LAB 2 / ASSIGNMENT 2

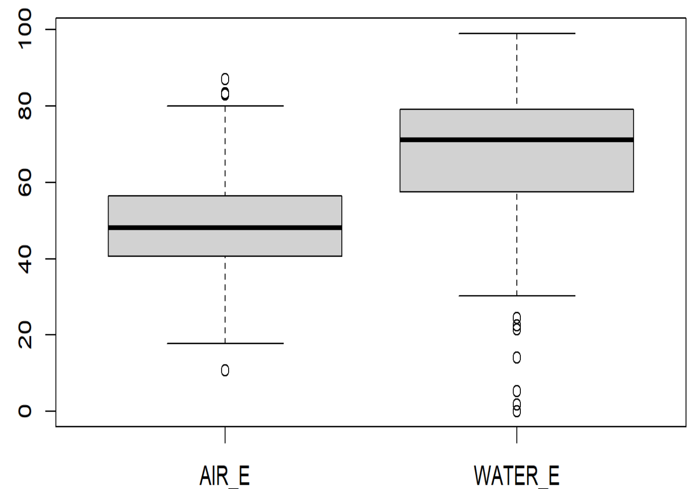
Data Analytics

Ethan Cruz

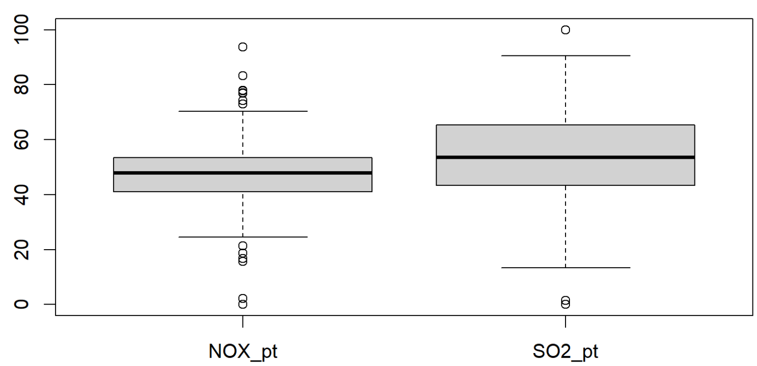
cruze6

Lab2 part 1a

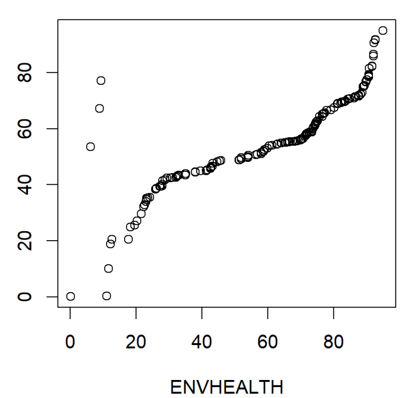
```
> # AIR & WATER PROCESSING
> AIR <- as.numeric(AIR_E[!tf])
> WATER <- as.numeric(WATER_E[!tf])
> mmm(AIR)
[1] "Mean: 49.460736196319"
[1] "Median: 48.24"
[1] "Mode: 44.69"
> mmm(WATER)
[1] "Mean: 67.4782208588957"
[1] "Median: 71.17"
[1] "Mode: 71.4"
```

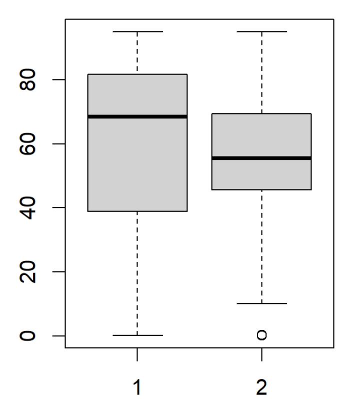



```
> mmm(NOX)
[1] "Mean: 47.5077721251779"
[1] "Median: 48.03794684"
[1] "Mode: 28.36468953"
> mmm(SO2)
[1] "Mean: 53.0528796301779"
[1] "Median: 53.72759976"
[1] "Mode: 20.63477718"
```



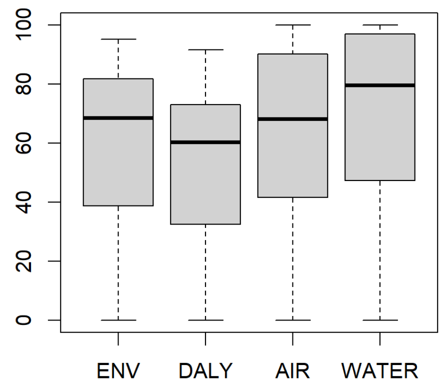

```
> mmm(CLIM)
[1] "Mean: 55.3337423312883"
[1] "Median: 55.43"
[1] "Mode: 60.74"
> mmm(AGRI)
[1] "Mean: 70.8582208588957"
[1] "Median: 75.29"
[1] "Mode: 54.55"
```





