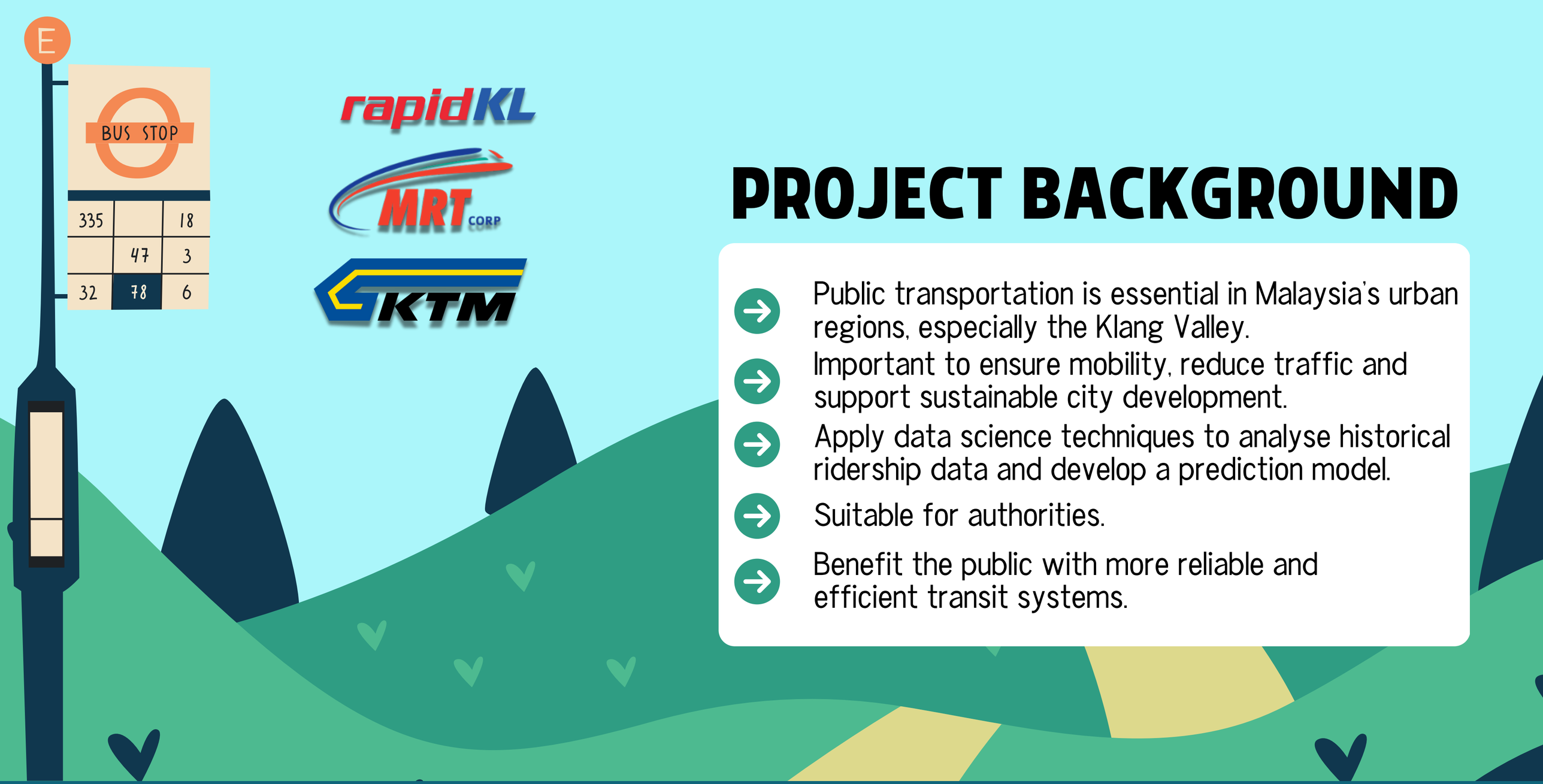
DAILY PUBLIC TRANSPORT RIDERSHIP PREDICTIONS USING MACHINE LEARNING

WIE2003 INTRODUCTION TO DATA SCIENCE

Group 6

- Amirul Hisyam Bin Amirruddin (23002578)
- Imran Haziq Bin Khairul Anuar (23001784)
- Mohamad Aiman Bin Ibhrahim (23001926)
- Mohamad Izzulqhaleeq Bin Mohd Faris (23002413)
- Wan Muhammad Hazwan Bin Wan Rozli (23002102)



PROJECT BACKGROUND

- Public transportation is essential in Malaysia's urban regions, especially the Klang Valley.
- Important to ensure mobility, reduce traffic and support sustainable city development.
- Apply data science techniques to analyse historical ridership data and develop a prediction model.
- Suitable for authorities.
- Benefit the public with more reliable and efficient transit systems.

PROBLEM STATEMENT

- Daily ridership varies greatly and is difficult to manage with static schedules.
- Many agencies lack automated forecasting systems despite having the data.
- Consequences of no predictive planning:
  - Under–servicing during peaks.
  - Wasted resources during off–peak periods.

PROJECT OBJECTIVES

Collect and prepare accurate historical ridership data from Malaysia's urban transport systems.

Develop a prediction system to estimate daily ridership numbers based on trends and historical patterns.

