Midterm Exam

Zixuan Mary Liu

11/2/2020

Instruction

This is your midterm exam that you are expected to work on it alone. You may NOT discuss any of the content of your exam with anyone except your instructor. This includes text, chat, email and other online forums. We expect you to respect and follow the GRS Academic and Professional Conduct Code.

Although you may NOT ask anyone directly, you are allowed to use external resources such as R codes on the Internet. If you do use someone's code, please make sure you clearly cite the origin of the code.

When you finish, please compile and submit the PDF file and the link to the GitHub repository that contains the entire analysis.

Introduction

In this exam, you will act as both the client and the consultant for the data that you collected in the data collection exercise (20pts). Please note that you are not allowed to change the data. The goal of this exam is to demonstrate your ability to perform the statistical analysis that you learned in this class so far. It is important to note that significance of the analysis is not the main goal of this exam but the focus is on the appropriateness of your approaches.

Data Description (10pts)

Please explain what your data is about and what the comparison of interest is. In the process, please make sure to demonstrate that you can load your data properly into R.

The data I collected is about what restaurant I usually go to. I want to make my own system of ranking to see what restaurant I like the best. I made this data since it contains the arrive time from my apartment to the specific restaurant and how frequent I go there. This data contains 8 variables, I used mostly Google map to create my data, I also went to my bank statement to see how the average cost for each meal and how often I go there.

```
# Load the data set and view it
data <- read.csv('C:/Users/49431/Downloads/678_my_data.csv')
head(data,10)
```

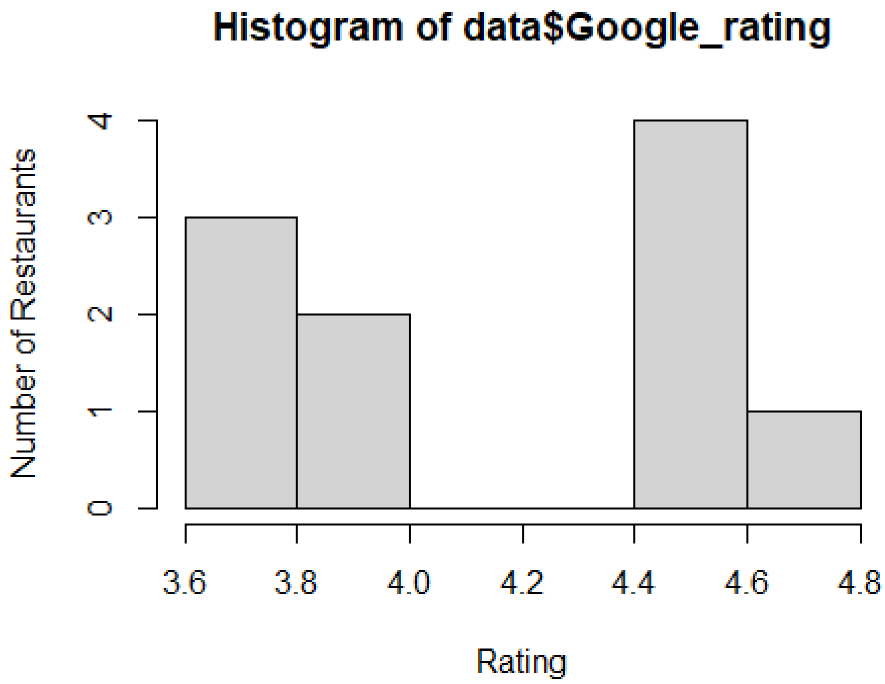
```
##    ï..Restaurant_id        Restaurant_name             Cuisines
Arrive_time
## 1                 1                  Tatte                CafÃ©
```

```
15
## 2                 2 Terra at Eataly Boston   Italian restaurant
19
## 3                 3             Happy Lamb   Hot pot restaurant
17
## 4                 4               Chipotle   Mexican restaurant
15
## 5                 5            Shake Shack Hamburger restaurant
16
## 6                 6         Six Po Hot Pot   Chinese restaurant
24
## 7                 7               Gyu-Kaku  Japanese restaurant
18
## 8                 8            Shinmio Tea     Bubble tea store
18
## 9                 9          Su Su Gourmet   Chinese restaurant
18
## 10               10                   Tora  Japanese Restaurant
16
##    Frequency_per_month Average_Cost Google_reviews Google_rating
## 1                    8           20           1225           4.6
## 2                    1           60            175           3.8
## 3                    5           50            756           4.5
## 4                   20           10            298           3.9
## 5                   15           15            208           3.8
## 6                    4           60            165           4.0
## 7                    2           60           1250           4.5
## 8                   24            7             41           4.8
## 9                    7           30             76           3.6
## 10                   9           44            299           4.5
```

