Forecasting Future Sales of Avocados

Presented By: Yina Li, Laury Tremblay-Trudel, Kevin Heng & Mouloud Seffal

Table of Contents

Introduction

- Goal: Build a model to **forecast future sales of avocados**
- Predictions will be helpful for avocado industry participants (companies, farmers, etc.)
- Predicting future sales will make key players more prepared

The Dataset

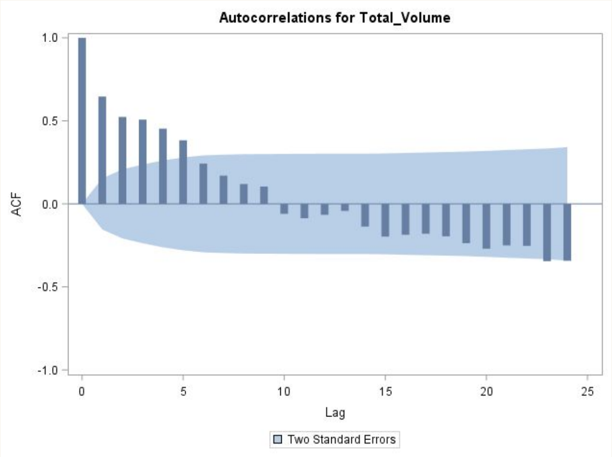
- Retrieved from **Kaggle**
- **18,250 observations** of avocado sales per state
- **3 years** of observations
- 13 numerical variables → **10 variables**
- PROC SQL → mean(variable)

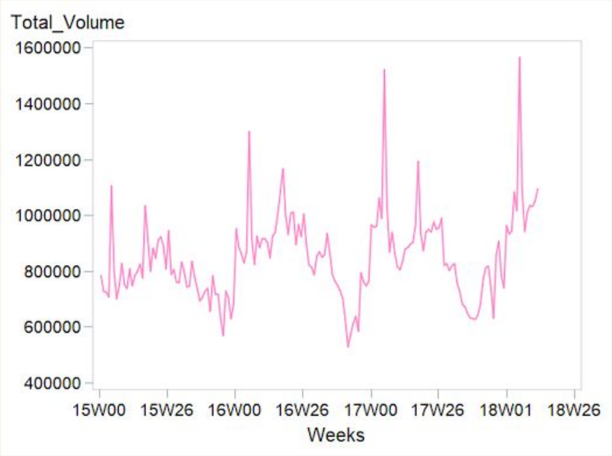
Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Informat
2	AveragePrice	Num	8	BEST4.	BEST4.
1	Date	Num	8	DDMMYY10.	DDMMYY10.
9	Large_Bags	Num	8	BEST10.	BEST10.
8	Small_Bags	Num	8	BEST11.	BEST11.
7	Total_Bags	Num	8	BEST11.	BEST11.
3	Total_Volume	Num	8	BEST11.	BEST11.
10	XLarge_Bags	Num	8	BEST9.	BEST9.
4	_4046	Num	8	BEST11.	BEST11.
5	_4225	Num	8	BEST11.	BEST11.
6	_4770	Num	8	BEST10.	BEST10.

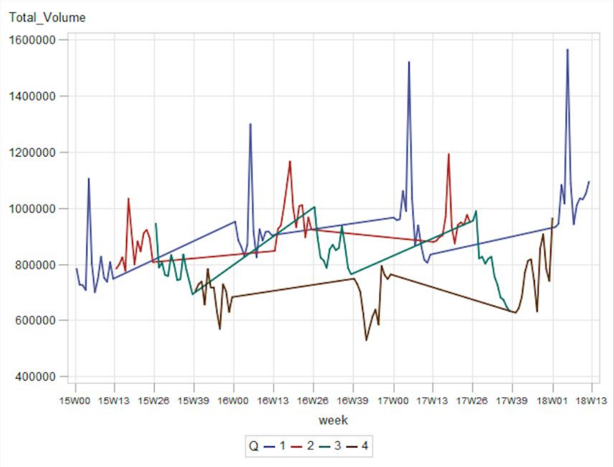
```
proc sql;
create table avocado_t as
select Date, mean(Total_Volume) as Total_Volume,
mean(_4046) as small_size, mean(_4225) as large_size,
mean(_4770) as xlarge_size,
mean(Total_Bags) as Total_Bags,
mean(Small_Bags) as Small_Bags,
mean(Large_Bags) as Large_Bags, mean(XLarge_Bags) as XLarge_Bags
from work.avacado
group by Date;
quit;
```

