VIRGINIA COMMONWEALTH UNIVERSITY

FORECASTING METHODS

ASSIGNMENT 5

SATYANARAYAN VENKAT NALDIGA

V01108247

SUBMITTED TO-

PROF.JASON MERRICK

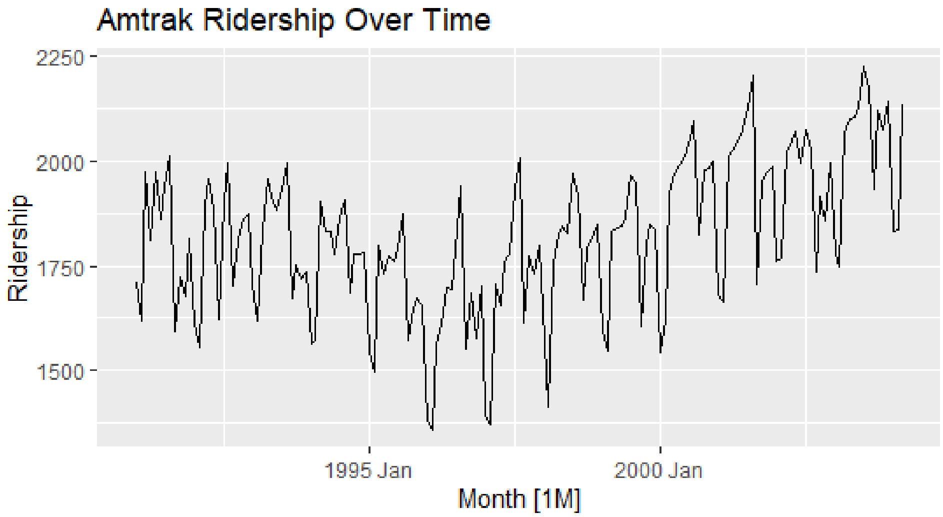
Date of Submission: 10-02-2024

Introduction

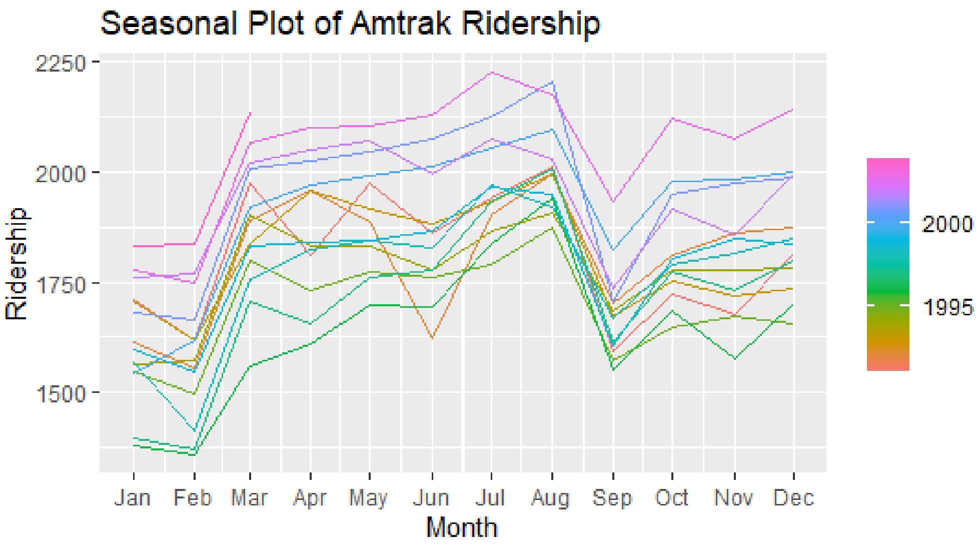
The Amtrak dataset provides a comprehensive look at the monthly ridership figures for Amtrak trains spanning from the early 1990s to the early 2000s. This dataset is invaluable for analyzing trends in public transportation usage, specifically in the context of national rail services in the United States. By examining Amtrak's ridership data, we can gain insights into how external factors such as seasons, economic conditions, and other socio-economic events influence public transport usage over time.

