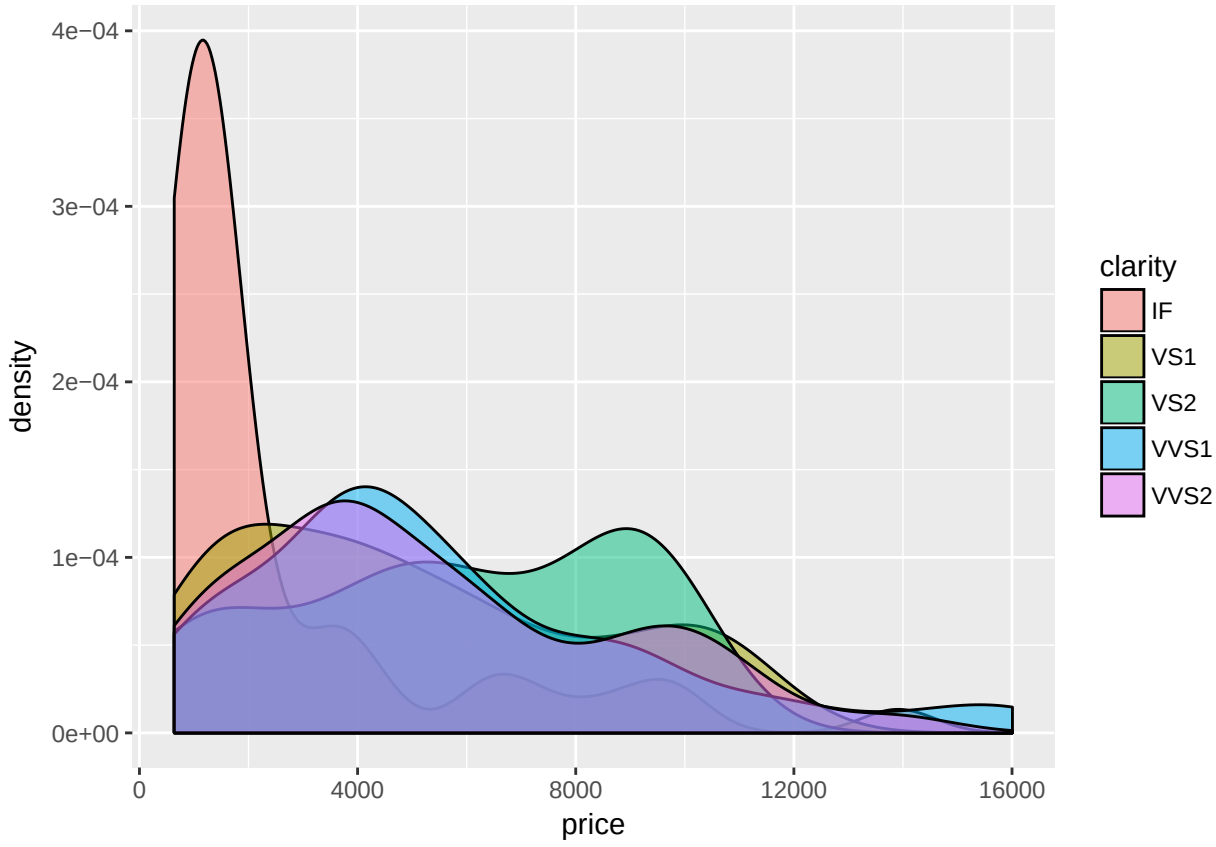
Homework 4 Solutions

```
suppressPackageStartupMessages(library(dplyr))
suppressPackageStartupMessages(library(ggplot2))
```

Exercise 1

```
diam <- read.csv("https://www.macalester.edu/~ajohns24/data/Diamonds.csv")
```

```
#a
ggplot(diam, aes(x=price, fill=clarity)) +
    geom_density(alpha=0.5)
```




```
#b
mod1 <- lm(price ~ clarity, diam)
summary(mod1)
```

```
##
## Call:
## lm(formula = price ~ clarity, data = diam)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5220.2 -1940.0  -990.9  2063.5 11218.2
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2694.8      494.2   5.453 1.03e-07 ***
## clarityVS1    2362.3      613.9   3.848 0.000145 ***
## clarityVS2    3163.4      668.5   4.732 3.42e-06 ***
## clarityVVS1   2872.9      671.4   4.279 2.52e-05 ***
## clarityVVS2   2661.8      618.0   4.307 2.24e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3278 on 303 degrees of freedom
## Multiple R-squared:  0.08428,    Adjusted R-squared:  0.0722
## F-statistic: 6.972 on 4 and 303 DF,  p-value: 2.216e-05
```

```
#c
#smallest price = IF
2694.8
```

```
## [1] 2694.8
```