Lab2 part 1b

```
> summary(lmENVH)

Call:
lm(formula = ENVHEALTH ~ DALY + AIR_H + WATER_H)

Residuals:
       Min         1Q     Median         3Q        Max 
-0.0073210 -0.0027069 -0.0000915  0.0022285  0.0053404 

Coefficients:
              Estimate Std. Error  t value Pr(>|t|)    
(Intercept) -1.458e-05  6.520e-04   -0.022    0.982    
DALY         5.000e-01  1.988e-05 25147.716   <2e-16 ***
AIR_H        2.500e-01  1.276e-05 19593.273   <2e-16 ***
WATER_H      2.500e-01  1.816e-05 13764.921   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.003015 on 159 degrees of freedom
Multiple R-squared:      1,	Adjusted R-squared:      1 
F-statistic: 3.77e+09 on 3 and 159 DF,  p-value: < 2.2e-16
```




```
> summary(Model1)

Call:
lm(formula = CLIMATE ~ DALY + ENVHEALTH + WATER_H)

Residuals:
    Min      1Q  Median      3Q     Max 
-37.218  -9.180   0.845   8.577  46.138 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 75.14318    2.96869  25.312   <2e-16 ***
DALY        -0.22501    0.16088  -1.399    0.164    
ENVHEALTH    0.06172    0.23239   0.266    0.791    
WATER_H     -0.16252    0.11501  -1.413    0.160    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.73 on 159 degrees of freedom
Multiple R-squared:  0.2919,	Adjusted R-squared:  0.2785 
F-statistic: 21.85 on 3 and 159 DF,  p-value: 6.709e-12
```

```
> summary(Model2)

Call:
lm(formula = CLIMATE ~ DALY + ENVHEALTH + WATER_H)

Residuals:
    Min      1Q  Median      3Q     Max 
-37.218  -9.180   0.845   8.577  46.138 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 75.14318    2.96869  25.312   <2e-16 ***
DALY        -0.22501    0.16088  -1.399    0.164    
ENVHEALTH    0.06172    0.23239   0.266    0.791    
WATER_H     -0.16252    0.11501  -1.413    0.160    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.73 on 159 degrees of freedom
Multiple R-squared:  0.2919,	Adjusted R-squared:  0.2785 
F-statistic: 21.85 on 3 and 159 DF,  p-value: 6.709e-12
```

For Shapiro-Wilk test:

The Sample size is 163, which is between 5 and 5000 so it’s valid to use the test. The p-value for every test is less than 0.05 and the W is high (~0.9) so they are all close to normal distribution!

Lab2 part 2 - Regression

```
> lmROLL <- lm(ROLL~UNEM+HGRAD)
> summary(lmROLL)

Call:
lm(formula = ROLL ~ UNEM + HGRAD)

Residuals:
    Min      1Q  Median      3Q     Max 
-2102.2  -861.6  -349.4   374.5  3603.5 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -8.256e+03  2.052e+03  -4.023  0.00044 ***
UNEM         6.983e+02  2.244e+02   3.111  0.00449 ** 
HGRAD        9.423e-01  8.613e-02  10.941 3.16e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1313 on 26 degrees of freedom
Multiple R-squared:  0.8489,	Adjusted R-squared:  0.8373 
F-statistic: 73.03 on 2 and 26 DF,  p-value: 2.144e-11
```

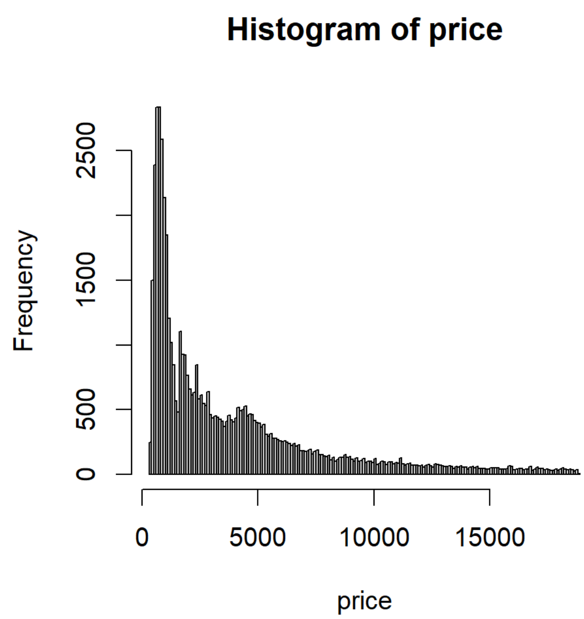
```
> # Predict ROLL if Unem = 7% and HGrad = 90,000
> Punemp <- 7
> Phgrad <- 90000
> newdat <- data.frame(Punemp,Phgrad)
> colnames(newdat) <- c('UNEM','HGRAD')
> predict(lmROLL,newdat)
       1 
81437.04 
```

