Avocado Time Series Analysis

Koby Arndt

1/11/2021

Introduction

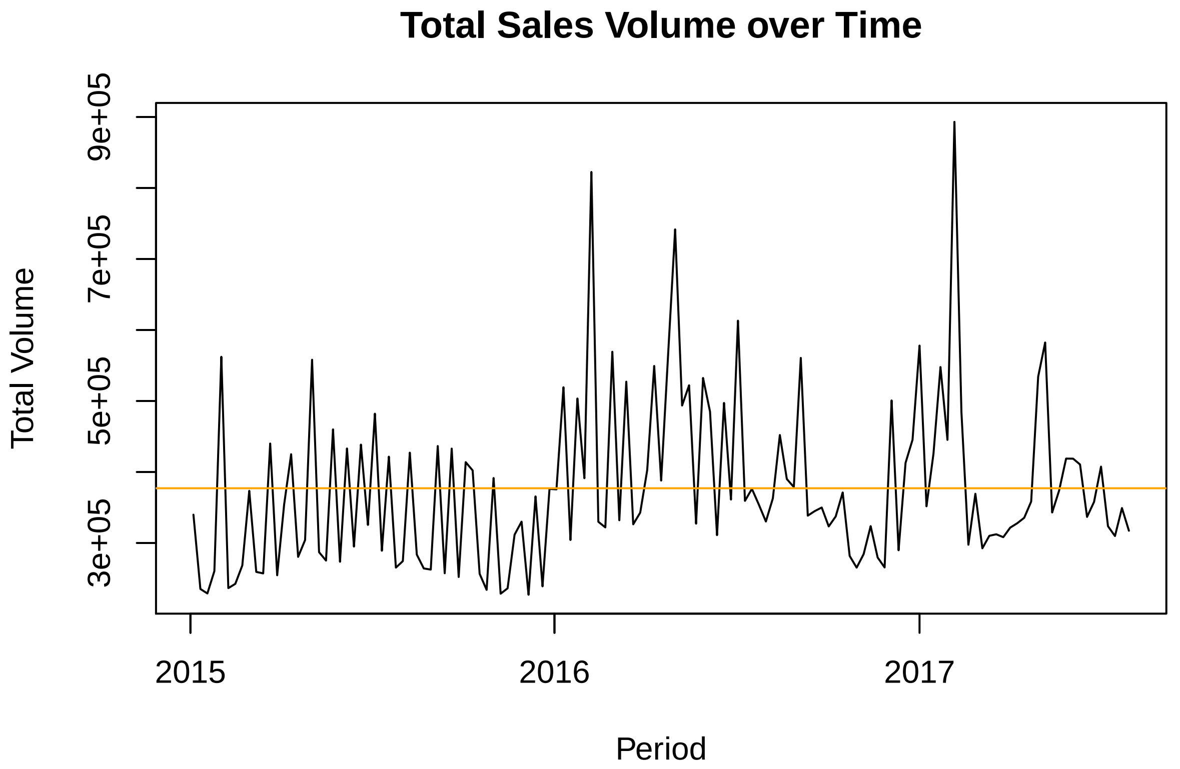
The raw data was obtained from the Kaggle website, where it had been uploaded by user Justin Kiggins. The data was sourced from the Hass Avocado Board and contains sales data for avocados in the United States from 2015 to 2018. A full data dictionary can be found in Appendix C. The analyst, Koby Arndt, has selected this data set in order to apply their knowledge of data science tools and techniques. This project is intended as both an exercise and an opportunity for self-education.

Data Wrangling

The first goal of this project was to implement the raw data into a data warehouse in SQL. Data warehouses are a preferred storage system data is being kept specifically for analytics purposes. Data wrangling was performed in Alteryx. Three dimensional tables were created; calendar, capturing the time aspect of the data, region, capturing the geographical aspect of the data, and type, capturing the conventional-organic dichotomy. The Alteryx Workflows and final SQL relational schema are pictured in Appendix B.

Initial Data Exploration

As an example of the sort of analysis that can be conducted using the Avocado data warehouse, the analyst queried time series data of total non-organic avocado sales volume in Tampa, Florida. The training data was plotted over a trendline equal to the mean total volume. The trendline is pictured in orange.



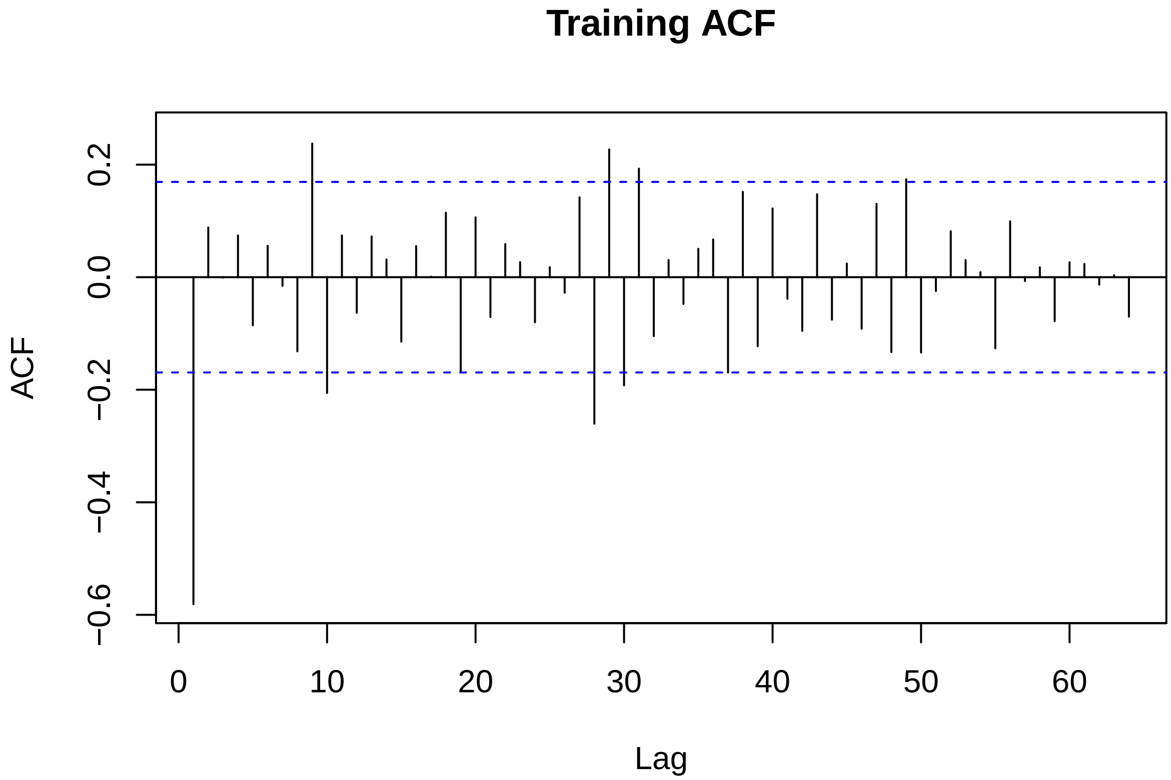
The total sales volume trendline crosses the mean frequently, but not consistently. A Unit Root KPSS test from the urca library was used to shed further light on whether the data was stationary. The test statistic returned as 0.4363, greater than the 10pct critical value of 0.347 but lower than the 5pct critical value of 0.463. The null hypothesis of the data being stationary could not be rejected at the 5% alpha level. This result is, again, inconclusive. The analyst ultimately chose to apply a logarithmic transformation to the data in order to manage its high variability and a one-period-lag difference transformation to the data in order to secure its stationarity. The difference transformation will remove patterns of autocorrelation in the data, so that constant patterns like seasonal variation can be clearly observed.

