SQL Report

Ryan Ranjiv Singh

December 4, 2016

Library Initialization

```r
library(RPostgreSQL)
```

```
## Loading required package: DBI
```

```r
library(ggplot2)
```

Initialization Code

```r
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)


#Run this to test connection
dbListTables(con)
```

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

```r
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"
```

```r
r1 <- dbGetQuery(con, statement = paste(
  "SELECT DISTINCT category_name, cast(proof as integer)",
  "FROM products",
  "WHERE cast(proof as integer) >= 85 and category_name is not null"))

#to see what r1 is
str(r1)
```

```
## 'data.frame':    268 obs. of  2 variables:
##  $ category_name: chr  "WHISKEY LIQUEUR" "SCOTCH WHISKIES" "HIGH PROOF BEER" "SINGLE BARREL BOURBON W
##  $ proof        : int  91 86 90 111 85 94 86 88 100 118 ...
```

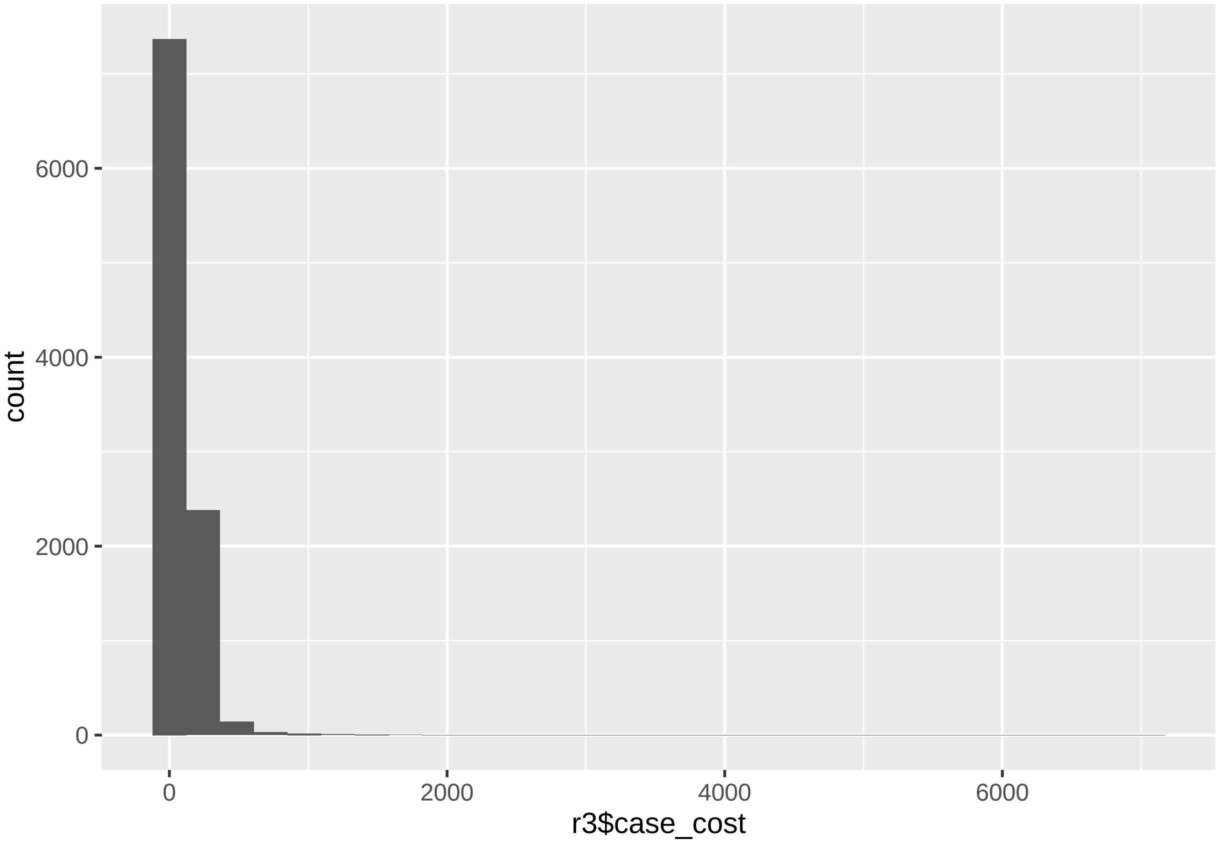
```
#Exploring case cost
r3 <- dbGetQuery(con, statement = paste(
  "SELECT case_cost",
  "FROM products"))
```