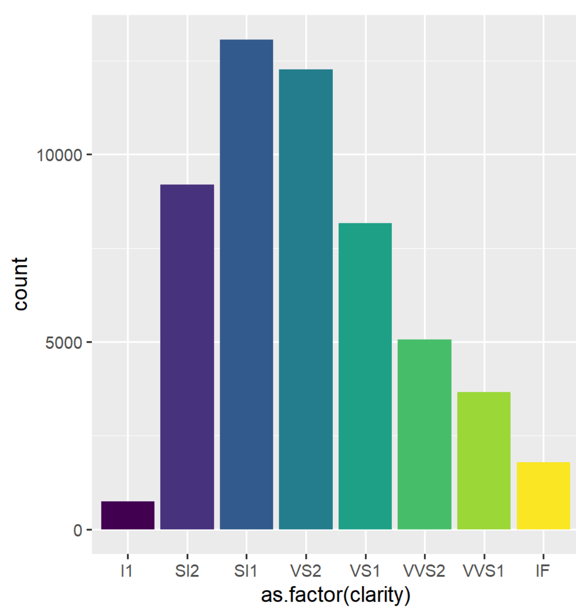
```
> lmROLLCPTA <- lm(ROLL~UNEM+HGRAD+INC)
> summary(lmROLLCPTA)

Call:
lm(formula = ROLL ~ UNEM + HGRAD + INC)

Residuals:
     Min       1Q   Median       3Q      Max 
-1148.84  -489.71    -1.88   387.40  1425.75 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -9.153e+03  1.053e+03  -8.691 5.02e-09 ***
UNEM         4.501e+02  1.182e+02   3.809 0.000807 ***
HGRAD        4.065e-01  7.602e-02   5.347 1.52e-05 ***
INC          4.275e+00  4.947e-01   8.642 5.59e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 670.4 on 25 degrees of freedom
Multiple R-squared:  0.9621,	Adjusted R-squared:  0.9576 
F-statistic: 211.5 on 3 and 25 DF,  p-value: < 2.2e-16
```

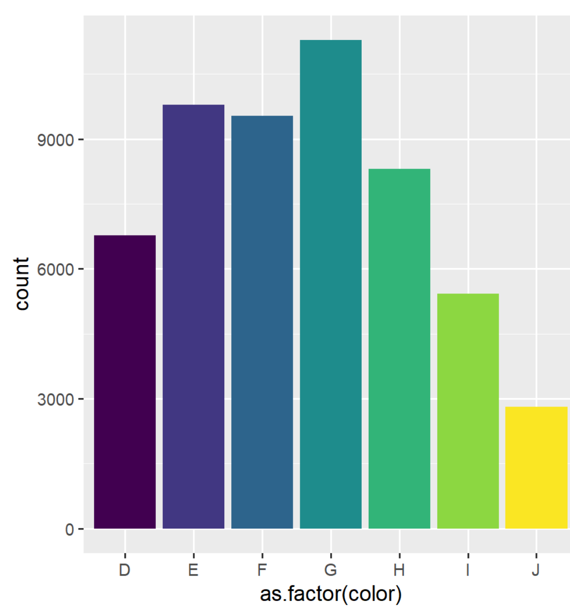
```
> # Predict ROLL if Unem = 7%, HGrad = 90k, and INC = 25k
> Pinc <- 25000
> newdat2 <- data.frame(Punemp,Phgrad,Pinc)
> colnames(newdat2) <- c('UNEM','HGRAD','INC')
> predict(lmROLLCPTA,newdat2)
       1 
137452.6 
```

