Traffic Flow Prediction in the City of Bandung

Data Science Mini Project

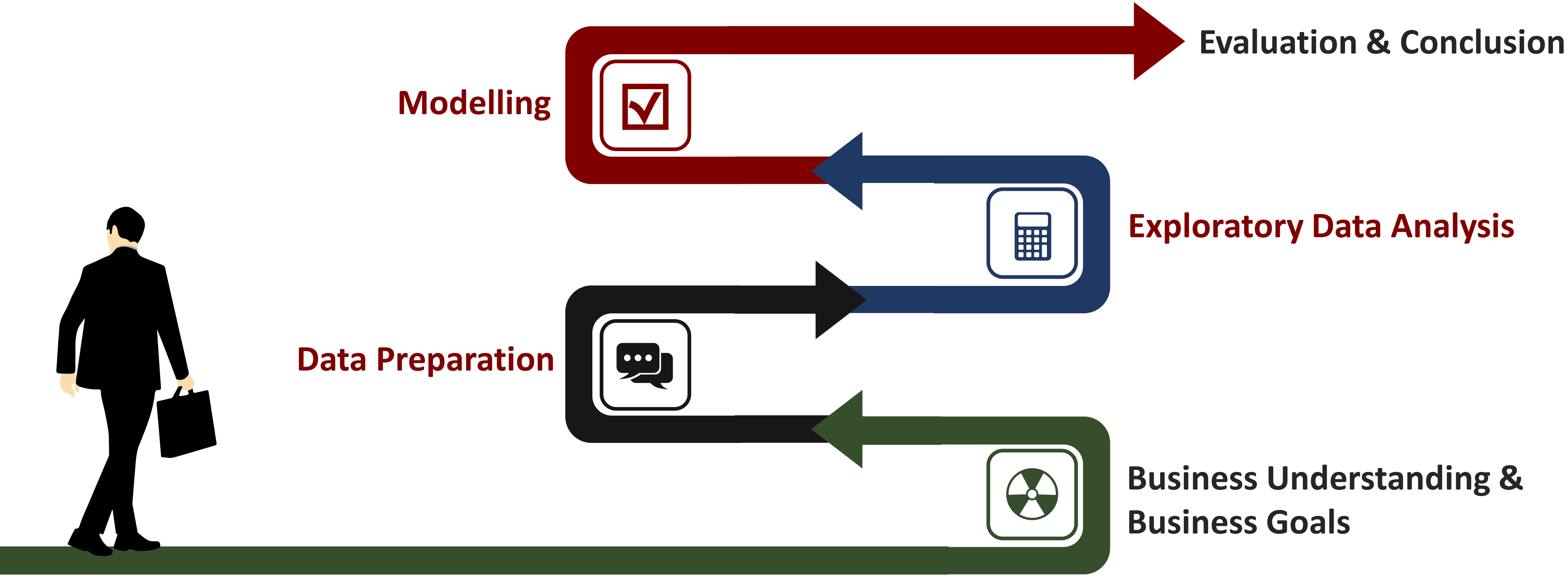
Data Consultant Bootcamp 2023

Microsoft

DATA SCIENCE INDONESIA

Data
Science
Learning
Studio

Outline



Evaluation & Conclusion

Modelling

Exploratory Data Analysis

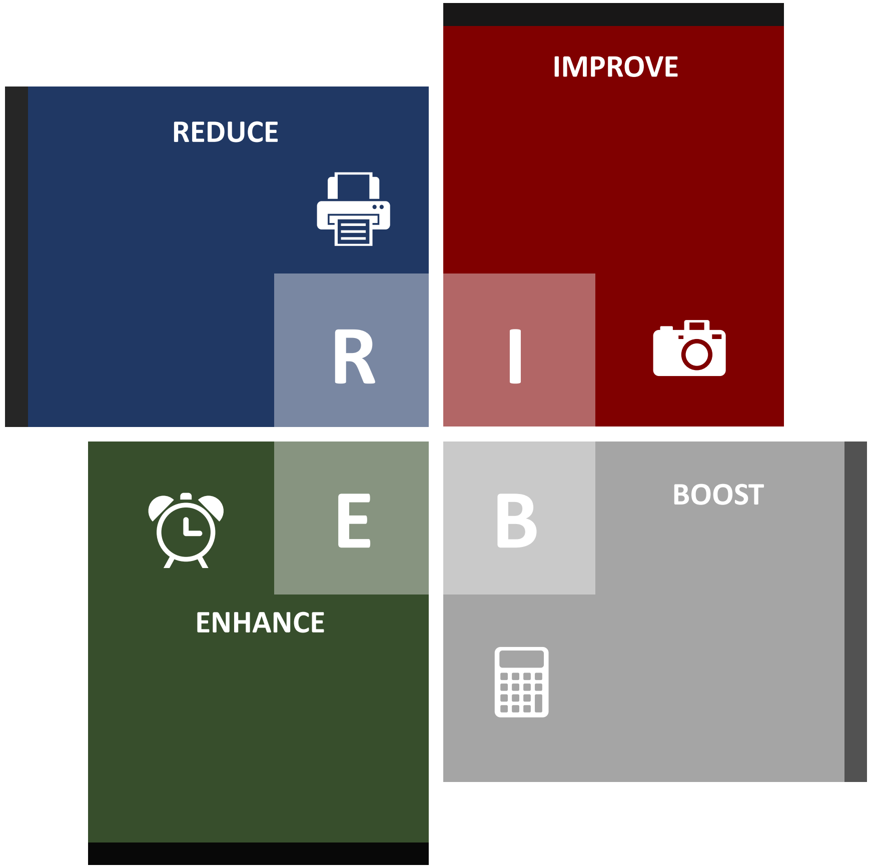
Data Preparation

Business Understanding &
Business Goals

Business **Understanding**

The city of Bandung is a highly populated area with high levels of **traffic congestion**. This congestion leads to various problems such as increased travel time, increased fuel consumption, and increased air pollution. This can have negative impacts on the local economy, quality of life, and public health. To address these issues, the local government has proposed a project to **predict traffic flow** patterns in the city.