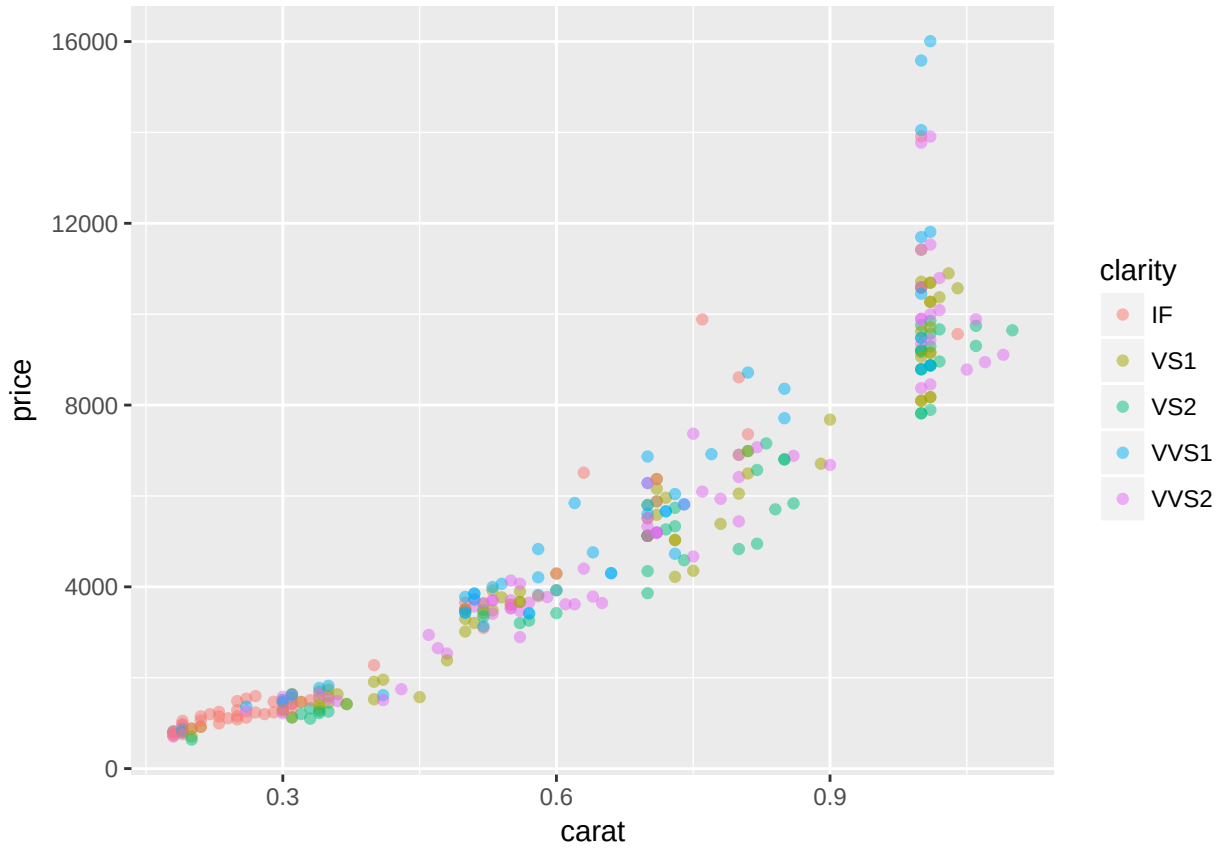
```
#highest price = VS2
2694.8 + 3163.4
```

```
## [1] 5858.2
```

```
#d
#The avg price for all other clarity levels (VS1,VS2,VVS1,VVS2)
#is sig. higher than for IF diamonds
```

Exercise 2

```
#a
ggplot(diam, aes(y=price, x=carat, color=clarity)) +
    geom_point(alpha=0.5)
```



