**Question 2**

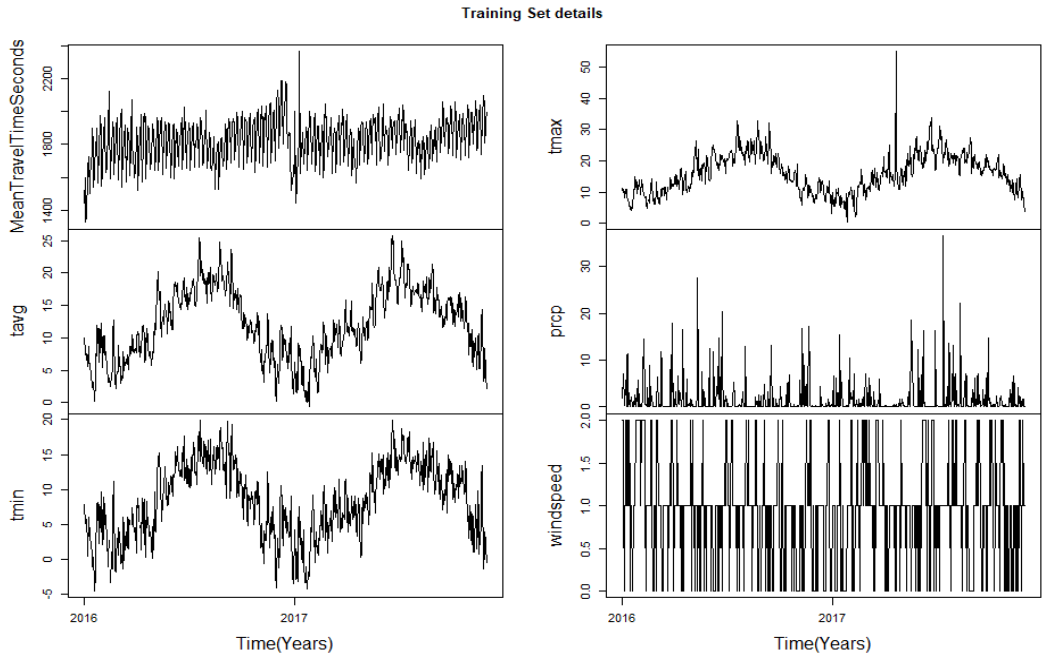



*Figure 1: Plot of variables in the dataset*

**Figure 1** shows plots of all the parameters in the dataset. We also see **yearly seasonality** in the temperature data. There are many spikes in the precipitation data which indicates many outliers. There are also big spikes in the data for maximum temperature which may indicate higher value outliers. **Windspeeds** range between **0,1,2.**

A closer inspection of the ACF of **mean travel times** showed **seasonal variations** at every **7 lags** even though they appear to be white noise in **Figure 1** (possibly due to the high frequency of observations). This is an indication of **weekly seasonality with frequency 7**. The ACF of both precipitation and windspeed appear stationary. On the other hand, the ACF of **maximum, minimum and average temperatures** shows **strong autocorrelation with yearly seasonality**.

| | MeanTravelTimeSeconds | tavg | tmin | tmax | prcp | windspeed |
|---|---|---|---|---|---|---|
| MeanTravelTimeSeconds | 1.000000000 | 0.02910464 | 0.04371287 | 0.01501888 | -0.006200695 | -0.05045250 |
| tavg | 0.029104639 | 1.00000000 | 0.95730917 | 0.94635904 | -0.016247747 | 0.08492698 |
| tmin | 0.043712868 | 0.95730917 | 1.00000000 | 0.86123185 | 0.044407086 | 0.13547456 |
| tmax | 0.015018877 | 0.94635904 | 0.86123185 | 1.00000000 | -0.051126207 | 0.01429939 |
| prcp | -0.006200695 | -0.01624775 | 0.04440709 | -0.05112621 | 1.000000000 | 0.06133438 |
| windspeed | -0.050452497 | 0.08492698 | 0.13547456 | 0.01429939 | 0.061334376 | 1.00000000 |

*Figure 2: Correlation between different variables*

In **Figure 2**, we see the correlation table between the different variables in the dataset. The **temperatures** appear **highly correlated** to each other. On the other hand, **precipitation and windspeed** have **less correlation** with other variables. **Minimum, maximum and average temperature** show **higher positive correlation** with mean travel times than the other variables. **Precipitation and windspeed** show **negative correlation** with mean travel times. We can try these correlated variables for models that support regression errors.

There **5** outliers for **mean travel times**, **103** for **precipitation** and **265** outliers for **windspeed**. The spike in maximum temperature at the end of 2016 corresponds to **1 outlier** we see in the **maximum temperature**. There are **0** outliers for **the minimum and average temperatures**.

