Pedestrian Counts Prediction

Executive Summary:

The objective of this project was to Forecast pedestrian counts for Melbourne Central Street in Sydney using time series analysis. The key findings, predictive model details, accuracy metrics, and insights from the heat map visualization are summarized below.

1. Introduction:

The analysis of pedestrian counts in October 2023 holds significant relevance within the broader context of urban planning, public safety, and resource allocation. As cities continue to evolve and grow, understanding pedestrian movement patterns becomes crucial for optimizing infrastructure, enhancing public spaces, and ensuring the well-being of urban communities.

In October 2023, several factors contribute to the importance of this analysis:

1. **Seasonal Variations:**
    - Analyzing pedestrian counts during this period allows for insights into how these seasonal variations impact foot traffic in different locations.
2. **Event Impact:**
    - Various events, such as festivals, public gatherings, or holidays, may influence pedestrian activity. Understanding these patterns in October helps local authorities plan and manage events effectively, ensuring public safety and enjoyment.
3. **Urban Mobility:**
    - The dynamics of pedestrian movement are interconnected with broader urban mobility trends. Analyzing pedestrian counts provides valuable information for optimizing transportation systems, planning pedestrian-friendly zones, and improving overall accessibility.

2. Data Collection and Preprocessing:

Data Source:

This data were collected from City of Melbourne Home page. It contains about 120 Datasets to be downloaded. Each dataset represents the pedestrian counts for each month from 2013 to 2023. We scraped the data and saved the DataFrame in a csv file for direct use.

Data Exploration:

Our data contained many problems. Street count columns had mixed data types rather than being all of type integer. Date column contained different time format that we had to unify and change column data type to datetime for further usage. 'Hour' column wasn't in time format and was separated from the Date column, so we had also to unify its formats and concatenate Date column with Hour column also for further usage.

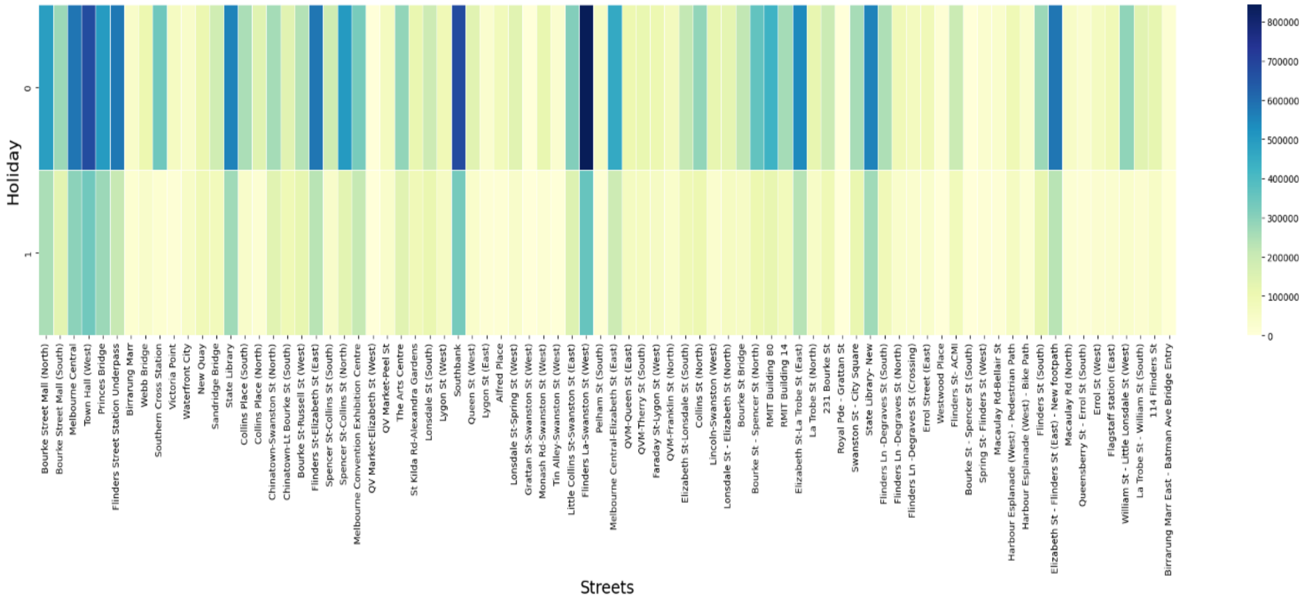
Preprocessing:

