Salaries and Anxiety

Read in the dataset: http://vicpena.github.io/sta9750/salary.csv

The dataset contains salaries ("salary"), anxiety levels ("anxiety") on a scale that goes from 0 (no anxiety) to 7 (very anxious), and education level ("education").

```
library(dplyr)
library(tidyr)
library(ggplot2)
```

Firstly, we load the important libraries we need for completing this exercise. Many of which would be useful for other questions in the exercise.

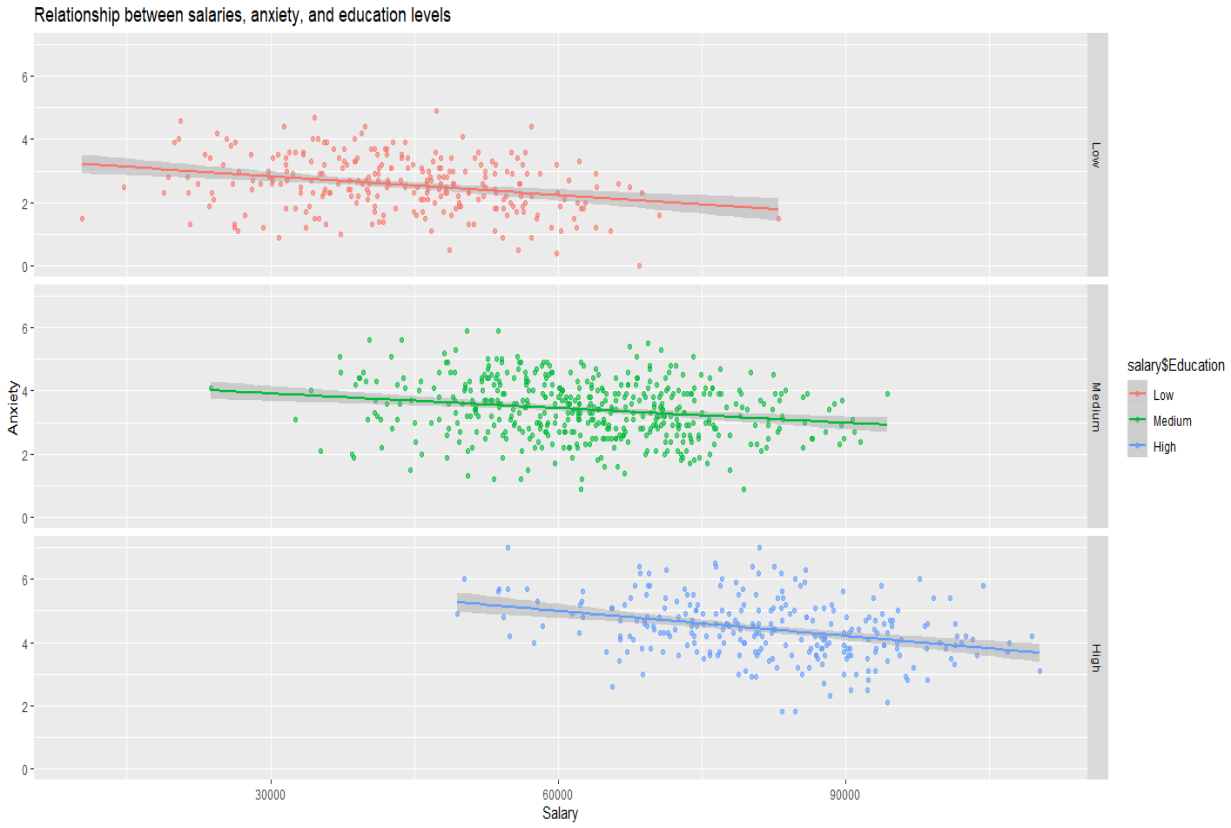
Then we also load the data to begin our exercise:

```
# Load the data
salary <- read.csv("C:\\Users\\its_t\\Documents\\CUNY Fall 2019\\9750 - Softw
are Tools and Techniques_Data Science\\HW2\\salary.csv")
```

1. Create a figure that displays the relationship between salaries, anxiety, and education levels. Your figure can contain more than one plot / facet / panel. Make sure that the labels and the title are interpretable. Interpret in detail the relationships that you see.