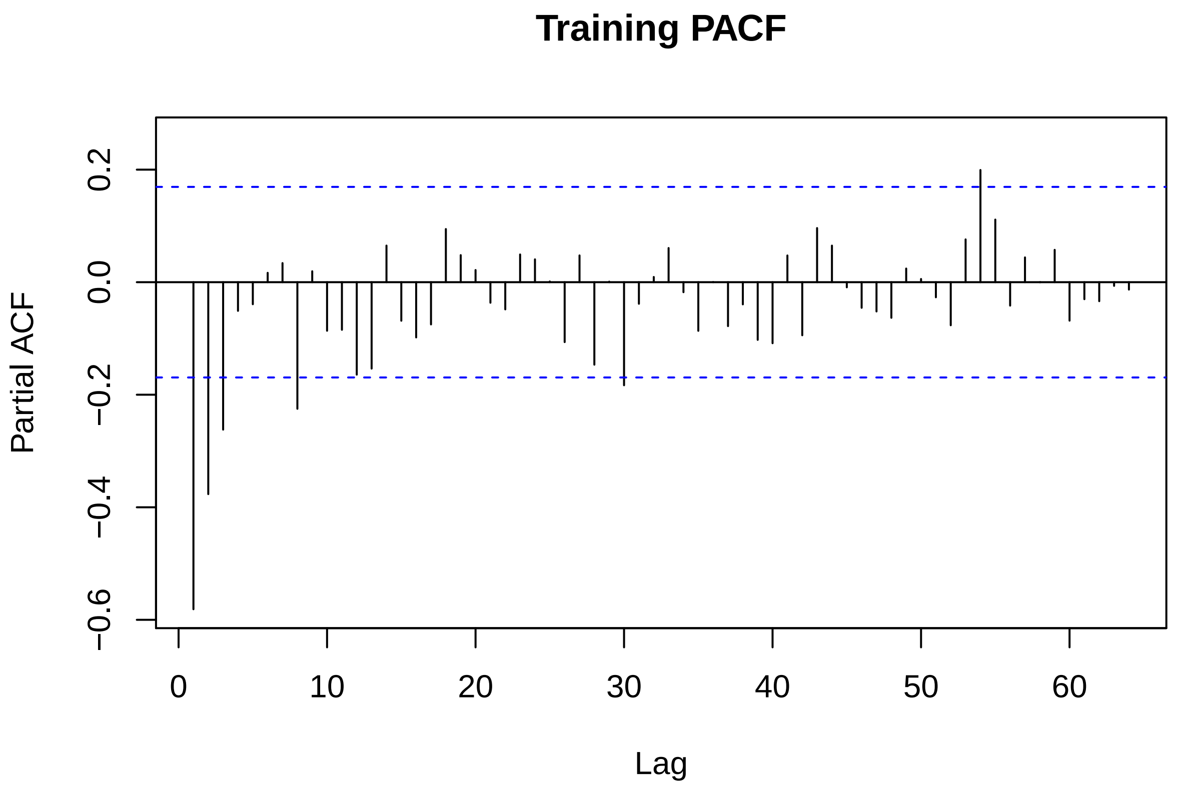
ACF and PACF

The auto- and partial auto- and cross-covariance and -correlation functions were plotted for the transformed training data in order to get a sense of trends or seasonality present. In ACF and PACF plots, spikes that cross the blue dotted lines represent significant variation.

In the ACF plot, the analyst noted a large negative spike at the first period, and large spikes roughly every 18 periods (weeks).

In the PACF plot, the analyst noted a trend of slow decay, three large negative spikes in the first three periods, and large spikes at periods 8, 27, and 50.


