Profits in Sneaker Re-Selling industry. Weeks since release date and Shoe Size.

Introduction

Sneaker culture has come a long way. Once the symbol of athleticism, sneakers have transcended their primary function to become fashionable object of desire. Flipping premium sneakers (buying and reselling sneakers at cost higher than retail price) has become a highly lucrative occupation for sneakerheads (those who collect shoes) with the global sneaker market predicted to reach US$120 billion by 2026. The process of buying premium sneakers is not as straightforward as it seems to be due to the fierce competition. Due to the limited supply, only a few lucky customers will be able to obtain a pair of rare shoes. Some may end up wearing them themselves, but other would hold on to the pair of newly acquired shoes with the goal to resell them later at a higher price. The rise of marketplace websites like StockX and GOAT streamlined such transaction for general public, encouraging people to sell their shoes more than ever before. These websites act as a middleman between buyers and sellers, making otherwise shady resale market transactions transparent and secure. In addition to convenience, StockX provides both buyers and sellers with usable data like current market value, number of items sold, and loss or gain on items. Nowadays, premium sneaker collectors readily utilize such platforms as the main approach to turning a rare pair of sneakers into cash.

The data for the project was provided by StockX, an online marketplace and sneakers reseller, as a part of their 2019 Data Contest. The data set consists of random sample of Off-White and Yeezy 350 sales from between 9/1/2017 and 2/13/2019 made exclusively in the US. According to organizations, the data sample was curated by taking a random fixes percentage of StockX sales for each colorway, on each day, since September 2017. For example, for each day Off-White Jordan 1 was present on the market, the random selected percentage of its sale was taken from each day. 99,956 cases (27,794 Off-White sales and 72,162 Yeezy sales) were classified according to 8 variable: Order date, Brand, Sneaker Name, Sale Price ($), Release Date, Shoe Size, and Buyer State.

The data set allows us to take a look at general trends of sneaker reselling industry to determine the optimal reselling timing, explore what state could have been considered the epicenter of the sneaker industry, investigate whether sneaker sizes matter in the resale game.

Since the data was taken from open-source available to anyone, some of the individuals might have posted their submission on the internet. The winners of the StockX Data Contest winners were announced on their website, which displayed visualizations of the best submission along with key takeaways from each one of them. The insights this project sheds light on do not build upon any of the visualizations showcased on this website.

Results

Data wrangling: Feature Engineering .

The data set presents multiple sneakers for which the release date varied a lot. In order to account for difference in release dates and the sale date, an additional variable *weeks_after_release* was added to allow cross-comparison of different sneakers over the same point in time after release. A rounding modification of this variable *time_elapsed* was added in order to make graphs more clean. It should be noted that some cases resulted in negative value for this variable, which means that some sneakers were sold pre-release. A new variable *profit_percentage* was added to indicate percent change profit in comparison to retail price, which is constant at $220 for every model of Yeezys no matter the colorway. This variable standarizes each shoe relative to its retail price and acts as a better indicator of the high demand.

Since weeks_after_release is a linear combination of **Order Date** and **Release Date**, these variables were dropped before the analysis. For similar reasons, **Sale Price** and **Retail Price** were dropped, as they are explained by the variable *weeks_after_release*. The **Buyer Region** variable was turned into lowercase to join with other data sets for further analysis.

The dimensions were congruent with the description of the dataset with no cases missing.

```
# load the data set
my_data <- read_excel("StockX-Data-Contest-2019-3.xlsx")
#check the dimensions
dim(my_data)
```

```
## [1] 99956     8
```

```
#data wrangling
sneakers_profits <- my_data %>%
  mutate(profit_percentage = (`Sale Price` - `Retail Price`)/`Retail Price` * 100,
         weeks_after_release = difftime(`Order Date`,`Release Date`, units = "weeks"),
         time_elapsed =  round(weeks_after_release,3)) %>%
         mutate(`Buyer Region` = tolower(`Buyer Region`)) %>%
         select(-c(`Order Date`, `Release Date`)) %>%
         select(-c(`Sale Price`, `Retail Price`))

head(sneakers_profits)
```

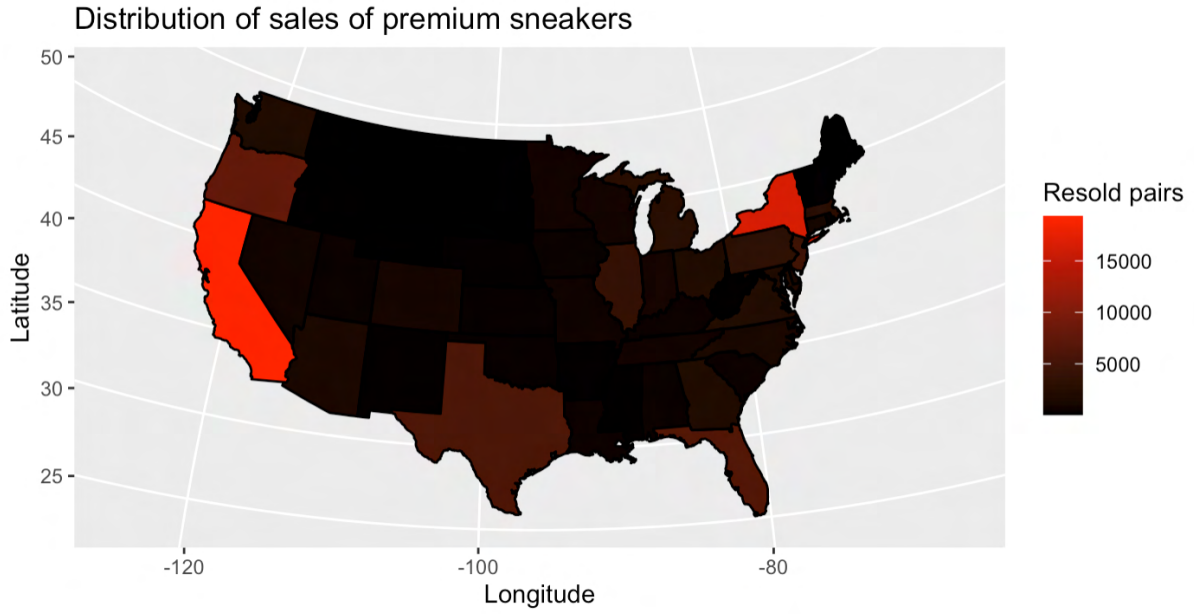
```
## # A tibble: 6 x 7
##   Brand `Sneaker Name`                `Shoe Size` `Buyer Region` profit_percenta~
##   <chr> <chr>                               <dbl> <chr>                     <dbl>
## 1 Yeezy Adidas-Yeezy-Boost-350-Low-~         11   california                 399.
## 2 Yeezy Adidas-Yeezy-Boost-350-V2-C~         11   california                 211.
## 3 Yeezy Adidas-Yeezy-Boost-350-V2-C~         11   california                 214.
## 4 Yeezy Adidas-Yeezy-Boost-350-V2-C~         11.5 kentucky                   389.
## 5 Yeezy Adidas-Yeezy-Boost-350-V2-C~         11   rhode island               276.
## 6 Yeezy Adidas-Yeezy-Boost-350-V2-C~          8.5 michigan                   263.
## # ... with 2 more variables: weeks_after_release <drtn>, time_elapsed <drtn>
```