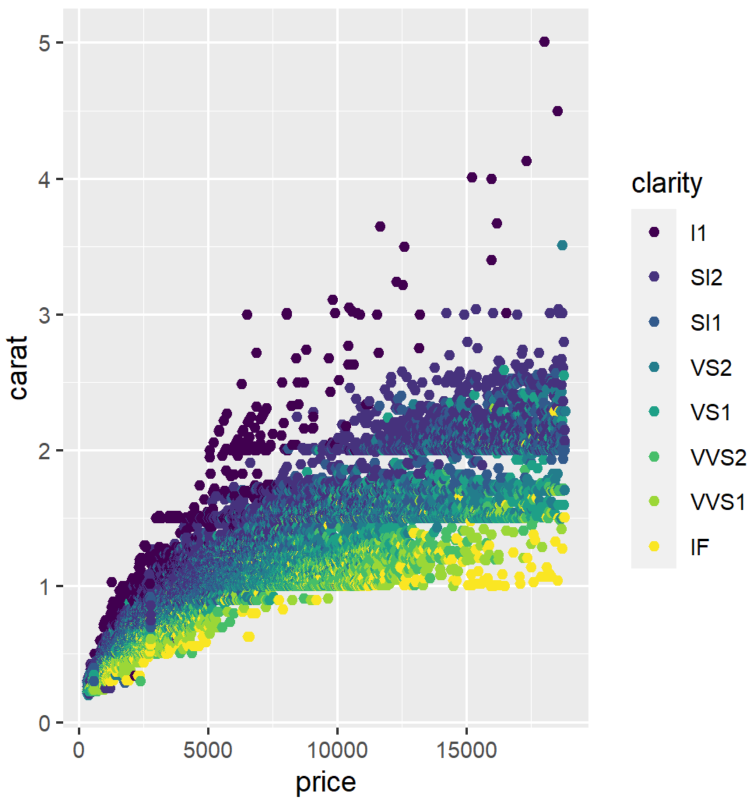
Lab2 part 2 - Diamonds

2a









What do you observe about the relationship between carat, clarity, and price?

Higher Carat tends to lead to higher prices. This is proven by the fact there isn't a 2-carat diamond cheaper than 5k dollars meanwhile, most diamonds less than 1-carat are <5k dollars.

Clarity will also affect price. A 1-carat diamond with imperfect clarity (I1) will go for <5000 while a 1-carat diamond which is near-flawless (IF) can go for >15000!!

2b

```
> # color by price and carat
> aggregate(x=price, by=list(color), FUN=mean)
  Group.1        x
1       D 3169.954
2       E 3076.752
3       F 3724.886
4       G 3999.136
5       H 4486.669
6       I 5091.875
7       J 5323.818
> aggregate(x=price, by=list(color), FUN=median)
  Group.1      x
1       D 1838.0
2       E 1739.0
3       F 2343.5
4       G 2242.0
5       H 3460.0
6       I 3730.0
7       J 4234.0
> aggregate(x=carat, by=list(color), FUN=mean)
  Group.1         x
1       D 0.6577948
2       E 0.6578667
3       F 0.7365385
4       G 0.7711902
5       H 0.9117991
6       I 1.0269273
7       J 1.1621368
> aggregate(x=carat, by=list(color), FUN=median)
  Group.1    x
1       D 0.53
2       E 0.53
3       F 0.70
4       G 0.70
5       H 0.90
6       I 1.00
7       J 1.11
```

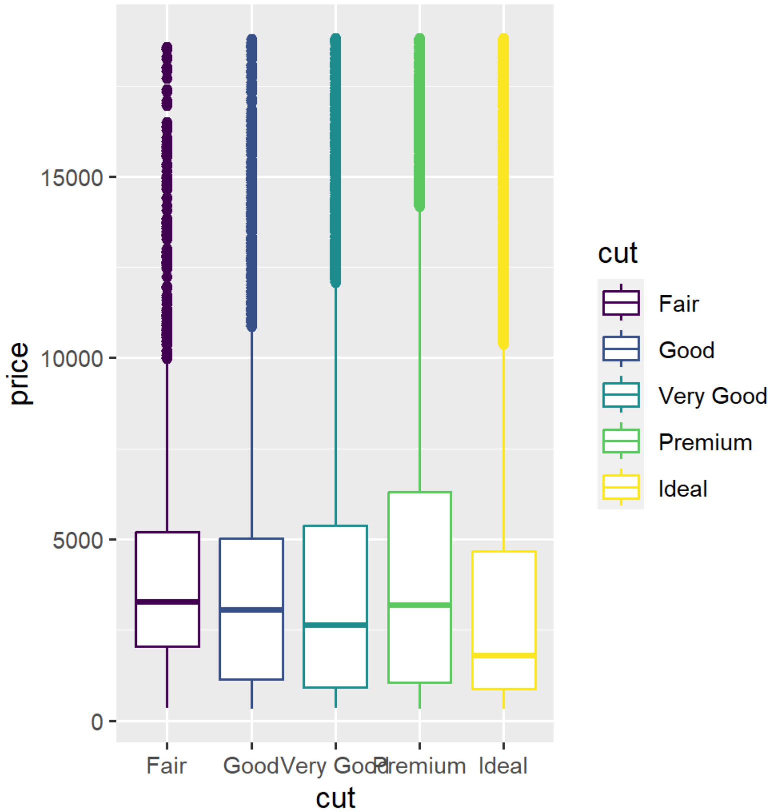
```
> # cut vs price
> aggregate(x=price, by=list(cut), FUN=mean)
    Group.1        x
1      Fair 4358.758
2      Good 3928.864
3 Very Good 3981.760
4   Premium 4584.258
5     Ideal 3457.542
```


```
> # color by price and carat
> aggregate(x=price, by=list(color), FUN=mean)
  Group.1        x
1       D 3169.954
2       E 3076.752
3       F 3724.886
4       G 3999.136
5       H 4486.669
6       I 5091.875
7       J 5323.818
```


The Premium Cut diamonds have the highest average price (odd!)