Analyzing such data helps in understanding the operational challenges and planning needs of rail services, facilitating better resource allocation and strategic planning to meet varying demand across different times of the year. The dataset also allows for the application of various statistical and forecasting models to predict future trends, which is crucial for improving service efficiency, scheduling, and overall customer satisfaction in public transport systems.

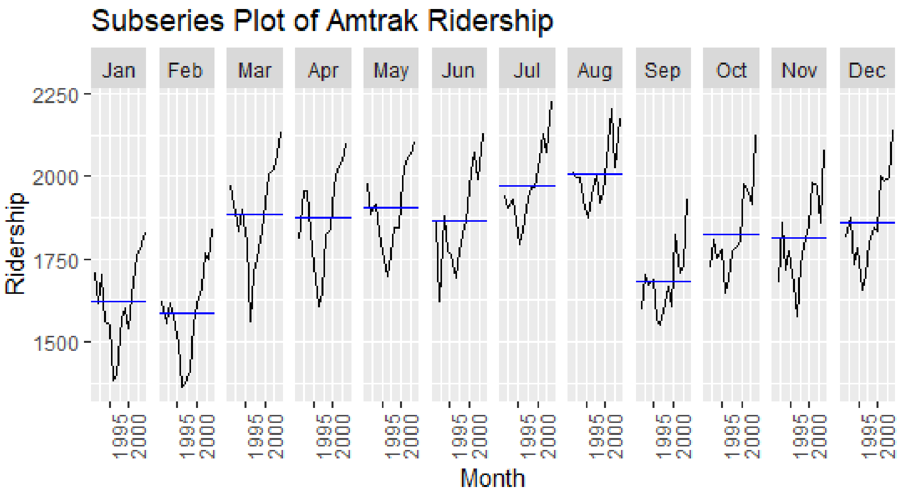
In the detailed analysis that follows, we delve into the ridership trends demonstrated in the dataset, highlighting seasonal variations and long-term shifts in ridership. We further explore the application of two different statistical models—a linear trend model and a quadratic trend model—to forecast future ridership patterns. This comparative analysis helps to determine which model more accurately captures the underlying trends in the data, thereby providing a better tool for forecasting future ridership needs.



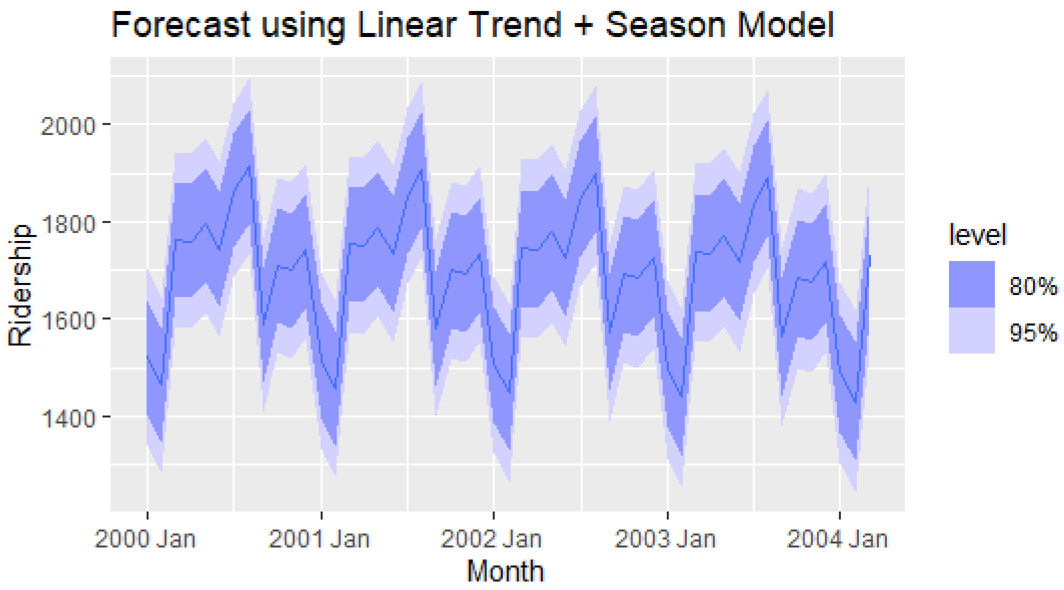
The series of graphs provide a detailed view of Amtrak's ridership patterns and forecast models. Starting with the "Amtrak Ridership Over Time" graph, it clearly shows the monthly ridership trends from early 1992 through early 2000, highlighting noticeable fluctuations that indicate strong seasonal influences. Peaks in ridership typically occur during the summer months, likely reflecting increased travel for vacations and leisure activities. Conversely, there is a notable dip during the winter months, which might be attributed to less favorable travel conditions and a decrease in tourist activity.