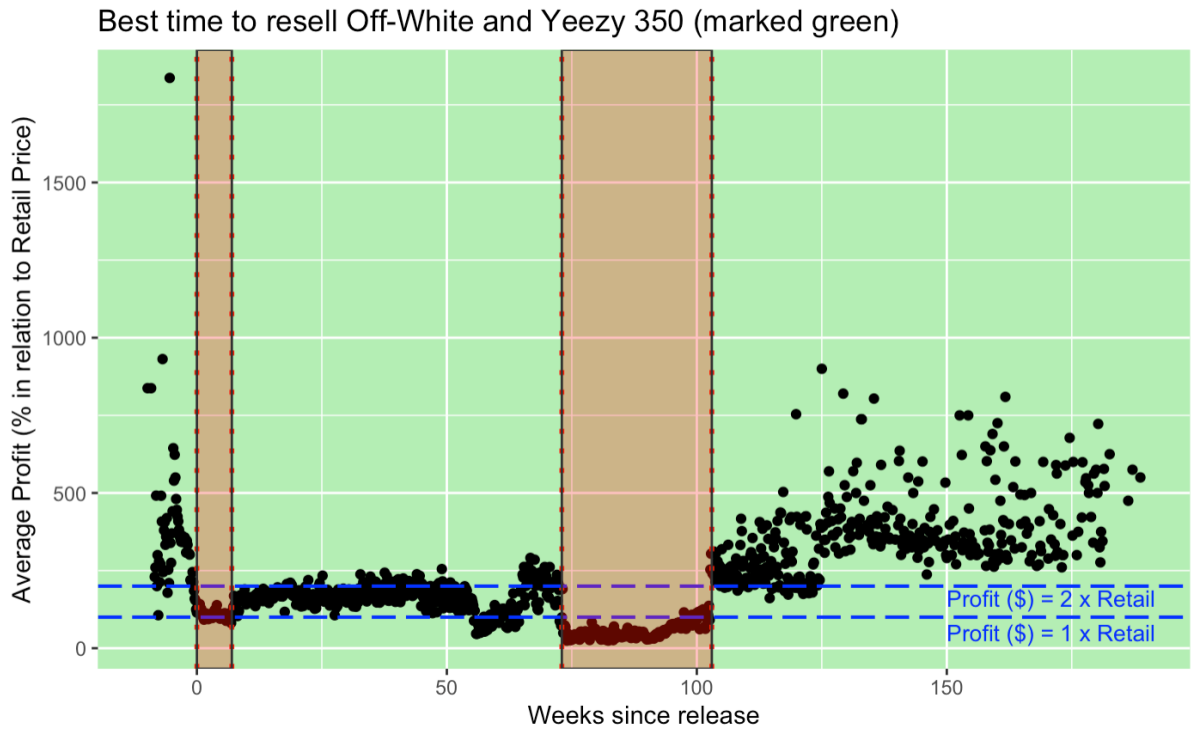
```
# check  whether data contains any missing data
sneakers_profits %>%  is.null() %>% sum()
```

```
## [1] 0
```

Visualize the data: Distribution of buyers across states and Best time to Re-sell

(All the code for visualizations is included in the Appendix)





Judging from the choropleth map, the epicenter of the sneakers industry in 2019 would be considered the state of California. Falling behind by 3000 units sold is New York. These two states by far exceed all other states in number of buyers of collectible sneakers. This interesting trend could be attributed to the general people's ability to pay. According to the average income by state in 2019 statistics, New York ($\$76,450$) and California ($\$72,430$) are among states with the highest average income (New York and California and fifth and sixth positions respectively).

According best reselling time graph, we can see that the most profitable timespan to sell a pair of premium sneakers is before its release, between 1.5 months and 1,4 years since its release date, and after 2 years of its release. The potential seller takes the risks of undervaluing the pair if one decides to sell them for the similar price as competitors in the first month and a half after the release date. It was surprising to find that after 1.5 years and 2 years after the release, there is a dip in the profits. This phenomenon could be explained by the fact that sneakers lose their value, as there is a surplus of older models that are gradually being replaced by their substitutes However, 2 years after the release date the profit skyrocket due to the shortage of older models on the market. A potential seller could expect, on average, to get more than triple of the retail price for the new pair of collectible sneakers 2 years after the release. Comparable profits, on average, could be gained when re-selling sneakers prior to their release date. When selling sneakers between 1.5 months and 1,4 years after their release, the seller could expect, on average, a profit between the initial retail price and twice the retail price of a pair.