R code and the output:

```
> View(salary)

> salary$Education = factor(salary$Education, levels = c("Low", "Medium", "Hi
gh"))

> ggplot(salary) +
+   aes(x = salary$Salary, y = salary$Anxiety, color = salary$Education) +
+   geom_point(alpha = 0.6)+geom_smooth(method = "lm")+
+   facet_grid(salary$Education ~ .) + xlab("Salary") + ylab("Anxiety") +
+   ggtitle("Relationship between salaries, anxiety, and education levels")
```



Explanation: First, we arrange the data for education column which is a factor variable in the right sequence to interpret the results properly.

We do see that with increase in education level, there is an increase in pay but the same cannot be concluded for anxiety. For example, cases for people with high education, as plotted above with blue scatter plot – the median salary is closer to what is probably the highest for people with low education.

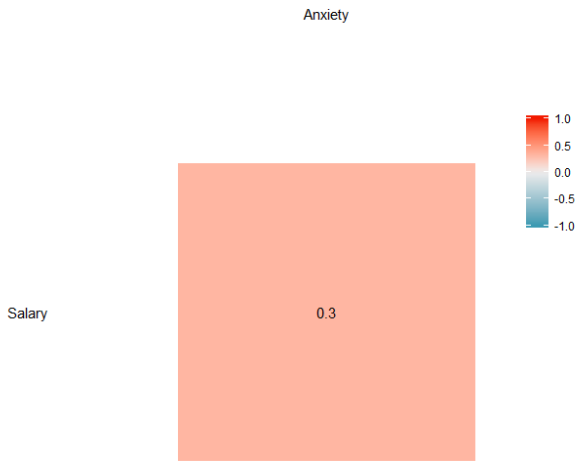
2. An article claims that higher salaries come at the cost of higher anxiety levels. Does the figure you created in part 1 support this claim?

Explanation: As we see from the above graph, we can say that the best fit line for the regression plot between salary and anxiety is not showing a positive slope as we move from left to right on the x-axis. This means, the increase in salary doesn't really always mean that the anxiety level will increase. We can see that for some cases it is true and there could be many reasons for the that, but it can neither be generalized not it can be solely attributed to high paycheck. Also, the same trend is observable for different level of education and thus we can conclude that there is no such direct relationship between increase in salary leading to increase in anxiety.

Also, we check for a relationship between salary and anxiety we see a low co-relation and for the anxiety to increase with increase in salary, the co-efficient should have a high positive value i.e., closer to 1.

R code and the output:

```
> library(GGally)
> ggcorr(salary, label = TRUE)
```



Italian Restaurants in NYC

Read in the dataset: http://vicpena.github.io/sta9750/spring19/nyc.csv

The variables are:

- Case: case-indexing variable
- Restaurant: name of the restaurant
- Price: average price of a meal and a drink per person
- Food: average Zagat rating of the quality of the food (from 0 to 25)
- Decor: same as above, but with quality of the decor
- Service: same as above, but with quality of service
- East: it is equal to 1 if the restaurant is on the East Side (i.e. east of Fifth Ave)

Answer the following questions:

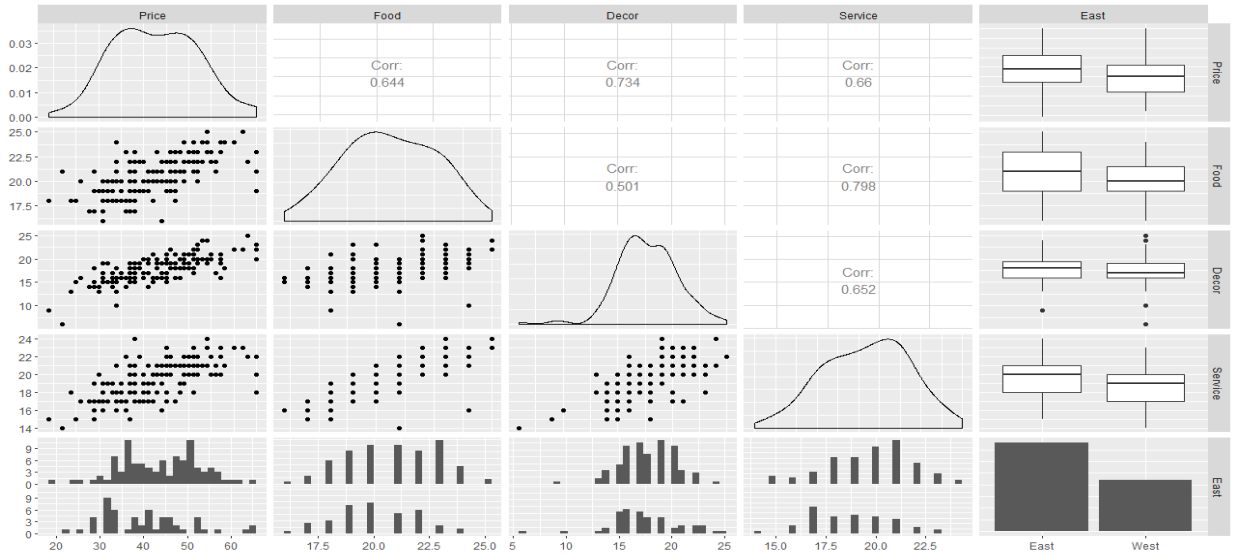
Firstly, we load the data that we need to answer the below questions:

```
> # Load the data
> ir_nyc <- read.csv("C:\\Users\\its_t\\Documents\\CUNY Fall 2019\\9750 - Sof
tware Tools and Techniques_Data Science\\HW2\\nyc.csv")

> View(ir_nyc)
```

