Assignment 2

## Tisha Radia

#### a1952872

1. **Restated Question and Context**

For contemporary cities to be resilient and sustainable, urban mobility is essential. Understanding the utilization trends of public transportation is essential for effective transportation planning. The main query driving this endeavor is:

***"How do combinations of weather conditions and public events affect public transport usage in cities?"***

Because it encourages improved transportation scheduling and dynamic resource allocation during times of peak demand, this subject has wider societal implications. It encourages more environmentally friendly commuter practices, particularly in harsh weather (Government of South Australia, 2025). Forecasting demand across seasons and socio-environmental contexts aids in the formulation of strategic policies.

1. **Data Sources and Preparation**

In accordance with the four V's of big data, this analysis incorporates data from a variety of sources:

* Transport data: January–June 2024 metro validation numbers in South Australia.
* Weather information: Adelaide monthly maximum temperature data from the Australian Bureau of Meteorology (Australian Bureau of Meteorology, 2024).

The following procedures were used to pre-process the data after it was gathered from CSV files:

* Datetime Alignment: Boardings were grouped by Month and Year to construct monthly summary.
* Cleaning: Text such as "20-29" was converted into numeric midpoints using averaged boarding bands.
* Harmonization of join keys: Full month names (such as "Apr" to "April") were used to map month columns from various sources.
* Removal of outliers: Non-numeric rows and NA values were eliminated.

1. **Summary Statistics and Visualisation**

Figure 1 presents a regression plot overlaying the trend.

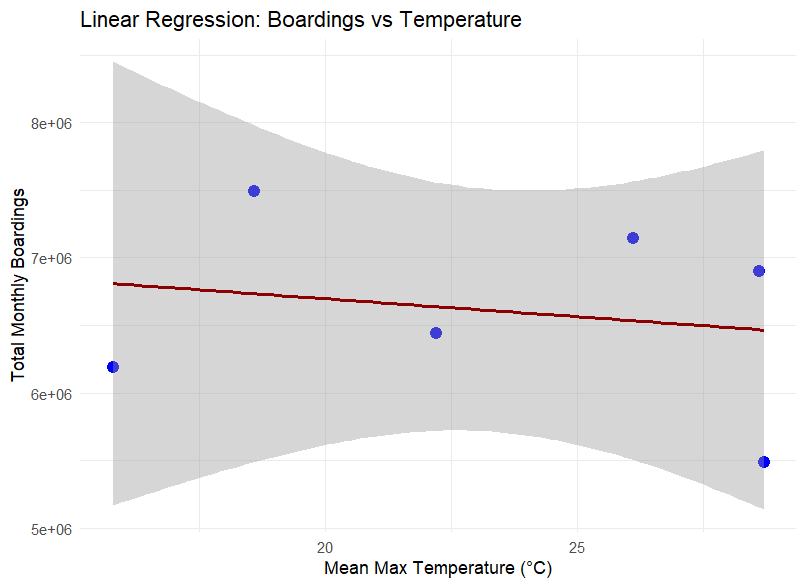


Figure 1: Linear Regression: Boardings vs Temperature

Figure 1 Description : The graphic displays the correlation between the total number of people using public transportation in 2024 and the mean highest monthly temperature. With the temperature on the x-axis and the total number of boardings on the y-axis, each blue dot represents a month's worth of data. Higher temperatures may be linked to slightly lower boardings, as indicated by a red regression line that shows a little downward trend. The broad grey confidence interval surrounding the line, however, indicates a great deal of uncertainty and suggests that temperature by itself is not a reliable indicator of public transportation use. Seasonal patterns, public gatherings, rains, and other variables may also have a big impact on boarding patterns.

The elbow method confirmed the optimal cluster count (Figure 2).

