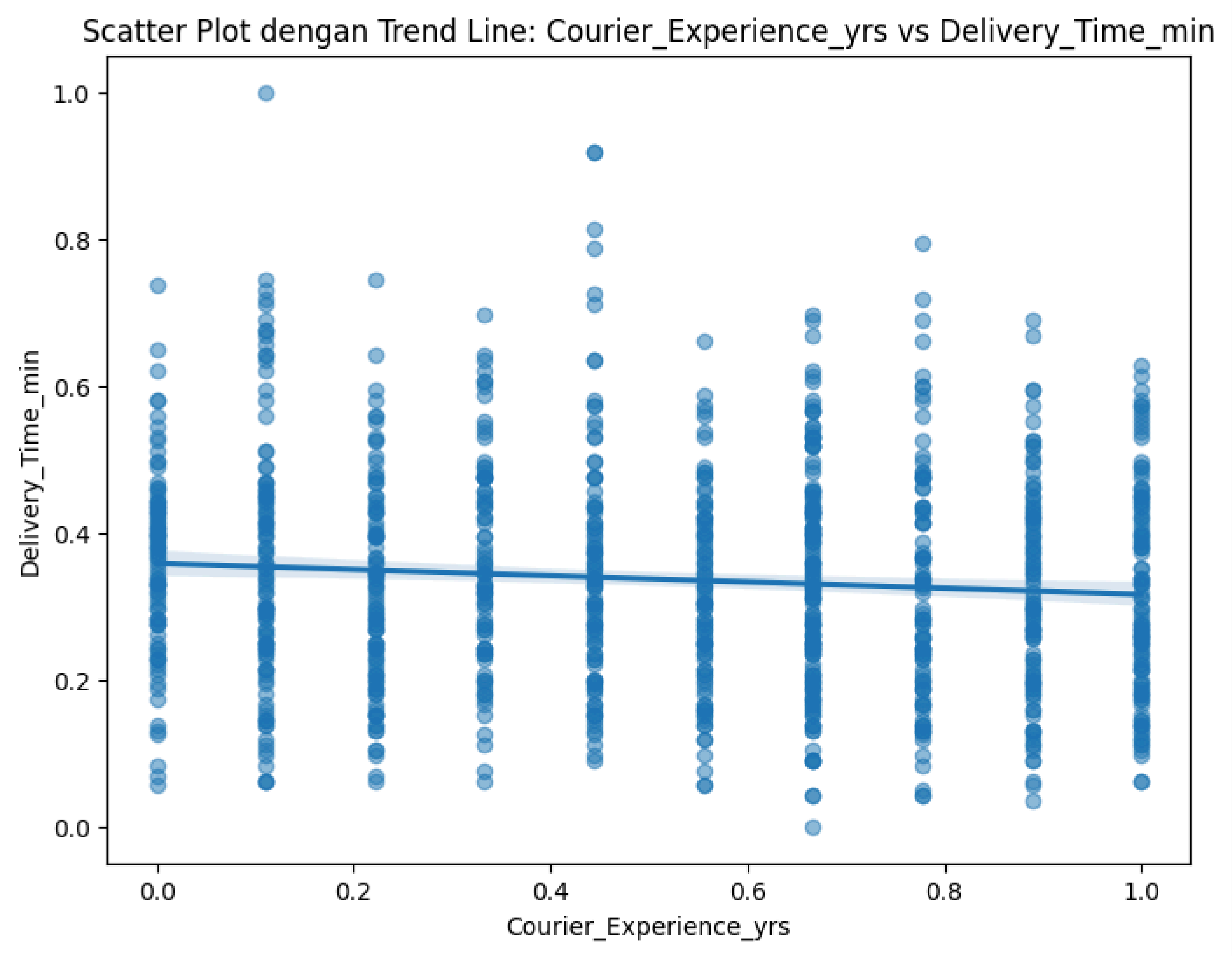
- Distance is a strong and reliable predictor of delivery time.
- A clear positive linear relationship is observed: as the distance increases, the delivery time also tends to increase.