EDA (10pts)

Please create one (maybe two) figure(s) that highlights the contrast of interest. Make sure you think ahead and match your figure with the analysis. For example, if your model requires you to take a log, make sure you take log in the figure as well.

```
# Plot the distribution of average ratings to see if analysis is applicable
hist(data$Google_rating, xlab = "Rating", ylab = "Number of Restaurants")
```



Shown in below is a correlation map for my data that describes the relationship between the different features. The heatmap below shows that all numeric variables have a positive correlation. average_cost, Arrive_time and Google_reviews, Google_rating have especially high positive correlation. Average_Cost and Frequency_per_month have high negative corrlation.

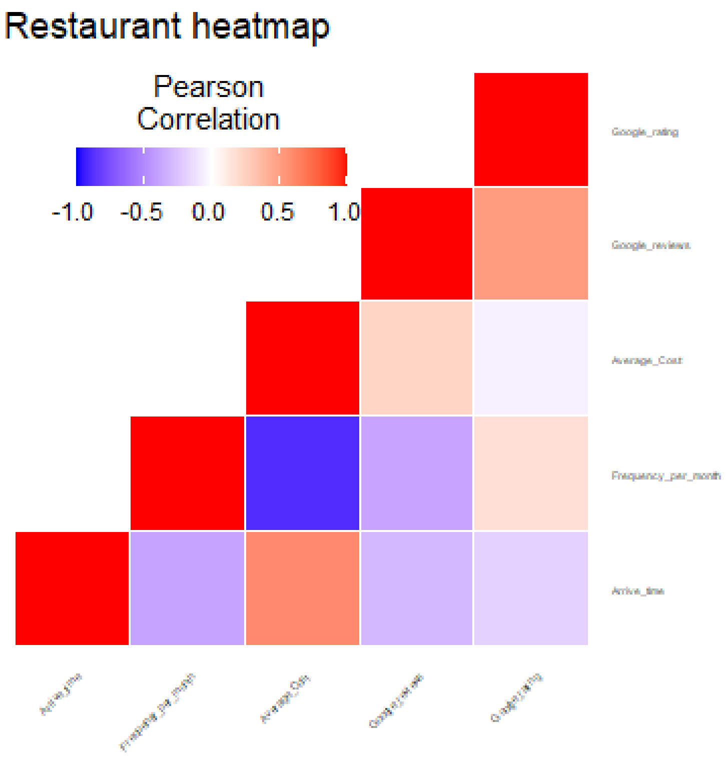
```
library(stats)
library(reshape2)

##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##     smiths

#heatmap plot data
temp <- data[4:8]
cormat <- round(cor(temp),2)
melted_cormat <- melt(cormat)
  # Get upper triangle of the correlation matrix
  get_upper_tri <- function(cormat){
    cormat[lower.tri(cormat)]<- NA
    return(cormat)
  }
upper_tri <- get_upper_tri(cormat)
# Melt the correlation matrix
melted_cormat <- melt(upper_tri, na.rm = TRUE)
# Create a ggheatmap
```

```r
ggheatmap <- ggplot(melted_cormat, 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()+ # minimal theme
 theme(axis.text.x = element_text(angle = 45, vjust = 1,
    size = 12, hjust = 1))+
 coord_fixed() + ggtitle("Restaurant heatmap")
# Print the heatmap
ggheatmap +
theme(axis.text.x = element_text(size=4),
      axis.text.y = element_text(size=4),
  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(0.6, 0.7),
  legend.direction = "horizontal")+
  guides(fill = guide_colorbar(barwidth = 7, barheight = 1,
              title.position = "top", title.hjust = 0.5)) +
  scale_y_discrete(position = "right")
```



