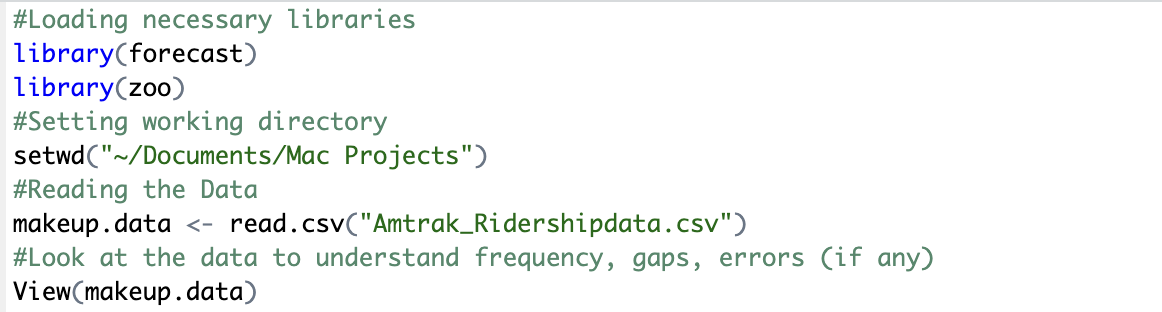
Amtrak Ridership Demand Forecast

Executive Summary: We deploy three models namely Linear, Quadratic and Holt Winter’s to see which one predicts the Amtrak ridership demand with least forecasting errors.

Data set: A csv file containing Demand values against dates from Jan 2000 to March 2022

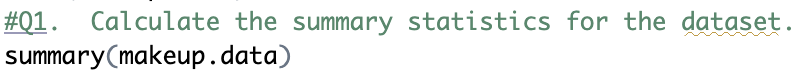
## Process Flow in Rstudio:



A screenshot of a calendar

Description automatically generated

We can see that the data is a repetitive calender month and year from 2000 Jan to 2022 March



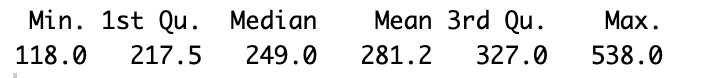
A white background with black text

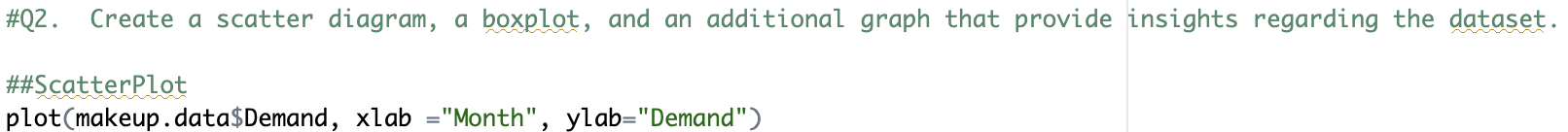
Description automatically generated

From above we can see that Demand values range (118,538) in the time interval and Median of the data points is 281.2

**A computer screen shot of a computer code

Description automatically generated**

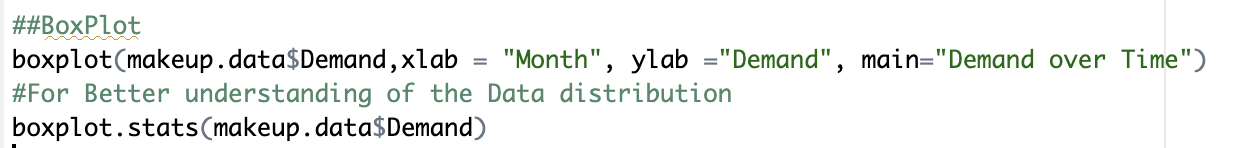




A graph of dots and numbers

Description automatically generated

We are not able to observe anything from this plot as data is all over the place, so let’s try using boxplot next.



A graph of a diagram

Description automatically generated with medium confidence

A white background with numbers

Description automatically generated

Boxplot and the stats following that provide us some vital information such as mean, median, min and max visually. We can observe the tails on the top seem to extend, but the data is concentrated on lower numbers. We can also observe the number of outliers in the data on the upper end of maximum value.