Business **Goals**



Reduce traffic congestion

By **predicting traffic flow** patterns, the local government can identify areas with high congestion and develop strategies to alleviate it.

Improve transportation efficiency

With **accurate traffic flow predictions**, transportation authorities can better plan routes and schedules for public transportation systems.

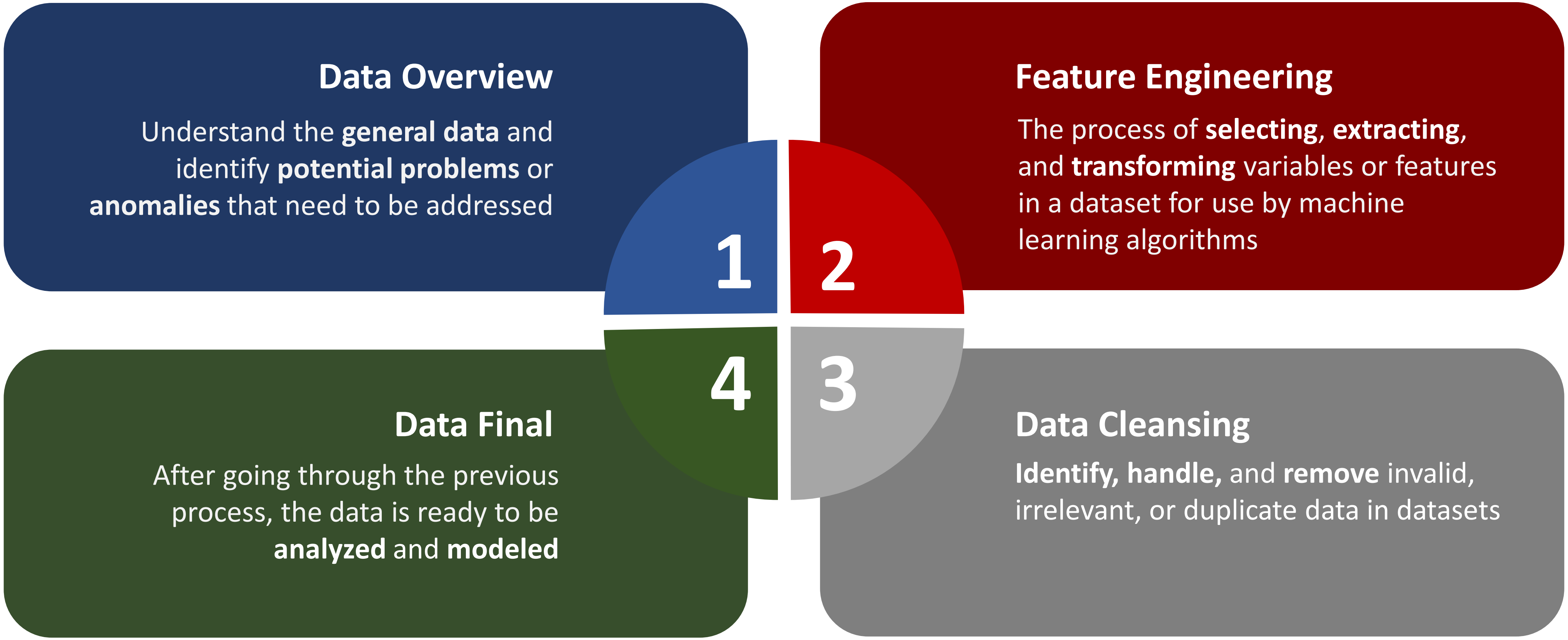
Enhance road safety

By **predicting traffic flow**, authorities can take proactive measures to prevent accidents and improve road safety.

Boost the local economy

Traffic congestion can have negative impacts on local businesses. By **improving traffic flow**, this project can increase the efficiency of the transportation network and boost economic activity in the city.

Data Preparation



1 Data Overview

Understand the **general data** and identify **potential problems** or **anomalies** that need to be addressed

2 Feature Engineering

The process of **selecting**, **extracting**, and **transforming** variables or features in a dataset for use by machine learning algorithms

3 Data Cleansing

Identify, handle, and **remove** invalid, irrelevant, or duplicate data in datasets

4 Data Final

After going through the previous process, the data is ready to be **analyzed** and **modeled**