- Handled inconsistent data such as negative values in pedestrian counts.
- Handles missing and Outliers values by dropping them from our target column 'Melbourne Central' as they weren't large compared to our data and I preferred dropping it rather than using statistical method such as mean and median as in some cases these values will be inconsistent to the date corresponding to.
  (E.g., Having missing value in data 07/10/2022 03:00:00 can't be handled by median=964 as the count of people in late hours is less than 100 in average)

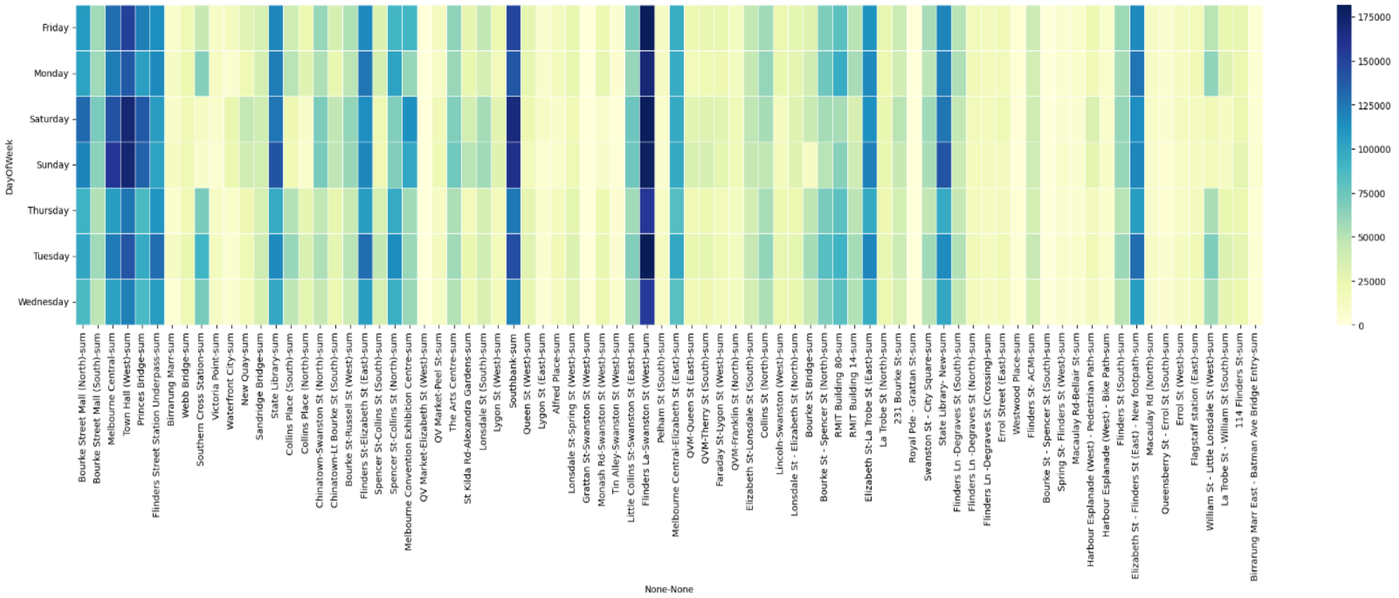
3. Exploratory Data Analysis (EDA):

Visualizations:

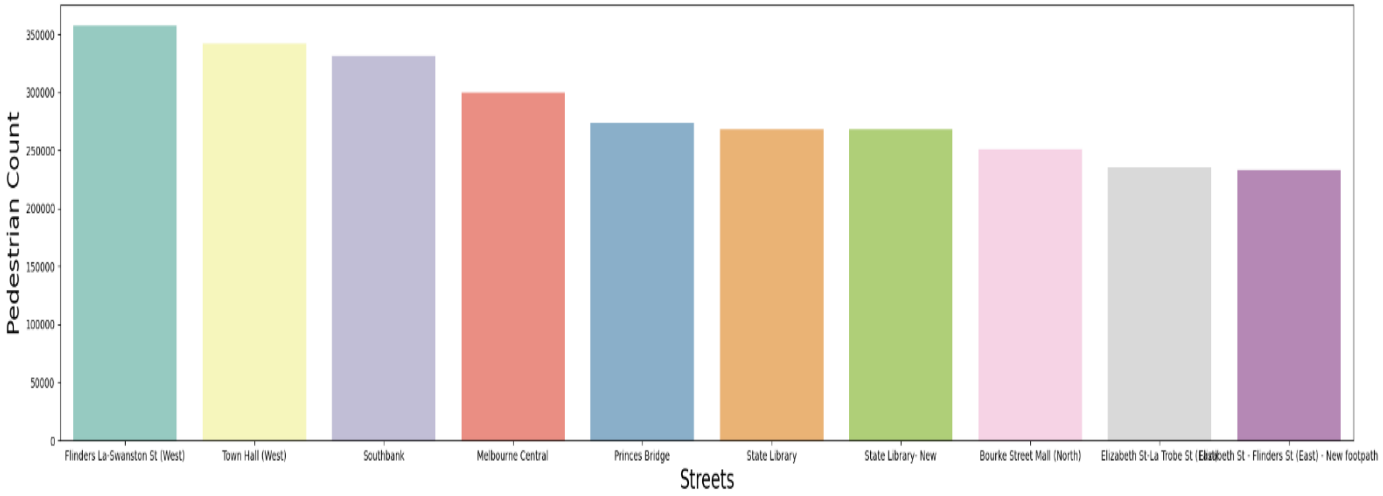
- Busiest streets on Weekends



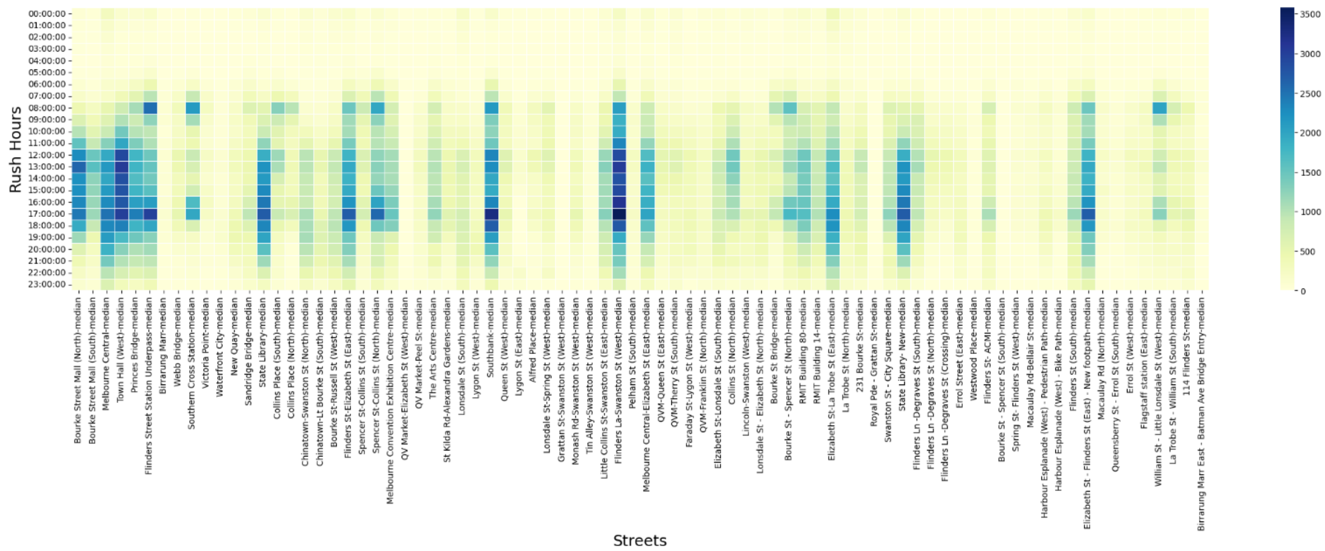
- Crowdedness of streets throughout the week.



- Top 10 busiest streets



- Rush Hours



Statistical Analysis:

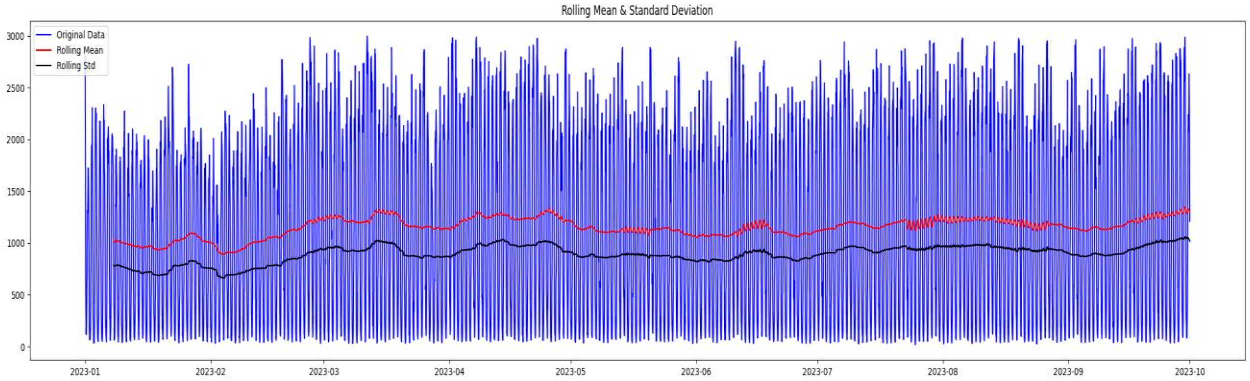
Used statistical method as (mean , std, median, min, max) to describe our data. Also Used correlation to observe connection and busyness of streets with each other.