EDA: Preparation Time and Delivery Time



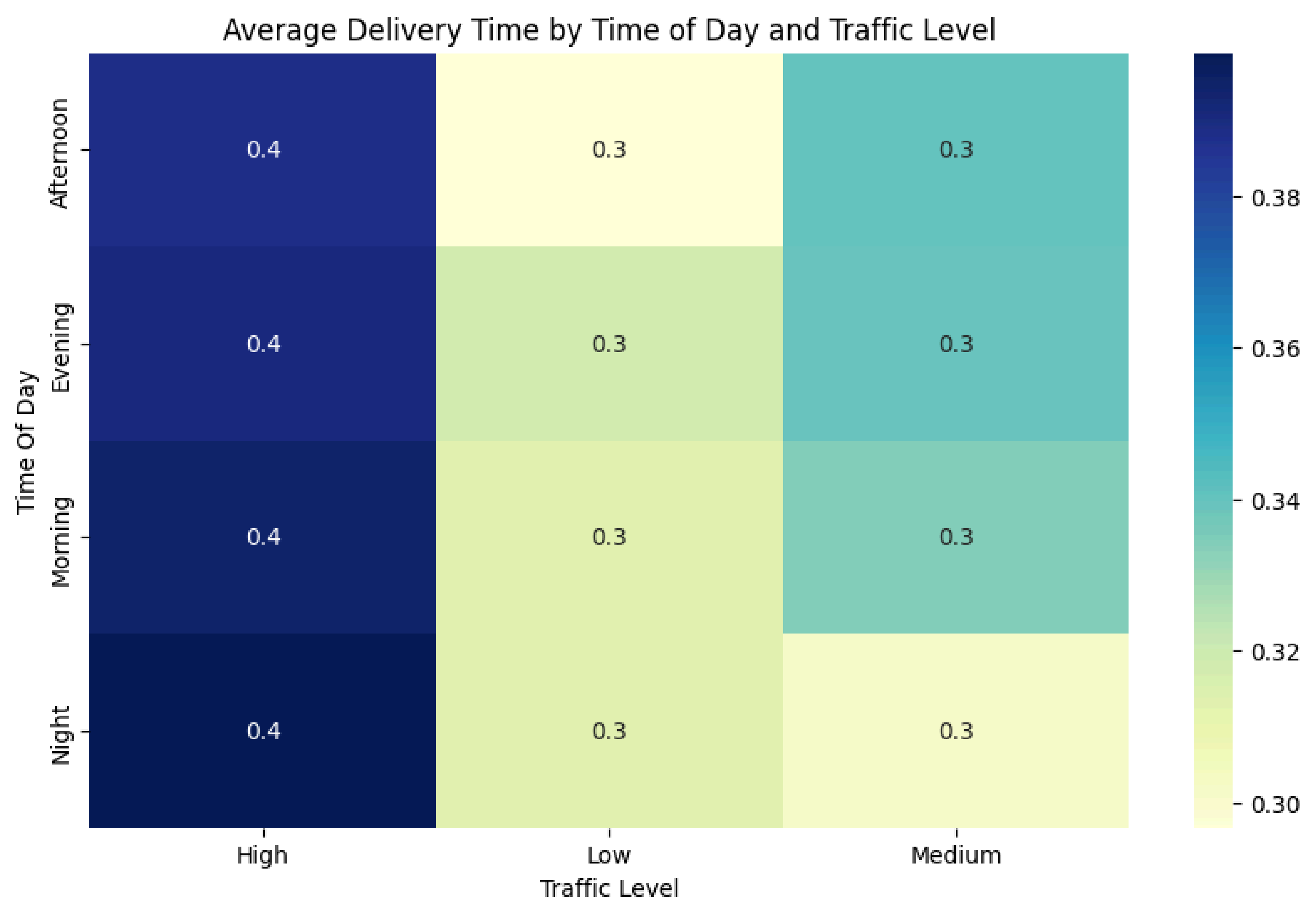
- There is a positive linear trend, indicating that an increase in preparation time tends to correspond with an increase in delivery time.
- However, the correlation appears weak, as the data points are widely scattered around the trend line.
- This suggests that while preparation time may have a slight impact on delivery time, it is not the sole or dominant factor influencing it.

EDA: Courier Experience and Delivery Time



- The trend line shows a slightly decreasing slope, indicating a weak negative relationship between courier experience and delivery time.
- As courier experience increases, delivery time tends to decrease slightly, suggesting that more experienced couriers may deliver slightly faster.