Analyses: Regression models and ANOVA analysis

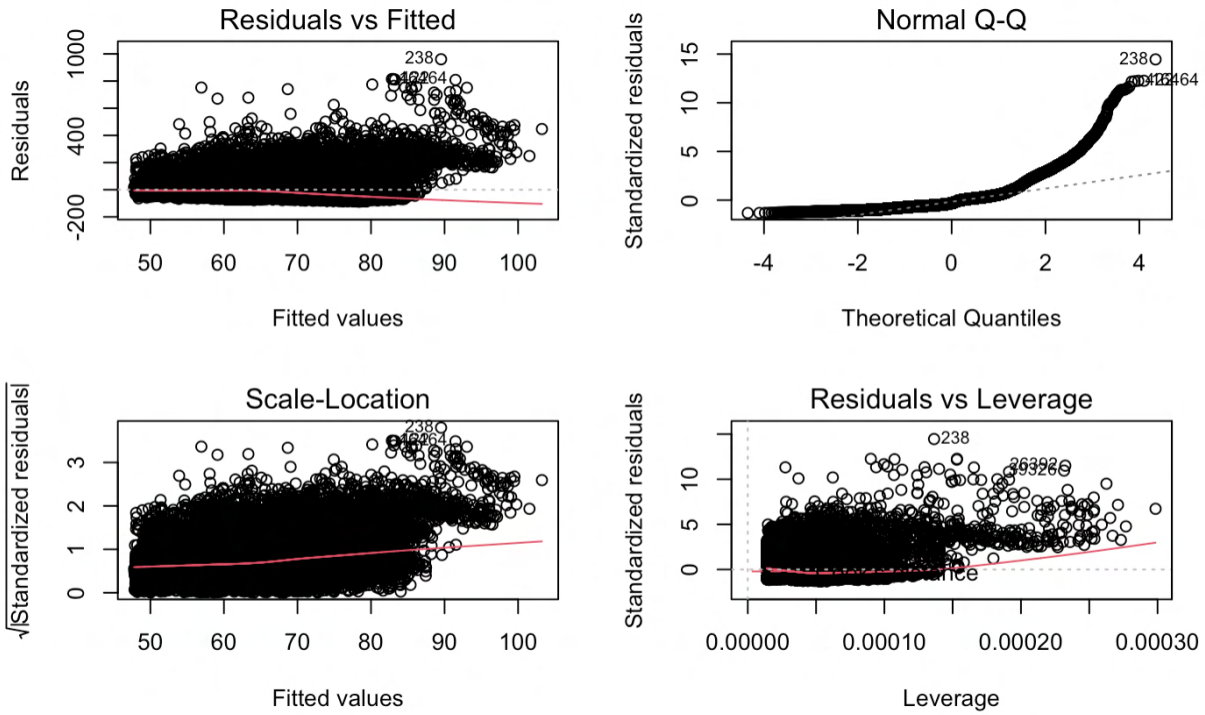
Observed the strong effect that time elapsed since the release of the sneakers has on profit, we would want to back up by factorial ANOVA model. In order to account for multiple models, the following analysis focuses primarily on Yeezy sneakers, which mostly differ in colorways only with fixed retail price. We are interested in modeling the response variable (profit) as a function of two variables (weeks after release date and shoe

size). Particularly,we are interested in whether there is a statistically significant difference in mean profits among different shoe sizes.

```
yeezy_sneakers_profits <- sneakers_profits %>%
    filter(Brand == "Yeezy")
```

After fitting the first regression model, the ANOVA assumptions were checked before proceeding to the analysis. The independence of cases condition was satisfied according to the description of how the data was collected. One of the testing assumptions includes that the population from which samples are drawn should be normally distributed. This assumptions was checked by displaying diagnostic plots. In the ideal case, residual points should follow the straight dashed line. However, this doesn't apply to our data. The Normal Q-Q plot shows that the upper end deviates from the straight line and the lower end follows the straight line, which indicates a positive skew. Since the normality condition is obviously violated, the data needs to be transformed to look more normally distributed.

```
# Create a main effects only model
fit_main_eff_yeezy <- lm(profit_percentage ~ weeks_after_release + `Shoe Size` , data = yeezy_sneakers_
```



In order to take a long transform of the response variable, only the data that resulted in positive profit was kept. The transformation technique was chosen to be the Box-Cox transformation. The basic idea behind this method is to find some value for λ such that the transformed data is as close to the normally distributed as possible, using the following formula.

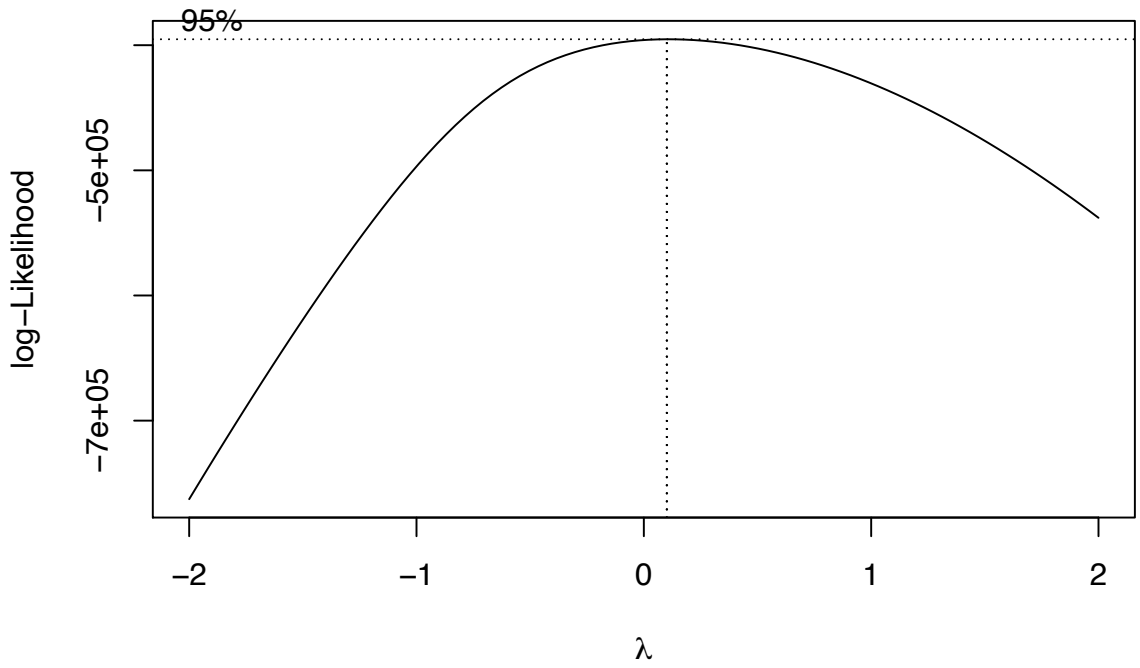
$y(\lambda) = \frac{y^{\lambda}-1}{\lambda}$, if $\lambda \neq 0$ $y(\lambda) = log(y)$, $\lambda = 0$

A box-cox transformation was performed by utilizing **boxcox()** function from the *MASS* library.

