Project 1

Akshay Kumar Varanasi (av32826)

Instructions

This knitted R Markdown document (as a PDF) *and* the raw R Markdown file (as .Rmd) should both be submitted to Canvas by 4:00pm on **Feb 26th, 2019**. These two documents will be graded jointly, so they must be consistent (as in, don't change the R Markdown file without also updating the knitted document!).

All results presented *must* have corresponding code. **Any answers/results given without the corresponding R code that generated the result will be considered absent.** To be clear: if you do calculations by hand instead of using R and then report the results from the calculations, **you will not receive credit** for those calculations. All code reported in your final project document should work properly. Please do not include any extraneous code or code which produces error messages. (Code which produces warnings is acceptable, as long as you understand what the warnings mean.)

For this project, you will be using the dataset `flavors_of_cacao`. This dataset contains expert ratings of over 1,700 individual chocolate bars, along with information on their regional origin, percentage of cocoa, the variety of chocolate bean used, and where the beans were grown.

```
flavors_of_cacao <-
  read_csv("https://raw.githubusercontent.com/clauswilke/dviz.supp/master/data-
raw/cacao/cacao_clean.csv") %>%
  extract(cocoa_percent, "cocoa_percent", regex = "([^%]+)%", convert = TRUE)
```

```
## Parsed with column specification:
## cols(
##   company = col_character(),
##   bean_origin_detailed = col_character(),
##   REF = col_integer(),
##   review_date = col_integer(),
##   cocoa_percent = col_character(),
##   location = col_character(),
##   rating = col_double(),
##   bean_type = col_character(),
##   bean_origin = col_character()
## )
```

```
head(flavors_of_cacao)
```

```
## # A tibble: 6 x 9
##   company bean_origin_det…  REF review_date cocoa_percent location rating
##   <chr>   <chr>            <int>      <int>         <dbl> <chr>     <dbl>
## 1 A. Mor… Agua Grande       1876       2016            63 France     3.75
## 2 A. Mor… Kpime             1676       2015            70 France     2.75
## 3 A. Mor… Atsane            1676       2015            70 France     3
## 4 A. Mor… Akata             1680       2015            70 France     3.5
## 5 A. Mor… Quilla            1704       2015            70 France     3.5
## 6 A. Mor… Carenero          1315       2014            70 France     2.75
## # … with 2 more variables: bean_type <chr>, bean_origin <chr>
```

The column contents are as follows:

- **company**: name of the company manufacturing the bar.
- **bean_origin_detailed**: the specific geo-region of origin of the bar.
- **REF**: a value linked to when the review was entered in the database. Higher = more recent.
- **review_date**: date of publication of review.
- **cocoa_percent**: cocoa percentage (darkness) of the chocolate bar being reviewed.
- **location**: manufacturer base country.
- **rating**: expert rating for the bar.
- **bean_type**: the variety (breed) of bean used, if provided.
- **bean_origin**: the broad geo-region of origin of the bean.

Problems

Problem 1: (10 pts) Write R code that counts the number of reviews for each company location and calculates a minimum and a maximum ratings of each company location. Filter your output for countries with more than 20 reviews, and order your output from highest to lowest number of reviews.

```
# We want to find number of reviews for each location
flavors_of_cacao %>%
  group_by(location) %>% # First we are grouping the data based on location lik
e "USA","Italy"
  summarize(number_reviews = n(), min_rate = min(rating), max_rate = max(rating
))->
  #summarize counts the number of reviews for each location and calculates a mi
nimum and a maximum ratings of each location
  flavors_of_cacao_location

# To see the outcome of the above
flavors_of_cacao_location
```

```
## # A tibble: 60 x 4
##     location   number_reviews min_rate max_rate
##     <chr>               <int>    <dbl>    <dbl>
##  1 Amsterdam               4     3.25     3.75
##  2 Argentina               9     2.75     3.75
##  3 Australia              49     2.5      4
##  4 Austria                26     2.75     3.75
##  5 Belgium                40     1        4
##  6 Bolivia                 2     2.75     3.75
##  7 Brazil                 17     2.75     4
##  8 Canada                125     2        4
##  9 Chile                   2     3.75     3.75
## 10 Colombia               23     2        4
## # … with 50 more rows
```

```
# Since we want only those places with total number of reviews greater than 20,
we filter the rows and arrange in descending order
flavors_of_cacao_location  %>%
  filter((number_reviews) > 20) %>%
  arrange(desc(number_reviews))-> flavors_of_cacao_new

# To see the resultant data after we filtered accordingly
flavors_of_cacao_new
```

```
## # A tibble: 14 x 4
##     location    number_reviews min_rate max_rate
##     <chr>                <int>    <dbl>    <dbl>
##  1 U.S.A.                 764     1.5      4
##  2 France                 156     1.5      4
##  3 Canada                 125     2        4
##  4 U.K.                    96     1.75     4
##  5 Italy                   63     1.5      5
##  6 Ecuador                 54     1.5      4
##  7 Australia               49     2.5      4
##  8 Belgium                 40     1        4
##  9 Switzerland             38     2        4
## 10 Germany                 35     1.5      4
## 11 Austria                 26     2.75     3.75
## 12 Spain                   25     2.5      4
## 13 Colombia                23     2        4
## 14 Hungary                 22     2.25     3.75
```