And used visualization to understand the data distributions.

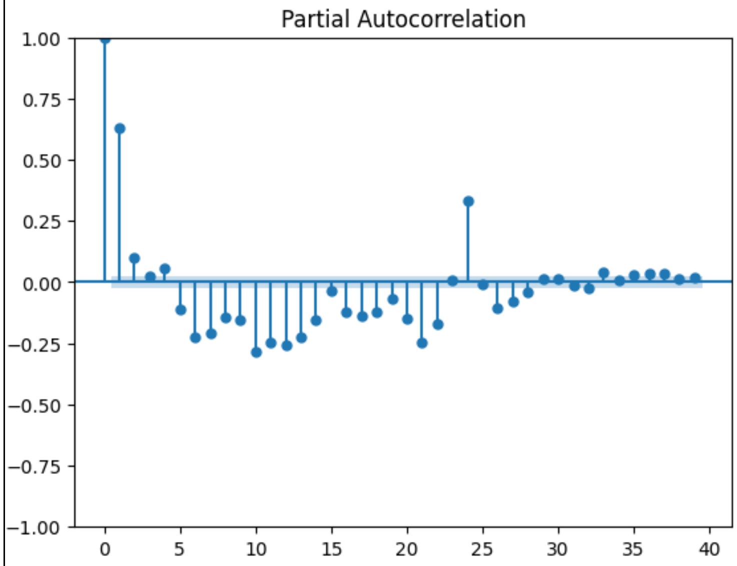
4. Predictive Model Details:

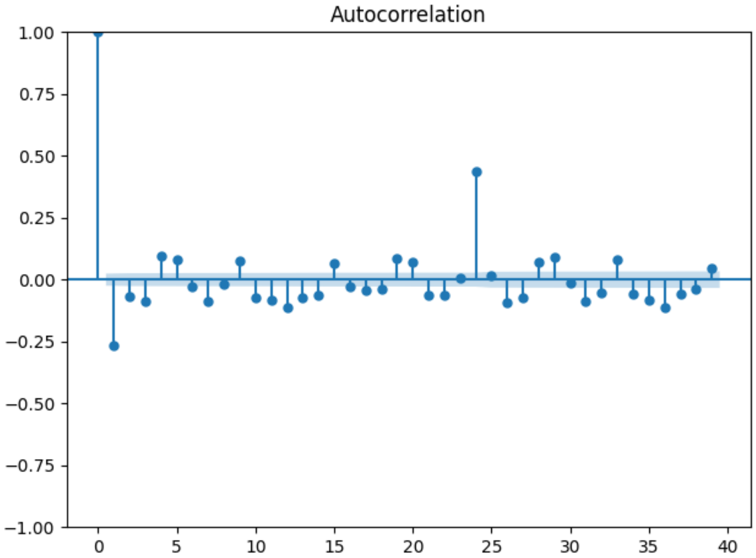
Analyzing time series data using an ARIMA (AutoRegressive Integrated Moving Average) model involves several steps. ARIMA is a widely used statistical method for time series forecasting. Here's a step-by-step guide on the analytical methods for using the ARIMA model:

- **Understand the Data:**
  - Examine the time series data to understand its characteristics, trends, and seasonality. This involves visual inspection of plots and summary statistics.
- **Stationarity:**
  - Check for stationarity in the time series. ARIMA assumes that the data is stationary, meaning its statistical properties (like mean and variance) do not change over time. If the data is not stationary, differencing may be required.



- **ACF and PACF:**
  - Examine the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots to identify potential values for the autoregressive
  - (AR) and moving average (MA) parameters.

