**Time Series Amtrak Ridership Analysis**

Jimmy Nguyen

Luke Awino

University of San Diego

ADS-506 Applied Time Series Analysis

4 December 2021

# TABLE OF CONTENTS

## Problem Statement 4

Justification 4

Literature Review 4

Data Exploration 6

Data Pre-processing 7

Data Splitting 9

Model Strategies 9

Validation and Testing 11

insert section name here 11

Evaluate Performance 11

Results and Final Model Selection 12

## Conclusion 11

## [Reference](#_44sinio) list 12

## [Appendix A](#_2s8eyo1) 13

Table A1. Information about the Data Set. 13

## 

## Table A2. Insert name here . 14

## Table A3. Insert name here 14

Figure A1. Insert a figure here for later 15 Figure A2. Insert figure here for later 16

[Appendix B](#_35nkun2) 17

Table B1. Insert name here 17

Appendix C 18

RMarkDown Notebook 18

# 

# Problem Statement

The Amtrak client would like to know the number of passengers for the next 3 months. Based on historical data, Amtrak would like to model the behavior of ridership from 1991 to 1999. This will allow Amtrak to understand how well the new high-speed rail service affected the northeast U.S. from 2000 to 2013. The ridership data is collected monthly, reporting the number of passengers. The task entails forecasting models to predict the number of passengers for the next 3 months in 2013. The forecasted number of passengers will determine the necessary adjustments to make improvements for the new service system.

# Justification

The purpose of this project is to forecast Amtrak ridership. With this information, Amtrak can better plan how to allocate funds, mainly for changing the routes and equipment management. “Deciding whether or not to introduce a new passenger train service or make changes to existing service comes down to one question: how many more people will ride the train after the changes are made? (Kenton, 2015). Being able to forecast which routes are close to their maximum ridership, Amtrak can prioritize these routes by adding rail cars and or more routes. This will increase profitability; using this information, routes and rail cars can be added to busy routes, increasing capacity for more paying customers who may have had to seek other means of transportation by mode or competitor. Amtrak, with this forecasting information, can reallocate resources from slower routes to busier routes, thereby maximizing their profits by either cutting down on routes or reallocating rail cars.

The outcome of this project will enable Amtrak to allocate vital resources for operations efficiently; Forecasting also helps plan maintenance during slower periods or seasons. Amtrak is also fine-tuning its analysis of the effects of a host of service changes on ridership, including on-time performance, station improvements, and changes in onboard amenities (Kenton, 2015).

# Literature Review

On December 11, 2000, Amtrak introduced the US’s only high-speed rail service nicknamed Acela, which is still the only high-speed rail service in the US (Waite, 2021). Having made a significant investment in the service, Amtrak needs demand forecasts to maximize revenue.

Demand forecast involves using past historical data to make a prediction or estimation of future demand. In this case, the goal is to make a prediction of future demand for the next 3 months of service, using time series analysis to make the forecast.

The goal of this literature review is to find other research that has been done for forecasting ridership across different transportation modes.

**Airline**

Over the past several years, airline travel has increased in popularity in Indonesia. To keep up with the demand, forecasting has become very crucial in planning for strategic growth. The Airline industry in Indonesia in 2016 reached the 2nd highest after China, and it’s predicted to have a high passenger surge in 2037 (Ramadhani et al., 2020). Forecasting also affects fleet planning, route planning, and annual operation planning; poor forecasting will positively impact the company’s revenue (Ramadhani et al., 2020).

Due to significant pattern components within the data, such as “trend, seasonality, cyclicity, and irregularity, time series was chosen as the forecasting method (Ramadhani et al., 2020). The time series model ARIMA was compared to the Neural network model to see which model would perform better; previous research has shown that neural networks had outperformed the ARIMA models. Monthly data from two of the most profitable routes over five years were used.

The metric for success was the model with the lowest minimum error; The results showed that the neural network model had the lowest minimum errors for both routes, the study has certain limitations, the parameters were manually combined using trial and error, and the data interval used was not focused on daily or weekly demand (Ramadhani et al., 2020).

**Bus**

Chinese urban public transportation has improved dramatically over the past decades; according to Ye et al. (2019), rapid economic development has played a role in this. Data from the China Academy of Transportation Science in 2017 shows that it has reached 55.9 million kilometers or four times its total highway mileage (Ye et al., 2019). Forecasting became critical after investigations on passenger demand showed that even with the demand, several popular routes were often in short supply; forecasting the demand will help the bus companies maximize the profit.