Analysis

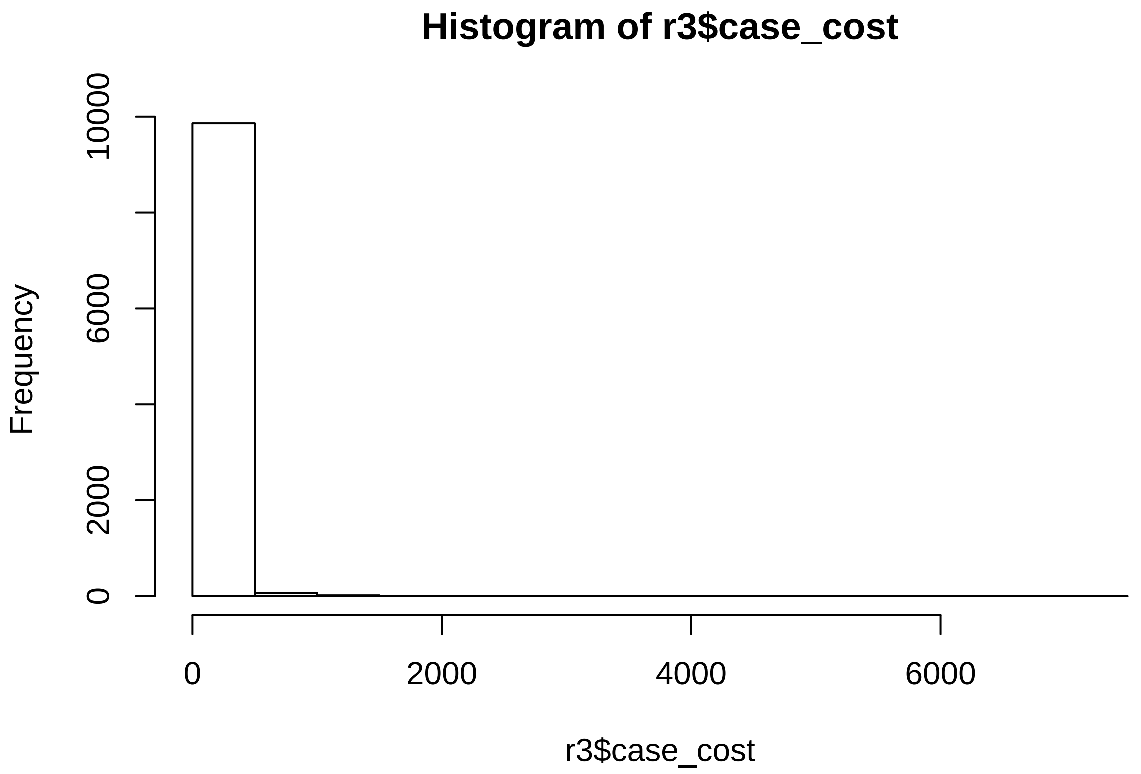
```
qplot(r3$case_cost)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 8 rows containing non-finite values (stat_bin).
```



```
hist(r3$case_cost)
```



```r
r4 <- r3$case_cost[!is.na(r3$case_cost)]

mean(r4)
```

```
## [1] 111.4349
```

```r
median(r4)
```

```
## [1] 83
```

```r
var(r4)
```

```
## [1] 25152.13
```

