Predictive Model for Average Avocado Prices

December 16, 2020
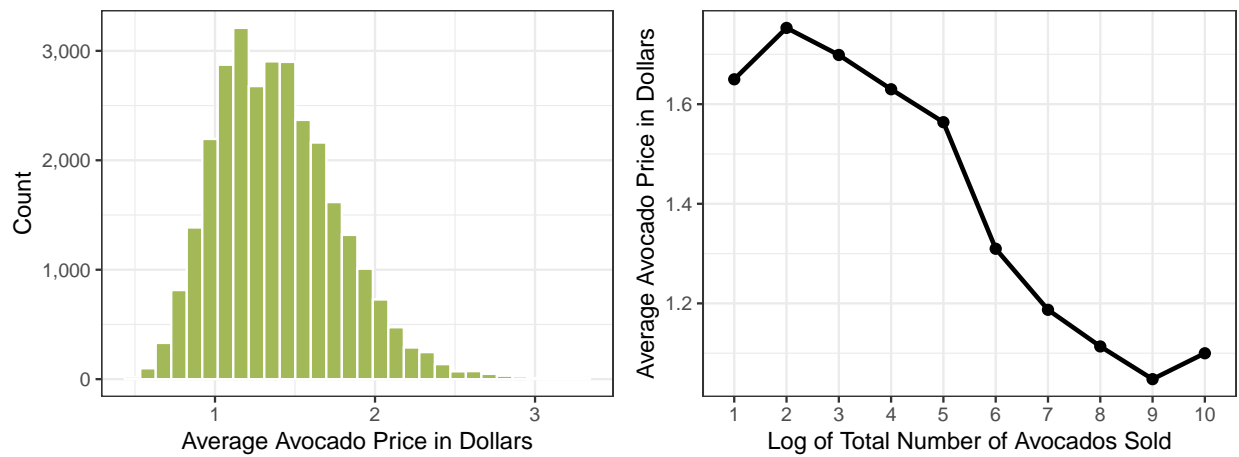
Executive Summary:

Introduction:

Methods:

Exploratory Data Analysis:
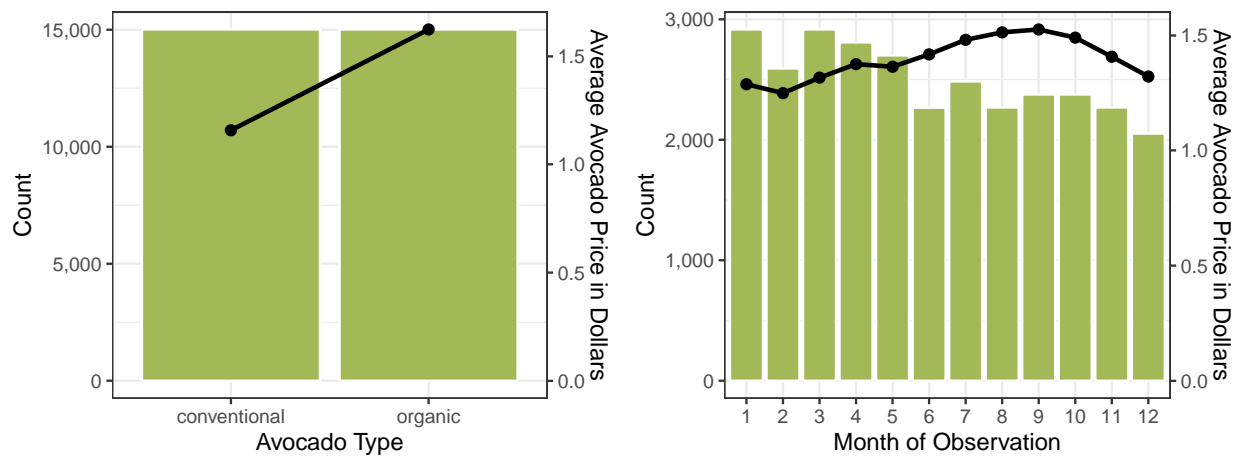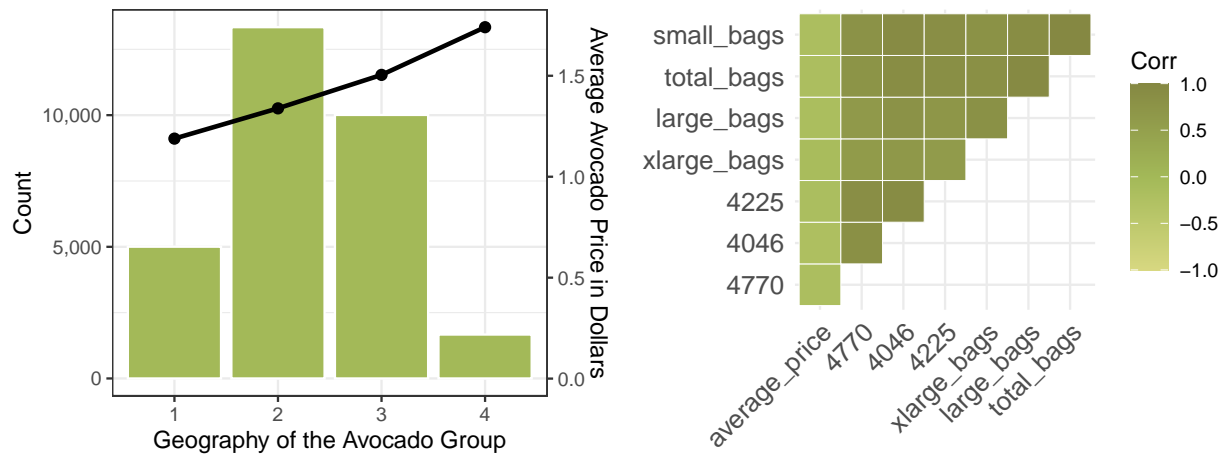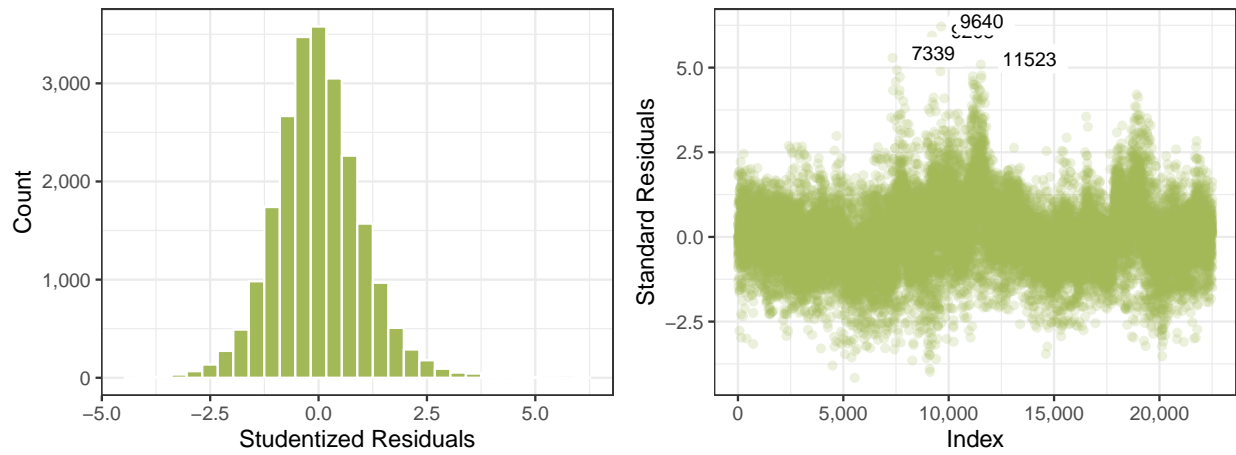