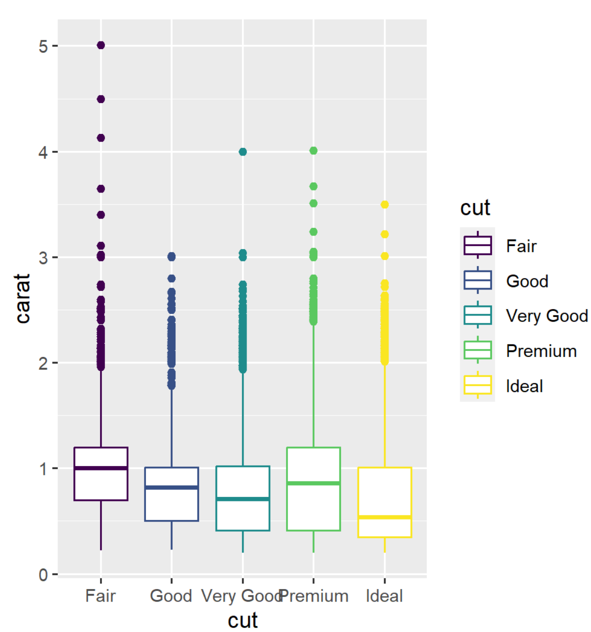
The color by Price vs Carat there is a slight difference. Mean price actually goes down from D to E before going up E to J. For Carat, the size just goes up the whole way. I believe there are people who want perfect no-yellow diamonds more than they want a bigger diamond (since D’s are usually 0.65 while J’s are usually 1.16). This means there is an odd increase in price from E to D, even though the trend would state average price for D is less than E.

2c

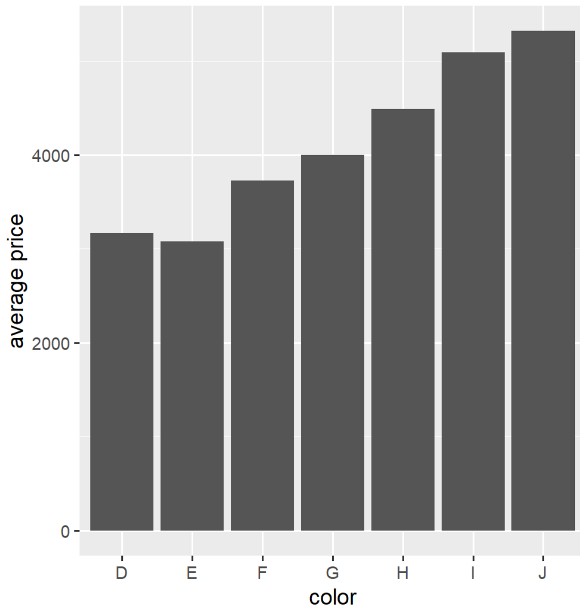


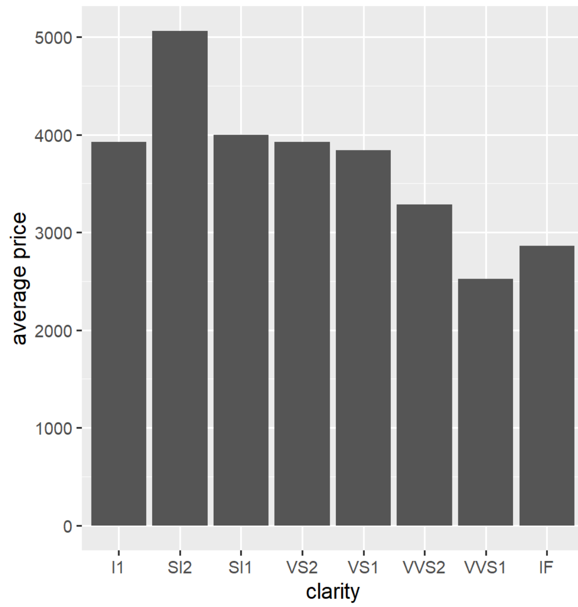
This actually tells me a lot about what kinds of diamonds are Premium vs Ideal. I believe large-carat diamonds are scary to cut, so they just decide to go for a premium cut vs an idea one. If they go for an ideal cut and mess up, a 2-carat diamond might hit 1.98 and now its useless. So they go for premium.