Data **Overview**

301995 rows

Average of Median Length
722.78

62 days & **1419** hours

Average of Median Delay
132.15

2022,6 July until **2022, 6 Sep**

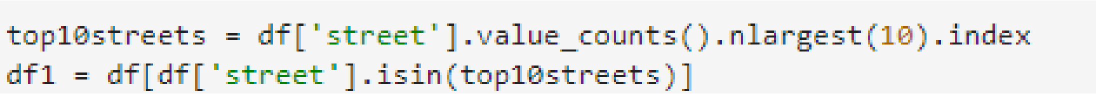
Average of Median Speed (kmh)
12.24

1161 street & **5** level

Average of Total Records
21.34

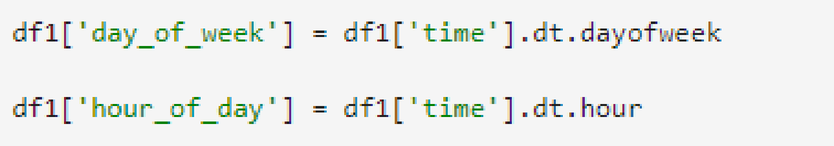
Feature Engineering

add **street category** variable column



```
top10streets = df['street'].value_counts().nlargest(10).index
df1 = df[df['street'].isin(top10streets)]
```

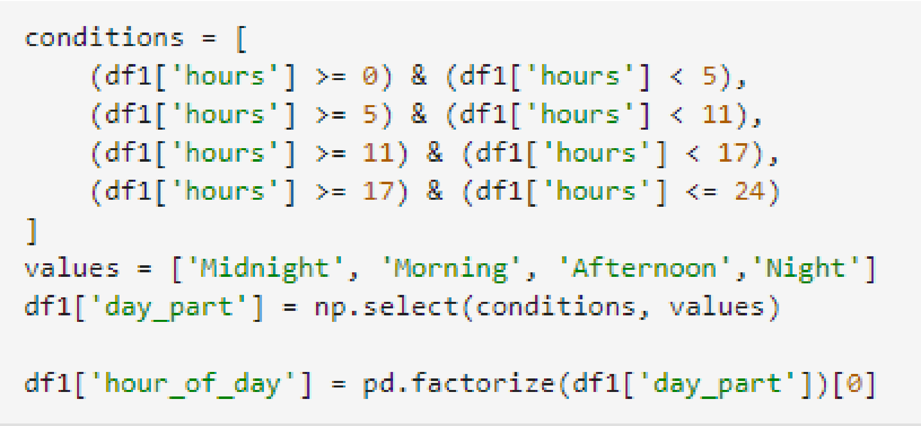
add **day of week** & **hour of day** variable column



```
df1['day_of_week'] = df1['time'].dt.dayofweek

df1['hour_of_day'] = df1['time'].dt.hour
```

add **day of part** & **change hour of day** variable column

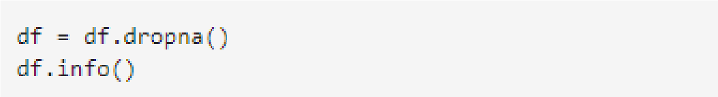


```
conditions = [
    (df1['hours'] >= 0) & (df1['hours'] < 5),
    (df1['hours'] >= 5) & (df1['hours'] < 11),
    (df1['hours'] >= 11) & (df1['hours'] < 17),
    (df1['hours'] >= 17) & (df1['hours'] <= 24)
]
values = ['Midnight', 'Morning', 'Afternoon','Night']
df1['day_part'] = np.select(conditions, values)

df1['hour_of_day'] = pd.factorize(df1['day_part'])[0]
```

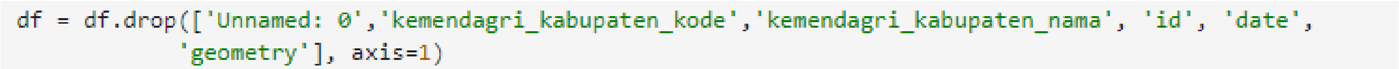
Data **Cleansing**

delete **rows** containing null values



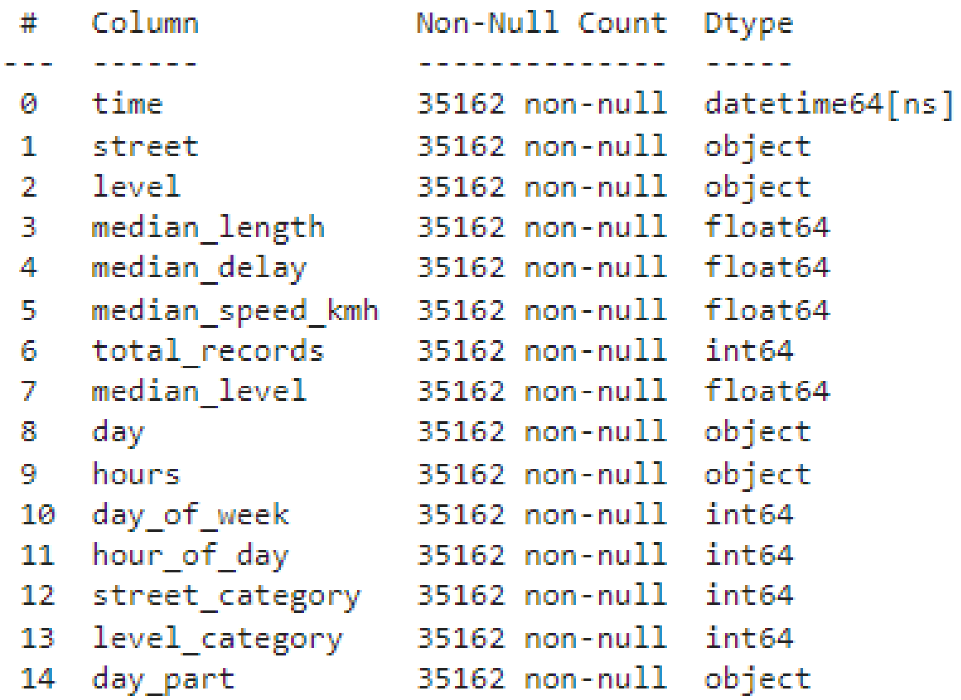
```
df = df.dropna()
df.info()
```

delete **variable column** that useless



```
df = df.drop(['Unnamed: 0','kemendagri_kabupaten_kode','kemendagri_kabupaten_nama', 'id', 'date',
            'geometry'], axis=1)
```