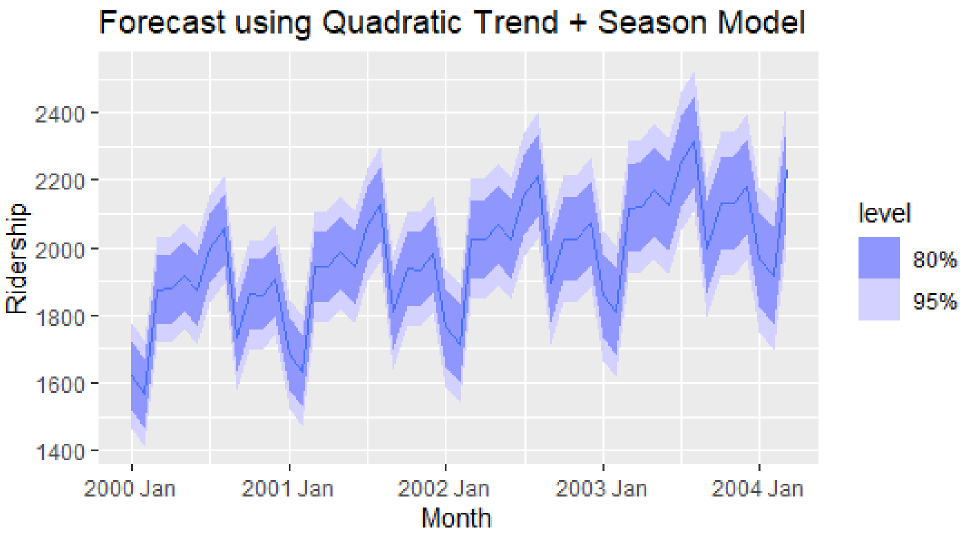
The "Seasonal Plot of Amtrak Ridership" graph further delineates these patterns, comparing monthly ridership across different years—1992, 1995, and 1997. This comparison shows that while the overall level of ridership varies each year, the seasonal trend of peaking in summer and dipping in winter remains consistent. This visualization underscores the predictable nature of travel habits influenced by seasonal variations.



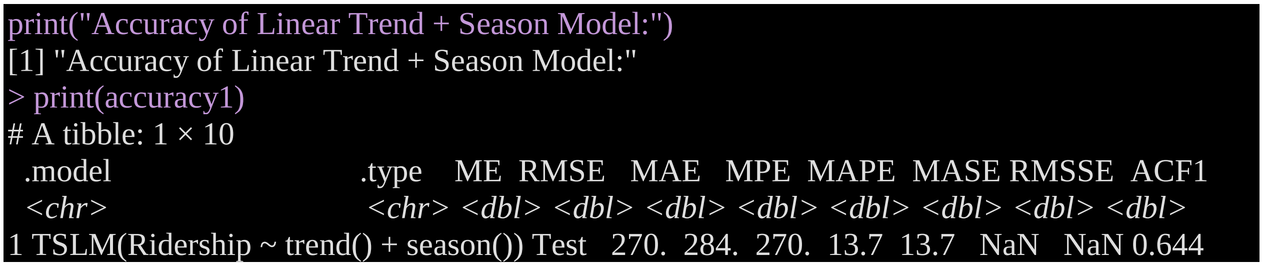
In the "Subseries Plot of Amtrak Ridership," the data are dissected further to show ridership distribution for each month over several years. This plot uses blue lines to indicate the mean ridership for each month, clearly illustrating the cyclical nature of travel with peaks in mid-year and troughs at the start and end of each year. Such a breakdown is helpful for identifying specific months where ridership significantly deviates from the average, indicating potential areas of interest for further analysis or targeted business strategies.