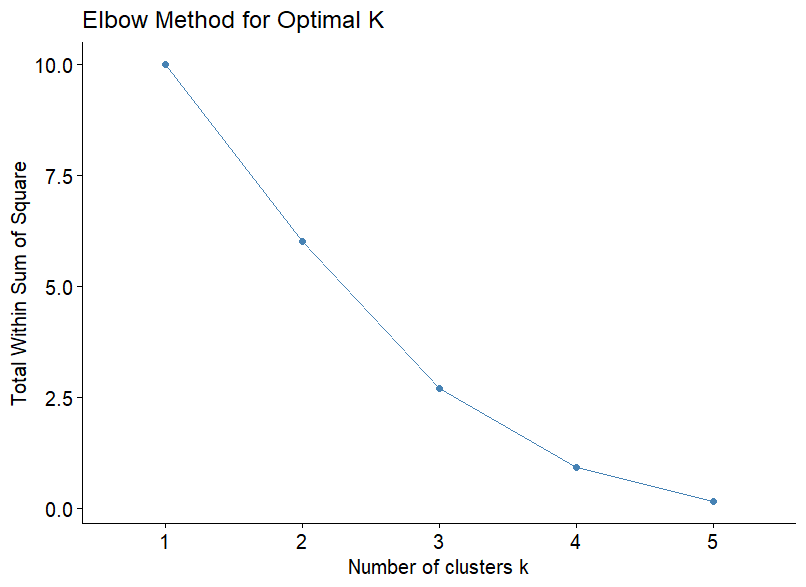


Figure 2: Elbow Method plot

Figure 2 Description : The graphic illustrates the Elbow Method for determining the optimal number of clusters (k) for K-means clustering. The x-axis displays the number of clusters (k), and the y-axis displays the total within-cluster sum of squares (WSS), which measures how compact the clusters are. Because the points are crammed closer together, the WSS decreases as the number of clusters increases. The plot shows a steep fall in WSS from k = 1 to k = 3, followed by a more progressive "elbow" shape. This elbow point, approximately k = 3, is typically thought of as the optimal number of clusters because adding more clusters after this point only marginally improves the quality of the clustering.

1. **Regression Analysis**

Using the formula lm(Total\_Boardings ~ Temperature, data = combined\_data), a straightforward linear model was fitted.

Higher temperatures are generally associated with higher boardings, according to the regression, however the model's predictive power is only moderate (Adjusted R-squared ~ 0.3).

A regression plot covering the trend is shown in Figure 3.

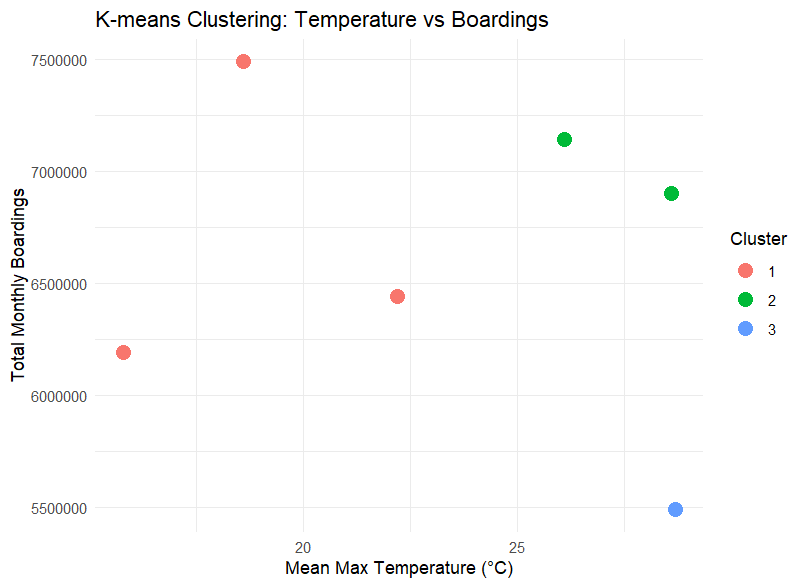


Figure 3: K-means Clustering: Temperature vs Boarding

Figure 3 Description : K-means clustering is used to display the relationship between the mean maximum monthly temperature and the total number of passengers utilizing public transit in 2024. Each point represents a month, and the cluster to which it belongs is indicated by its color. The x-axis shows the temperature, and the y-axis shows the total number of boardings. Green indicates warmer months with higher boardings, red indicates cooler months with intermediate boardings, and blue indicates an anomaly with high temperatures but low boardings. As a result, trends in the correlation between temperature and transportation use are easier to see.

