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 extremaly health and have a lot of benefits and nutritions. Individuals can use avocados with any other ingridients 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 signficant 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 splitted 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 interquatile 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.

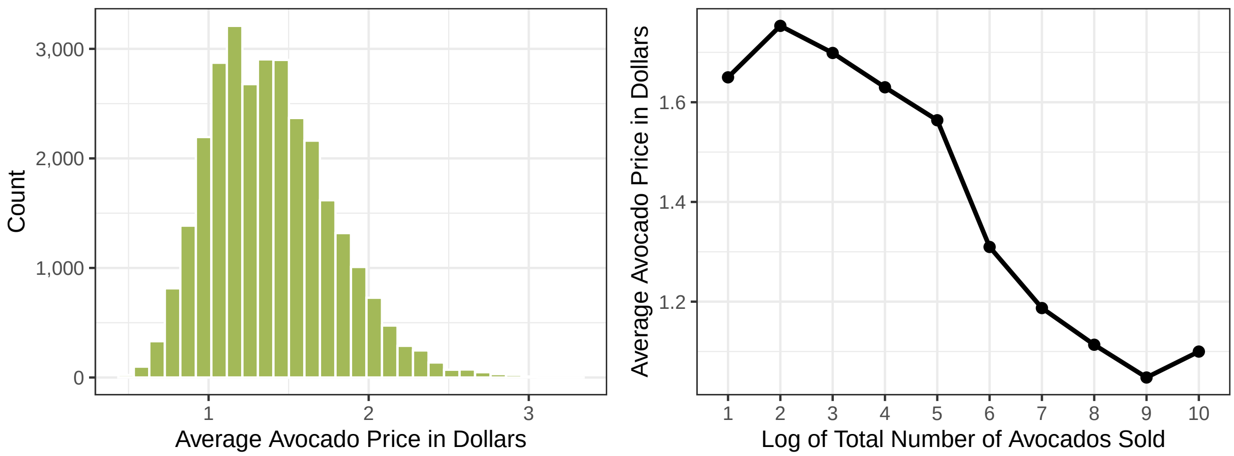


Figure 1

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 solds. 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 consumpsion. 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 growed 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 2.2 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.

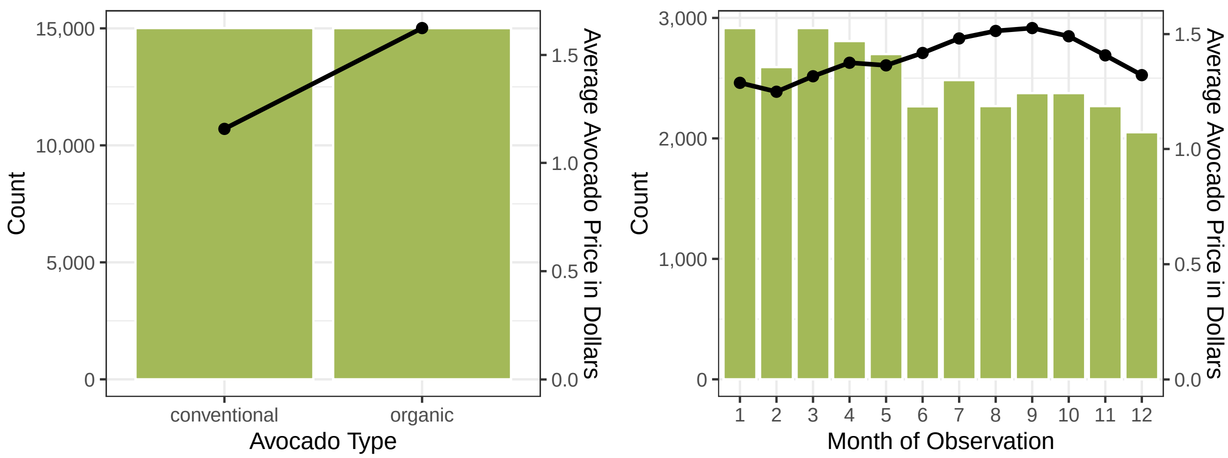
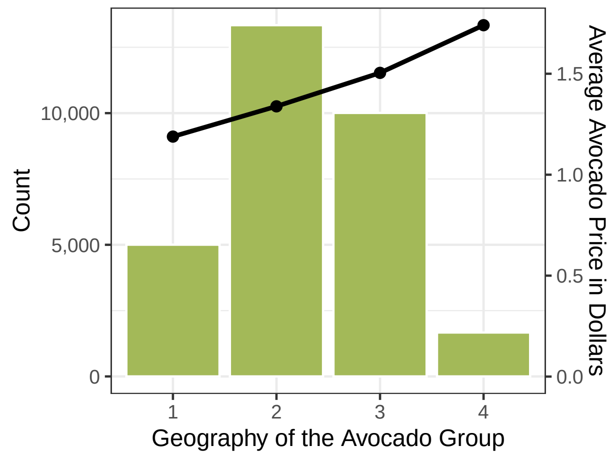


Figure 2

Month of observation is the time where the avocado is observed. Time of the year is important for fruits and vegitables 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 2.2 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.

