577 – Time Series

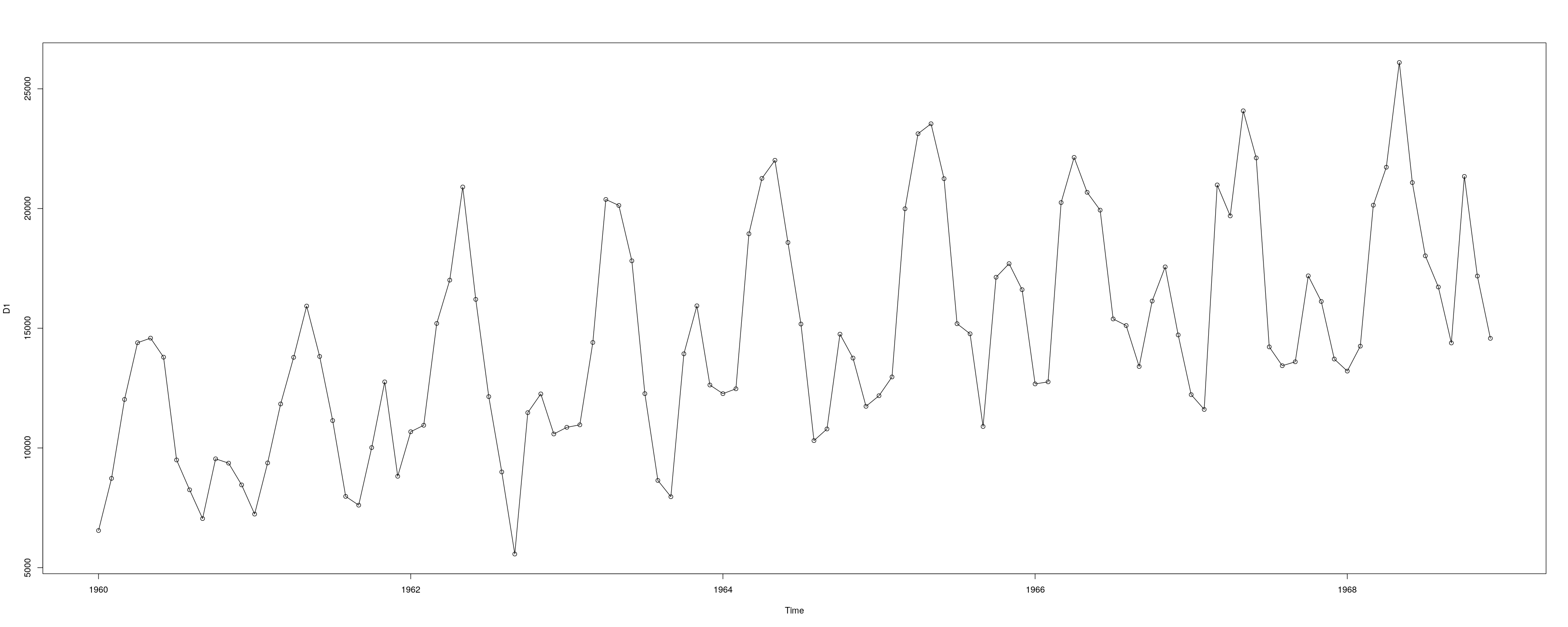
Final Project – Evaluating Car Data

Naeem Sonia

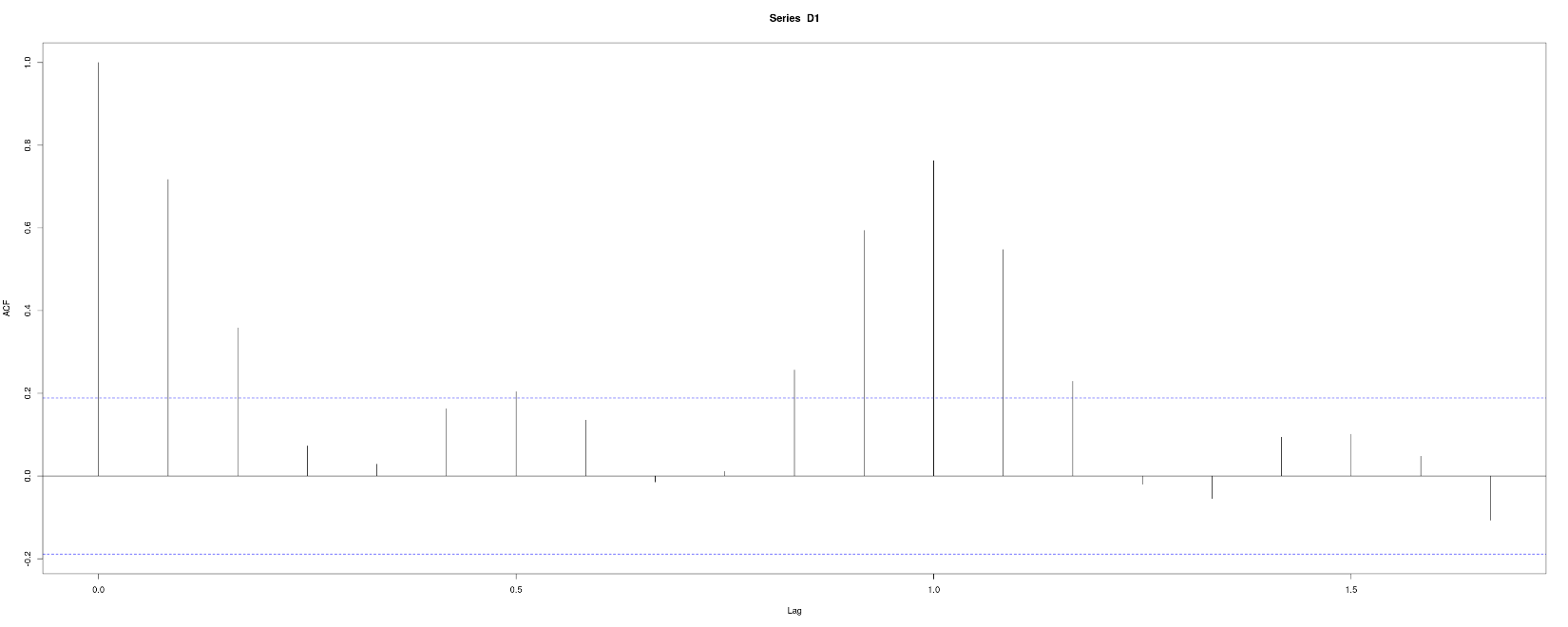
Eric Maibach

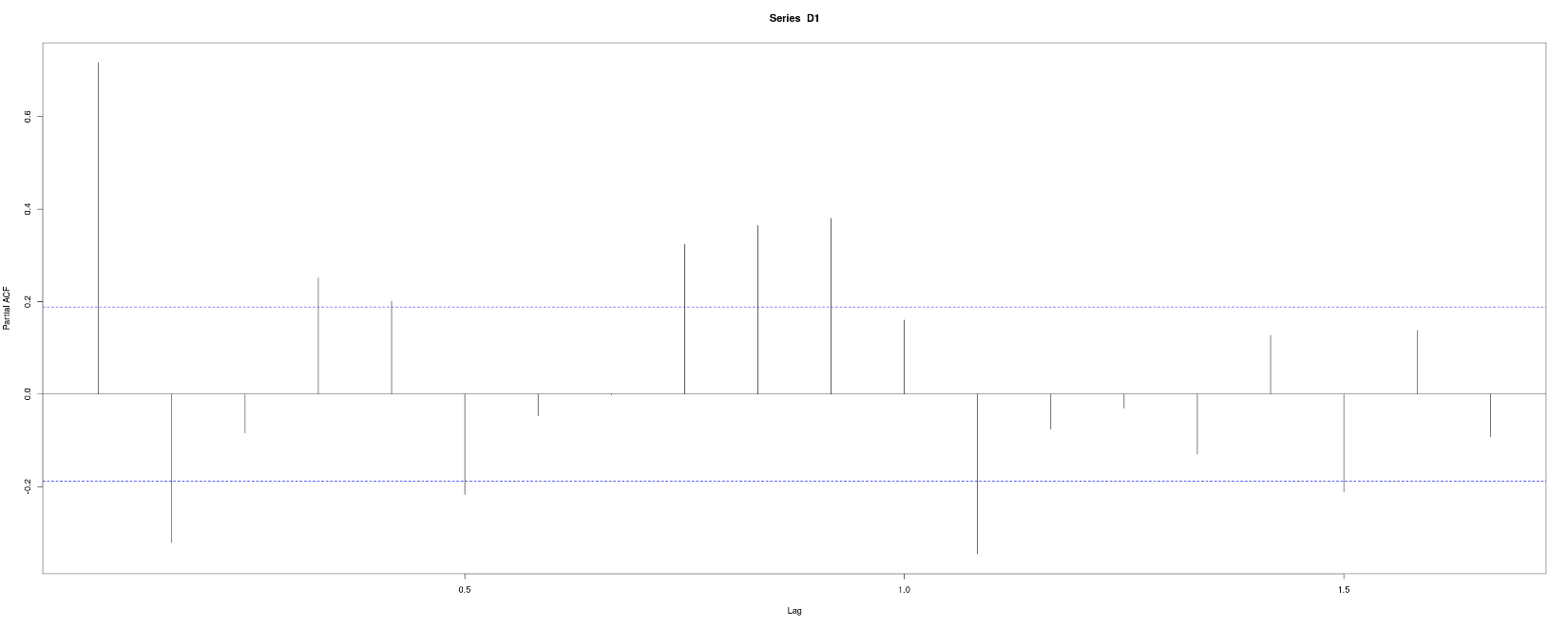
# Overview of Data

We will be evaluating a data set consisting of monthly car sales in Quebec from 1960 to 1968. The data was