Problem 2: (20 pts) Use the data-frame you generated in Problem 1 to find a location with the highest maximum rating and a location with the lowest minimum ratings. Perform a statistical test to determine whether there is a significant difference in ratings between these two locations.

```
# Finding the location with highest maximum rating
flavors_of_cacao_new  %>%
  filter(max_rate==max(flavors_of_cacao_new$max_rate))
```

```
## # A tibble: 1 x 4
##   location number_reviews min_rate max_rate
##   <chr>             <int>    <dbl>    <dbl>
## 1 Italy                63      1.5        5
```

```
# Finding the location with lowest minimum rating
flavors_of_cacao_new  %>%
  filter(min_rate==min(flavors_of_cacao_new$min_rate))
```

```
## # A tibble: 1 x 4
##   location number_reviews min_rate max_rate
##   <chr>             <int>    <dbl>    <dbl>
## 1 Belgium              40        1        4
```

```
# To do the statistical we need to get data of each location into two different
variables

# Knowing that Italy has the highest max rating, we filter data with location I
taly
flavors_of_cacao %>%
  filter(location == 'Italy') -> italy

# Knowing that Belgium has the lowest min rating, we filter data with location
Belgium
flavors_of_cacao %>%
  filter(location == 'Belgium') -> belgium

# Doing t-test between the ratings of the two locations
t.test(italy$rating,belgium$rating)
```

```
##
##   Welch Two Sample t-test
##
## data:  italy$rating and belgium$rating
## t = 1.5475, df = 65.278, p-value = 0.1266
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   -0.06727618  0.53056983
## sample estimates:
## mean of x mean of y
##   3.325397  3.093750
```

From t-test, we can clearly see that they are significantly different as the p-value is small. Even looking at the mean values, mean rating of Italy is 3.325 and mean rating of Belgium is 3.094, so they are different.

Problem 3: (40 pts) Make one plot that visualizes the relationship between the number of reviews and maximum and minimum ratings. Use the data-frame you created in Problem 1. Your code should be well-commented and describe the various steps you take to create this figure. **HINT**: Convert your dataset to a tidy format before you plot.