The main research objective was to predict the daily passenger volume in Jiaozuo city in China based on the ticket sales records; an experiment was performed to see if using just the weekday data would result in higher prediction accuracy (Ye et al., 2019).

The results showed that the data from the weekend was necessary for accuracy prediction, Ye et al., (2019) arrived at this conclusion after observing 3 months’ worth of data. Using 3 months of data limited the study by not being long enough to analyze the impact of patterns such as trends and seasonality on the dataset (Ye et al., 2019).

**Train**

Guo et al., (2010) studied the ARIMA and Holt-Winters model to forecast China’s monthly freight. Railway freight has shown volatility due to international fluctuations in the industrial product sector; most of the predictions are focused on the medium- and long-term forecast, and as such, the goal of this article is to build a model to forecast the short-term freight (Guo et al.,2010).

The data had 96 samples and it showed seasonal fluctuations with a steady growth trend, which could also influence the short-term forecast results.Both models performed well in the short-term forecast. The Arima model performed better due to its lower minimum error and the estimates being closer to the actual real values than the Holt-Winters model (Guo et al.,2010).

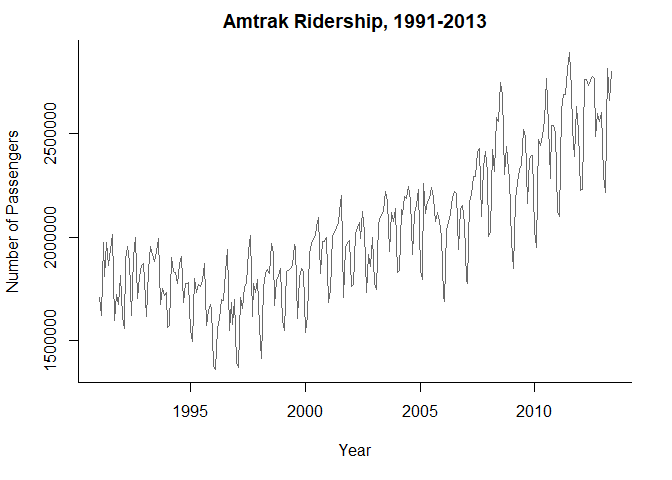
In conclusion, most of the existing research has been done using the ARIMA model, and it has made valuable predictions when it comes to forecasting. Trends, seasonality need to be taken into account when modeling.

# Data Exploration

The data set was obtained fromthe Bureau of Transportation Statistics which includes monthly Amtrak ridership data from January 1991 until May 2013. There are 269 rows with only two columns containing the timestamp of each month in that respective year and the total counts of ridership. During the exploratory data analysis, a time series trend was visible over the years. This series also shows signs of seasonality where this may be hypothesized as non-stationary.

**Figure 1.**

*Amtrak Ridership (1991-2013).*

****

*Note.* This figure shows the time series plot of the ridership data Amtrak has collected monthly over the time periods of January 1991 to May 2013.

# Data Pre-processing

The first steps were renamed to the original columns of the data set to Dates and Number\_of\_Passengers. Then the data was converted to a time series object to allow for further time series analysis. The frequency of the time series object was set to 12 due to the observations being monthly each year.

## Simple Smoothing Methods

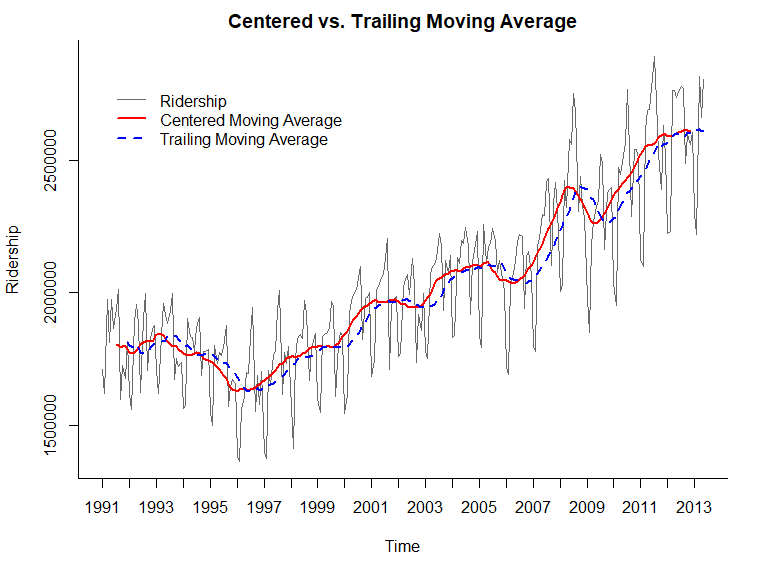
At the beginning of the data exploration, time components such as seasonality and trend are present over time. Thus, a good approach may be data-driven methods because it deals with data without a predetermined structure. An example of using a simple smoothing method is using the moving average. The moving average is a simple smoothing method, containing average values across a time window, *w*, specified by the user. Moving average works by suppressing seasonality and noise in a dataset. However, there are two types of moving averages: the centered moving average and the trailing moving average.

