StockX Data Analysis

Ben Annor-Adjaye

11/4/2021

Setting up my environment

Notes: setting up my R environment by loading the 'tidyverse' for data modelling and visualization,'here' for file referencing, 'skimr' and 'janitor' for data cleaning, 'dplyr' for data manipulation.

```
install.packages('tidyverse')
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.1'
## (as 'lib' is unspecified)
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
install.packages("here")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.1'
## (as 'lib' is unspecified)
```

```
library("here")
```

```
## here() starts at /cloud/project
```

```
install.packages("skimr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.1'
## (as 'lib' is unspecified)
```

```
library("skimr")

install.packages("janitor")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.1'
## (as 'lib' is unspecified)
```

```
library("janitor")
```

```
##
## Attaching package: 'janitor'
```

```
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
install.packages("dplyr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.1'
## (as 'lib' is unspecified)
```

```
library("dplyr")

install.packages("ggpubr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.1'
## (as 'lib' is unspecified)
```

```
library("ggpubr")
```

Summary of dataset

```
StockX <-read.csv("Sneaker2.0.csv", header = TRUE, sep = ",")

#Fix a typo in a column name
StockX <- StockX%>% rename(retail_price_usd=retai_price_usd)

#Adding column to categorize the buyer_region (i.e. the states where resale orders where made) into the
StockX <- StockX %>% mutate(US_Region = case_when(buyer_region == "California" ~ "West",buyer_region ==

#Adding column to categorize the sneaker_names into the silhouettes
StockX <- StockX %>% mutate(Silhouette = case_when(grepl("350",sneaker_name) ~ "Yeezy 350",grepl("Force

#Adding column for the quarterly cycle
StockX <- StockX %>% mutate(Quarter = case_when(grepl("Q1",order_period..q.yyyy.) ~ "Q1",grepl("Q2",ord

#Checking the data types
str(StockX)
```

```
## 'data.frame':    99956 obs. of  14 variables:
##  $ X                    : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ order_date           : chr  "2017-09-01" "2017-09-01" "2017-09-01" "2017-09-01" ...
##  $ order_period..q.yyyy.: chr  "Q3 2017" "Q3 2017" "Q3 2017" "Q3 2017" ...
##  $ brand                : chr  "Adidas" "Adidas" "Adidas" "Adidas" ...
##  $ designer             : chr  "Yeezy" "Yeezy" "Yeezy" "Yeezy" ...
##  $ sneaker_name         : chr  "Adidas-Yeezy-Boost-350-Low-V2-Beluga" "Adidas-Yeezy-Boost-350-V2-Cor
##  $ resale_price_usd     : int  1097 685 690 1075 828 798 784 460 465 465 ...
##  $ retail_price_usd     : int  220 220 220 220 220 220 220 220 220 220 ...
##  $ release_date         : chr  "2016-09-24" "2016-11-23" "2016-11-23" "2016-11-23" ...
##  $ shoe_size            : num  11 11 11 11.5 11 8.5 11 10 11 11 ...
##  $ buyer_region         : chr  "California" "California" "California" "Kentucky" ...
##  $ US_Region            : chr  "West" "West" "West" "South" ...
##  $ Silhouette           : chr  "Yeezy 350" "Yeezy 350" "Yeezy 350" "Yeezy 350" ...
##  $ Quarter              : chr  "Q3" "Q3" "Q3" "Q3" ...
```

```
#Summary of the data set and data structure
summary(StockX)
```