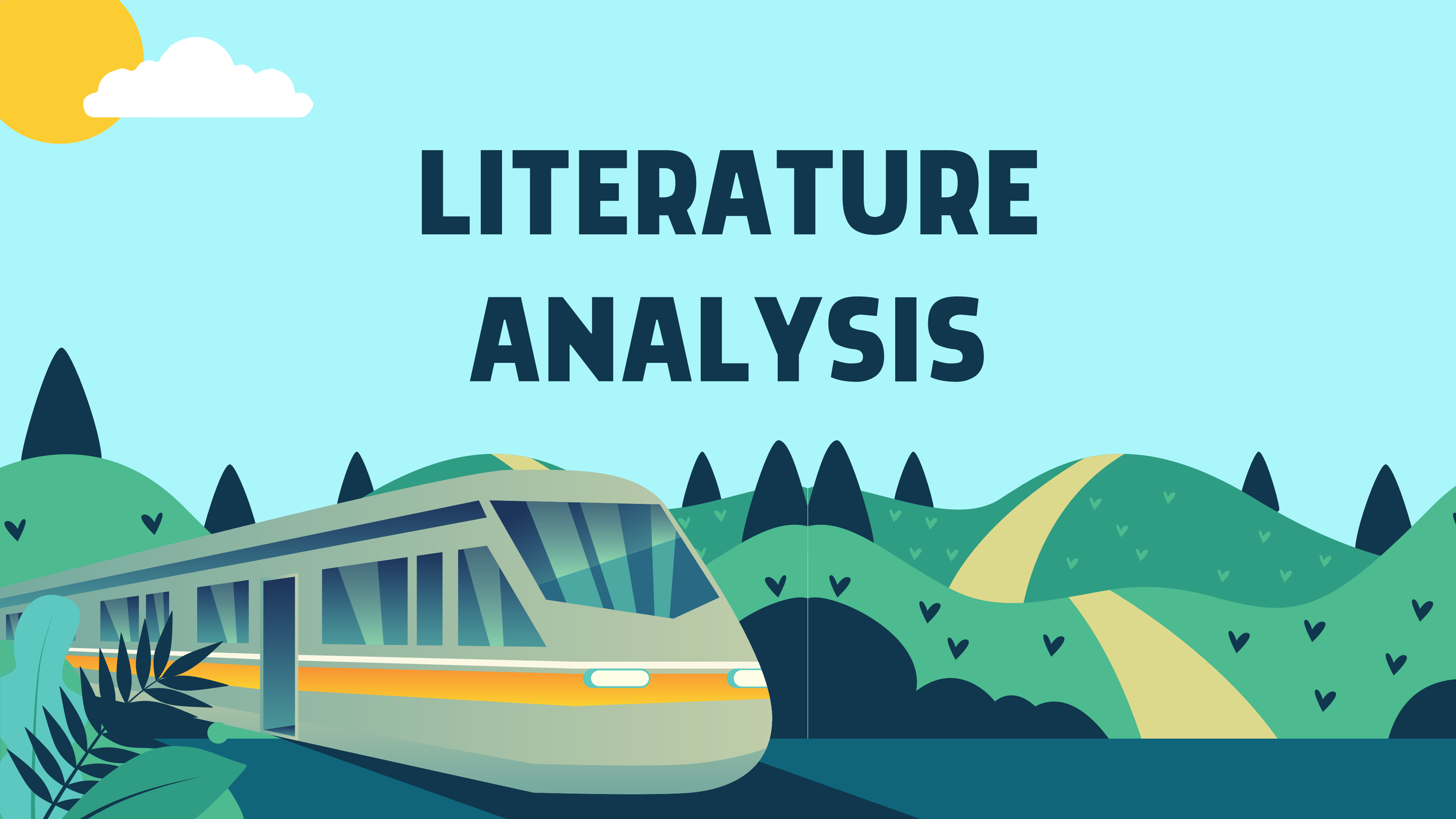
Analyse and evaluate the prediction results for their accuracy, stability and relevance to real-world transport planning.

PROJECT DOMAIN

- The project is based in the **transportation** domain, focusing on daily ridership forecasting for urban public transport in Kuala Lumpur.

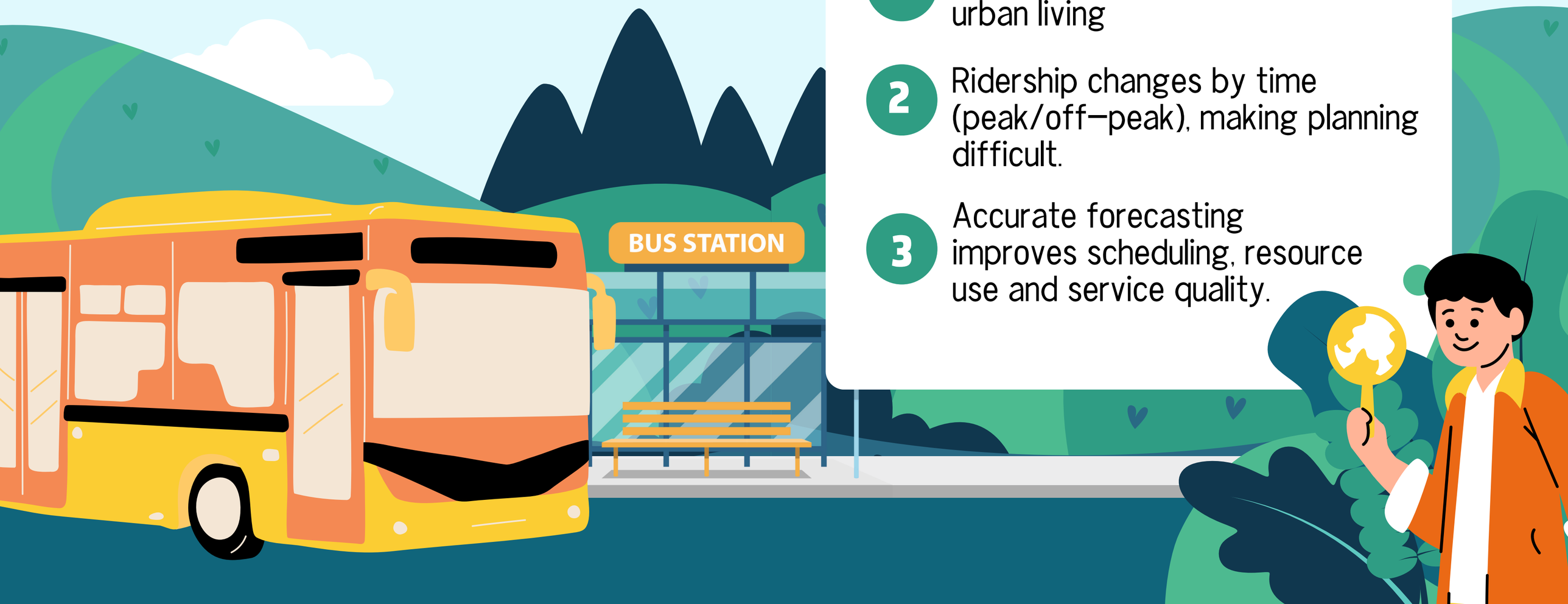
- Why **transportation**? Accurate ridership prediction is vital for:
  → Improved service planning.
  → Reduced congestion.
  → Efficient urban mobility.