Identifying Patterns

- Trend component
- Seasonal component
- Autocorrelation → lags decrease slowly towards 0

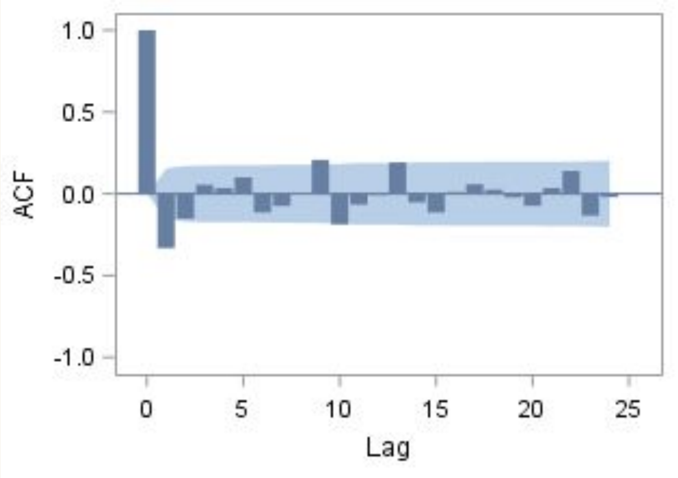






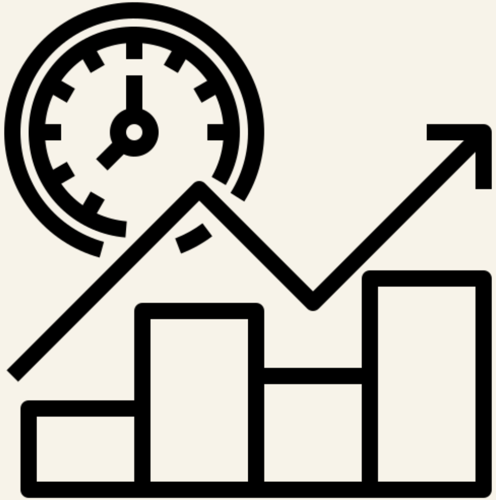
Identifying Stationarity

- Differenced dataset until we reach a **positive Lag 1** followed by a **sudden drop**
- This indicates that the dataset is now **stationary**



Forecast Horizon

- Forecast 1 year into the future
- As **dataset is in weeks**, we will forecast for approximately **52 weeks**



Accuracy Measures

- **MAD**: Mean absolute deviation → average distance between each data value and the mean → a way of measuring variation in a dataset

- **MAPE**: Mean absolute percentage error → a measure of prediction accuracy → measures how accurate the forecasting method is

- **RMSE**: Root mean square error → the standard deviation of the residuals → measures difference between observed and predicted values

Naive Method

- Follows closely actual data
- Consistent **under- and over-estimation**
- Validation values higher than training → not suitable for forecasting
- Will be used as **benchmark**

Naive training error term

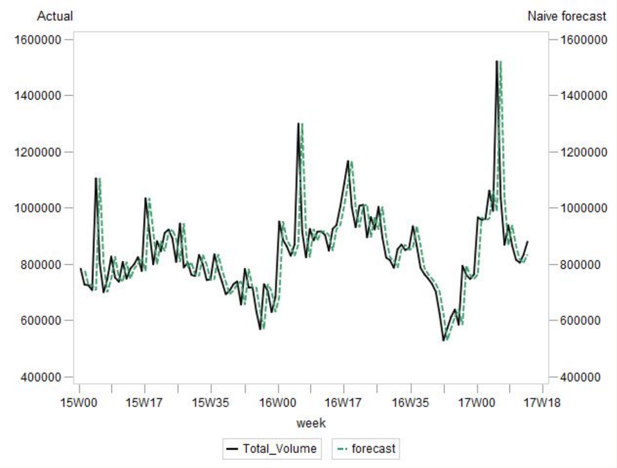
MAD	MSE	RMSE	MAPE	MPE
79232.50	15343841306	123870.26	(0.71877%)	8.94677%