obtained from <https://nmimoto.github.io/datasets/car.csv>.



Looking at a graph of the data over time, it appears the data may have a seasonal component to it and may be increasing over time. The ACF and PACF graphs indicate some correlation, and seem to indicate a seasonal component to the time series.





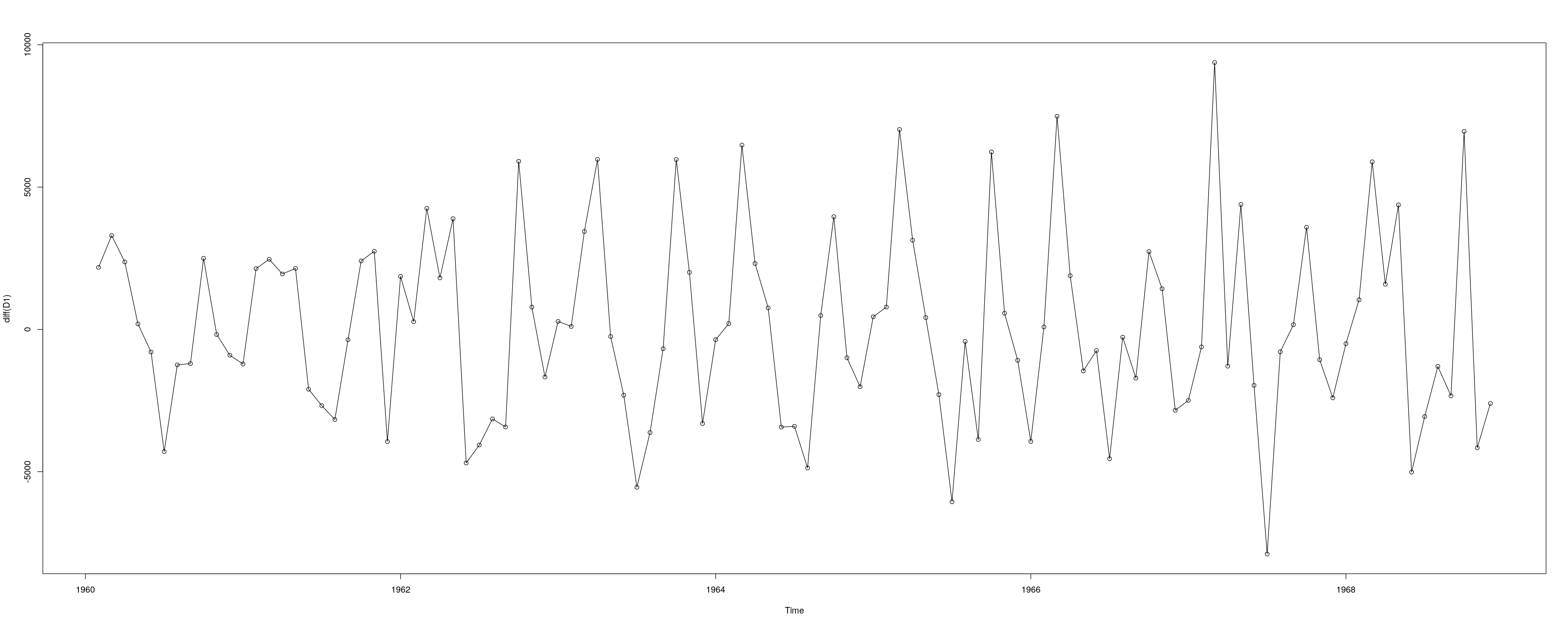
# Seasonality

Viewing graphs of the data, and viewing the ACF and PACF graphs, the data does appear to have a seasonal component to it. We tested models with and with out the seasonal component, and found that the seasonal models did perform better, and did a better job of modeling the significant correlation at lag 12.