Data **Final**

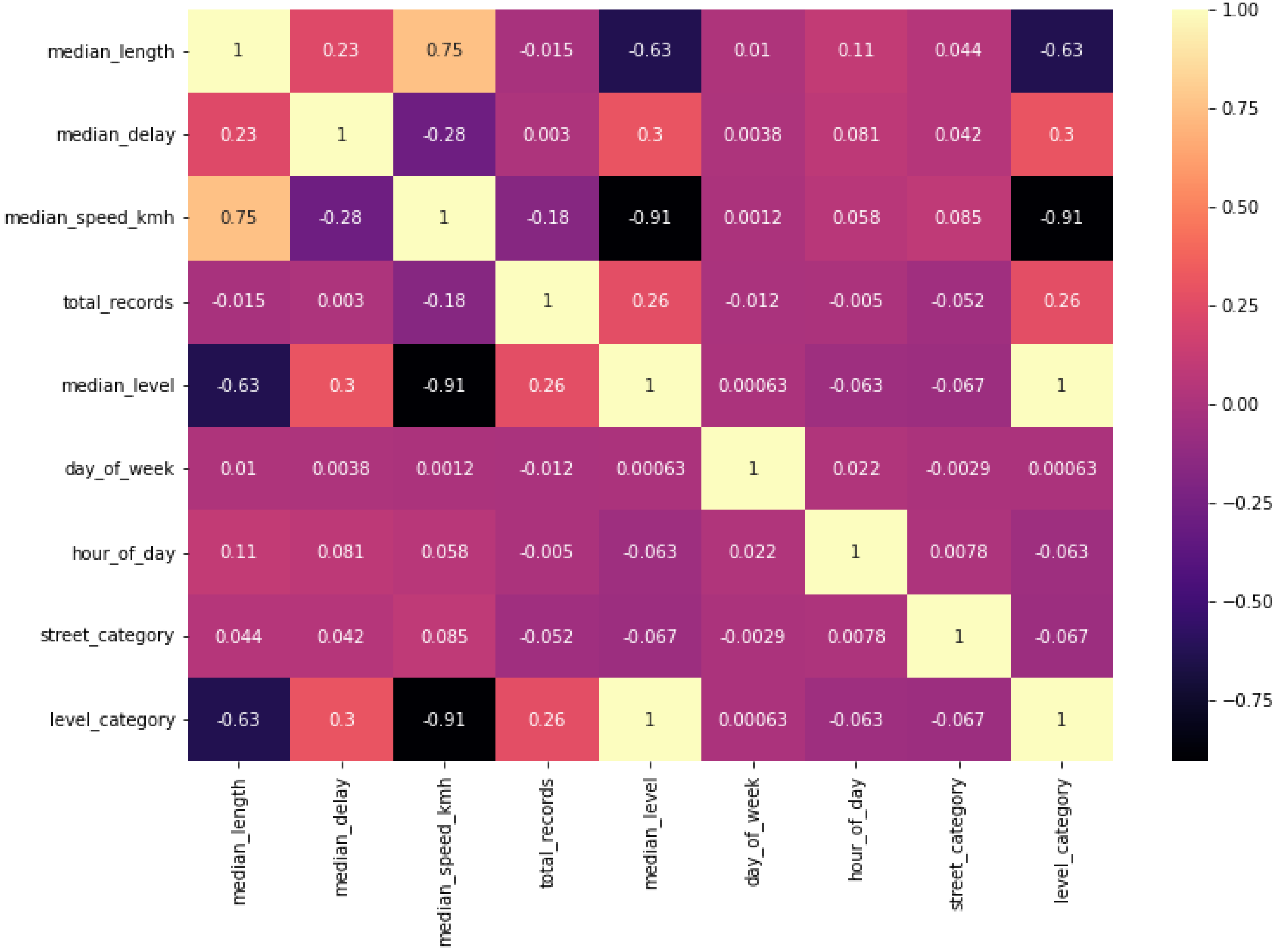


```
 #   Column            Non-Null Count  Dtype         
---  ------            --------------  -----         
 0   time              35162 non-null  datetime64[ns]
 1   street            35162 non-null  object        
 2   level             35162 non-null  object        
 3   median_length     35162 non-null  float64       
 4   median_delay      35162 non-null  float64       
 5   median_speed_kmh  35162 non-null  float64       
 6   total_records     35162 non-null  int64         
 7   median_level      35162 non-null  float64       
 8   day               35162 non-null  object        
 9   hours             35162 non-null  object        
 10  day_of_week       35162 non-null  int64         
 11  hour_of_day       35162 non-null  int64         
 12  street_category   35162 non-null  int64         
 13  level_category    35162 non-null  int64         
 14  day_part          35162 non-null  object        
```

The data frame that has additional information will later be used for data modeling in **predicting traffic flow** in the city of Bandung. The target variable used is **Median Speed** (kmh)

