

Predictive Model for Average Avocado Prices

December 16, 2020

Executive Summary:

Introduction: Avocado is a fruit that is originated from southern America. This fruit is extremely health and have a lot of benefits and nutritions. Individuals can use avocados with any other ingredients to complete a meal such as toast and salad. In addition, avocados can also be used to make healthy oil or desserts like avocado ice cream or smoothie. Knowing the fact that avocados are health for a human body, avocados have been the rise of American's new favorites fruits since the last decade. The amount of avocados have been sold in America are higher and higher everyday. With this being said, knowing the variables that cause the price of avocados to go up and down would benefits all consumers and restaurants owner. For example, a restaurants that have avocado toast or guacamole on their menu would have a better understanding of where to get cheaper avocados to maximize their profit. Knowing the cost of avocado can also help restaurants owner create budget for their restaurant and cost of a dish on their menu. In addition, individuals who like avocado would also know where and when to get type of avocado they desire. There are many questions asked in favor of these issues including 1) What factor impact the average price of avocados? 2) Are characteristics of an avocado important in pricing decision? 3) How good is the model? 4) How can we improve the model? and 5) How can we implement the model for the consumer to gain easy access? This analysis will attempt to answer all these questions by starting with a variable data analysis, developing the model using a multiple linear regression, assessing the quality of the model, and providing significant results of the model. The purpose of this model is to predict the average avocado prices using various variables provided in the dataset.

Methods: A dataset was retrieved from Kaggle, a website that have input and output from scientists and college students. This dataset has a large sample size of 30,021 observations from 2015 to 2020. The data was originally collected from the Hass Avocado Broad (HAB) website. There are no missing values for all the variables in the dataset. This dataset has historical data of avocado prices and characteristics. The dataset contains two time series columns including date of observation and year of observation, one characteristic variable including type of the avocado, and one geographical variable. In addition, there are eight quantitative predictors including total number of avocados sold, total number of avocados with Price Look-Up (PLU) code 4046 sold, total number of avocados with PLU code 4226 sold, total number of avocados with PLU code 4770 sold, total number of bags sold, total number of small bags sold, total number of large bags sold, and total number of extra large bags sold. There are 54 distinct geographical regions of where the avocados are from with different average prices for different time of the year. The goal of this analysis is to find the relationship between predictors and average avocado prices. To build a model with average avocado price, a multiple linear regression with significant predictors will be used. Before building a model, we will explore how each variable impact the average of avocado prices. Then, using the "best" model, we will predict the avocado price to validate our model. The data will be split-

ted into a .75/.25 proportion train/test set. All analysis will be done in R Studio with version 3.6.2.

Exploratory Data Analysis: Before building a model, it is important to explore the distribution of average avocado prices and the relationship of it with each of the predictor. The target variable average price have an approximately normal distribution, see Figure 1.1. With a normal distribution assumption, a multiple linear regression can be applied. The price of avocados range from \$0.44 to \$3.25 with an average of \$1.35 for each avocado. In terms of predictors, a descriptive statistics was constructed for all continuous variable in the dataset, see Table A2 in the Appendix. All continuous variable are right skewed with an extremely large maximum value compare to the means and the third interquartile ranges. This signifies a log-transformation is needed for all continous predictors to minimize skewness effect. In addition, the values of the continuous variables will be smaller and easier to analyze.