Training ACF
ACF
0.2
0.0
−0.2
−0.4
−0.6
0
10
20
30
40
50
60
Lag

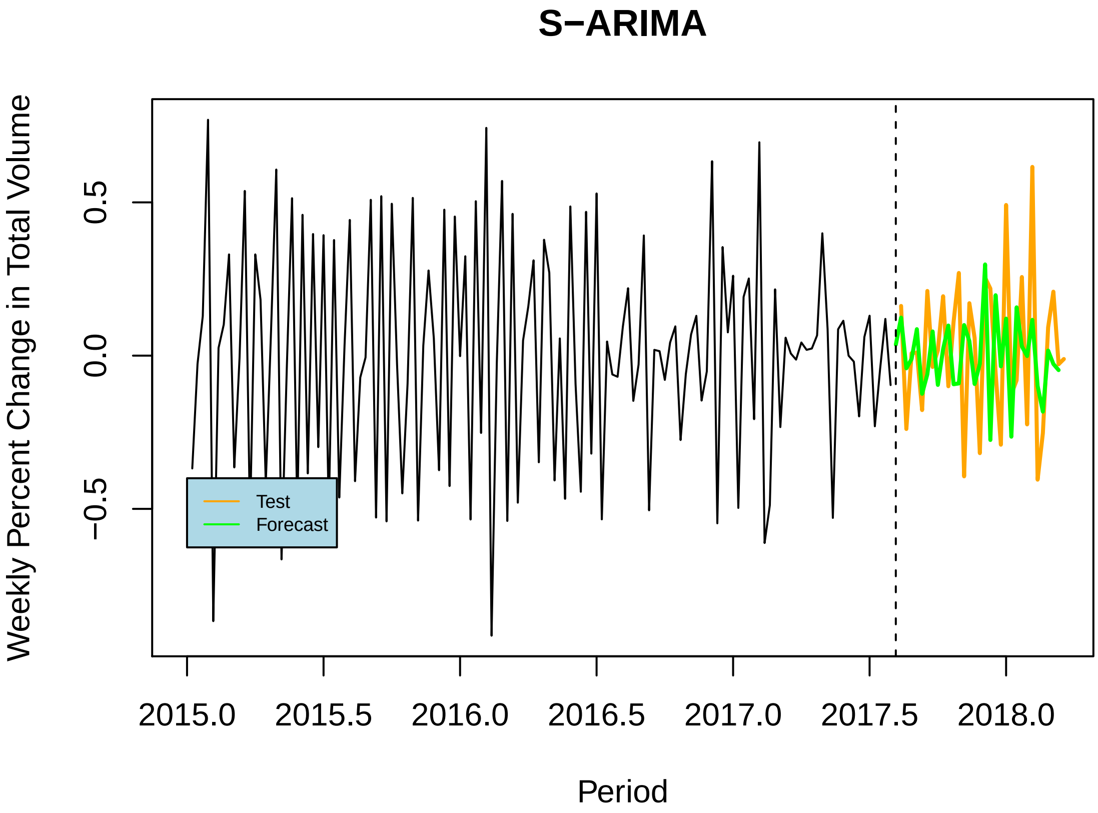


Model Building

Initially an AR(1)MA(1) ARIMA model was fitted to the training data using the auto.arima function in the forecast package. This autogenerated model achieved an AICc score of 30.74. A second manually determined model with a seasonal MA(1) component was found to achieve a superior AICc of 24.7.

Residual ACF and PACFs were plotted for both models; in both plots for both models, one significant negative spike persisted at period 10. The Box-Pierce test function from the stats packaged was applied to both sets of residuals, each resulting in a test statistic of 0.9429. The null hypothesis that the residual data is independently distributed cannot be rejected. This indicates that the residual variation is random, i.e. noise. ACF and PACFs can be found in Appendix A.

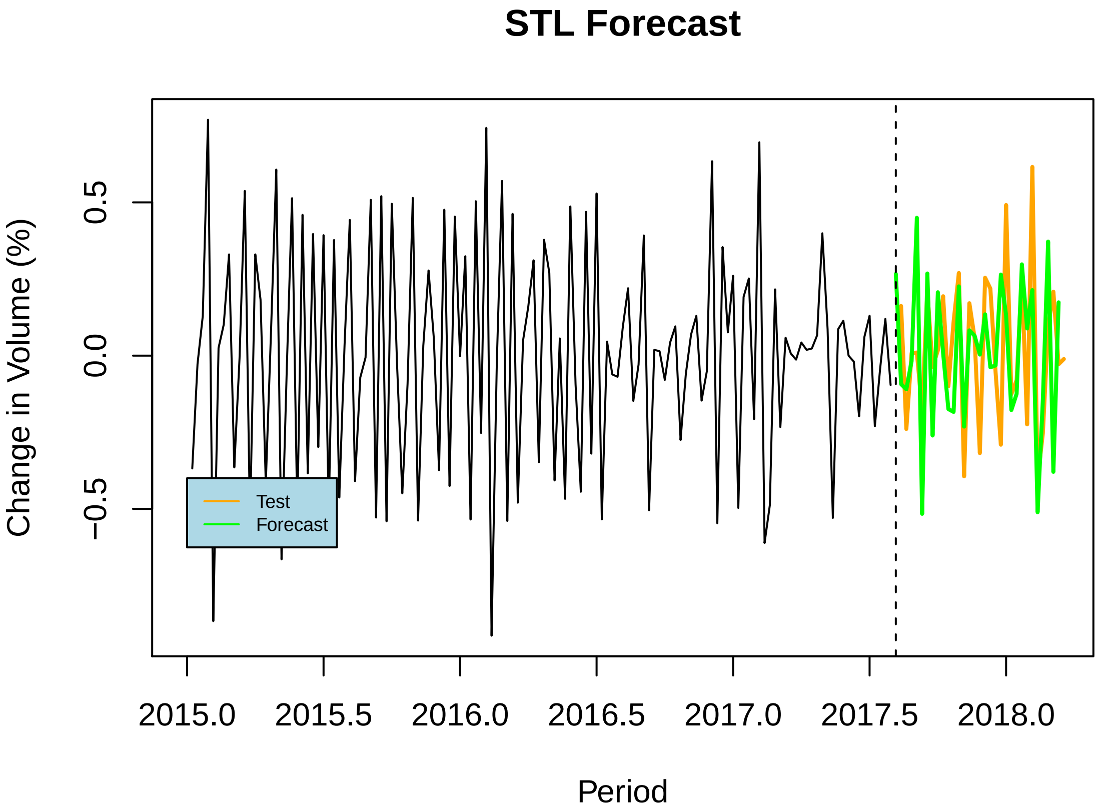
The S-MA(1) AR(1)MA(1) was preferred based on AICc score, and was therefore used to forecast the testing data.



A Seasonal and Trend decomposition using Loess Forecasting model was fitted to the training data using the stlf function in the forecast package.

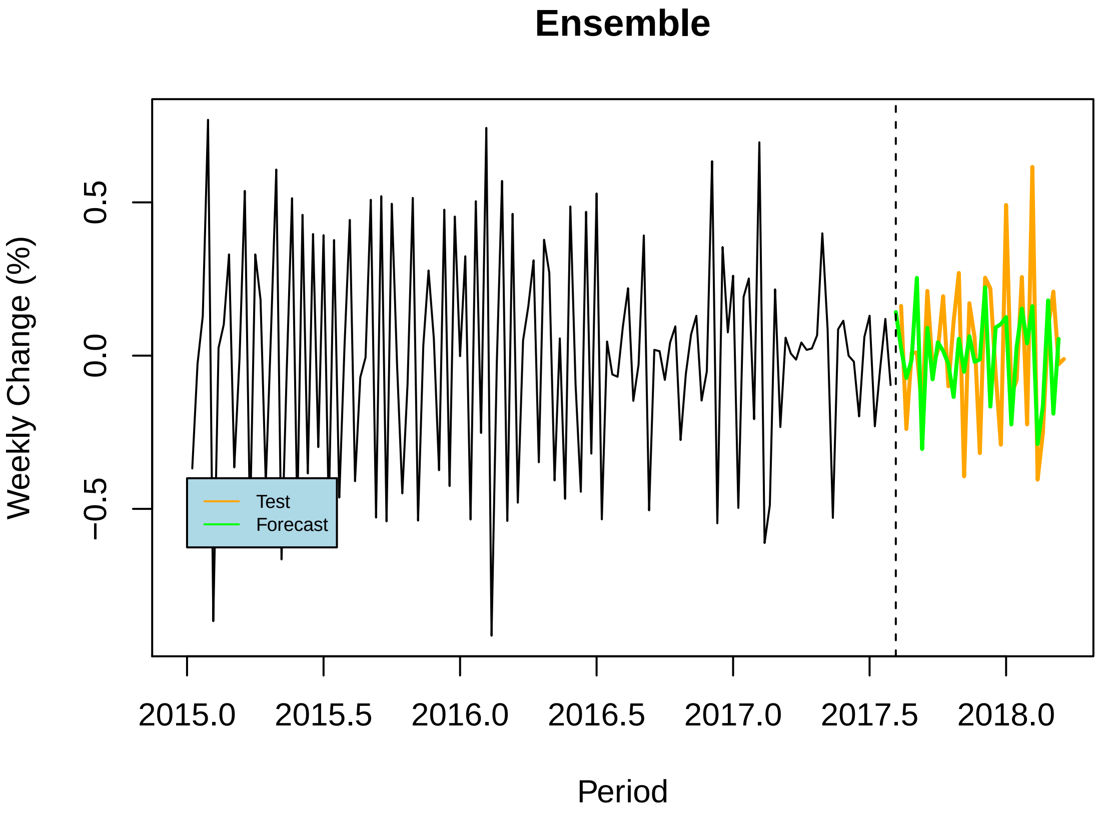
As with previous models, ACF and PACF of the residuals were plotted and Box-Pierce test was applied in order to determine whether residuals were independently distributed. The plots are pictured in Appendix A. In this case, several significant spikes persisted in residual ACF and PACF plots. Accordingly, the Box-Pierce test statistic had a p-value less than 0.05, indicating that the residuals are not independently distributed. This means that the model does not account for all trends in the data.