1. Create a figure that contains plots for all the pairs of variables in the dataset, except Case (i.e., a figure that contains plots for Restaurant vs Price, Food vs Price, Decor vs Service, etc.). Describe what you see in the plots. What are the strongest and weakest relationships you see?

R code and the output:

```
> ir_nyc <- ir_nyc %>% select(3:7)
> ggpairs(ir_nyc)
```



Explanation: We use the ggpairs function we learned during our class that gives us a good understanding of the variables involved in a dataset and what relationship they carry.

In general, we think of variables to have a strong relationship if they have more than 80% relationship (irrespective of the sign of the coefficient), but that's not a universal truth and it can differ with the industry and the size of the dataset. Since, we don't have a benchmark shared to us or any comparison study from other types of restaurants in NYC, we decide to use this sample as the only available info.

Having said that, we can see that the strongest relationship is between food and service that is 0.798. This is followed by décor and price which has a relationship coefficient of 0.734

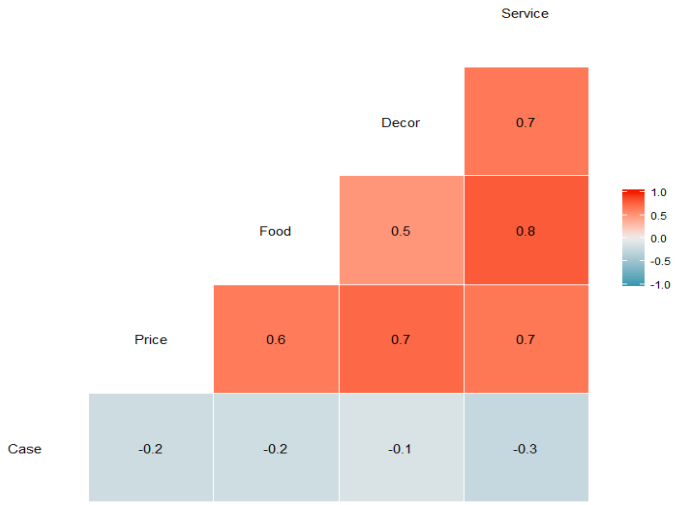
The weakest relationship is observed between food and décor with a coefficient of 0.501. This can be understood as higher rating with décor has nothing to do with quality of food served to the audience.

Density plots for every numeric continuous variable also help us to identify skewness, kurtosis and distribution information. The boxplot can help us identify the outliers and the distribution of quartile values and in case of 'décor' we can identify few of them.

2. Provide a heatmap for the correlation between the numerical variables in the dataset. What can you see?

R code and the output:

```
> ggcorr(ir_nyc, label = TRUE)
```



Explanation: We use the gg_corr function from GGally package to create a heatmap of the correlation between different numerical values. There is no need to filter out categorical data variables because gg_corr function identifies them and only plots heatmap for numerical values.