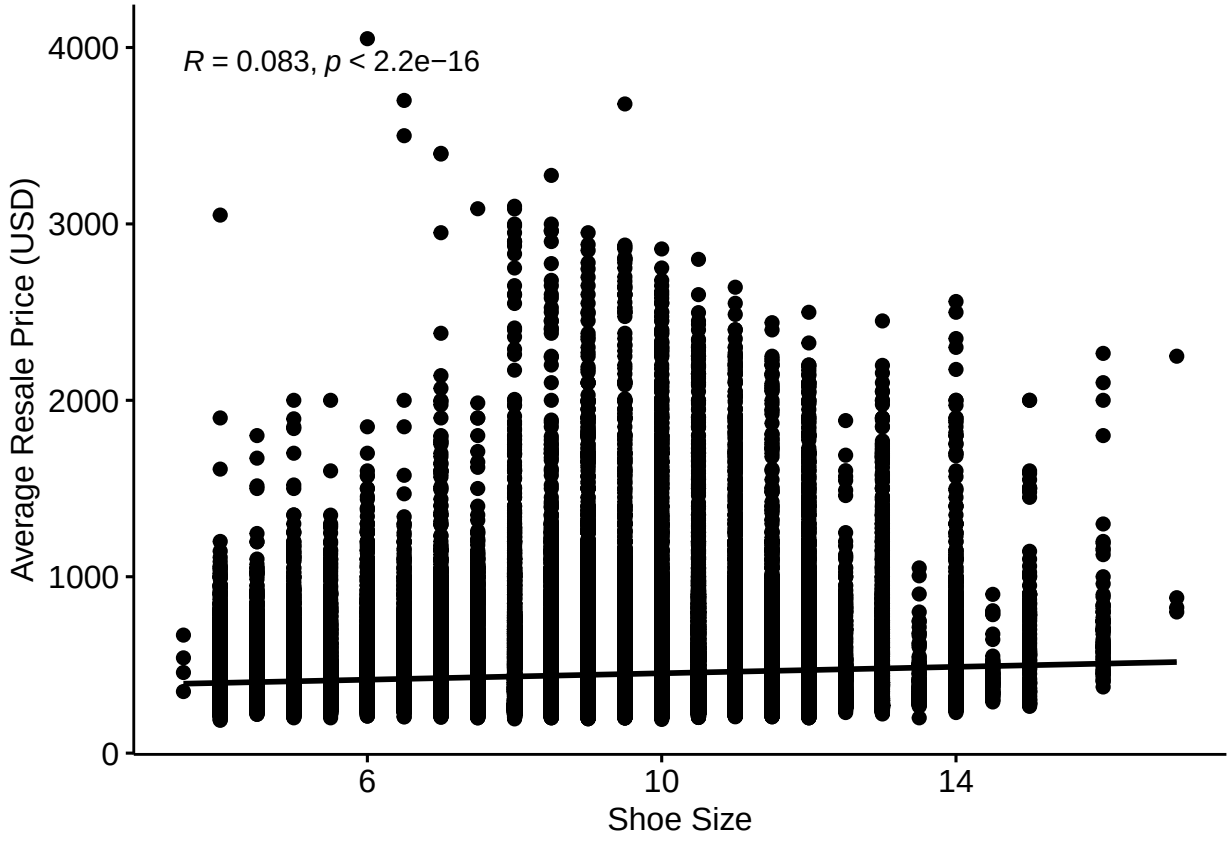
```
##        X          order_date        order_period..q.yyyy.    brand
##  Min.   :    1   Length:99956       Length:99956          Length:99956
##  1st Qu.:24990   Class :character   Class :character      Class :character
##  Median :49978   Mode  :character   Mode  :character      Mode  :character
##  Mean   :49978
##  3rd Qu.:74967
##  Max.   :99956
##    designer         sneaker_name       resale_price_usd retail_price_usd
##  Length:99956       Length:99956       Min.   : 186.0   Min.   :130.0
##  Class :character   Class :character   1st Qu.: 275.0   1st Qu.:220.0
##  Mode  :character   Mode  :character   Median : 370.0   Median :220.0
##                                        Mean   : 446.6   Mean   :208.6
##                                        3rd Qu.: 540.0   3rd Qu.:220.0
##                                        Max.   :4050.0   Max.   :250.0
##  release_date         shoe_size      buyer_region        US_Region
##  Length:99956       Min.   : 3.500   Length:99956       Length:99956
##  Class :character   1st Qu.: 8.000   Class :character   Class :character
##  Mode  :character   Median : 9.500   Mode  :character   Mode  :character
##                     Mean   : 9.344
##                     3rd Qu.:11.000
##                     Max.   :17.000
##   Silhouette          Quarter
##  Length:99956       Length:99956
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

Regression Analysis

Note:to determine which attributes influence the average resale price of a sneaker