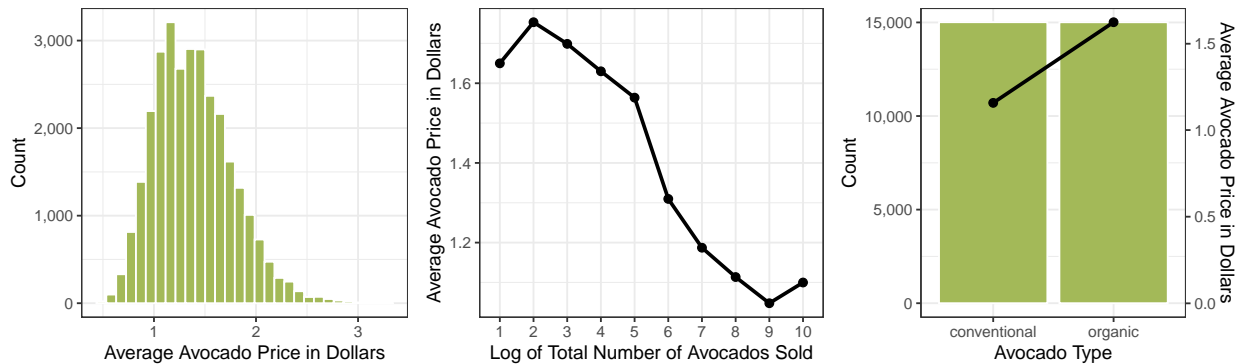


Figure 1

Total number of avocados sold is the number that have been sold in each location and each time of the year. Economically, the price of avocado will be lower if there are more supply. On the other hand, the price will be higher if there is a shortage in avocado. On a consumer side, a cheaper avocado would be more like to be bought than a more expensive one. This leads to a higher sells in avocados if the price is cheaper. Figure 1.2 shows a negative relationship between average avocado price and the log-transformation of total number of avocados sold. The rest of the article will use log of total number of avocados sold, unless otherwise specify. In Figure 1.2, the total number of avocados sold were sorted from smallest to highest and divided into 10 equal buckets. The first bucket contains the smaller number of avocados sold. The last bucket contains the highest number of avocados sold. A negative relationship indicates the average price will be cheaper if there are more avocado sold. However, the price will be more expensive if there are less avocado sold. This have proven the points of less supply lead to higher price. Likewise, higher price of avocados leads to less consupsion. Therefore, the amount of avocado sold will be less. This variable would contribute significant impact in the model to predict the averge of avocado prices.

The type of the avocado is also a signicant factor that would change the price of an avocado. In this dataset, there are two types of avocados including conventional and organic. Conventional avocados are traditoinal growing method that majority of the countries do. The price of conventional avocado would be more likely to be cheaper since the cost of growing conventional avocados are cheaper. On the other hand, organic avocados are grown without using any chemical, fertilizers, pesticides, or other artificial agents. Without a booster, organic avovados take longer to grow

and easier to be eaten by insects. Therefore, the price of organic avocado would be considered more expensive. Figure 1.3 depicts the relationship of average avocado price with avocado types. As seen in the plot shown using a black line, organic avocados have a higher average price than conventional avocado. This indicates that an organic avocado would cause the price to be higher compared to a conventional avocado. Furthermore, this variable, type of avocado, have a balance proportion (shown using two bars). With a balance proportion, the predictive model would be more unbiased. Therefore, type of avocados would contribute a significant impact in predicting the price of avocados.

Month of observation is the time where the avocado is observed. Time of the year is important for fruits and vegetables because majority of the fruit only grow in a certain months or their season. Therefore, we would expect to have more avocados during avocado season and less avocados otherwise. As mentioned before, the price of avocado would be more like to increase if there are less avocados in supply. This means that, the average price would increase during non-avocado seasons and decrease during avocado seasons. Figure 2.1 shows a plot of average avocado price by month of observation with its frequency. As one can see, avocado seasons can be assume to be between January and May with the highest number of observations. On the other hand, non-avocado seasons are likely to be in between the month of June to December with lower number of avocados observed. In addition, a black line depicts lower average prices from January to May and higher prices from June to December. This indicates that during avocados season, the price of avocados are lower because there are more avocados in supply. On the other hand, the low supply of avocados from June to December causes the average price to go up. Therefore, this variable month of the year could contribute significant relationship with average price in the linear regression model.

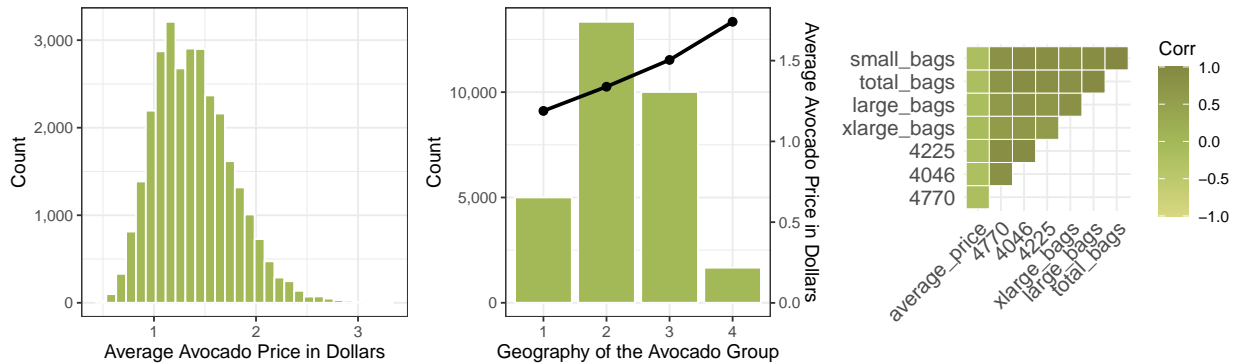


Figure 2

Last but not least, location of the avocados sold could also be significant to the model. The geography of the avocado is the region or city the avocado is sold. Different cities and states have different living expenses resulting in different prices of the avocados. For example, the living expenses in California is higher than Texas. Thus, the price of avocados in California cities would be higher than the prices of avocados in Texas. Figure 2.2 depicts a relationship between average avocado prices and geographical location. This variable is sorted from smallest to highest average price and assigned four geographical groups. The first geographical group sells the cheapest avocados. The regions in the first group include Cincinnati, Columbus, Dallas, Houston, Nashville, New Orleans, Phoenix, Roanoke, and South Central. Furthermore, the last geographical group sells the most

expensive avocados. Those regions include Hartford, New York, and San Francisco. Knowing the location of the market, the model can provide more variation of the average price.