LITERATURE ANALYSIS



WHY PUBLIC TRANSPORT FORECASTING MATTERS ?

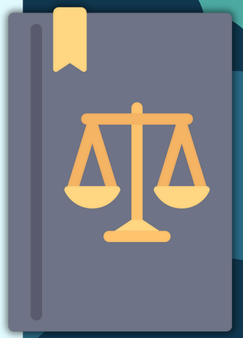
1. Public transport reduces traffic, pollution and supports sustainable urban living
2. Ridership changes by time (peak/off–peak), making planning difficult.
3. Accurate forecasting improves scheduling, resource use and service quality.



TRADITIONAL FORECASTING MODELS

- Traditional models like ARIMA (Auto–Regressive Integrated Moving Average) and SARIMA (Seasonal ARIMA) are widely used in time series forecasting.
- They are good at modeling trends, seasonality and cyclical patterns in data.
- Useful when ridership patterns are regular and consistent over time.
- ARIMA models are best for datasets with strong autocorrelation and no seasonality.
- SARIMA adds a seasonal component, making it suitable for weekly or monthly ridership patterns.
- However, both ARIMA and SARIMA struggle with:

1. Sudden shifts in data caused by external factors such holidays, weather and school breaks.
2. Unpredictable events like national celebrations or route changes.



GLOBAL PRACTICES VS MALAYSIA

- Cities like Singapore, Seoul, and London use real-time data to adjust public transport dynamically.
- These cities apply forecasting in intelligent transport systems (ITS). Forecasting is part of daily operations, not just research.
- Malaysia already has open ridership data (LRT, MRT, Monorail, RapidKL).
- But the data is mostly used for reporting and not for predictive planning.
- Forecasting tools like SARIMAX or ML are not yet widely applied in real-world transit management.

GOAL OF THIS STUDY

- Use SARIMAX to forecast public transport ridership in Malaysia.
- Include external factors like holidays, weather, and school breaks.
- Improve accuracy of predictions using real Malaysian data (e.g., RapidKL, MRT).
- Help planners make better decisions for scheduling and operations.
- Support smarter, data-driven public transport like in global cities.

METHODOLOGY → OSEMN

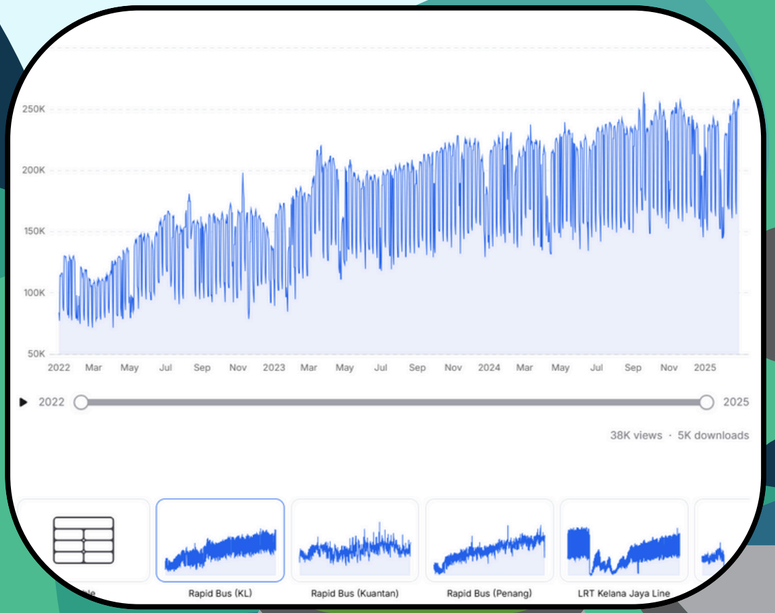
1. **Stage 1**: Obtain
2. **Stage 2**: Scrub
3. **Stage 3**: Explore
4. **Stage 4**: Model
5. **Stage 5**: Interpret

OBTAIN: DATA COLLECTION

1. Sourced from Malaysia's official open data portal
2. Covers multiple public transport services used in Kuala Lumpur like MRT, LRT, Monorail and RapidKL Bus
3. Downloaded in CSV format and loaded using Python (pandas) for analysis

data.gov.my

MALAYSIA'S OFFICIAL OPEN DATA SITE



rapidKL

BUS LRT LRT MONOREL

	bus_rkl	bus_rkn	bus_rpn	rail_lrt_ampang	rail_mrt_kajang	rail_lrt_kj	rail_monorail	rail_mrt
2019-01-01				113,357	114,173	139,634	35,804	
2019-01-02				182,715	169,316	274,224	31,859	
2019-01-03				187,904	175,304	286,469	31,893	
2019-01-04				198,420	187,891	304,755	34,121	
2019-01-05				120,773	112,660	145,036	29,950	
2019-01-06				101,145	95,913	120,032	25,342	
2019-01-07				197,569	184,365	301,290	31,988	
2019-01-08				196,879	185,920	304,680	31,792	
2019-01-09				197,314	188,770	307,069	32,305	
2019-01-10				198,876	189,818	310,510	32,057	
2019-01-11				209,859	204,378	326,471	33,265	
2019-01-12				124,608	123,932	152,720	30,671	
2019-01-13				106,718	108,346	127,805	27,235	
2019-01-14				199,162	192,036	306,455	32,172	
2019-01-15				195,405	189,880	305,742	31,260	

SCRUB: DATA CLEANING

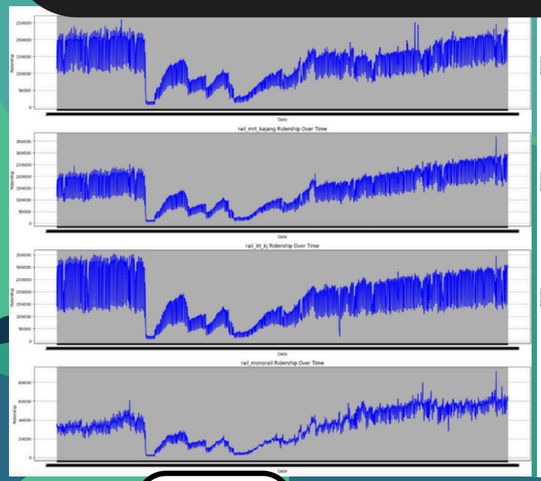
1 Dropping irrelevant services

Services outside Kuala Lumpur:

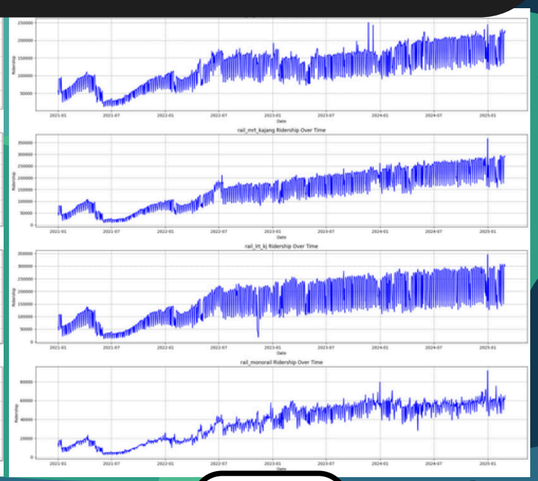
- Rapid Penang
- Rapid Kuantan
- KTM Komuter Utara
- KTM Intercity
- KTMB ETS
- KTM Shuttle Tebrau

```
#Dropping services that are not provided in or around KL
columns_to_drop = [
    'bus_rkn',              # Kuantan
    'bus_rpn',              # Penang
    'rail_ets',             # Ets
    'rail_intercity',       # Intercity
    'rail_komuter_utara',   # Komuter Utara
    'rail_tebrau'           # Johor-Singapore
]
df = df.drop(columns=columns_to_drop)
```

```
#Remove COVID-19 MCO timeframe
df['date'] = pd.to_datetime(df['date'])  # convert to datetime
df = df[(df['date'].dt.year >= 2021)]
```



Before



After

2 Removing COVID-19 periods

- Data from the year 2021 and earlier were excluded.
- Prevent noise and anomalies in model training.

SCRUB: DATA CLEANING

```
df['date'] = pd.to_datetime(df['date'])  # convert to datetime
df = df[(df['date'].dt.year >= 2021)]
```

	date	bus_rkl	rail_lrt_ampang	rail_mrt_kajang	rail_lrt_kj	rail_monorail	rail_mrt_pjy	rail_komuter
731	2021-01-01	NaN	54211	49130	56441	14651	NaN	NaN
732	2021-01-02	NaN	57689	51302	59116	13541	NaN	NaN
733	2021-01-03	NaN	46277	40364	46212	11088	NaN	NaN
734	2021-01-04	NaN	90205	80414	105722	17337	NaN	NaN
735	2021-01-05	NaN	90728	79556	107425	17631	NaN	NaN
...	...	...	...	...	...	...	...	...
2246	2025-02-24	249536.0	220881	272982	287593	60522	170853.0	32450.0
2247	2025-02-25	253188.0	221891	288166	303996	62582	178369.0	31604.0
2248	2025-02-26	258599.0	224522	292944	307859	62213	180403.0	30992.0
2249	2025-02-27	252569.0	220964	290119	303152	60832	174956.0	32905.0
2250	2025-02-28	254793.0	228126	292777	299468	65148	177707.0	38422.0

1520 rows × 8 columns

3 Datetime conversion

- The date column was converted from string to datetime format.
- Enables time–based operations.

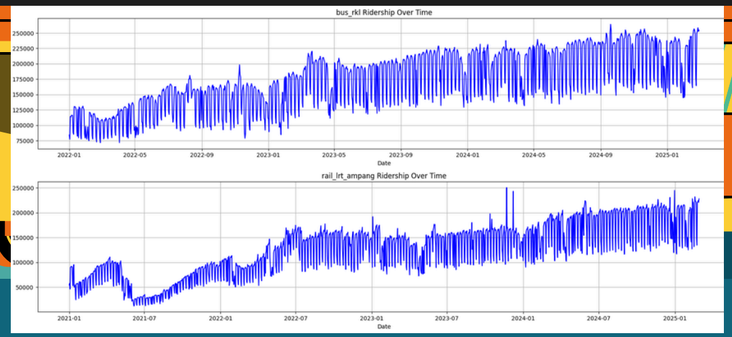
4 Handling missing values

- Missing ridership values were preserved (NaN).
- Our modeling and plotting methods are designed to skip them.
- This avoids introducing bias.

```
for col in df.select_dtypes(include='number').columns:

    # 1. Get first valid index (from cleaned df)
    first_valid_index = df[col].first_valid_index()
    if first_valid_index is None:
        continue  # skip if column is fully NaN

    # Slice the data from the first valid index onward
    trimmed_data = df.loc[first_valid_index:, ['date', col]].copy()
    trimmed_data = trimmed_data.dropna()
```