```
#Checking the relationship between shoe size and resale price
ggscatter(StockX, x = "shoe_size", y = "resale_price_usd", add = "reg.line", conf.int = TRUE, cor.coef =
```

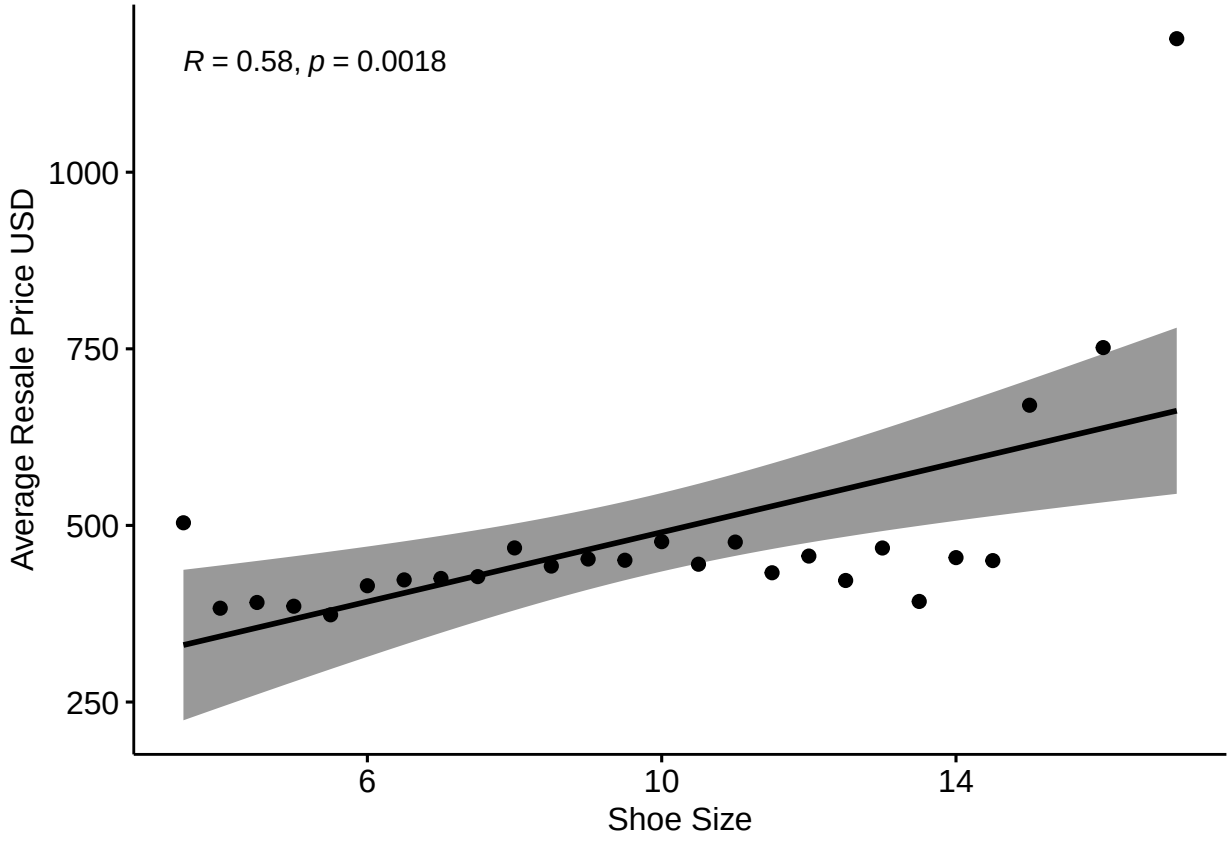
```
## `geom_smooth()` using formula 'y ~ x'
```




```
#Creating a data frame to check the relationship between shoe size and average resale price
df1 <- group_by(StockX, shoe_size)
df2 <- summarise(df1, average_resale_price = mean(resale_price_usd))