Power Analysis (10pts)

Please perform power analysis on the project. Use 80% power, the sample size you used and infer the level of effect size you will be able to detect. Discuss whether your sample size was enough for the problem at hand. Please note that method of power analysis should match the analysis. Also, please clearly state why you should NOT use the effect size from the fitted model.

```
pwr.anova.test(k=8, n=10, sig.level = 0.05, power = 0.8)
```

```
## 
##      Balanced one-way analysis of variance power calculation 
## 
##               k = 8
##               n = 10
##               f = 0.4443177
##       sig.level = 0.05
##           power = 0.8
## 
## NOTE: n is number in each group
```

```
pwr.anova.test(k=8, f = 0.25, sig.level = 0.05, power = 0.8)
```

```
## 
##      Balanced one-way analysis of variance power calculation 
## 
##               k = 8
##               n = 29.59154
##               f = 0.25
##       sig.level = 0.05
##           power = 0.8
## 
## NOTE: n is number in each group
```

For my dataset, I have 8 observations in each group, therefore I choose to use power anova test since I have more than 3 variables. The effect size is 0.4443177 which is a bit large. I choose to use 0.25 for variance comparison, since it is an appropriate value for effect size. Then I found out that I have to have around 30 observations in each group to have 0.25 effect size. However, it might cause Type M error if the effect size is too big.

Modeling (10pts)

Please pick a regression model that best fits your data and fit your model. Please make sure you describe why you decide to choose the model. Also, if you are using GLM, make sure you explain your choice of link function as well.

I made a rating system that's on top of my head, which does not require look into the data too much. The scoring algorithm used to rank the restaurants takes into account the ratings and number of reviews of each restaurant and uses the True Bayesian Estimate method. This method will account for situations such as when a restaurant has a high rating but very few number of reviews, which will not be as reliable as a restaurant having a lower rating but a lot of reviews. The method uses the formula:

weighted rating (WR) = $(v \div (v+m)) \times R + (m \div (v+m)) \times C$ where:

R = Rating

v = number of reviews

m = minimum reviews required

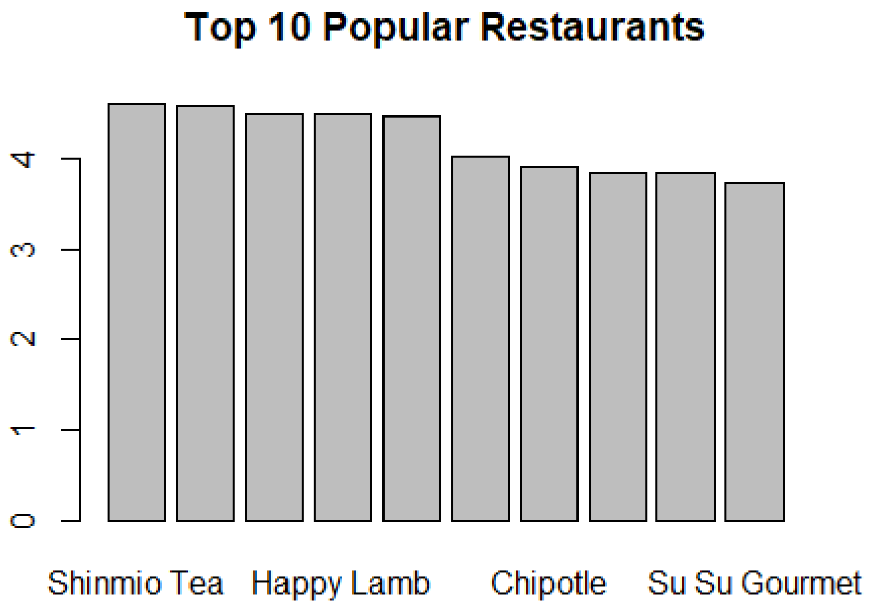
C = the mean rating across the whole report

