Group Work with Diamonds and Beavers

Noah Ferrel

Group Work:

Use the diamonds data set to answer the following questions:

1. Are flawless diamonds on average smaller than those with inclusions? (Ch 6)
2. What is the distribution of price per carat of the diamonds in this set? (Ch 7)
3. What proportion of diamonds in the set are of each cut? Does this change with clarity? (Ch 10)
4. What determines a diamond's cost? (Ch 12,14)

Use the beavers data set to answer the following questions:

1. Does the body temperature of a beaver vary with time? (Ch 13)
2. How much variation is associated with beaver temperature measurements? (Ch16)

https://www.kaggle.com/shivam2503/diamonds

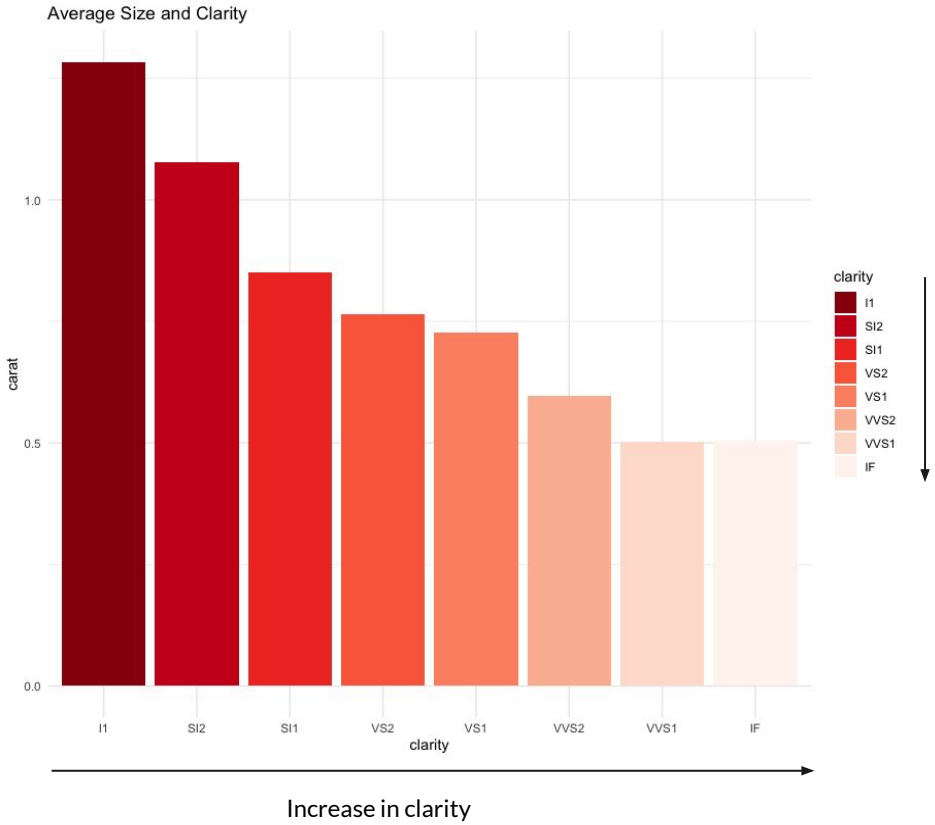
Bring several graphs with you for Wed 3/3 to present in class. Be prepared to present:

Examples of your graphics answering each question. Explain the graphing choices you made for each (how does it follow OCAR?).

Provide code to the rest of the class reproducing your graphs.

Diamonds

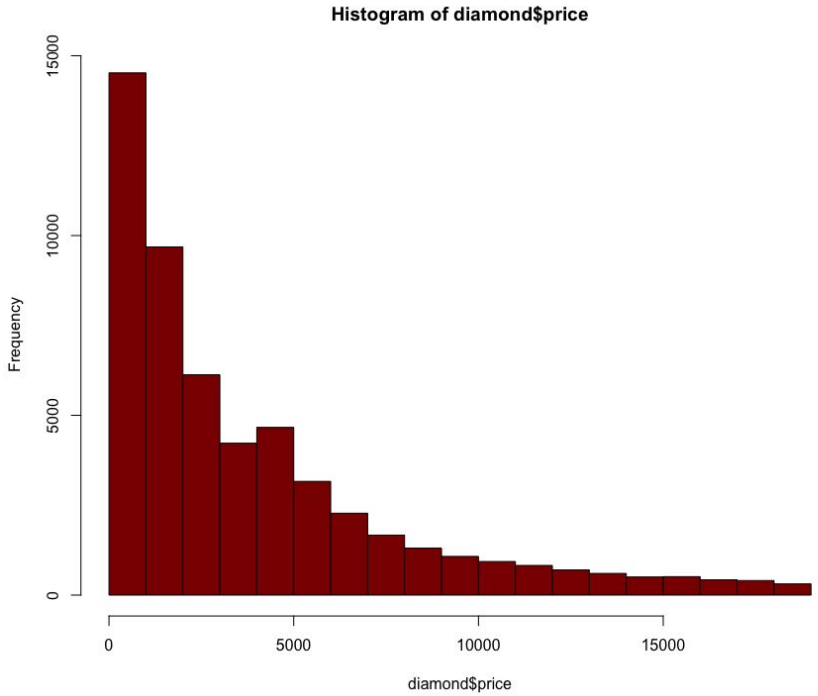
Are flawless diamonds on average smaller than those with inclusions?



Code for plot

```
#Are flawless diamonds on average smaller than those with inclusions?

new_counts = unlist(lapply(unique(diamond$clarity), function(x)
                    round(mean(diamond$carat[diamond$clarity==x]),3)))

cat = unique(cat)

new_counts = as.data.frame(cbind(cat,new_counts))

colnames(new_counts)= c("clarity","carat")

count_ordered = rbind(new_counts[7,],new_counts[1:2,],new_counts[4,],new_counts[3,],
             new_counts[5:6,],new_counts[8,])

rownames(count_ordered)=1:8

count_ordered$clarity = factor(count_ordered$clarity,
                               levels = c("I1", "SI2", "SI1", "VS2", "VS1", "VVS2", "VVS1", "IF"))
count_ordered$carat = as.double(count_ordered$carat)

ggplot(count_ordered,aes(x=clarity,y=carat,fill=clarity))+
  geom_bar(stat='identity')+
  theme_minimal()+scale_fill_brewer(type = "seq",palette="Reds",direction = -1)+ggtitle("Average Size and Clarity")

```

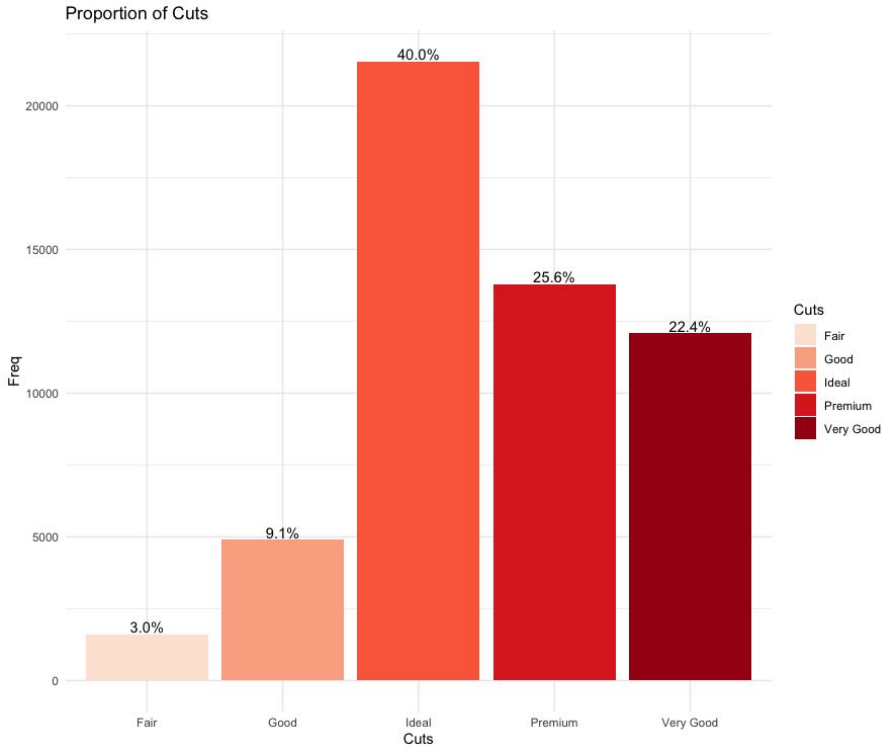
What is the distribution of price per carat of the diamonds in this set?



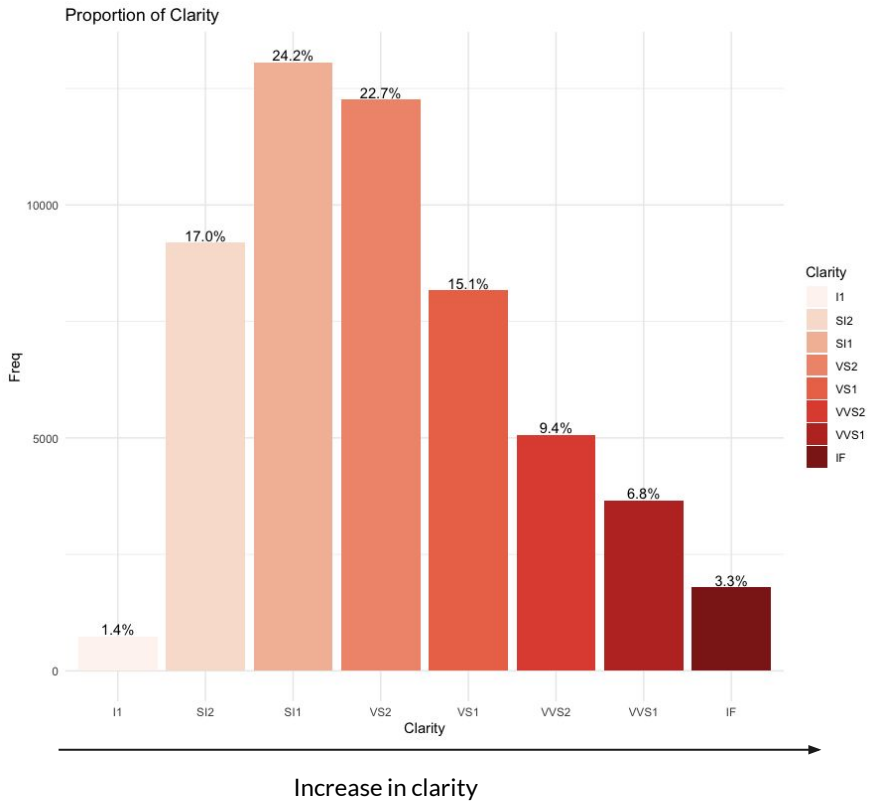
Code for plot

```
46  #What is the distribution of price per carat of the diamonds in this set?
47
48  hist(diamond$price, col = 'darkred')
49
```

What proportion of diamonds in the set are of each cut?



Does this change with clarity?



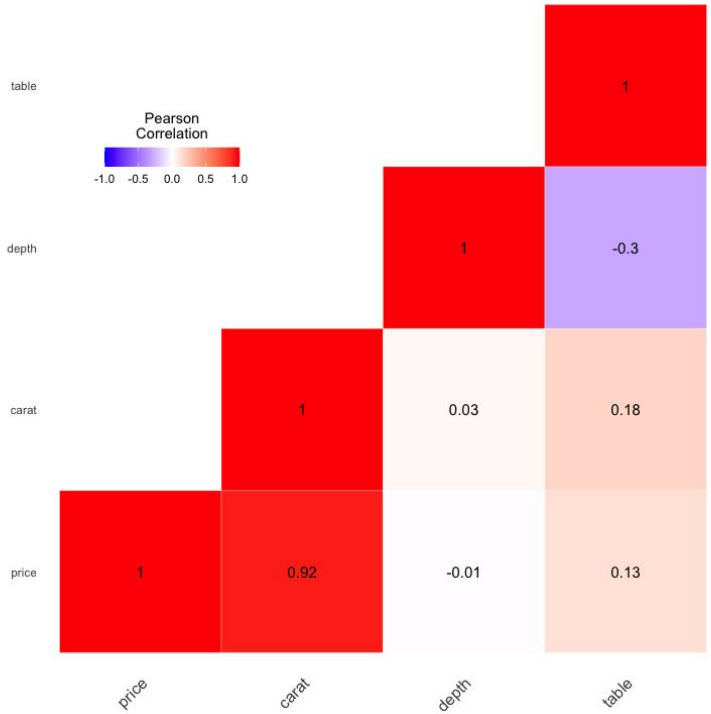
Code for plot

```
#What proportion of diamonds in the set are of each cut? Does this change with clarity?

cuts = as.data.frame(table(diamond$cut))

colnames(cuts) = c("Cuts","Freq")
cuts %>%
  arrange(desc(Freq)) %>%
  mutate(prop = percent(Freq / sum(Freq))) -> cuts



ggplot(cuts,aes(x=Cuts,y=Freq,fill=Cuts))+
  geom_bar(stat='identity')+
  geom_bar(stat='identity')+
  theme_minimal()+scale_fill_brewer(type = "seq",palette="Reds",direction = 1)+ggtitle("Proportion of Cuts")+
  geom_text(aes(label=prop,vjust = -0.2))+
  guides(fill = guide_legend(title = "Cuts"))

```

Code for plot

```
clarity = as.data.frame(table(diamond$clarity))
clarity$Clarity = factor(clarity$Clarity,
                         levels = c("I1", "SI2", "SI1", "VS2", "VS1", "VVS2", "VVS1", "IF"))

colnames(clarity) = c("Clarity","Freq")
clarity %>%
  arrange(desc(Freq)) %>%
  mutate(prop = percent(Freq / sum(Freq))) -> clarity



ggplot(clarity,aes(x=Clarity,y=Freq,fill=Clarity))+
  geom_bar(stat='identity')+
  geom_bar(stat='identity')+
  theme_minimal()+scale_fill_brewer(type = "seq",palette="Reds",direction = 1)+ggtitle("Proportion of Clarity")+
  geom_text(aes(label=prop,vjust = -0.2))+
  guides(fill = guide_legend(title = "Clarity"))

```

What determines a diamond’s cost?



Code for plot

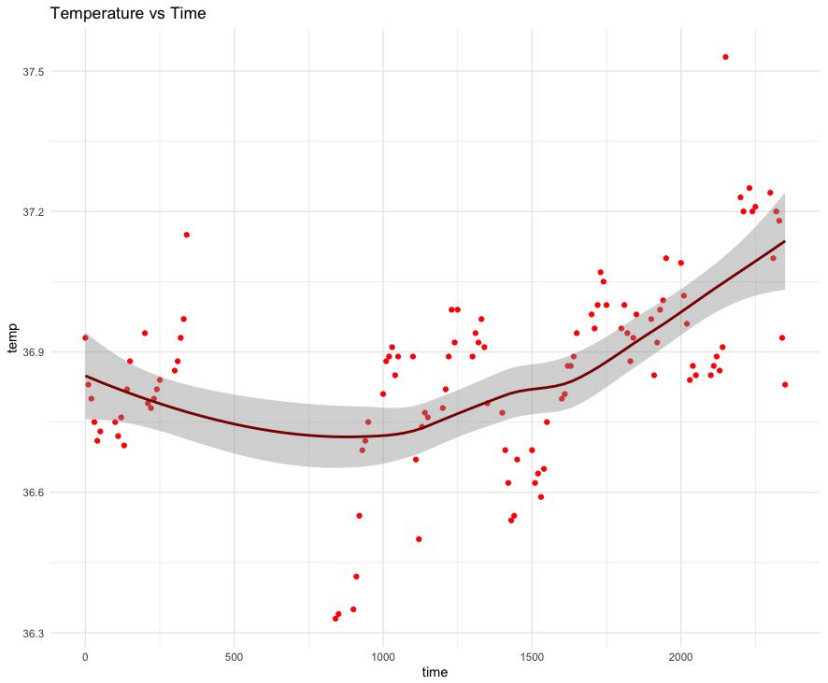
```
90  # What determines a diamonds cost? (Ch 12,14)
91
92  plot(diamond , pch=20 , cex=1.5 , col="#69b3a2")
93
94  pairs(diamond, lower.panel = NULL)
95
96  cor(diamond)
97  ggpairs(diamond, title="correlogram with ggpairs()",method = c("everything", "pearson"))
98
99  library("MASS")
100
101 diamond$X =NULL
102
103 diamond_base = lm(diamond$price~., data = diamond)
104 diamond_final = stepAIC(diamond_base,trace = F,direction = c("both"))
105
106 dia = diamond[,c("price","carat","depth","table")]
107 diacormat <- round(cor(dia),2)
108
109 # Get upper triangle of the correlation matrix
110 get_upper_tri <- function(cormat){
111   cormat[lower.tri(cormat)]<- NA
112   return(cormat)
113 }
114
115 upper_tri <- get_upper_tri(diacormat)
116
117 melted_dia <- melt(upper_tri,na.rm = TRUE)
118 head(melted_dia)
```

Code for plot

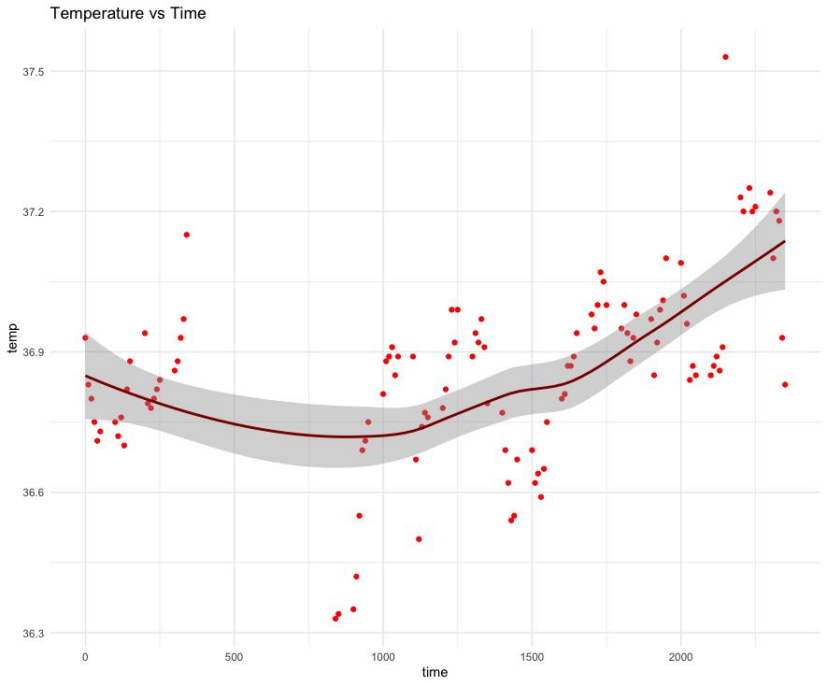
```
ggplot(data = melted_dia, aes(Var2, Var1, fill = value))+
  geom_tile(color = "white")+
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1,1), space = "Lab",
                       name="Pearson\nCorrelation") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 45, vjust = 1,
                                   size = 12, hjust = 1))+
  coord_fixed()+
  geom_text(aes(Var2, Var1, label = value), color = "black", size = 4) +
  theme(
    axis.title.x = element_blank(),
    axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank(),
    axis.ticks = element_blank(),
    legend.justification = c(1, 0),
    legend.position = c(0.3, 0.7),
    legend.direction = "horizontal")+
  guides(fill = guide_colorbar(barwidth = 7, barheight = 1,
                               title.position = "top", title.hjust = 0.5))
```

Beavers

Does the body temperature of a beaver vary with time?



How much variation is associated with beaver temperature measurements?



Code for plot

```
144  #Does the body temperature of a beaver vary with time? (Ch 13)
145
146
147  head(beaver1)
148
149  ggplot(data = beaver1,aes(time,temp)) +
150    geom_point(color="Red")+
151    theme_minimal()+
152    geom_smooth(color="Darkred")+ggtitle("Temperature vs Time")
153
```