Megan Fantes

SQL Assignment

November 25, 2016

First, we must connect to the Postgres database:

```
library(RPostgreSQL)

## Set up the connection parameters you will need
host <- "analyticsga-east2.c20gkj5cvu3l.us-east-1.rds.amazonaws.com"
port <- "5432"
username <- "analytics_student"
password <- "analyticsga"

## Use the name of the specific database you will access
dbname <- "iowa_liquor_sales_database"

## Specify the PostreSQL driver
drv <- dbDriver("PostgreSQL")

## Now establish the connection
con <- dbConnect(drv, user = username, password = password,
dbname = dbname, port = port, host = host)
```

Now we make sure that we have established the proper connection with the database:

```
## get a list of the tables in the database
dbListTables(con)
```

```
## [1] "products" "stores"   "counties" "sales"
```

```
## get a list of the variable names in the "products" table
dbListFields(con, "products")
```

```
##  [1] "item_no"          "category_name"    "item_description"
##  [4] "vendor"           "vendor_name"      "bottle_size"
##  [7] "pack"             "inner_pack"       "age"
## [10] "proof"            "list_date"        "upc"
## [13] "scc"              "bottle_price"     "shelf_price"
## [16] "case_cost"
```

We can test the RpostgresSQL package by recreating the commands we did in class:

Which items comes in packs larger than 12?

```
query <- dbGetQuery(con, statement = paste(
  "SELECT DISTINCT item_description, pack",
  "FROM products",
  "WHERE pack > 12",
  "ORDER BY pack"))

## We use the SELECT DISTINCT command so we only get unique entries in our
## final table. We do not care if one product is entered 10 times in the
## original table, we only want it to appear once in our final table

head(query)
```

```
##                       item_description pack
## 1                         Esrum Kloster   15
## 2           Jeanne D'arc Belzebuth Ale   15
## 3          Bourbon Heritage Multi-Pack   15
## 4                              Draupnir   15
## 5               Pripp's Carnegie Porter   15
## 6 Porfidio Tripl-Distilled Plata Minis   15
```

Which items have a case cost of less than $70?

```
query <- dbGetQuery(con, statement = paste(
  "SELECT DISTINCT item_description, case_cost",
  "FROM products",
  "WHERE case_cost < 70",
  "ORDER BY case_cost"))

head(query)
```

```
##                   item_description case_cost
## 1                  Littlemill 12yr      0.00
## 2        Ypioca Cachaca Ouro(gold)      0.00
## 3              Hammer Sickle Vodka      0.00
## 4                       Cachaca 21      7.50
## 5 Hiram Walker Peach 3pak Schnapps     10.25
## 6                 Gaetano Kamikaze     11.50
```

Which items come is packs larger than 12 AND have a case cost of less than $70?

```
query <- dbGetQuery(con, statement = paste(
  "SELECT DISTINCT item_description, pack, case_cost",
  "FROM products",
  "WHERE pack > 12 AND case_cost < 70",
  "ORDER BY case_cost, pack"))

head(query)
```

```
##                     item_description pack case_cost
## 1                   Gaetano Kamikaze   24     11.50
## 2             Neuzeller Golden Abbot   20     11.70
## 3 Neuzeller Porter Black Strong Case   20     11.70
## 4           Samuel Adams Double Bock   24     17.00
## 5                 Kulmbacher Pilsner   24     18.50
## 6               North Coast Acme Ipa   24     19.16
```

Which items have a proof of 85 or more?

```
query <- dbGetQuery(con, statement = paste(
  "SELECT DISTINCT item_description, proof",
  "FROM products",
  "WHERE cast(proof as integer) >= 85",
  "ORDER BY proof"))

## NOTICE: we needed to CAST the variable proof as an INTEGER in order to
## manipulate it like a number, because in the original table it is TEXT

head(query)
```

```
##                     item_description proof
## 1             E.H. Taylor Tornado HA   100
## 2           Southern Comfort 100 Prf   100
## 3            Midnight Moon Cranberry   100
## 4 J.w. Dant 100prf Bond Bourbon 54mo   100
## 5        Firewater Cinnamon Schnapps   100
## 6     John J Bowman Str. Bbn.Why. HA   100
```