# Stationality

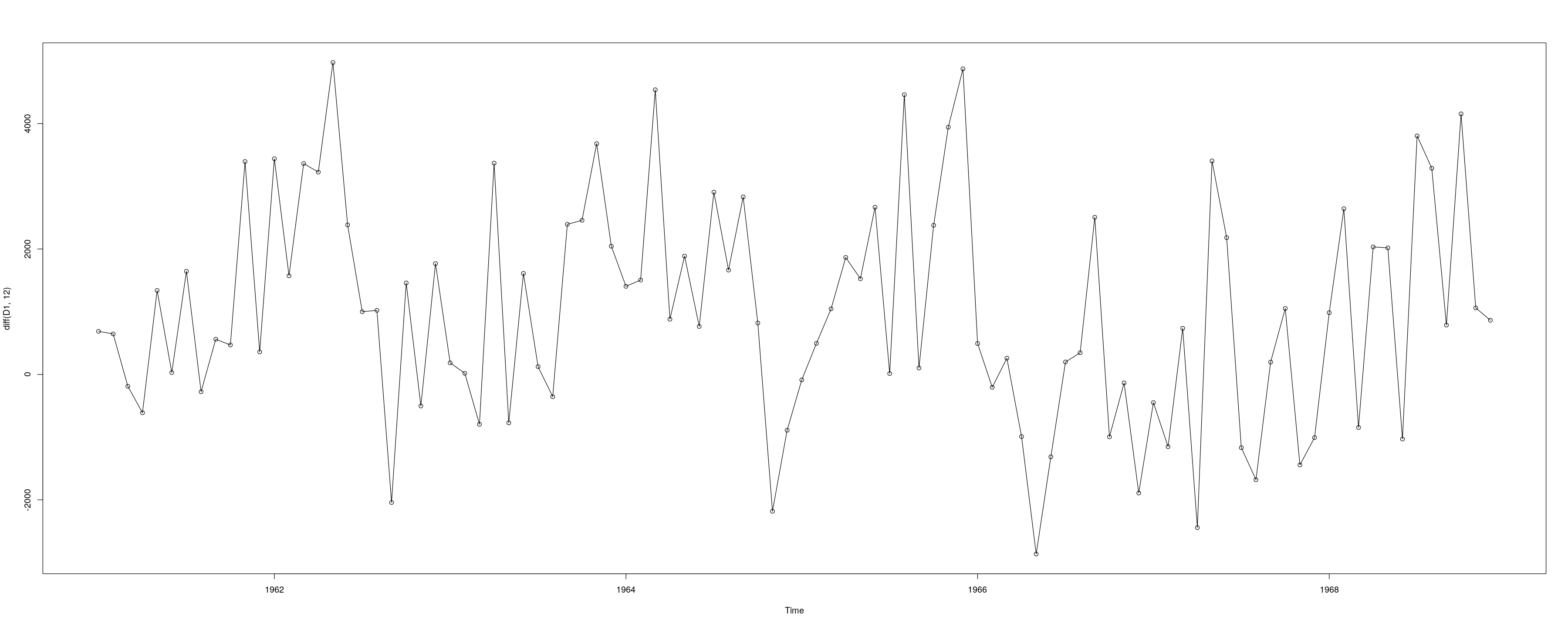
If you look at the graph of the raw data, it does not appear stationary. Performing statistical tests confirmed these suspicions. While the ADF and PP tests both had p values of 0.01 indicating that it is stationary, the KPSS test had a p value of 0.01, indicating that it is not stationary. No there is concern.

Taking a difference of d=1 makes the data appear more stationary.



Performing statistical tests confirms this, with the KPSS, ADF, and PP p values all indicating that it is stationary.

In addition, if a seasonal difference of D=1 is taken, then that also makes the data appear stationary.



Performing statistical test confirms this, with the KPSS, ADF, and PP p values all indicating that it is stationary. So, the data became stationary with d=1 and became stationary with D=1.