## Centered vs. Trailing Moving Average

The centered-moving average is useful for visualizing trends and the trailing moving average is useful for forecasting. These two moving averages will be compared against each other in Figure 2.

**Figure 2.**

*Centered vs. Trailing Moving Average*

****

*Note.* Twelve was the selected time window and this indicates the length of the seasonal cycle. This figure shows somewhat of a global U-shape, but the moving average looks to increase as the year progresses.

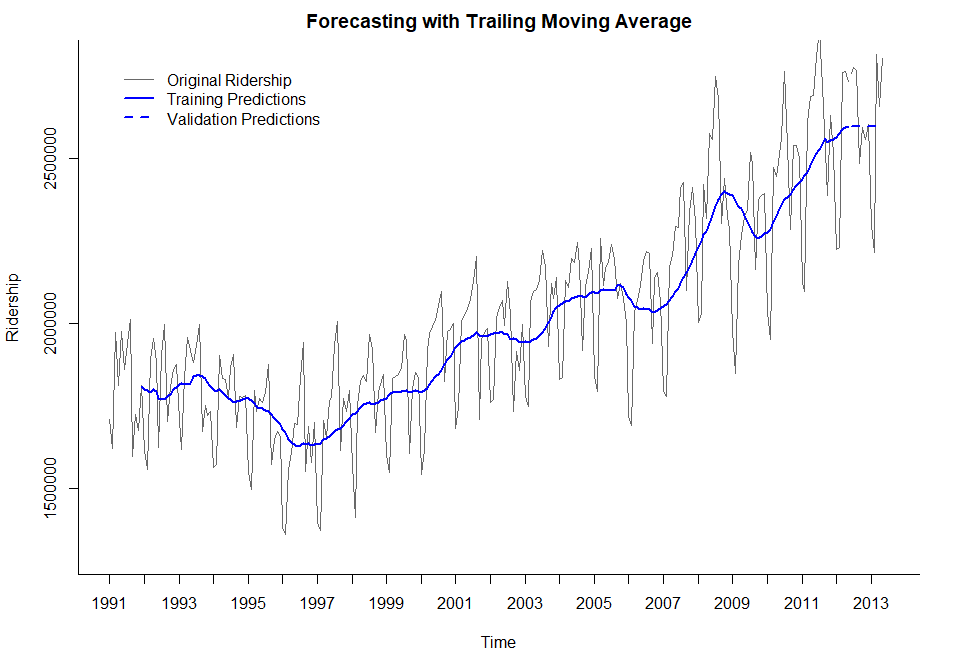
Since the centered-moving average uses past and future data, this is a limitation for making forecasts when future information is unavailable. Therefore, the better approach is to move forward with the trailing average for forecasting. This method selects the window that is placed over the most recent values.

## Trailing Moving Average Forecasting

In Figure 3 below, the trailing moving average does not perform well with forecasting on this time series data. The first thing to notice is that the forecasts for all the months in the validation period denoted in blue dashes are identical because this method is not roll-forward next month forecasts. It is clear that the trailing moving average forecaster is inadequate for the Amtrak monthly forecast task.

**Figure 3.**

*Forecasting with a Trailing Moving Average*



*Note.* The forecast is constant during the validation period, which means it does not capture the trend and seasonality of the data.

The reason why is because it does not capture the seasonality in the data. The forecaster predicted seasons with high ridership with lower ridership and seasons with low ridership with high ridership. This occurs because the moving average lags behind when forecasting a time series with a trend. Therefore, over-forecasting and under-forecasting in the presence of increasing and decreasing trends. Between the smoothing methods of moving averages, it should only be used for forecasting when a series lacks seasonality and trend, which is not true here for the Amtrak ridership data.