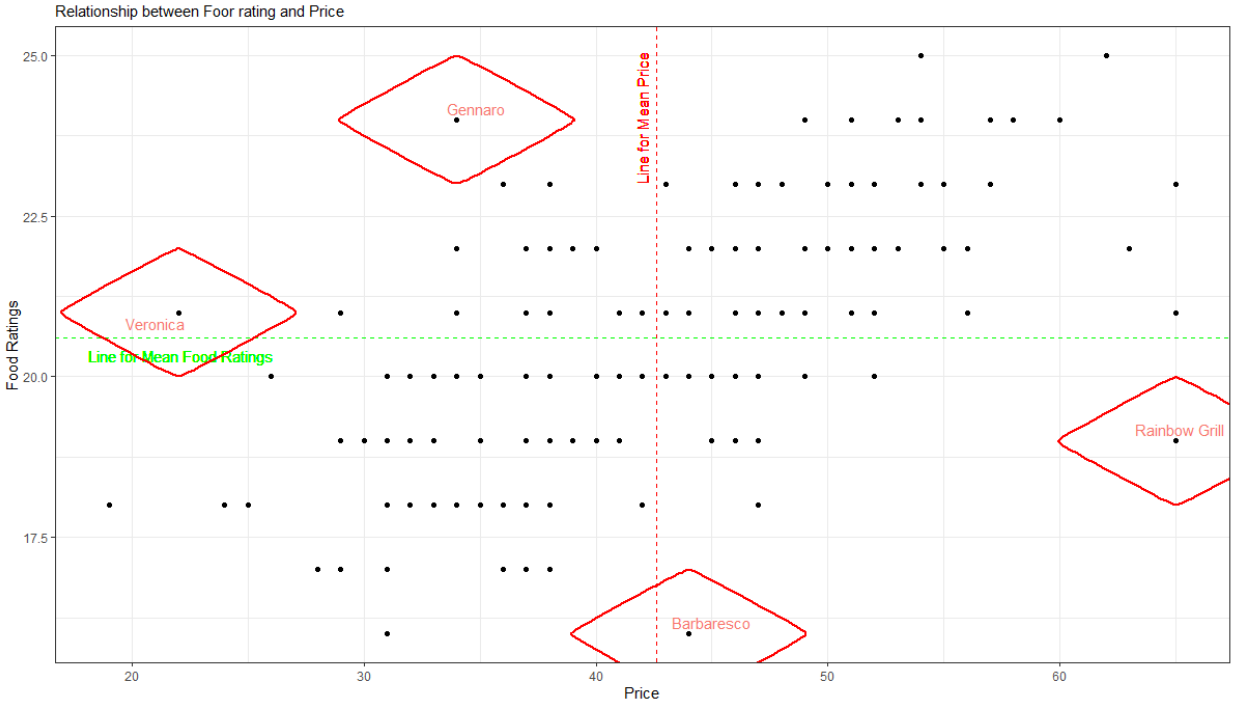
The strongest relationship is between food and service as observed in the previous question and it shows the same result as 0.8. The weakest is seen for case and décor at -0.1. If we exclude 'case' from the analysis then Food and Décor has the lowest co-relation with a co-efficient equal to 0.5

3. Find 2 examples of cheap restaurants that have relatively good food and 2 examples of expensive restaurants that have relatively bad food.

R code and the output:

```
> summary(ir_nyc)

# Mean Price = 43
# Mean Food Rating = 21

> library(ggalt)
> library(ggrepel)

> highlight <- ir_nyc %>% subset(Price < 25 & Food > 20 | Price  <35 & Food >
22.5| Price > 64 & Food < 20 | Price > 40 & Food < 17)

> h1 <- highlight %>% filter(row_number()==1)
> h2 <- highlight %>% filter(row_number()==2)
> h3 <- highlight %>% filter(row_number()==3)
> h4 <- highlight %>% filter(row_number()==4)

> theme_set(theme_bw())

>   ggplot(ir_nyc, aes(x=ir_nyc$Price, y = ir_nyc$Food)) + geom_point() +
+   labs(subtitle="Relationship between Foor rating and Price", y="Food Ratin
gs", x="Price") +
+   geom_vline(xintercept = mean(ir_nyc$Price), color = "red", linetype = "da
shed") +
+   geom_hline(yintercept = mean(ir_nyc$Food), color = "green", linetype = "d
ashed") +
+   geom_text(aes(x=41.5, label="\n Line for Mean Price", y=24), colour="red"
, angle=90, text=element_text(size=5)) +
+   geom_text(aes(x=22, label="\n Line for Mean Food Ratings", y=20.5), colou
r="green", angle=360, text=element_text(size=5)) +
+   geom_text(aes(x=22, label="\n Line for Mean Food Ratings", y=20.5), colou
r="green", angle=360, text=element_text(size=5)) +
+   geom_text_repel(data = highlight, aes(label=Restaurant, x=Price,y=Food, c
olor = "blue"))+
+   geom_encircle(aes(x=Price,y=Food),data=h1,color="red", size=2,expand=1/10
0000, s_shape=0.9) +
```

```
+    geom_encircle(aes(x=Price,y=Food),data=h2,color="red", size=2,expand=1/10
0000, s_shape=0.9) +
+    geom_encircle(aes(x=Price,y=Food),data=h3,color="red", size=2,expand=1/10
0000, s_shape=0.9) +
+    geom_encircle(aes(x=Price,y=Food),data=h4,color="red", size=2,expand=1/10
0000, s_shape=0.9)
```



Explanation: We first look for the summary function to understand what the mean value for the food ratings and prices variables that we are going to use to identify 2 examples of cheap restaurant and 2 examples of expensive restaurant.