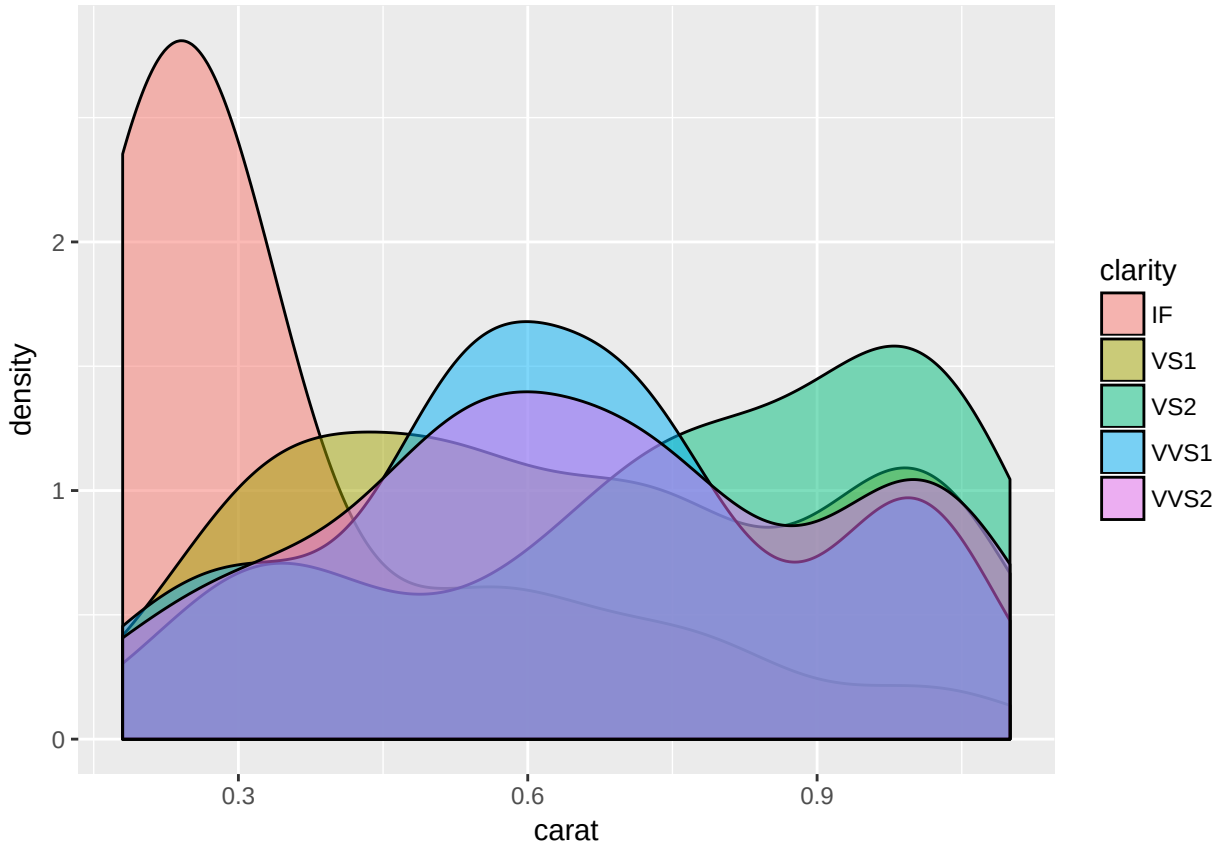

```
#b
mod2 <- lm(price ~ clarity + carat, diam)
summary(mod2)
```

```
##
## Call:
## lm(formula = price ~ clarity + carat, data = diam)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1962.4  -584.1   -62.5   434.8  5914.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1851.2      177.5 -10.429  < 2e-16 ***
## clarityVS1   -1001.2      203.0  -4.932 1.35e-06 ***
## clarityVS2   -1561.9      228.2  -6.844 4.30e-11 ***
## clarityVVS1   -403.7      219.7  -1.837   0.0672 .
## clarityVVS2   -958.8      205.8  -4.659 4.78e-06 ***
## carat        12226.4      232.2  52.664  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1029 on 302 degrees of freedom
## Multiple R-squared:  0.9101, Adjusted R-squared:  0.9086
## F-statistic: 611.3 on 5 and 302 DF,  p-value: < 2.2e-16
```

```
#c
#practice

#d
#highest = IF (all other clarity coefs are negative)
#lowest = VS2 - on average it gets $1561.9 less than an IF diamond
#of the same size

#e
#IF diamonds tend to be small (the bigger the diamond the more room for flaws)
#and small diamonds get less money
ggplot(diam, aes(x=carat, fill=clarity)) +
    geom_density(alpha=0.5)
```



