Project 1

Caroline Cweren

3/1/2020

```
#Data sets
library(dplyr)
library(ggplot2)
Fertility1 <- read.csv("Fertility1 - Sheet1.csv")
Fertility2 <- read.csv("Fertility2 - Sheet1.csv")

Fertility1<-Fertility1%>%mutate(ID=1:333)
glimpse(Fertility1)
```

```
## Observations: 333
## Variables: 7
## $ Age     <int> 40, 37, 40, 40, 30, 29, 31, 33, 36, 35, 25, 39, 35, 30, 37,...
## $ LowAFC  <int> 40, 41, 38, 36, 36, 35, 24, 28, 30, 32, 27, 32, 31, 18, 29,...
## $ MeanAFC <dbl> 51.5, 41.0, 41.0, 37.5, 36.0, 35.0, 35.0, 34.0, 33.0, 32.0,...
## $ FSH     <dbl> 5.3, 7.1, 4.9, 3.9, 4.0, 3.9, 3.8, 4.3, 4.9, 3.7, 5.0, 5.3,...
## $ E2      <int> 45, 53, 40, 26, 49, 67, 49, 20, 60, 36, 20, 37, 30, 33, 40,...
## $ MaxE2   <int> 1427, 802, 4533, 1804, 2526, 3812, 1087, 1615, 1879, 2009, ...
## $ ID      <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, ...
```

```
Fertility2<-Fertility2%>%mutate(ID=1:333)%>%select(-c(MaxE2))
glimpse(Fertility2)
```

```
## Observations: 333
## Variables: 5
## $ MaxDailyGn <dbl> 300.0, 225.0, 450.0, 300.0, 150.0, 150.0, 262.5, 375.0, ...
## $ TotalGn    <dbl> 2700.0, 1800.0, 4850.0, 2700.0, 1500.0, 975.0, 2512.5, 3...
## $ Oocytes    <int> 25, 7, 27, 9, 19, 19, 13, 15, 23, 26, 22, 22, 7, 27, 12,...
## $ Embryos    <int> 13, 6, 15, 4, 12, 16, 9, 9, 10, 8, 13, 18, 5, 18, 9, 2, ...
## $ ID         <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 1...
```

Introduction: For this project, I chose to combine two datasets pertaining to fertility in women by the variable, ID. I obtained both datasets from the available Rdatasets on https: //vincentarelbundock.github.io/Rdatasets/datasets.html. The first dataset I chose contained data on variables pertaining to ovarian follicles, such as age, smallest antral follicle count(LowAFC), average antral follicle count (Mean AFC), maximum follicle stimulating hormone level (FSH), and the maximum fertility level (E2). As a pre-med student particularly interested in becoming a pediatritian or OBG-YN, I thought it would provide me with useful knowledge and be really interesting to analyze the factors that contribute to a woman's fertility. For example, the antral follicle count is a very important measurement that counts the number of eggs containing follicles,fliud-filled sacs that contain immature eggs, developing in a woman's ovaries. FSH variable is important for stimulating the growth in follicles before an egg can be release during ovuilation.

The second dataset contained data on maximum fertility level (MaxE2), maximum daily gonadtropin level (MaxDailyGn), total gonadotropin level (TotalGn), number of oocytes, and number of embryos. The variable of gonadtropin hormone level is necessary for the stimulation of the female gonads. The variable of number of oocytes, immature eggs, which are later fertilized to form an embryo. All of these variables play are dependent on one another and play a vital role in fertility. Based in this, I thought it would be interested to compare important variables, such as the correlation between FSH levels and the number of viable embryos, that effect fertility in women. I am expecting to see a positive correlation between the variables from the Fertility 1 dataset and the Fertility 2 dataset.

```
#Tidying
library(tidyverse)
Fertility_join<-Fertility1%>%full_join(Fertility2, by="ID")%>%glimpse()
```

```
## Observations: 333
## Variables: 11
## $ Age        <int> 40, 37, 40, 40, 30, 29, 31, 33, 36, 35, 25, 39, 35, 30, ...
## $ LowAFC     <int> 40, 41, 38, 36, 36, 35, 24, 28, 30, 32, 27, 32, 31, 18, ...
## $ MeanAFC    <dbl> 51.5, 41.0, 41.0, 37.5, 36.0, 35.0, 35.0, 34.0, 33.0, 32...
## $ FSH        <dbl> 5.3, 7.1, 4.9, 3.9, 4.0, 3.9, 3.8, 4.3, 4.9, 3.7, 5.0, 5...
## $ E2         <int> 45, 53, 40, 26, 49, 67, 49, 20, 60, 36, 20, 37, 30, 33, ...
## $ MaxE2      <int> 1427, 802, 4533, 1804, 2526, 3812, 1087, 1615, 1879, 200...
## $ ID         <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 1...
## $ MaxDailyGn <dbl> 300.0, 225.0, 450.0, 300.0, 150.0, 150.0, 262.5, 375.0, ...
## $ TotalGn    <dbl> 2700.0, 1800.0, 4850.0, 2700.0, 1500.0, 975.0, 2512.5, 3...
## $ Oocytes    <int> 25, 7, 27, 9, 19, 19, 13, 15, 23, 26, 22, 22, 7, 27, 12,...
## $ Embryos    <int> 13, 6, 15, 4, 12, 16, 9, 9, 10, 8, 13, 18, 5, 18, 9, 2, ...
```

```
Mutatedata<- Fertility_join%>% mutate(Age_Group = case_when(Age>40 ~ "Forties",30<=Age & Age<=40 ~ "Thi

Tidydata<-Mutatedata%>% group_by(Age_Group) %>% summarize(mean_oocytes=mean(Oocytes), sd_FSH=sd(FSH), c

untidyfert<-Tidydata%>% pivot_wider(names_from = "Age_Group", values_from = "mean_oocytes")
untidyfert
```

```
## # A tibble: 3 x 6
##   sd_FSH count se_FSH Forties Thirties Twenties
##    <dbl> <int>  <dbl>   <dbl>    <dbl>    <dbl>
## 1   2.34    50  0.330     9.3     NA       NA
## 2   1.84   248  0.117    NA       12.3     NA
## 3   1.37    35  0.231    NA       NA       11.9
```

```
tidyfert <-untidyfert%>%pivot_longer(c("Forties","Thirties", "Twenties"), names_to="Age_Group", values_
tidyfert
```

```
## # A tibble: 3 x 5
##   sd_FSH count se_FSH Age_Group mean_Oocytes
##    <dbl> <int>  <dbl> <chr>            <dbl>
## 1   2.34    50  0.330 Forties            9.3
## 2   1.84   248  0.117 Thirties          12.3
## 3   1.37    35  0.231 Twenties          11.9
```

Initially both of my datasets were tidy; however, to demonstrate my tidying skills I used pivot_wider to untidy my summary statistics of mean oocytes, the standard deviation of FSH, and the standard error of