This is also shown in the plot where the top line of the box plot price goes much higher for premium vs ideal. To back up my theory, here's a box plot of carat vs cut. Avg carat for Premium is higher than Ideal

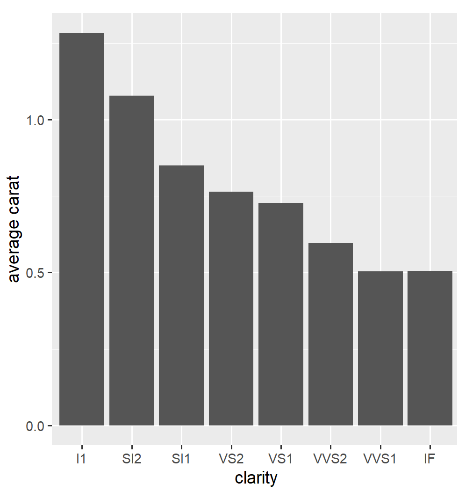


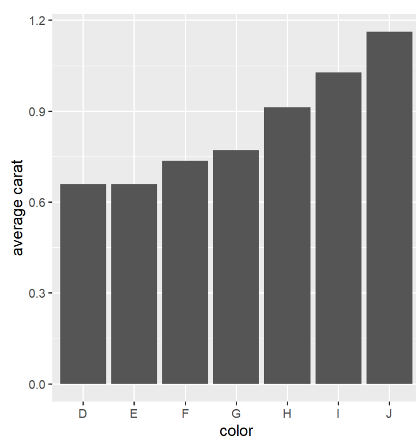
2c continued





The average price vs color trend I noticed in 2b is still here. Clarity is more interesting. Near-perfect diamonds (I1) are cheaper than the grade below (SI2). I again believe this has to do with carat quality. I think SI2 will be higher average carat size vs I1. To check if this is true lets make a new plot:

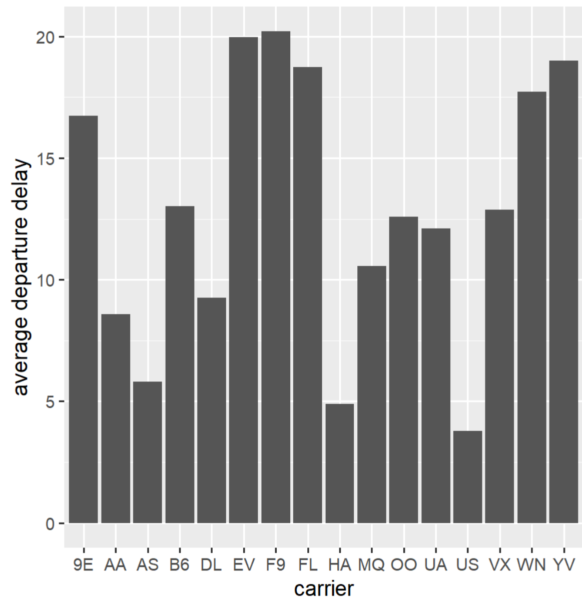




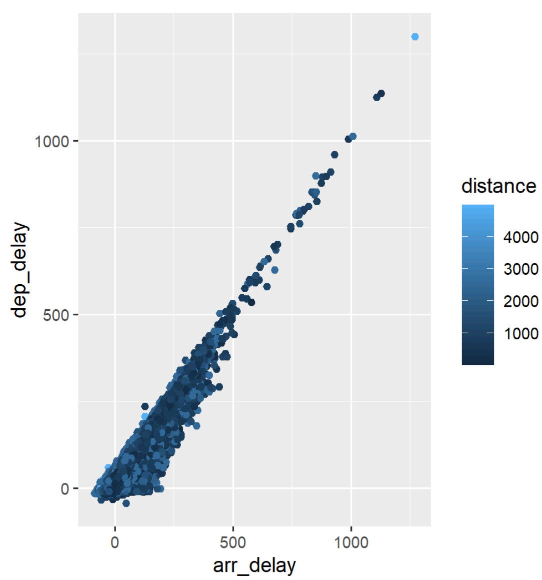
I was proven wrong!! This surprises me. My theory on worse color = higher carat was true though!