**Alternative methods for Non-Stationary Series**

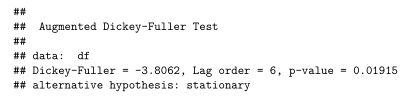
Differencing is a popular method for removing trend or seasonality patterns by taking the difference between two values in a series. For example, the lag-1 difference takes the difference between every two executive values in time. On the other hand, differencing at lag-k takes the difference from every value from k-periods back. However, before using differencing as a pre-processing step, a Dickey-Fuller test should be used to check if the assumptions for a series are actually non-stationary.

## Dickey-Fuller Test

The Dickey-Fuller test checks if a time series data is stationary or not. Although it was clear during the data exploration that this series was not stationary. Figure 4 below shows the results of a Dickey-Fuller test.

**Figure 4.**

*Results of the Dickey-Fuller Test*



*Note.* The results here indicate that a p-value was less than the chosen level of significance at 0.05. This means that the Dickey-Fuller test rejects the null hypothesis stating that it is actually statistically significant as a stationary series instead of non-stationary.

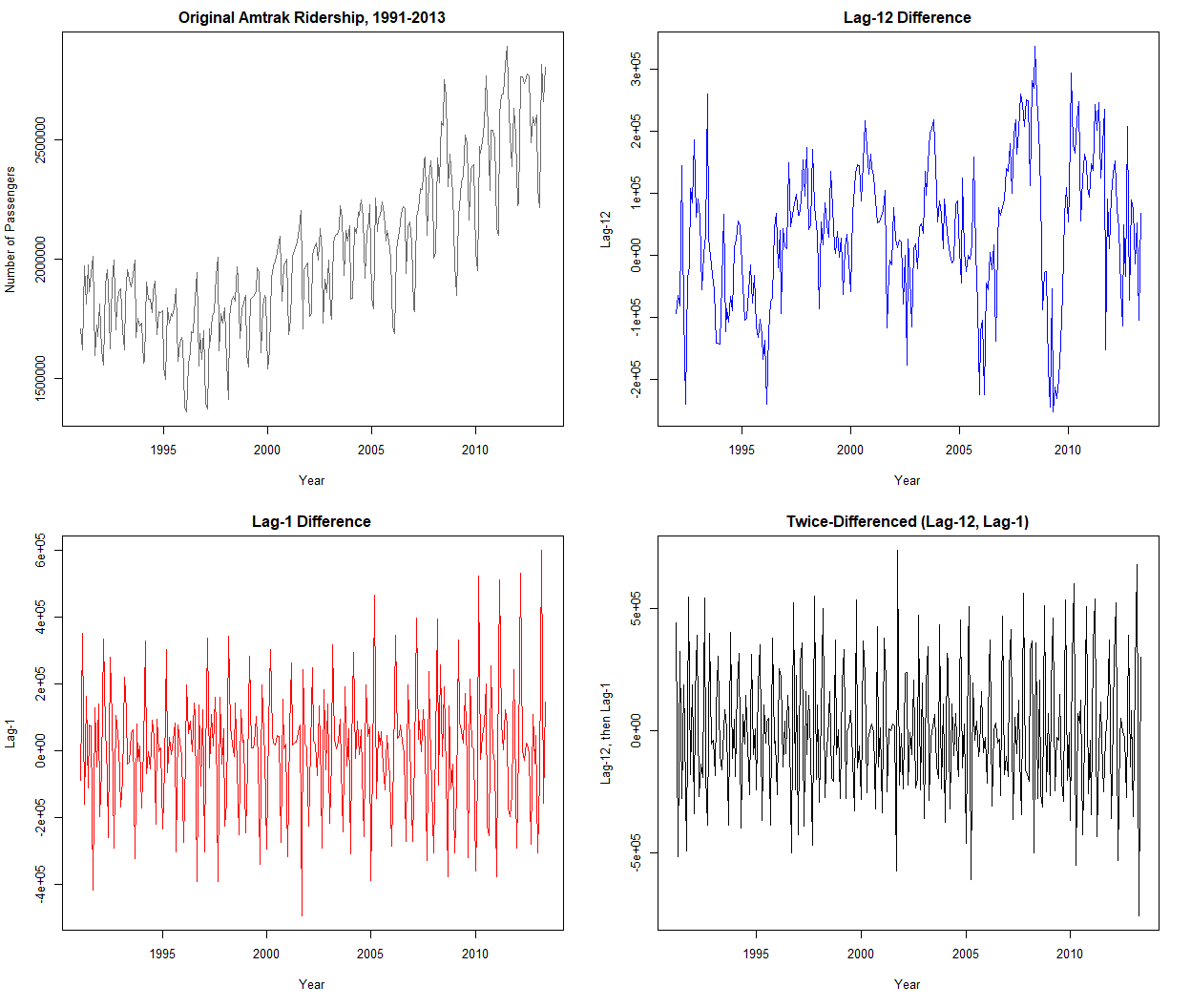
Although the dataset visually appears to be non-stationary, the Dickey-Fuller test rejects the null hypothesis in favor of the alternative hypothesis that the dataset is indeed stationary. However, these results are from a biased test with a type one error. A stationary time series has a constant mean, minimal seasonality effect, and no variance increase over time. Meanwhile, the Amtrak ridership data set has no constant mean with signs of seasonality and an increasing trend as the years progress. Therefore, the next step is to continue with differencing for pre-processing the series.