I will choose the minimum reviews required as 20 which is reasonable for a restaurant.

```
C <- mean(data$Google_rating, na.rm=TRUE)
m <- 20
w <- data$Google_reviews/(data$Google_reviews + m)
data$Score <- w*data$Google_rating + (1-w)*C
x = data[order(data$Score, decreasing= T),]
top10 <- head(x, 10)
top10$Restaurant
```

```
##  [1] "Shinmio Tea"             "Tatte"                    "Gyu-Kaku"
##  [4] "Happy Lamb"              "Tora"                     "Six Po Hot Pot"
##  [7] "Chipotle"                "Terra at Eataly Boston" "Shake Shack"
## [10] "Su Su Gourmet"
```

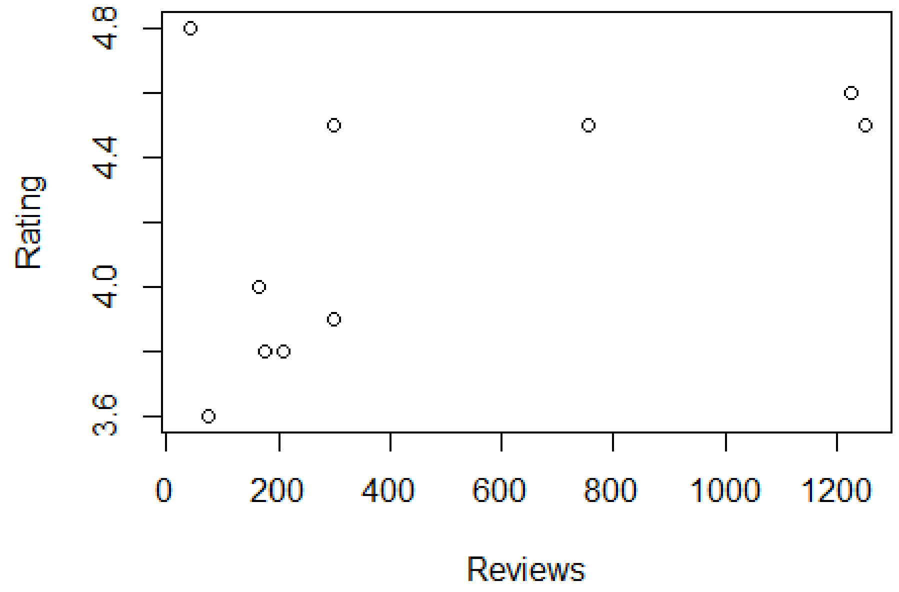
```
barplot(top10$Score, main="Top 10 Popular Restaurants",
names.arg=as.character(top10$Restaurant))
```



The ranking of my 10 Restaurants are shown together with a barplot of their weighted score. It turned out that I like to drink Shinmio Tea the most. Milktea > any other food to me.

```r
# Convert cleaned Rating and Votes column into numbers from factor level
numericRatings = as.numeric(as.character(data$Google_rating))
numericReviews = as.numeric(as.character(data$Google_reviews))

# Plot Rating vs Reviews to see if linear regression is applicable
plot(numericRatings ~ numericReviews, xlab = "Reviews", ylab = "Rating")
```