Other variables like total number of avocados with Price Look-Up (PLU) code 4046 sold, total number of avocados with PLU code 4226 sold, total number of avocados with PLU code 4770 sold, total number of bags sold, total number of small bags sold, total number of large bags sold, and total number of extra large bags sold also have a strong negative relationship with the average prices. However, the correlation of all the continuous independent variable are very high, majority above .80, see Figure 2.3. If all high correlated variables are used within the same model, multicollinearity issues will arise. This leads the coefficients of each predictors to be unstable. Therefore, only total number of avocados sold will be use in the model. Total volume would describe total number of all different categories added up. In other words, the variables total number of avocados with Price Look-Up (PLU) code 4046 sold, total number of avocados with PLU code 4226 sold, total number of avocados with PLU code 4770 sold, total number of bags sold, total number of small bags sold, total number of large bags sold, and total number of extra large bags sold are a component of total volume. Using just one variable total volume alone would be good enough to cover the details of each components. Other variables did not mentions in the exploratory analysis have little to no relationship with average avocado prices.

Model Fitting/Inferences: Now that we have found all the significant variables to predict the average price of avocados, predictive models can be achieved. As mentioned before, a train set of .75 of the proportion of the data was used to train the models. The data splitting process was random partition to select random observations and stratified sampling to balance the frequency of target variable. An initial model with all significant variables was built including total number of avocados sold, month of observation, avocado type, and geographical group. Here, month of observation, avocado type, and geographical group were treated as categorical variables and total volume was treated as a continuous variable. This model has an adjusted R-square of .57 indicating a 57% of variation in average price can be explained by all these predictors. As seen in Table 1, all coefficients have the right magnitude consistent with the exploratory analysis. In addition, the p-values for all variables are approximately zeros and their confident intervals do not contain zeros. This signifies that all variables in this model are significant and needed in the model.

To further select variables for the final model, a stepwise variable selection process was ran using Akaike information criterion (AIC). Three validation metrics were used to compare the initial model with the stepwise model including mean square error (MSE), AIC, and adjusted R-square, see Table A3. After a stepwise function was ran, the results suggest the model to keep the same variables. Table A3 depicts that model 1 and model 2 have the same features and the same values for all the validation metrics. This indicates that the four variables total number of avocados sold, month of observation, avocado type, and geography group would form the best model to predict avocado price.

An interaction model was also constructed using all combination of variables in the previous model. A stepwise selection method was also applied to the interaction model to select the best variables among them using AIC methods. Looking at Table A3, model with interaction terms and stepwise

Term	Coef	SdError	F-Stat	pValue	2.5% CI	97.5% CI
(Intercept)	1.228	0.018	70.111	0.000	1.193	1.262
log_total_volume	-0.029	0.001	-23.237	0.000	-0.031	-0.026
month2	-0.037	0.008	-4.712	0.000	-0.052	-0.021
month3	0.031	0.008	4.066	0.000	0.016	0.045
month4	0.094	0.008	12.402	0.000	0.079	0.109
month5	0.080	0.008	10.502	0.000	0.065	0.095
month6	0.134	0.008	16.547	0.000	0.118	0.149
month7	0.198	0.008	25.137	0.000	0.182	0.213
month8	0.223	0.008	27.667	0.000	0.207	0.239
month9	0.242	0.008	30.527	0.000	0.227	0.258
month10	0.192	0.008	24.207	0.000	0.176	0.207
month11	0.113	0.008	13.961	0.000	0.097	0.129
month12	0.026	0.008	3.202	0.001	0.010	0.043
typeorganic	0.367	0.005	67.379	0.000	0.356	0.377
geography_bins2	0.160	0.005	33.397	0.000	0.150	0.169
geography_bins3	0.309	0.005	61.764	0.000	0.299	0.318
geography_bins4	0.563	0.008	69.124	0.000	0.547	0.579

Table 1: Summary regression of final model

model with interaction terms have much lower AIC. However, adjusted R-squared is only a slightly better, 1.6% to be exact. In addition, MSE values for interaction models are also slight better than models without interaction terms. With a slightly better in MSE and adjusted R-square, it would be better off to choose the initial model. In addition, models with interaction terms have 77 coefficients, which is too many variables to deal with in a model. The models without interaction terms only have 16 coefficients. Therefore, this analysis will choose the initial model to be the final model with four variables total volume, month of the year, avocado type, and location.

After choosing a model, it is important to assess the robustness of the model. In the model diagnostics stage, there were four criterion used to assess the quality of the model including studentized residuals, Cook's distance, and variance inflation factor (VIF) scores. First of all, studentize residuals gives the residuals values between actual and predicted average avocado price. Studentize residuals can give us the outliers with respect to the target variable, average price. As seen in Figure 3.1, the studentized residuals are approximately normal with the mode at 0. This passed the mean assumption for studentized residuals of a model. Moreover, looking at Figure B4, the residuals seem to be random with an similar amount of positive and negative residuals. In addition, this plots depicts that residuals are more likely to be within absolute value of 2.5. There are some values that have extremely high residuals including observations 7339, 9608, 9640, and 11523. These observations have very high average avocado price compare to other, see Table A4.