## Detrending and Dseasonalizing

There are different approaches to making a time series stationary, such as detrending and deseasonalizing. Detrending can be used by the lag-1 difference of a series. This would remove the somewhat U-shape of the Amtrak ridership series. Deasonalizing would remove the seasonality by using a lag-12 difference series. A k-period of 12 was used because of monthly data over the years. When both of these components exist, double differencing can be applied to the series. Figure 5 below shows the comparison between the original data, the lag-12 difference, the lag-1 difference, and lastly the double-differenced data.

**Figure 5.**

*Comparisons between Different Differenced Data*

*Note.* Since there are both seasonality and trend, the double differencing effect on the bottom right resulted in a series without trend or monthly seasonality.

The first plot on the top left shows the original Amtrak ridership with trend and seasonality. The lag-1 plot on the bottom left has been differenced by a month to remove the trend. While the lag-12 different plot on the top right was adjusted to remove monthly seasonality. Lastly, the bottom right panel has been double-differenced to remove both trend and monthly seasonality.

# Data Splitting

## Fixed-Partitioning

Forecasting with time series data also requires the appropriate partitioning to evaluate performance of the models. Similar to cross-sectional, there will be a training set, a validation set, and the test set. Although partitioning data into these sets are usually done randomly, this is not the case for a time series. The first step is to decide the different training and validation periods. Meaning, the time window will need to be specified before partitioning into different sets. The training set consists of the earlier periods, while the validation set is the periods the model has not yet been trained on. This will be the later period. For the test period, this partitioning set will be the forecasted months the client has requested. Thus, the data splitting strategy for the Amtrak ridership data will be a fixed-partitioning method. The training period will be from January 1991 to May 2012. The validation period will be from June 2012 to May 2013. While the test period Amtrak requested is from June 2013 to August 2013.

# Model Strategies

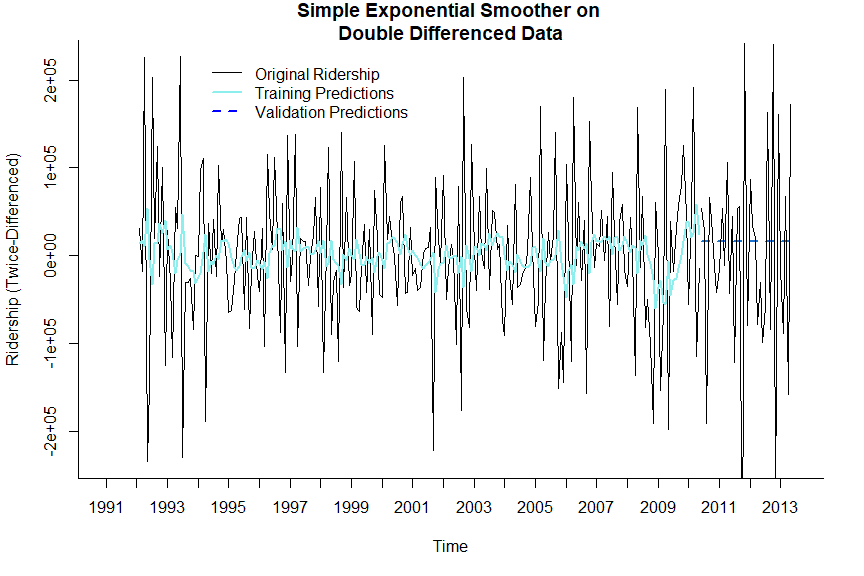
## Simple Exponential Smoothing

Exponential smoothing works in a similar manner to forecasting with a moving average, although it differs by taking the weighted average of all past values of a series. The weights will decrease exponentially into the past and this is valuable as it gives more weight to recent data more than past data. This is a popular method due to its low computational costs, easy automation, and good performance. Exponential smoothing works well with data without trend or seasonality, so it is a logical model for the double-differenced series.