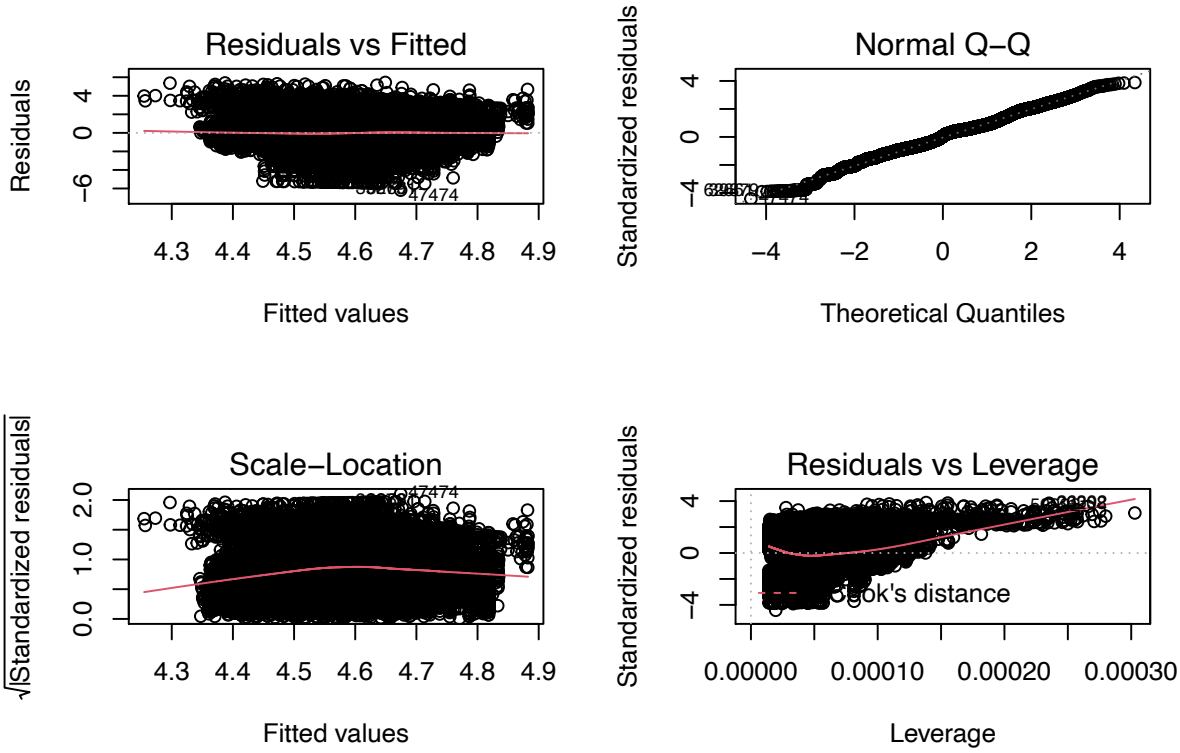
```
# leave only data that results in positive profit
yeezy_boxcox <- yeezy_sneakers_profits %>% filter(profit_percentage > 0)
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```
#find optimal lambda for Box-Cox transformation
bc <- boxcox(profit_percentage ~ weeks_after_release + `Shoe Size` , data = yeezy_boxcox)
```




```
# chooose lambda value that maximizes log likelihood
lambda <- bc$x[which.max(bc$y)]

# create a new main effects only model
new_model <- lm(((profit_percentage^lambda-1)/lambda) ~ weeks_after_release + `Shoe Size` , data = yeezy
# look at diagnostic plots
par(mfrow = c(2, 2))
plot(new_model)
```



A new set of diagnostic plots shows the improvement in several ways. The Residual vs Fitted plot displays a horizontal line, without distinct patterns, which is an indication for a linear relationship. The Normal Q-Q plot now shows all data points following a straight line, which points to the normal distribution of the residuals. The homogeneity of residual variance (homoscedascity) accounted for by the examination of the *scale-location* plot. In ideal case, there should be a horizontal line with equally spread residuals along the range of predictors. The deviation from this ideal scenario points to the possibility of non-constant variances in the residual errors (heteroscedasticity). Nevertheless, the ANOVA analysis was still conducted, with the key insights being additionally checked by the Kruskal-Wallis test, a non-parametric equivalent of the ANOVA parametric test.

```
# check for main effects of the weeks after release and shoe size
anova(new_model)
```

```
## Analysis of Variance Table
##
## Response: ((profit_percentage^lambda - 1)/lambda)
##                         Df Sum Sq Mean Sq F value    Pr(>F)
## weeks_after_release      1    173  173.48  88.498 < 2.2e-16 ***
## `Shoe Size`              1    423  423.14 215.863 < 2.2e-16 ***
## Residuals            71273 139713    1.96
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Kruskal-Wallis test
kruskal.test(profit_percentage ~ `Shoe Size`, data = yeezy_sneakers_profits)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  profit_percentage by Shoe Size
## Kruskal-Wallis chi-squared = 823.22, df = 24, p-value < 2.2e-16
```