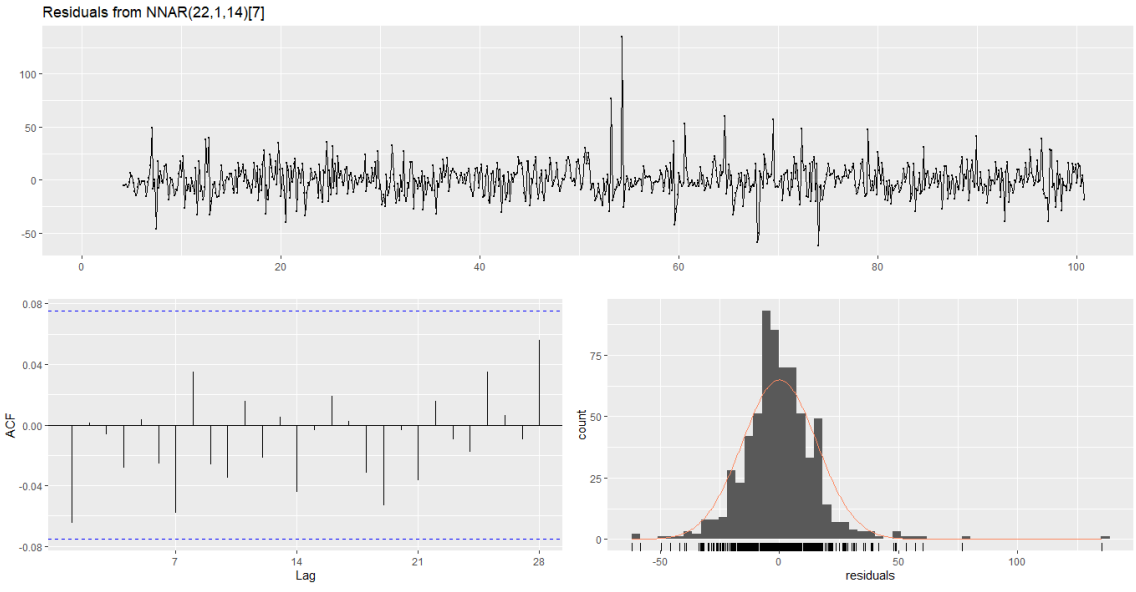
Possible issues with the data include inadequate data of windspeeds which can take only 3 values between low, medium, high. Seasonal variations of the data have not been provided beforehand, although we see yearly seasonality in temperatures and weekly seasonality in mean travel times on close observation. There are also NA values for some rows of the weather data. Precipitation and windspeeds also look highly unpredictable and this might affect forecasting.

Summary statistics for the training set show that the date range is from **02-01-2016 to 30-11-2017**. There are **5** NA values in **tmin, 6** NA values in **tmax and 1** NA value in the **prcp** column. Linear interpolation can be used to fill these columns. Windspeeds range between 0,1,2 with the median at 1. Analysing the dataset, we see observations in the dataset with **similar values** for mean travel times, average temperature and windspeed but with **no NA values**. It may be possible that the NA values occurred due to **incorrect collection of data.**

**Question 3**

a) For choosing the model, the **ACF of the mean travel times** was analysed. It showed **seasonal variations at every 7 lags**. This means the daily observations might have a **weekly seasonality.** After converting the mean travel times to a time series and **setting the frequency to 7**, the decompose function displayed a **complex trend and weekly seasonal variations.**
**NNAR(22,1,14)[7] with all other weather data as predictors** is the **best model** among the models we have trained for the time series of mean travel times.
As for the choice of other models, a **SARIMA(3,0,5)(1,1,1)[7]** model was chosen based on the characteristics of the **ACF and PACF of the training set**. Multiple **SARIMA(3,0,5)(1,1,1)[7]** models were tested with various predictors added in. Additionally, the **holt-winters model** was chosen because of the **complex trend** seen when the **mean travel times** was decomposed. The ACF of both models resulted in white noise but the RMSE score was lower for **NNAR(22,1,14)[7]**.

b) There are some assumptions made for the model chosen:
- The weather conditions might have an impact on the ride times of our model since we saw some correlation between the other variables and mean travel time.
- We have assigned at frequency of 7 to the **mean travel times** time series since we saw weekly seasonality in the ACF. It is also assumed that the temperatures have yearly seasonality based on the ACF data.




```
        Box-Ljung test

data:  nn_fit_all$residuals
X-squared = 8.3282, df = 10, p-value = 0.5968
```

*Figure 3: Residuals of the fit with statistical tests*

We check the plot of the residuals, their ACF and histogram. The histogram of the **residuals approximately fits a normal distribution**. The ACF displays the residuals, and the all the **spikes are within the significance bound**. This is a good indication of **stationarity**. This is further confirmed by a statistical test like **Box-Ljung test** showing a p-value of **0.5968.** Setting a frequency of 7 has allowed the **NNAR(P,p,k)** model to fit the seasonality as well with **P=22, p=1 and k=14 and frequency=7**. There appears to be a **better** fit due to the addition of the weather data as ARIMA errors for our model, although we can only verify if this **translates to**