Lab2 part 2 - NYCFlights13

3a

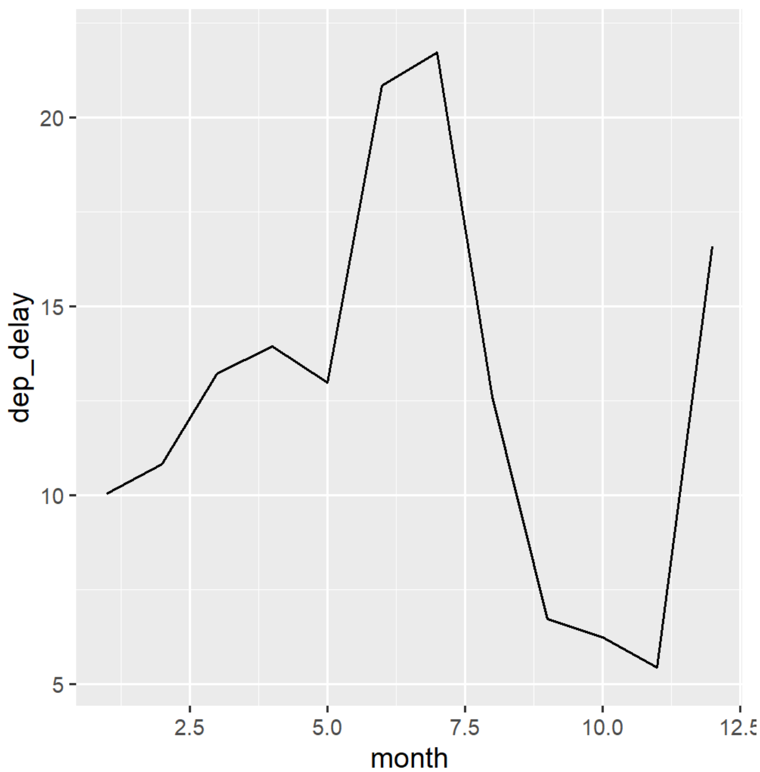


F9 has the highest delay. This is Frontier Airlines.



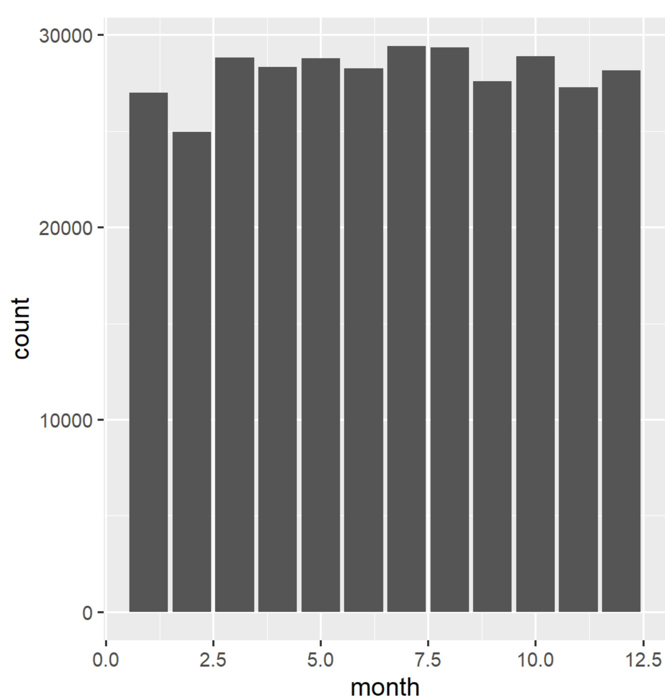
If you depart late, you’re probably going to arrive late. The correlation of this is very, very strong. That being said, there are some times you depart on time but arrive late. There is rarely a moment where you depart late but arrive early. You would have to be a speeding pilot to do that.

3b

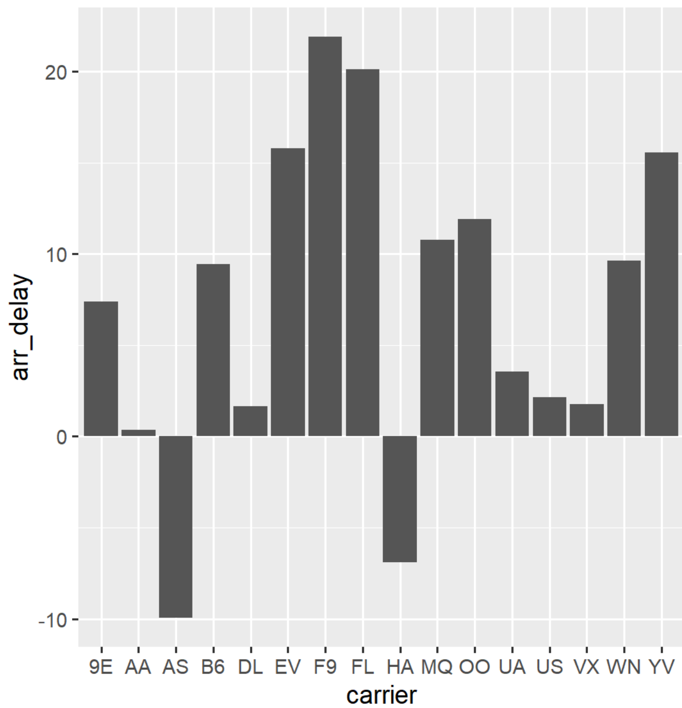


Departures have high delays between June-July. This could be weather related (hurricane season?). Also there is a jump in delay in December. This is most likely the holiday rush!