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 5.1 depicts a relationship between average avocado prices and geographical location. This variables is sorted from smallest to highest average price and assign four geographical group. The first geographical group sells 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.



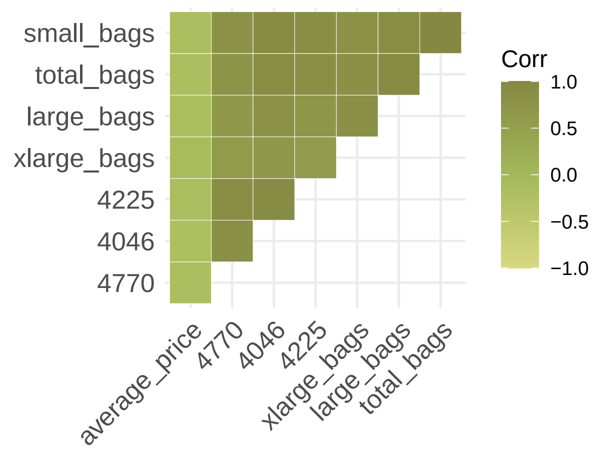


Figure 3

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 5.2. If all high correlated variables are used within the same model, multicolinarity 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:

Conclusion:

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

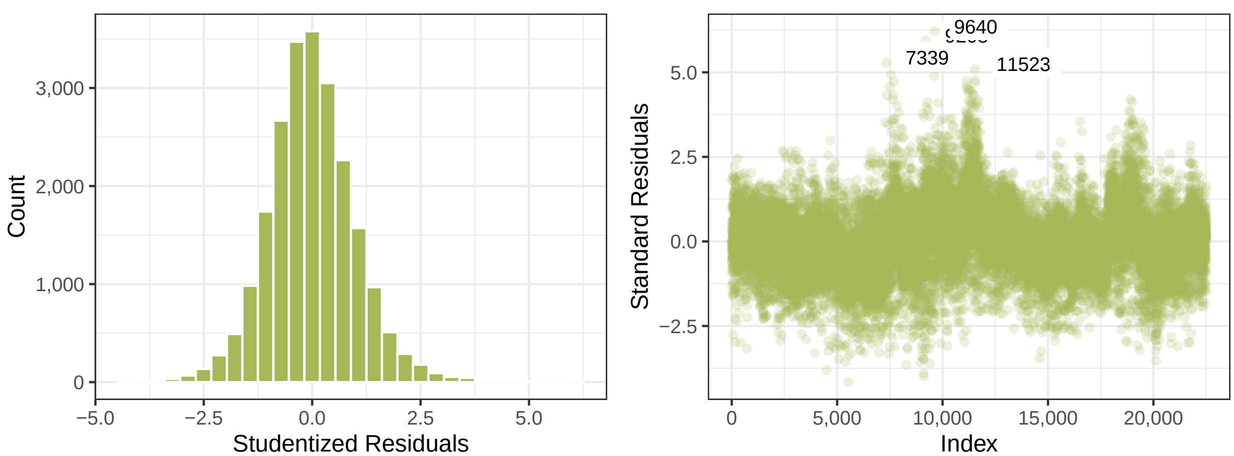


Figure 4

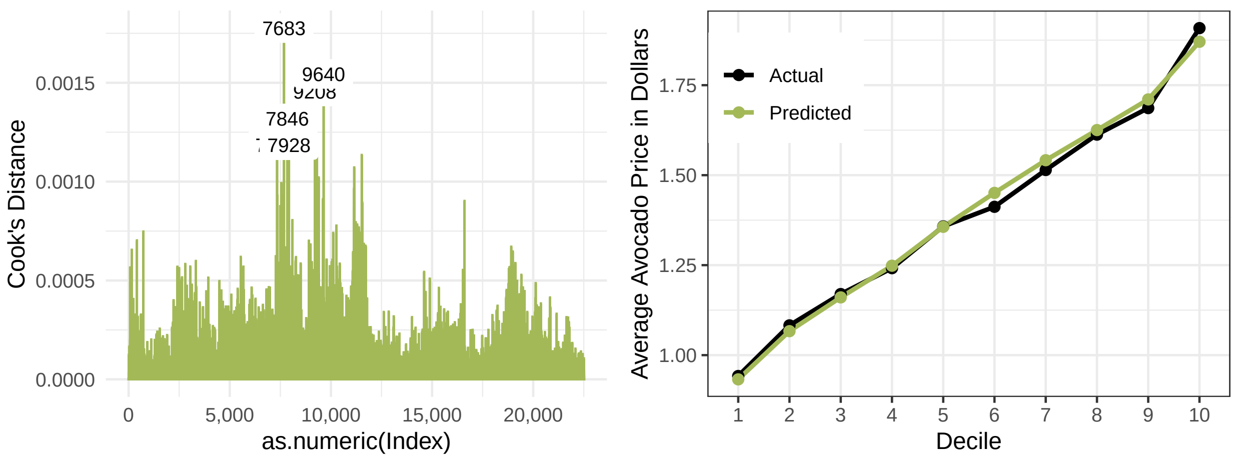

7683
9640
9208
7846
7928
Cook's Distance
0.0015
0.0010
0.0005
0.0000
0
5,000
10,000
15,000
20,000
as.numeric(Index)
Average Avocado Price in Dollars
Actual
Predicted
1.75
1.50
1.25
1.00
1
2
3
4
5
6
7
8
9
10
Decile

Figure 5

Introduction:

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

	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 4

Appendix B: R Code

```r
#### Packages ####

library(magrittr)
library(tidymodels)
library(lubridate)
library(corrplot)
library(MASS)
library(broom)
library(car)
library(tidyverse)


#### Parameters ####

avocado_color <- "#A3B958"


#### Functions ####

ExploreVariable <- function(df, xvar, count = TRUE, x_axis){

  group_df <- df %>%
    group_by({{xvar}}) %>%
    summarise(average_price = mean(average_price),
              count = n(),
              .groups = "drop")

  ratio <- max(group_df$count) / max(group_df$average_price)

  if(count == TRUE){
    p <- ggplot(group_df, aes(x = factor({{xvar}}), group = 1)) +
      geom_bar(aes(y = count), stat = "identity", fill = avocado_color, col = "white") +
      geom_point(aes(y = average_price * ratio), size = 2, color = "black") +
      geom_line(aes(y = average_price * ratio), size = 1, color = "black") +
      scale_y_continuous(sec.axis = sec_axis(~./ratio, name = "Average Avocado Price in Dollars"),
                         label = scales::comma)

    y_lab <- "Count"

  } else {
    p <- ggplot(group_df, aes(x = {{xvar}}, group = 1)) +
      geom_point(aes(y = average_price), size = 2, color = "black") +
      geom_line(aes(y = average_price), size = 1, color = "black")

    y_lab <- "Average Avocado Price in Dollars"
  }
  p +
    labs(y = y_lab, x = x_axis) +
    theme_bw()
}


ValidationTable <- function(fit, model_type){
```

```
  mod <- fit
  fit_summary <- tibble(Model = model_type,
                        "Number of Features" = length((coef(mod) %>% names())[-1]),
                        MSE = mean(mod$residuals^2),
                        Adj.R.squared = summary(mod)$adj.r.squared,
                        F.statistics = summary(mod)$fstatistic[[1]],
                        AIC = AIC(mod))
  return(fit_summary)
}

#### Data ####

avocado <- read_csv("data/avocado-updated-2020.csv")


# descriptive statistics of the data

summary(avocado %>%
          select_if(is.numeric))


# add deciles to continuous response

avocado %<>%
  mutate(
    total_volume_bins = as_factor(cut(total_volume, breaks = 10,
                                      include.lowest = TRUE, labels = FALSE)),
    "4046_bins" = as_factor(cut(`4046`, breaks = 10,
                                include.lowest = TRUE, labels = FALSE)),
    "4225_bins" = as_factor(cut(`4225`, breaks = 10,
                                include.lowest = TRUE, labels = FALSE)),
    "4770_bins" = as_factor(cut(`4770`, breaks = 10,
                                include.lowest = TRUE, labels = FALSE)),
    total_bags_bins = as_factor(cut(total_bags, breaks = 10,
                                    include.lowest = TRUE, labels = FALSE)),
    small_bags_bins = as_factor(cut(small_bags, breaks = 10,
                                    include.lowest = TRUE, labels = FALSE)),
    large_bags_bins = as_factor(cut(large_bags, breaks = 10,
                                    include.lowest = TRUE, labels = FALSE)),
    xlarge_bags_bins = as_factor(cut(xlarge_bags, breaks = 10,
                                     include.lowest = TRUE, labels = FALSE)))

# add month

avocado %<>%
  mutate(month = factor(month(date)))


# feature engineer

price_by_location <- avocado %>%
  group_by(geography) %>%
  summarise(average_price = mean(average_price),
            .groups = "drop") %>%
  mutate(average_price_bins = cut(average_price, breaks = 4,
                                  include.lowest = TRUE, labels = FALSE)) %>%
```

```
  dplyr::select(geography, geography_bins = average_price_bins) %>%