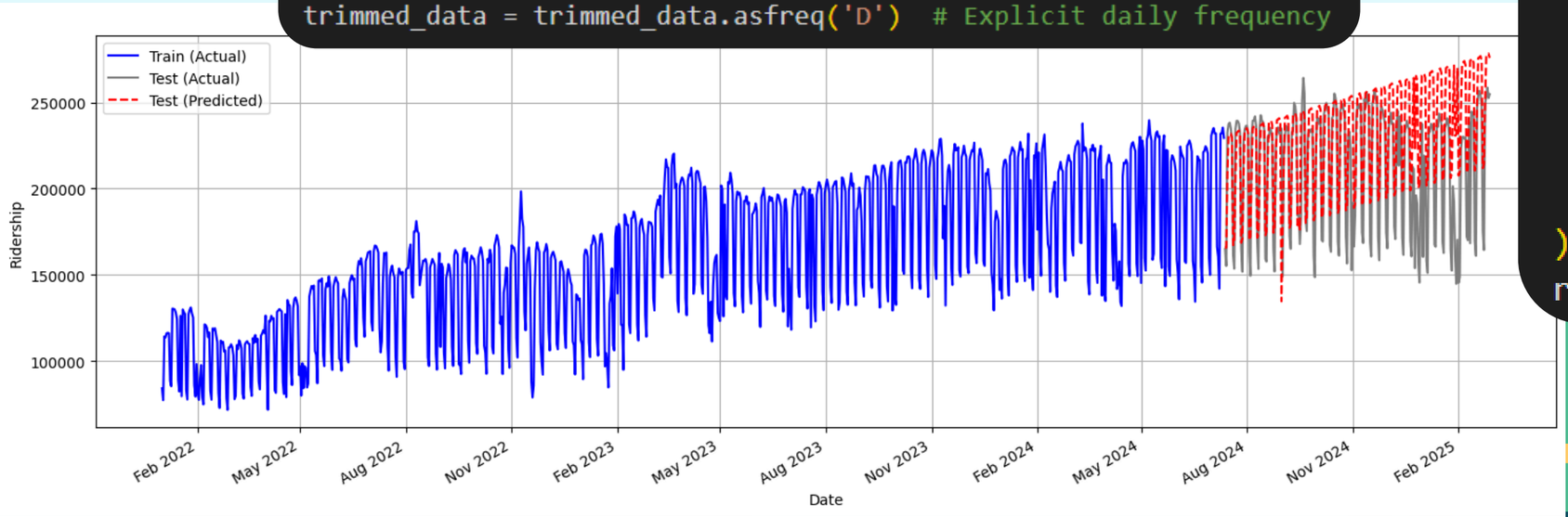
SCRUB: DATA CLEANING

5 Frequency setting

- The dataset was explicitly set to daily frequency to support time series modeling with SARIMAX.
- This ensures consistent date indexing.

```python
# 2. Convert 'date' to datetime and set as index
trimmed_data['date'] = pd.to_datetime(trimmed_data['date'])
trimmed_data.set_index('date', inplace=True)

# Declare daily frequency explicitly
trimmed_data = trimmed_data.asfreq('D')  # Explicit daily frequency
```



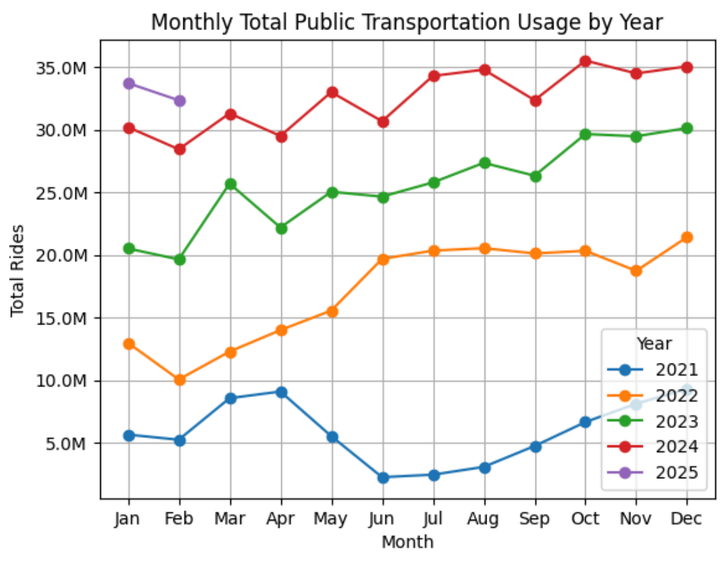

```python
# 4. Fit the SARIMAX model
model = SARIMAX(
    train[col],
    order=(1, 1, 1),
    seasonal_order=(1, 1, 1, 7),
    enforce_stationarity=False,
    enforce_invertibility=False
)
results = model.fit(disp=False)
```

