Andrew Wolfe
Individual Project
MAS 640
March 2, 2024

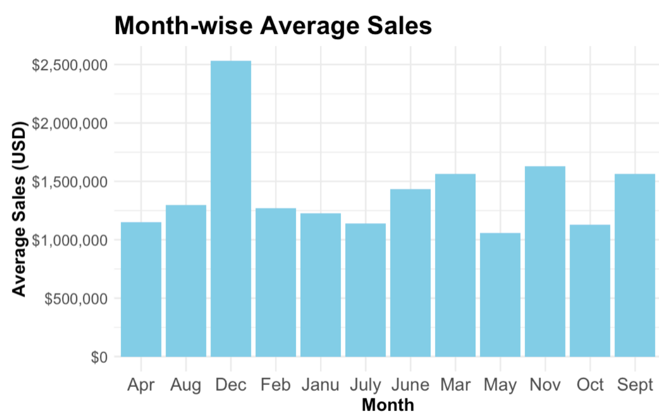
I. Data Introduction

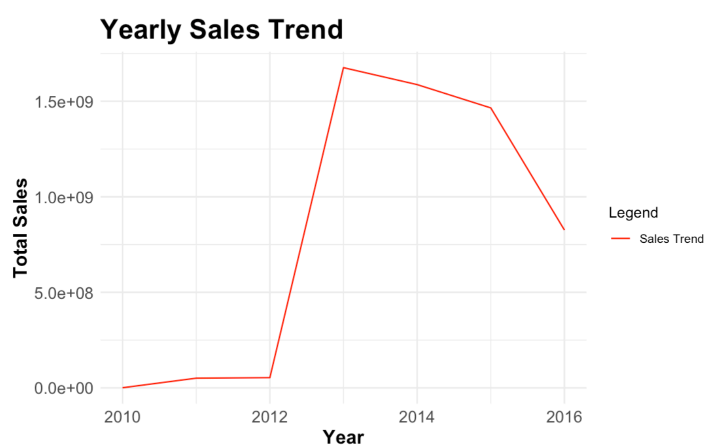
The data was pivoted to ensure that the 'Categories' were separated into three different columns, along with 'Month' and 'Sales'. All NA values and rows with 0 in the 'Sales' column were deleted. The Category group that I looked into when conducting the forecasting part of this project was:

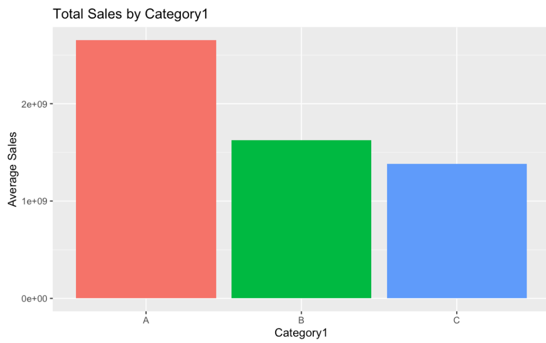
Category 1: A, Category 2: A, Category 3: B

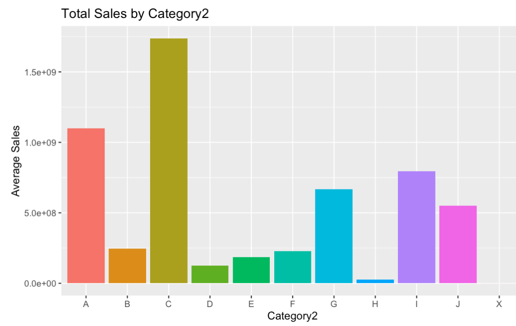
II. Data Exploration

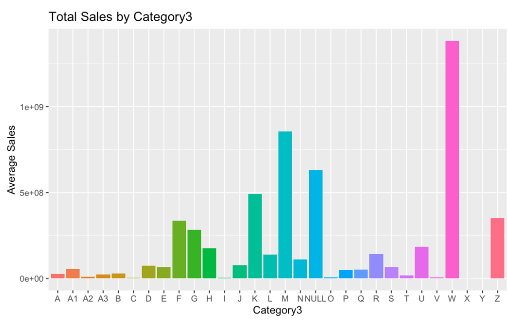
Upon initial data exploration, I created a couple of bar graphs showing the total sales by Category and Month. By looking at the figures below, we see that December had the highest average sale price. Moreover, when looking at the yearly sales growth rate, we see that the price increased exponentially from 2012-2013, but has been declining ever since. Also, for Category 1, Category 'A' had the highest average sale price. For Category 2, Category 'C' had the highest average sale price. For Category 3, Category 'W' had the highest average sale price.