Figure 1

Figure 2

Figure 3

Figure 4
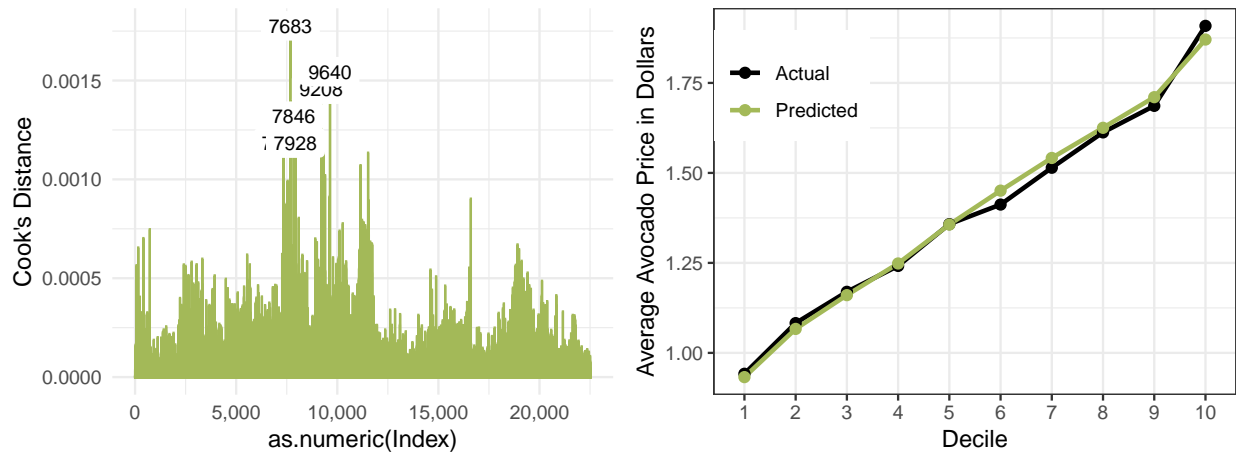
Figure 5

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 Fitting/Inferences:

Conclusion:

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
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 in Dollars"),
36                          label = scales::comma)
37
38     y_lab <- "Count"
39
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){
```

```r
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)) %>%
```

```r
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 tranformation 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")
```

```r
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    corrplot(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
```

```r
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))
```

```r
275
276 diag <- augment(final_fit) %>%
277   mutate(Index = 1:nrow(.))
278
279 diag %<>%
280   mutate(high_cooksd = case_when(
281     .cooksd > 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 = .cooksd)) +
292   geom_bar(stat = "identity", col = avocado_color) +
293   labs(y = "Cook's Distance") +
294   theme_minimal() +
295   geom_label(data = diag %>% filter(.cooksd > 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 | .cooksd > 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
```

```r
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