FSH. The summary statistics were taken from my joined dataset, and I converted the numeric "Age" variable to become a categorical "Age_Group" variable in order to group the data by the three age groups of women feritility, "twenties", "thirties","forties". Next, I used pivot_longer to tidy my dataset of the summary statistics. I named the newly tidy dataset "tidyfert".

```
#Joining
Fertility_join<-Fertility1%>%full_join(Fertility2, by="ID")%>%glimpse()
```

```
## Observations: 333
## Variables: 11
## $ Age         <int> 40, 37, 40, 40, 30, 29, 31, 33, 36, 35, 25, 39, 35, 30, ...
## $ LowAFC      <int> 40, 41, 38, 36, 36, 35, 24, 28, 30, 32, 27, 32, 31, 18, ...
## $ MeanAFC     <dbl> 51.5, 41.0, 41.0, 37.5, 36.0, 35.0, 35.0, 34.0, 33.0, 32...
## $ FSH         <dbl> 5.3, 7.1, 4.9, 3.9, 4.0, 3.9, 3.8, 4.3, 4.9, 3.7, 5.0, 5...
## $ E2          <int> 45, 53, 40, 26, 49, 67, 49, 20, 60, 36, 20, 37, 30, 33, ...
## $ MaxE2       <int> 1427, 802, 4533, 1804, 2526, 3812, 1087, 1615, 1879, 200...
## $ ID          <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 1...
## $ MaxDailyGn <dbl> 300.0, 225.0, 450.0, 300.0, 150.0, 150.0, 262.5, 375.0, ...
## $ TotalGn     <dbl> 2700.0, 1800.0, 4850.0, 2700.0, 1500.0, 975.0, 2512.5, 3...
## $ Oocytes     <int> 25, 7, 27, 9, 19, 19, 13, 15, 23, 26, 22, 22, 7, 27, 12,...
## $ Embryos     <int> 13, 6, 15, 4, 12, 16, 9, 9, 10, 8, 13, 18, 5, 18, 9, 2, ...
```

First, I had to match the sample number of each woman observed in each datasets by mutating a new column in both datasets labeled as "ID" for each observation. Next, I used the select function to delete an extraneous column of MaxE2 in the second dataset.These alterations are shown in the section of my importation of the datasets above. After cleaning the datasets, I decided to join the two fertility datasets by using full_join by "ID". By using full_join, I was able to retain all all rows and columns of both datasets. After joining, my new dataset contained all 11 unique variables. I thought that both datasets contained equally valuable data pertaining to different variables of fertility. For this reason, I did not use a right_join or left_join that would prioritize just one of the datasets. The completely newly joined dataset was named "Fertility_join".

```
#Wrangling
Fertility_join%>%filter(Age=="37")
```

```
##   Age LowAFC MeanAFC FSH E2 MaxE2  ID MaxDailyGn TotalGn Oocytes Embryos
## 1  37     41      41 7.1 53   802   2        225  1800.0       7       6
## 2  37     29      29 4.9 40  1840  15        225  1425.0      12       9
## 3  37     20      28 5.2 47   692  18        225  1800.0       9       4
## 4  37     18      27 5.7 46  2427  19        225  2300.0      25      15
## 5  37     22      22 3.6 19  1869  39        225  1125.0      20      10
## 6  37     14      18 6.3 34  1116  76        450  4050.0       7       5
## 7  37     16      17 3.6 31  1357  79        225  1650.0      10       8
## 8  37     16      16 6.3 53  2737  90        150  1162.5      17       5
## 9  37     13      15 3.7 27  1956 110        300  2700.0      14      13
##  [ reached 'max' / getOption("max.print") -- omitted 11 rows ]
```

```
Fertility_join %>% select(MaxE2, Embryos)%>%glimpse()
```

```
## Observations: 333
## Variables: 2
## $ MaxE2   <int> 1427, 802, 4533, 1804, 2526, 3812, 1087, 1615, 1879, 2009, ...
## $ Embryos <int> 13, 6, 15, 4, 12, 16, 9, 9, 10, 8, 13, 18, 5, 18, 9, 2, 8, ...
```

```
Fertility_join %>% arrange( desc(Age))%>%glimpse()
```

```
## Observations: 333
## Variables: 11
## $ Age        <int> 46, 46, 45, 44, 44, 44, 44, 44, 43, 43, 43, 43, 43, 43, ...
## $ LowAFC     <int> 7, 2, 9, 20, 20, 6, 5, 3, 16, 13, 5, 10, 10, 10, 9, 9, 8...
## $ MeanAFC    <dbl> 7.0, 2.0, 10.0, 20.0, 20.0, 9.0, 5.5, 5.0, 16.0, 13.0, 1...
## $ FSH        <dbl> 3.9, 9.4, 7.6, 3.8, 4.8, 5.4, 9.3, 6.5, 7.6, 5.7, 6.8, 6...
## $ E2         <int> 63, 65, 32, 48, 50, 35, 27, 43, 31, 27, 28, 54, 42, 62, ...
## $ MaxE2      <int> 1541, 1151, 290, 1667, 1554, 408, 1108, 2011, 862, 2091,...
## $ ID         <int> 291, 330, 211, 48, 49, 223, 311, 319, 99, 149, 202, 212,...
## $ MaxDailyGn <dbl> 450, 450, 525, 225, 225, 450, 450, 450, 525, 450, 450, 4...
## $ TotalGn    <dbl> 3600, 4050, 5625, 1800, 2175, 3600, 4500, 5850, 4875, 36...
## $ Oocytes    <int> 7, 7, 5, 15, 19, 7, 7, 6, 6, 12, 12, 5, 13, 7, 7, 6, 2, ...
## $ Embryos    <int> 4, 2, 3, 10, 8, 5, 5, 3, 3, 7, 4, 3, 6, 3, 5, 5, 1, 2, 1...
```

```
Fertility_join %>% group_by(Age, Oocytes)%>%summarize(mean_emb=mean(Embryos))%>%glimpse()
```

```
## Observations: 223
## Variables: 3
## Groups: Age [25]
## $ Age      <int> 21, 23, 23, 24, 25, 25, 26, 27, 27, 27, 27, 27, 27, 27, 27...
## $ Oocytes  <int> 11, 9, 12, 13, 12, 22, 7, 2, 7, 10, 11, 12, 13, 15, 18, 5,...
## $ mean_emb <dbl> 5.0, 5.0, 1.0, 7.0, 6.5, 13.0, 6.0, 2.0, 3.0, 7.0, 5.0, 7....
```

```
Fertility_join %>% mutate_if(is.numeric,round)%>%glimpse()
```

```
## Observations: 333
## Variables: 11
## $ Age        <dbl> 40, 37, 40, 40, 30, 29, 31, 33, 36, 35, 25, 39, 35, 30, ...
## $ LowAFC     <dbl> 40, 41, 38, 36, 36, 35, 24, 28, 30, 32, 27, 32, 31, 18, ...
## $ MeanAFC    <dbl> 52, 41, 41, 38, 36, 35, 35, 34, 33, 32, 32, 32, 31, 31, ...
## $ FSH        <dbl> 5, 7, 5, 4, 4, 4, 4, 4, 5, 4, 5, 5, 5, 4, 5, 5, 5, 5, 6,...
## $ E2         <dbl> 45, 53, 40, 26, 49, 67, 49, 20, 60, 36, 20, 37, 30, 33, ...
## $ MaxE2      <dbl> 1427, 802, 4533, 1804, 2526, 3812, 1087, 1615, 1879, 200...
## $ ID         <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 1...
## $ MaxDailyGn <dbl> 300, 225, 450, 300, 150, 150, 262, 375, 300, 225, 125, 3...
## $ TotalGn    <dbl> 2700, 1800, 4850, 2700, 1500, 975, 2512, 3075, 4800, 127...
## $ Oocytes    <dbl> 25, 7, 27, 9, 19, 19, 13, 15, 23, 26, 22, 22, 7, 27, 12,...
## $ Embryos    <dbl> 13, 6, 15, 4, 12, 16, 9, 9, 10, 8, 13, 18, 5, 18, 9, 2, ...
```

```
Fertility_join %>% mutate(`Embryos_pctile` = ntile(Embryos,100))%>%glimpse()
```

```
## Observations: 333
## Variables: 12
## $ Age            <int> 40, 37, 40, 40, 30, 29, 31, 33, 36, 35, 25, 39, 35, ...
## $ LowAFC         <int> 40, 41, 38, 36, 36, 35, 24, 28, 30, 32, 27, 32, 31, ...
## $ MeanAFC        <dbl> 51.5, 41.0, 41.0, 37.5, 36.0, 35.0, 35.0, 34.0, 33.0...
## $ FSH            <dbl> 5.3, 7.1, 4.9, 3.9, 4.0, 3.9, 3.8, 4.3, 4.9, 3.7, 5....
## $ E2             <int> 45, 53, 40, 26, 49, 67, 49, 20, 60, 36, 20, 37, 30, ...
## $ MaxE2          <int> 1427, 802, 4533, 1804, 2526, 3812, 1087, 1615, 1879,...
## $ ID             <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 1...
## $ MaxDailyGn     <dbl> 300.0, 225.0, 450.0, 300.0, 150.0, 150.0, 262.5, 375...
## $ TotalGn        <dbl> 2700.0, 1800.0, 4850.0, 2700.0, 1500.0, 975.0, 2512....
## $ Oocytes        <int> 25, 7, 27, 9, 19, 19, 13, 15, 23, 26, 22, 22, 7, 27,...
## $ Embryos        <int> 13, 6, 15, 4, 12, 16, 9, 9, 10, 8, 13, 18, 5, 18, 9,...
## $ Embryos_pctile <int> 92, 44, 95, 22, 88, 97, 72, 73, 78, 66, 93, 98, 31, ...
```

```
#summary statistics
Fertility_join%>% summarize_all(mean)
```

```
##        Age   LowAFC  MeanAFC      FSH       E2    MaxE2  ID MaxDailyGn  TotalGn
## 1 35.33333 12.28829 13.52943 5.935135 41.24625 1546.102 167   310.7748 2830.797
##    Oocytes  Embryos
## 1 11.83784 6.727727
```

```
Fertility_join%>% summarize_all(sd)
```

```
##        Age   LowAFC  MeanAFC      FSH       E2    MaxE2       ID MaxDailyGn
## 1 4.698115 6.920879 7.427507 1.942827 15.22441 780.8228 96.27305   115.7717
##    TotalGn  Oocytes  Embryos
## 1 1371.787 5.912322 4.081302
```

```
Fertility_join%>% summarize_all(var)
```

```
##        Age   LowAFC  MeanAFC      FSH       E2    MaxE2     ID MaxDailyGn
## 1 22.07229 47.89857 55.16787 3.774575 231.7826 609684.3 9268.5   13403.08
##   TotalGn  Oocytes  Embryos
## 1 1881799 34.95555 16.65702
```

```
Mutatedata<- Fertility_join%>% mutate(Age_Group = case_when(Age>40 ~ "Forties",30<=Age & Age<=40 ~ "Thi

Mutatedata%>% group_by(Age_Group) %>% summarize(mean_oocytes=mean(Oocytes), sd_embryos=sd(Embryos), cou
```

```
## # A tibble: 3 x 7
##   Age_Group mean_oocytes sd_embryos count se_Embryos min_Embryos max_Embryos
##   <chr>            <dbl>      <dbl> <int>      <dbl>       <int>       <int>
## 1 Forties            9.3       3.29    50      0.466           1          16
## 2 Thirties          12.3       4.23   248      0.269           0          23
## 3 Twenties          11.9       3.20    35      0.541           1          16
```

```
Mutatedata%>% group_by(Age_Group) %>% summarize(mean_MaxE2=mean(MaxE2), sd_MaxE2=sd(MaxE2), count=n(),s
```

```
## # A tibble: 3 x 5
##   Age_Group mean_MaxE2 sd_MaxE2 count se_MaxE2
##   <chr>          <dbl>    <dbl> <int>    <dbl>
## 1 Forties        1388.     714.    50    101.
## 2 Thirties       1560.     805.   248     51.1
## 3 Twenties       1669.     674.    35    114.
```

```
Mutatedata%>% group_by(Age_Group) %>% summarize(mean_E2=mean(E2), sd_E2=sd(E2), count=n(),se_E2=sd_E2/s
```

```
## # A tibble: 3 x 5
##   Age_Group mean_E2 sd_E2 count se_E2
##   <chr>       <dbl> <dbl> <int> <dbl>
## 1 Forties      40.4  16.4    50 2.32
## 2 Thirties     41.4  15.2   248 0.963
## 3 Twenties     41.3  14.3    35 2.42
```

```
Mutatedata%>% group_by(Age_Group) %>% summarize(mean_FSH=mean(FSH), sd_FSH=sd(FSH), count=n(),se_FSH=sd
```

```
## # A tibble: 3 x 5
##   Age_Group mean_FSH sd_FSH count  se_FSH
##   <chr>        <dbl>  <dbl> <int>  <dbl>
## 1 Forties       6.85   2.34    50  0.330
## 2 Thirties      5.91   1.84   248  0.117
## 3 Twenties      4.77   1.37    35  0.231
```

```
Mutatedata%>% group_by(Age_Group) %>% summarize(mean_meanAFC=mean(MeanAFC), sd_meanAFC=sd(MeanAFC), cou
```

```
## # A tibble: 3 x 5
##   Age_Group mean_meanAFC sd_meanAFC count se_meanAFC
##   <chr>            <dbl>      <dbl> <int>      <dbl>
## 1 Forties           9.63       4.56    50      0.646
## 2 Thirties         14.0        7.72   248      0.490
## 3 Twenties         15.7        6.83    35      1.15
```

```
Mutatedata%>% group_by(Age_Group) %>% summarize(mean_lowAFC=mean(LowAFC), sd_lowAFC=sd(LowAFC), count=n
```

```
## # A tibble: 3 x 5
##   Age_Group mean_lowAFC sd_lowAFC count se_lowAFC
##   <chr>           <dbl>     <dbl> <int>     <dbl>
## 1 Forties          8.74      4.52    50     0.639
## 2 Thirties        12.6       7.12   248     0.452
## 3 Twenties        14.8       6.60    35     1.12
```

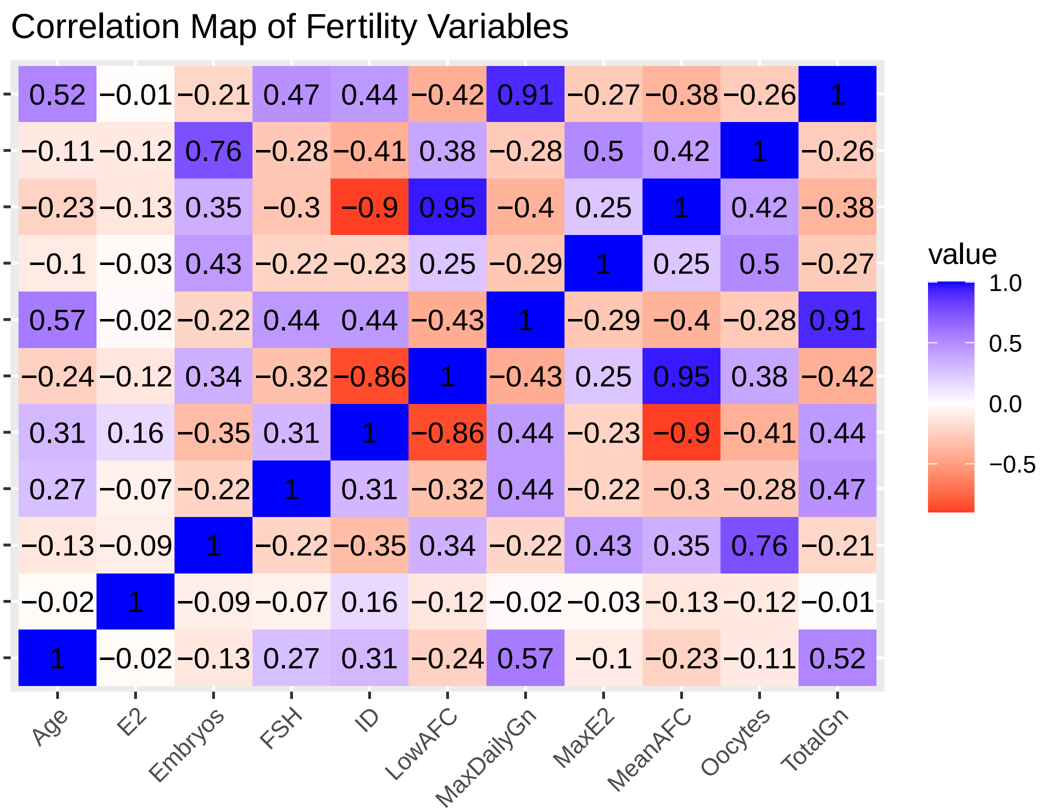
```
Mutatedata%>% group_by(Age_Group) %>% summarize(min_totalGn=min(TotalGn), max_totalGn=max(TotalGn))
```

```
## # A tibble: 3 x 3
##   Age_Group min_totalGn max_totalGn
##   <chr>           <dbl>       <dbl>
## 1 Forties          1800        5850
## 2 Thirties          925        7275
## 3 Twenties          825        4125
```

Within this section, I generated summary statistics. I first used the filter function to view the data of each variable for only individuals of 37 years of age. From this, I was able to see that there were a total of twenty observations from 37 year old women, and 9 out od 20 of those women had a maximum daily gonadtropin level of 255. The next dyplr function I used was the select function to choose data from only the “MaxE2”" and “Embryos” variables. I then used the arrange function to arrange the rows by descending age, starting with the eldest age of 46. After, I was curious to see what the mean number of embryos was based on age and the number of oocytes. In order to do this, I used the group_by function to group the data by the "Age" variable and the "Oocytes" variable and used the summarize function to calculate the mean number of embryos.From this data, I was able to see that the highest mean number of embryos of 18 occured in an individual of age 30 and who contained 27 oocytes. From this, I used the mutate function to generate the percentiles of the "Embryos" variable and determine that an embryo number value of 18 was in the 98th percentile. Additonally, I used the mutate_if function to round all decimal values within the dataset to the nearest whole number.

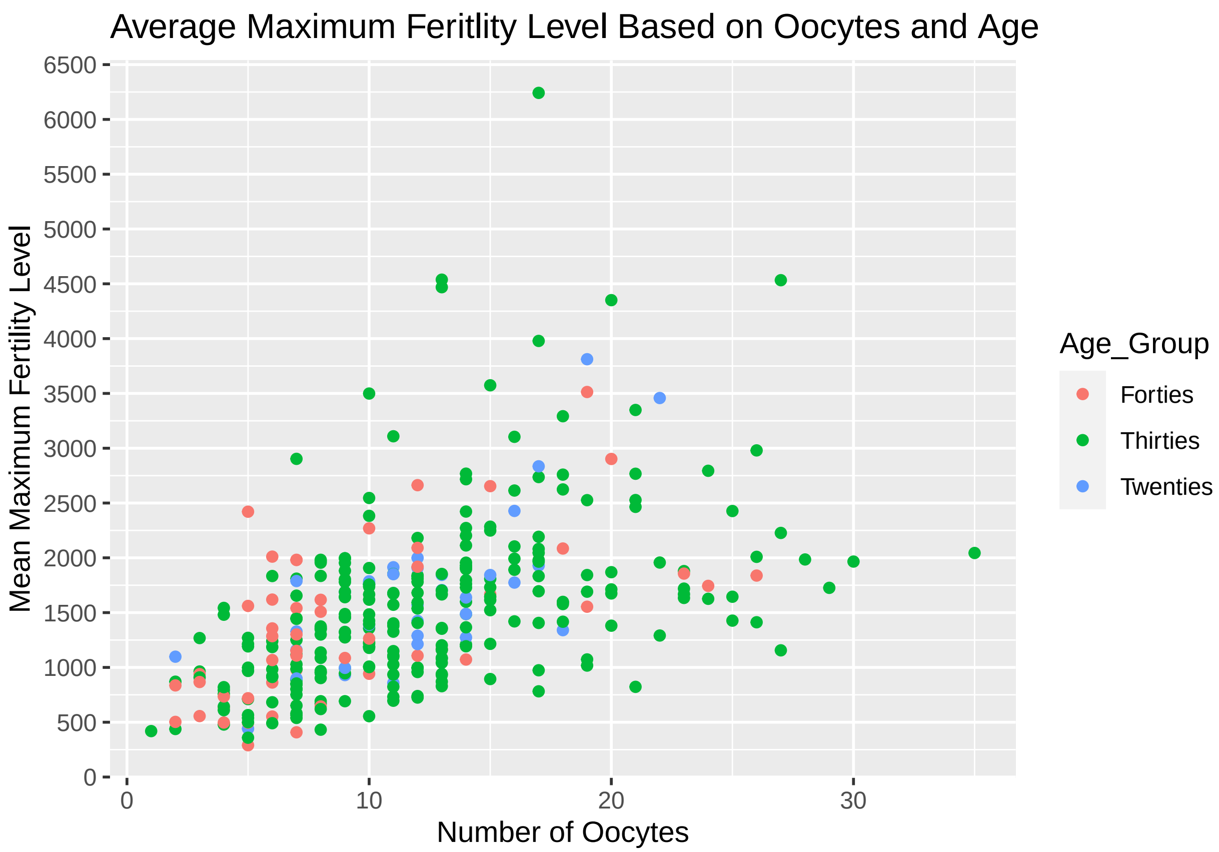
Next, I used the summarize_all function to calculate the mean, standard deviation, and the variance of all numeric variables within my joined dataset. Looking at the "MeanAFC" variable, the mean is 13.52943, the standard deviation is 7.427507, and the varaiance is 55.16787. Additionally, from these statistics, it can be determined that the mean age is about 35.3 years and the deviation from that age value within the group is about 4.69 years difference. Since all my variables were numeric, I used the mutate function to convert my numeric "Age" variable into a categorical variable, called "Age_Group", which sorted the age of women into three categories: "twenties", "thirties", and "forties". Using my new mutated dataframe, I was able to group summary statistics by the "Age_Group" categorical variable. I conducted the the mean of "Oocytes", the standard deviation, the count, the standard error,the minimum, and maximum number of "Embryos" based on each "Age_Group". From this, I determined that the "Thirties" age group had the greatest mean oocytes of 12.34677 and greatest standard deviation of 4.229445 embryos, count of 248 embryos, and the maximum number of embryos of 23. The lowest mimimum number of embryos of 0 was also found in the "Thirties" age group. I continued to do the same summary analyzation of mean, standard deviation, and standard error for each of my other numeric variables. In almost all summary statistics analyzed by age group, the "Thirties" age group had the highest mean values for all numeric variables, except FSH. The highest mean FSH value of 6.848 was found in the "Forties" group. I also computed the maximum and mimum total gonadtropin levels in each "Age_Group" category. The "Thirties" age group had the highest maximum total gonadtrpin level of 7275 and the "Twenties" age group had the lowest minimum total gonadtropin level of 825.

```
#Visualizing
Fertility_join %>% select_if(is.numeric) %>% cor %>% as.data.frame %>% rownames_to_column %>% pivot_long
```

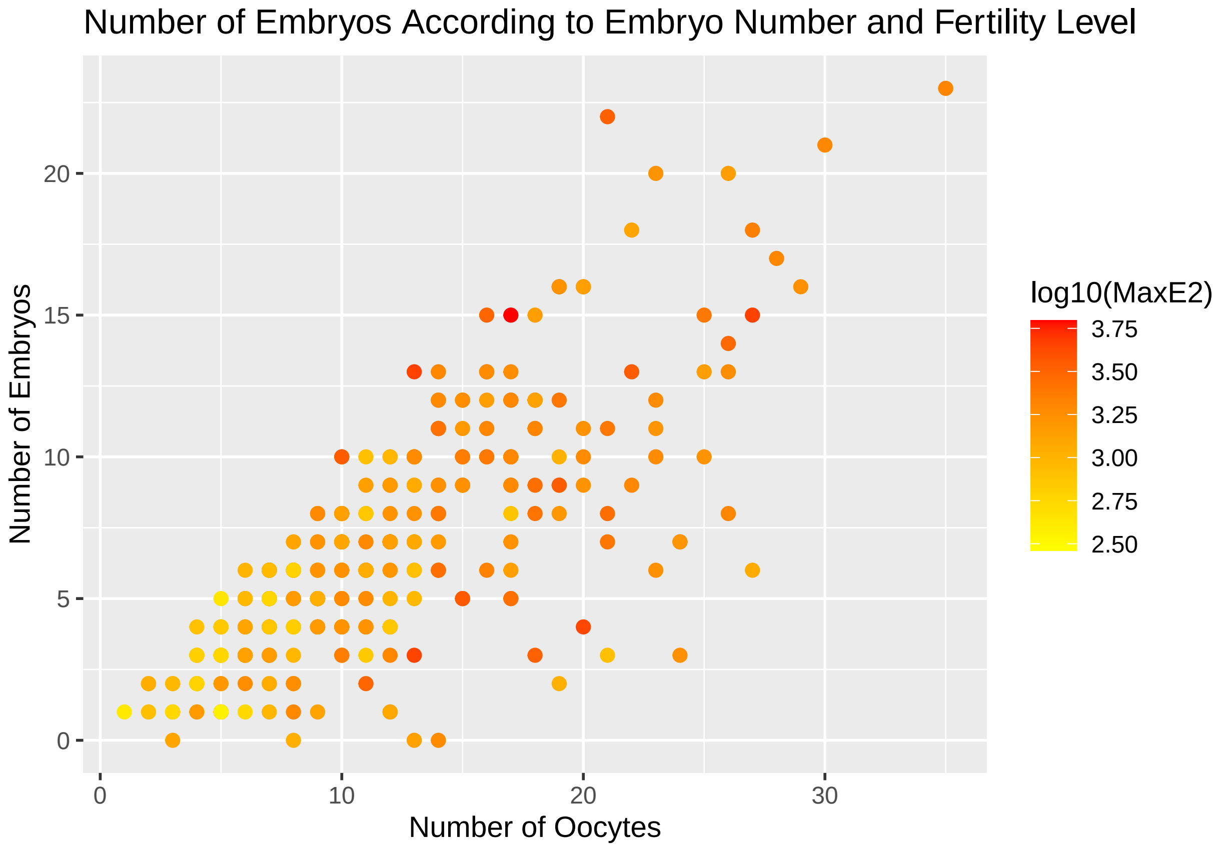

Correlation Map of Fertility Variables

	Age	E2	Embryos	FSH	ID	LowAFC	MaxDailyGn	MaxE2	MeanAFC	Oocytes	TotalGn
TotalGn	0.52	−0.01	−0.21	0.47	0.44	−0.42	0.91	−0.27	−0.38	−0.26	1
Oocytes	−0.11	−0.12	0.76	−0.28	−0.41	0.38	−0.28	0.5	0.42	1	−0.26
MeanAFC	−0.23	−0.13	0.35	−0.3	−0.9	0.95	−0.4	0.25	1	0.42	−0.38
MaxE2	−0.1	−0.03	0.43	−0.22	−0.23	0.25	−0.29	1	0.25	0.5	−0.27
MaxDailyGn	0.57	−0.02	−0.22	0.44	0.44	−0.43	1	−0.29	−0.4	−0.28	0.91
LowAFC	−0.24	−0.12	0.34	−0.32	−0.86	1	−0.43	0.25	0.95	0.38	−0.42
ID	0.31	0.16	−0.35	0.31	1	−0.86	0.44	−0.23	−0.9	−0.41	0.44
FSH	0.27	−0.07	−0.22	1	0.31	−0.32	0.44	−0.22	−0.3	−0.28	0.47
Embryos	−0.13	−0.09	1	−0.22	−0.35	0.34	−0.22	0.43	0.35	0.76	−0.21
E2	−0.02	1	−0.09	−0.07	0.16	−0.12	−0.02	−0.03	−0.13	−0.12	−0.01
Age	1	−0.02	−0.13	0.27	0.31	−0.24	0.57	−0.1	−0.23	−0.11	0.52

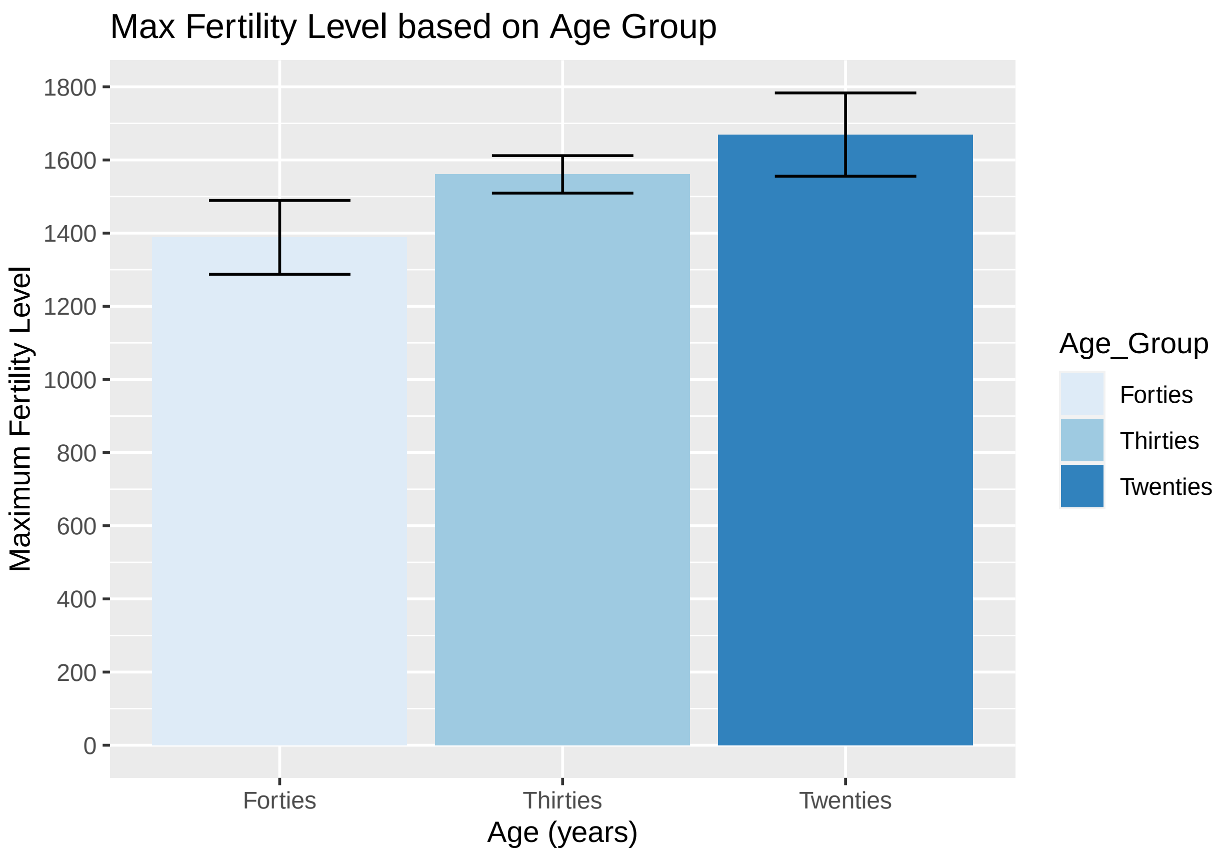
```
ggplot(data=Mutatedata, aes(x=Oocytes, y=MaxE2, color=Age_Group),stat = "summary", fun.y="mean")+ scal
```



```
ggplot(Mutatedata, aes(Oocytes,Embryos))+
geom_point(size=2,aes(color=log10(MaxE2)))+
scale_color_gradient(low="yellow", high="red")+ggtitle("Number of Embryos According to Embryo Number anc
```



```
ggplot(Mutatedata, aes(x = Age_Group, y= MaxE2))+ geom_bar(aes(y=MaxE2, fill= Age_Group),stat = "summary
```



The first plot shows the correlation heatmap of all the numeric variables in the dataset. In the correlation Map, it appears that the two numeric variables with the stongest correlation of 0.95 is betweenMeanAFC and LowAFC. It is also shown that the "ID" variable and "MeanAFC" have the weakest correlation of -0.9. A correlation trend I found interesting was the low correlation of -0.43 between "MaxDailyGn" and "LowAFC". This means that the maximum amount of gonadtropin, which stimulates the activity of the ovaries, does not really effect the smallest antral follicle count. Addtionally, the number oocytes and embryos is highly correlated. This makes sense because oocytes, if fertilized and viable, will become embryos. The second graph represents the average maximum fertility level(MaxE2) by the number of oocytes and what age group determines this value. It appear that as the number of oocytes increases, the mean maximum fertility level increases slightly, suggesting a slight positive correlation between these two variables.The color of the points describes which age group, whether in the twenties, thirties, or forties, for these points. From observing the graph, much of the higher higher points for mean maximum fertility are in the green color, indicating that the thirties age group has some of the highest mean maximum fertility levels. Additionally, most of the red dots are located towards the bottom left of the graph, indicating that many women in their forties have lower numbers of oocytes and lower mean maximum fertility levels. Overall, there are more green dots than than blue or red, revealing that most fertility observations occurred in women in their thirties. I made this graph by using ggplot and geom_point. I changed the tick-marks along the y-axis for the graph, and used "stat=summary" to give only the mean values of maximum fertility levels.

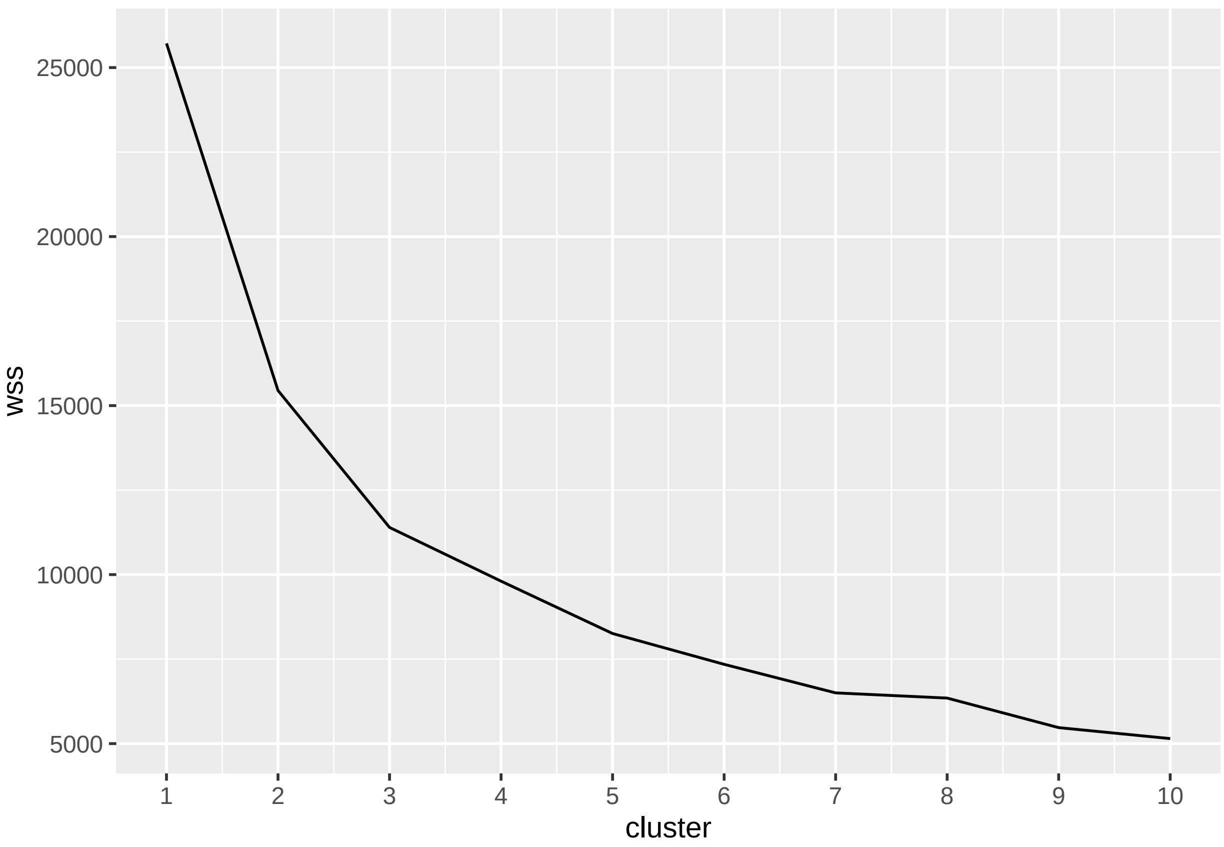
The third graph describes the relationship between the number of oocytes and the number of embryos, along with the log of the maximum fertility level(MaxE2) value. It appears that there is a strong correlation between the number of oocytes and embryos. In the graph, as the number of oocytes increases, the value of embryos also increases. In addition, the color of the points indicates the logarithmic value of MaxE2. It appears that the more yellow points, indicating lower log(MaxE2), are located in the bottom left corner of the graph. This represents that lower numbers of oocytes and lower number of embryos indicate lower maximum fertility levels. The red and orange points are located mostly to the right of graph, where there are either high numbers of oocytes or embryos. I made this graph using ggplot and geom_point. I also changed the size and the color of the points based on log(MaxE2).

Lastly, I provided a visualization of mean maximum fertility levels based on the three different age groups by using ggplot and geom_bar. From this, I was able to visualize, as indicated from my summary statistics, that the twenties age group had the highest mean maximum fertility level. Additionally, I was able to view the mean standard error of each age group by using geom_errorbar, revealing that the thirties age group had the lowest standard error. I changed the tick-marks along the y-axis for the graph, and used “stat=summary” to give only the mean values of maximum fertility levels, and altered the width of the standard error bar.

```r
#Dimentionality Reduction
library(cluster)
kmeandata<-Fertility_join%>%select(-LowAFC, -MeanAFC, -E2, -ID,-MaxE2,-MaxDailyGn, -TotalGn)
#Embryos and FSH levels
wss<-vector()
for(i in 1:10){
tempfit <- Fertility_join%>%select(-LowAFC, -MeanAFC, -E2, -ID,-MaxE2,-MaxDailyGn, -TotalGn)%>%kmeans(.
wss[i] <- tempfit$tot.withinss%>%glimpse()
}
```

```
##  num 25717
##  num 15447
##  num 11396
##  num 9806
##  num 8257
##  num 7345
##  num 6501
##  num 6347
##  num 5473
##  num 5148
```

```r
ggplot()+geom_line(aes(x=1:10,y=wss))+xlab("cluster")+scale_x_continuous(breaks=1:10)
```



```
kmeans1<-kmeandata%>%scale%>%kmeans(2)%>%glimpse()
```

```
## List of 9
##  $ cluster     : int [1:333] 2 1 2 1 2 2 2 2 2 2 ...
##  $ centers     : num [1:2, 1:4] 0.22 -0.323 0.334 -0.489 -0.601 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:2] "1" "2"
##   .. ..$ : chr [1:4] "Age" "FSH" "Oocytes" "Embryos"
##  $ totss       : num 1328
##  $ withinss    : num [1:2] 516 392
##  $ tot.withinss: num 907
##  $ betweenss   : num 421
##  $ size        : int [1:2] 198 135
##  $ iter        : int 1
##  $ ifault      : int 0
##  - attr(*, "class")= chr "kmeans"
```

```
kmeans1
```

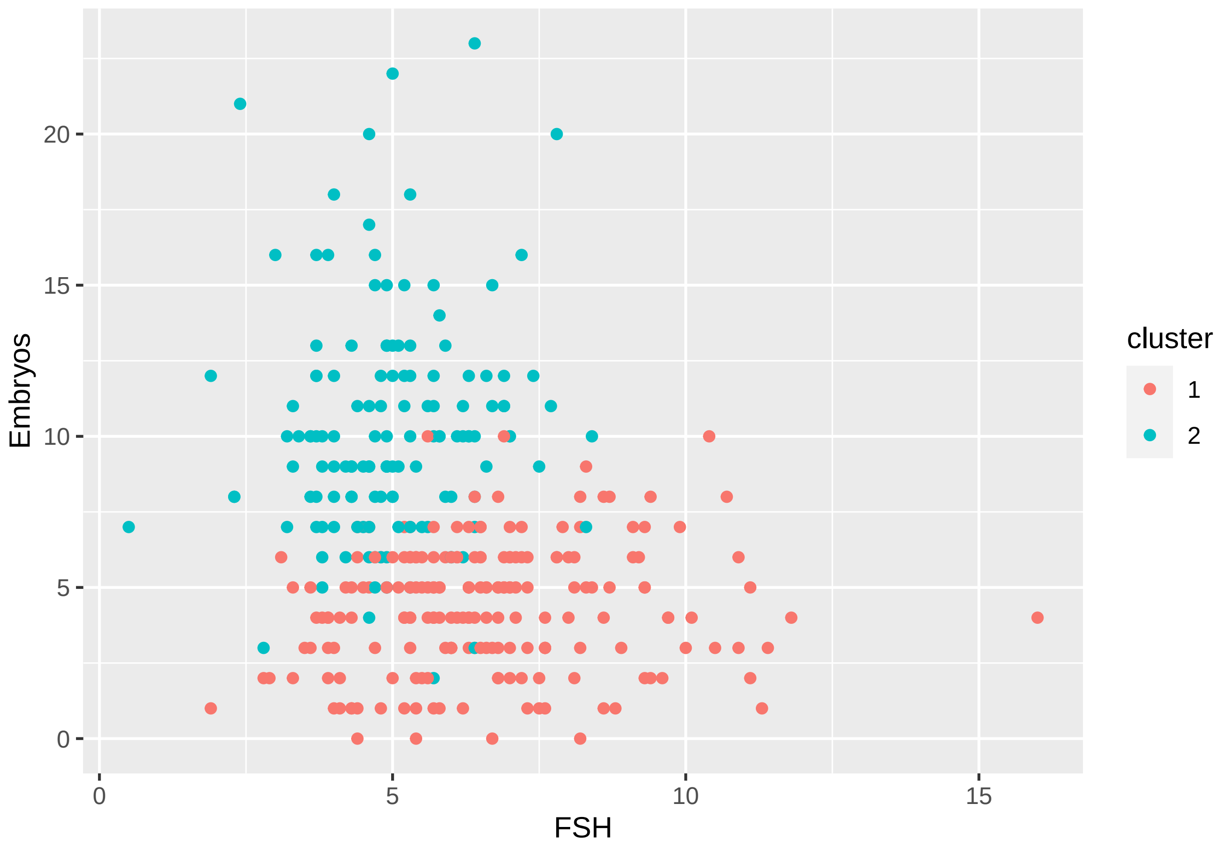
```
## K-means clustering with 2 clusters of sizes 198, 135
##
## Cluster means:
##          Age        FSH    Oocytes    Embryos
## 1  0.2203763  0.3336367 -0.6012882 -0.5827165
## 2 -0.3232186 -0.4893338  0.8818894  0.8546509
```

```
##
## Clustering vector:
##   [1] 2 1 2 1 2 2 2 2 2 2 2 2 1 2 2 1 2 1 2 2 2 2 2 2 2 2 1 2 2 2 2 2 1 1 1 2 2
##  [38] 2 2 2 2 2 2 1 2 2 2 2 2 1 2 1 1 2 1 2 1 1 1 2 2 2 1 2 1 2 2 1 2 2 2 1 2 1
##  [75] 2 1 2 2 2 2 1 1 1 1 2 1 2 2 2 1 2 2 2 2 2 1 2 1 1 1
##  [ reached getOption("max.print") -- omitted 233 entries ]
##
## Within cluster sum of squares by cluster:
## [1] 515.7006 391.7943
##  (between_SS / total_SS =  31.7 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
kmeansclust<-kmeandata%>%mutate(cluster=as.factor(kmeans1$cluster))%>%glimpse
```

```
## Observations: 333
## Variables: 5
## $ Age     <int> 40, 37, 40, 40, 30, 29, 31, 33, 36, 35, 25, 39, 35, 30, 37,...
## $ FSH     <dbl> 5.3, 7.1, 4.9, 3.9, 4.0, 3.9, 3.8, 4.3, 4.9, 3.7, 5.0, 5.3,...
## $ Oocytes <int> 25, 7, 27, 9, 19, 19, 13, 15, 23, 26, 22, 22, 7, 27, 12, 11...
## $ Embryos <int> 13, 6, 15, 4, 12, 16, 9, 9, 10, 8, 13, 18, 5, 18, 9, 2, 8, ...
## $ cluster <fct> 2, 1, 2, 1, 2, 2, 2, 2, 2, 2, 2, 2, 1, 2, 2, 1, 2, 1, 2, 2,...
```

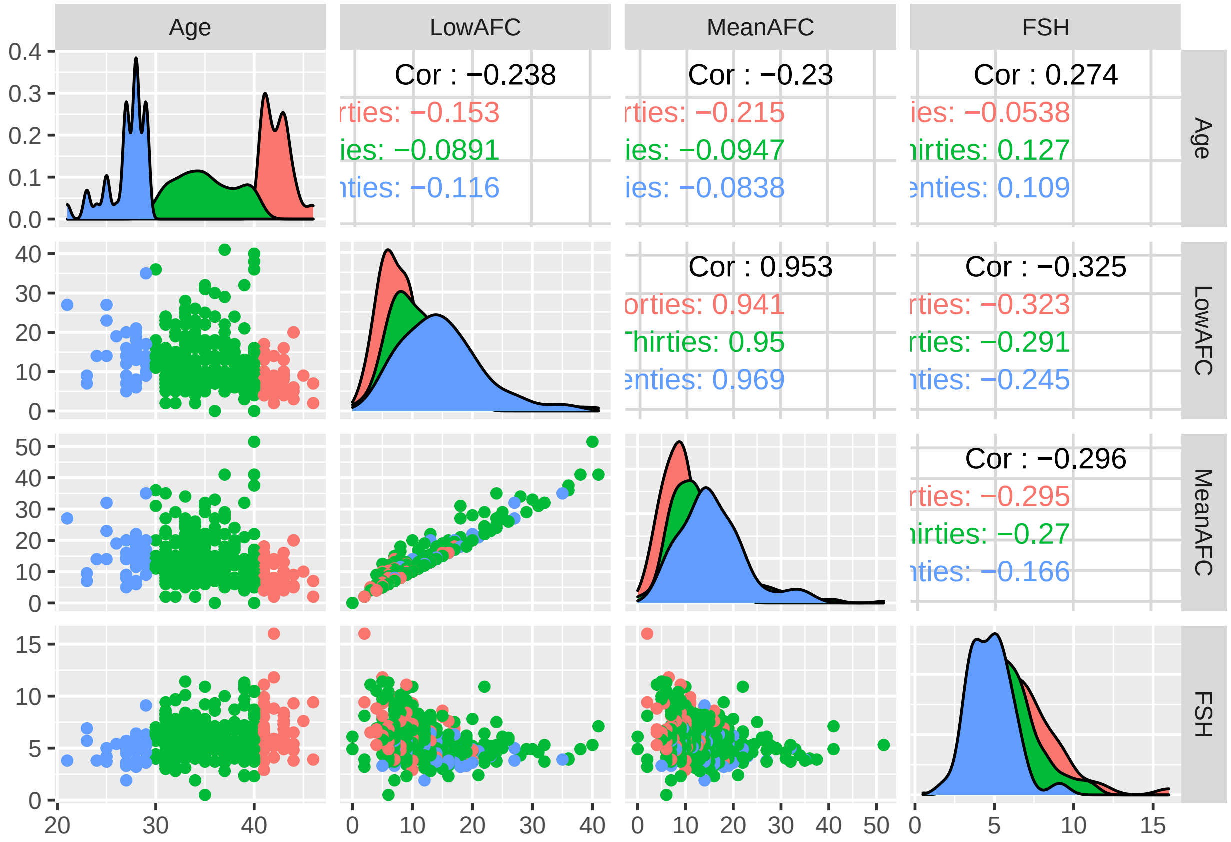
```
kmeansclust%>%ggplot(aes(FSH, Embryos, color=cluster))+geom_point()
```



```
fertility_clus<-Mutatedata%>%mutate(cluster=as.factor(kmeans1$cluster))
format<-fertility_clus%>%group_by(Age_Group)%>%count(cluster)%>%arrange(desc(n))%>%
pivot_wider(names_from="cluster",values_from="n",values_fill = list('n'=0))
format
```

```
## # A tibble: 3 x 3
## # Groups:   Age_Group [3]
##   Age_Group   `1`   `2`
##   <chr>     <int> <int>
## 1 Thirties    144   104
## 2 Forties      42     8
## 3 Twenties     12    23
```

```
library(GGally)
ggpairs(fertility_clus, columns=1:4, aes(color=Age_Group))
```



For my data, I performed a k-means cluster for five numeric variables, FSH levels, number of Embryos, number of oocytes, and Age. First, I made a new dataset, named “kmeandata” to include only these 5 variables. Next, I determined the number of clusters by making a plot of clusters vs. wss. I chose to do 2 clusters based on the elbow shown in the grapoh generated. At cluster number 2, the graph started to decrease gradually. I then used ggplot to create a plot by using k-means cluster. Within the plot, I am able to see two distinct groups. From this, I am able to observe that generally women with higher numbers of embryos had lower FSH levels. Additinally I was able to visualize all pairwise combinations of the variables:age, lowAFC, MeanAFC, and FSH by using ggpairs function to create a scatterplot matrix. From this, I am able to determine the correlation between each of the 4 variables, revealing that MeanAFC and LowAFC have the highest correlation among the variables analyzed. LowAFC and FSH have the lowest correlation.The variable distribution based on 3 distinct age groups can be visualized along the diagonal. It is shown that the twenties age group has a wider distribution of LowAFC and MeanAFC.The forties age group seems to have high density for lower levels of LowAFC and MeanAFC.However in the twenties age group, there is high density for lower levels of FSH.