Exercise 3

```
foods <- read.csv("https://www.macalester.edu/~ajohns24/Data/BabyBoyFood.csv")
head(foods)
```

```
##                                                    Type
## 1 Beef, ground, regular, pan-cooked
## 2 Shredded wheat cereal
## 3 BF, cereal, rice w/apples, dry, prep w/ water
## 4 BF, sweet potatoes
## 5 Pear, raw (w/ peel)
## 6 Celery, raw
```

```r
#a
babies <- data.frame(sex=c("male","female"))
set.seed(2017)
births <- data.frame(nmales=rep(0,132), nfemales=rep(0,132))
for(i in 1:132){
    samp <- sample_n(babies, size=50, replace=TRUE)
    births[i,] <- table(samp)
}
births$food <- foods$Type

#b
births %>% filter(nmales < 18)
```

```
##   nmales nfemales
## 1     17       33
## 2     17       33
##                                                                food
## 1 BF, applesauce
## 2 Tomato, raw
```

```r
births %>% filter(nfemales < 18)
```

```
##   nmales nfemales
## 1     33       17
##                                                                food
## 1 Brown gravy, canned or bottled
```

```r
#c
#Conclude that a food is linked to birth sex when it's not.

#d
set.seed(2017)
pvals <- rep(0,132)
for(i in 1:132){
    nmales <- births$nmales[i]
    pvals[i] = prop.test(x=nmales, n=50)$p.value
}
births$pvals <- pvals

#e
births %>% filter(pvals < 0.05)
```

```
##   nmales nfemales
## 1     17       33
## 2     33       17
## 3     17       33
##                                                                food
## 1 BF, applesauce
## 2 Brown gravy, canned or bottled
## 3 Tomato, raw
##        pvals
## 1 0.03389485
## 2 0.03389485
## 3 0.03389485
```

Exercise 4

a. 0.05

b. $1 - 0.95^2 = 0.0975$

c. $1 - 0.95^{100} = 0.9940795$

d. .

```
births %>% filter(pvals*132 < 0.05)
```

```
## [1] nmales   nfemales food     pvals
## <0 rows> (or 0-length row.names)
```

e. By making it harder to make Type I errors, we increase the probability of making Type II errors (missing real results).

Exercise 5

```
bf <- read.csv("https://www.macalester.edu/~ajohns24/data/bodyfatsub.csv")
bfmod <- lm(BodyFat ~ Hip + Weight, bf)
summary(bfmod)
```

```
##
## Call:
## lm(formula = BodyFat ~ Hip + Weight, data = bf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.1981  -4.1935  -0.1629   4.5087  18.4368
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -48.52400   10.85559  -4.470 1.19e-05 ***
## Hip           0.56325    0.17501   3.218  0.00146 **
## Weight        0.06417    0.04171   1.538  0.12525
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.418 on 247 degrees of freedom
## Multiple R-squared:  0.406,  Adjusted R-squared:  0.4012
## F-statistic:  84.4 on 2 and 247 DF,  p-value: < 2.2e-16
```

a. `Weight`: When holding `Hip` constant, there's a 0.06 percentage point increase in body fat percent for every extra 1 lb of weight.

b. p = 0.12525.