Cook's distance is also one of the most important criteria to check the influential points with respect to the predictors. If Cook's distance for an observation is relative high, this observation could be an influential point. Figure 3.2 depicts a Cook's distance plot for each observation point in the training data. All points are mostly smaller than 0.0005. However, there are five extremely high Cook's distance that needed attention include 7928, 7846, 7683, 9208, and 9640. These values are influential points that originated from independent variables. As seen in Table A4, total number of avocados sold should be higher in the first half of the year. However, none of the observations

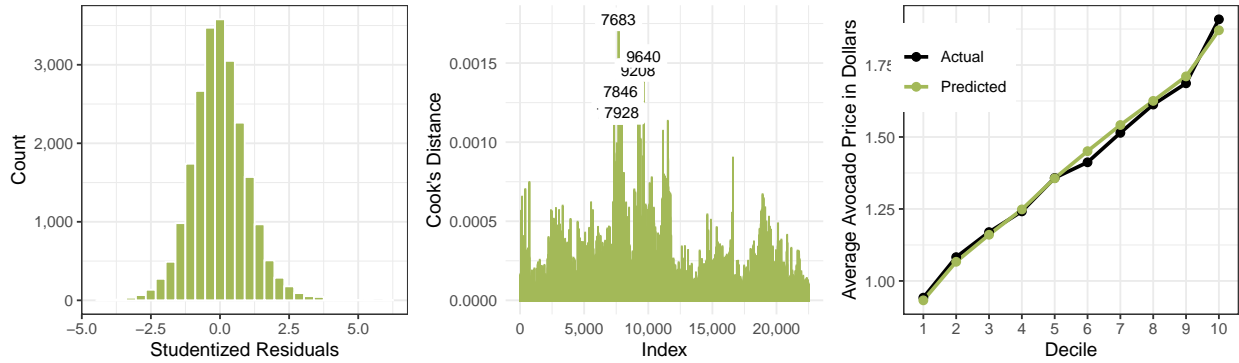


Figure 3

were decided to remove from the data since there are no sufficient reasons to remove them. Last but not least, the VIF scores was generated to check if there is any multicollinearity issue presented in the model. However, there are no signal saying that there would be multicollinearity issues in the model. Table A7 shows all VIF values are below 2 which indicate there are no multicollinearity issues.

A .25 proportion of data was hold out to use as a test set to validate the model. After generating prediction, an MSE score was computed to be approximately 0. This indicates the model is predicting really well with an overall mean square error of nearly 0. To further prove this point, a lift chart was constructed. The predicted average avocado price was sorted from smallest to higher then divided into 10 equal deciles. The actual and predicted average price was computed by decile to compare. Figure 3.3 shows a lift chart of empirical and indicated values of average avocado price. In this plot, the two lines are very closed to each other consistent with the model is robust. Therefore, this model can be use to predict the price of avocados.

Now that we have built the model and checked the quality of the model, an interpretation of the model will be present in this section. All values mentioned in this section are from Table 1. The final model is consisted of four variables including total number of avocado sold, month of observation, avocado type, and geographical group. This model has an adjusted R-squared of 0.57 indicating that 57% of the data in the target variable can be explained by four predictors. Total volume has a negative relationship with total average price, according to the model. On average, if total number of avocado sold increases by one percent, the average price will decreases by \$0.0003. Moreover, the month of observation also have significant impact on the price of avocado. As mentioned before, avocados would be cheaper during avocado seasons. Different months of the year would generate a different revenue for avocados. For example, February would have lower avocado price than January by 0.037 times, while March has higher avocado price than January by 0.031 times. The values for the rest of the months can be retrieved at Table 1. Moving on to the next variable avocado type. How the avocados grow have greatly impacted the price of the avocados. An organic avocado would increase the avocado price by 0.367 times compared to conventional avocado. This makes sense because the money goes with the quality. Last but not least, the region of where the avocados were sold are also a significant factor in the model. Regions and cities for each geographical group can be found in Table A6. In general, if an avocado is sold from region 2, the price of avocado increases by 0.160 times compared to region 1. In addition, the price would increase by 0.309 and 0.0563 times if an avocado is sold from region 3 and 4, respectively, compared to region 1.

Conclusion:

Bibliography:

Appendix A: Supplemental Tables

Table 2: Summary Statistics for all numerical independent features

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
total_volume	30,021	939,255	3,813,519	85	14,299	489,803	63,716,144
4046	30,021	299,107	1,289,108	0	783	115,156	22,743,616
4225	30,021	284,901	1,169,078	0	2,814	140,947	20,470,573
4770	30,021	21,629	100,919	0	0	5,424	2,546,439
total_bags	30,021	333,534	1,415,618	0	8,374	159,174	31,689,189
small_bags	30,021	232,126	950,503	0	5,956	112,938	20,550,407
large_bags	30,021	95,185	467,210	0	352	36,068	13,327,601
xlarge_bags	30,021	6,223	38,137	0	0	560	1,022,564