#Checking the relationship between shoe size and resale price
ggscatter(df2, x = "shoe_size", y = "average_resale_price", add = "reg.line", conf.int = TRUE, cor.coef
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
#Converting the attributes "brand", "designer", "sneaker_name", "order_period" and "buyer_region" to a
StockX$brand= as.factor(StockX$brand)
StockX$order_period..q.yyyy.= as.factor(StockX$order_period..q.yyyy.)
StockX$designer= as.factor(StockX$designer)
StockX$sneaker_name= as.factor(StockX$sneaker_name)
StockX$buyer_region= as.factor(StockX$buyer_region)
StockX$US_Region= as.factor(StockX$US_Region)
StockX$Silhouette= as.factor(StockX$Silhouette)
StockX$shoe_size= as.factor(StockX$shoe_size)

#Updated data set structure
str(StockX)
```

```
## 'data.frame':    99956 obs. of  14 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ order_date          : chr  "2017-09-01" "2017-09-01" "2017-09-01" "2017-09-01" ...
##  $ order_period..q.yyyy.: Factor w/ 7 levels "Q1 2018","Q1 2019",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ brand               : Factor w/ 2 levels "Adidas","Nike": 1 1 1 1 1 1 1 1 1 1 ...
##  $ designer            : Factor w/ 2 levels "Off-White","Yeezy": 2 2 2 2 2 2 2 2 2 2 ...
##  $ sneaker_name        : Factor w/ 50 levels "Adidas-Yeezy-Boost-350-Low-Moonrock",..: 6 10 11 12 13
##  $ resale_price_usd    : int  1097 685 690 1075 828 798 784 460 465 465 ...
##  $ retail_price_usd    : int  220 220 220 220 220 220 220 220 220 220 ...
##  $ release_date        : chr  "2016-09-24" "2016-11-23" "2016-11-23" "2016-11-23" ...
##  $ shoe_size           : Factor w/ 26 levels "3.5","4","4.5",..: 16 16 16 17 16 11 16 14 16 16 ...
##  $ buyer_region        : Factor w/ 51 levels "Alabama","Alaska",..: 5 5 5 18 40 23 5 33 17 10 ...
##  $ US_Region           : Factor w/ 4 levels "Midwest","Northeast",..: 4 4 4 3 2 1 4 2 1 3 ...
##  $ Silhouette          : Factor w/ 10 levels "Air_Force_1_Low",..: 9 9 9 9 9 9 9 9 9 9 ...
```

```
##  $ Quarter              : chr  "Q3" "Q3" "Q3" "Q3" ...
```

```
#Regression model
reg_output1=lm(resale_price_usd~shoe_size+Quarter+US_Region+brand+Silhouette,data = StockX)
summary(reg_output1)
```

```
##
## Call:
## lm(formula = resale_price_usd ~ shoe_size + Quarter + US_Region +
##     brand + Silhouette, data = StockX)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -598.89  -89.96  -38.90   35.96 3038.52
##
## Coefficients: (1 not defined because of singularities)
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                283.144     88.665   3.193  0.00141 **
## shoe_size4                  41.284     88.721   0.465  0.64170
## shoe_size4.5                43.538     88.780   0.490  0.62385
## shoe_size5                  66.860     88.692   0.754  0.45095
## shoe_size5.5                70.421     88.711   0.794  0.42730
## shoe_size6                  71.159     88.687   0.802  0.42235
## shoe_size6.5                60.648     88.723   0.684  0.49425
## shoe_size7                  59.583     88.689   0.672  0.50170
## shoe_size7.5                74.863     88.710   0.844  0.39872
## shoe_size8                 100.498     88.676   1.133  0.25708
## shoe_size8.5               104.789     88.675   1.182  0.23732
## shoe_size9                 101.351     88.661   1.143  0.25298
## shoe_size9.5               108.585     88.662   1.225  0.22069
## shoe_size10                105.481     88.658   1.190  0.23414
## shoe_size10.5               93.781     88.662   1.058  0.29018
## shoe_size11                100.310     88.662   1.131  0.25790
## shoe_size11.5               98.641     88.682   1.112  0.26602
## shoe_size12                 93.745     88.668   1.057  0.29040
## shoe_size12.5              130.958     88.928   1.473  0.14085
## shoe_size13                 96.929     88.681   1.093  0.27440
## shoe_size13.5              129.404     89.844   1.440  0.14978
## shoe_size14                110.789     88.745   1.248  0.21189
## shoe_size14.5              189.338     90.732   2.087  0.03691 *
## shoe_size15                168.724     89.998   1.875  0.06083 .
## shoe_size16                464.995     90.863   5.118 3.10e-07 ***
## shoe_size17                754.737    125.342   6.021 1.73e-09 ***
## QuarterQ2                    1.352      2.129   0.635  0.52537
## QuarterQ3                  -39.841      1.797 -22.171  < 2e-16 ***
## QuarterQ4                  -38.673      1.431 -27.029  < 2e-16 ***
## US_RegionNortheast           8.406      1.884   4.462 8.11e-06 ***
## US_RegionSouth               4.704      1.937   2.429  0.01513 *
## US_RegionWest               17.408      1.858   9.371  < 2e-16 ***
## brandNike                  142.502      3.637  39.179  < 2e-16 ***
## SilhouetteAir_Max_90        70.118      5.366  13.067  < 2e-16 ***
## SilhouetteAir_Max_97       219.193      5.958  36.790  < 2e-16 ***
## SilhouetteAir_Presto       264.077      4.586  57.589  < 2e-16 ***
## SilhouetteBlazer_Mid       106.633      4.635  23.007  < 2e-16 ***
## SilhouetteJordan_1_High    513.326      4.384 117.102  < 2e-16 ***
```