  distinct()

avocado %<>%
  left_join(price_by_location, by = "geography") %>%
  modify_at("geography_bins", as_factor)


# log tranformation of total_volume

avocado %<>%
  mutate(log_total_volume = log(total_volume)) %>%
  mutate(log_total_volume_bins = as_factor(cut(log_total_volume, breaks = 10,
                                               include.lowest = TRUE, labels = FALSE)))


#### Explore ####

# target variable: average_avocado price

ggplot(avocado, aes(average_price)) +
  geom_histogram(fill = avocado_color, bins = 30, col = "white") +
  theme_bw() +
  labs(x = "Average Avocado Price in Dollars", y = "Count") +
  scale_y_continuous(label = scales::comma)


# continuous predictors

ExploreVariable(avocado, total_volume_bins, count = FALSE,
                x_axis = "Total Number of Avocados Sold")
ExploreVariable(avocado, log_total_volume_bins, count = FALSE,
                x_axis = "Log of Total Number of Avocados Sold")

ExploreVariable(avocado, `4046_bins`, count = FALSE,
                x_axis = "Total Number of Avocados with PLU 4046 Sold")
ExploreVariable(avocado, `4225_bins`, count = FALSE,
                x_axis = "Total Number of Avocados with PLU 4225 Sold")
ExploreVariable(avocado, `4770_bins`, count = FALSE,
                x_axis = "Total Number of Avocados with PLU 4770 Sold")
ExploreVariable(avocado, total_bags_bins, count = FALSE,
                x_axis = "Total Number of Bags Sold")
ExploreVariable(avocado, small_bags_bins, count = FALSE,
                x_axis = "Total Number of Small Bags Sold")
ExploreVariable(avocado, large_bags_bins, count = FALSE,
                x_axis = "Total Number of Large Bags Sold")
ExploreVariable(avocado, xlarge_bags_bins, count = FALSE,
                x_axis = "Total Number of Extra Large Bags Sold")

# categorical predictors

ExploreVariable(avocado, type, count = TRUE,
                x_axis = "Avocado Type")
ExploreVariable(avocado, year, count = TRUE,
                x_axis = "Year of Observation")
```

```r
ExploreVariable(avocado, month, count = TRUE,
                x_axis = "Month of Observation")
ExploreVariable(avocado, geography, count = FALSE,
                x_axis = "Geography of The Avocado") +
  coord_flip()
ExploreVariable(avocado, geography_bins, count = TRUE,
                x_axis = "Geography of the Avocado Group")


# correlation

corr_table <- avocado %>%
  dplyr::select(total_volume, `4046`, `4225`, `4770`, total_bags, small_bags,
                large_bags, xlarge_bags, average_price) %>%
  cor()

corr_table %>%
  {.[order(abs(.[, 1]), decreasing = TRUE),
     order(abs(.[, 1]), decreasing = TRUE)]} %>%
  corrplot(method = "number", type = "upper")


#### Model Development ####

# select significant variables

avocado %<>%
  dplyr::select(average_price,
                log_total_volume,
                month,
                type,
                geography_bins)


# split data into train and test set

set.seed(123)

avocado_split <- initial_split(avocado, strata = average_price)

avocado_train <- training(avocado_split)

avocado_test <- testing(avocado_split)


# Initial model with all predictors

init_fit <- lm(average_price ~ .,
               data = avocado_train)


# variable selection using stepAIC

step_fit <- stepAIC(init_fit, direction = "both", trace = FALSE)
```

```r
# interaction

int_fit <- lm(average_price ~ .*.,
              data = avocado_train)


# variable selection for interaction model

int_step_fit <- stepAIC(int_fit, direction = "both", trace = FALSE)


# generate iteration log

init_fit_summary <- ValidationTable(init_fit, "Initial Model")