Which items are in the whiskey or whiskies category OR are over 85 proof?

```
query <- dbGetQuery(con, statement = paste(
  "SELECT DISTINCT item_description, category_name, proof",
  "FROM products",
  "WHERE category_name LIKE '%WHISK%' OR cast(proof as integer) >= 85"))

head(query)
```

```
##                       item_description             category_name proof
## 1                      Whiskey Liqueur           WHISKEY LIQUEUR    76
## 2                   Old Taylor Bourbon STRAIGHT BOURBON WHISKIES    80
## 3 Pebble Beach 12yr Single Malt Scotch        SINGLE MALT SCOTCH    86
## 4               Johnnie Walker Gold DNO           SCOTCH WHISKIES    80
## 5                          Ron Burgundy          BLENDED WHISKIES    80
## 6             Woodford Reserve Bourbon STRAIGHT BOURBON WHISKIES    86
```

Which items are over 90 proof?

```
query <- dbGetQuery(con, statement = paste(
  "SELECT DISTINCT item_description, proof",
  "FROM products",
  "WHERE cast(proof as integer) > 90",
```

```
  "ORDER BY proof"))

head(query)
```

```
##                     item_description proof
## 1    Phillips Wintergreen 100 P.e.t.   100
## 2               Zambello Red Sambuca   100
## 3                  Dt Blackbull 12yr   100
## 4                Lead Mine Moonshine   100
## 5       J W Dant Bond 100prf Bourbon   100
## 6 J.w. Dant 100prf Bond Bourbon 54mo   100
```

Which items have a case cost of less than $60?

```
query <- dbGetQuery(con, statement = paste(
  "SELECT DISTINCT item_description, case_cost",
  "FROM products",
  "WHERE case_cost < 60",
  "ORDER BY case_cost"))

head(query)
```

```
##                   item_description case_cost
## 1                  Littlemill 12yr      0.00
## 2              Hammer Sickle Vodka      0.00
## 3        Ypioca Cachaca Ouro(gold)      0.00
## 4                      Cachaca 21      7.50
## 5 Hiram Walker Peach 3pak Schnapps     10.25
## 6                 Gaetano Kamikaze     11.50
```

Which items are either Single Malt Scotches or Canadian Whiskies?

```
query <- dbGetQuery(con, statement = paste(
  "SELECT DISTINCT item_description, category_name",
  "FROM products",
  "WHERE category_name = 'SINGLE MALT SCOTCH' OR category_name LIKE 'CANADIAN WHISK%'"))

head(query)
```

```
##                       item_description      category_name
## 1           Balvenie 12yr Single Barrel SINGLE MALT SCOTCH
## 2             Windsor Canadian Traveler  CANADIAN WHISKIES
## 3 Glenfiddich 15yr Cask Strength Scotch SINGLE MALT SCOTCH
## 4 Glenmorangie 10 Yr Single Malt Scotch SINGLE MALT SCOTCH
## 5    Matisse Pure Malt 12 Yr Old Scotch SINGLE MALT SCOTCH
## 6  Glen Garioch 8 Yr Single Malt Scotch SINGLE MALT SCOTCH
```

Which items are in the whiskey (or whiskies) category?

```
query <- dbGetQuery(con, statement = paste(
  "SELECT DISTINCT item_description, category_name",
  "FROM products",
  "WHERE category_name LIKE '%WHISK%'"))

head(query)
```

```
##                              item_description
## 1          Old Forester Birthday Bourbon 2010
## 2                            Laphroaig 18YR HA
## 3                     Chivas Regal Scotch 12yr
## 4 Wild Turkey Kentucky Spirit Bourbon Whiskey
## 5                                  Ginos White
## 6                        Jefferson's Reserve
##                     category_name
## 1      STRAIGHT BOURBON WHISKIES
## 2                SCOTCH WHISKIES
## 3                SCOTCH WHISKIES
## 4 SINGLE BARREL BOURBON WHISKIES
## 5               BLENDED WHISKIES
## 6      STRAIGHT BOURBON WHISKIES
```