As with the S-ARIMA model, the STL forecast was graphed against the actual testing data. The analyst noted that the STL forecast appeared to match the level of variation present in the testing data more adequately than the S-ARIMA model.



Finally, an ensemble model of the S-ARIMA and STL models was considered in order to balance the strengths and weaknesses of both. A weight alpha was selected with a precision of two decimal places through an iterative trial-and-error process. The alpha that produced the minimum RMSE was selected. The minimum RMSE of 0.218 was found at an alpha of 0.54.

Finally, the ensemble model forecast was plotted over the actual testing data. Overall, the ensemble model outperformed all other models in the RMSE, MAE, and MPE error metrics.



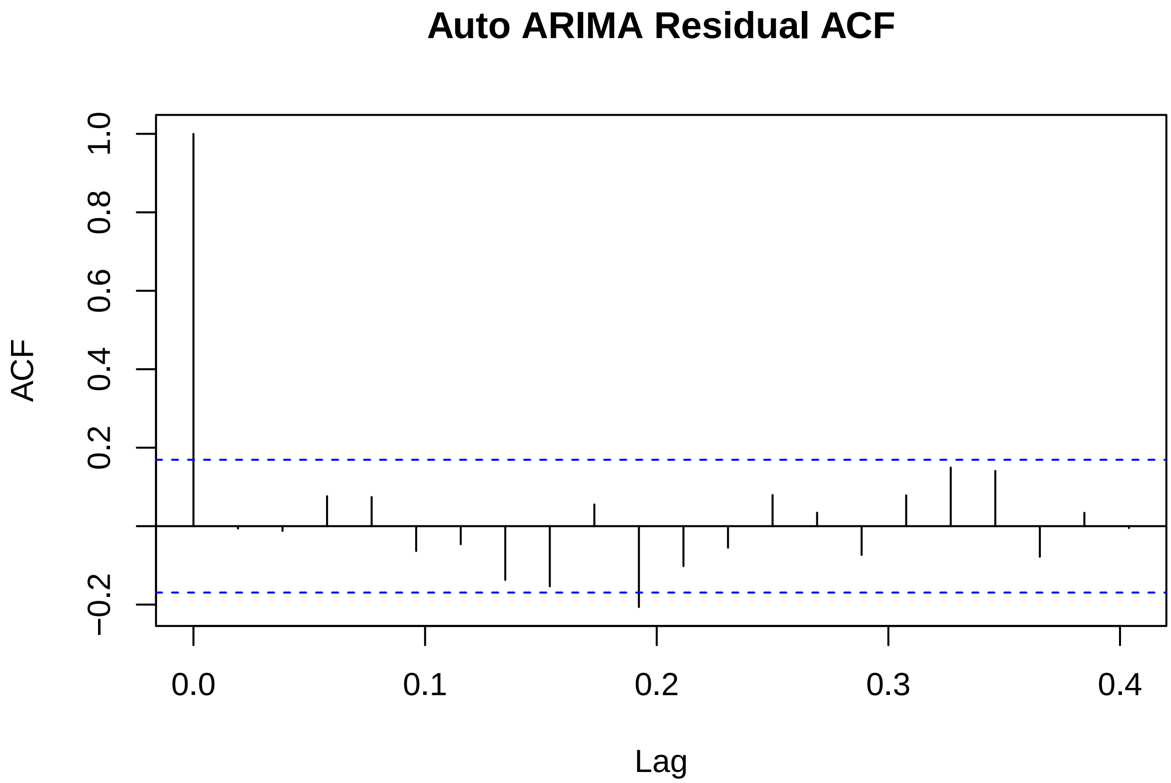