The ANOVA analysis confirmed that not only does the time elapsed since the release of sneakers, but also the size plays a significant role in their evaluation in the future. Both Main Effects were showed to be statistically significant. A significant Kruskal-Wallis test indicates that at least one sample (differentiated by a shoe size) dominates the other sample. Kruskal-Wallis backs up the previous conclusion since it has no assumptions, unlike the ANOVA test.

```
new_model2 <- lm(((profit_percentage^lambda-1)/lambda) ~ weeks_after_release * `Shoe Size` , data = yee
# type III sum of squares the order that variables are added does not matter
car::Anova(new_model2, type = "III")
```

```
## Anova Table (Type III tests)
##
## Response: ((profit_percentage^lambda - 1)/lambda)
##                                 Sum Sq    Df    F value    Pr(>F)
## (Intercept)                      47843     1 24413.2887 < 2.2e-16 ***
## weeks_after_release                  8     1     3.8696   0.04917 *
## `Shoe Size`                        391     1   199.3270 < 2.2e-16 ***
## weeks_after_release:`Shoe Size`     41     1    20.7976 5.113e-06 ***
## Residuals                       139672 71272
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The interaction effect was then added to the analysis. Type three sum of squares was utilized to compute sum of squares for the p-value due to the constraints of the unbalanced design. In this case, the order that that the terms were entered into the model didn't matter. Besides previous conclusion, the new insight showed the interaction effect to be statistically significant. This notable observation shows more pairs of bigger shoe sizes were held on to for the longer period of times, which resulted in the inflated sale prices.

In order to gain additional insights into what sizes of sneakers resulted in more profitable transactions, the data was grouped by the sizes, and the average profit (in % relation to the retail price) was plotted for two most profitable periods defined above (7-73 weeks and >103 weeks). Based on the plot, we can make an interesting observation. Average she sizes that should be sold more often do not bring high profits. It seems that rare sizes (below and above average male sizes in the US) tend to result in slightly higher profits due to their limited supply (sizes 5.5-6.5 & >13). It should be noted that both ends of the trend lines might not be representative of average profit due to the little number of cases for these sizes present in the data set.

Conclusion

The analysis was able to give answers to multiple question posed at the beginning of the project. Two states, California and New York, were identified as states where the sneaker industry thrived the most. This conclusion was pointed to out to have a potential relation to the comparably high average income, which gives direction for further exploration. The bestselling time for premium collectible sneakers (in our case, Off-White x Nike collaborations and Yeezys only) was shown to be two years after the release date (still high demand with shortage of supply on the market). It was shown by the visualization that it is in the

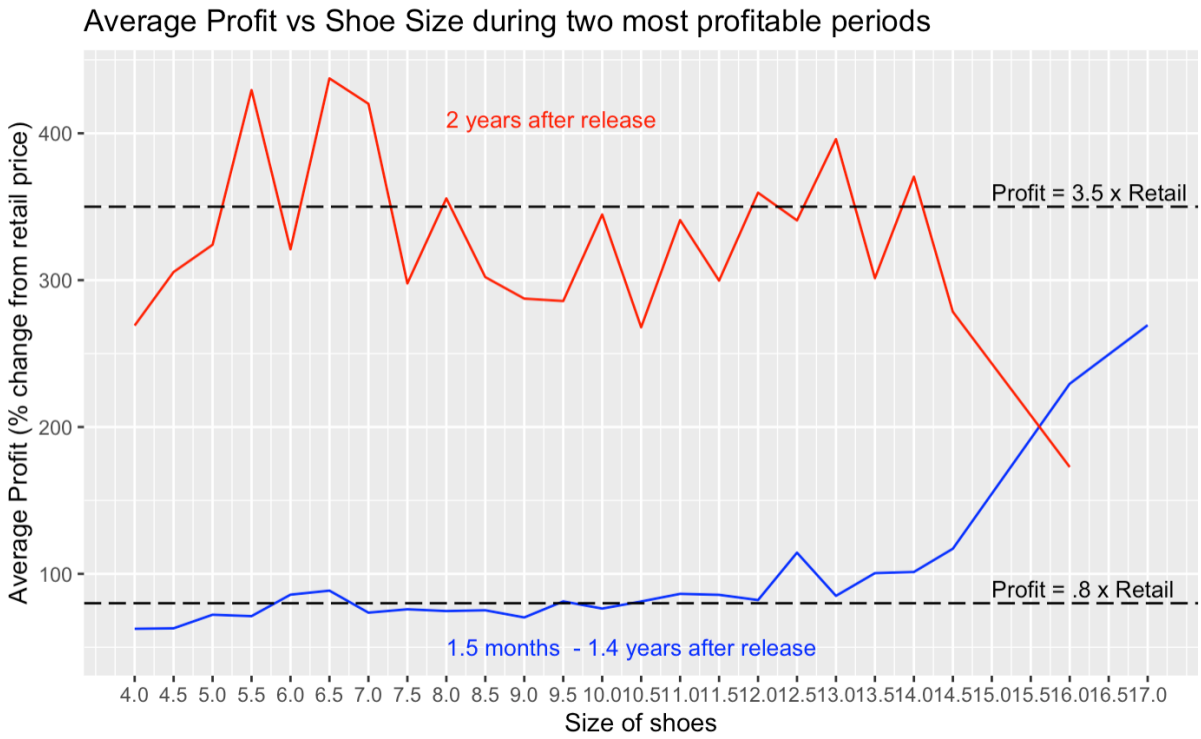


Figure 1: Profit vs Shoe Size

best interest to not re-sell sneakers a first couple of months after their release date, as a potential seller has a risk of undervaluing a pair. According to the visualization, there also exists a trend in time (1.5 - 2 years) where sneakers loose their popularity with a drop in average profits. Taking these insights into account may help potential sellers to correctly assess the price of collectible sneakers to not miss out on any value. The analysis showed that mean profits differ significantly for different sizes of shoes. It was determined that bigger shoe sizes were held on to more often with the goal to re-sell them in the future. Nevertheless, it was shown that not only larger sizes (>13) brought in the most money, but also small sizes (5.5-6.5) were slightly more profitable due to their limited supply.

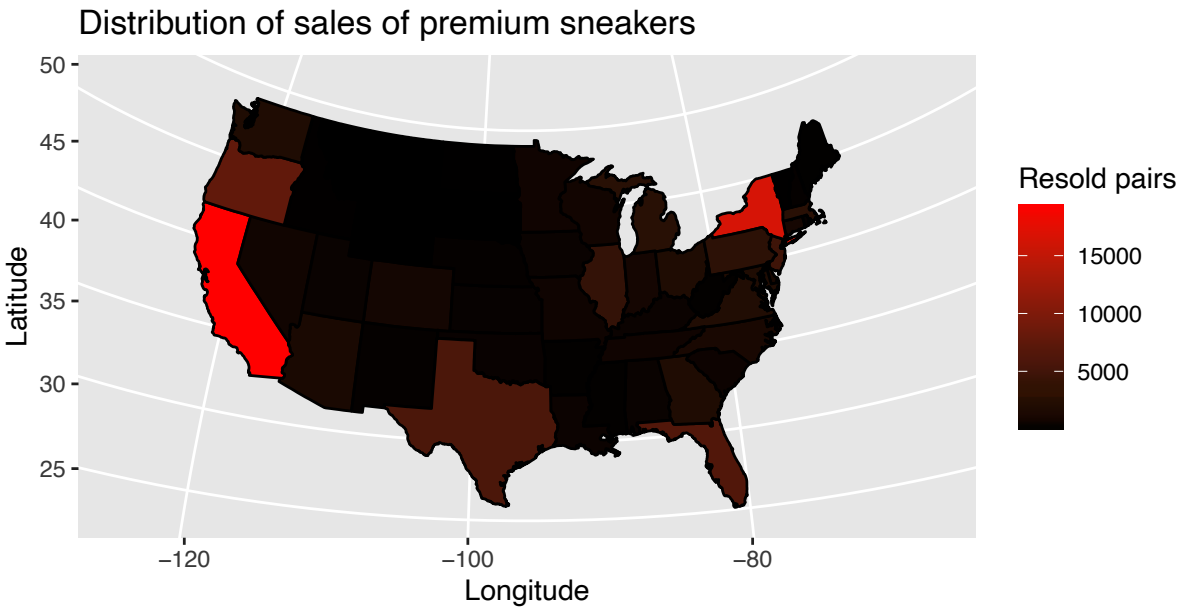
Reflection

The knowledge of dplyr and ggplot tools came very handy for creating visualizations to answer questions of interest, making the process of making inferences pretty straightforward. The analysis part of the project turned out to be very confusing, since there are a lot of groups for each of the variables. Fortunately, I was able to resolve all the issues myself with results turning out to be very insightful.

Additional analysis: Polynomial regression analysis was performed to analyze the relationship between the time elapsed since the release date and the average profit modeled as an n-th degree polynomial of Yeezy sneakers. An interesting observation was that the regression of third degree, which represented the general trend for most premium sneakers but didn't overfit to the particular trends of Yeezys, reminded of the Nike swoosh logo. Also, different transformations were tried on the data with almost no improvement in distribution of the data (various log transformations).

Appendix