Page Break

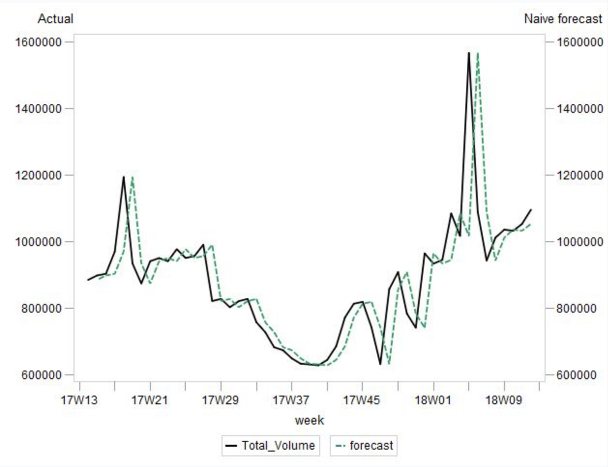
Naive validation error term

MAD	MSE	RMSE	MAPE	MPE
75900.02	17783983071	133356.60	(0.34686%)	7.88173%

Training



Validation



Simple Average

- Method **does not forecast accurately**
- Forecasted values are largely **underestimated**
- Is not representative of trend and seasonality
- MAD & RMSE better than Naive method

Training

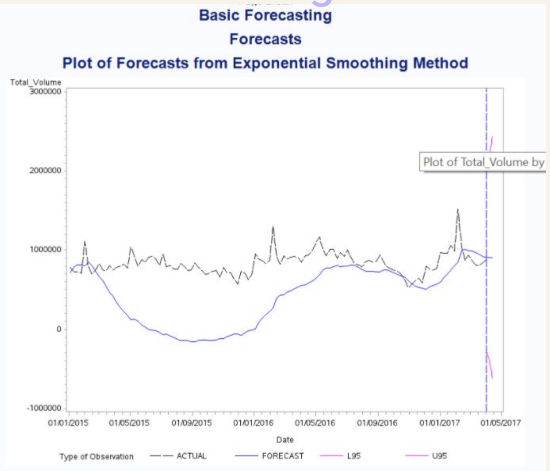
Page Break

Obs	n	MAD	MSE	RMSE	MAPE	MPE
1	129	24972.85	1232546510.4	35107.64	(6.67%)	13.73%

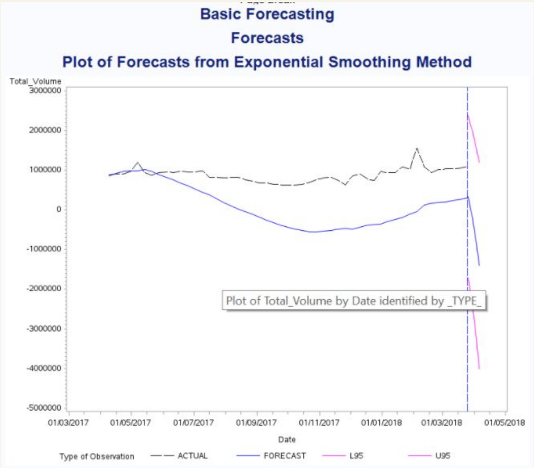
Validation

Obs	n	MAD	MSE	RMSE	MAPE	MPE
1	64	60786.93	5460657554	73896.26	21.21%	(18.75%)

Training



Validation



Moving Average (3-MA)