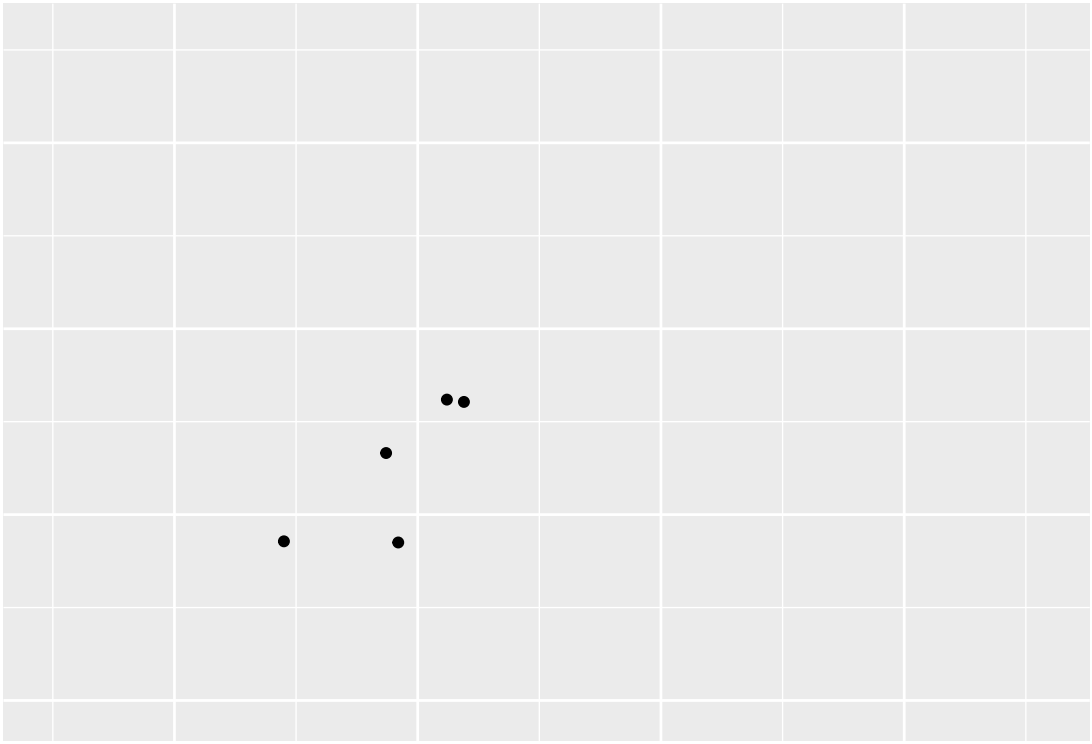
```
bfmod2 <- lm(BodyFat ~ Weight, bf)
summary(bfmod2)
```

```
##
## Call:
## lm(formula = BodyFat ~ Weight, data = bf)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -18.412  -4.744  -0.023   4.950  20.720
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -14.69314    2.76045  -5.323 2.29e-07 ***
## Weight        0.18938    0.01533  12.357  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.538 on 248 degrees of freedom
## Multiple R-squared:  0.3811, Adjusted R-squared:  0.3786
## F-statistic: 152.7 on 1 and 248 DF,  p-value: < 2.2e-16
```

c. Weight and hip circumference are *multicollinear*. On top of what's already explained by hip about body fat, weight doesn't add a significant amount of information:

```
ggplot(bf, aes(x=Hip,y=Weight)) +
    geom_point()
```



Exercise 6

```
Rides <- read.csv("https://www.macalester.edu/~ajohns24/Data/NiceRide2016sub.csv")
dim(Rides)
```

```
## [1] 40000     8
```

```
head(Rides, 3)
```