	Model	Number of Features	MSE	Adj.R.squared	F.statistics	AIC
1	Initial Model	16.000	0.062	0.572	1879.873	1421.841
2	Stepwise Model	16.000	0.062	0.572	1879.873	1421.841
3	Model with Interaction Terms	78.000	0.060	0.588	413.437	598.507
4	Stepwise Model with Interaction Terms	77.000	0.060	0.588	418.808	597.053

Table 3: Regression validation metrics including MSE, R-squared adjusted, and AIC

	Index	average_price	log_total_volume	month	type	geography_bins
1	7339	3.030	8.220	10	organic	2
2	7683	3.250	9.723	10	organic	4
3	7846	2.990	9.849	11	organic	4
4	7928	2.940	9.766	11	organic	4
5	9208	3.050	7.634	3	organic	2
6	9640	3.170	8.013	4	organic	2
7	11523	2.990	7.944	10	organic	2

Table 4

	GVIF	Df	GVIF ^{1/(2*Df)}
log_total_volume	2.713	1.000	1.647
month	1.007	11.000	1.000
type	2.673	1.000	1.635
geography_bins	1.031	3.000	1.005

Table 5

	geography	geography_bins
1	Cincinnati/Dayton	1
2	Columbus	1
3	Dallas/Ft. Worth	1
4	Houston	1
5	Nashville	1
6	New Orleans/Mobile	1
7	Phoenix/Tucson	1
8	Roanoke	1
9	South Central	1
10	Atlanta	2
11	Buffalo/Rochester	2
12	Denver	2
13	Detroit	2
14	Great Lakes	2
15	Harrisburg/Scranton	2
16	Indianapolis	2
17	Jacksonville	2
18	Las Vegas	2
19	Los Angeles	2
20	Louisville	2
21	Miami/Ft. Lauderdale	2
22	Midsouth	2
23	Orlando	2
24	Pittsburgh	2
25	Plains	2
26	Portland	2
27	Richmond/Norfolk	2
28	South Carolina	2
29	Southeast	2
30	Tampa	2
31	Total U.S.	2
32	West	2
33	West Tex/New Mexico	2
34	Albany	3
35	Baltimore/Washington	3
36	Boise	3
37	Boston	3
38	California	3
39	Charlotte	3
40	Chicago	3
41	Grand Rapids	3
42	Northeast	3
43	Northern New England	3
44	Philadelphia	3
45	Raleigh/Greensboro	3
46	Sacramento	3
47	San Diego	3
48	Seattle	3
49	Spokane	3
50	St. Louis	3
51	Syracuse	3
52	Hartford/Springfield	4
53	New York	4

	GVIF	Df	$\text{GVIF}^{1/(2 \cdot \text{Df})}$
log_total_volume	2.713	1.000	1.647
month	1.007	11.000	1.000
type	2.673	1.000	1.635
geography_bins	1.031	3.000	1.005