EDA: Average Delivery time from Time of Day and Traffic Level



- Cars have the longest delivery times in the afternoon (~0.4), likely due to being affected by traffic.
- Scooters and bicycles tend to have more consistent and faster delivery times at all times.
- Evenings show the relatively fastest and most stable delivery times, regardless of vehicle type.

MACHINE LEARNING

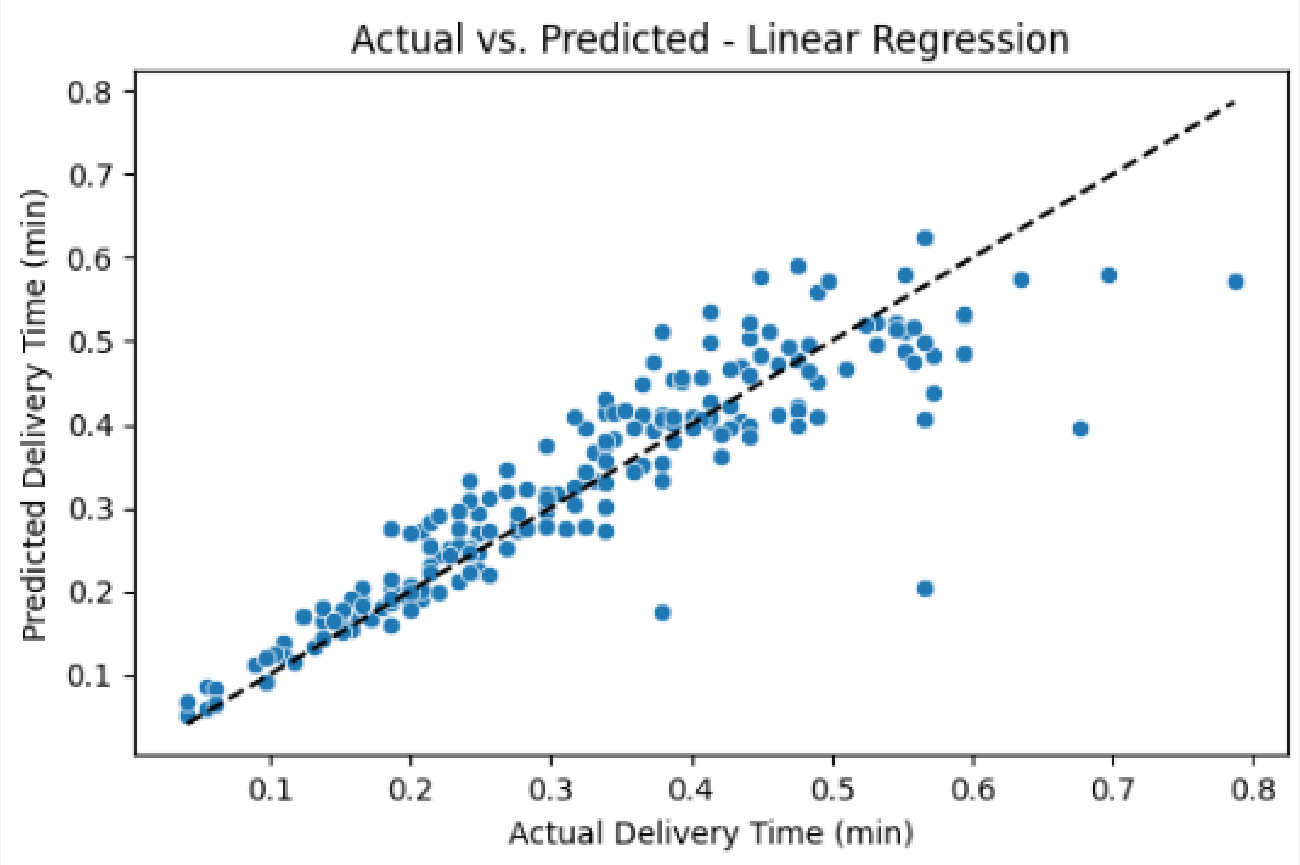
Model Selection

- Linear Regression
- Decision Tree
- Random Forest
- XGBoost

Model Linear Regression

```
Model: Linear Regression
  Mean Absolute Error (MAE): 0.04
  Mean Squared Error (MSE): 0.00
  Root Mean Squared Error (RMSE): 0.06
  R-squared (R2): 0.83
```

- Mean Absolute Error (MAE): The average prediction error is only 0.04.
- Root Mean Squared Error (RMSE): Indicates predictions are fairly consistent, with little deviation.
- R-squared (R^2): The model explains 83% of the variation in the data. This indicates the model is very accurate and reliable.

