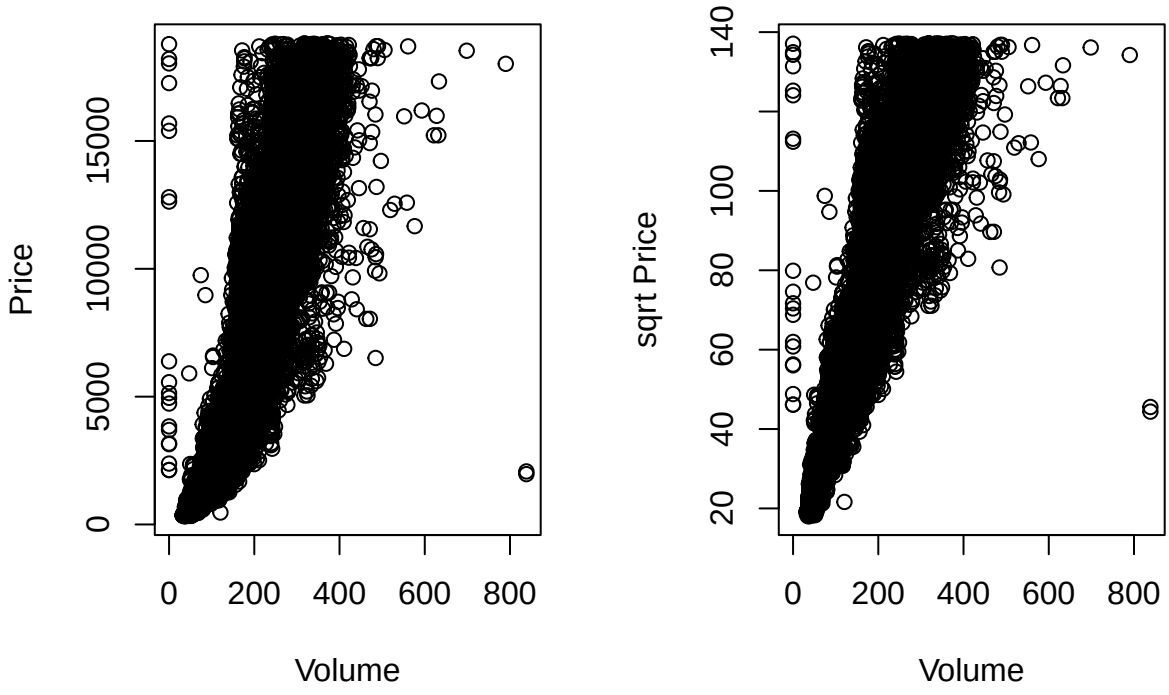
3302 Final Project

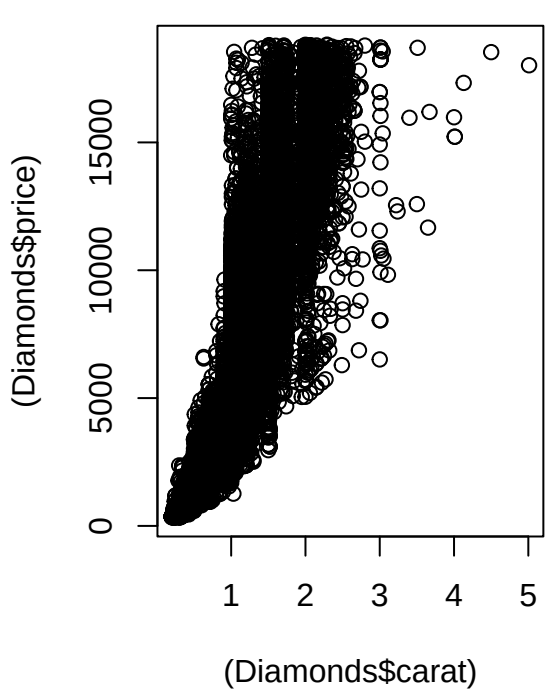
Charles Costanzo

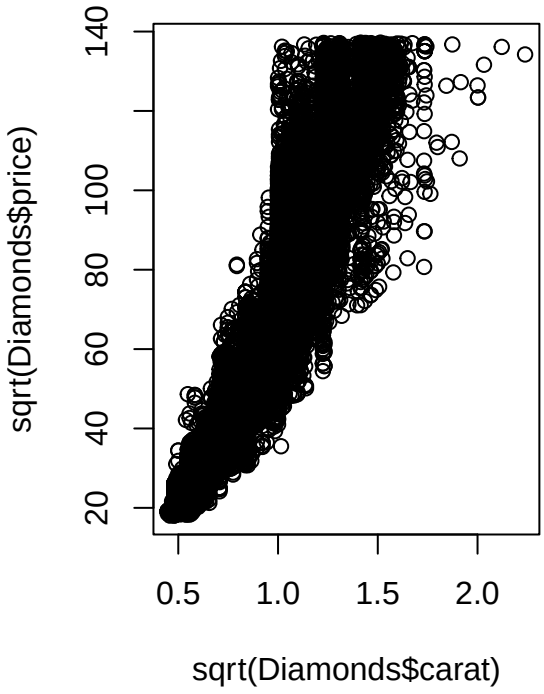
2023-04-11

```
Diamonds <- diamonds %>%
  mutate(volume = x * y * z)

Diamonds <- Diamonds %>%
  select(-c(x,y,z))

par(mfrow = c(1,2))
plot((Diamonds$volume[Diamonds$volume < 3000]),(Diamonds$price[Diamonds$volume < 3000]),
     xlab = "Volume", ylab = "Price")
plot((Diamonds$volume[Diamonds$volume < 3000]),sqrt(Diamonds$price[Diamonds$volume < 3000]),
     xlab = "Volume", ylab = "sqrt Price")
```

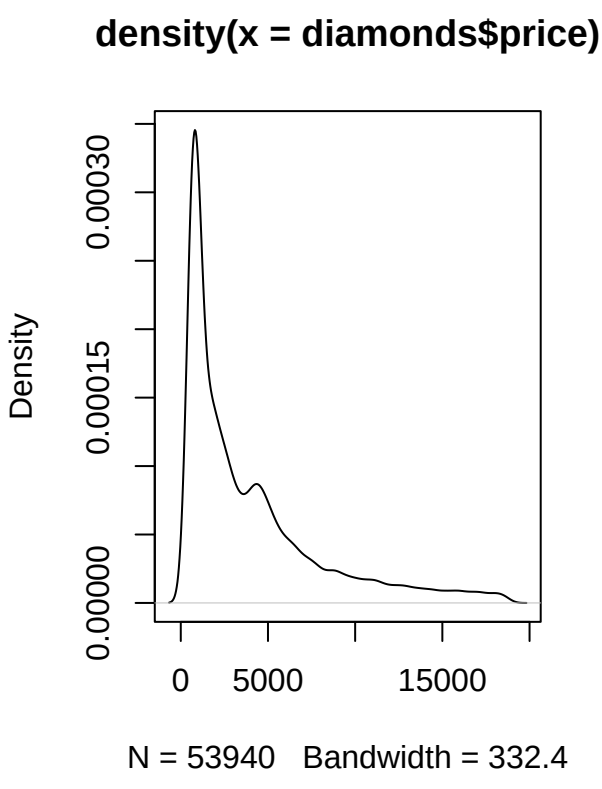



```
par(mfrow = c(1,2))
plot((Diamonds$carat),(Diamonds$price))
plot(sqrt(Diamonds$carat),sqrt(Diamonds$price))
```



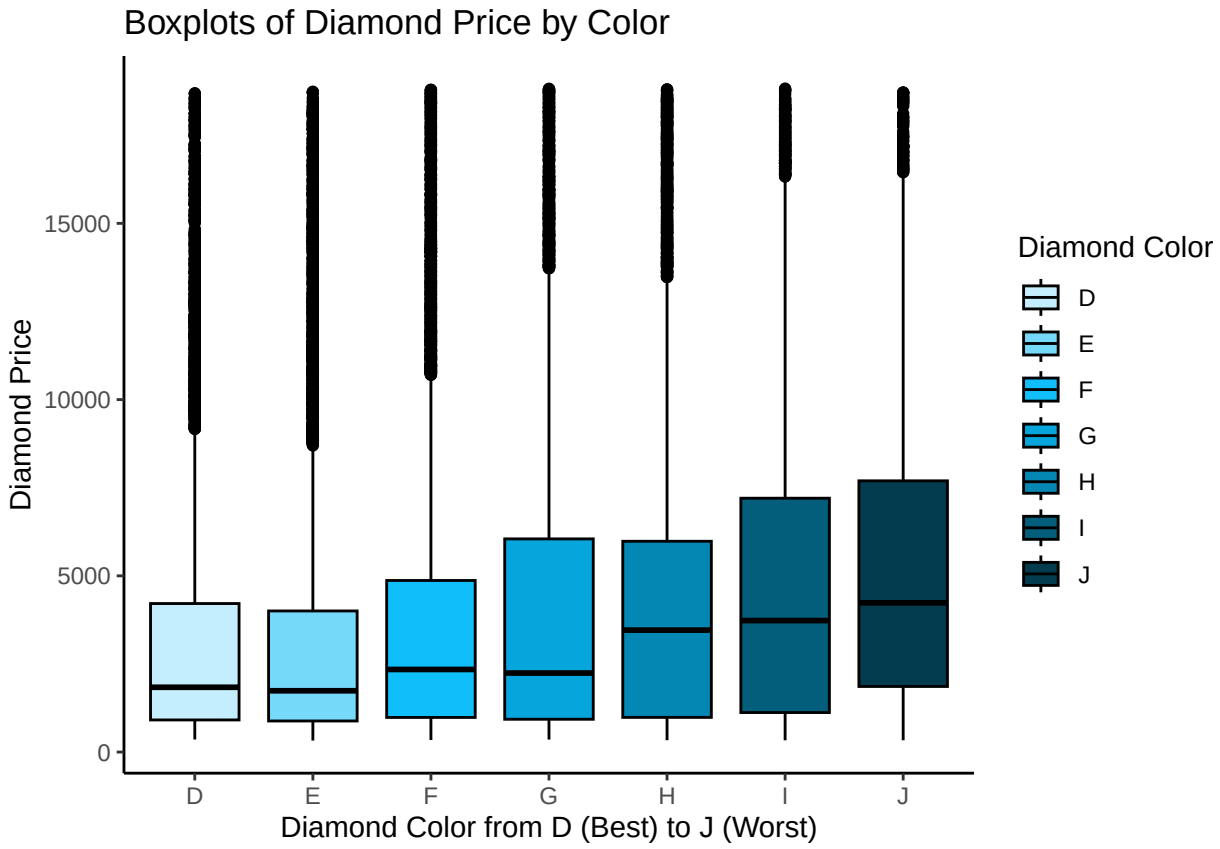



```
plot(density(diamonds$price))
```



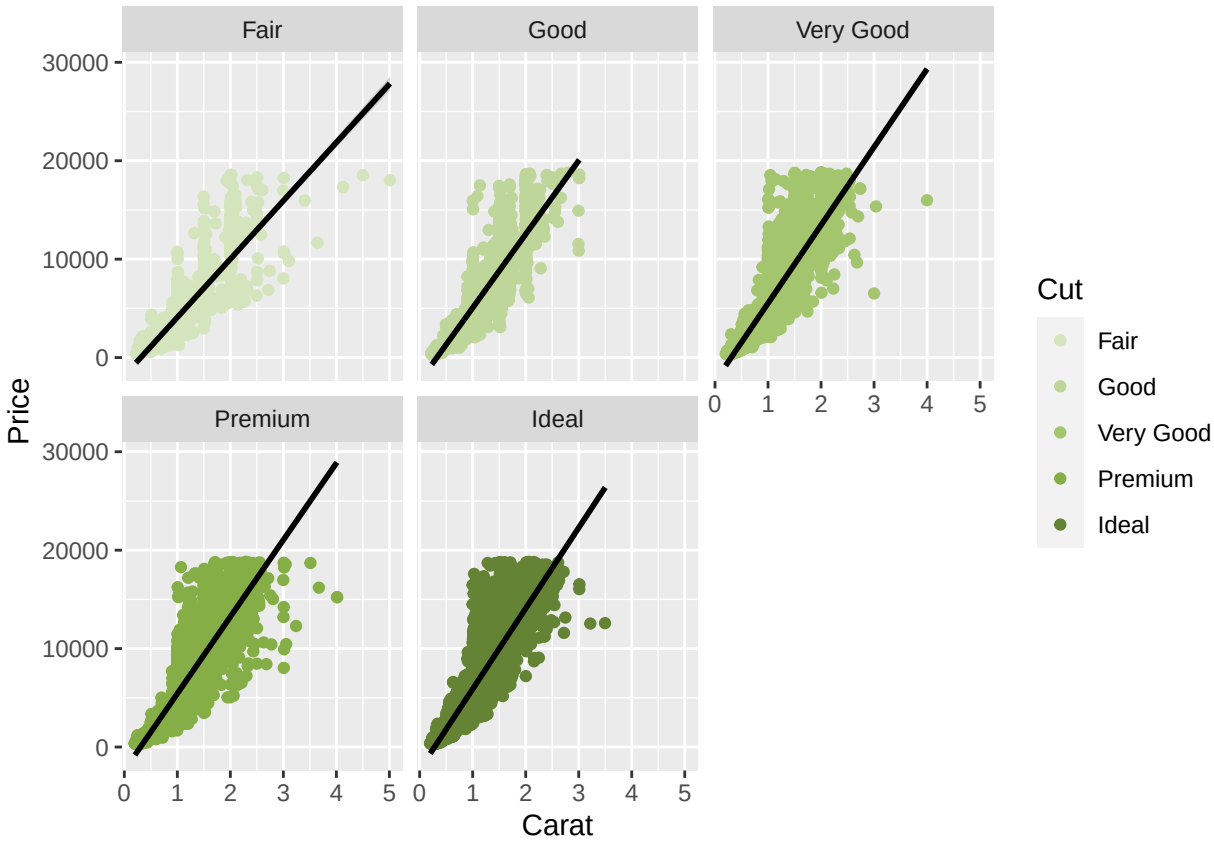

```
Diamonds %>%
  ggplot() +
  geom_boxplot(aes(x = color, y = price, fill = color), color = "black") +
  theme_classic() +
```

```