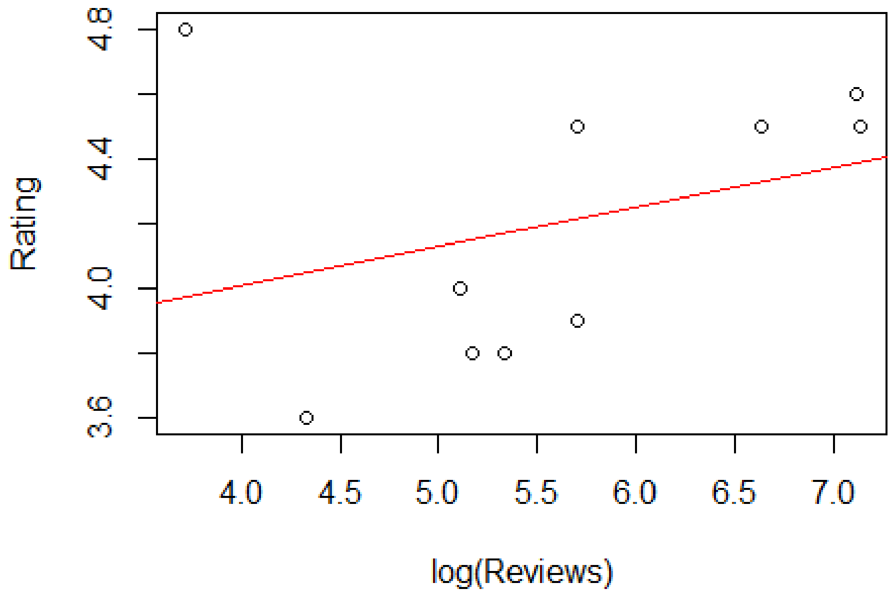


```
# Transform the data by taking the log(Reviews) to standardize data and
remove large values
logReviews = log(numericReviews)
plot(numericRatings ~ logReviews, xlab = "log(Reviews)", ylab = "Rating")

# Build linear regression model
model1 = lm(numericRatings ~ logReviews)
summary(model1)

##
## Call:
## lm(formula = numericRatings ~ logReviews)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4469 -0.3393 -0.0139  0.2053  0.8280
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.5212     0.7143   4.930  0.00115 **
## logReviews    0.1214     0.1255   0.967  0.36167
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4231 on 8 degrees of freedom
## Multiple R-squared:  0.1047, Adjusted R-squared:  -0.007175
## F-statistic: 0.9359 on 1 and 8 DF,  p-value: 0.3617
```

```
abline(model1, col = "red")
```



```
# Add logReviews to the data set for easier analysis
data$logReviews= logReviews

# Try more linear regression models
model2 = lm(numericRatings ~ numericReviews)
summary(model2)

##
## Call:
## lm(formula = numericRatings ~ numericReviews)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42639 -0.26173 -0.07008  0.12783  0.78989
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.9910449  0.1749771  22.809 1.45e-08 ***
## numericReviews 0.0004651  0.0002794   1.664    0.135
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3854 on 8 degrees of freedom
## Multiple R-squared:  0.2572, Adjusted R-squared:  0.1644
## F-statistic:  2.77 on 1 and 8 DF,  p-value: 0.1346
```

```
# Try more linear regression models
model3 = lm(data$Score ~ numericReviews + data$Frequency_per_month +
data$Average_Cost+ data$Arrive_time + numericRatings)
summary(model3)

##
## Call:
## lm(formula = data$Score ~ numericReviews + data$Frequency_per_month +
##     data$Average_Cost + data$Arrive_time + numericRatings)
##
## Residuals:
##         1         2         3         4         5         6         7
8
## -0.003478 -0.043606  0.009242  0.010205 -0.011453  0.016977 -0.005658 -
0.018224
##         9        10
##  0.020128  0.025868
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               8.368e-01  1.466e-01   5.707  0.00466 **
## numericReviews            2.409e-05  4.326e-05   0.557  0.60721
## data$Frequency_per_month -6.965e-03  4.638e-03  -1.502  0.20755
## data$Average_Cost        -9.303e-04  1.535e-03  -0.606  0.57707
## data$Arrive_time         -5.586e-03  6.022e-03  -0.928  0.40611
## numericRatings            8.454e-01  3.994e-02  21.166 2.95e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03147 on 4 degrees of freedom
## Multiple R-squared:  0.9966, Adjusted R-squared:  0.9924
## F-statistic:   235 on 5 and 4 DF,  p-value: 5.017e-05
```

Looking at the outputs, we can see that it is not statistically significant that the natural log(reviews), but it's statistically significant for numericReviews and the r square for my model 2 and model 3 is 0.1644, 0.0024 respectably. But it has a positive effect on the ratings. This is reflected in the regression line plotted in red.