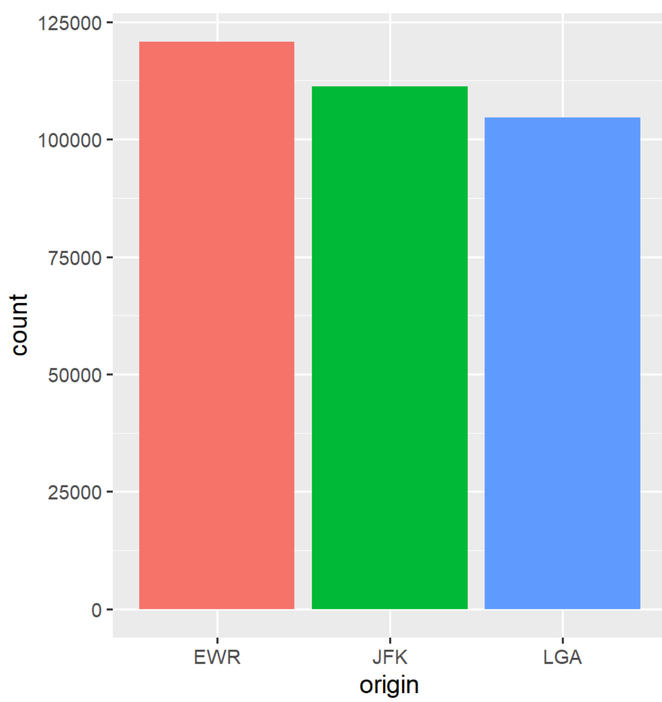
February has the lowest number of flights
July has the highest amount of flights



3c



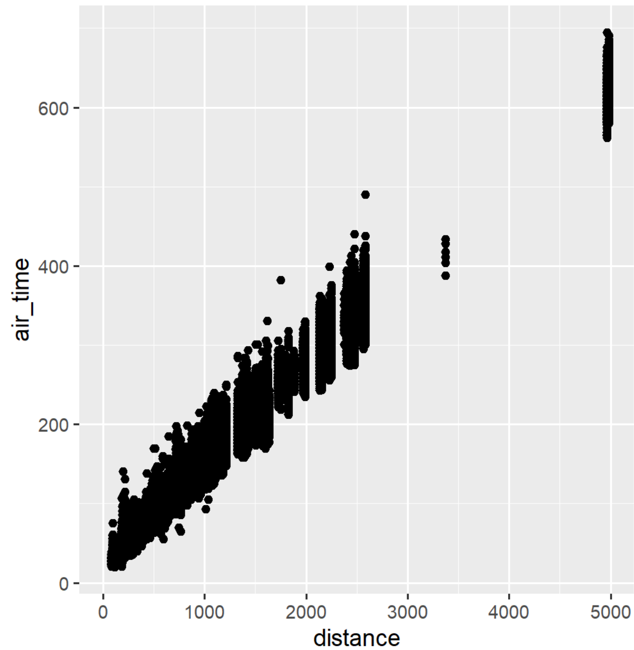
20
10
0
-10
arr_delay
9E AA AS B6 DL EV F9 FL HA MQ OO UA US VX WN YV
carrier



125000
100000
75000
50000
25000
0
count
EWR JFK LGA
origin

3d

```
> summary(lmDist)

Call:
lm(formula = flights$distance ~ flights$air_time)

Residuals:
     Min       1Q   Median       3Q      Max 
-1102.71   -52.15     4.71    53.57   714.03 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      -1.242e+02  3.323e-01  -373.7   <2e-16 ***
flights$air_time  7.781e+00  1.873e-03  4154.4   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 100.4 on 327344 degrees of freedom
  (9430 observations deleted due to missingness)
Multiple R-squared:  0.9814,	Adjusted R-squared:  0.9814 
F-statistic: 1.726e+07 on 1 and 327344 DF,  p-value: < 2.2e-16
```



The plot (and model) show there is a high correlation between the distance of the flight and the duration of the flight

HA has the highest airtime (Hawaiian Airlines Inc.)

YV has the lowest airtime (Mesa Airlines Inc.)