  xlab("Diamond Color from D (Best) to J (Worst)") + ylab("Diamond Price") +
  ggtitle("Boxplots of Diamond Price by Color") +
  scale_fill_manual(name = "Diamond Color",
                    values = c("#C4EEFD","#75D9FA","#10BFF9","#06A5DB","#0587B3","#035E7B",
                               "#033C4F"))
```




```
ggplot(data = diamonds) +
  geom_jitter(mapping = aes(
    x = carat,
    y = price,
    color = cut
  )) +
  geom_smooth(mapping = aes(
    x = carat,
    y = price,
  ),
  color = "black",
      method = "glm") +
scale_color_manual(values = c("#D4E4BC","#BED699","#A2C56D","#85AF46","#648334"),
                   name = "Cut") +
  xlab("Carat") + ylab("Price") +
facet_wrap(~cut)
```

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


Fair
Good
Very Good
Premium
Ideal
Price
Carat
30000
20000
10000
0
0
1
2
3
4
5
Cut
Fair
Good
Very Good
Premium
Ideal

```r
quantile(diamonds$price)
```

```
##       0%      25%      50%      75%     100%
##   326.00   950.00  2401.00  5324.25 18823.00
```

```r
Diamonds <- diamonds

Q1 <- quantile(diamonds$price, .25)

Q3 <- quantile(diamonds$price, .75)

IQR <- IQR(diamonds$price)

Diamonds <- Diamonds %>%
  mutate(volume = x * y * z,
         expensive = case_when(
           diamonds$price > median(diamonds$price) ~ "1",
           diamonds$price <= median(diamonds$price) ~ "0"
           ),
         cut = case_when(
           cut == "Ideal" ~ "1",
           cut == "Premium" ~ "2",
           cut == "Very Good" ~ "3",
           cut == "Good" ~ "4",
           cut == "Fair" ~ "5"
         ),
         color = case_when(
           color == "D" ~ "1",
           color == "E" ~ "2",
           color == "F" ~ "3",
           color == "G" ~ "4",
           color == "H" ~ "5",
           color == "I" ~ "6",
           color == "J" ~ "7"
         ),
         clarity = case_when(
           clarity == "IF" ~ "1",
           clarity == "VVS1" ~ "2",
           clarity == "VVS2" ~ "3",
           clarity == "VS1" ~ "4",
           clarity == "VS2" ~ "5",