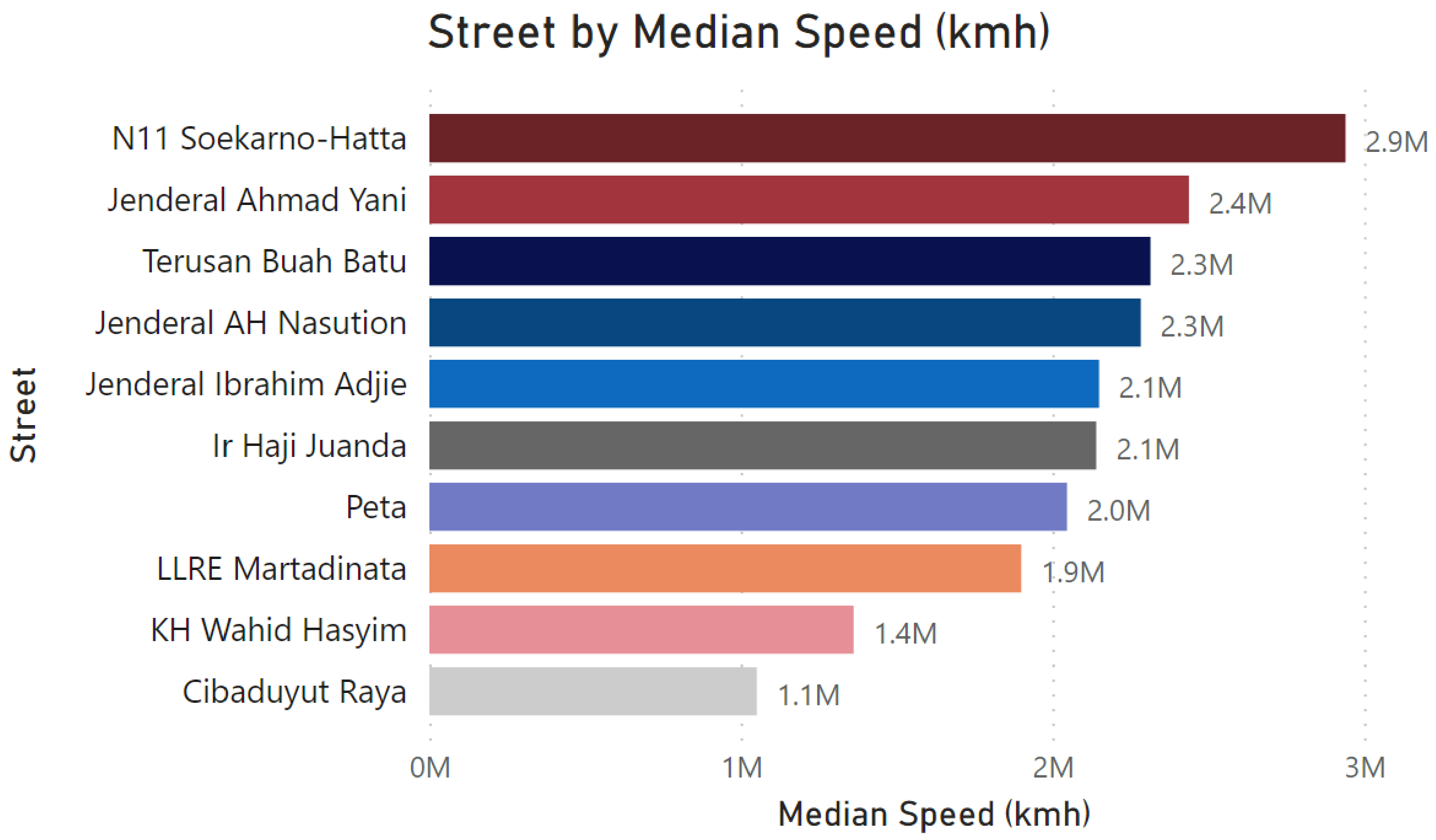
Category Level has a Strong Correlation with the **Median Speed** (kmh)

The **level category** has the strongest correlation to the **median speed**, which is equal to **0.91**. After that the **median length** is **0.75** and then the **median delay** is **0.28**