What is the most exensive purchase of Svedka?

```
query <- dbGetQuery(con, statement = paste(
  "SELECT DISTINCT description, total",
  "FROM sales",
  "WHERE description LIKE '%Svedka%'",
  "ORDER BY total DESC",
  "LIMIT 1"))

query
```

```
##    description   total
## 1 Svedka Vodka 34522.8
```

Which unique items in the "WHISK" category have a proof over or equal to 70?

```
query <- dbGetQuery(con, statement = paste(
  "SELECT DISTINCT item_description AS Product, category_name, proof",
  "FROM products",
  "WHERE category_name LIKE '%WHISK%' AND cast(proof as integer) >= 70"))

head(query)
```

```
##                              product             category_name proof
## 1                   Whiskey Liqueur           WHISKEY LIQUEUR    76
## 2                Old Taylor Bourbon STRAIGHT BOURBON WHISKIES    80
## 3            Johnnie Walker Gold DNO           SCOTCH WHISKIES    80
## 4        Virginia Gentleman 80 Proof STRAIGHT BOURBON WHISKIES    80
## 5 Old Forester Birthday Bourbon 2010 STRAIGHT BOURBON WHISKIES    94
## 6      Mccormick Str Bourbon Whiskey STRAIGHT BOURBON WHISKIES    80
```

How many items are available per vendor name?

```
query <- dbGetQuery(con, statement = paste(
  "SELECT vendor_name, COUNT(item_no)",
  "FROM products",
  "GROUP BY vendor_name",
  "ORDER BY count(item_no) DESC"))

head(query)
```

```
##                       vendor_name count
## 1                 Jim Beam Brands   925
## 2                 Diageo Americas   907
## 3 Pernod Ricard Usa/austin Nichols   599
## 4        Yahara Bay Distillers Inc   579
## 5    Heaven Hill Distilleries Inc.   388
## 6              Bacardi U.s.a. Inc.   357
```

What are the top 5 stores with the highest sales?

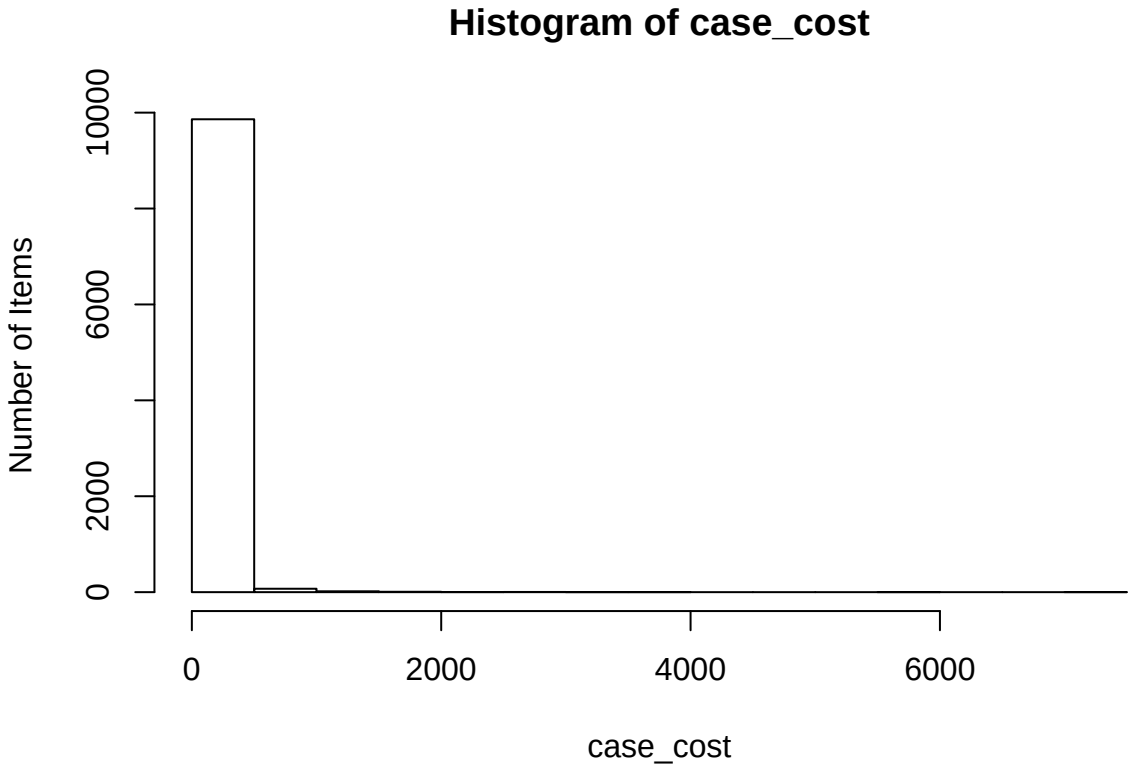
```
query <- dbGetQuery(con, statement = paste(
  "SELECT stores.name, SUM(sales.total)",
  "FROM sales LEFT OUTER JOIN stores ON sales.store = stores.store",
  "GROUP BY stores.name",
  "ORDER BY sum(sales.total) DESC",
  "LIMIT 5"))

query
```

```
##                                 name      sum
## 1       Hy-vee #3 / Bdi / Des Moines 13920087
## 2                     Central City 2 11942400
## 3  Sam's Club 6344 / Windsor Heights  6159480
## 4     Sam's Club 8162 / Cedar Rapids  5734722
## 5 Hy-vee Wine and Spirits / Iowa City  5665144
```