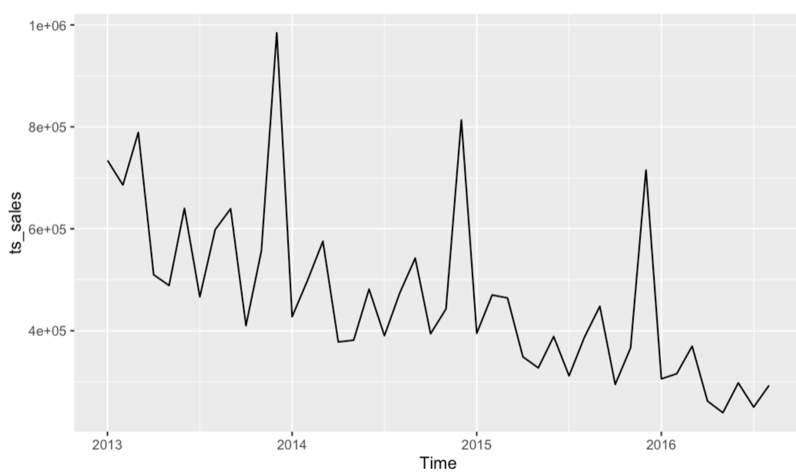




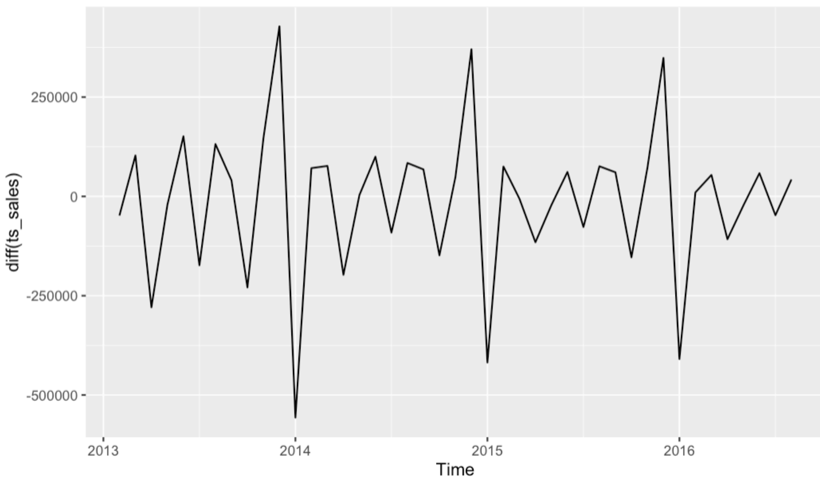
III. Forecasting Model

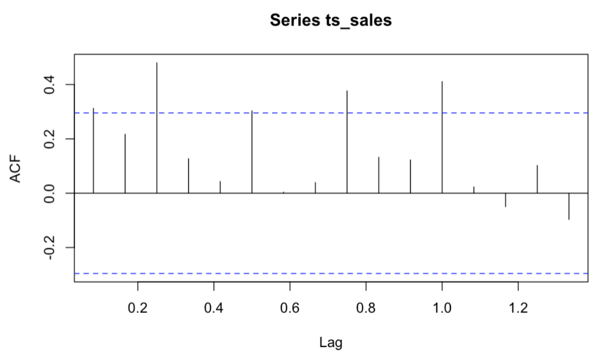
Using the Arima Model, I created a model to forecast sales for the Category Group: A, A, B during the next 12 months. Using basic functions, such as autoplot, acf, pacf, and auto.arima, I created a model using a nonseasonal MA(1) with one difference, forecasting with h = 12 (next 12 months).

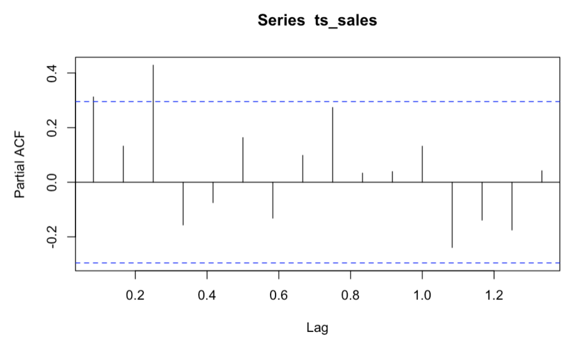
Upon this model discovery, the original time series plot showed a decreasing trend in terms of sales for this category group as depicted below.