```r
# get summary on total number of sneakers re-sold
sneakers_per_state <- sneakers_profits %>% group_by(`Buyer Region`) %>% summarize(n())
names(sneakers_per_state)[2] <- 'Resold pairs'

# get the map of states
states_map <- map_data("state")

# join sneakers_per_state and states_map data frames
states_map <- map_data("state") %>%
  left_join(sneakers_per_state, by = c("region" = "Buyer Region"))


# order the data
states_map <- arrange(states_map, group, order)

# create cholorpleth map
ggplot(states_map,  aes(x = long, y = lat, group = group,  fill = `Resold pairs`)) +
  ggtitle("Distribution of sales of premium sneakers") +
  xlab("Longitude") +
  ylab("Latitude") +
    geom_polygon(color = "black") +
    coord_map("polyconic") +
  scale_fill_gradient(low = "black", high = "red")
```



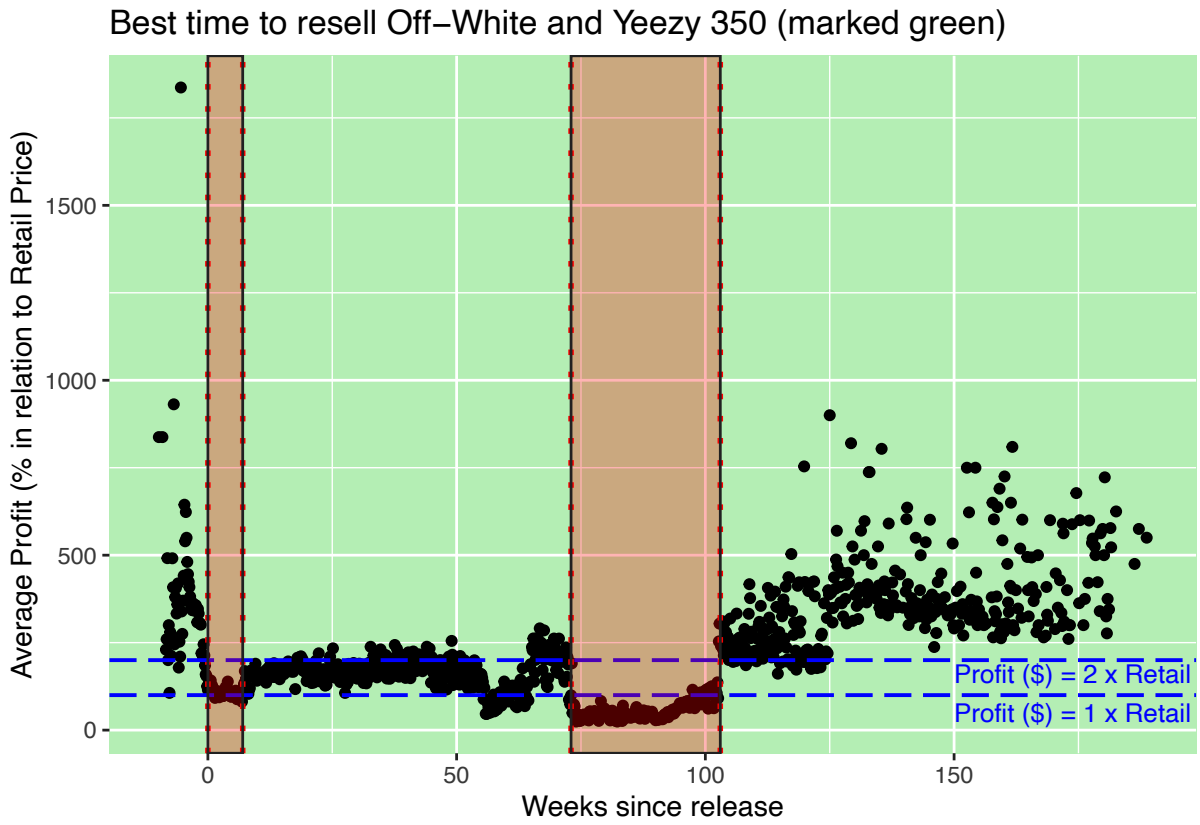