Validation (10pts)

Please perform a necessary validation and argue why your choice of the model is appropriate.

I think linear regression is the best fit for my model.

```
summary(model3)

##
## Call:
## lm(formula = data$Score ~ numericReviews + data$Frequency_per_month +
##     data$Average_Cost + data$Arrive_time + numericRatings)
```

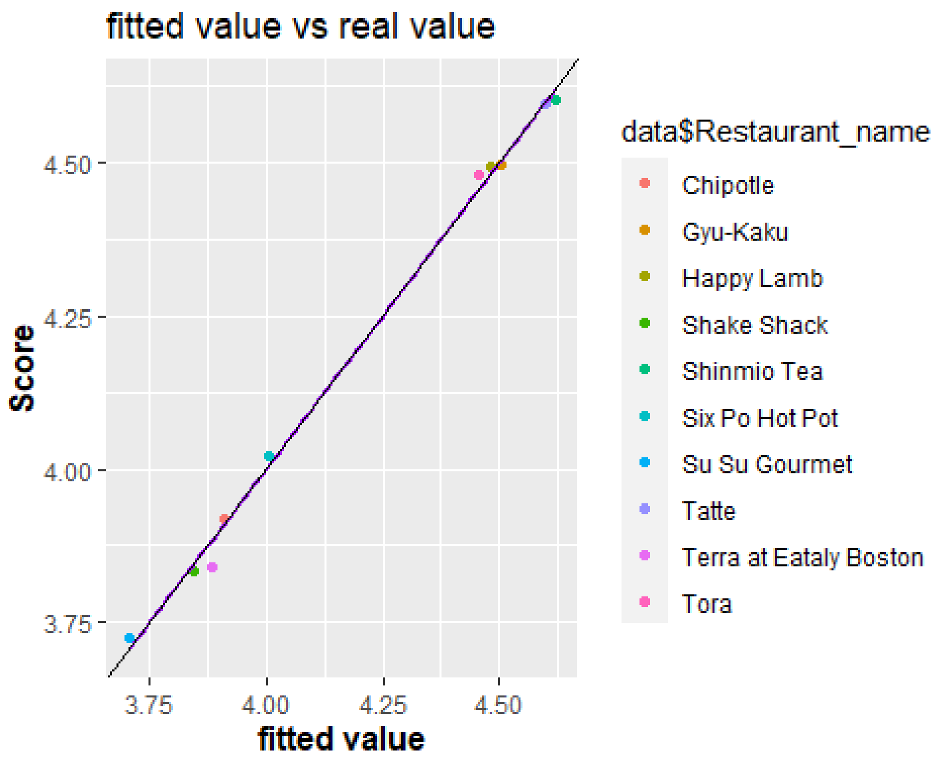
```
##
## Residuals:
##         1         2         3         4         5         6         7
8
## -0.003478 -0.043606  0.009242  0.010205 -0.011453  0.016977 -0.005658 -
0.018224
##         9        10
##  0.020128  0.025868
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               8.368e-01  1.466e-01   5.707  0.00466 **
## numericReviews            2.409e-05  4.326e-05   0.557  0.60721
## data$Frequency_per_month -6.965e-03  4.638e-03  -1.502  0.20755
## data$Average_Cost        -9.303e-04  1.535e-03  -0.606  0.57707
## data$Arrive_time         -5.586e-03  6.022e-03  -0.928  0.40611
## numericRatings            8.454e-01  3.994e-02  21.166 2.95e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03147 on 4 degrees of freedom
## Multiple R-squared:  0.9966, Adjusted R-squared:  0.9924
## F-statistic:   235 on 5 and 4 DF,  p-value: 5.017e-05
```