a. (30 points)

```
# Prints out the Untidy dataset
flavors_of_cacao_new
```

```
## # A tibble: 14 x 4
##    location    number_reviews min_rate max_rate
##    <chr>                <int>    <dbl>    <dbl>
##  1 U.S.A.                 764     1.5      4
##  2 France                 156     1.5      4
##  3 Canada                 125     2        4
##  4 U.K.                    96     1.75     4
##  5 Italy                   63     1.5      5
##  6 Ecuador                 54     1.5      4
##  7 Australia               49     2.5      4
##  8 Belgium                 40     1        4
##  9 Switzerland             38     2        4
## 10 Germany                 35     1.5      4
## 11 Austria                 26     2.75     3.75
## 12 Spain                   25     2.5      4
## 13 Colombia                23     2        4
## 14 Hungary                 22     2.25     3.75
```

```
# Since there are two columns min rate and max rate which is acutually a variab
le so there should
# be one column saying whether it is max or min. There should be another column
with the value of ratings.

# convert to tidy dataset using gather
flavors_of_cacao_new %>%
  gather(min_max,ratings,min_rate:max_rate)-> flavors_of_cacao_tidy

# Prints out the tidy dataset
flavors_of_cacao_tidy
```

```
## # A tibble: 28 x 4
##    location     number_reviews min_max   ratings
##    <chr>                 <int> <chr>       <dbl>
##  1 U.S.A.                  764 min_rate      1.5
##  2 France                  156 min_rate      1.5
##  3 Canada                  125 min_rate      2
##  4 U.K.                     96 min_rate      1.75
##  5 Italy                    63 min_rate      1.5
##  6 Ecuador                  54 min_rate      1.5
##  7 Australia                49 min_rate      2.5
##  8 Belgium                  40 min_rate      1
##  9 Switzerland              38 min_rate      2
## 10 Germany                  35 min_rate      1.5
## # … with 18 more rows
```

```
ggplot(flavors_of_cacao_tidy, aes(x=number_reviews, y=ratings, color=min_max))+
geom_point()+geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```
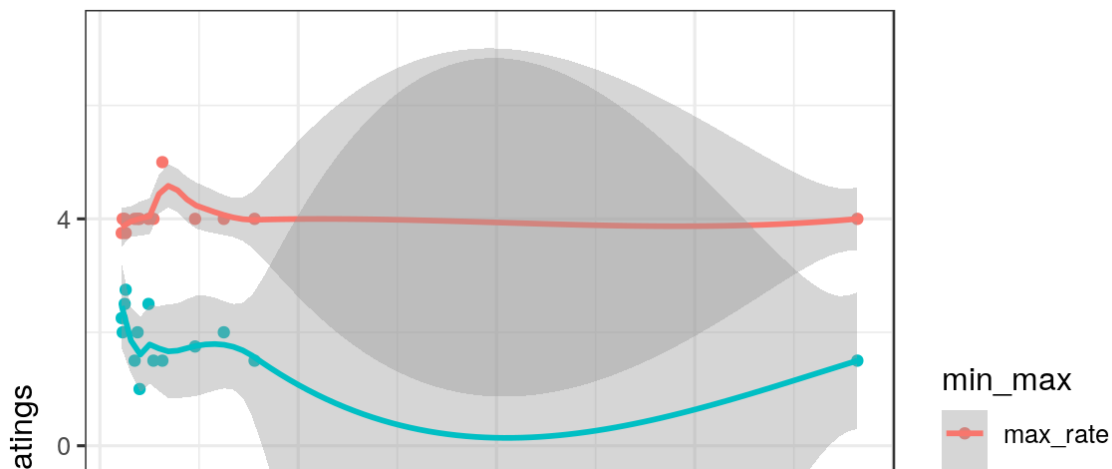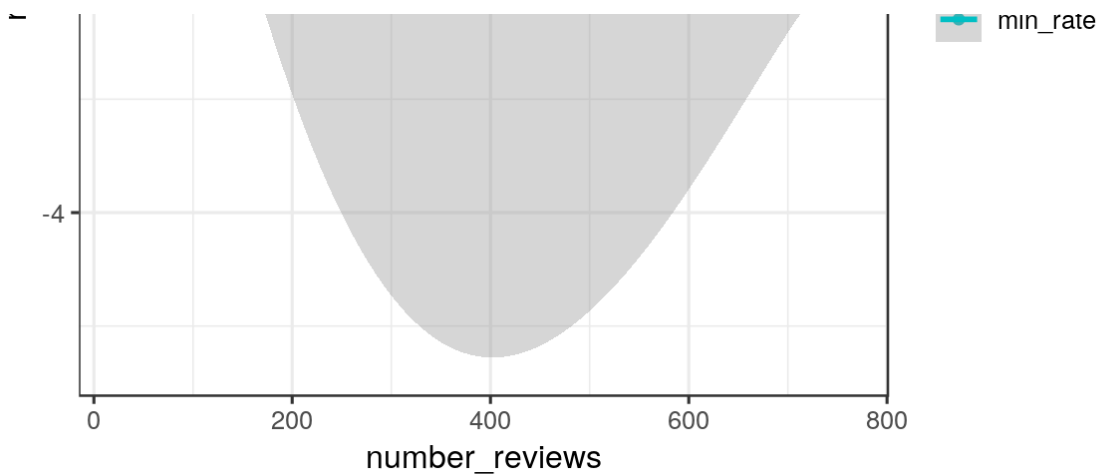
b. (10 points) Discuss the information (overarching trends, patterns, etc.) your plot reveals. Be sure to include in your discussion the similarities/differences among minimum and maximum ratings. Your discussion should also explain the results of the t-test in Problem 2 in the context of this plot. Be sure to also include a clear, logical justification for why you selected the particular geom(s) used to represent this data. Please limit your full response to a maximum of 10 sentences.

Answer We used geom_points to plot them as there are lot of data points and we want to see how they are spread. Along with that we use geom_smooth to see the trend within them. We colored points describing min rating and max rating differently, as they are two different things plotted in same graph and we are interested to study trends within them.

From the plot, we can see that there is very little variation in max rating compared to min rating. As you can see, for min rating even though the number of reviews is almost same, ratings are quite different. While in max rating, the ratings don't change with number of reviews. So both min and max ratings are not really effected by number of reviews.

In previous question, we did t-test between Italy and Belgium with number of reviews being 63 and 40 respectively. We found out that the distribution of ratings for these locations are quite different even though number of reviewzs are close enough. So number of reviews does not tell us anything about the distribution, maybe it depends on the location rather than number of reviews or some other thing.

Problem 4: (30 pts) Think of **one** (and only one!) conceptual question to ask about the data set `flavors_of_cacao` . Clearly state your question in the space provided below. Use the ggplot2 library to create a plot that can help you find an answer to the question. For the plot, provide a clear explanation as to why this type of plot (e.g. boxplot, barplot, histogram, etc.) is best for providing the information you are asking about. Answer your question by interpreting your plot and identifying any trends it reveals, or does not reveal, as the case may be. Please limit the discussion to 4-6 sentences.

To receive full credit for Problem 4, we look for the following for a question:

- *A clear, coherent question about the data. (Questions end in a question mark!)*
- *The question should be conceptual and **should not** prompt a specific analysis or plot.*
- *A plot that helps answer your proposed question, with a justification for why you chose to make the type of plot that you made.*

- *An interpretation of your plot and a response to your proposed question.*
- *Statistical analysis **is not** necessary. Just interpret your plot.*

You cannot reuse the questions about the `flavors_of_cacao` data set from the previous problems.

Question

How different are the ratings for the company with higest and lowest number of reviews considering only the companies with minimum 20 reviews?