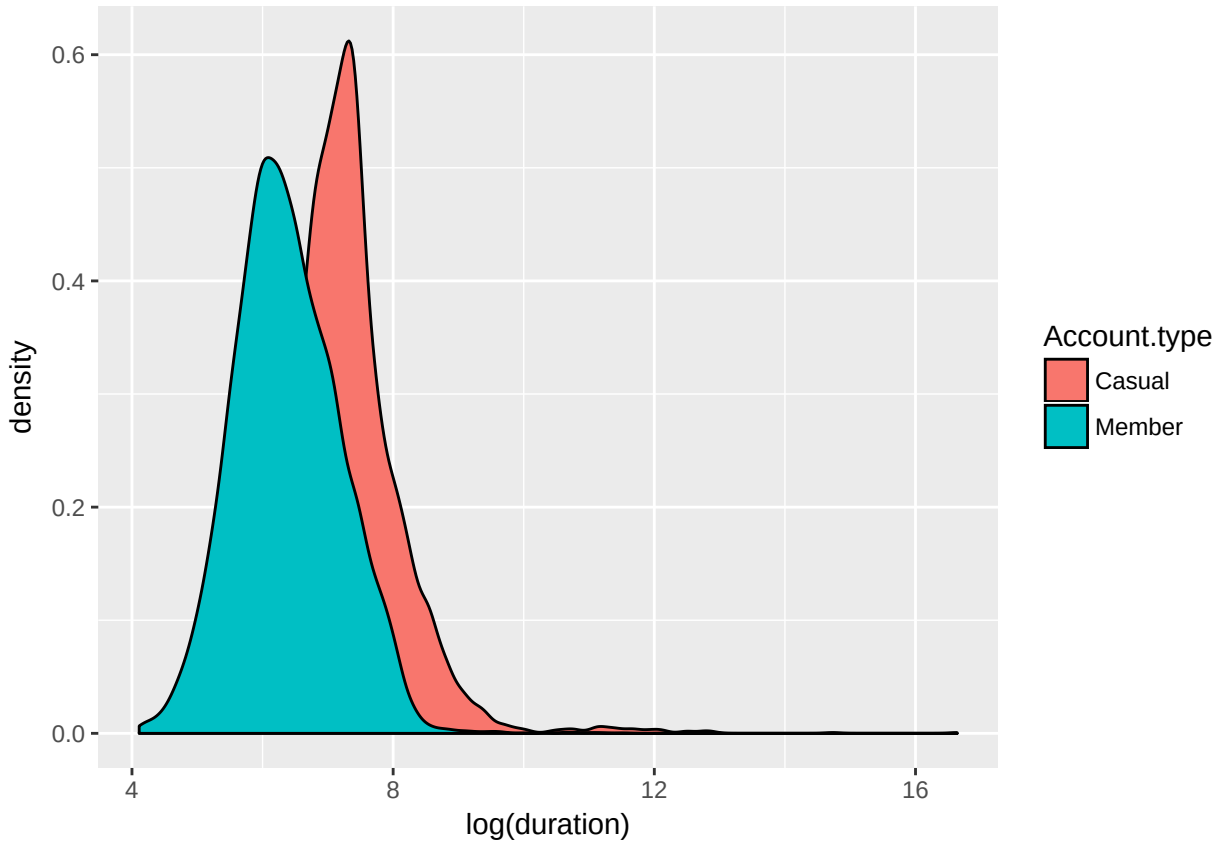
```
##        Start.date             Start.station Start.station.number
## 1
## 2
## 3 8/21/2016 11:49 15th Ave SE & Como Ave SE                30110
##          End.date              End.station End.station.number
## 1
## 2
## 3 8/21/2016 11:55 4th Street & 13th Ave SE              30009
##   Total.duration..seconds. Account.type
## 1                        0
## 2                        0
## 3                      343       Member
```

```
suppressPackageStartupMessages(library(lubridate))


Rides <- Rides %>%
    rename(duration="Total.duration..seconds.") %>%
    select(c(Start.date, Start.station, End.station, duration, Account.type)) %>%
    filter(duration > 0) %>%
    mutate(hours=as.factor(hour(mdy_hm(Start.date))), months=as.factor(month(mdy_hm(Start.date))))
```

Exercise 6

```
#a
ggplot(Rides, aes(x=log(duration), fill=Account.type)) +
    geom_density()
```




```
#b
summary(lm(log(duration) ~ Account.type, Rides))
```

```
##
## Call:
## lm(formula = log(duration) ~ Account.type, data = Rides)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0957 -0.5390 -0.0505  0.4954  9.4142
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         7.22288    0.01108  651.93   <2e-16 ***
## Account.typeMember -0.84616    0.01373  -61.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8373 on 16376 degrees of freedom
## Multiple R-squared:  0.1883, Adjusted R-squared:  0.1883
## F-statistic:  3799 on 1 and 16376 DF,  p-value: < 2.2e-16
```