The scatter plot helps us identify the outliers here and we use the value of those points to create a different data frame to identify the 4 restaurants that will satisfy the condition.

	Case	Restaurant	Price	Food	Decor	Service	East
21	117	Veronica	22	21	6	14	West
39	168	Gennaro	34	24	10	16	West
62	130	Rainbow Grill	65	19	23	18	West
132	50	Barbaresco	44	16	16	16	East

We use the above table and try to identify them in the graph and for that we use geom_text_repel function. We don't label all the restaurants to avoid labels getting

overlapped. I also have encircled the points using geom_encircle function to focus the area of the graph which answers our question.

My final answers are:
2 cheap restaurants: Veronica and Gennaro
2 expensive restaurants: Rainbow Grill and Barbaresco

Alternatively, there is one more straightforward solution. We subtract the food ratings from the price and see the top two and the bottom two but the problem with that logic is that there could be a case where the ratings are moderate but the prices are too high and hence the difference skews unfavorably. I though using median will be a better criteria to identify the best and the worst in terms of value for money for given set of restaurants.

Having said that, the alternate solution will lead to following results:
2 cheap restaurants: Veronica and Lamarca
2 expensive restaurants: Rainbow Grill and Harry Cipriani

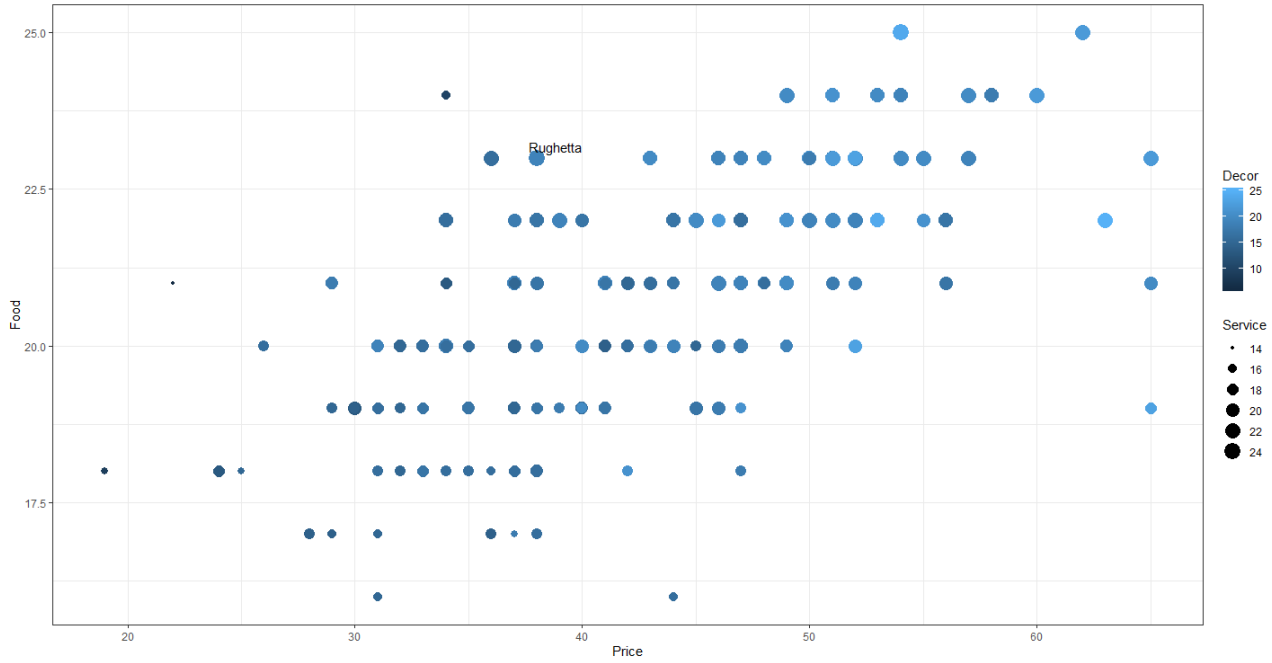
R code and the output:

```
> new %>% arrange(new$diff) %>%  head(2)
  Case     Restaurant Price Food Decor Service East diff
1  130  Rainbow Grill    65   19    23      18 West  -46
2   30 Harry Cipriani    65   21    20      20 East  -44
> new %>% arrange(-new$diff) %>%  head(2)
  Case Restaurant Price Food Decor Service East diff
1  117   Veronica    22   21     6      14 West   -1
2  115    Lamarca    19   18     9      15 East   -1
```

4. Suppose you're going on a date and want to use the information in this dataset to pick where to go. Assume your budget is at most $40. Assuming that you can get a table anywhere you want, where would you go and why?