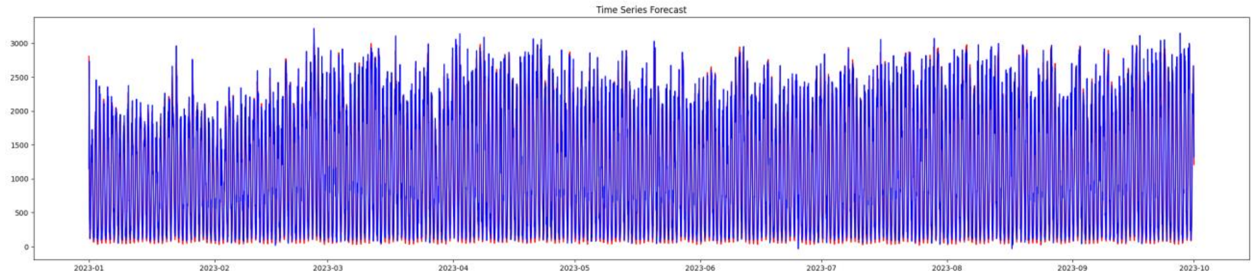




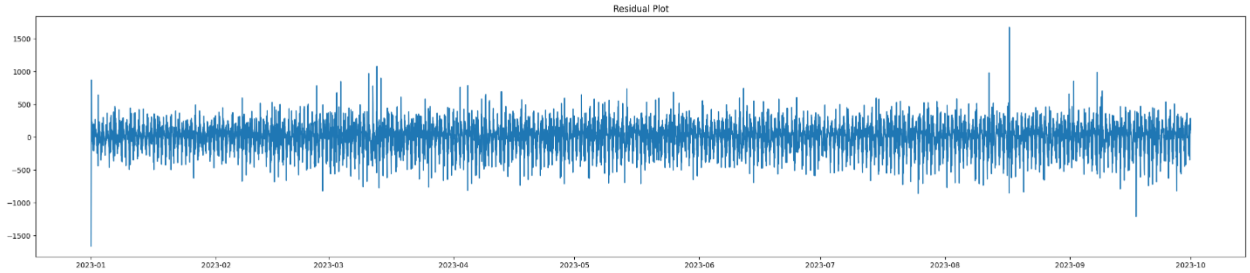
Identification of Model Order (p, d, q):

- Based on the ACF and PACF plots, determine the orders (p, d, q) for the ARIMA model:
- **p (AR Order):** The number of lag observations included in the model.
- **d (Differencing):** The number of times the raw observations are differenced.
- **q (MA Order):** The size of the moving average window.

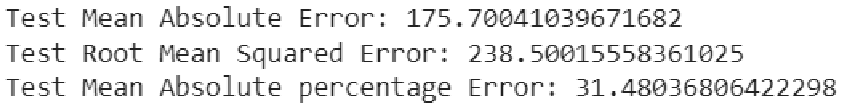
- **Training the ARIMA Model:**
  - Split the sorted data(by time) into training and testing sets.
  - Fit the ARIMA model to the training data using the identified orders (p, d, q).



- **Prediction:**
    - Generate predictions for future time points using the trained ARIMA model.

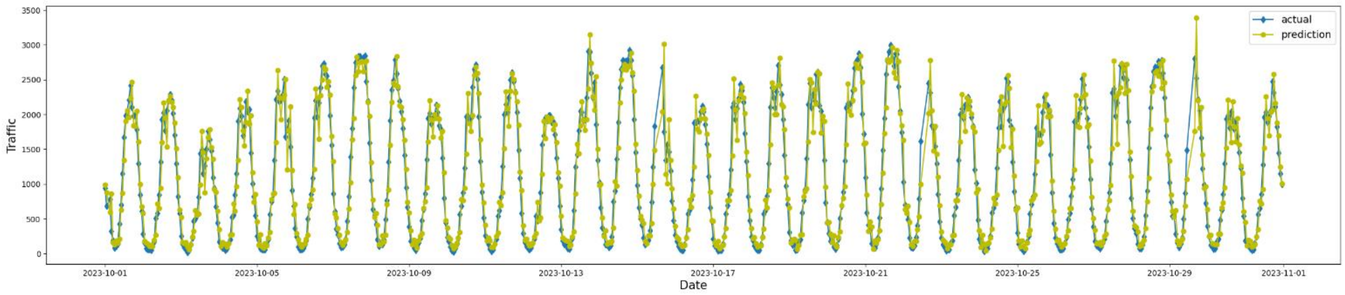


- **Model Evaluation:**
    - Evaluate the model's performance on the testing set.
    - Used metrics : Root Mean Squared Error (RMSE) , Mean Absolute Error (MAE) and Mean Absolute percentage Error.



```
Test Mean Absolute Error: 175.70041039671682
Test Root Mean Squared Error: 238.50015558361025
Test Mean Absolute percentage Error: 31.48036806422298
```

- **Visualization:**
    - Visualized the predicted values against the actual values to assess the model's performance and capture any patterns.


actual
prediction
Traffic
Date