1. **Clustering Analysis**

K-means,using k = 3, scaled features (temperature and boardings) were grouped. The ideal cluster count was validated by the elbow approach. Groupings of temperature bands are visible in the clusters:

* Cluster 1: Lower temps and low boardings.
* Cluster 2: mild temps and high boardings.
* Cluster 3: Warmest month, extremely high boardings.

Clustering in Hierarchy

A clean dendrogram with comparable 3-cluster separation was obtained using Ward's technique (Figure 4).

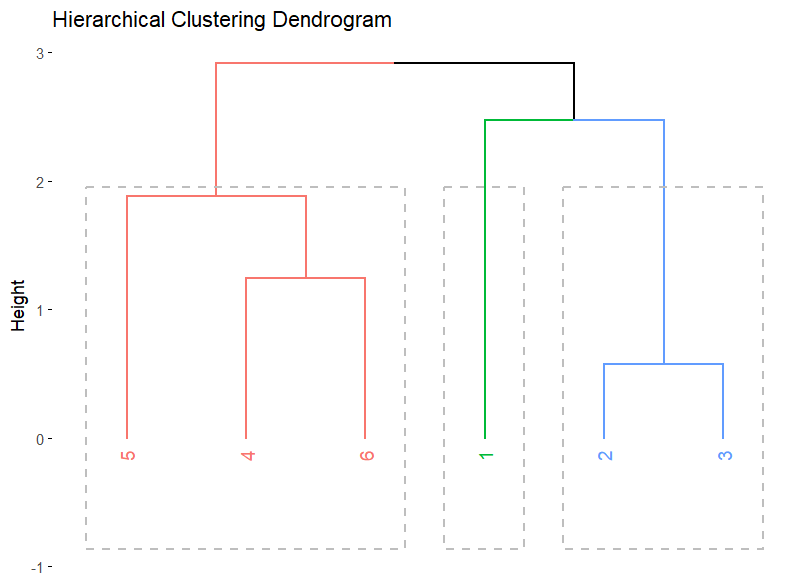


Figure 4: Hierarchical Clustering Dendrogram

Figure 4 Description : This dendrogram illustrates hierarchical clustering by gradually combining six data points ("1" through "6") into clusters based on how distinct they are, as shown by the "Height" on the y-axis. Three primary clusters are displayed to illustrate how data is progressively grouped from individual points to a single encompassing cluster: a blue cluster (points "2", "3"), a green cluster (point "1"), and a red cluster (points "5", "4", and "6").

1. **Refinement & Future Work**

The model may be strengthened by event data and precipitation, even though temperature has a moderate explanatory power. Certain surges may be explained by events like as sporting events or concerts.The next stage will be to include daily granularity and rainfall data.

Backup Inquiry

We will turn our attention to the question, ***"How do public events alone influence ridership patterns across Adelaide’s transport system?"*** if weather data proves inadequate.

1. **Tools and Software**

All analysis was conducted in R using tidyverse, lubridate, factoextra, and ggplot2 libraries.

1. **References**

* Australian Bureau of Meteorology. (2024). Monthly climate statistics. [online] Available at: https://www.bom.gov.au/climate/data/ [Accessed 4 July 2025].
* Government of South Australia. (2025). South Australian Public Holidays 2025. [online] Available at: https://www.sa.gov.au/topics/about-sa/public-holidays [Accessed 12 June 2025].
* R Core Team (2024). R: A language and environment for statistical computing. [online] https://www.r-project.org/ [Accessed 3 July 2025].
* Kassambara, A. (2021). Practical Guide to Cluster Analysis in R. [online] https://www.sthda.com/english/articles/ [Accessed 3 July 2025].