Explanation: We filter for the rows that adjusts with our budget. Here, it is $40 and thus we filter for observations for price >= 40. Then we do a rowwise addition for each column and see what is the highest sum for any restaurant to calculate the highest overall rating inclusive of all parameters – food, décor, service and price.

When we do that we find, **Rughetta** to be the best solution and if I happen to go on a date, this will be my preferred choice.



R code and the output:

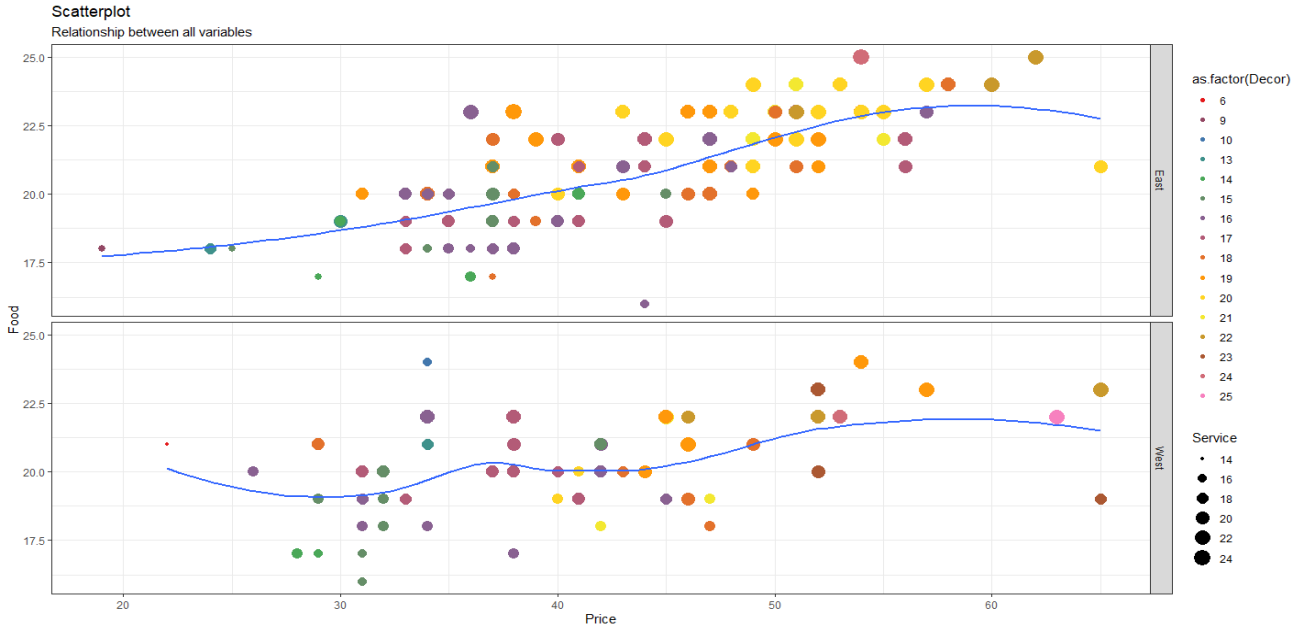
```
> ir_date <- ir_nyc %>% filter(Price <= 40) %>% select(3:6)
> ir_date <- cbind(ir_date, Total = rowSums(ir_date))
> ir_date <- ir_nyc %>% filter(Price <= 40)

> ggplot(ir_nyc, aes(x=Price, y = Food)) +
+   geom_point(aes(col = Decor, size = Service)) +
+   geom_text_repel(data = subset(ir_date, Restaurant== "Rughetta"), aes(labe
l=Restaurant, x=Price,y=Food))
```

5. Create a figure that displays the relationship between price, food, decor, service, and the East / West indicator. Your figure can contain more than one plot / facet / panel. Make sure that the labels and the title are interpretable. Interpret in detail the relationships that you see.