For my finalized model, the R-squared is 0.9924 which is pretty close to 1, but the p-value is significant only for numericRatings. The r-squared implied it's a good fit, but p-value and power test imply that this is not.

```
fitted <- 8.368e-01+ 2.409e-05*numericReviews-6.965e-
03*data$Frequency_per_month-9.303e-04*data$Average_Cost-5.586e-
03*data$Arrive_time+8.454e-01*numericRatings
ggplot(data)+
geom_point(mapping=aes(x=fitted,y=Score,color=data$Restaurant_name))+
geom_smooth(mapping=aes(x=fitted,y=Score),method =
"lm",color="purple",se=F)+geom_abline(intercept = 0,slope = 1)+
theme(axis.text = element_text(size = 9),
axis.title = element_text(size = 12, face = "bold")) +
labs(title = "fitted value vs real value",y = "Score",x = "fitted value")

## Warning: Use of `data$Restaurant_name` is discouraged. Use
`Restaurant_name`
## instead.

## `geom_smooth()` using formula 'y ~ x'
```



This plot shows that when I compared with the line y=x, my fitted line(the purple one) is almost overlapped with the black line (y=x),and all my dots (restaurant_name) is around my fitted line.

Inference (10pts)

Based on the result so far please perform statistical inference to compare the comparison of interest.

```
confint(model3)
```

```
##                                  2.5 %      97.5 %
## (Intercept)                4.296785e-01 1.243956575
## numericReviews            -9.600203e-05 0.000144190
## data$Frequency_per_month -1.984070e-02 0.005911061
## data$Average_Cost         -5.190750e-03 0.003330174
## data$Arrive_time          -2.230633e-02 0.011133933
## numericRatings             7.345031e-01 0.956295572
```

```
coef(model3)
```

```
##              (Intercept)           numericReviews data$Frequency_per_month
##             8.368175e-01             2.409397e-05            -6.964822e-03
##        data$Average_Cost         data$Arrive_time           numericRatings
##            -9.302877e-04            -5.586201e-03             8.453994e-01
```

I found the confidence interval for the slopes. On the 97.5% confidence interval, the true slope of reviews is -9.600203e-05 to 0.000144190, -1.984070e-02 to 0.005911061 for

Frequency_per_month, -5.190750e-03 to 0.003330174 for Average_Cost, -2.230633e-02 to 0.011133933 for arrive time and 7.345031e-01 to 0.956295572 for ratings.

For coefficient, if all variables is null, then my score is 0.84, and if one unit increased in reviews, then my score increase 0.00002. if one unit increased in Frequency_per_month, then my score decrease by 0.00696. if one unit increased in Average_Cost, then my score decrease by 0.00093. if one unit increased in Arrive_time, then my score decrease by 0.00559.if one unit increased in numericRatings, then my score increased by 0.85

Discussion (10pts)

Please clearly state your conclusion and the implication of the result.

The model I use is:$Score = 8.368e - 01 + 2.409e - 05 * Reviews - 6.965e - 03 * FrequencyPerMonth - 9.303e - 04 * AverageCost - 5.586e - 03 * ArriveTime + 8.454e - 01 * Ratings$. This is a linear regression model that I found out is the best fit for my dataset. By looking at the heatmap in EDA part, we can see that all numeric variables does not have a high correlation between each other. Even though r-square is close to 1 but the p-value is not so significant, I dont think that it's an optimal model for the effect size. I need to find more restaurant to increase my data size, in order to have a better effect size of my model. One reason to cause my model turn out this way might be that I did not chose the best scoring algorithm. If given more time, I will choose another scoring function to better fit my model.

Limitations and future opportunity. (10pts)

Please list concerns about your analysis. Also, please state how you might go about fixing the problem in your future study.

Most of the p-value is greater than 0.05 which means the variables of linear regression is not significant.
I should collect more observations and do ANOVA as well for future work. And also try some other scoring algorithm for my data.

Comments or questions

If you have any comments or questions, please write them here.