We did test that taking a seasonal difference was appropriate for this data set. We fitted a ARIMA(0,0,0)(0,1,1)[12] with drift model and confirmed that the sma1 term was not close to 1, and we fitted a ARIMA(0,0,12)(0,1,0)[12] with drift model and confirmed that none of the ma terms where close to 1.

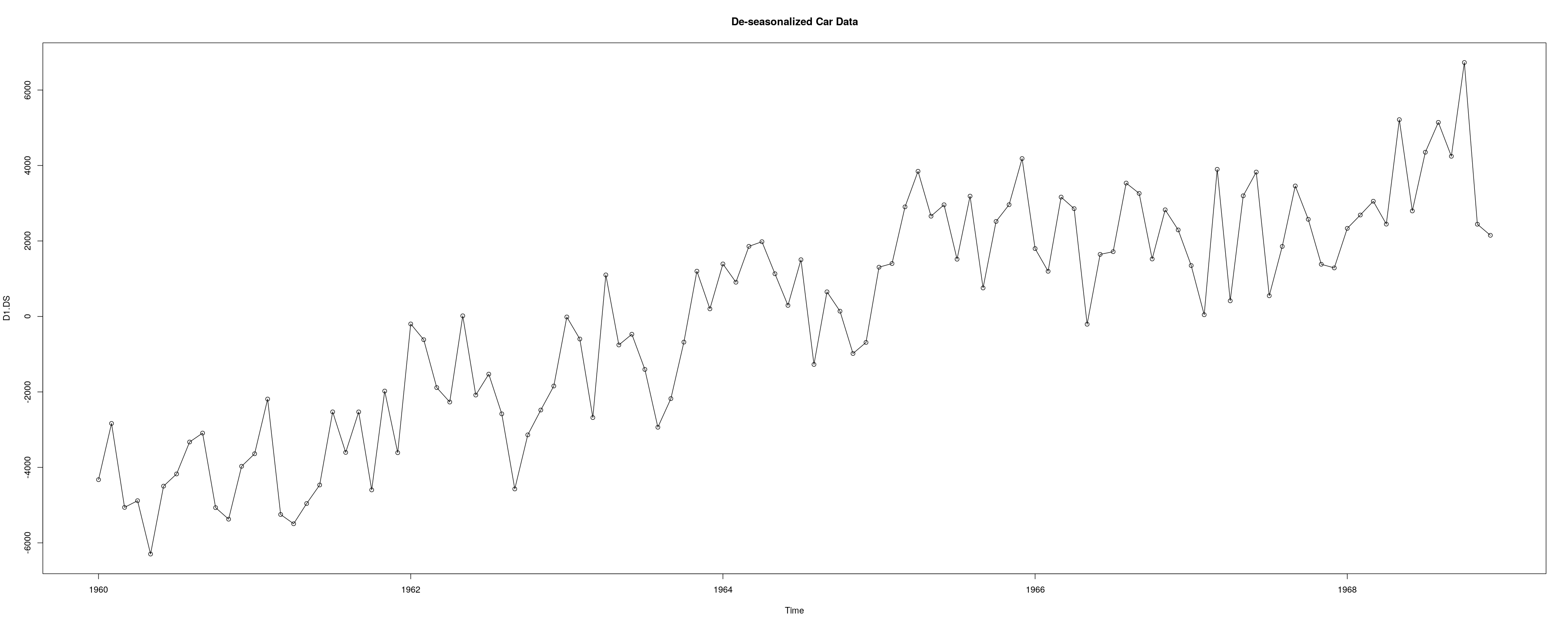
After testing several models, we found that the models with d=1 where performing better than the models with D=1.

# Transformations

Our dataset contains the sale price of cars over time, so it is financial data. Often it is appropriate to take the log of financial data in order fix the issue of variance increasing as values increase. However, the graph of our data did not indicate this was a significant problem. We also do not run into any specific problems that needed to be fixed with the log transformation. We did test some models with the log transformation; however, these models did not seem to be performing significantly better. Without a clear benefit of the log transformation, we went without it in order to have a simpler model.

# Linear Trend or Random Trend

We calculated the seasonal average, and then subtracted that from the data. A chart of this de-seasonlized data does appear to have a linear trend.