- Training line plot very similar to original dataset
- Validation set **over-smoothed** and does not
- Model appears to be improvement, but **errors are still very high**

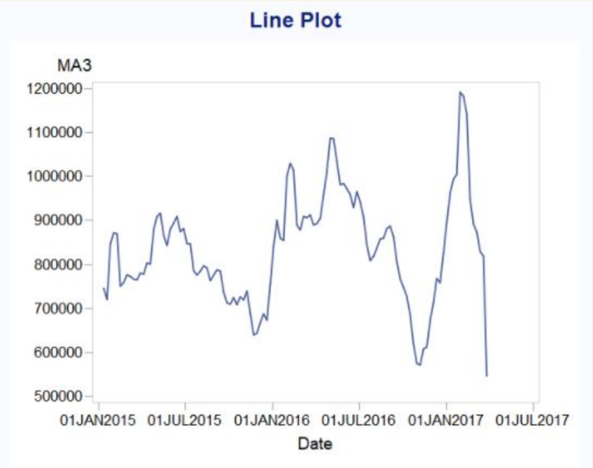
Training

Obs	n	MAD	MSE	RMSE	MAPE	MPE
1	169	24160.99	2029637457.8	45051.50	2.96%	(0.34%)

Validation

Obs	n	MAD	MSE	RMSE	MAPE	MPE
1	52	40235.27	4986904690.8	70618.02	4.68%	(0.73%)

Training

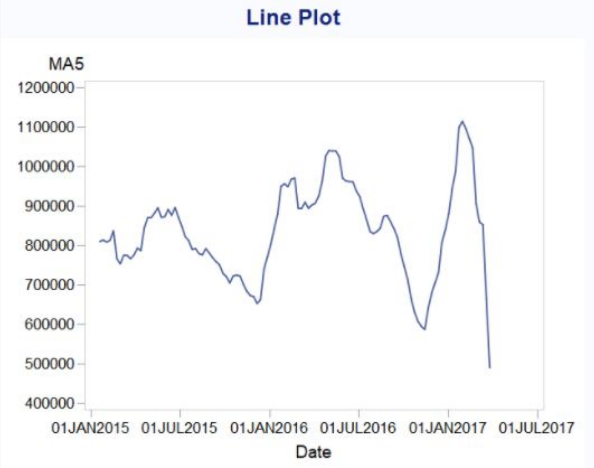


Validation

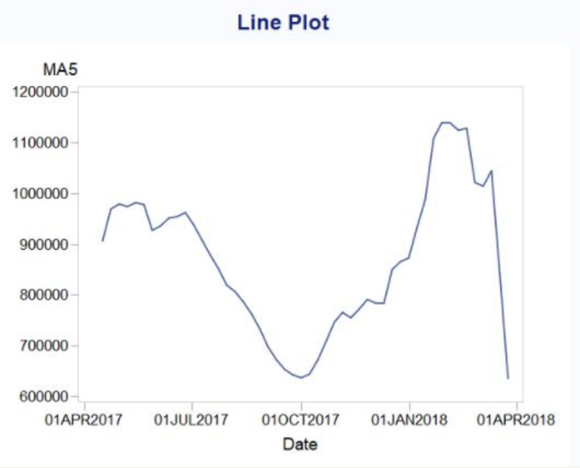
Moving Average (5-MA)

- Higher order made curve **smoother**
- MAD, RMSE, MAPE → **best forecasting model yet**

Training



Validation



Training

Obs	n	MAD	MSE	RMSE	MAPE	MPE
1	169	17609.00	1096608977.2	33115.09	2.23% (	0.39%)

Validation

Obs	n	MAD	MSE	RMSE	MAPE	MPE
1	52	31006.34	2764138280.3	52575.07	3.72% (	0.88%)

Holt's Linear Exponential Smoothing

- Advanced smoothing technique
- **Multiplicative** method → seasonal variations increase
- MAD, RMSE, MAPE decreased → **better model** for medium-term forecasts
- Peak seasonality **not as sharp** as original dataset