Table 7

Appendix B: Supplemental Figures

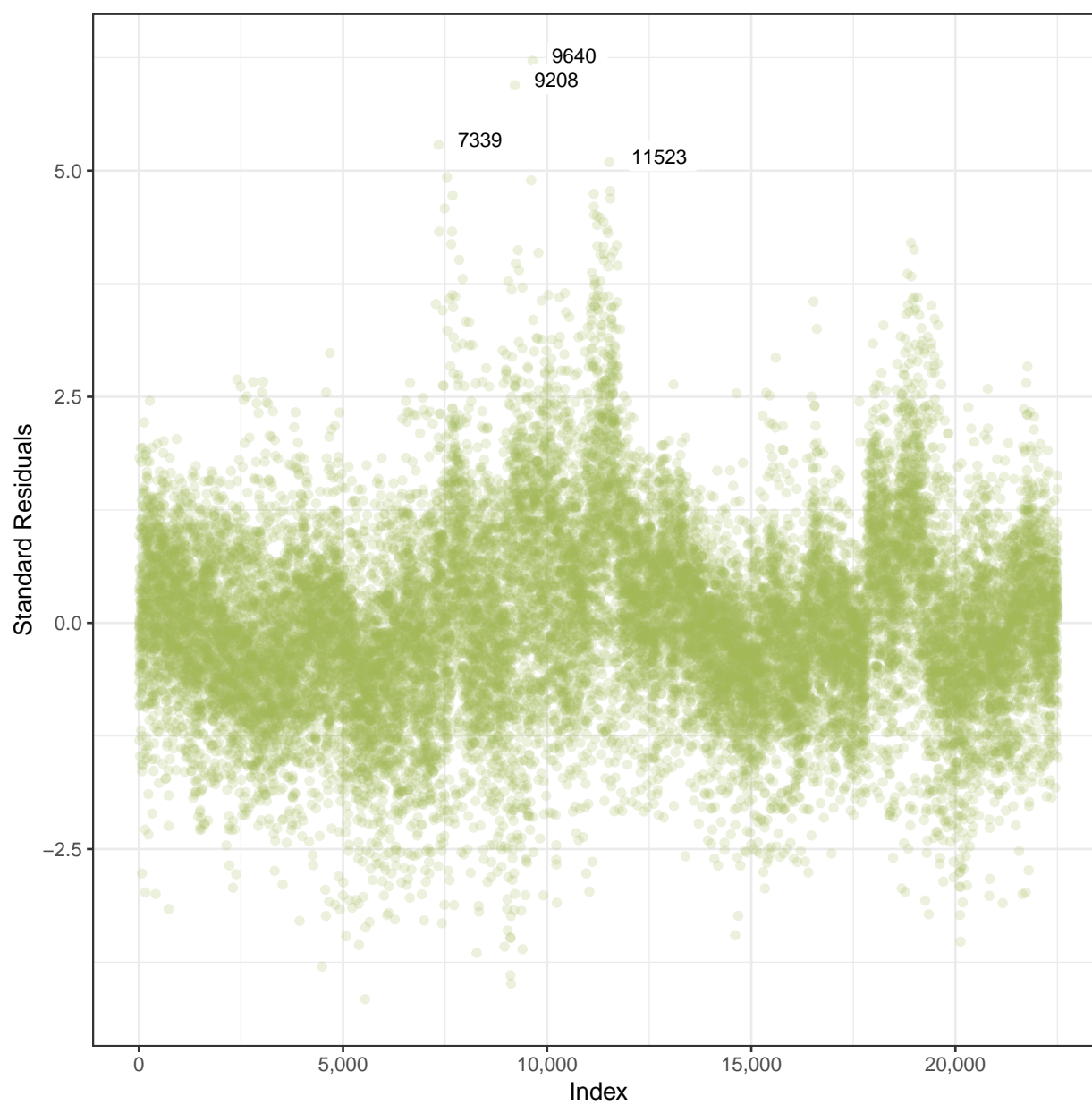


Figure 4

Appendix C: R Code

```

1 ##### Packages #####
2
3 library(magrittr)
4 library(tidymodels)
5 library(lubridate)
6 library(corrplot)
7 library(MASS)
8 library(broom)
9 library(car)
10 library(tidyverse)
11
12
13 ##### Parameters #####
14
15 avocado_color <- "#A3B958"
16
17
18 ##### Functions #####
19
20 ExploreVariable <- function(df, xvar, count = TRUE, x_axis){
21
22   group_df <- df %>%
23     group_by({{xvar}}) %>%
24     summarise(average_price = mean(average_price),
25               count = n(),
26               .groups = "drop")
27
28   ratio <- max(group_df$count) / max(group_df$average_price)
29
30   if(count == TRUE){
31     p <- ggplot(group_df, aes(x = factor({{xvar}}), group = 1)) +
32       geom_bar(aes(y = count), stat = "identity", fill = avocado_color, col = "white") +
33       geom_point(aes(y = average_price * ratio), size = 2, color = "black") +
34       geom_line(aes(y = average_price * ratio), size = 1, color = "black") +
35       scale_y_continuous(sec.axis = sec_axis(~./ratio, name = "Average Avocado Price
36         in Dollars"),
37                           label = scales::comma)
38
39     y_lab <- "Count"
40   } else {
41     p <- ggplot(group_df, aes(x = {{xvar}}, group = 1)) +
42       geom_point(aes(y = average_price), size = 2, color = "black") +
43       geom_line(aes(y = average_price), size = 1, color = "black")
44
45     y_lab <- "Average Avocado Price in Dollars"
46   }
47   p +
48     labs(y = y_lab, x = x_axis) +
49     theme_bw()
50 }
51
52
53 ValidationTable <- function(fit, model_type){

```

```

54 mod <- fit
55 fit_summary <- tibble(Model = model_type,
56                        "Number of Features" = length((coef(mod) %>% names())[1]),
57                        MSE = mean(mod$residuals^2),
58                        Adj.R.squared = summary(mod)$adj.r.squared,
59                        F.statistics = summary(mod)$fstatistic[[1]],
60                        AIC = AIC(mod))
61 return(fit_summary)
62 }
63
64 ##### Data #####
65
66 avocado <- read_csv("data/avocado-updated-2020.csv")
67
68
69 # descriptive statistics of the data
70
71 summary(avocado %>%
72         select_if(is.numeric))
73
74
75 # add deciles to continuous response
76
77 avocado %>%
78   mutate(
79     total_volume_bins = as_factor(cut(total_volume, breaks = 10,
80                                       include.lowest = TRUE, labels = FALSE)),
81     "4046_bins" = as_factor(cut(`4046`, breaks = 10,
82                                 include.lowest = TRUE, labels = FALSE)),
83     "4225_bins" = as_factor(cut(`4225`, breaks = 10,
84                                 include.lowest = TRUE, labels = FALSE)),
85     "4770_bins" = as_factor(cut(`4770`, breaks = 10,
86                                 include.lowest = TRUE, labels = FALSE)),
87     total_bags_bins = as_factor(cut(total_bags, breaks = 10,
88                                     include.lowest = TRUE, labels = FALSE)),
89     small_bags_bins = as_factor(cut(small_bags, breaks = 10,
90                                     include.lowest = TRUE, labels = FALSE)),
91     large_bags_bins = as_factor(cut(large_bags, breaks = 10,
92                                     include.lowest = TRUE, labels = FALSE)),
93     xlarge_bags_bins = as_factor(cut(xlarge_bags, breaks = 10,
94                                     include.lowest = TRUE, labels = FALSE)))
95
96 # add month
97
98 avocado %>%
99   mutate(month = factor(month(date)))
100
101
102 # feature engineer
103
104 price_by_location <- avocado %>%
105   group_by(geography) %>%
106   summarise(average_price = mean(average_price),
107             .groups = "drop") %>%
108   mutate(average_price_bins = cut(average_price, breaks = 4,
109                                   include.lowest = TRUE, labels = FALSE)) %>%