Conclusions

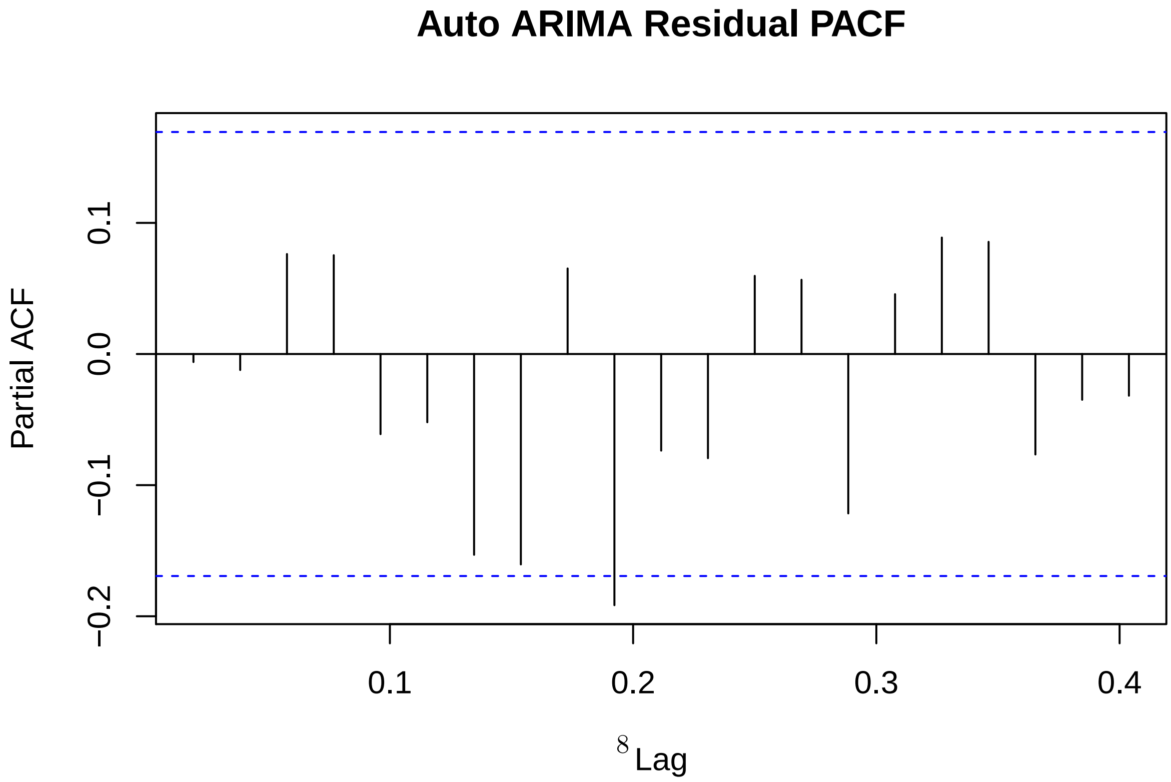
The data warehouse storage system proved very amenable to thorough and directed analytics. Further insights can be gained from the data, but in order to strengthen those insights, the analyst would recommend further data collection in order to capture a greater sense of seasonal patterns.

In the analysis performed, variation in total sales volume of conventional avocados in the Tampa, Florida region was most successfully forecasted by an ensemble model combining an S-MA(1) AR(1)MA(1) ARIMA model and an automatically calculated STL forecast.

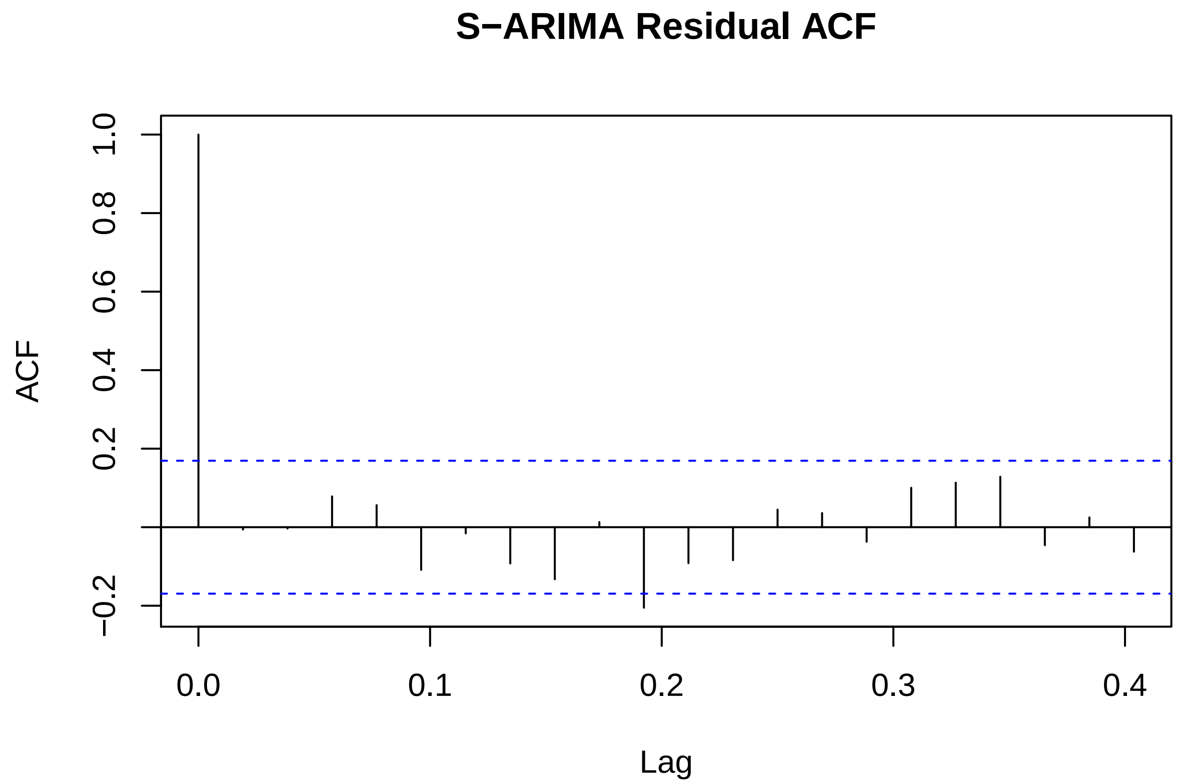
Appendix A: Descriptive Statistics and Visualizations

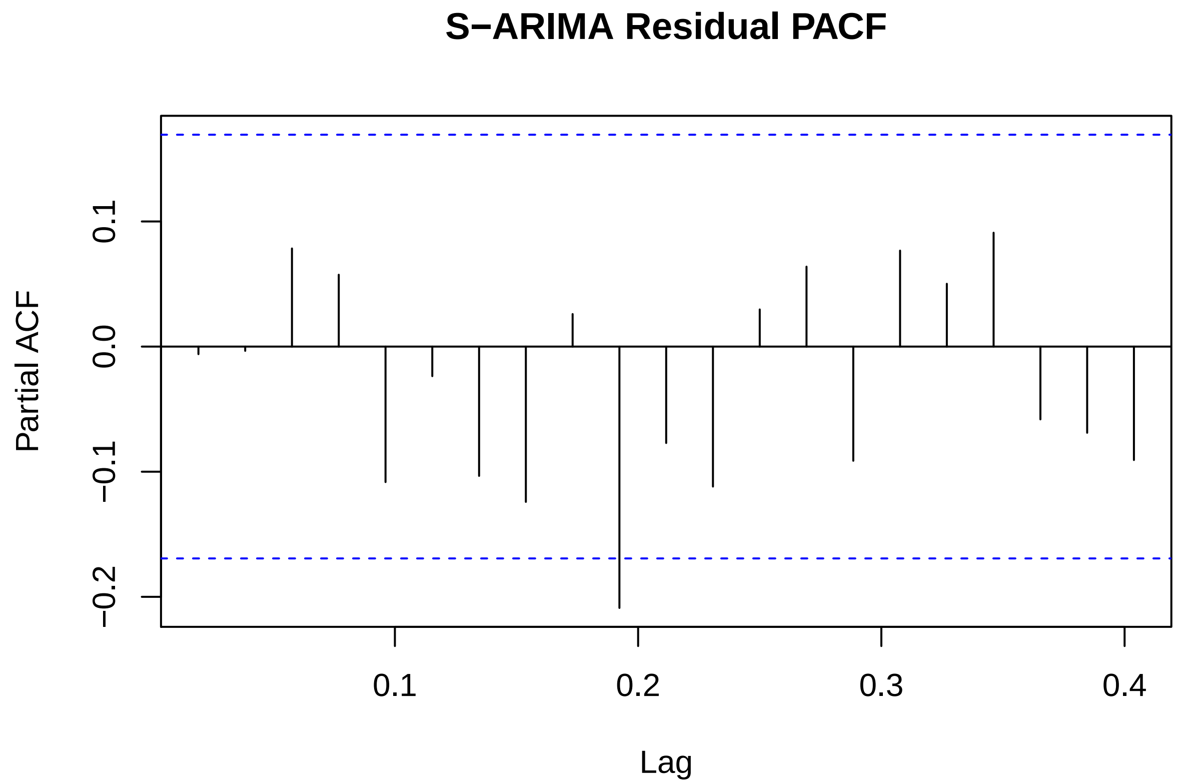
Item 1: Auto AR(1)MA(1) ARIMA Residual ACF and PACF