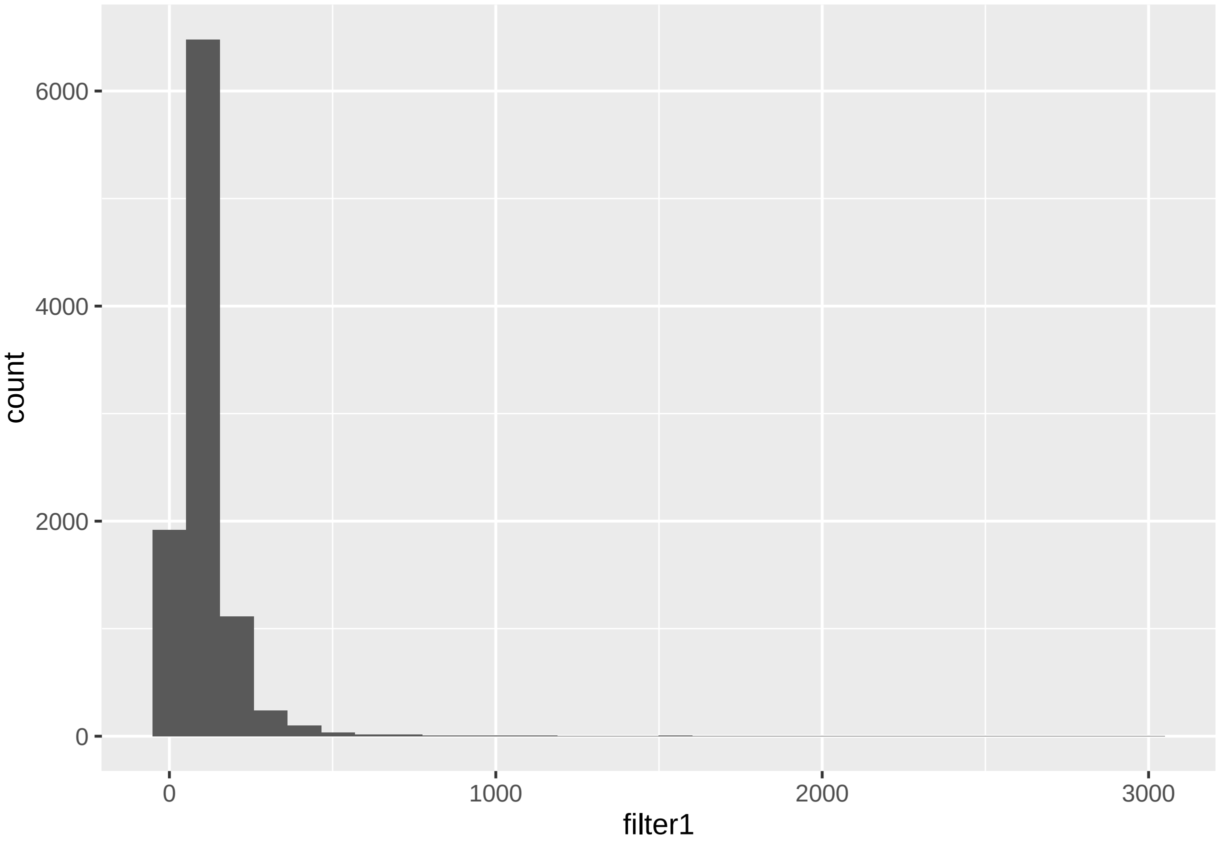
Looking at the dot chart and the histogram, we can see that most of the data is clustered under 3000, leading to an assumption that the data point above 3000 are outliers.

As such, we should remove those data points.

```r
filter1 <- subset(r3, case_cost < 3000)
qplot(filter1)
```

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

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Analysis 2

Now I attempt to run a regression using the variables. However, I need to change the class types for certain variables, and also remove unwanted characters like '$' from the bottle price column.

The regression conducted will not include the names of the vendor or category, as that generates too many individual levels that complicates the simple regression. Variable selection methods will have to be conducted in order to better analyze the effect of the vendor/category on the regression.

```
r2 <- dbGetQuery(con, statement = paste(
  "SELECT category_name, vendor_name, bottle_size, pack, inner_pack, proof, bottle_price, shelf_price, 
  "FROM products"))
```

```
## Warning in postgresqlExecStatement(conn, statement, ...): RS-DBI driver
## warning: (unrecognized PostgreSQL field type money (id:790) in column 6)
```