```

```

110   dplyr::select(geography, geography_bins = average_price_bins) %>%
111   distinct()
112
113   avocado %>%
114     left_join(price_by_location, by = "geography") %>%
115     modify_at("geography_bins", as_factor)
116
117
118   # log transformation of total_volume
119
120   avocado %>%
121     mutate(log_total_volume = log(total_volume)) %>%
122     mutate(log_total_volume_bins = as_factor(cut(log_total_volume, breaks = 10,
123                                                    include.lowest = TRUE, labels = FALSE
124                                                    )))
125
126   ##### Explore #####
127
128   # target variable: average_avocado price
129
130   ggplot(avocado, aes(average_price)) +
131     geom_histogram(fill = avocado_color, bins = 30, col = "white") +
132     theme_bw() +
133     labs(x = "Average Avocado Price in Dollars", y = "Count") +
134     scale_y_continuous(label = scales::comma)
135
136
137   # continuous predictors
138
139   ExploreVariable(avocado, total_volume_bins, count = FALSE,
140                   x_axis = "Total Number of Avocados Sold")
141   ExploreVariable(avocado, log_total_volume_bins, count = FALSE,
142                   x_axis = "Log of Total Number of Avocados Sold")
143
144   ExploreVariable(avocado, `4046_bins`, count = FALSE,
145                   x_axis = "Total Number of Avocados with PLU 4046 Sold")
146   ExploreVariable(avocado, `4225_bins`, count = FALSE,
147                   x_axis = "Total Number of Avocados with PLU 4225 Sold")
148   ExploreVariable(avocado, `4770_bins`, count = FALSE,
149                   x_axis = "Total Number of Avocados with PLU 4770 Sold")
150   ExploreVariable(avocado, total_bags_bins, count = FALSE,
151                   x_axis = "Total Number of Bags Sold")
152   ExploreVariable(avocado, small_bags_bins, count = FALSE,
153                   x_axis = "Total Number of Small Bags Sold")
154   ExploreVariable(avocado, large_bags_bins, count = FALSE,
155                   x_axis = "Total Number of Large Bags Sold")
156   ExploreVariable(avocado, xlarge_bags_bins, count = FALSE,
157                   x_axis = "Total Number of Extra Large Bags Sold")
158
159   # categorical predictors
160
161   ExploreVariable(avocado, type, count = TRUE,
162                   x_axis = "Avocado Type")
163   ExploreVariable(avocado, year, count = TRUE,
164                   x_axis = "Year of Observation")

```

```

165 ExploreVariable(avocado, month, count = TRUE,
166                 x_axis = "Month of Observation")
167 ExploreVariable(avocado, geography, count = FALSE,
168                 x_axis = "Geography of The Avocado") +
169   coord_flip()
170 ExploreVariable(avocado, geography_bins, count = TRUE,
171                 x_axis = "Geography of the Avocado Group")
172
173
174 # correlation
175
176 corr_table <- avocado %>%
177   dplyr::select(total_volume, `4046`, `4225`, `4770`, total_bags, small_bags,
178                 large_bags, xlarge_bags, average_price) %>%
179   cor()
180
181 corr_table %>%
182   {.[order(abs(.[, 1]), decreasing = TRUE),
183      order(abs(.[, 1]), decreasing = TRUE)]} %>%
184   corplot(method = "number", type = "upper")
185
186
187 ##### Model Development #####
188
189 # select significant variables
190
191 avocado %>%
192   dplyr::select(average_price,
193                 log_total_volume,
194                 month,
195                 type,
196                 geography_bins)
197
198
199 # split data into train and test set
200
201 set.seed(123)
202
203 avocado_split <- initial_split(avocado, strata = average_price)
204
205 avocado_train <- training(avocado_split)
206
207 avocado_test <- testing(avocado_split)
208
209
210 # Initial model with all predictors
211
212 init_fit <- lm(average_price ~ .,
213               data = avocado_train)
214
215
216 # variable selection using stepAIC
217
218 step_fit <- stepAIC(init_fit, direction = "both", trace = FALSE)
219
220