step_fit_summary <- ValidationTable(step_fit, "Stepwise Model")
int_fit_summary <- ValidationTable(int_fit, "Model with Interaction Terms")
int_step_fit_summary <- ValidationTable(int_step_fit, "Stepwise Model with Interaction Terms")

bind_rows(init_fit_summary,
          step_fit_summary,
          int_fit_summary,
          int_step_fit_summary) %>%
  modify_if(is.numeric, round, 3)


# final model

final_fit <- init_fit

final_fit %>%
  tidy() %>%
  modify_if(is.numeric, round, 3)


#### Model Diagnostic ####

# add rownames

avocado_train %<>%
  rownames_to_column()

# Residuals

avocado_train %<>%
  mutate(predict = predict(final_fit),
         rstudent = rstudent(final_fit))


## Influential Observations
## Cook's D plot
## identify D values > 4/(n-p-1) as a guide;
## Cook and Weisberg recommend 0.5 and 1 (R uses these guides in default diagnostic plots below)

cutoff <- 4/((nrow(avocado_train) - length(final_fit$coefficients) - 2))
```

```
diag <- augment(final_fit) %>%
  mutate(Index = 1:nrow(.))

diag %<>%
  mutate(high_cooksd = case_when(
    .cooksd > cutoff ~ 1, TRUE ~ 0),
    col_stdresid = case_when(
      .std.resid > 0 ~ 1,
      .std.resid < 0 ~ 0),
    high_hat = case_when(
      .hat > .1 ~ 1,
      TRUE ~ 0))

## cook's distant ggplot

ggplot(diag, aes(x = as.numeric(Index), y = .cooksd)) +
  geom_bar(stat = "identity", col = avocado_color) +
  labs(y = "Cook's Distance") +
  theme_minimal() +
  geom_label(data = diag %>% filter(.cooksd > cutoff + .001),
             aes(label = Index), label.size = NA, size = 3) +
  scale_x_continuous(label = scales::comma)


## Normality of Residuals

ggplot(diag, aes(x = .std.resid)) +
  geom_histogram(bins = 30, col = "white", fill = avocado_color) +
  theme_bw() +
  labs(x = "Studentized Residuals",
       y = "Count") +
  scale_y_continuous(label = scales::comma)


# studentize residual plot

ggplot(diag, aes(x = Index, y = .std.resid)) +
  geom_point(alpha = 0.2, col = avocado_color) +
  labs(y = "Standard Residuals", x = "Index") +
  theme_bw() +
  theme(legend.position = "none") +
  geom_label(data = diag %>% filter(.std.resid > 5 | .std.resid < -5),
             aes(label = Index), label.size = NA, size = 3, hjust=-.25, vjust=.3) +
  scale_x_continuous(label = scales::comma)


# outlier and influential points

diag %>%
  filter(.std.resid > 5 | .std.resid < -5 | .cooksd > cutoff + .001) %>%
  dplyr::select(average_price, log_total_volume, month, type, geography_bins)


## VIF
## vif score seem very close to 1
```

```
vif(final_fit) # closer to 1 the better; 5-10 is moderate


#### Predictions

# Account for sigma^2

sd_fit <- sd(final_fit$resid)


# predict

avocado_test %<>%
  mutate(average_price_preds = predict(final_fit, newdata = .))


# errors

avocado_test %<>%
  mutate(average_price_error = average_price - average_price_preds)

mse <- mean(avocado_test$average_price_error)^2


# create lift chart

avocado_test %<>%
  mutate(average_price_decile = ntile(average_price_preds, n = 10))

decile_price <- avocado_test %>%
  group_by(average_price_decile) %>%
  summarise(Actual = mean(average_price),
            Predicted = mean(average_price_preds),
            .groups = "drop") %>%
  gather(key, price, -average_price_decile)

ggplot(decile_price,
       aes(x = factor(average_price_decile), y = price,
           group = key, color = key)) +
  geom_line(size = 1) +
  geom_point(size = 2) +
  theme_bw() +
  scale_color_manual(values = c("black", avocado_color)) +
  labs(x = "Decile",
       y = "Average Avocado Price in Dollars") +
  theme(legend.title = element_blank(), legend.position = c(0.15,.8))
```

Listing 1: Appendix of Code