```
#output top five states
states_map %>%
  arrange(desc(`Resold pairs`)) %>%
  distinct(region,`Resold pairs`) %>%
  head()
```

```
##       region Resold pairs
## 1 california        19349
## 2   new york        16525
## 3     oregon         7681
## 4    florida         6376
## 5      texas         5876
## 6 new jersey         4720
```

```
# rectangular areas to mark worst times to sell
rect1 <- data.frame(xmin = 0, xmax = 7, ymin=-Inf, ymax=Inf)
rect2 <- data.frame(xmin = 73, xmax = 103, ymin=-Inf, ymax=Inf)

# data set with average profits across different timings
avg_profits <- sneakers_profits %>% group_by(time_elapsed) %>%
  summarize(avg_premium_profit <- sum(profit_percentage)/n())
# change the name of variable for better clarity
names(avg_profits)[2] <- 'avg_profit'
# graph for optimal re-selling time
plot <- ggplot(avg_profits, aes(time_elapsed,avg_profit)) + geom_point() +
```

```
  geom_vline(xintercept = 7, linetype = "dotted", color = "red", size = 1) +
  geom_vline(xintercept = 0, linetype = "dotted", color = "red", size = 1) +
  geom_vline(xintercept = 73, linetype = "dotted", color = "red", size = 1) +
  geom_vline(xintercept = 103, linetype = "dotted", color = "red", size = 1) +
  geom_hline(yintercept = 100, linetype = "longdash", color = "blue", size = .7) +
  geom_hline(yintercept = 200, linetype = "longdash", color = "blue", size = .7)
plot +
  geom_rect(data = rect1,aes(xmin=xmin, xmax=xmax, ymin=ymin, ymax=ymax),
              color="grey20",
              alpha=0.3,
              fill = "red",
              inherit.aes = FALSE) +
  geom_rect(data = rect2,aes(xmin=xmin, xmax=xmax, ymin=ymin, ymax=ymax),
              color="grey20",
              alpha=0.3,
              fill = "red",
              inherit.aes = FALSE) +
  theme(panel.background = element_rect(fill = "darkseagreen2", color = "darkseagreen2", size = 0.5, li
  annotate(geom = "text", x = 150, y = 50,
           label = "Profit ($) = 1 x Retail", hjust = 0, col = "blue", size = 3.5) +
  annotate(geom = "text", x = 150, y = 160,
           label = "Profit ($) = 2 x Retail", hjust = 0, col = "blue", size = 3.5) +
  ggtitle("Best time to resell Off-White and Yeezy 350 (marked green)") +
  xlab("Weeks since release") +
  ylab("Average Profit (% in relation to Retail Price)")
```

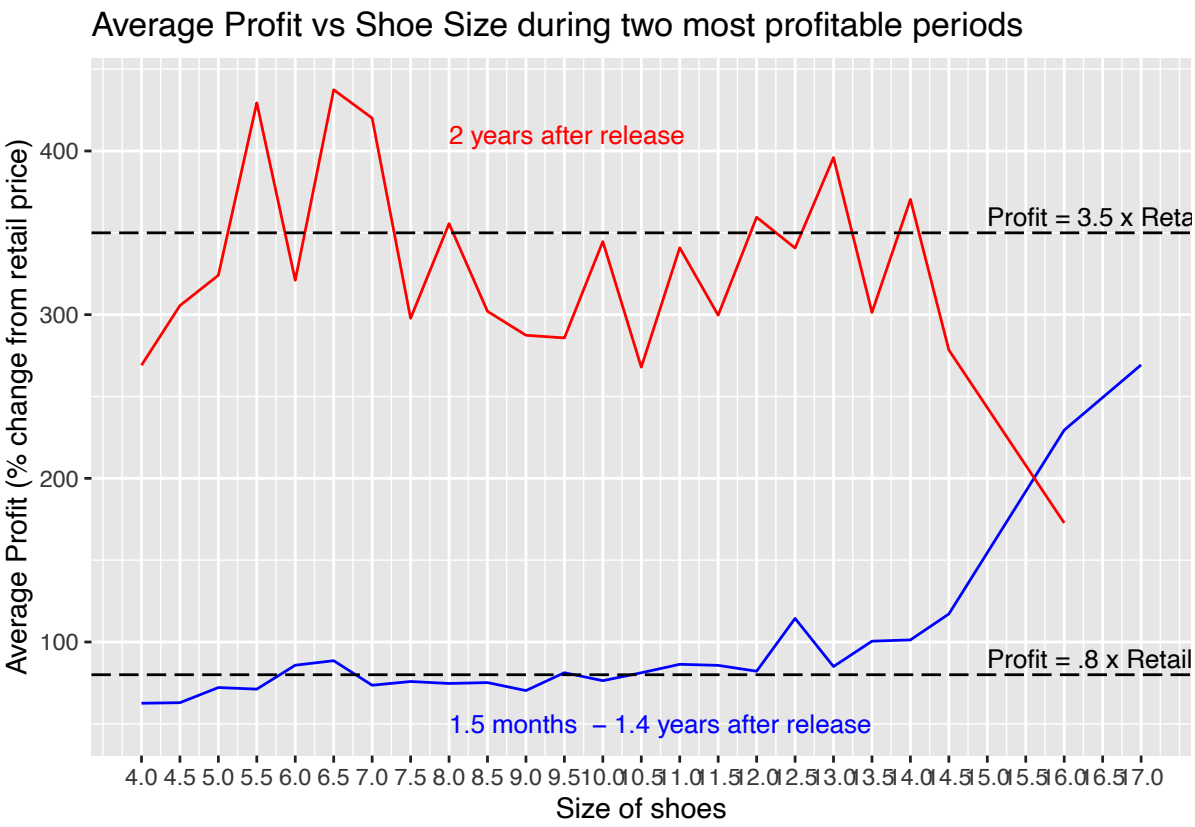
```
## Don't know how to automatically pick scale for object of type difftime. Defaulting to continuous.
```



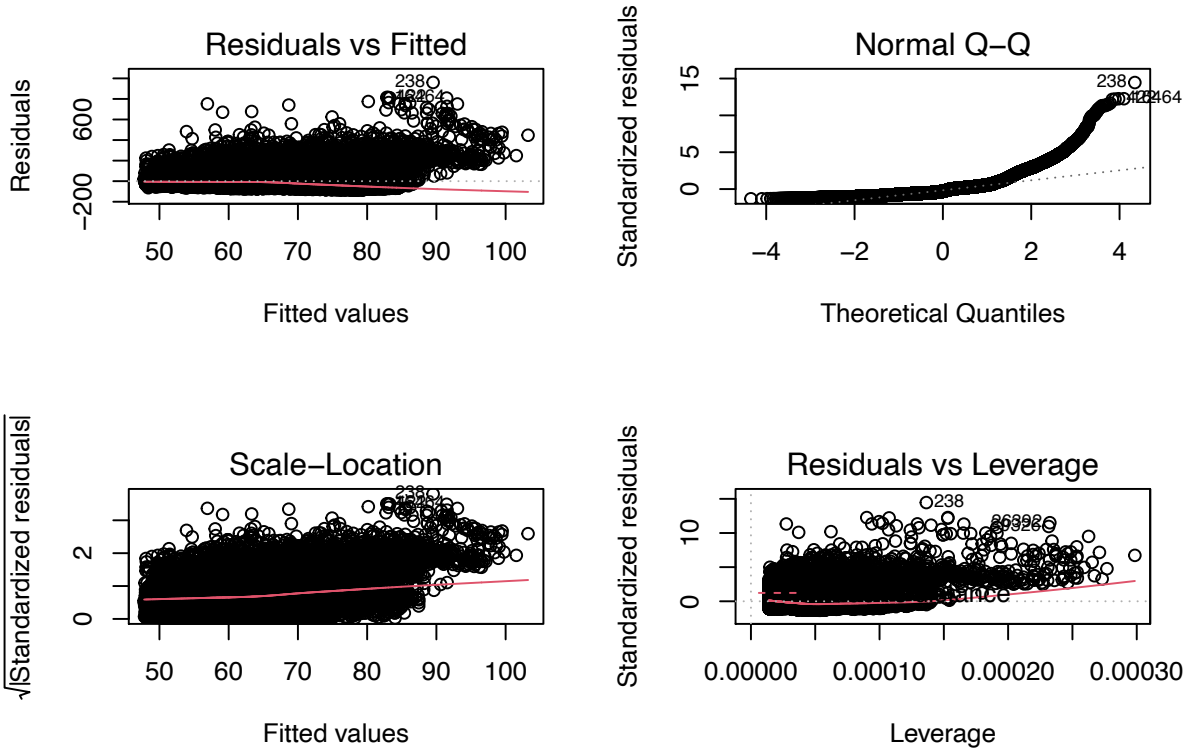