**better forecasting** with the RMSE score. With the addition of predictors to our model, the RMSE score for our forecast decreased from **188.17** to **165.20**. The **lower RMSE score** indicates a better fit for the model.

c) The forecast looks reasonable. The red line indicates fitted values, and the model appears to fit the data well. The black line indicates the actual values, and the blue line indicates the forecasted values. The model closely predicts the downward trend for the month of December and is also able to **predict the weekly variations** in average ride times. The forecast appears to be close to the actual values. The **prediction interval is also minimal** which indicates a good fit. We can further confirm this with the RMSE scores of various models. The RMSE of **the best SARIMA model** is **188.18**, Holt-Winters model is **206.29** and **NNAR(22,1,14)[7]** with weather data as predictors is **165.20.** All of these models result in white noise for the residuals, but **NNAR(22,1,14)[7**] is the model with the lowest RMSE score. For this model, precipitation, windspeed, maximum, minimum and average temperatures were combined as the predictors. There was a significant decrease in the RMSE score for **NNAR(22,1,14)[7]** with the addition of weather data as **ARIMA errors.** Thus, we can confirm that the overall weather for the day influences the forecast of our model.

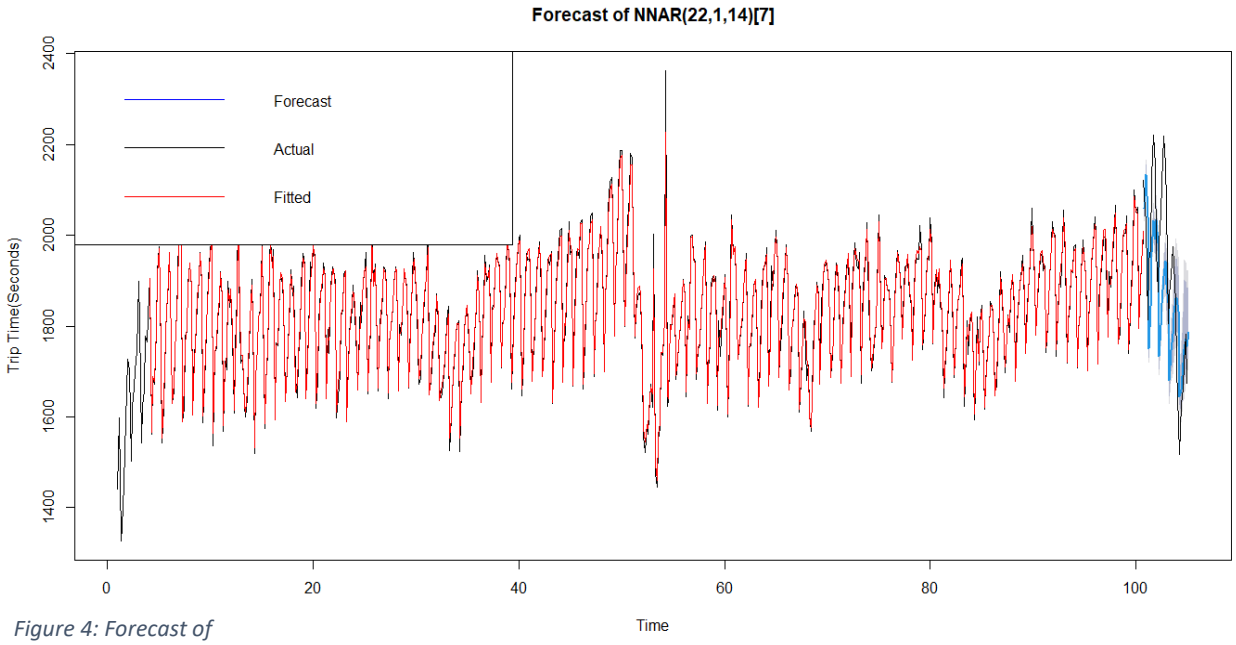



*Figure 4: Forecast of NNAR(22,1,14)[7] with ARIMA errors*

d) There are a few key limitations to our analysis:

- There is wider scope when choosing a model, but only a few key models were chosen depending on the ACF and PACF of the mean travel times. The outliers and NA values also affect the forecast results.
- The NNAR(P,p,k) model automatically chooses the values for P, p and k. Only the addition of weather data as the predictors was done manually. There is further room for improvement on this model since we may be able to find **optimal values for P,p and k** manually instead.
- The weather data has improved forecasts on the NNAR model, but we are only limited to using weather data to fit regression errors. There may be many other external factors affecting the data.
- The predictors(weather data) seem to improve forecasting(RMSE value) only for the NNAR model and not the SARIMA model. This will require further analysis.