Exploring case-cost

```
## Warning: package 'knitr' was built under R version 3.2.5
```

We can start our exploration of case_cost by creating a histogram and calculating summary statistics:




```
##   mean variance median
## 1   NA       NA     NA
```

Hmmmmmm, that is not very helpful. We can't immediately calculate the mean, median, and variance – there is likely missing data that is preventing their calculation. And that histogram is not very pretty, it is strongly right-skewed. It seems like there are A LOT of cases with a very low cost (comparatively), and only a few with a very high cost.

Let's figure out how many missing data points there are:

number.of.missing.values
8

OK, so there are 8 missing values in the data set. Let's try removing these missing data points and *then* calculating summary statistics:

mean	variance
111.43	25152.13

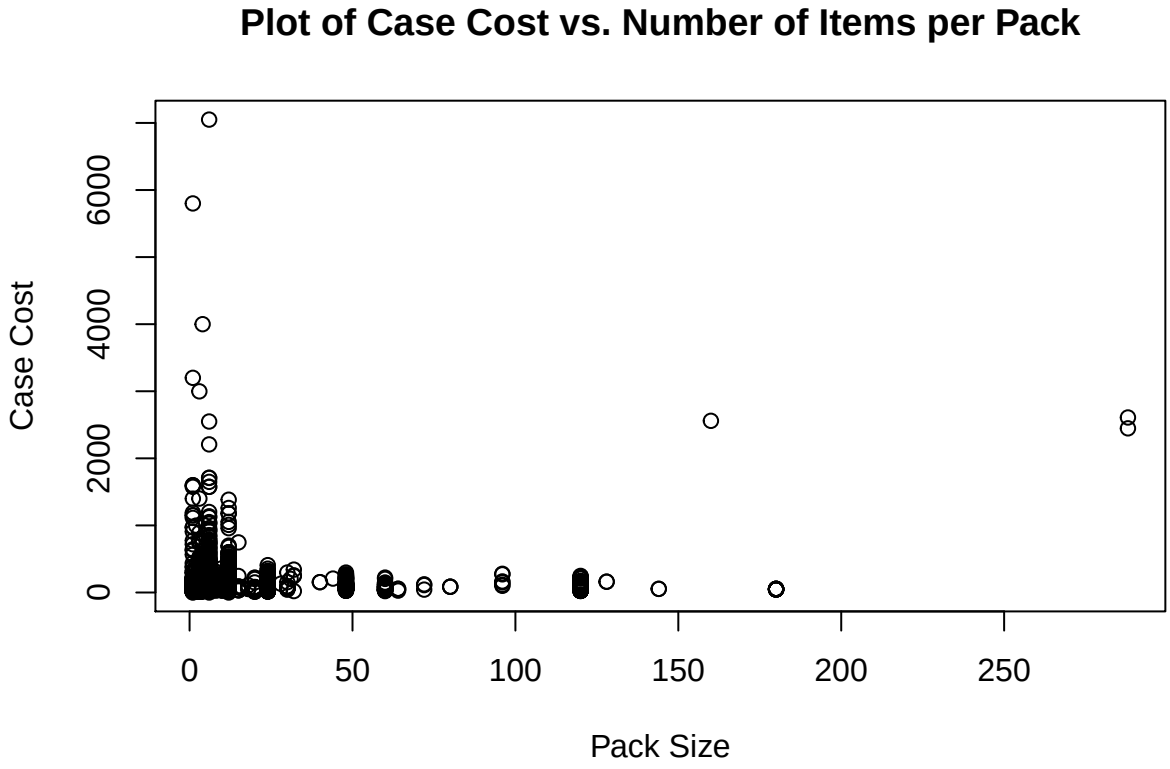
minimum	median	maximum
0	83	7049.7