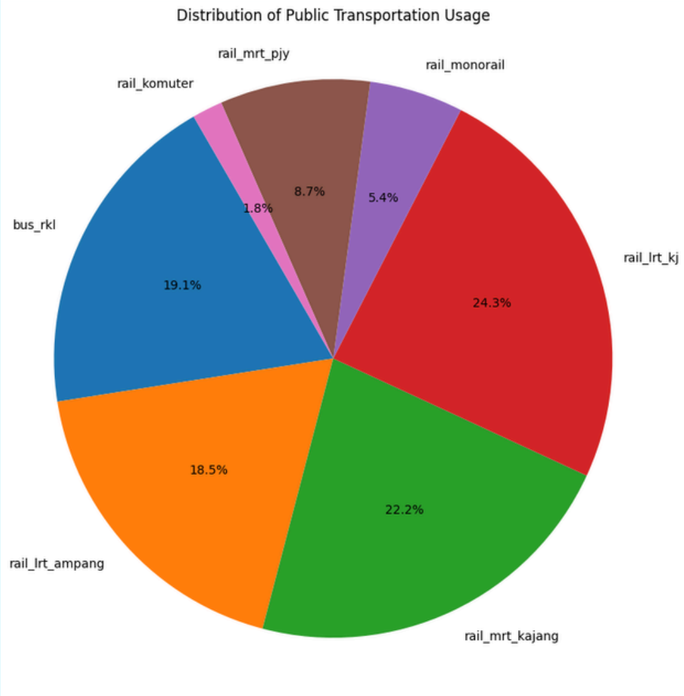
EXPLORE: EDA

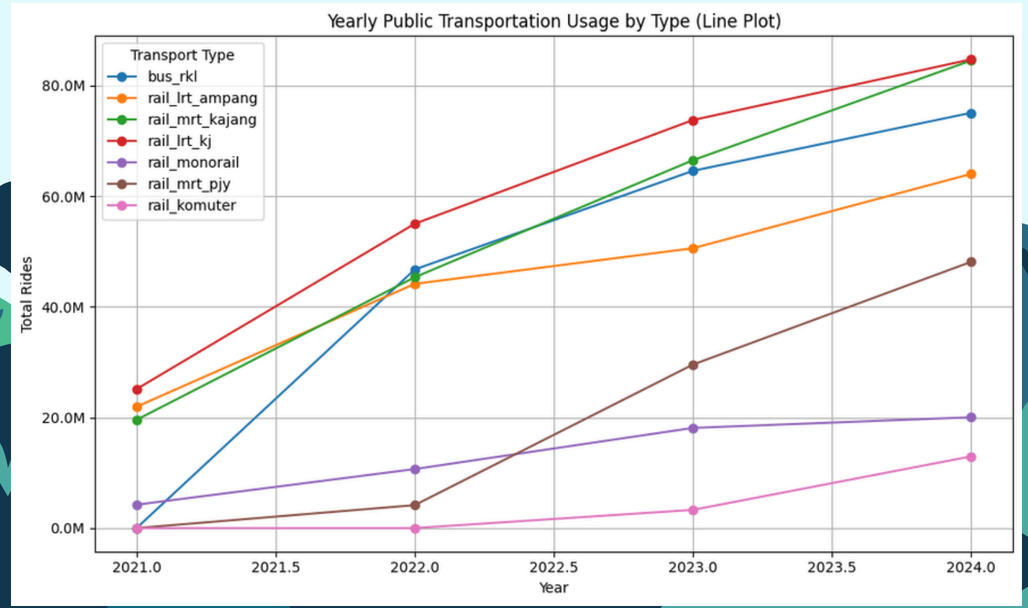
Visualization

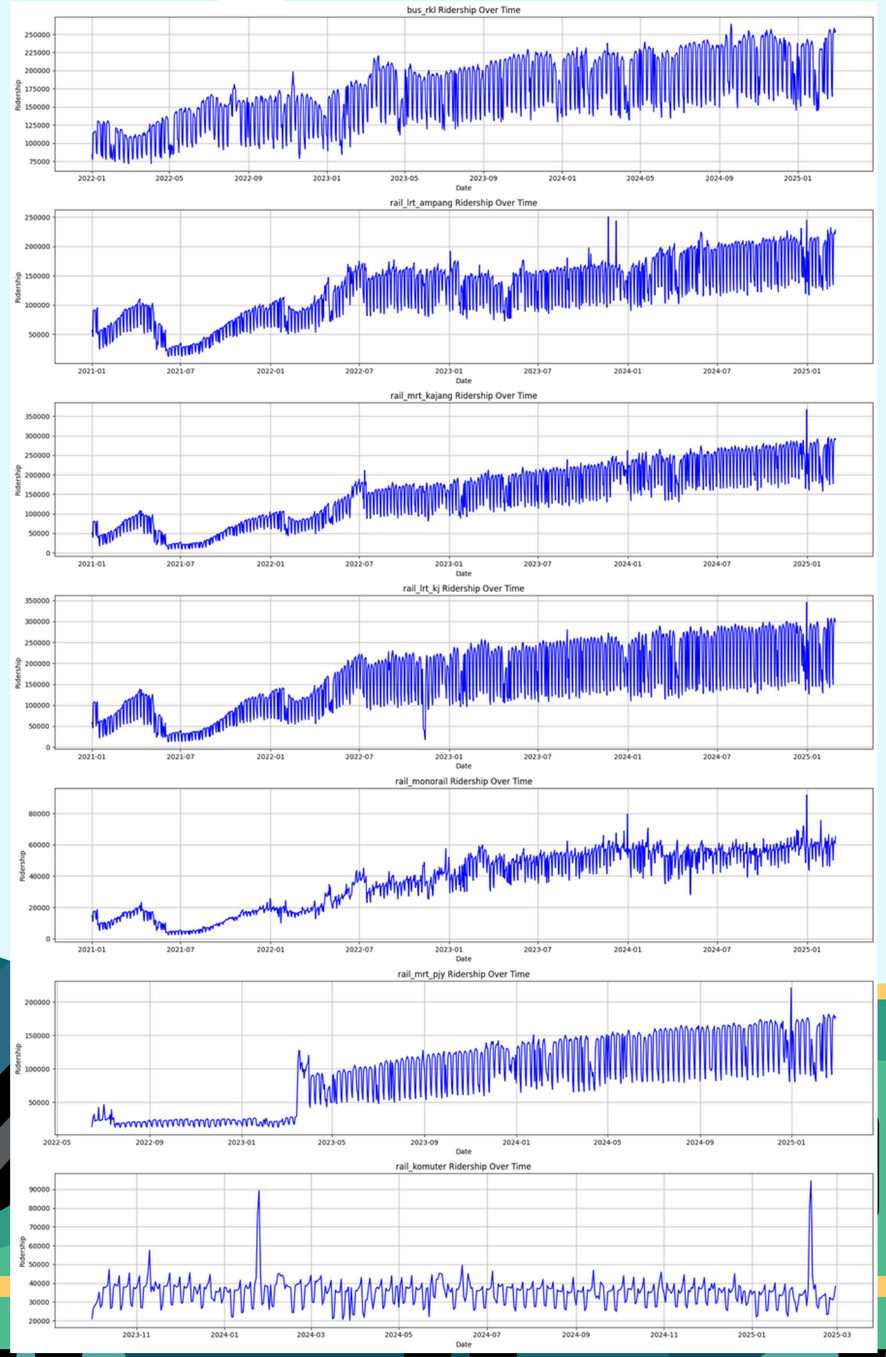
Domain Insights

1 VISUALIZATION









2 DOMAIN INSIGHTS

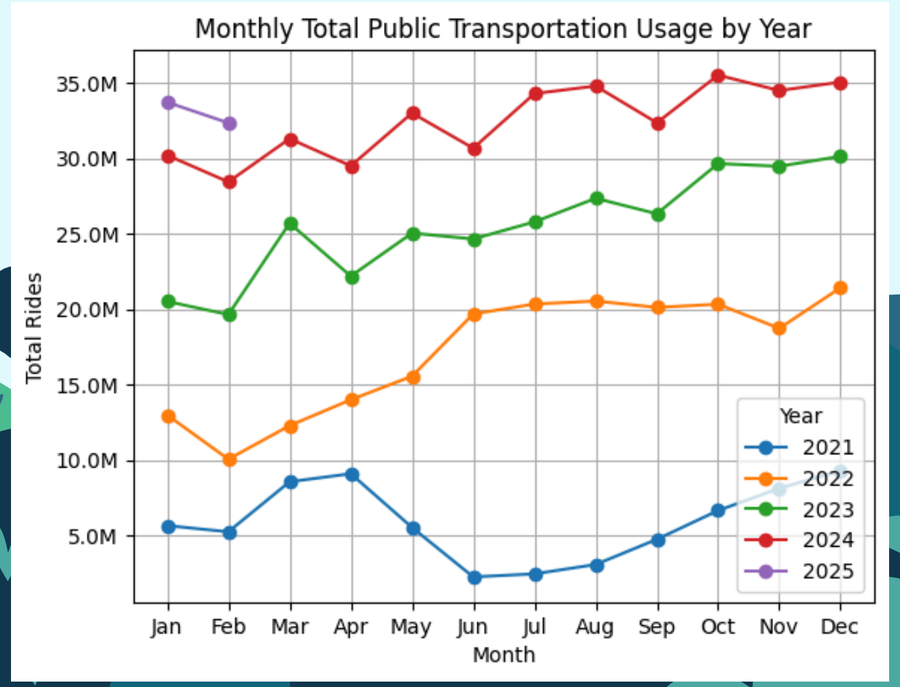
Clear Trends in Public Transport Usage

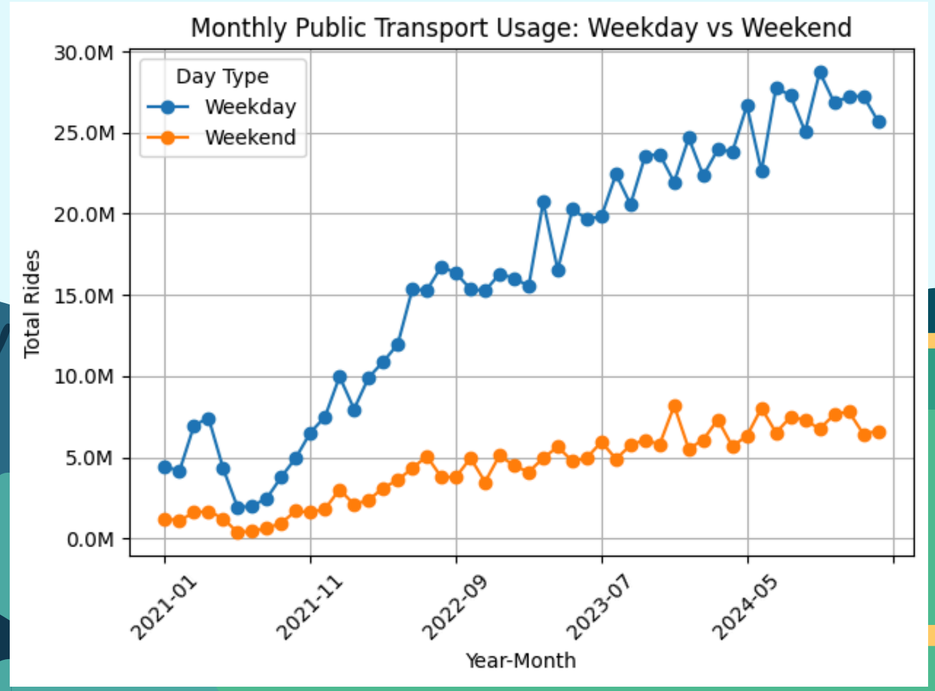
Transport Mode Distribution

Impact of Real-World Events