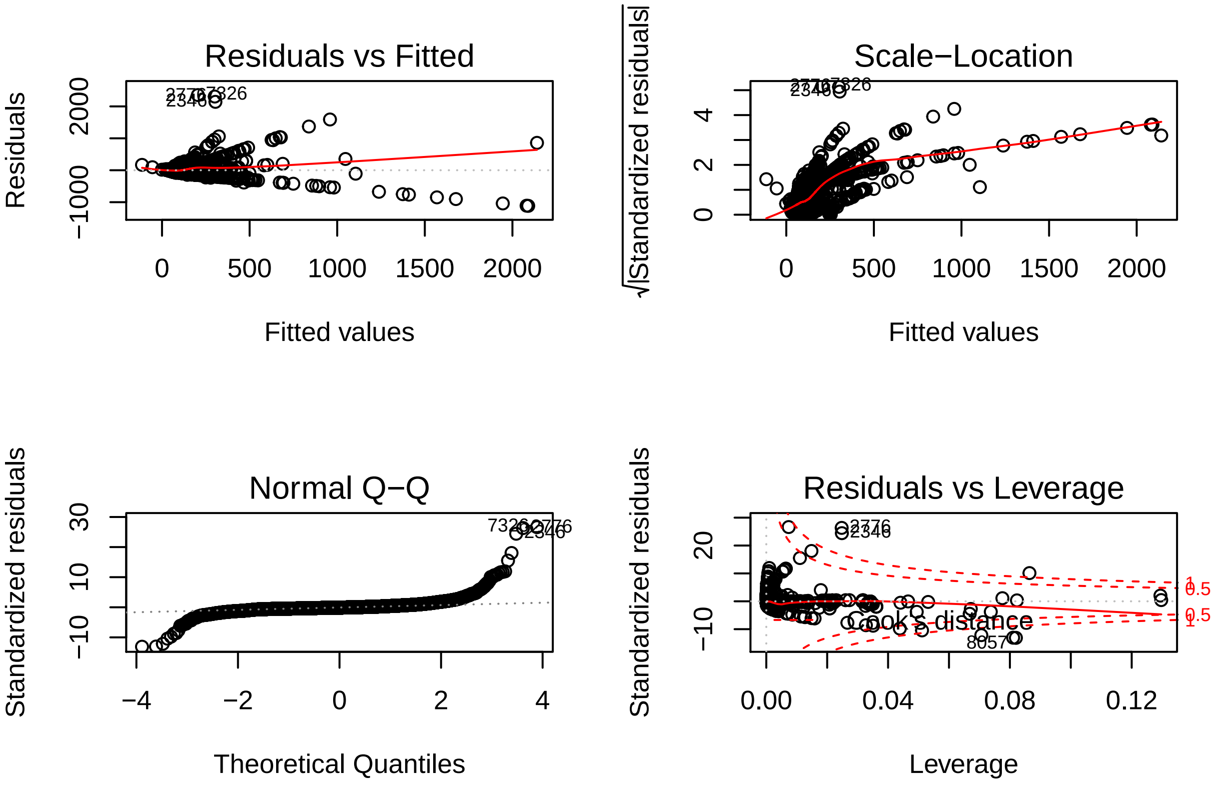
```
r2$proof <- as.integer(r2$proof)
r2$bottle_price <- sub('.','', r2$bottle_price)
r2$bottle_price <- as.numeric(r2$bottle_price)
```

```
## Warning: NAs introduced by coercion
```

```
reg1 <- lm(r2$case_cost ~ r2$bottle_size + r2$pack + r2$inner_pack + r2$proof + r2$bottle_price + r2$sh
summary(reg1)
```

```
##
## Call:
## lm(formula = r2$case_cost ~ r2$bottle_size + r2$pack + r2$inner_pack +
##     r2$proof + r2$bottle_price + r2$shelf_price)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1115.13   -30.19   -10.46    17.06  2354.48
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     37.9682029  2.4620588  15.421  < 2e-16 ***
## r2$bottle_size  -0.0037524  0.0008745  -4.291 1.80e-05 ***
## r2$pack          0.7535894  0.0502223  15.005  < 2e-16 ***
## r2$inner_pack   -1.4244845  0.2691695  -5.292 1.23e-07 ***
## r2$proof         0.5031635  0.0311243  16.166  < 2e-16 ***
## r2$bottle_price -2.0700531  1.0314476  -2.007   0.0448 *
## r2$shelf_price   2.7567077  0.6886989   4.003 6.31e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 88.76 on 9950 degrees of freedom
##   (20 observations deleted due to missingness)
## Multiple R-squared:  0.4102, Adjusted R-squared:  0.4098
## F-statistic:  1153 on 6 and 9950 DF,  p-value: < 2.2e-16
```

```r
layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
plot(reg1)
```