A screenshot of a computer code

Description automatically generated

A graph showing time and time

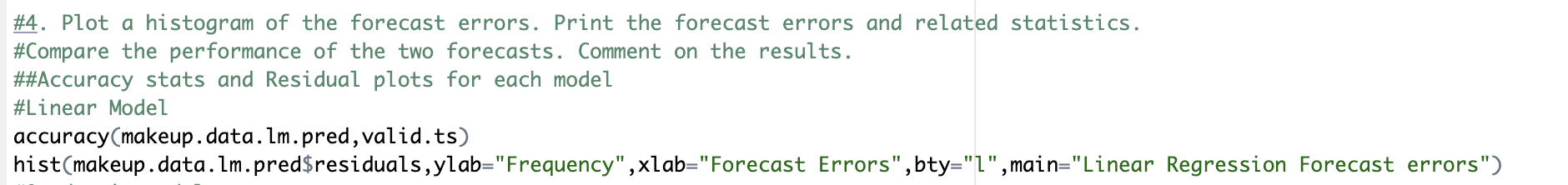
Description automatically generated

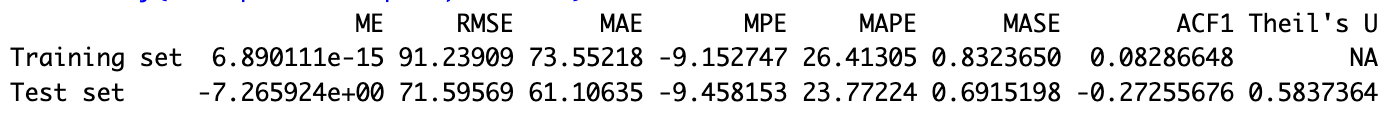
By creating this time series object and plot, we can visualize the data and make some observations related to the trend, seasonality etc. There seems to be a slight upward trend with time, seasonality is evident as there is lot of crests and troughs in graph and all of them seem to be happening cyclical in a calendar year.

A screenshot of a computer code

Description automatically generated

I have chosen three models namely: Linear regression, quadratic and Holt winter’s for the forecast.

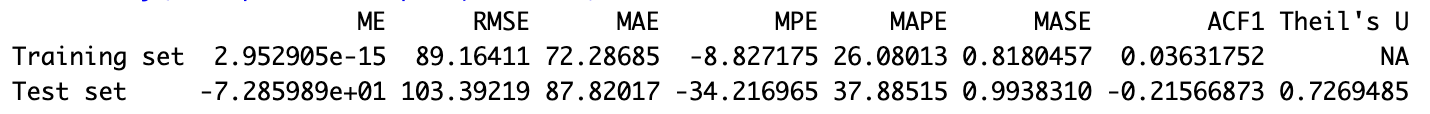




A graph of a graph

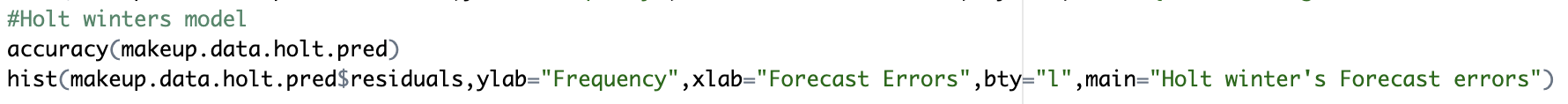
Description automatically generated

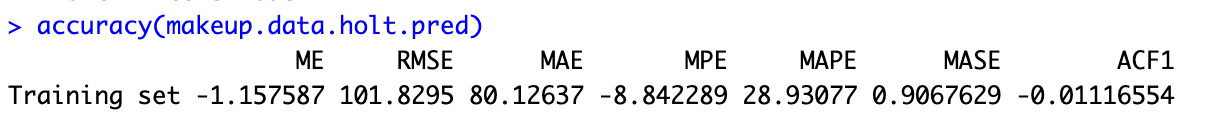




A graph of a graph

Description automatically generated





A graph of a forecast error

Description automatically generated

Looking at the accuracy stats and Residual plots of three models we have employed we can make some inferences.

1. Mean Error seems to be lowest for Holt winters model.
2. RMSE value of Linear regression seems to be the least, but we know that its not the best model as it doesn’t capture seasonality. Which also shows how a standalone parameter cannot decide a model’s performance.
3. ACF value is very low for Holt’s showing very less correlation to the immediately preceding values, but MASE is closest to 1 among the three showing that its comparable to Naïve Forecast, meaning there is still scope for improvement and we need to refine the model for optimal results.
4. Looking at the residual plots using histograms shows us that none of them seem to be normally distributed, which means they are not random noise and there seems to be some trend embedded in them. Primarily all three are left skewed or positively skewed which indicates that the models seem to be underestimating the values and might produce lower forecasts than actual observations.

**Conclusion:** Among the three Holt winters gives better results but looking at the residual plots, we must try to normalize the residuals by implementing some nonlinear models to capture the relationship of the observations in a better way, which might have an impact on the performance and give us optimal results.

Green Box: R code, Blue Box: Output