1 CLEAR TRENDS IN PUBLIC TRANSPORT USAGE

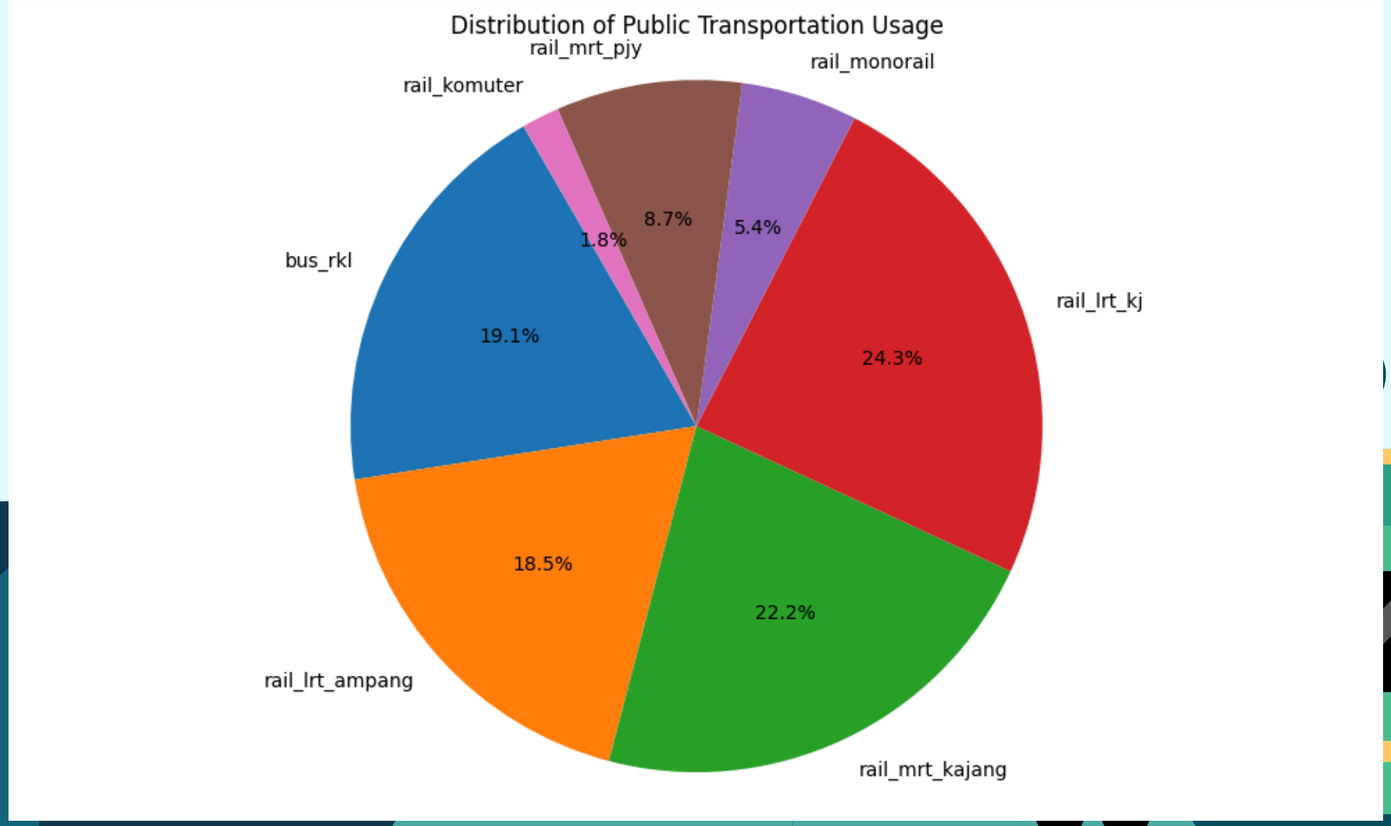
- Weekly ridership is consistently higher than weekends, confirming strong commuter pattern.
- Certain months like January and December show seasonal surge likely due to school breaks and holidays.
- MRT and LRT systems have higher volumes of ridership compared to buses and monorails.