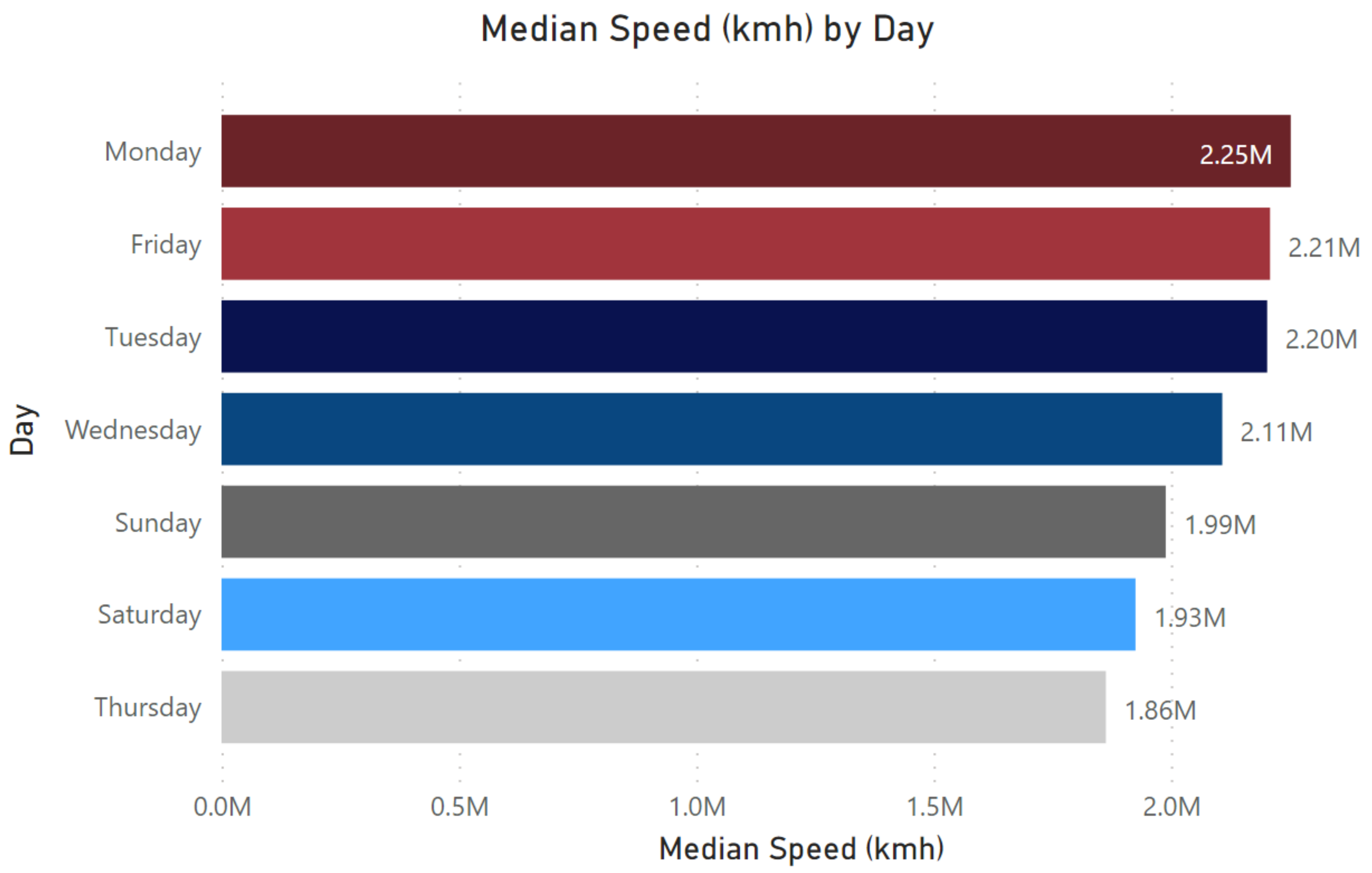


	median_length	median_delay	median_speed_kmh	total_records	median_level	day_of_week	hour_of_day	street_category	level_category
median_length	1	0.23	0.75	-0.015	-0.63	0.01	0.11	0.044	-0.63
median_delay	0.23	1	-0.28	0.003	0.3	0.0038	0.081	0.042	0.3
median_speed_kmh	0.75	-0.28	1	-0.18	-0.91	0.0012	0.058	0.085	-0.91
total_records	-0.015	0.003	-0.18	1	0.26	-0.012	-0.005	-0.052	0.26
median_level	-0.63	0.3	-0.91	0.26	1	0.00063	-0.063	-0.067	1
day_of_week	0.01	0.0038	0.0012	-0.012	0.00063	1	0.022	-0.0029	0.00063
hour_of_day	0.11	0.081	0.058	-0.005	-0.063	0.022	1	0.0078	-0.063
street_category	0.044	0.042	0.085	-0.052	-0.067	-0.0029	0.0078	1	-0.067
level_category	-0.63	0.3	-0.91	0.26	1	0.00063	-0.063	-0.067	1

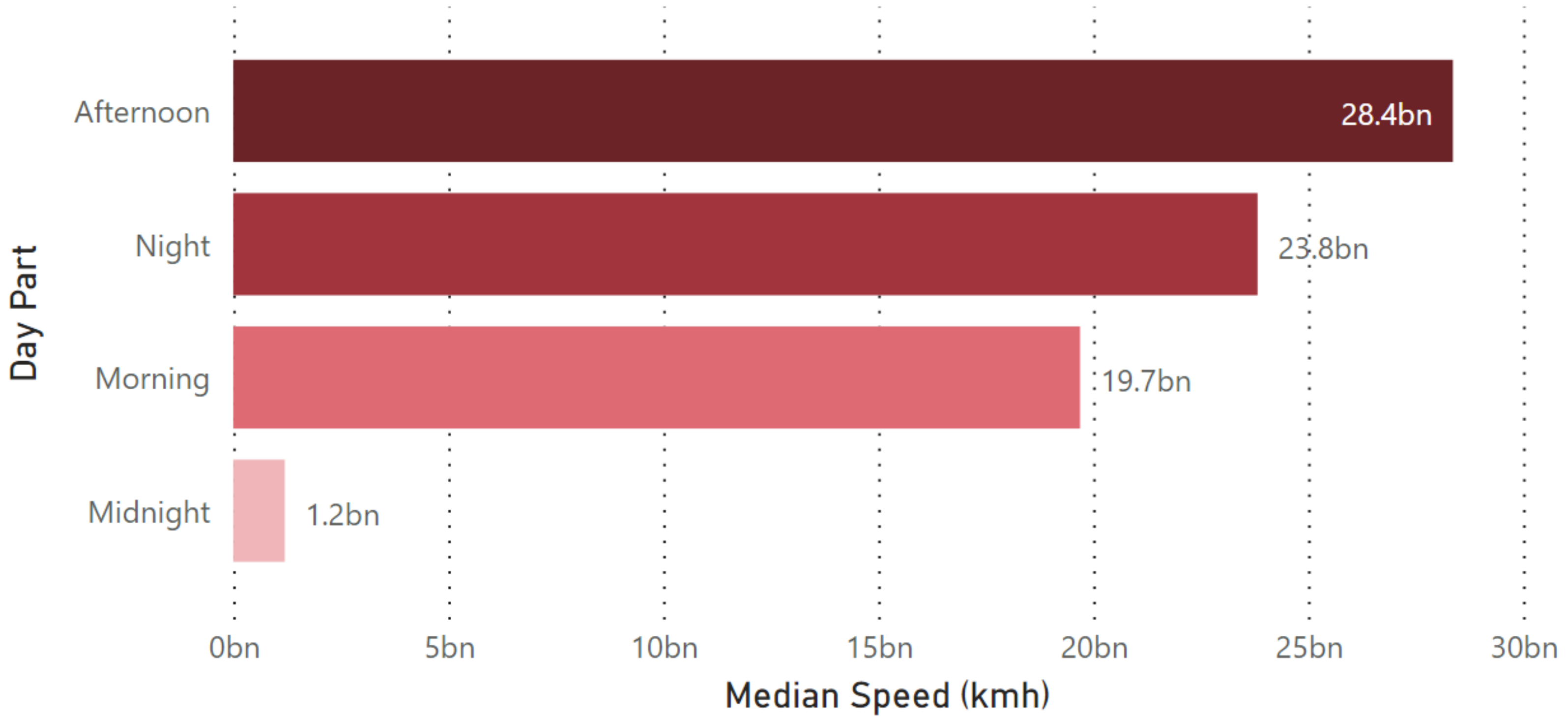
N11 Soekarno Hatta Street has the Highest Average Median Speed of the 10 Street Categories that have the Most Data