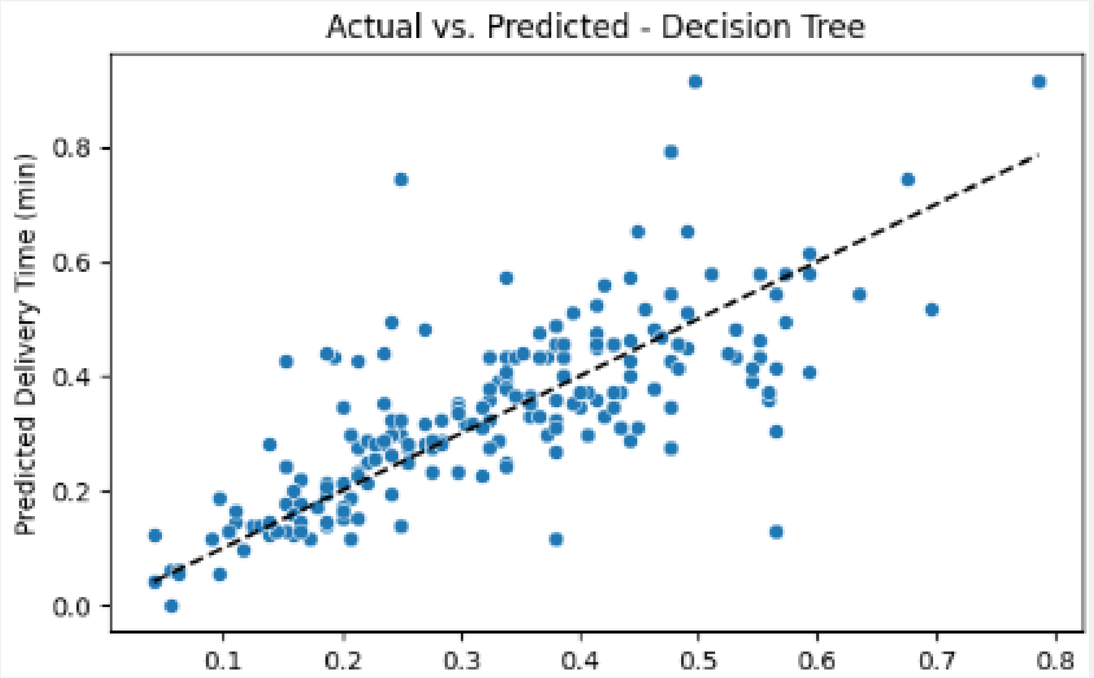


The Linear Regression model is able to predict the delivery time with high accuracy and small error.

Model Decision Tree

```
Model: Decision Tree
  Mean Absolute Error (MAE): 0.07
  Mean Squared Error (MSE): 0.01
  Root Mean Squared Error (RMSE): 0.11
  R-squared (R2): 0.46
```

- Mean Absolute Error (MAE): Average prediction error of 0.07 minutes (about 4 seconds).
- Root Mean Squared Error (RMSE): There is a larger deviation in the prediction compared to other models.
- R-squared (R^2): The model only explained 46% of the variation in the data, indicating low prediction accuracy.