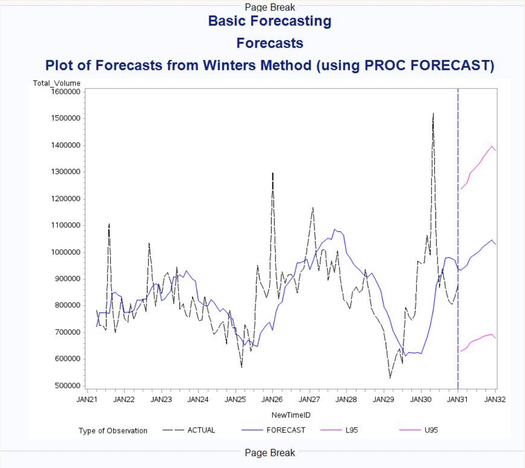
Training

Obs	n	MAD	MSE	RMSE	MAPE	MPE
1	130	19638.41	743988808.55	27276.16	2.37%	0.22%

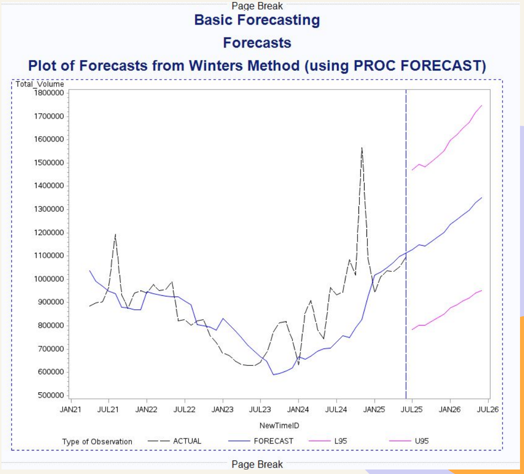
Validation

Obs	n	MAD	MSE	RMSE	MAPE	MPE
1	63	24019.72	1033644091.9	32150.34	2.78%	0.35%

Training

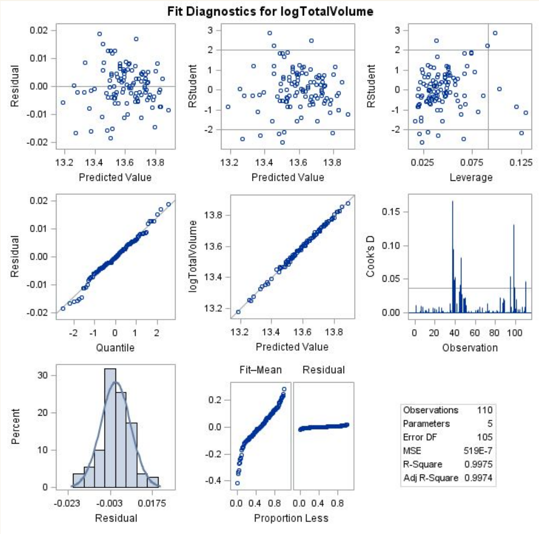


Validation



Multiple Regression

- Performed 2 transformations:
  1. Removed variables (multiple collinearity issues)
  2. Logged remaining variables → to remove extreme values



$\text{Total\_Volume} = 1.14799 + 0.23965_{\log(\text{total\_bags})} + 0.35277_{\log(\text{small})} + 0.39498_{\log(\text{large})} + 0.01356_{\log(\text{xlarge})}$