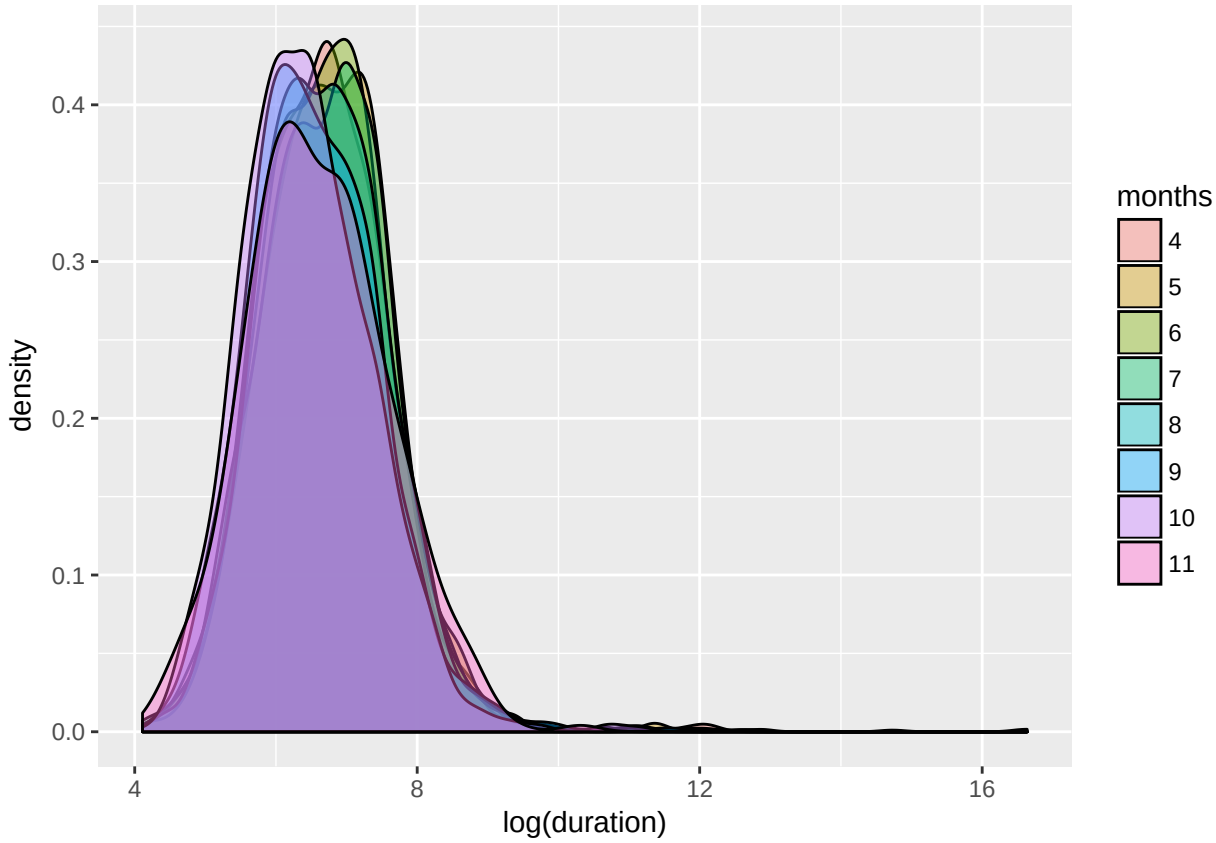
```
#interpretation:
#the typical duration of a member's ride is 13 min
#shorter than the typ. duration of a non-member's ride
(exp(7.22288) - exp(7.22288-0.84616))/60
```

```
## [1] 13.04057
```

```
#c
#yes. the p-value for the Account.typeMember coef is < 0.05
```

Exercise 7

```
#a
ggplot(Rides, aes(x=log(duration), fill=months)) +
    geom_density(alpha=0.4)
```




```
#b
summary(lm(log(duration) ~ months, Rides))
```

```
##
## Call:
## lm(formula = log(duration) ~ months, data = Rides)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6494 -0.6422 -0.0228  0.5857  9.9506
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.686459   0.025855 258.614  < 2e-16 ***
## months5      0.073814   0.032084   2.301 0.021423 *
## months6      0.026429   0.031286   0.845 0.398272
## months7      0.047594   0.030611   1.555 0.120016
```

```
## months8      0.006052   0.031488   0.192 0.847579
## months9     -0.119706   0.032645  -3.667 0.000246 ***
## months10    -0.227514   0.034588  -6.578 4.92e-11 ***
## months11    -0.081611   0.055184  -1.479 0.139191
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.925 on 16370 degrees of freedom
## Multiple R-squared:  0.009602,   Adjusted R-squared:  0.009178
## F-statistic: 22.67 on 7 and 16370 DF,  p-value: < 2.2e-16
```

```
#interpretation:
#the typical duration of a ride in May is 1 min
#longer than the typ. duration of a ride in April
(exp(6.686459) - exp(6.686459+0.073814))/60
```

```
## [1] -1.023309
```

```
#c
#yes, the p-value for `months5` < 0.05
```

Exercise 8

a. sample size. when n is large, standard error is small, thus even small, unmeaningful e□ect sizes are statistically significant

Exercise 9

```
#a
Stations <- read.csv("https://www.macalester.edu/~ajohns24/Data/NiceRideStations.csv")

#join the Stations and Rides
MergedRides <- Rides %>%
    left_join(Stations, by=c(Start.station = "Station")) %>%
    rename(start_lat=Latitude, start_long=Longitude) %>%
    left_join(Stations, by=c(End.station = "Station")) %>%
    rename(end_lat=Latitude, end_long=Longitude)
```

```
## Warning: Column `Start.station`/`Station` joining factors with different
## levels, coercing to character vector
```

```
## Warning: Column `End.station`/`Station` joining factors with different
## levels, coercing to character vector
```