```



```

221 # interaction
222
223 int_fit <- lm(average_price ~ .*.,
224             data = avocado_train)
225
226
227 # variable selection for interaction model
228
229 int_step_fit <- stepAIC(int_fit, direction = "both", trace = FALSE)
230
231
232 # generate iteration log
233
234 init_fit_summary <- ValidationTable(init_fit, "Initial Model")
235 step_fit_summary <- ValidationTable(step_fit, "Stepwise Model")
236 int_fit_summary <- ValidationTable(int_fit, "Model with Interaction Terms")
237 int_step_fit_summary <- ValidationTable(int_step_fit, "Stepwise Model with
    Interaction Terms")
238
239 bind_rows(init_fit_summary,
240           step_fit_summary,
241           int_fit_summary,
242           int_step_fit_summary) %>%
243   modify_if(is.numeric, round, 3)
244
245
246 # final model
247
248 final_fit <- init_fit
249
250 final_fit %>%
251   tidy() %>%
252   modify_if(is.numeric, round, 3)
253
254
255 ##### Model Diagnostic #####
256
257 # add rownames
258
259 avocado_train %>%
260   rownames_to_column()
261
262 # Residuals
263
264 avocado_train %>%
265   mutate(predict = predict(final_fit),
266          rstudent = rstudent(final_fit))
267
268
269 ## Influential Observations
270 ## Cook's D plot
271 ## identify D values > 4/(n-p-1) as a guide;
272 ## Cook and Weisberg recommend 0.5 and 1 (R uses these guides in default diagnostic
    plots below)
273
274 cutoff <- 4/((nrow(avocado_train) - length(final_fit$coefficients) - 2))

```

```

275
276 diag <- augment(final_fit) %>%
277   mutate(Index = 1:nrow(.))
278
279 diag %>%
280   mutate(high_cooksd = case_when(
281     .cooks_d > cutoff ~ 1, TRUE ~ 0),
282     col_stdresid = case_when(
283       .std.resid > 0 ~ 1,
284       .std.resid < 0 ~ 0),
285     high_hat = case_when(
286       .hat > .1 ~ 1,
287       TRUE ~ 0))
288
289 ## cook's distant ggplot
290
291 ggplot(diag, aes(x = as.numeric(Index), y = .cooks_d)) +
292   geom_bar(stat = "identity", col = avocado_color) +
293   labs(y = "Cook's Distance") +
294   theme_minimal() +
295   geom_label(data = diag %>% filter(.cooks_d > cutoff + .001),
296             aes(label = Index), label.size = NA, size = 3) +
297   scale_x_continuous(label = scales::comma)
298
299
300 ## Normality of Residuals
301
302 ggplot(diag, aes(x = .std.resid)) +
303   geom_histogram(bins = 30, col = "white", fill = avocado_color) +
304   theme_bw() +
305   labs(x = "Studentized Residuals",
306        y = "Count") +
307   scale_y_continuous(label = scales::comma)
308
309
310 # studentize residual plot
311
312 ggplot(diag, aes(x = Index, y = .std.resid)) +
313   geom_point(alpha = 0.2, col = avocado_color) +
314   labs(y = "Standard Residuals", x = "Index") +
315   theme_bw() +
316   theme(legend.position = "none") +
317   geom_label(data = diag %>% filter(.std.resid > 5 | .std.resid < -5),
318             aes(label = Index), label.size = NA, size = 3, hjust=-.25, vjust=.3) +
319   scale_x_continuous(label = scales::comma)
320
321
322 # outlier and influential points
323
324 diag %>%
325   filter(.std.resid > 5 | .std.resid < -5 | .cooks_d > cutoff + .001) %>%
326   dplyr::select(average_price, log_total_volume, month, type, geography_bins)
327
328
329 ## VIF
330 ## vif score seem very close to 1

```

```

331 vif(final_fit) # closer to 1 the better; 5-10 is moderate
332
333
334 ##### Predictions
335
336 # Account for sigma^2
337
338 sd_fit <- sd(final_fit$resid)
339
340
341 # predict
342
343 avocado_test %<>%
344   mutate(average_price_preds = predict(final_fit, newdata = .))
345
346
347 # errors
348
349 avocado_test %<>%
350   mutate(average_price_error = average_price - average_price_preds)
351
352 mse <- mean(avocado_test$average_price_error)^2
353
354
355 # create lift chart
356
357 avocado_test %<>%
358   mutate(average_price_decile = ntile(average_price_preds, n = 10))
359
360 decile_price <- avocado_test %<>%
361   group_by(average_price_decile) %>%
362   summarise(Actual = mean(average_price),
363             Predicted = mean(average_price_preds),
364             .groups = "drop") %>%
365   gather(key, price, -average_price_decile)
366
367 ggplot(decile_price,
368        aes(x = factor(average_price_decile), y = price,
369            group = key, color = key)) +
370   geom_line(size = 1) +
371   geom_point(size = 2) +
372   theme_bw() +
373   scale_color_manual(values = c("black", avocado_color)) +
374   labs(x = "Decile",
375        y = "Average Avocado Price in Dollars") +
376   theme(legend.title = element_blank(), legend.position = c(0.15,.8))

```

Listing 1: Appendix of Code