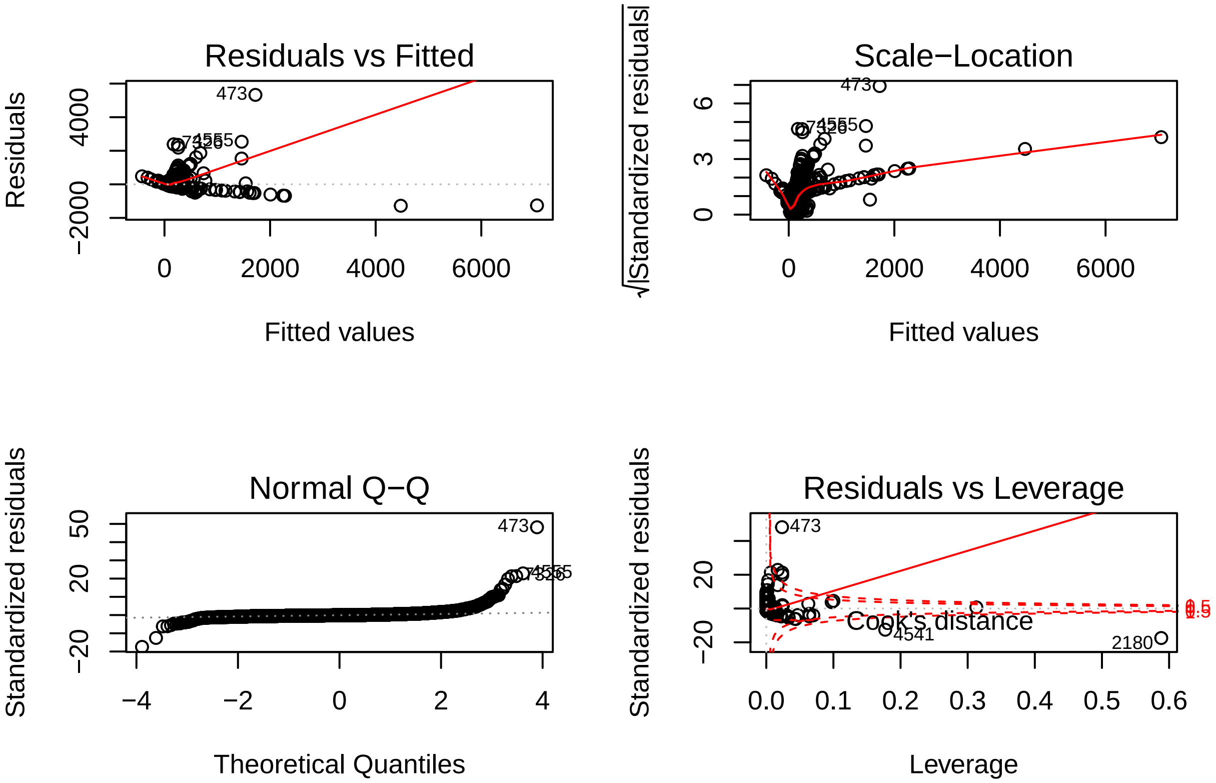
From this regression analysis, it seems like every factor is strongly correlated to the case cost variable, with bottle price having the largest p value. However, the p value is still <0.05.

I will attempt to remove the factor bottle price, since its p value was the lowest, and also has the 'common sense' relationship, as the case cost should be derived from number of bottles in case * bottle price.

```
reg2 <- lm(r2$case_cost ~ r2$bottle_size + r2$pack + r2$inner_pack + r2$proof + r2$shelf_price)
summary(reg2)
```

```
##
## Call:
## lm(formula = r2$case_cost ~ r2$bottle_size + r2$pack + r2$inner_pack +
##     r2$proof + r2$shelf_price)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1275.1   -36.2   -13.3    14.9  5327.5
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    45.2622133  3.0147281  15.014  < 2e-16 ***
## r2$bottle_size -0.0029556  0.0004483  -6.593 4.53e-11 ***
## r2$pack         0.5600500  0.0623539   8.982  < 2e-16 ***
## r2$inner_pack  -4.1850553  0.2975766 -14.064  < 2e-16 ***
## r2$proof        0.6481380  0.0382481  16.946  < 2e-16 ***
## r2$shelf_price  0.9127577  0.0098754  92.427  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 111.9 on 9963 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.502,  Adjusted R-squared:  0.5017
## F-statistic:  2008 on 5 and 9963 DF,  p-value: < 2.2e-16
```

```
layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
plot(reg2)
```



The summary shows that all the remaining factors are significant, as demonstrated in the first regression. The p values of bottle size, inner pack and shelf price all became significantly smaller, as a result of removing bottle price. This suggests that there may have been some correlation between case cost and bottle price.

```
cor.test(r2$case_cost, r2$bottle_price)
```

```
##
##  Pearson's product-moment correlation
##
## data:  r2$case_cost and r2$bottle_price
## t = 74.057, df = 9955, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5831895 0.6085229
## sample estimates:
##       cor
## 0.5960045
```

The correlation test shows that my suspicion is true. Almost 60% of the variation in case cost is represented by bottle price, and since the t value is large, p value is small, we reject the null hypothesis, and can conclude that there is correlation between the variables.