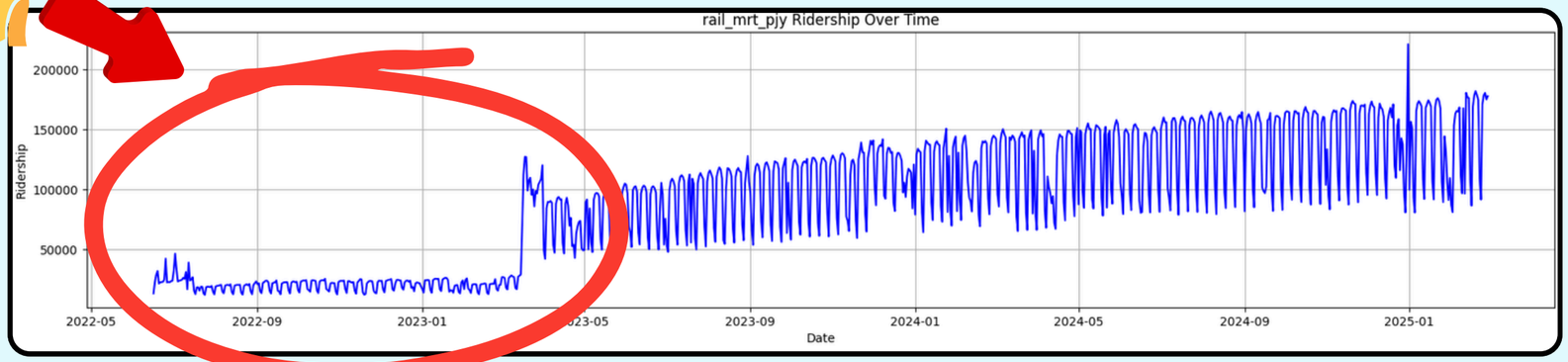
2 TRANSPORT MODE DISTRIBUTION

- MRT Kajang and LRT Kelana Jaya lines have the highest share of ridership.
- Bus and Komuter usage is significantly lower, indicating they may serve more niche and less dense route.

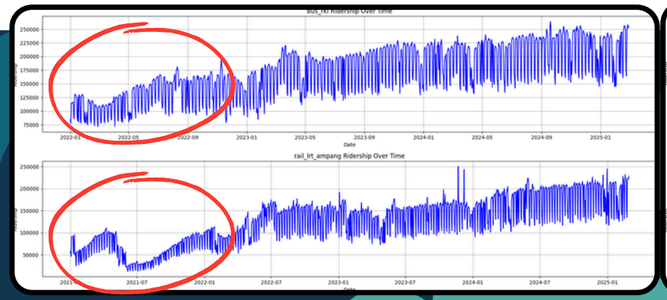


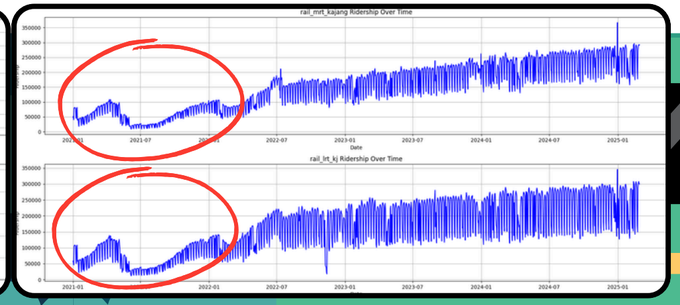
3 IMPACT OF REAL-WORLD EVENTS

- MRT Putrajaya Line was not fully open in late 2022 causing the ridership frequency to be low during the timeframe. After it fully launched in March 2023, more people started using it and the ridership began to significantly rise.



- Post-COVID recovery can be seen as Malaysia moved past the MCO in 2022. The ridership slowly increased as people returned to work and daily activities after lockdowns.





ETHICAL CONSIDERATION

- **Purpose Alignment**

Making sure project goals aligns with public interest, as well as open data policy by government.

- **Avoiding Plagiarism**

Use open data by the government and give proper credit when referencing others' example.

- **Handling Abnormal Events (COVID–19)**

Treat this period differently e.g., truncate the dataset at the time where MCO is over to prevent skewed analysis.

- **Project and Model Security**

Use safe and secure platform to help maintain data integrity and prevent altering and manipulation from others.

- **Maximizing Benefits and Minimizing Harms**

Frame the model to support decision–making, not to replace it. Explain the limitation so the model is interpreted correctly.

IMPACT OF OUR DATA PRODUCT

Improved Service Planning

Transit authorities can use the forecasts to adjust train frequency, timing or maintenance more efficiently.

Reduced Overcrowding

Riders enjoy safer, and more comfortable travels after planners preemptively increase the capacity on the busy days.

Encouraging Public Transport Use.

By making the public transport more reliable and comfortable, this can encourage people to use it more regularly over other transport options.

Support for Infrastructure Planning

The model helps identify long term growth trends on certain lines or areas and support decisions on future rail expansions, station upgrades, etc.

REFERENCE

1. Department of Statistics Malaysia. (n.d.). *Daily Public Transport Ridership*. [Data set]. https://data.gov.my/data-catalogue/ridership_headline

2. Melanie. (2024, March 14). *SARIMAX model: What is it? How can it be applied to timeseries?* DataScientest. https://datascientest.com/en/sarimax-model-what-is- it-how-can-it-be-applied-to-time-series#:~:text=Among%20the%20various%20approaches%20available,into%20the%20analysis%20to%20improve

3. Hotz, N. (2024, March 31). *OSEMN Data Science Life Cycle*. Data Science PM. https://www.datascience-pm.com/osemn/

4. Penn State. (n.d.). *Forecasting using ARIMA/SARIMAX models*. https://online.stat.psu.edu/stat510/lesson/6

5. Obaid, H. (n.d.). *Forecasting demand in public transportation system* [Code notebook]. Kaggle. https://www.kaggle.com/code/hamzaobaid/forecasting-demand-in-public-transportation-system

THANK YOU