Figure 5 below shows the exponential smoother on the double-differenced data. This smoothing method is under the *ets* framework using the model called additive (A), no trend (N), and no seasonality (N), denoted as “ANN”.

**Figure 6.**

*Simple Exponential smoothing*

**

*Note.* The forecasts for the validation set remained a singular value similar to the moving average forecast model.

Both moving average and exponential smoothing models were not able to have accurate forecasts, even though the pre-processing was already performed. Another solution is to use a more complex and sophisticated exponential smoothing that is able to model data with both trend and seasonality instead of removing them.

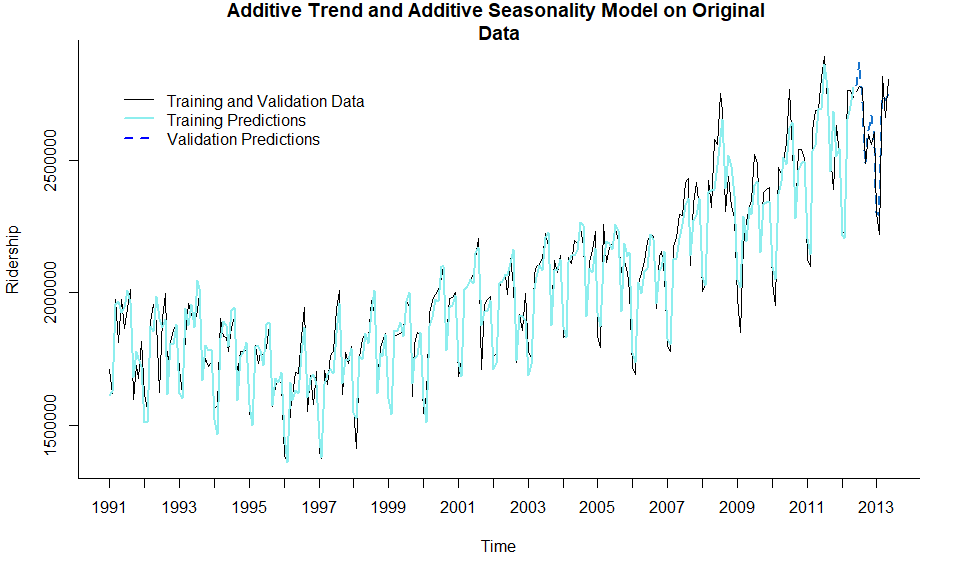
## Models with Additive or Multiplicative Trends

Double exponential smoothing can be used on a series that contains an additive trend. This is also called Holt's linear trend model where the local trend is estimated and is updated as more data comes in. Although additive trend models assume the level changes from one period to the next by a fixed amount, multiplicative trends assume it changes by a factor instead.

A further extension of the double exponential smoothing where the k-step-ahead forecasts also takes into consideration the seasonality of the current period. While the trend is considered from the additive and multiplicative seasonality. This is an adaptive method that allows the components (levels, trends, and seasonality) to change over time. Therefore, the Amtrak ridership series would not need the pre-processing performed previously. Figure 7 shows this model fitted to the training periods and made forecasts on the validation periods.

**Figure 7.**

*Additive Trend and Additive Seasonality Model*

**

*Note.* This approach uses Holt-Winters method to build an additive trend and seasonality model with multiplicative error on the original data without a second order difference.

This falls under the ETS framework using the M.A.A model. However, what if the choice was to use a multiplicative trend and seasonality model or try different combinations of these components.