```
# average profit for sales between 1.5 months
# and 1.4 years after the release date
yeezy_shoesizes_weeks_7_73 <- yeezy_sneakers_profits %>%
  filter(weeks_after_release > 7, weeks_after_release < 73 ) %>%
  group_by(`Shoe Size`) %>%
  summarize(avg_profit = sum(profit_percentage)/n())
# average profit for sales 2 years after release
yeezy_shoesizes_weeks_greater103 <- yeezy_sneakers_profits %>%
  filter(weeks_after_release > 103) %>%
  group_by(`Shoe Size`) %>%
  summarize(avg_profit = sum(profit_percentage)/n())
```

```
# plot average profit over shoe sizes
ggplot() +
  geom_line(data = yeezy_shoesizes_weeks_7_73,
            aes(`Shoe Size`, avg_profit), color = "blue") +
  geom_line(data = yeezy_shoesizes_weeks_greater103,
            aes(`Shoe Size`, avg_profit), color = "red") +
  scale_x_continuous(breaks = round(seq(min(yeezy_shoesizes_weeks_7_73$`Shoe Size`),
                                        max(yeezy_shoesizes_weeks_7_73$`Shoe Size`),
                                        by = 0.5),1)) +
  ggtitle("Average Profit vs Shoe Size during two most profitable periods") +
  xlab("Size of shoes") +
  ylab("Average Profit (% change from retail price)") +
  annotate(geom = "text", x = 8, y = 50,
           label = "1.5 months  - 1.4 years after release", hjust = 0, col = "blue", size = 3.5) +
```

```
  annotate(geom = "text", x = 8, y = 410,
           label = "2 years after release", hjust = 0, col = "red", size = 3.5) +
  geom_hline(yintercept = 80, linetype = "longdash", color = "black", size = .5) +
  geom_hline(yintercept = 350, linetype = "longdash", color = "black", size = .5) +
  annotate(geom = "text", x = 15, y = 90,
           label = "Profit = .8 x Retail", hjust = 0, col = "black", size = 3.5) +
  annotate(geom = "text", x = 15, y = 360,
           label = "Profit = 3.5 x Retail", hjust = 0, col = "black", size = 3.5)
```




```
# diagnostic plot for initial model that  violates assumptions
par(mfrow = c(2, 2))
plot(fit_main_eff_yeezy)
```


Residuals vs Fitted
Residuals
Fitted values
Normal Q–Q
Standardized residuals
Theoretical Quantiles
Scale–Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Leverage
238