Turning to the forecasts, the "Forecast using Linear Trend + Season Model" graph shows predictions from 2000 to early 2004, incorporating both a linear trend and seasonal factors. The model's predictions, surrounded by 80% and 95% prediction intervals, suggest that while the model captures the general seasonal pattern, it projects a fairly conservative view of ridership changes over time. The intervals provide a statistical range indicating the reliability of the forecasts and highlighting the expected variability in ridership.



The "Forecast using Quadratic Trend + Season Model" presents a similar seasonal pattern but includes a quadratic trend to account for potential non-linear growth in ridership figures. This model predicts higher peaks in ridership compared to the linear model, suggesting an anticipation of growth or more pronounced seasonal impacts. The wider prediction intervals in this model point to greater uncertainty, possibly reflecting the increased complexity of predicting future trends with non-linear elements.



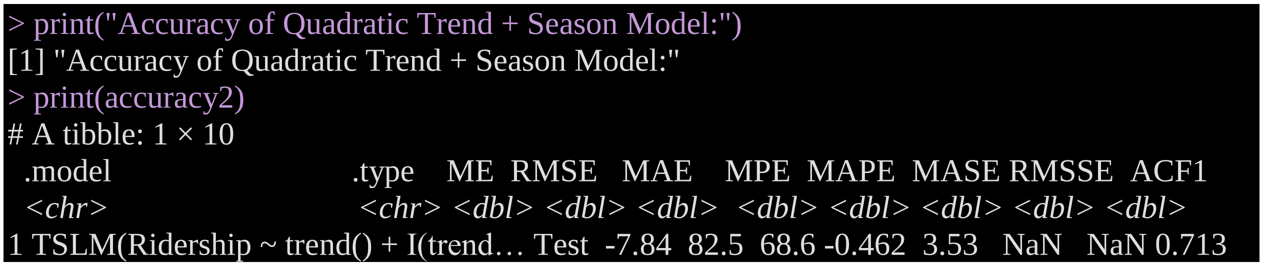
```
print("Accuracy of Linear Trend + Season Model:")
[1] "Accuracy of Linear Trend + Season Model:"
> print(accuracy1)
# A tibble: 1 × 10
  .model                          .type    ME  RMSE   MAE   MPE  MAPE  MASE RMSSE  ACF1
  <chr>                           <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 TSLM(Ridership ~ trend() + season()) Test   270.  284.  270.  13.7  13.7   NaN   NaN 0.644
```

Summary of the accuracy metrics for the "Linear Trend + Season Model" applied to the Amtrak ridership data. Here's a breakdown of each metric and what it indicates about the model's performance:

- **model**: This specifies the model used, which in this case includes a linear trend and seasonal components (TSLM(Ridership ~ trend() + season())).
- **.type**: Indicates that these statistics are calculated on the test dataset.
- **ME (Mean Error)**: The Mean Error is 270. This value indicates that, on average, the model's predictions are 270 units higher than the actual ridership values. A positive ME suggests a systematic overestimation by the model.

- **RMSE (Root Mean Squared Error)**: The RMSE is 284, which measures the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. A lower RMSE value is preferred as it indicates closer fit to the data.
- **MAE (Mean Absolute Error)**: The MAE is also 270, similar to the ME, indicating the average magnitude of the errors in a set of predictions, without considering their direction (positive or negative errors are treated the same). It's a measure of accuracy in a linear regression model.
- **MPE (Mean Percentage Error)**: The MPE is 13.7, suggesting that, on average, the model's predictions overestimate the actual values by 13.7%. This percentage error provides context for the ME by showing the proportion of error relative to actual values.
- **MAPE (Mean Absolute Percentage Error)**: The MAPE is also 13.7, similar to MPE, but it takes the absolute value of the percentage errors. This metric averages the absolute percentage errors by the actual values, giving insight into the accuracy without the direction of the errors, similar to MAE.
- **MASE (Mean Absolute Scaled Error)** and **RMSSE (Root Mean Squared Scaled Error)**: Both these values are shown as NaN, which typically indicates that there is insufficient data to compute the scale or that the denominator in the calculation of these metrics is zero. In practice, these would compare the MAE and RMSE to those of a naïve benchmark model.
- **ACF1 (First Autocorrelation of Errors)**: The ACF1 value is 0.644, indicating a moderate positive autocorrelation in the model residuals at lag 1. This suggests that the model's errors are correlated from one prediction to the next, which can be indicative of a trend or seasonal effect not fully captured by the model.