With this in mind, it was necessary to take the difference of the time series and plot the pacf, acf, and auto.arima graphs. These graphs showed a model of nonseasonal MA(1) with one difference was most necessary for this forecasting model.







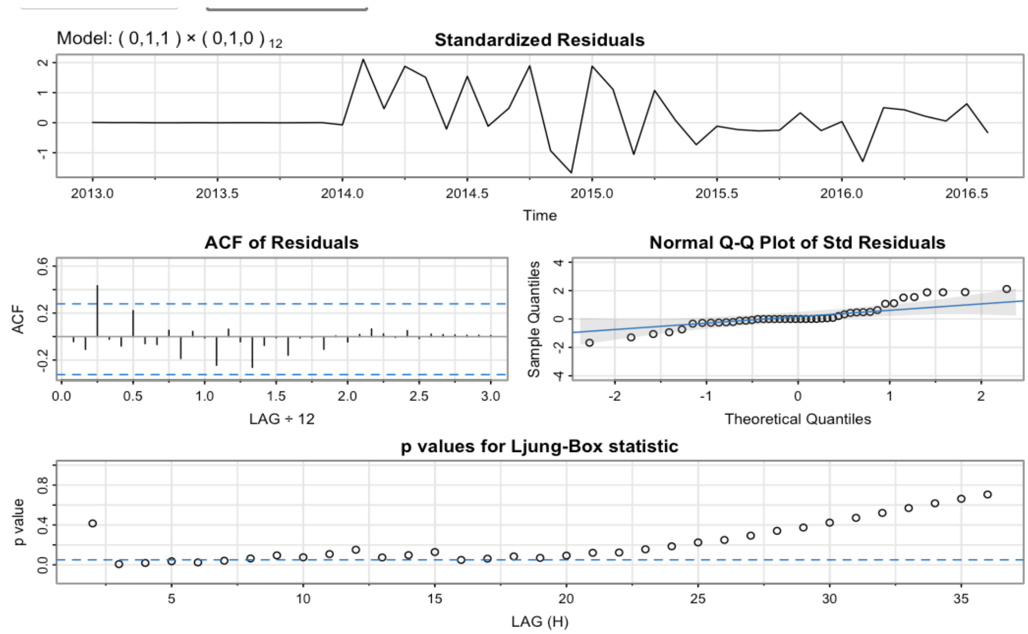

```
Series: ts_sales
ARIMA(0,1,1)(0,1,0)[12]

Coefficients:
          ma1
      -0.5474
s.e.   0.1459

sigma^2 = 2.542e+09:  log likelihood = -379.33
AIC=762.65   AICc=763.08   BIC=765.52
```

IV. Residuals Check

For residual diagnostics, we look for four components: random standardized residuals, the ACF of residuals remaining within the boundary, a normal Q-Q plot, and p-value points above the boundary line in the Ljung-Box test. The model fit summary showed that it was a good fit as the residuals appear randomly over time, meaning there was no seasonality left unexplained by the model. The Ljung-Box test p-values were all mostly above 0.05, showing no significant autocorrelation in the residuals. Moreover, the Q-Q plot shows the residuals are normally distributed.



V. Model Fitted

Once the model was properly chosen, I needed to fit it and create a summary of it. The figure below shows an AIC value of 762.65. The AIC is a measure showing how well a model fits, where lower scores are better. Moreover, the RMSE for the training set is 41,633.67, which may appear high at first, but when compared to the scale of the data, this number seems fine for our model.

```
Series: ts_sales
ARIMA(0,1,1)(0,1,0)[12]

Coefficients:
          ma1
      -0.5474
s.e.   0.1459

sigma^2 = 2.542e+09:  log likelihood = -379.33
AIC=762.65   AICc=763.08   BIC=765.52

Training set error measures:
                   ME     RMSE      MAE      MPE     MAPE      MASE        ACF1
Training set 9784.424 41633.67 26814.51 2.757701 6.635316 0.2564602 -0.04321908
```

VI. Forecasting Graph

After the assumptions have been checked and the model is properly fitted, we can finally forecast the model for Sales for the Category Group A, A, B in the next 12 months. When looking at the forecasted graph below, we see there is an expectation from the graph that Sales for this category group will continue to decrease in the next 12 months. In the blue, there does appear to be room for error, but generally and most likely, the Sales for this Category Group will decrease.