Auto ARIMA Residual ACF

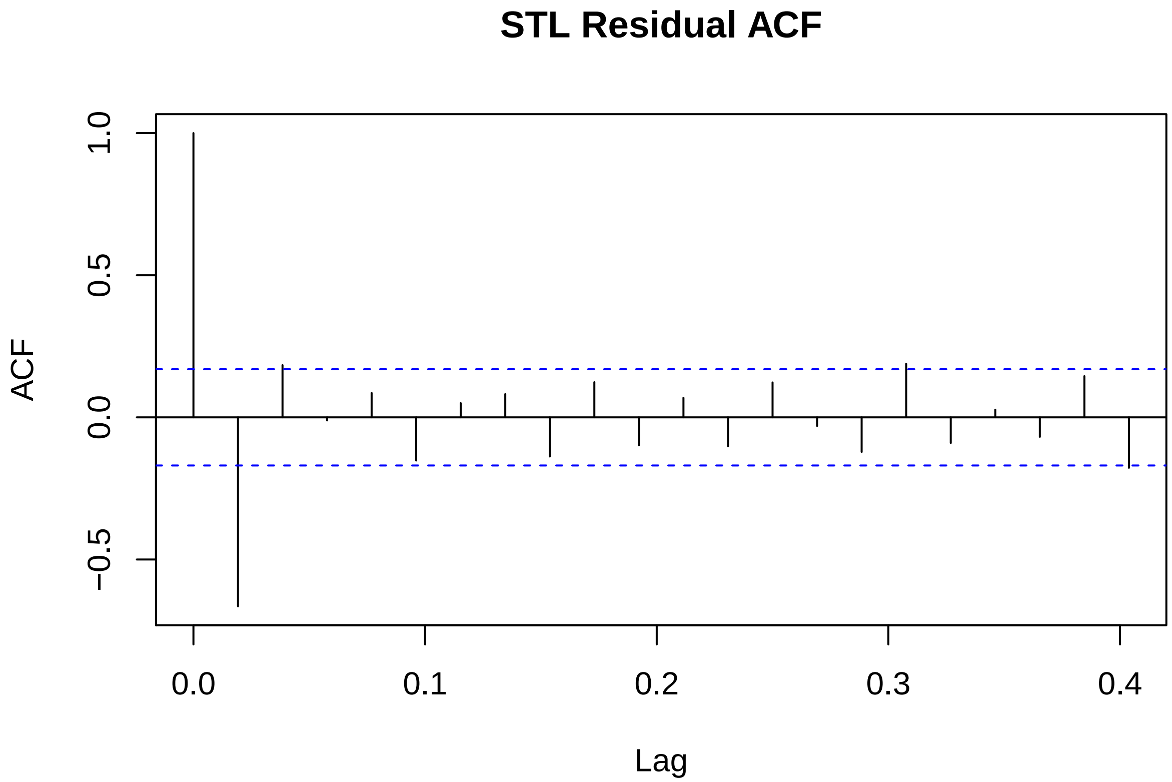

Auto ARIMA Residual PACF

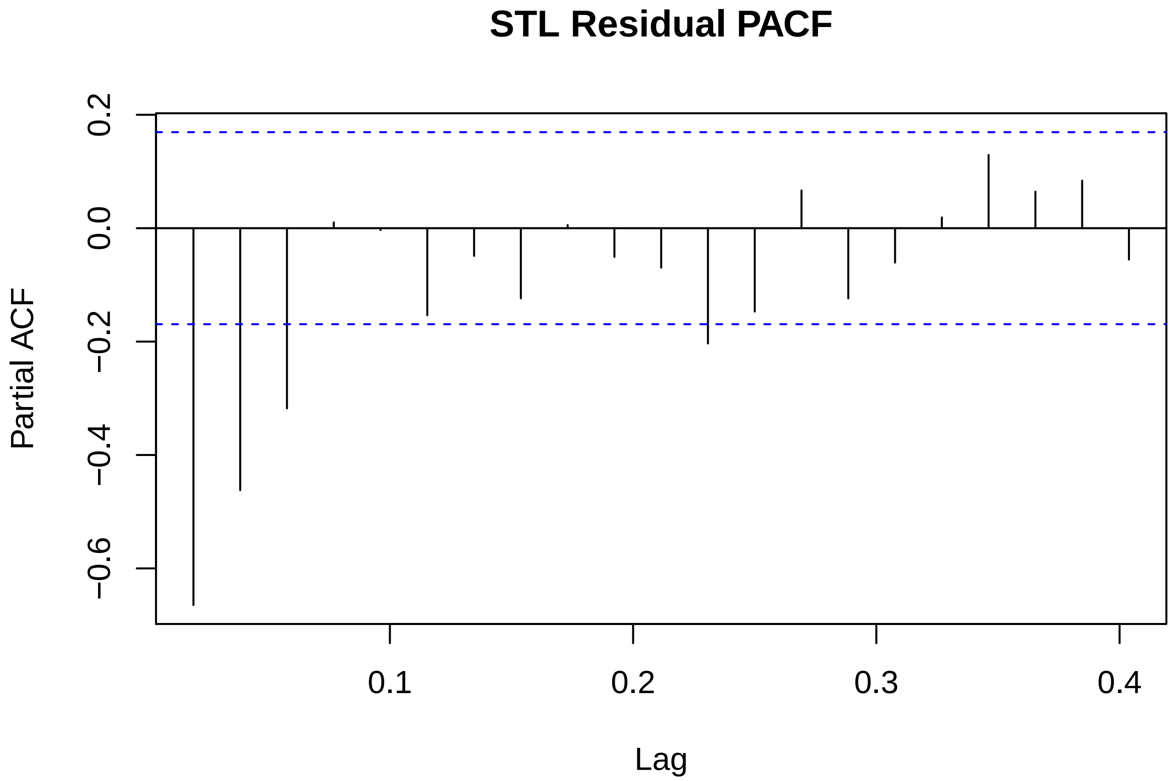
Item 2: S-AR(1) AR(1)MA(1) ARIMA Residual ACF and PACF