## Automated Model Selection for the Optimal Exponential Model

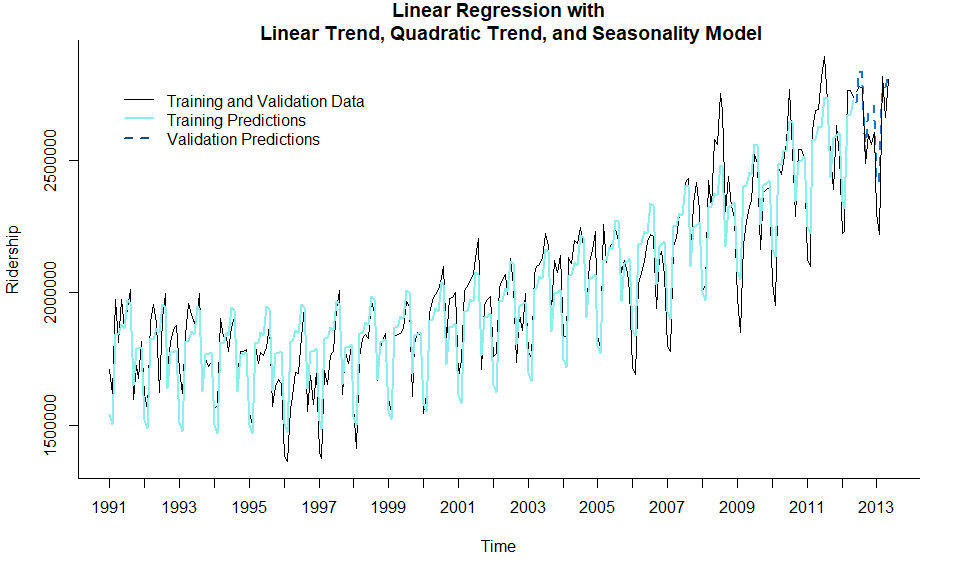
The optimal model chosen was a multiplicative error, additive trend, and multiplicative seasonality with an improved AIC score of 7155.16 which was lower than the previous MAA model with an AIC of 7181.09.

## Regression-Based Models

There are different types of common trends and seasonality that can be modeled by regression-based models estimated during the training period and used to forecast on future data. Such trends include linear, exponential, or polynomial, while the different seasonality are additive and multiplicative seasons. For example, a new model can be built with both linear and quadratic trends. Then a monthly seasonality can be added which is a factor of twelve, however it is redundant to use all twelve. In total, we will have thirteen predictors, with two for the linear and quadratic trends and the eleven dummy variables for each month, except January. Figure 8 shows what this model looks like.

**Figure 8.**

*Regression Model with Linear and quadratic trends and Monthly Seasonality.*

**

*Note.* This model uses a linear and quadratic trend along with the 11 monthly seasons as dummy variables, excluding January.

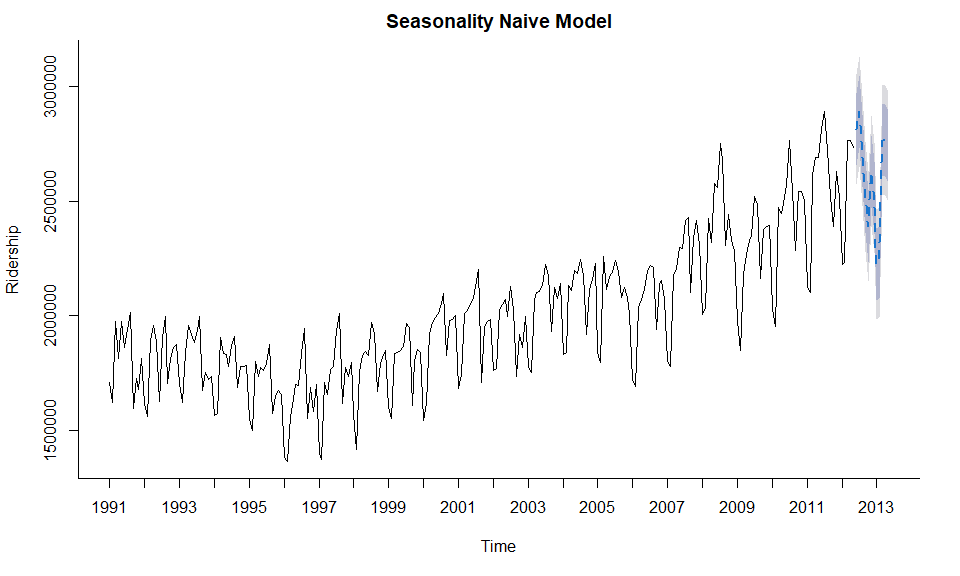
The AIC score has dropped significantly compared to the model with only linear trend. In comparison with the exponential smoothing model, this score has dropped drastically than before.

## Naive Forecast Models

Instead of sophisticated modeling, naive forecast models output the most recent values of the series. While a seasonal naive forecast model uses the recent value from the most identical season. For example, the forecast for August 2012 would be the value from August 2011 instead. Also, naive forecast models actually have great performance and it's easy to understand and implement. Although this seasonality naive model in Figure 9 will be used as a baseline model to compare other models predictive performances.