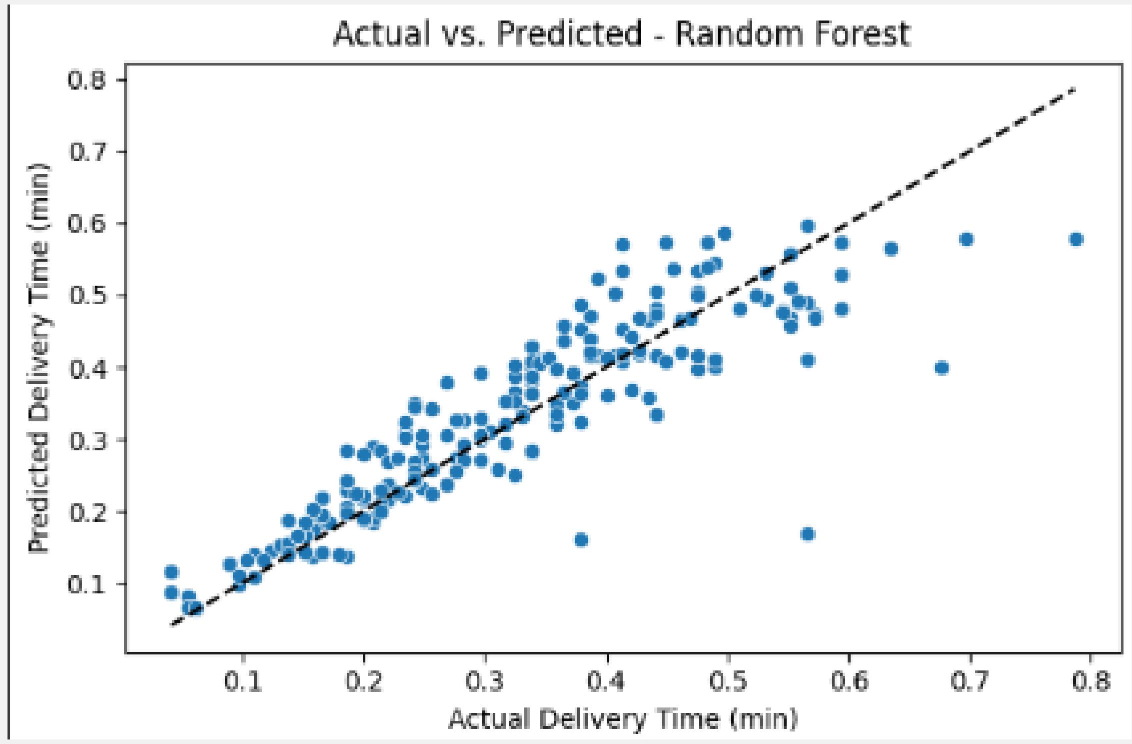


The ability of the model to explain the data is very limited

Model Random Forest

```
Model: Random Forest
  Mean Absolute Error (MAE): 0.05
  Mean Squared Error (MSE): 0.00
  Root Mean Squared Error (RMSE): 0.07
  R-squared (R2): 0.79
```

- Mean Absolute Error (MAE): The average prediction error is small, only 0.05 minutes.
- Root Mean Squared Error (RMSE): The predictions are consistent, with small deviations from the actual values.
- R-squared (R^2): The model explained 79% of the variation in the data. This indicates a fairly high accuracy.