```
head(flavors_of_cacao)
```

```
## # A tibble: 6 x 9
##   company bean_origin_det…   REF review_date cocoa_percent location rating
##   <chr>   <chr>            <int>       <int>         <dbl> <chr>     <dbl>
## 1 A. Mor… Agua Grande       1876        2016            63 France     3.75
## 2 A. Mor… Kpime             1676        2015            70 France     2.75
## 3 A. Mor… Atsane            1676        2015            70 France     3
## 4 A. Mor… Akata             1680        2015            70 France     3.5
## 5 A. Mor… Quilla            1704        2015            70 France     3.5
## 6 A. Mor… Carenero          1315        2014            70 France     2.75
## # … with 2 more variables: bean_type <chr>, bean_origin <chr>
```

Answer

First we need to group by company to calculate number of reviews for each company to find number of reviews for each company.

```
# We want to find number of reviews for each company
flavors_of_cacao %>%
  group_by(company) %>% # First we are grouping the data based on company like
"Soma","Bonnat"
  summarize(number_reviews = n(), min_rate = min(rating), max_rate = max(rating
)) %>%
  #summarize counts the number of reviews for each location and calculates a mi
nimum and a maximum ratings of each location
  arrange(desc(number_reviews))-> # arranging them in descending order of numbe
r of reviews
  flavors_of_cacao_company
flavors_of_cacao_company
```

```
## # A tibble: 416 x 4
##    company                    number_reviews min_rate max_rate
##    <chr>                               <int>    <dbl>    <dbl>
##  1 Soma                                   47     2.75        4
##  2 Bonnat                                 27     1.5         4
##  3 Fresco                                 26     2.75        4
##  4 Pralus                                 25     2           4
##  5 A. Morin                               23     2.75        4
##  6 Arete                                  22     2.75        4
##  7 Domori                                 22     3           4
##  8 Guittard                               22     2.5      3.75
##  9 Valrhona                               21     1.5         4
## 10 Hotel Chocolat (Coppeneur)             19     2.5       3.5
## # … with 406 more rows
```

Then we need to filter company with reviews more than 20 as to compare the distribution we need good number of points. (Here it is 20)

```
# We want only companies with minimum 20 number of reviews so filtering compani
es with less reviews than 20
flavors_of_cacao_company %>%
  filter((number_reviews) > 20) %>%
  arrange(desc(number_reviews))-> flavors_of_cacao_company
flavors_of_cacao_company
```

```
## # A tibble: 9 x 4
##   company  number_reviews min_rate max_rate
##   <chr>             <int>    <dbl>    <dbl>
## 1 Soma                 47     2.75        4
## 2 Bonnat               27     1.5         4
## 3 Fresco               26     2.75        4
## 4 Pralus               25     2           4
## 5 A. Morin             23     2.75        4
## 6 Arete                22     2.75        4
## 7 Domori               22     3           4
## 8 Guittard             22     2.5      3.75
## 9 Valrhona             21     1.5         4
```

From the above data frame, we can see that company with highest number of reviews is "Soma" and lowest is "Valrhona". We can make sure if this is correct by filtering based on highest and lowest number of ratings. We want to see how different are the distribution of ratings. We need to store the ratings data for each company different dataframe so that we can compare them.