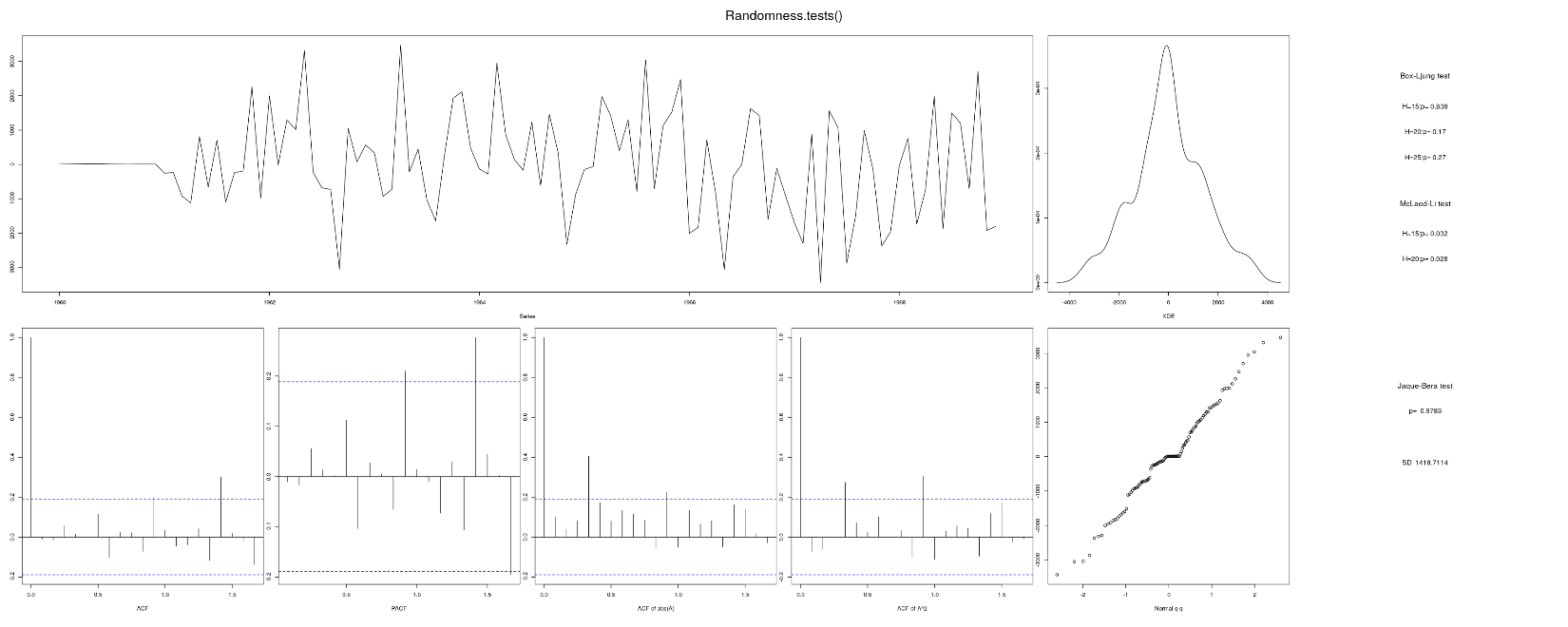
We verified this by performing a regression on the data and confirming the p-value for the regression was significant. We also performed a stationary test on the residuals, and confirmed that the residuals are stationary. The regression line has a slope of 1,000.1.

This verified that it was valid to try a model with a linear trend. However, in testing our models, we found that a model with drift was actually the best performing model.

# Order Selection Of ARMA

The ACF and PACF graphs of the data show correlation at the lower lags (1, 2, and possibly 3), and you also see the seasonal component with correlation at lag 12. However, you will also notice some significant correlation at lag 11 and lag 13. The correlation at lag 11 keep showing up.

If you put the seasonal time series into auto.arima() it will suggest of model of ARIMA(2,0,0)(0,1,2)[12] with drift. The sma2 term in this model is not significant. If you remove it, and look at the ARIMA(2,0,0)(0,1,1)[12] model, and test the residuals, you get:

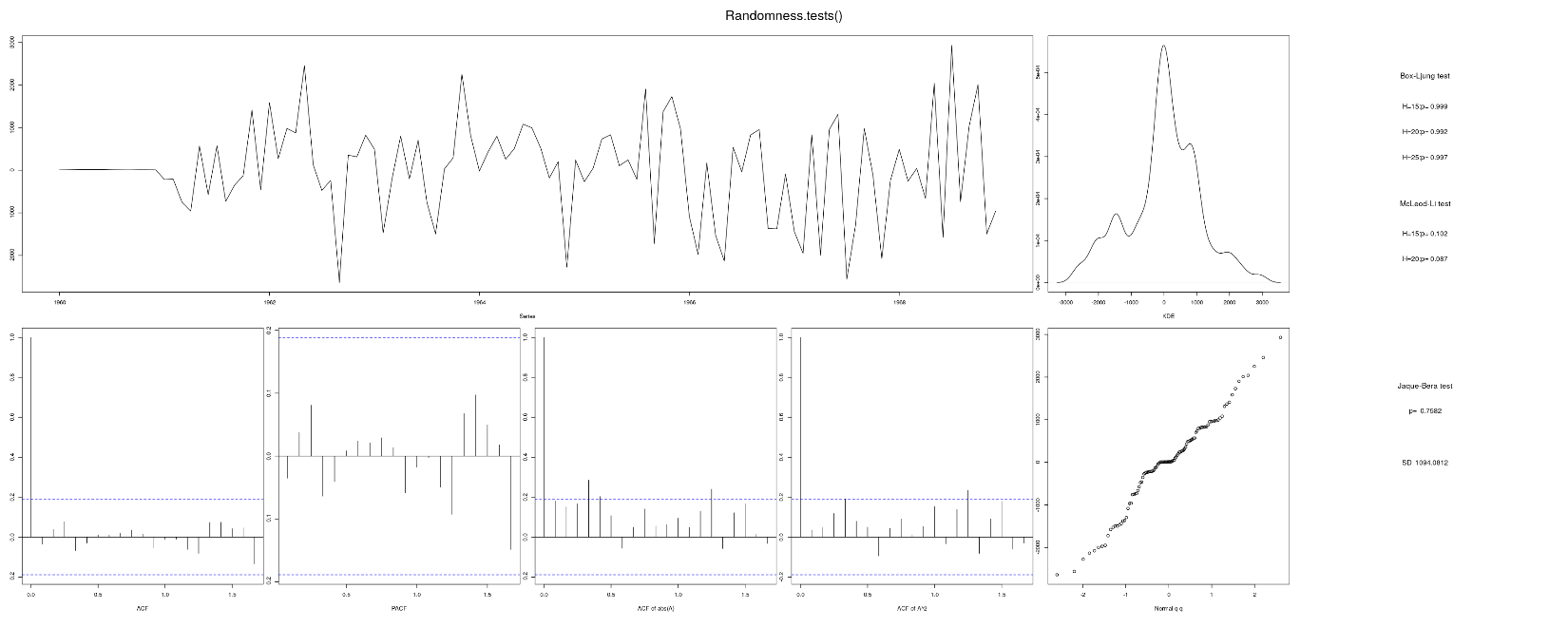


There correlation tests are indicating there is still correlation in the data. The ACF and PACF charts are indicating significant correlation at lag 11 and at lag 17. Because of this we tested a number of models with the p and q terms forced up higher in a attempt to capture this correlation at lag 11 and 17. What we found was that pushing the p term to 11 and the q term to 10 eliminated the correlation at lag 11 and lag 17.

Since auto.arima() was suggesting a model with D=1, we decided to manually test models with d=1. Again, we ran into the issues with correlation at lag 11 and lag 17, and once again forcing the p and q values to 11 and 10 eliminated this correlation in the residuals. We also found that the models with d=1 where performing better than the models with D=1.

# Final Model

The final model we selected was ARIMA(11,0,10)(0,1,1)[12] with drift. If we look at the residuals, we see the correlation tests look good, and the tests for normality look good. In addition, the ACF and PACF graphs are not showing any correlation.



We tested this model and found that the 95% confidence interval for the next prediction is 11,200.41 to 16,998.19. The rMSE was 1,688.845 which was close to the sigma value of 1,434.85.

# Additional Documents

Along with this report we submitted the following documents:

* 577TimeSeriesFinalProject-V2.pdf – this is a PDF of the Jupyter Notebook where we did our work. It is a combination of our notes and our R code.
* 577TimeSeriesFinalProject-V2.r – this is the R code from the Jupyter Notebook.
* 577TimeSeriesFinalProject-Log-V2.pdf – tis is a PDF of the Jupyter Notebook where we tested models with the log transformation.
* 577TimeSeriesFinalProject-Log-V2.r – this is the R code from the Jupyter Notebook where we tested models with the log transformation.