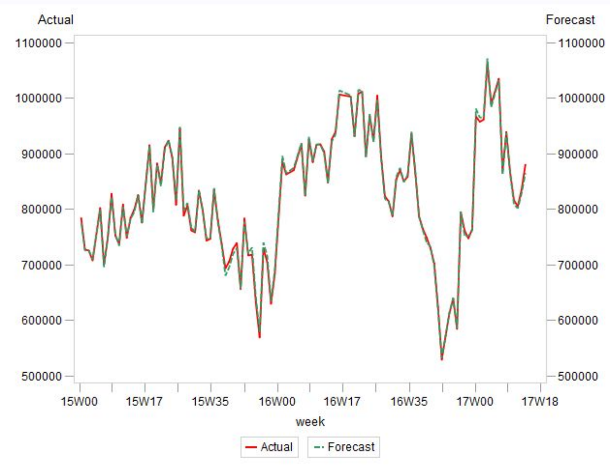
Pearson Correlation Coefficients, N = 118
Prob > |r| under H0: Rho=0

	small_size	large_size	xlarge_size	Small_Bags	Total_Bags	Large_Bags	XLarge_Bags
small_size	1.00000	0.57505	0.27412	0.14996	0.08956	-0.05492	-0.05257
		<.0001	0.0027	0.1051	0.3348	0.5547	0.5718
large_size	0.57505	1.00000	0.70750	0.08926	0.11094	0.15336	0.03543
	<.0001		<.0001	0.3364	0.2317	0.0973	0.7033
xlarge_size	0.27412	0.70750	1.00000	0.13567	0.16291	0.18577	0.31814
	0.0027	<.0001		0.1430	0.0780	0.0440	0.0004
Small_Bags	0.14996	0.08926	0.13567	1.00000	0.98802	0.84097	0.61159
	0.1051	0.3364	0.1430		<.0001	<.0001	<.0001
Total_Bags	0.08956	0.11094	0.16291	0.98802	1.00000	0.91342	0.57907
	0.3348	0.2317	0.0780	<.0001		<.0001	<.0001
Large_Bags	-0.05492	0.15336	0.18577	0.84097	0.91342	1.00000	0.36307
	0.5547	0.0973	0.0440	<.0001	<.0001		<.0001
XLarge_Bags	-0.05257	0.03543	0.31814	0.61159	0.57907	0.36307	1.00000
	0.5718	0.7033	0.0004	<.0001	<.0001	<.0001	

Multiple Regression

- **Lower MAD, RMSE, MAPE** than Holt's
- Best model thus far

Training



Validation

Training

Variable	Mean
square	1.0000990
abs	4340.34
proportion	1.249526E-6
abs_proportion	0.0054476

Page Break

Multiple Regression training error term

MAD	MSE	RMSE	MAPE	MPE
4340.34	1.00010	1.00005	0.00012%	0.54476%

Validation

Variable	Mean
square	224139273
abs	11263.80
proportion	0.0060003
abs_proportion	0.0121212

Page Break

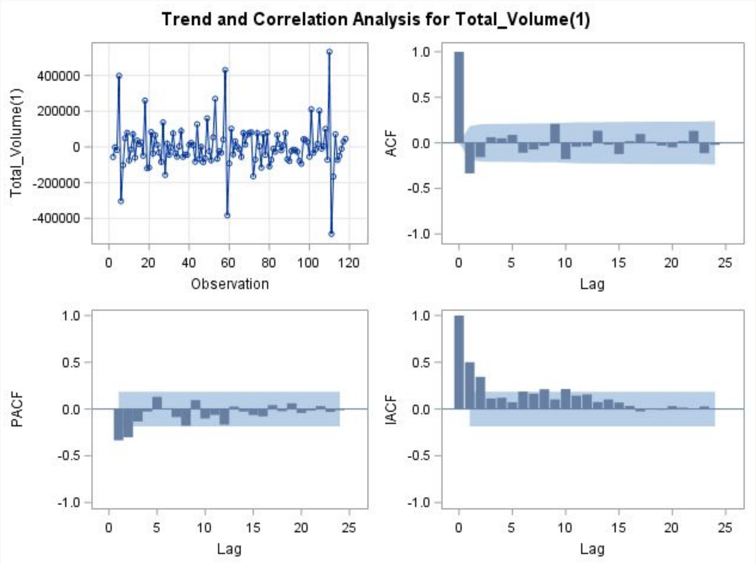
Multiple Regression validation error term

MAD	MSE	RMSE	MAPE	MPE
11263.80	224139273.41	14971.28	0.60003%	1.21212%

Box-Jenkins (ARIMA)