```
# Finding the company with highest num of reviews
flavors_of_cacao_company  %>%
  filter(number_reviews==max(flavors_of_cacao_company$number_reviews))
```

```
## # A tibble: 1 x 4
##   company number_reviews min_rate max_rate
##   <chr>            <int>    <dbl>    <dbl>
## 1 Soma                47     2.75        4
```

```
# Finding the company with lowest num of reviews
flavors_of_cacao_company  %>%
  filter(number_reviews==min(flavors_of_cacao_company$number_reviews))
```

```
## # A tibble: 1 x 4
##   company  number_reviews min_rate max_rate
##   <chr>             <int>    <dbl>    <dbl>
## 1 Valrhona             21      1.5        4
```

```
# Storing ratings for each company
flavors_of_cacao %>%
  filter(company == 'Soma') -> Soma

flavors_of_cacao %>%
  filter(company == 'Valrhona') -> Valrhona
```

To compare their distribution we can either plot boxplot or do a t-test.

```
# T test
t.test(Soma$rating,Valrhona$rating)
```

```
##
##  Welch Two Sample t-test
##
## data:  Soma$rating and Valrhona$rating
## t = 1.6947, df = 24.846, p-value = 0.1026
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   -0.05429264  0.55783874
## sample estimates:
## mean of x mean of y
##  3.585106  3.333333
```

Seeing that the p-value is low, we can say that distribution is quite different for both of them. We can

infer the same by looking at the means, we can see that values of mean of ratings for Soma is 3.58 and for Valrhona is 3.33, since the means are quite different we infer that the distribution is also quite different. T-test only says this but it does not say how they actually look or what is the reason behind it. For this boxplot gives better picture as it tell us more about distribution itself.

```
boxplot(Soma$rating,Valrhona$rating,names = c('Soma','Valrhona'))
```
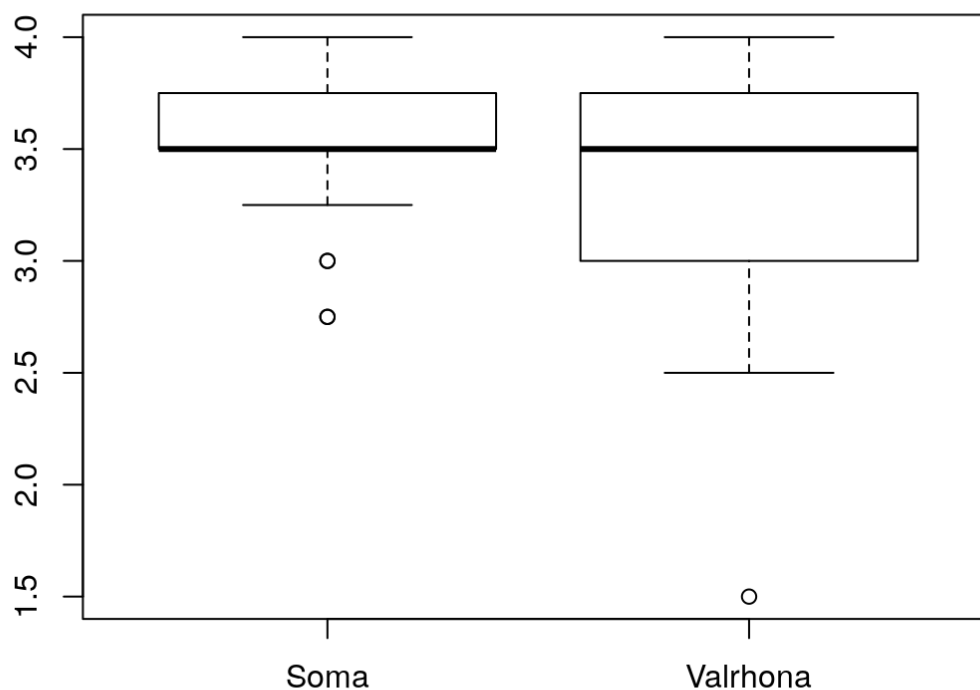

When we plot the box-plot for the two companies ratings, we see that Soma's distribution is not spread over like Valrohna. Both have outliers as we can see above. It may seem both are similar seeing the line in the plot but that is median, but it does not give us information of distribution correctly. We need to look at both mean and distribution itself. The t-test and box plot each give tell us different aspect of distribution. Seeing both the t-test and box-plot, we can infer that they are quite different.