**Figure 9.**

*Seasonal Naive Forecasting Model*

**

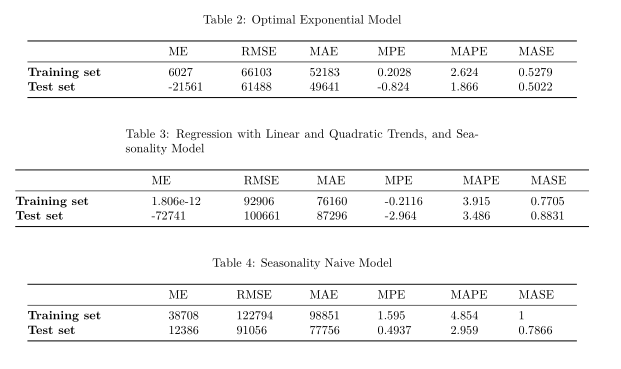
*Note.* This naive model will only be used as a baseline model to compare performance with the other models built previously, such as the optimal automated model selection and the regression-based model.

# Results and Final Model Selection

Based on the results, there are a handful of different performance metrics to evaluate across the models. The following models are competing against each other: the seasonal naive forecast model, the optimal exponential model, and the regression model with a monthly seasonality and linear and quadratic trends. The selected performance metric is the root mean square error (RMSE). This measure with lower values indicates a model with higher performance. A benefit of using this performance metric over the others is because the RMSE will be in the same units as the original data. The following tables in Table 1. displays the different performance metrics of the models stated previously.

**Table 1.**

*Model Performance Results*

**

*Note.* The primary performance metric that will be evaluated on is the RMSE.

The key performance is RMSE in the test set. The baseline performance will be the standard score to beat if there is a lower RMSE value. So the seasonal naive model has a RMSE score of 91,056 in the test set. For the optimal exponential model, the test set’s RMSE score is 61,488. On the other hand, the regression-based model has the highest RMSE score of 100,061, indicating the worst performance of the three. The final model selected to forecast the months of June to August of 2013 will be the optimal exponential model.

# Conclusion

## Findings

insert here.

## Suggestions

insert here

# References

*Amtrak Ridership*. Bts.gov. (2013). Retrieved 4 December 2021, from <https://www.bts.gov/archive/publications/multimodal_transportation_indicators/2013_08/passenger/amtrak_ridership>.

Waite, M. (n.d.). *Why the US needs to get on track with high-speed rail*. Greenbiz. Retrieved November 21, 2021, from <https://www.greenbiz.com/article/why-us-needs-get-track-high-speed-rail>.

Ramadhani, S., Dhini, A., & Laoh, E. (2020). Airline Passenger Forecasting using ARIMA and Artificial Neural Networks Approaches. *2020 International Conference on ICT for Smart Society (ICISS), ICT for Smart Society (ICISS), 2020 International Conference On*, *CFP2013V*–*ART*, 1–5.<https://doi-org.sandiego.idm.oclc.org/10.1109/ICISS50791.2020.9307571>

Ye, Y., Chen, L., & Xue, F. (2019). Passenger Flow Prediction in Bus Transportation System using ARIMA Models with Big Data. *2019 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2019 International Conference On*, 436–443. <https://doi-org.sandiego.idm.oclc.org/10.1109/CyberC.2019.00>

Yu-hua Guo, Xiu-peng Shi, & Xiao-dong Zhang. (2010). A study of short term forecasting of the railway freight volume in China using ARIMA and Holt-Winters models. *2010 8th International Conference on Supply Chain Management and Information, Supply Chain Management and Information Systems (SCMIS), 2010 8th International Conference On*, 1–6.

Kenton, M. (n.d.). *How do you predict how many will ride a train?* How do you predict how many will ride a train? - Trains Magazine - Trains News Wire, Railroad News, Railroad Industry News, Web Cams, and Forms. Retrieved November 30, 2021, from https://cs.trains.com/trn/b/observation-tower/archive/2015/01/31/how-do-you-predict-how-many-will-ride-a-train.aspx.

[Aditya Kadiwal](https://www.kaggle.com/adityakadiwal)

[Aditya Kadiwa](https://www.kaggle.com/adityakadiwal)

# Appendix A

## Table A1

*Information about the Data Set.*

## 

## Table A2

*Insert name here.*

## Table A3

*Insert name here.*

## Figure A1

*Insert figure title here*

*Note*.

## Figure A2

*Insert figure title here*

*Note*.

# Appendix B

## Table B1

*Insert name here.*

*Note.*

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

**Appendix C**