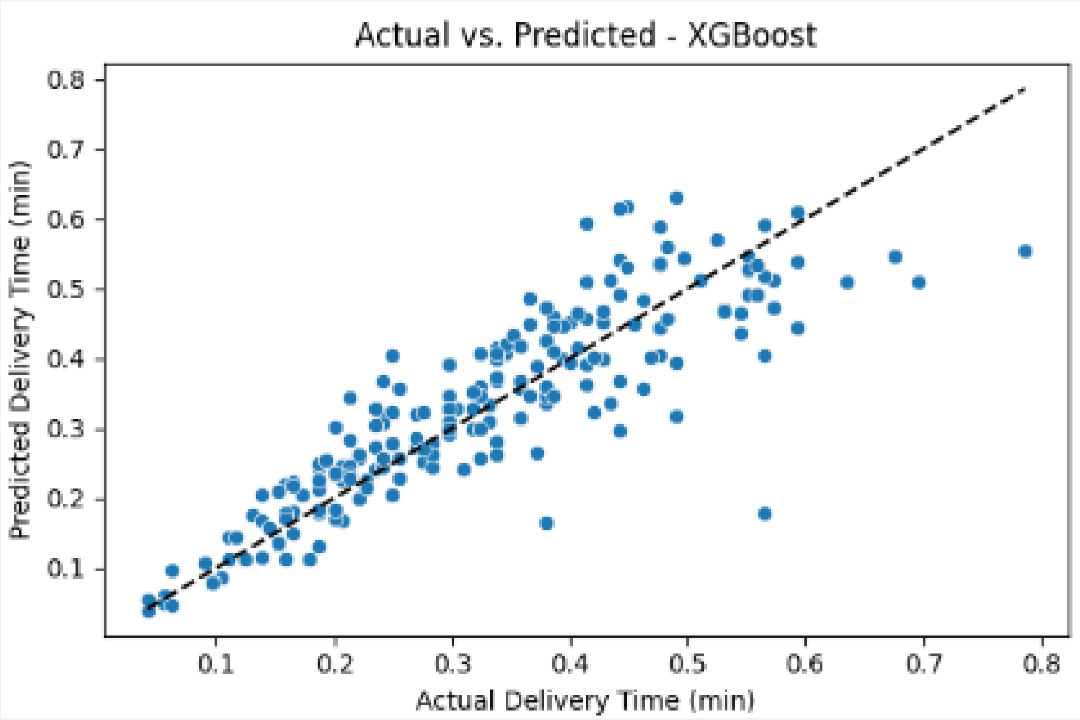


The Random Forest model is able to predict the delivery time well, with a consistent data distribution and only a small deviation.

Model XGBoost

```
Model: XGBoost
  Mean Absolute Error (MAE): 0.05
  Mean Squared Error (MSE): 0.01
  Root Mean Squared Error (RMSE): 0.07
  R-squared (R2): 0.75
```

- Mean Absolute Error (MAE): The average prediction error is low.
- Root Mean Squared Error (RMSE): Shows fairly consistent predictions, although there is a slight deviation.
- R-squared (R^2): The model explains 75% of the variation in the actual data, which is good enough but still below Linear Regression.



XGBoost is able to make fairly accurate and stable predictions, although there are still some deviations, especially at extreme values.

Model Conclusion

Model	MAE	MSE	RMSE	R^2	Conclusion
Linear Regressi on	0.04	0.0	0.06	0.83	Best accuracy and highly consistent
Decision Tree	0.07	0.01	0.11	0.46	Low performance, prone to overfitting
Random Forest	0.05	0.00	0.07	0.79	Accurate and stable, close to linear regression
XGBoost	0.05	0.01	0.07	0.75	Good and stable, slightly less than RF

- Linear Regression gives the best performance overall (highest R^2, lowest MAE & RMSE).
- Random Forest and XGBoost are strong alternatives with good balance.
- Decision Tree is the weakest model due to lower accuracy and higher error.

Model Conclusion (Tuned)

Model	MAE	MSE	RMSE	R^2	Conclusion
Decision Tree	0.06	0.01	0.08	0.67	Low performance and tendency to overfitting on training data.
Random Forest	0.05	0.00	0.07	0.79	Accurate and stable, close to the best linear regression performance. Suitable for use in prediction.
XGBoost	0.05	0.01	0.07	0.80	Good and consistent performance, even slightly superior to Random Forest in terms of R^2.

- XGBoost gives the best results with an R^2 of 0.80 and the lowest MAE/MSE.
- Random Forest is a strong alternative with almost equivalent performance.
- Decision Tree should be avoided for the final prediction due to overfitting.

Tuning VS Non-Tuning Best Model (Linear Regression)

Model	MAE	MSE	RMSE	R^2	Kesimpulan
GridSearch CV	0.05	0.00	0.07	0.79	High and consistent accuracy, suitable for final implementation.
Lasso	0.04	0.00	0.06	0.83	Performance is excellent and efficient, suitable for data with possibly non-essential features.
Ridge	0.04	0.00	0.06	0.82	Stable and robust, almost equivalent to Lasso, slightly lower in R^2.
Non-Tuning	0.04	0.00	0.06	0.83	Although the evaluation results are quite good, there is a risk of overfitting without further tuning.

- Lasso and Ridge showed improved accuracy thanks to regularization (L1 and L2).
- Tuning (GridSearchCV) helps select the best parameters for a more stable model.
- Models without tuning, despite high scores, tend to overfit the training data.

Recommendation

MODEL

- Use Linear Regression when simplicity and high accuracy are priorities.
- Consider Random Forest or XGBoost if robustness is more important, especially for noisy data.
- Always prefer tuned models for better generalization.

DELIVERY TIME

- Optimize Delivery Management
- Optimize Preparation Process
- Improve Courier Dispatching
- Use Real-Time Prediction
- Reorganize High-Demand Areas
- Build a Monitoring Dashboard

Thank you