Every Monday the Traffic on the Street Experiences High Speed



Traffic Flow Speed is Highest in the Afternoon & the Low Flow Speed in the Midnight

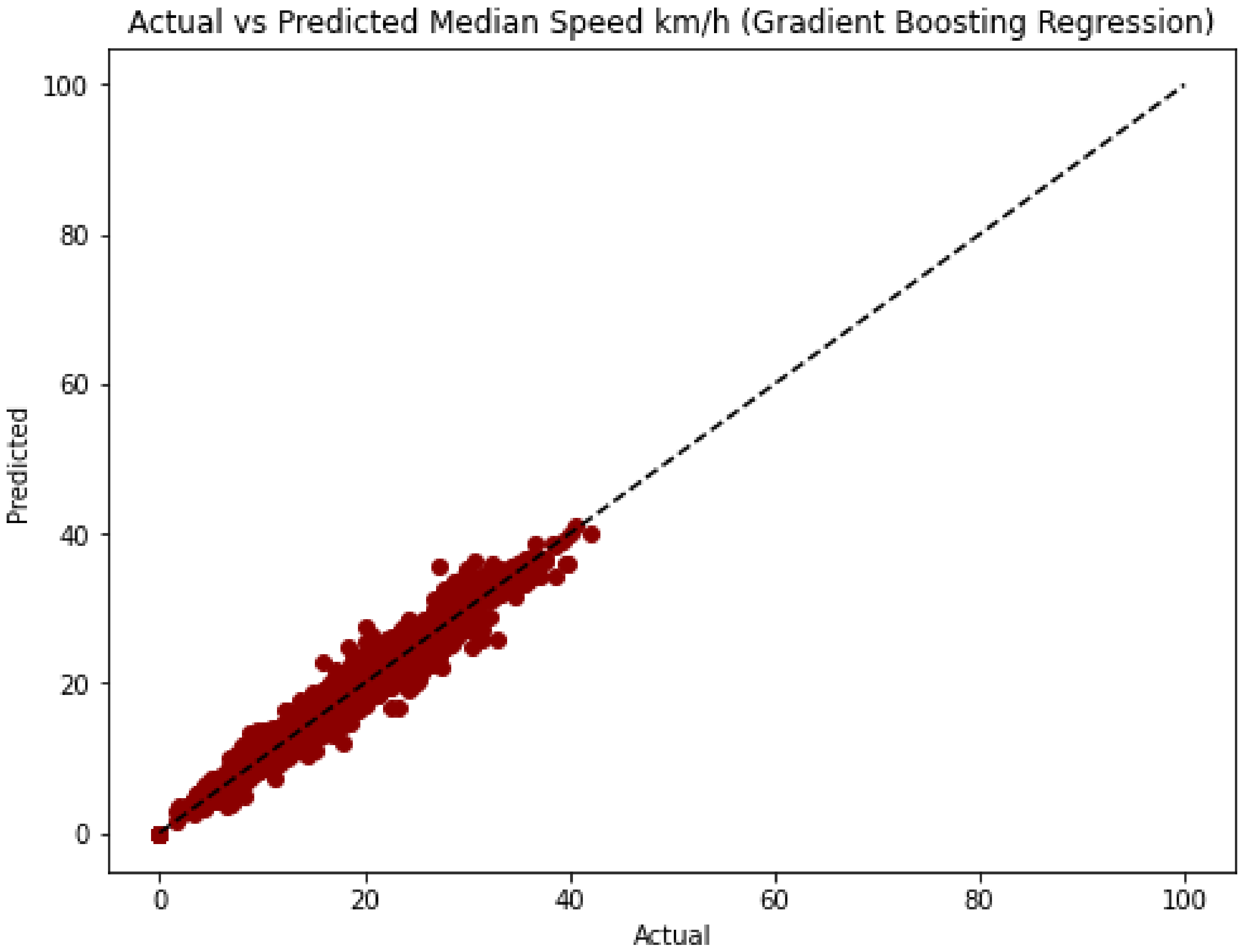


Gboost became the Best Model for Traffic Flow Prediction based on Accuracy and MSE value