Overall, the positive ME and MPE values suggest that the model tends to overestimate ridership. While the RMSE and MAE indicate the model fits the data to a certain extent, the significant autocorrelation in the errors (ACF1) suggests that there may be additional patterns in the data that the model is not capturing, which could potentially be addressed by considering more complex or different types of models.



```
> print("Accuracy of Quadratic Trend + Season Model:")
[1] "Accuracy of Quadratic Trend + Season Model:"
> print(accuracy2)
# A tibble: 1 × 10
  .model                          .type    ME  RMSE   MAE    MPE  MAPE  MASE RMSSE  ACF1
  <chr>                           <chr> <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>
1 TSLM(Ridership ~ trend() + I(trend… Test  -7.84  82.5  68.6 -0.462  3.53   NaN   NaN 0.713
```

summary of the accuracy metrics for the "Quadratic Trend + Season Model" applied to the Amtrak ridership data. Let's break down each metric to understand the performance of this model:

- **model**: Specifies the model used, incorporating both a quadratic trend (I(trend()^2)) and seasonal components, which suggests a more complex model that accounts for non-linear changes in ridership over time.
- **.type**: Indicates that these statistics are calculated on the test dataset.
- **ME (Mean Error)**: The Mean Error is -7.84, indicating that, on average, the model's predictions are 7.84 units below the actual ridership values. Unlike the first model, this quadratic model tends to slightly underestimate the actual values.
- **RMSE (Root Mean Squared Error)**: The RMSE is 82.5. This value is significantly lower than that of the linear model, suggesting a better fit to the data. RMSE is crucial for understanding the spread of residuals; a lower value here indicates that the model's predictions are closer to the actual data points.
- **MAE (Mean Absolute Error)**: The MAE is 68.6, which is also lower than in the linear model. This metric provides a clear indication of the average magnitude of the errors in predictions, irrespective of their direction. The lower MAE value reflects more accurate predictions overall.
- **MPE (Mean Percentage Error)**: The MPE is -0.462%, which reflects a slight average underestimation of the actual values in percentage terms. This is a marked improvement in terms of bias compared to the linear model.
- **MAPE (Mean Absolute Percentage Error)**: The MAPE is 3.53%, much lower than the 13.7% of the linear model. This metric gives an average of the absolute percentage errors and is a good indicator of the model's accuracy in percentage terms, showing a significant improvement in the accuracy of predictions.
- **MASE (Mean Absolute Scaled Error)** and **RMSSE (Root Mean Squared Scaled Error)**: Both values are again shown as NaN, possibly due to calculation issues similar to the linear model. These would typically be useful for comparing the model's performance to a naïve benchmark.
- **ACF1 (First Autocorrelation of Errors)**: The ACF1 value is 0.713, which is higher than in the linear model. This indicates a stronger positive autocorrelation in the model residuals at lag 1. While this suggests that some error patterns are still being carried over from one prediction to the next, the overall lower error metrics indicate that the quadratic model captures more complexity and dynamics in the data.

In conclusion, the "Quadratic Trend + Season Model" provides a significant improvement in forecasting accuracy over the "Linear Trend + Season Model" as demonstrated by lower values of RMSE, MAE, and MAPE, despite the higher autocorrelation in residuals. This suggests that incorporating a quadratic term helps to better capture underlying trends and variations in the ridership data, though it might also introduce some dependency in the forecast errors.