- Ran MA(1) and differentiation order 1 on a non-seasonal component → realized there is a seasonal component present at lag 9

- The arima model fitted to ARIMA (0,1,1)(0,1,1)9



Conditional Least Squares Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MA1,1	0.45087	0.06992	6.45	<.0001	1
MA1,2	0.52991	0.07502	7.06	<.0001	9

Box-Jenkins (ARIMA)

- Not a great model in comparison other forecasting model
- **Higher MAD** than Holt's and Multiple Regression

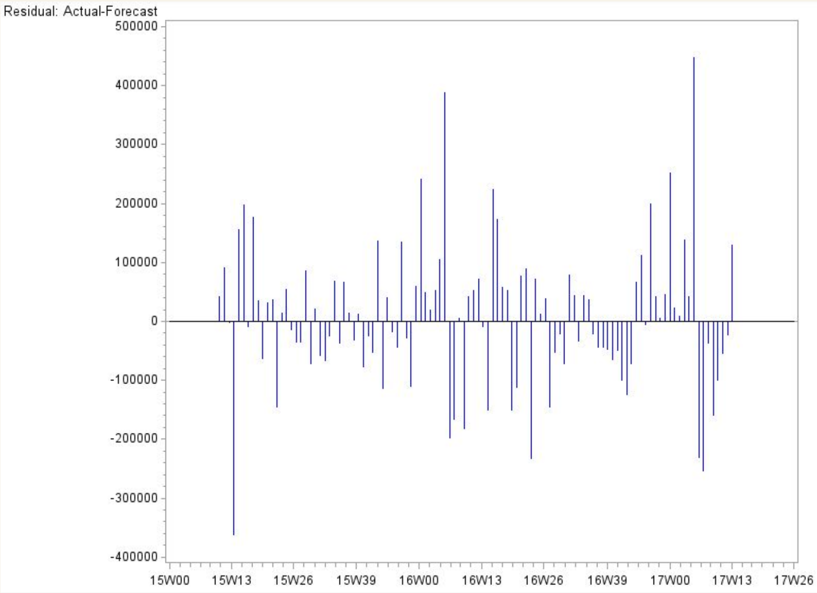
ARIMA training error term

MAD	MSE	RMSE	MAPE	MPE
79530.10	13066396354	114308.34	(0.01509%)	9.18725%

Page Break

ARIMA validation error term

MAD	MSE	RMSE	MAPE	MPE
181334.35	52845335073	229881.13	(5.53429%)	19.7540%



Best Forecasting Model

Forecast Method	MAD	MAPE	RMSE
Naive	75900.02	0.34686%	133356.60
Simple Average	24972.85	6.67%	35107.64
Moving average (3)	40235.27	4.68%	70618.02
Moving average (5)	31006.34	3.72%	52575.07
Holt's Linear Exponential Smoothing	24019.72	2.78%	32150.34
Multiple Linear Regression	11263.80	0.60003%	14971.28
ARIMA	181334.35	5.534%	229881.13

Thank you!

Any Questions?

Bibliography

Kiggins, J. (2018, June 06). Avocado Prices. Retrieved from https://www.kaggle.com/neuromusic/avocadoprices?fbclid=IwAR0nPspy_aWItqKz9d5wPOivdYsKrovAx9jdyHbPYX40tPMuLHJs4i kfQ9w

Fresh Avocado (2021). How to identify Hass Avocados. Retrieved from https://loveonetoday.com/how-to/identify-hass-avocados/

Weaver Street Market (2018). Peak Season for Avocados. Retrieved from https://www.weaverstreetmarket.coop/peak-season-for-avocados/#:~:text=Avocados%20are%20available%20year%20round,texture%20that%20we%20all%20love.

Hyndman, R.J., & Athanasopoulos, G. (2018). Forecasting: principles and practice, 2nd edition,

OTexts: Melbourne, Australia. Retrieved from: OTexts.com/fpp2.