Great! Now we have an idea of what our data looks like.
Our initial observations appear to be correct: The maximum value is over $7000 dollars, but the median cost is only $83, indicating a strong right-skew. An extremely large variance of $25,000 (much larger than the range of the data) confirms a strongly-skewed dataset as well.

Let's try to figure out if case cost is related to any other variables:

Maybe case cost is related to pack size?

First let's plot case cost vs. pack size:



There does not apppear to be a relationship, but let's run a quick regression on the data to check:

p.value.of.pack.size	R.squared
0.694	0

Pack size is NOT a significant predictor of case cost ($p = 0.694$, $R^2 < 0.0001$).

Maybe case cost is related to shelf price?

Plot case cost vs. shelf price:

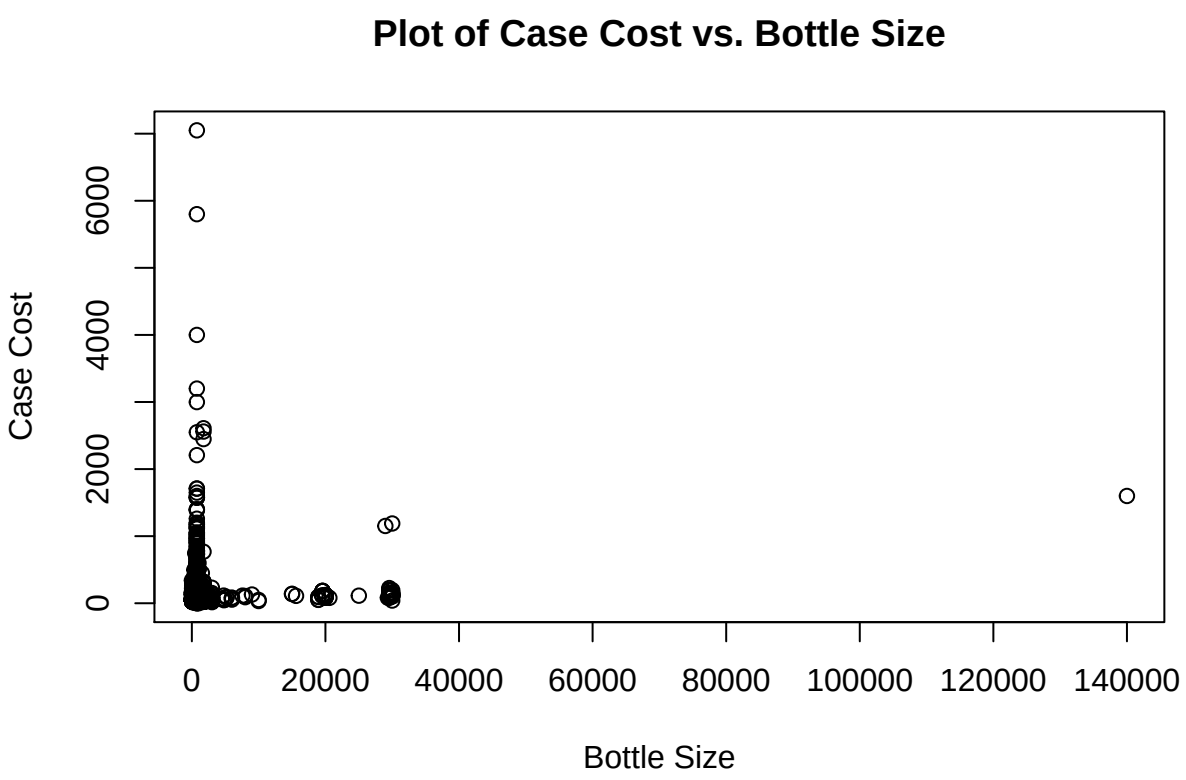
It appears as if there *could be* a relationship here. Let's run a regression through the datapoints to see:

shelf.price.coefficient	p.value.of.shelf.price	R.squared
0.85	0	0.47

Huzzah! We found a variable that is significantly related to case cost. As shelf price increases by \$1, case cost increases by \$0.85, with $p < 0.001$, and nearly 50% of the variation in case cost is explained by shelf price ($R^2 = 0.47$).

Maybe case cost is related to bottle size?

Plot case cost vs. bottle size:



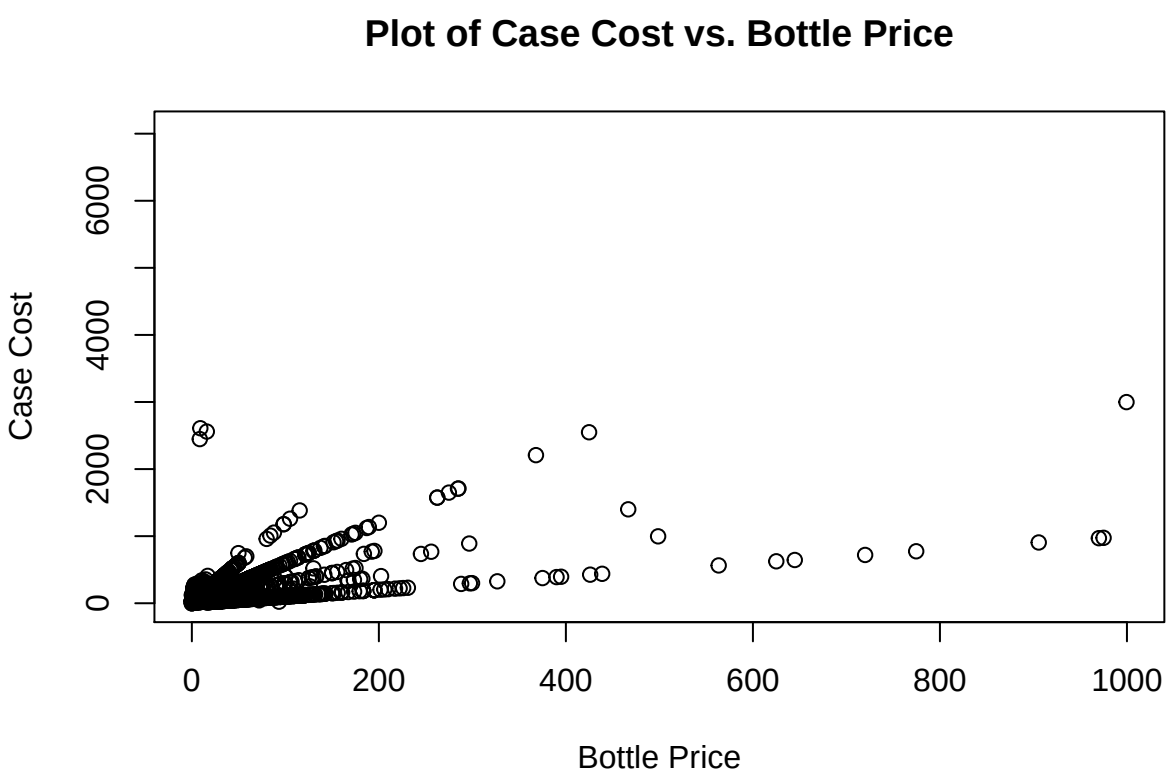
There does not apppear to be a relationship, because the bottle sizes seem to be right-skewed, with many small bottle and one VERY LARGE bottle. But let's run a quick regression on the data to check:

p.value.of.bottle.size	R.squared
0	0.005

According to this regression, bottle size is a significant predictor of case cost ($p < 0.0001$). However, The R^2 value for the regression is 0.005, which means that $< 1\%$ of the variation in Case Cost is explained by Bottle Size. Bottle Size appears to be an significant predictor of case cost because of one influential point, where bottle size = 140000 mL. Without this influential point, there does not appear to be a significant relationship between bottle size and case cost, thus we conclude that bottle size and case cost are not significantly correlated.

Maybe case cost is related to bottle price?

Plot case cost vs. bottle price:



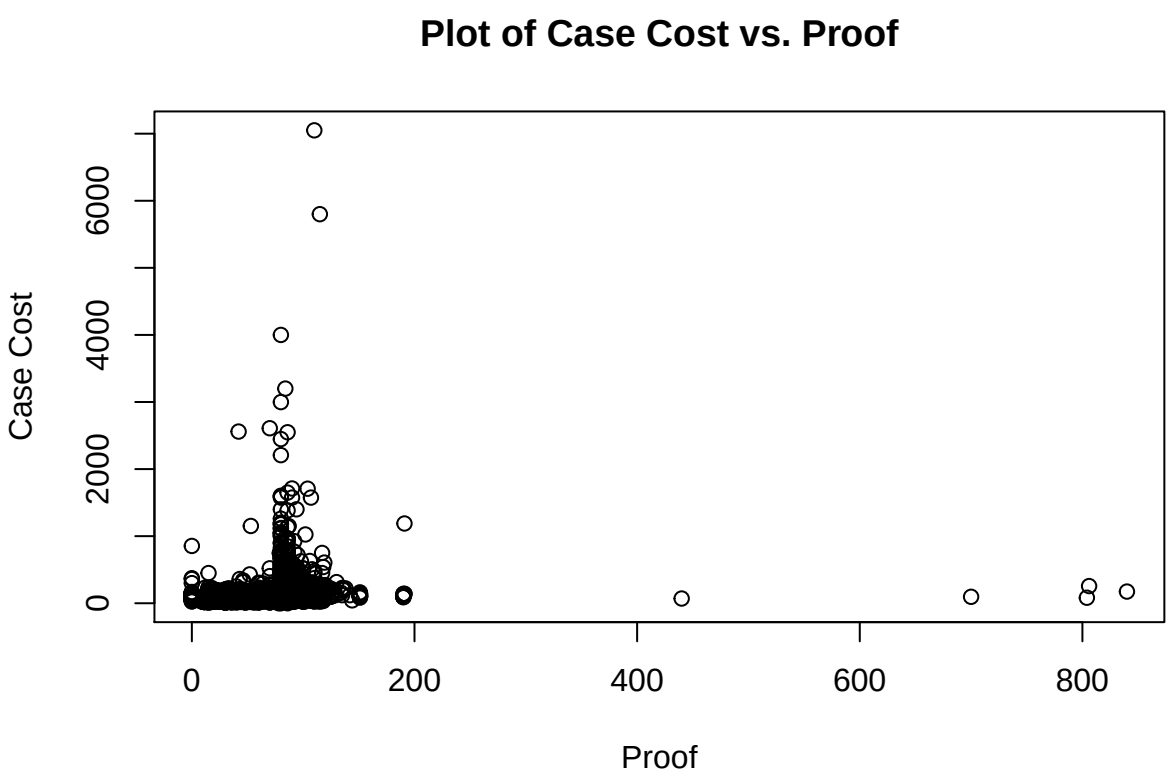
It looks like there might be a relationship here. Let's run a regression to check:

bottle.price.coefficient	p.value.of.bottle.price	R.squared
1.94	0	0.36

Huzzah! Another variable significantly related to case cost! As Bottle Price increases by \$1, Case Cost increases by \$1.94, with $p < 0.001$, and 36% of the variation in Case Cost is explained by Bottle Price ($R^2 = 0.36$).