R code and the output:

```
> colourCount <- length(unique(ir_nyc$Decor))
> colourCount
[1] 16

> getPalette <- colorRampPalette(brewer.pal(8, "Set1"))(colourCount)
> getPalette
 [1] "#E41A1C" "#934864" "#4277AD" "#3F918B" "#4AA858" "#658E67" "#896191" "#
B35B77" "#E3712B"
[10] "#FF980A" "#FFD422" "#F3E831" "#C9992C" "#AB5832" "#D16C78" "#F781BF"

> theme_set(theme_bw())
> ir_plot <- ggplot(ir_nyc, aes(x=ir_nyc$Price, y = ir_nyc$Food)) +
+   geom_point(aes(col = as.factor(Decor), size = Service)) +
```

```
+    geom_smooth(method="loess", se=F) + facet_grid(ir_nyc$East ~ .) +
+    scale_color_manual(values = getPalette) +
+    labs(subtitle="Relationship between all variables",
+         y="Food", x="Price", title="Scatterplot")

> plot(ir_plot)
```



Explanation: We use the ggplot function to plot the relationship between all the variables in the NYC database for Italian restaurants. We use all the aesthetics possible to make the graph look better. We assign variables to size and color to get the differences highlighted. However, because of the no. of different values possible for décor which we use for our color aesthetics, we had to call a separate color palette to get the color differentiator appear more clearly.

When we interpret the data, we see that East has more restaurants with higher food rating compared to West. Also, the restaurants from either East or West which have high ratings for 'décor' not necessarily leads to high food rating but that does lead to higher price. Additionally, cases with poor food ratings also have good poor ratings for service and that's true on both sides – East or West, and thus we can't conclude a relationship.

If we observe the curve line which tries to depict the relationship between the x and y variable, we see that with increase in price, the food quality has increased with a more dominant factor for East. The increase was true for West too, but it was less influential.

Interfaith dating data

Consider the interfaith dating dataset

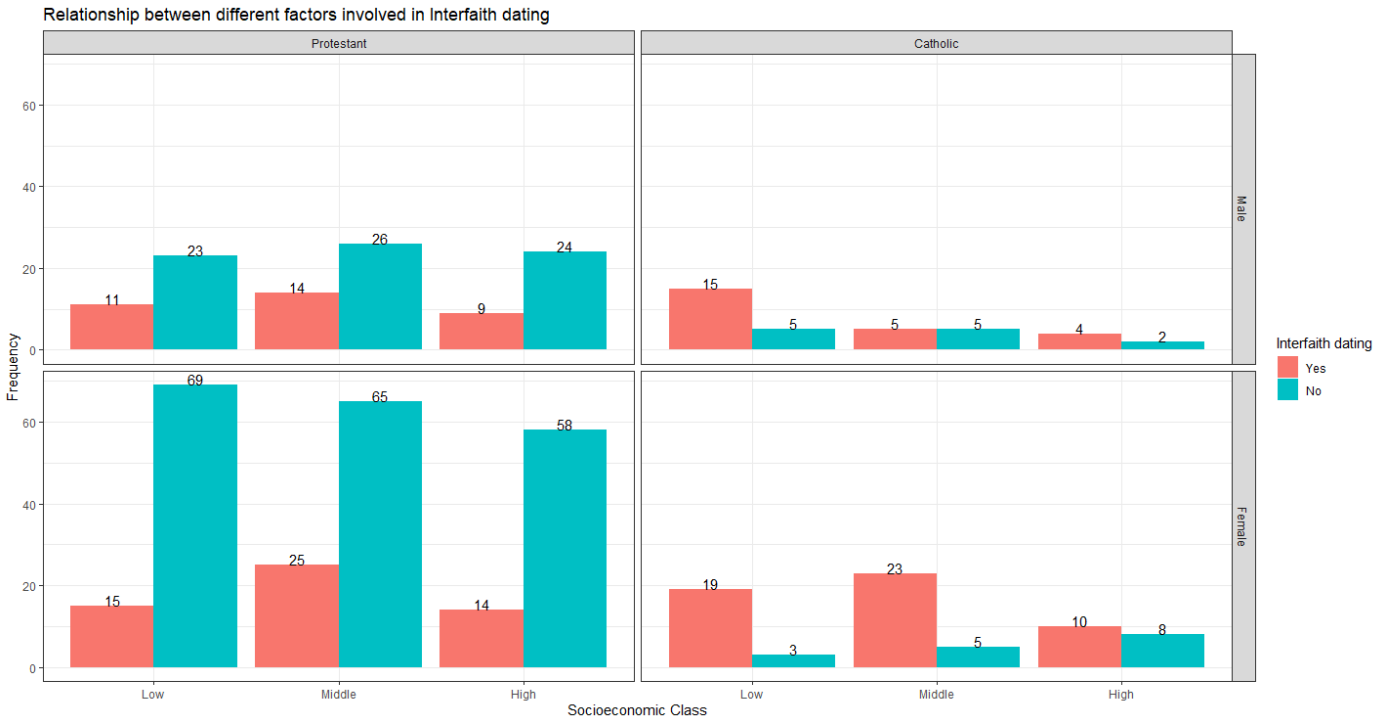
Description: http://users.stat.ufl.edu/~winner/data/interfaith.txt