Item 3: STL Model Residual ACF and PACF





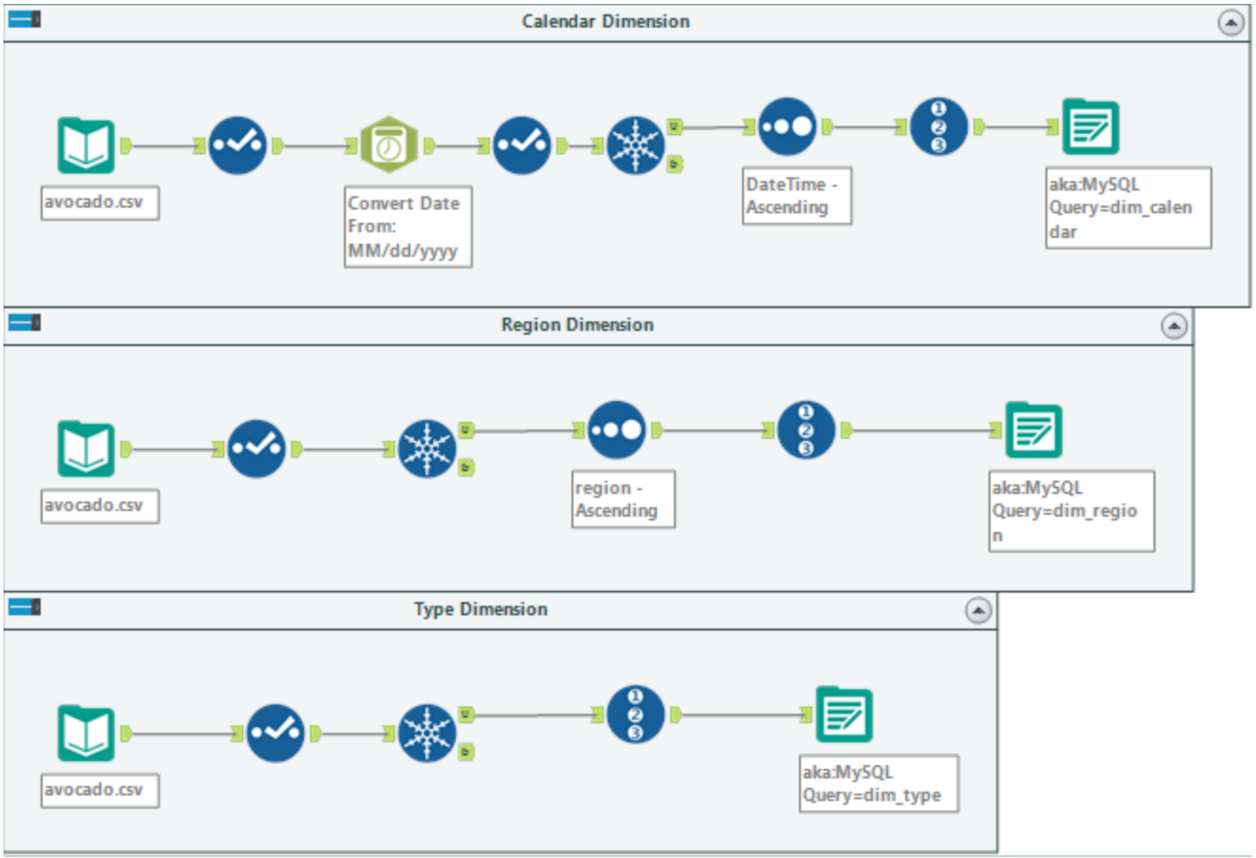
Item 4: Model Error Metric Table

```
knitr::kable((as.data.frame(errorMetrics)))
```

	Auto ARIMA	S-ARIMA	STL	Ensemble
ME	0.0204866	0.0224394	0.0201995	0.0214091
RMSE	0.2430614	0.2446395	0.2551956	0.2179406
MAE	0.1968394	0.2040852	0.2029650	0.1791265
MPE	100.7683757	109.2988567	-127.9942363	0.1440339
MAPE	100.7683757	170.2238799	310.1806819	187.5399513

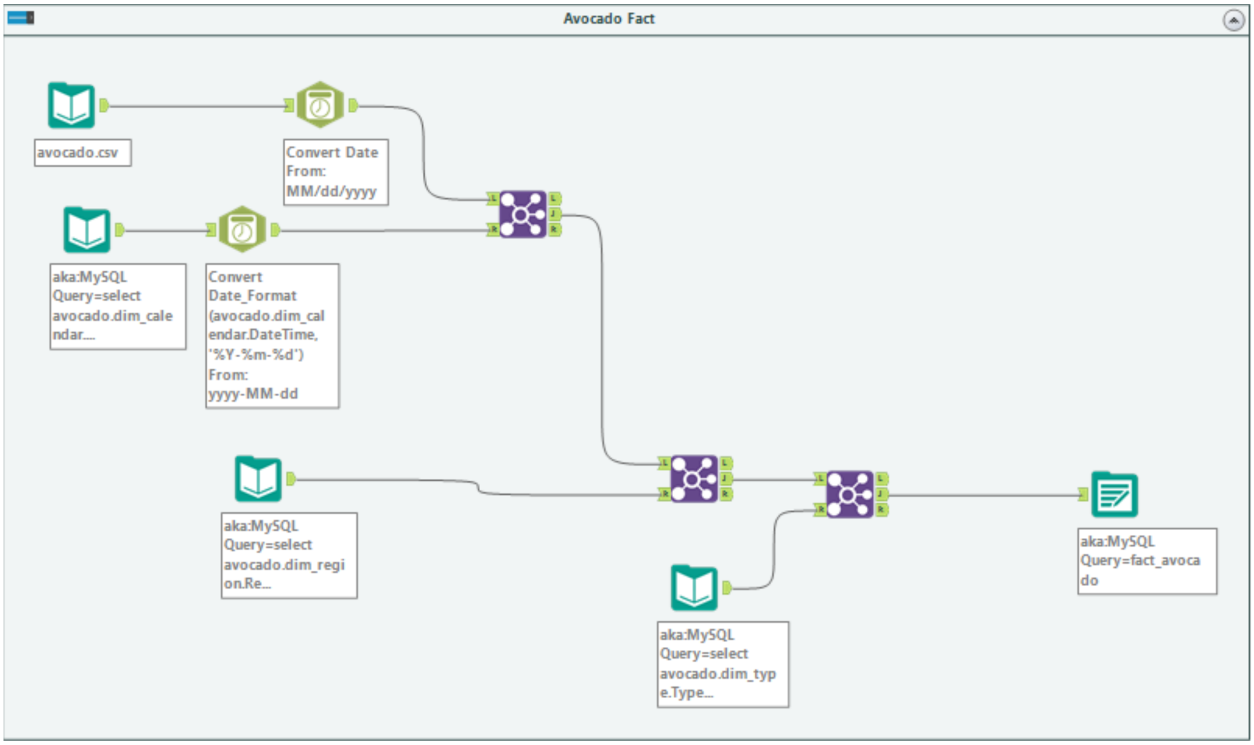
Appendix B: Data Preparation and Star Schema