           clarity == "SI1" ~ "6",
           clarity == "SI2" ~ "7",
           clarity == "I1" ~ "8"
         ),
         outlier = case_when(
           price < Q1 ~ "1",
           price > Q3 ~ "1",
           TRUE ~ "0"
         ))

Diamonds <- Diamonds %>%
  mutate(cut = as.factor(cut),
```

```
         color = as.factor(color),
         clarity = as.factor(clarity))

Diamonds <- Diamonds %>%
  mutate(cut = relevel(Diamonds$cut, ref = "5"),
         color = relevel(Diamonds$color, ref = "7"),
         clarity = relevel(Diamonds$clarity, ref = "8"))

Diamonds_no_outlier <- Diamonds %>%
  filter(outlier == "0")
```

```r
model1 <- glm(as.factor(expensive) ~ (carat) + clarity,
              family = binomial(link = "logit"),
              data = Diamonds_no_outlier)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
summary(model1)
```

```
##
## Call:
## glm(formula = as.factor(expensive) ~ (carat) + clarity, family = binomial(link = "logit"),
##     data = Diamonds_no_outlier)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -28.0335     0.4752  -59.00   <2e-16 ***
## carat        28.2667     0.4472   63.20   <2e-16 ***
## clarity1     12.0036     0.2997   40.05   <2e-16 ***
## clarity2     11.5019     0.2742   41.95   <2e-16 ***
## clarity3     10.8982     0.2607   41.80   <2e-16 ***
## clarity4      9.1439     0.2232   40.96   <2e-16 ***
## clarity5      8.7366     0.2172   40.22   <2e-16 ***
## clarity6      7.8126     0.2079   37.59   <2e-16 ***
## clarity7      6.5436     0.2036   32.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 37391.1  on 26971  degrees of freedom
## Residual deviance:  9687.1  on 26963  degrees of freedom
## AIC: 9705.1
##
## Number of Fisher Scoring iterations: 8
```

```r
exp(coef(model1))
```

```
##  (Intercept)        carat     clarity1     clarity2     clarity3     clarity4
## 6.686854e-13 1.888274e+12 1.633371e+05 9.889961e+04 5.408062e+04 9.357532e+03
##     clarity5     clarity6     clarity7
## 6.226681e+03 2.471662e+03 6.947833e+02
```