Model	Mean Squared Error (MSE)	Accuracy
Linear Regression	5.794	0.906
Polynomial Regression	2.925	0.952
Decision Tree	1.781	0.971
Random Forest	1.366	0.977
Gradient Boosting	**1.024**	**0.983**

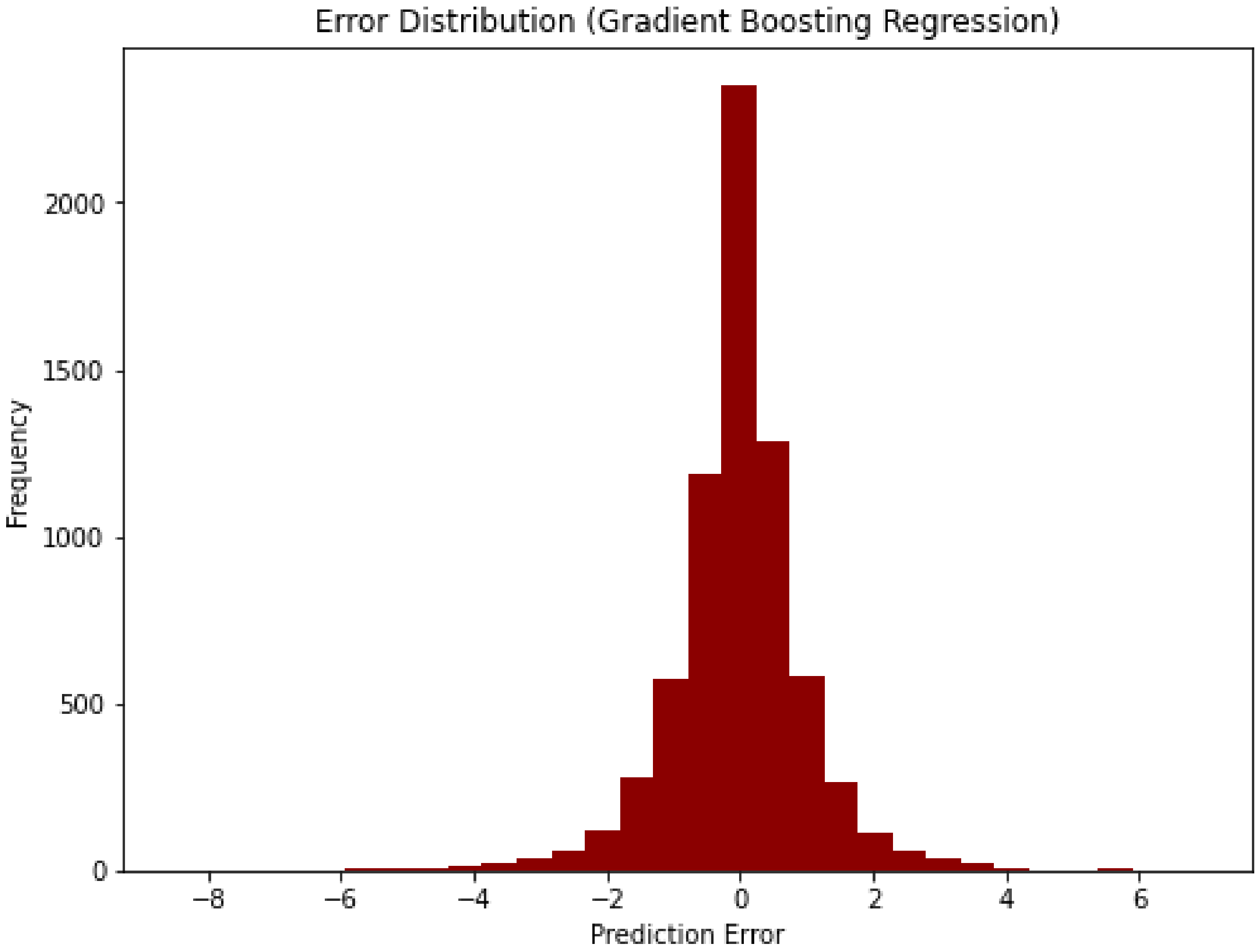
Based on the results of the hyperparameter turning, it can be concluded that the prediction of traffic flow uses the **Gradient Boosting** model

Positive Correlation between the **Predicted Outcome** and the **Target Variable**



Based on the visual correlation, the model can **better estimate** the target value than random guesses

The **Error** Distribution of the Model is a **Normal Distribution**



Error values are evenly distributed around the **average** error value. This shows that the model has a **consistent** level of accuracy in predicting target values

Cross-validation

R Square	Score
Mean	0.9753
Standard Deviation	0.0024

Conclusion

Based on the table, the average **r2 score** across all folds was **0.9753**, indicating that the **Gradient Boosting Regression model** is fairly accurate in predicting the target values in the dataset. Additionally, the **standard deviation** of the **r2 scores** across the folds was **0.0024**, indicating that the cross-validation results are relatively stable. Therefore, it can be concluded that the **Gradient Boosting Regression model** is suitable for predicting the target values

Hendra Kuswantoro

Thank You!

Feedback or suggestions are welcome