Item 1: Dimension data wrangling in Alteryx.



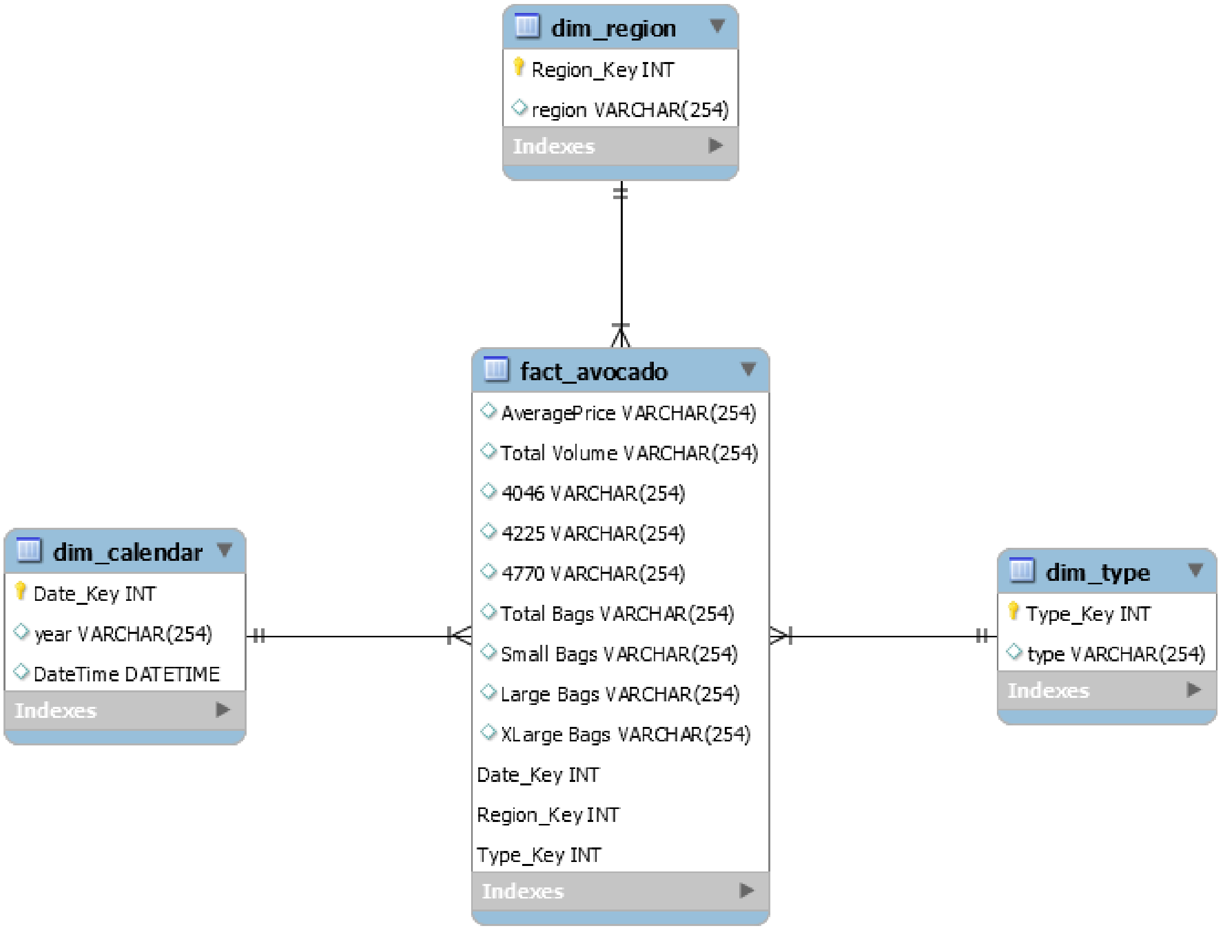
Unique dates, regions, and fruit types were given record IDs and reinserted into dimensional tables in MySQL.

Item 2: Fact data wrangling in Alteryx.


Avocado Fact
avocado.csv
Convert Date From: MM/dd/yyyy
aka:MySQL Query=select avocado.dim_calendar....
Convert Date_Format (avocado.dim_calendar.DateTime, '%Y-%m-%d') From: yyyy-MM-dd
aka:MySQL Query=select avocado.dim_region.Re...
aka:MySQL Query=select avocado.dim_type.Type...
aka:MySQL Query=fact_avocado

Avocado sales data read from the original csv file was transformed, related to the Calendar, Region, and Type dimensional tables, and reinserted into the fact table in MySQL.

Item 3: Final Star Schema



Relational star schema extracted from MySQL data warehouse.

Appendix C: Data Dictionary

Column	Type	Description
Week	Integer	Current date week
Date	Date	Current date date
AveragePrice	Numerical	Average price across outlets
Total Volume	Numerical	Average of individual avocados sold across outlets
4046	Numerical	Average of avocados with PLU 4046 sold across outlets
4225	Numerical	Average of avocados with PLU 4225 sold across outlets
4770	Numerical	Average of avocados with PLU 4770 sold across outlets
Total Bags	Numerical	Average of bags of avocados sold
Small Bags	Numerical	Average of small bags sold
Large Bags	Numerical	Average of large bags sold
Xlarge Bags	Numerical	Average of xlarge bags sold
type	Binary	0 if conventional, 1 if organic
year	Integer	Current date year
region	Character	Sales outlet location