Data: http://users.stat.ufl.edu/~winner/data/interfaith.dat

Create a figure that shows the relationship between socioeconomic class, religion, gender, and the indicator of interfaith dating . Your figure can contain more than one plot / facet / panel. Interpret in detail the relationships that you see in the plots. Make sure that the labels and the title are interpretable.

R code and the output:

```
> idd <- read.table("http://users.stat.ufl.edu/~winner/data/interfaith.dat",
header = FALSE)

> names(idd)[1:5] <- c("SEC", "Religion", "Gender", "Interfaith dating", "Cou
nt")

> names <- c(1:4)
> idd[,names] <- lapply(idd[,names],factor)

> str(idd)
'data.frame':  24 obs. of  5 variables:
 $ SEC              : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 2 2 ...
 $ Religion         : Factor w/ 2 levels "1","2": 1 1 1 1 2 2 2 2 1 1 ...
 $ Gender           : Factor w/ 2 levels "1","2": 1 1 2 2 1 1 2 2 1 1 ...
 $ Interfaith dating: Factor w/ 2 levels "1","2": 1 2 1 2 1 2 1 2 1 2 ...
 $ Count            : int  11 23 15 69 15 5 19 3 14 26 ...

> levels(idd$SEC) = c("Low", "Middle", "High")
> levels(idd$Religion) = c("Protestant", "Catholic")
> levels(idd$Gender) = c("Male", "Female")
> levels(idd$`Interfaith dating`) = c("Yes", "No")

> str(idd)
'data.frame':  24 obs. of  5 variables:
 $ SEC : Factor w/ 3 levels "Low","Middle",..: 1 1 1 1 1 1 1 1 2 2 ...
 $ Religion: Factor w/ 2 levels "Protestant","Catholic": 1 1 1 1 2 2 2 2 1 1
 $ Gender  : Factor w/ 2 levels "Male","Female": 1 1 2 2 1 1 2 2 1 1 ...
 $ Interfaith dating: Factor w/ 2 levels "Yes","No": 1 2 1 2 1 2 1 2 1 2 ...
 $ Count : int  11 23 15 69 15 5 19 3 14 26 ...

> ggplot(idd) +
+   aes(x= SEC, y = Count, fill = `Interfaith dating`, labels = TRUE) +
+   geom_col(position = "dodge") +
+   facet_grid(Gender ~ Religion) +
+   ggtitle("Relationship between different factors involved in Interfaith da
ting")+
+   xlab("Socioeconomic Class") + ylab("Frequency") +
+   geom_text(aes(label = Count, Count = Count + 0.05), position = position_d
odge(0.9),vjust = 0)
```



Interpretation: If we look at the plot above for each socio economic class we find that by gender female seems to have lesser cases of interfaith dating for Protestant sect. The % is less than 50 for all three – low, middle and high class. That being said, the class is not helping at all to improve the interfaith dating as high class seems to have a poor conversion. For males, it is slightly better but it is still not significant. However, for middle class, we see the ratio of 'yes to no' for interfaith dating is above 50%. Which is the only such case for protestant sect.

Comparatively, for Catholic sect the cases of interfaith dating is always either equal to or higher than the 'no' cases. Having said that, one must keep in mind that the sample data used here has a low volume for this sect. May be we should take a bigger sample size to see if the trend continues.

Overall if we have to identify one sect and one class which proves the ultimate exception for interfaith dating, then it would be females in Catholic sect under low class with a 'yes to no' ratio being 86% i.e., (19/22). This is followed by Catholic sect under low class with a 'yes to no' ratio being 82% i.e., (23/28). On the contrary, the worst 'ye to no' ratio was observed for Protestant sect under low class with a ratio of 18% (15/84).