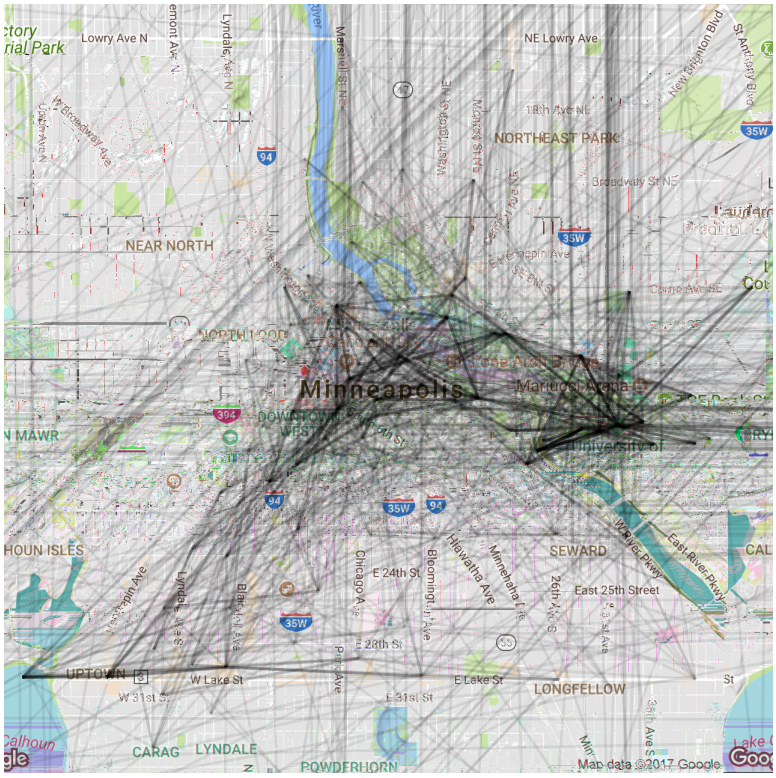
```
#b
suppressPackageStartupMessages(library(ggmap))
MN <- get_map("Minneapolis", zoom=13)
```

```
## Source : https://maps.googleapis.com/maps/api/staticmap?center=Minneapolis&zoom=13&size=640x640&scal
```

```
## Source : https://maps.googleapis.com/maps/api/geocode/json?address=Minneapolis
```

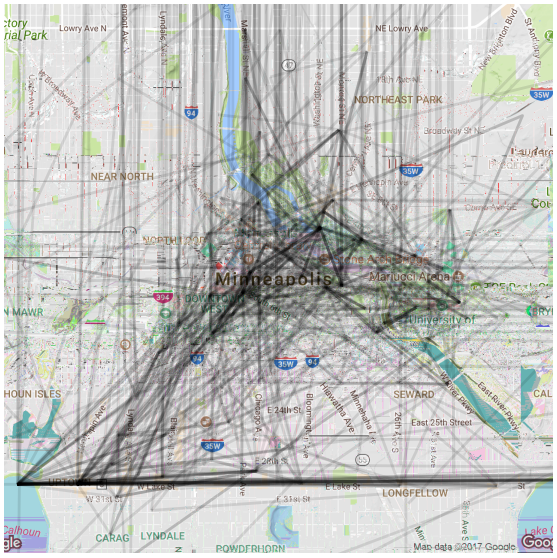
```
ggmap(MN) +
    geom_segment(data=MergedRides, aes(x=start_long, y=start_lat,
    xend=end_long, yend=end_lat), alpha=0.07)
```

```
## Warning: Removed 3366 rows containing missing values (geom_segment).
```




```
#c
ggmap(MN) +
    geom_segment(data=MergedRides, aes(x=start_long, y=start_lat,
    xend=end_long, yend=end_lat), alpha=0.1) +
    facet_wrap(~Account.type)
```

```
## Warning: Removed 3366 rows containing missing values (geom_segment).
```




```
#d (incomplete answer)
ggmap(MN) +
    geom_segment(data=MergedRides, aes(x=start_long, y=start_lat,
    xend=end_long, yend=end_lat), alpha=0.12) +
    facet_wrap(~hours)
```

```
## Warning: Removed 3366 rows containing missing values (geom_segment).
```