Finally, let's check if proof is related to case cost

Plot case cost vs. proof:



There appears to be a few influential points, but not a strong underlying a relationship. Let's run a regression:

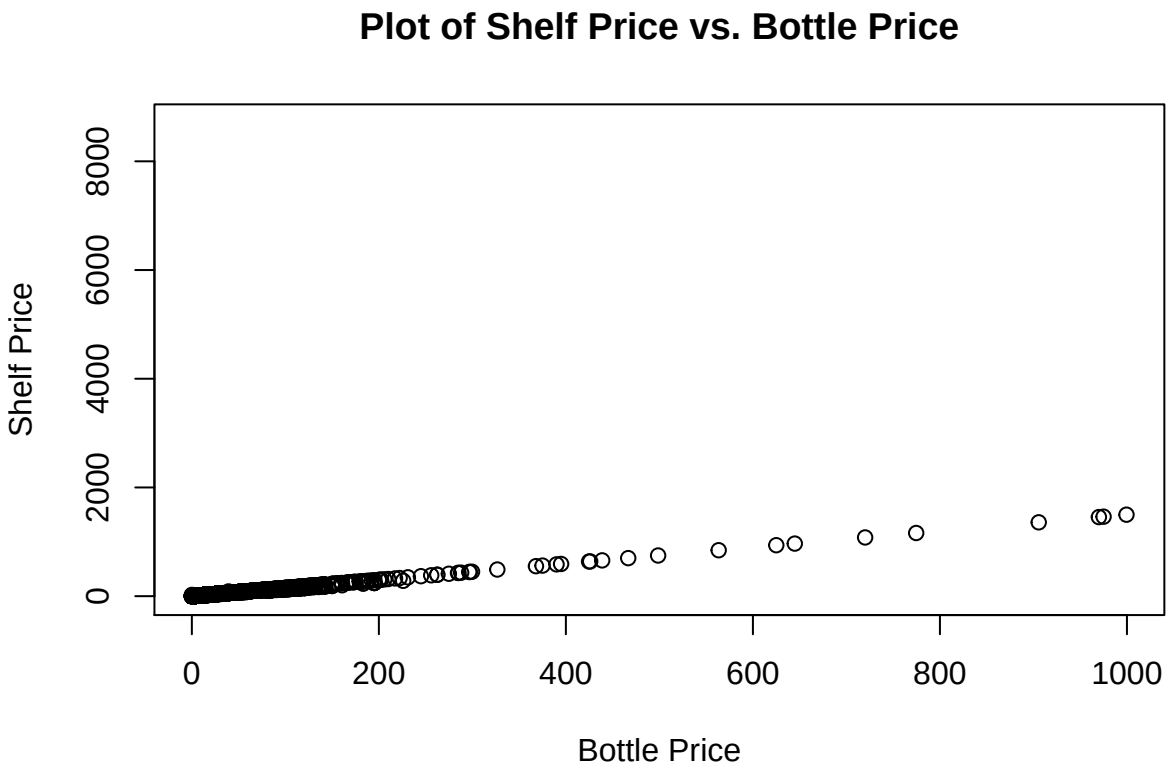
p.value.of.proof	R.squared
0.32	0.09

Proof is NOT significantly related to case cost ($p = 0.32$, $R^2 = 0.09$).

Summary

We found 2 variables that are significantly related to case cost: shelf price and bottle price. Shelf price explains nearly 50% of the variation in case cost, and bottle price explains 35% of the variation in case cost.

However, shelf price and bottle price seem like they could be correlated with one another: shouldn't the price of an individual bottle predict how much the bottle is sold for? Let's do a quick plot of shelf price vs. bottle price and run a quick regression:



p.value	R.squared
0	0.9975

Our hypothesis was correct: shelf price and bottle price are nearly perfectly correlated – essentially all of the variation in shelf price can be explained by bottle price ($R^2 = 0.9975$).

Becaue shelf price and bottle price are so perfectly correlated, we would not want to use BOTH of them to predict case cost, because we would be using redundant information. Since shelf price is more closely correlated with case cost, we would want to use only shelf price to predict case cost (if we were choosing between shelf price and bottle price).