```
## SilhouetteReact          -16.396      8.822  -1.859  0.06308 .
## SilhouetteVapor_Max      130.078      4.787  27.171  < 2e-16 ***
## SilhouetteYeezy 350           NA         NA      NA       NA
## SilhouetteZoom_Fly      -178.377      4.486 -39.762  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 177.2 on 99478 degrees of freedom
##   (437 observations deleted due to missingness)
## Multiple R-squared:  0.5207, Adjusted R-squared:  0.5205
## F-statistic:  2702 on 40 and 99478 DF,  p-value: < 2.2e-16
```

```
#Revised regression model
reg_output2=lm(resale_price_usd~Quarter+US_Region+brand+Silhouette,data = StockX)
summary(reg_output2)
```

```
##
## Call:
## lm(formula = resale_price_usd ~ Quarter + US_Region + brand +
##     Silhouette, data = StockX)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -620.58  -89.39  -43.53   36.37 3013.53
##
## Coefficients: (1 not defined because of singularities)
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              377.457      1.890 199.763  < 2e-16 ***
## QuarterQ2                  3.589      2.141   1.676   0.0937 .
## QuarterQ3                -39.837      1.809 -22.024  < 2e-16 ***
## QuarterQ4                -37.827      1.439 -26.294  < 2e-16 ***
## US_RegionNortheast         5.078      1.890   2.687   0.0072 **
## US_RegionSouth             3.896      1.949   1.999   0.0456 *
## US_RegionWest             14.115      1.858   7.598 3.03e-14 ***
## brandNike                145.016      3.657  39.650  < 2e-16 ***
## SilhouetteAir_Max_90      73.246      5.399  13.565  < 2e-16 ***
## SilhouetteAir_Max_97     221.825      5.997  36.991  < 2e-16 ***
## SilhouetteAir_Presto     262.996      4.585  57.362  < 2e-16 ***
## SilhouetteBlazer_Mid     105.991      4.666  22.715  < 2e-16 ***
## SilhouetteJordan_1_High  510.406      4.409 115.774  < 2e-16 ***
## SilhouetteReact          -11.042      8.877  -1.244   0.2136
## SilhouetteVapor_Max      128.376      4.818  26.645  < 2e-16 ***
## SilhouetteYeezy 350           NA         NA      NA       NA
## SilhouetteZoom_Fly      -178.435      4.516 -39.515  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 178.4 on 99503 degrees of freedom
##   (437 observations deleted due to missingness)
## Multiple R-squared:  0.5139, Adjusted R-squared:  0.5139
## F-statistic:  7014 on 15 and 99503 DF,  p-value: < 2.2e-16
```

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this: