Statistical Foundations of Data Analytics

LAB 2: MULTIPLE LINEAR REGRESSION I

Please print this booklet and answer all questions within the space provided. Please type the assignment and use R.

YOUR TEAM

Last name: Page
First name: Skylar
Student ID: 260711809

Last name: Li
First name: Meredith
Student ID: 260705801

Last name: Sandoval
First name: Santiago
Student ID: 260806090

Last name: Toronga
First name: Nicholas
Student ID: 260715831

Lab 2: What makes a restaurant more popular? A multiple regression approach

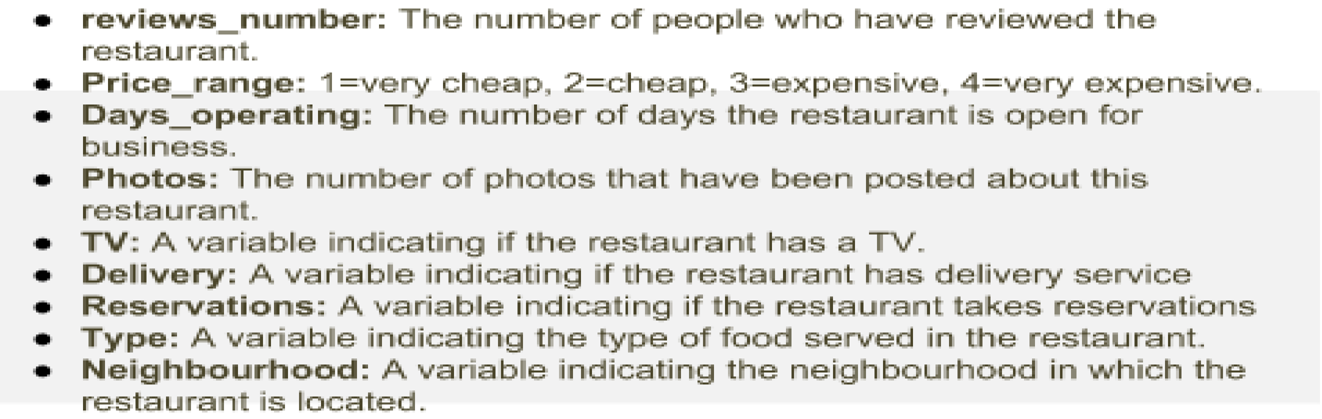
In this lab, we will continue exploring the question of: what makes a restaurant popular? Last time, we used the simple linear regression to do separate regressions. Today, we will use the techniques from Lecture 3, and use multiple linear regression on our Yelp.com data. You need to use *R* (any other software is not acceptable). We will use the following variables in our regression. Please attach your code at the end of the assignment.

Response Variable

Avg_star_rating: The average rating of the restaurant. This variable ranges from one to five.

One= Restaurant is terrible; Five= restaurant is amazing.

Predictors



- **reviews_number:** The number of people who have reviewed the restaurant.
- **Price_range:** 1=very cheap, 2=cheap, 3=expensive, 4=very expensive.
- **Days_operating:** The number of days the restaurant is open for business.
- **Photos:** The number of photos that have been posted about this restaurant.
- **TV:** A variable indicating if the restaurant has a TV.
- **Delivery:** A variable indicating if the restaurant has delivery service
- **Reservations:** A variable indicating if the restaurant takes reservations
- **Type:** A variable indicating the type of food served in the restaurant.
- **Neighbourhood:** A variable indicating the neighbourhood in which the restaurant is located.

:

1. Individual regressions (3 points)

Run four simple linear regressions ($Y=b_o+b_1x$), one for each of the four predictors. For each regression, find the t-statistic and p-value of the coefficient (for the predictor variable).

Note: the dataset is different from the one used in Lab 1.

avg_star_rating $=b_o+b_1$(reviews_number)

- b_o:3.8513588
- b_1:0.0008757
- r-squared: 0.02612
- t-test statistic:2.379
- p-value:0.01825

avg_star_rating $=b_o+b_1$(price_range)

- b_o: 4.01845
- b_1:-0.04794
- r-squared:0.003497
- t-test statistic:-0.86
- p-value:0.3905

avg_star_rating$=b_o+b_1$(days_operating)

- b_o:4.60995
- b_1:-0.10517
- r-squared:0.0226
- t-test statistic:-2.209
- p-value:0.02825

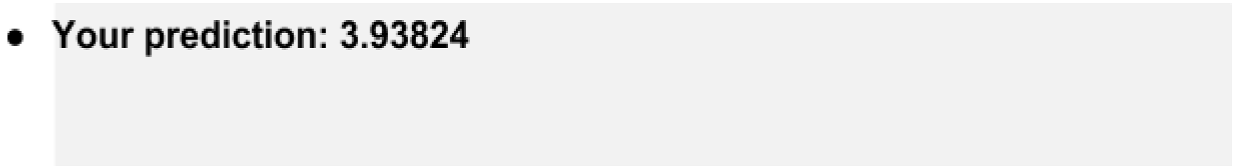
avg_star_rating$=b_o+b_1$(photos)

- b_o:3.8374660
- b_1:0.0006842
- r-squared:0.04209
- t-test statistic:3.045
- p-value:0.002623

2. Predictions (5 points)

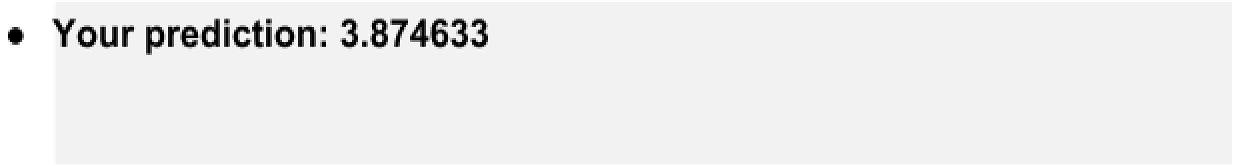
Based on the above results, predict the number of stars that a restaurant will receive if:

1) Restaurant has 100 reviews:



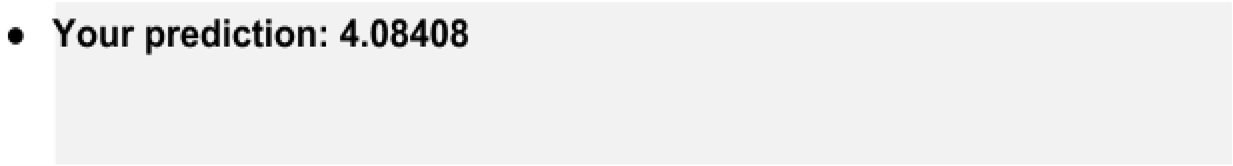
- **Your prediction: 3.93824**

2) Price range=3 ($$$):



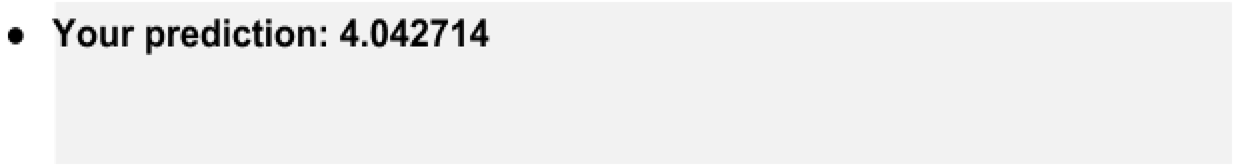
- **Your prediction: 3.874633**

3) It is open five days a week:



- **Your prediction: 4.08408**

4) People have published 300 photos:



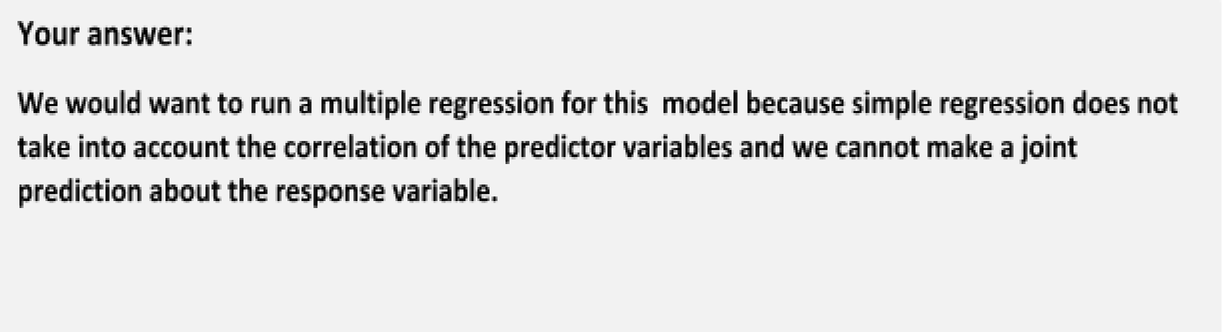
- **Your prediction: 4.042714**

3. Multiple regression (5 points)

Suppose we are thinking of running the following multiple regression:

$$avg\_star\_rating = b_o + b_1(reviews\_number) + b_2(price\_range) + b_3(days\_operating) + b_4(\text{photos})$$

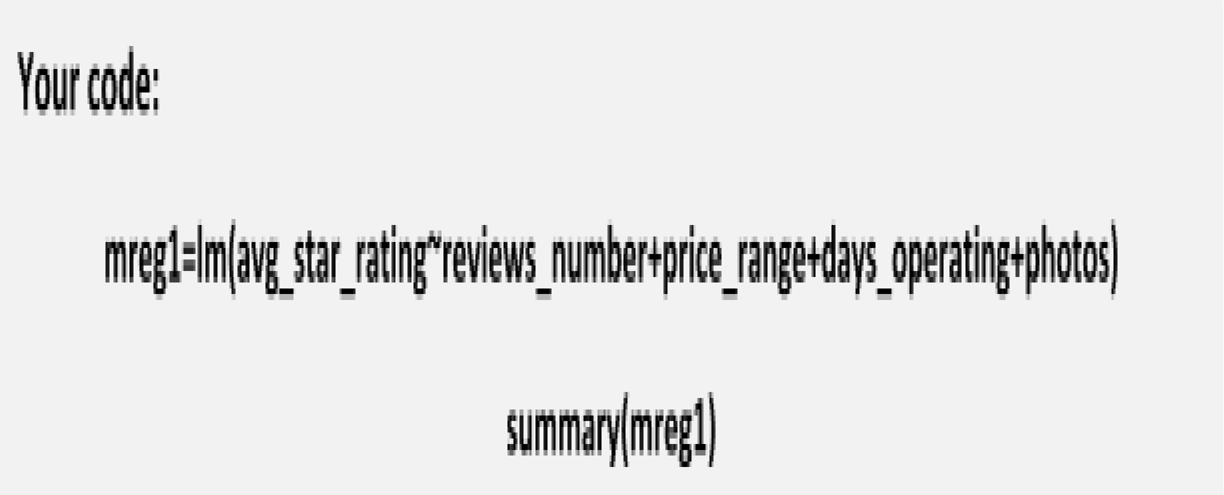
a) Why would we want to do this, as opposed to four separate simple linear regressions? Answer in the space provided below.



Your answer:

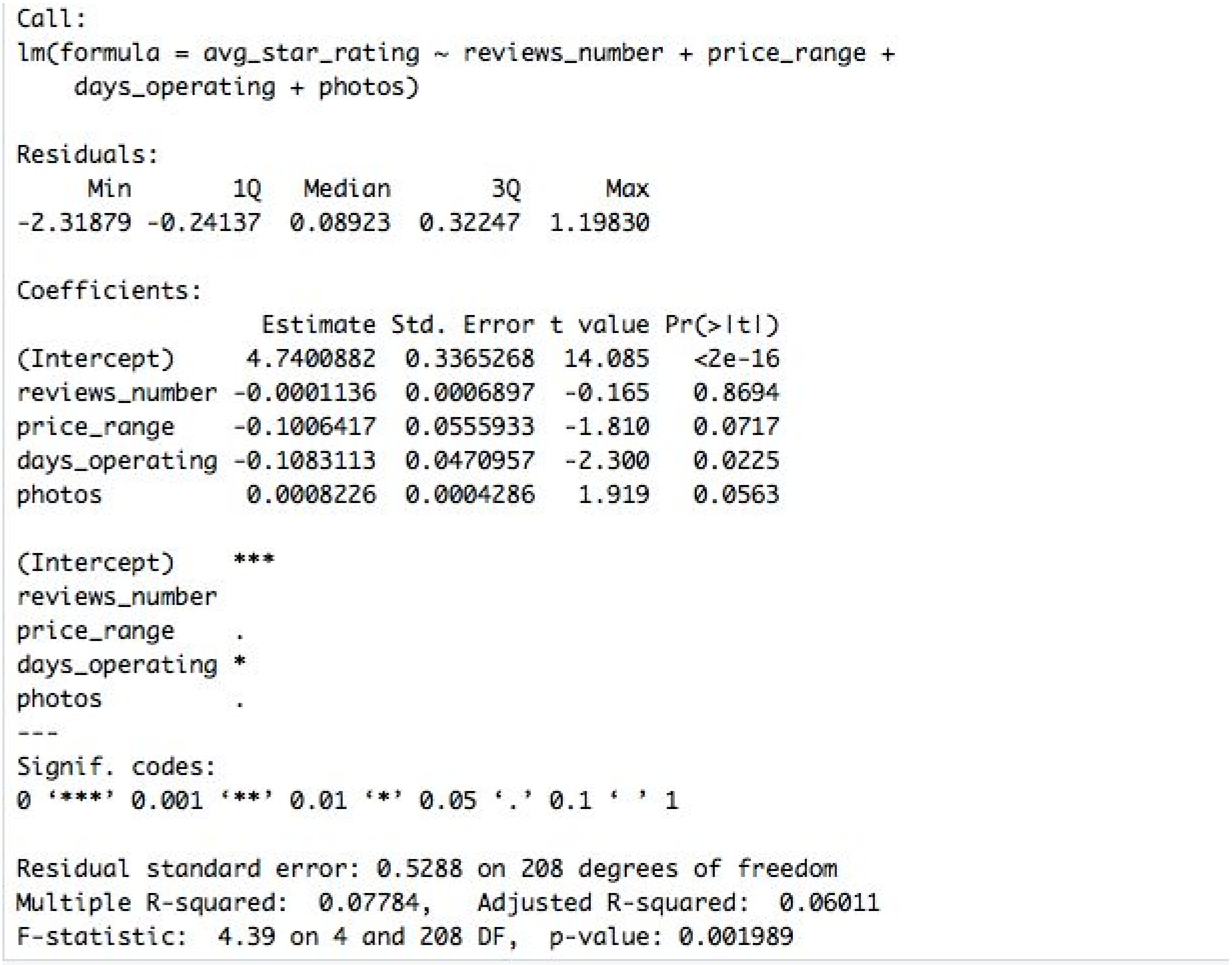
We would want to run a multiple regression for this model because simple regression does not take into account the correlation of the predictor variables and we cannot make a joint prediction about the response variable.

Now, run the multiple regression and provide the R-output below:



Your code:

```
mreg1=lm(avg_star_rating~reviews_number+price_range+days_operating+photos)

summary(mreg1)
```

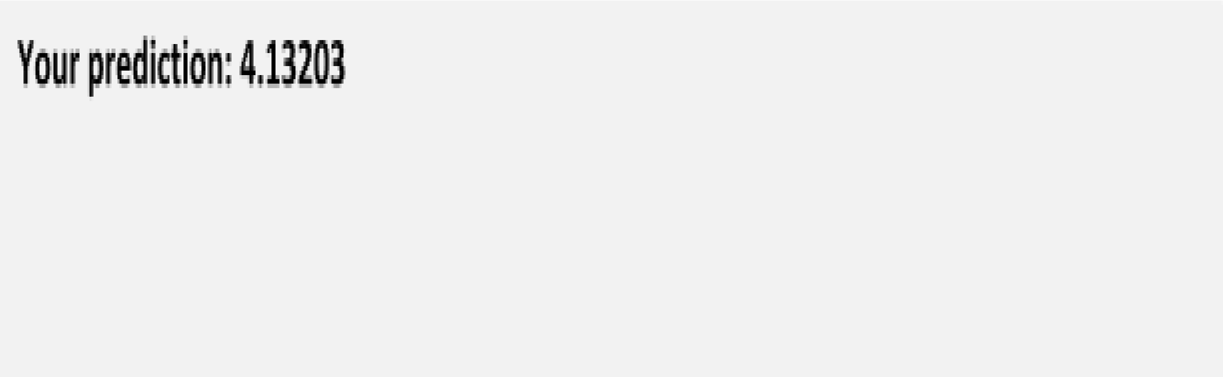


```
Call:
lm(formula = avg_star_rating ~ reviews_number + price_range +
    days_operating + photos)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.31879 -0.24137  0.08923  0.32247  1.19830 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.7400882  0.3365268  14.085   <2e-16
reviews_number -0.0001136  0.0006897  -0.165   0.8694
price_range    -0.1006417  0.0555933  -1.810   0.0717
days_operating -0.1083113  0.0470957  -2.300   0.0225
photos          0.0008226  0.0004286   1.919   0.0563
                  
(Intercept)    ***
reviews_number    
price_range    .  
days_operating *  
photos         .  
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5288 on 208 degrees of freedom
Multiple R-squared:  0.07784,	Adjusted R-squared:  0.06011 
F-statistic:  4.39 on 4 and 208 DF,  p-value: 0.001989
```

Which of the above coefficients are statistically significant at the 95% level?

Your answer: intercept and days_operating

Based on the above results (multiple regression), predict the number of stars a restaurant would have if it: (i) has 100 reviews, (ii) a price range=3, (iii) it is open for five days, and (iv) has 300 photos.



Your prediction: 4.13203

3. Categorical Variables (7 points)

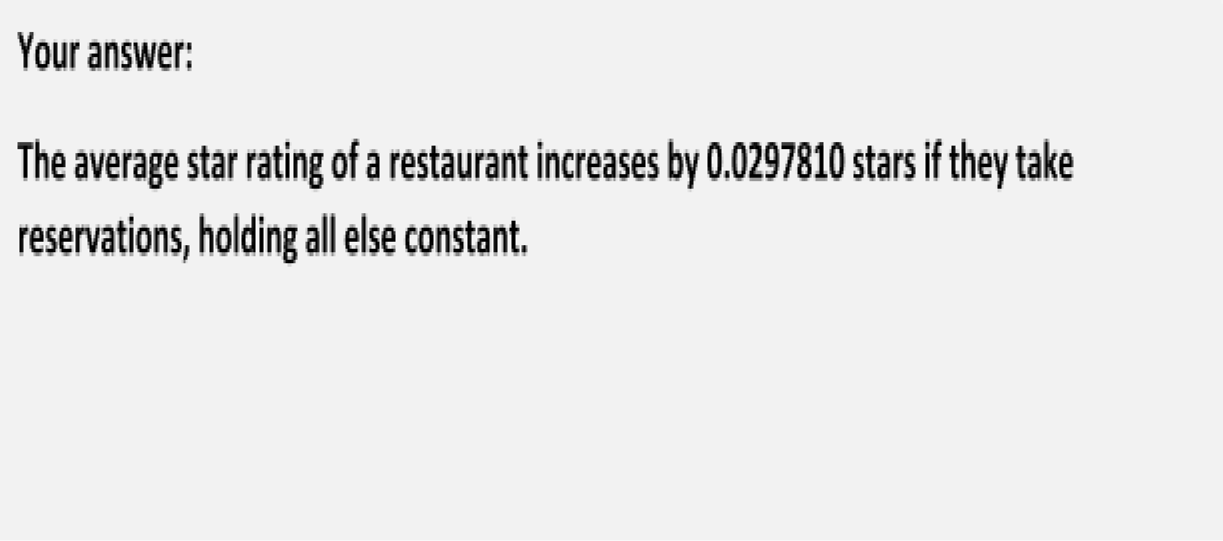
Run the following model

$$avg\_star\_rating = b_o + b_1(reviews\_number) + b_2(reservations)$$

Your results:

- b_o:3.8345791
- b_1:0.0008653
- b_2:0.0297810

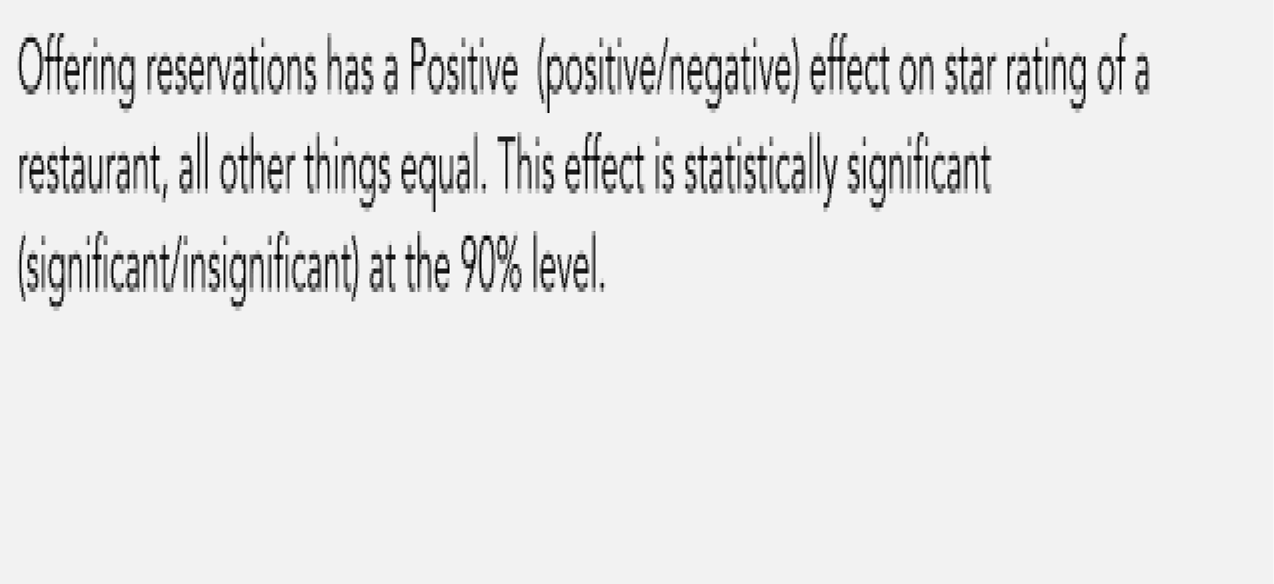
Now, interpret the coefficient b_2:re



Your answer:

The average star rating of a restaurant increases by 0.0297810 stars if they take reservations, holding all else constant.

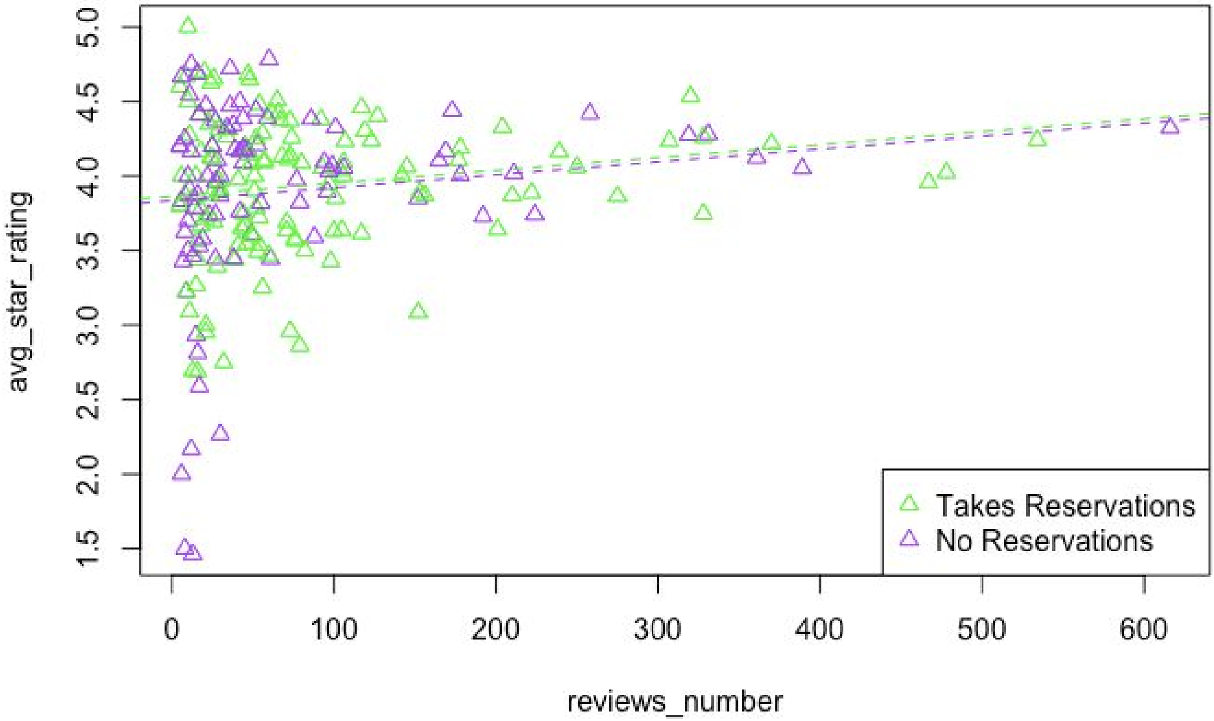
Fill in the blanks



Offering reservations has a Positive (positive/negative) effect on star rating of a restaurant, all other things equal. This effect is statistically significant (significant/insignificant) at the 90% level.

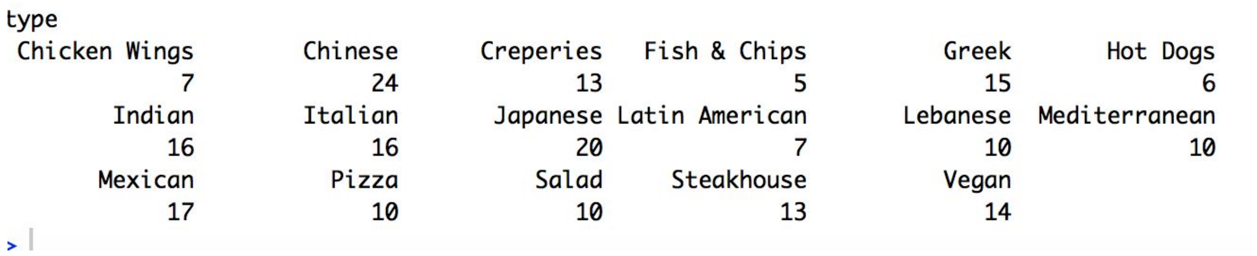
Draw the regression line (on scatter plot) for (i) restaurants that take reservations and (ii) restaurants that do not take reservations. Make sure you follow the instructions below:

- Restaurants that take reservations should be in green, restaurants that don’t should be purple.
- Make sure you create a legend.
- Instead of circles, I want triangles in the dots of the scatter plot (you will need to figure this on your own)
- I want the regression lines to be dashed lines (you will need to figure how to make them dashed on your own)



4. Categorical Variables with multiple categories (7 points)

a. How many restaurants of each type are there? Please paste the name of the categories and the number of observations per category below, using the *table()* function:



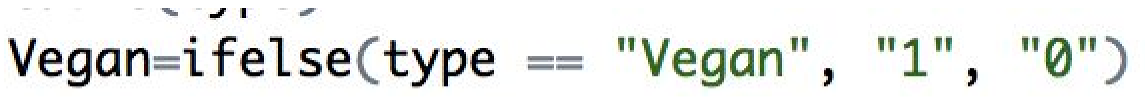
```
type
 Chicken Wings        Chinese      Creperies   Fish & Chips          Greek       Hot Dogs 
             7             24             13              5             15              6 
        Indian        Italian       Japanese Latin American       Lebanese  Mediterranean 
            16             16             20              7             10             10 
       Mexican          Pizza          Salad     Steakhouse          Vegan 
            17             10             10             13             14 
> 
```

b. Now, I want to determine if Vegan restaurants receive better ratings than non-Vegan. Run the following model

$$Avg\_stars\_rating = b_o + b_1(reviews\_number) + b_2(Vegan)$$

Where *Vegan*=1 if *type*= "Vegan", and 0 otherwise. You will need to figure out how to create this variable in *R* yourself. Hint: One way to do this is using the *ifelse()* command, but there are other ways.

Your command to create the *Vegan* variable:

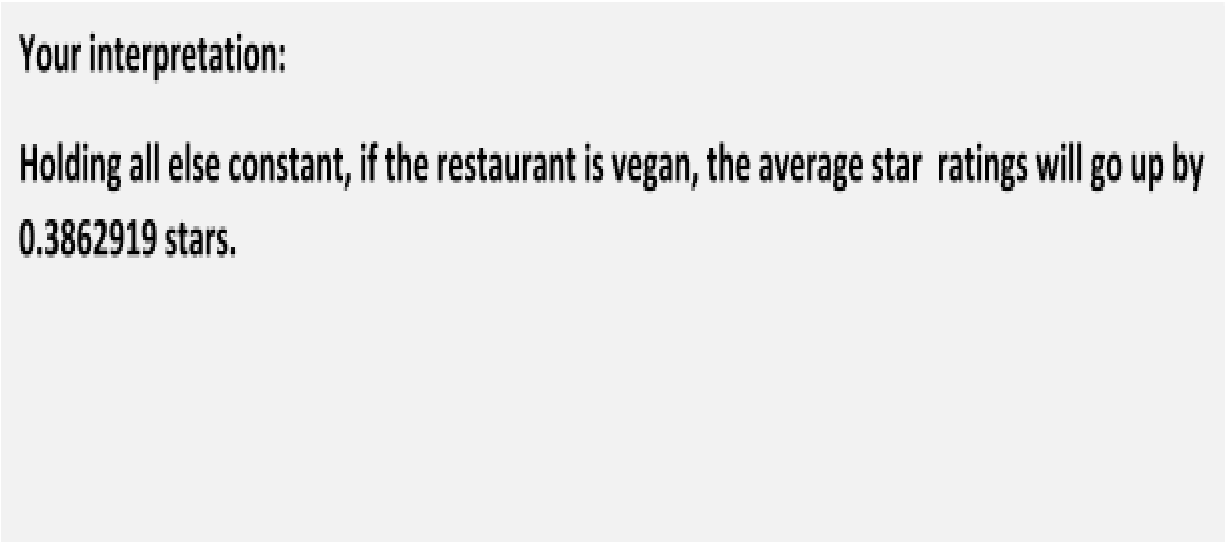


```
Vegan=ifelse(type == "Vegan", "1", "0")
```

Your results:

- b_o:3.8295076
- b_1:0.0008319
- b_2:0.3862919

Now, interpret the coefficient b_2:



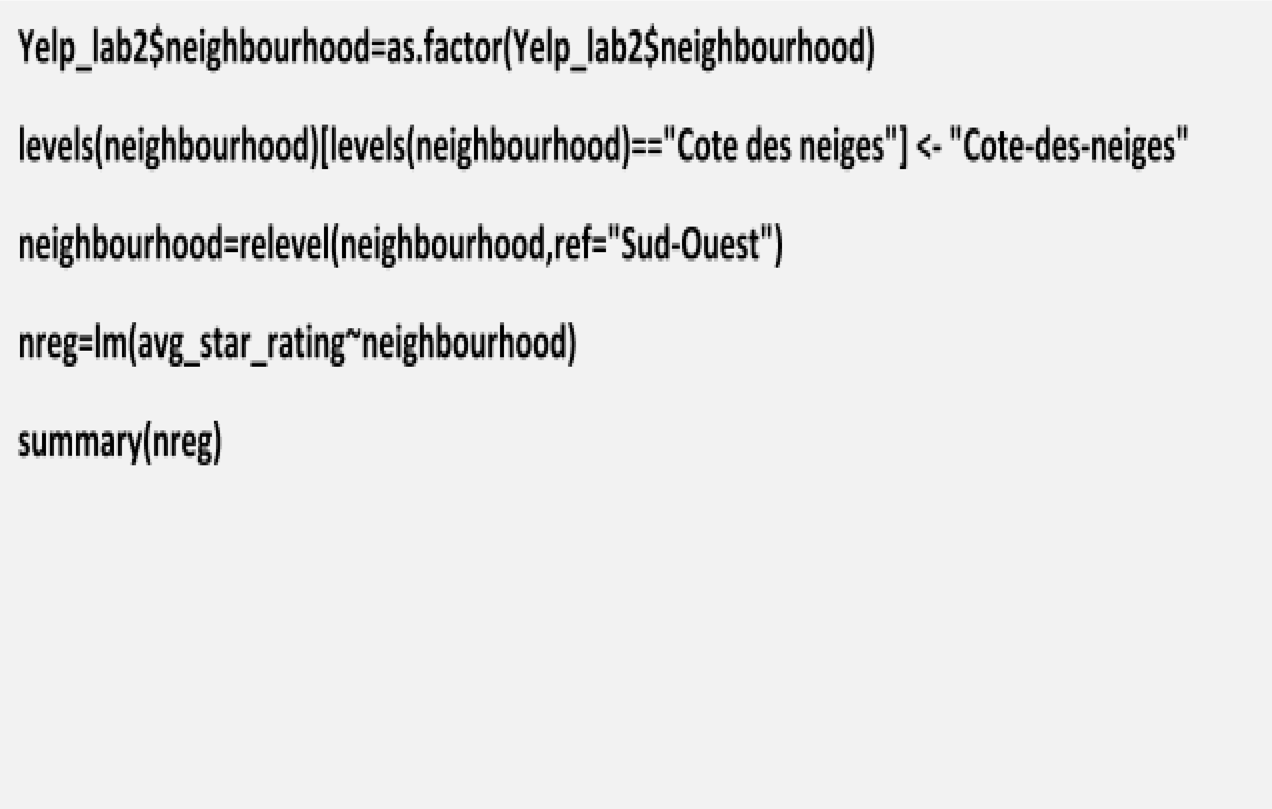
Your interpretation:

Holding all else constant, if the restaurant is vegan, the average star ratings will go up by 0.3862919 stars.

c. I want to know which neighbourhood (Plateau, Ville-Marie, Cote-de-neiges, etc) has restaurants with better reviews, on average. To test this, run a model where (i) the dependent variable is *avg_star_rating*, and (ii) the predictors are the categories found in *Neighbourhood* (no other predictors). Create a multiple linear regression, where the excluded dummy is *Sud-Ouest*.

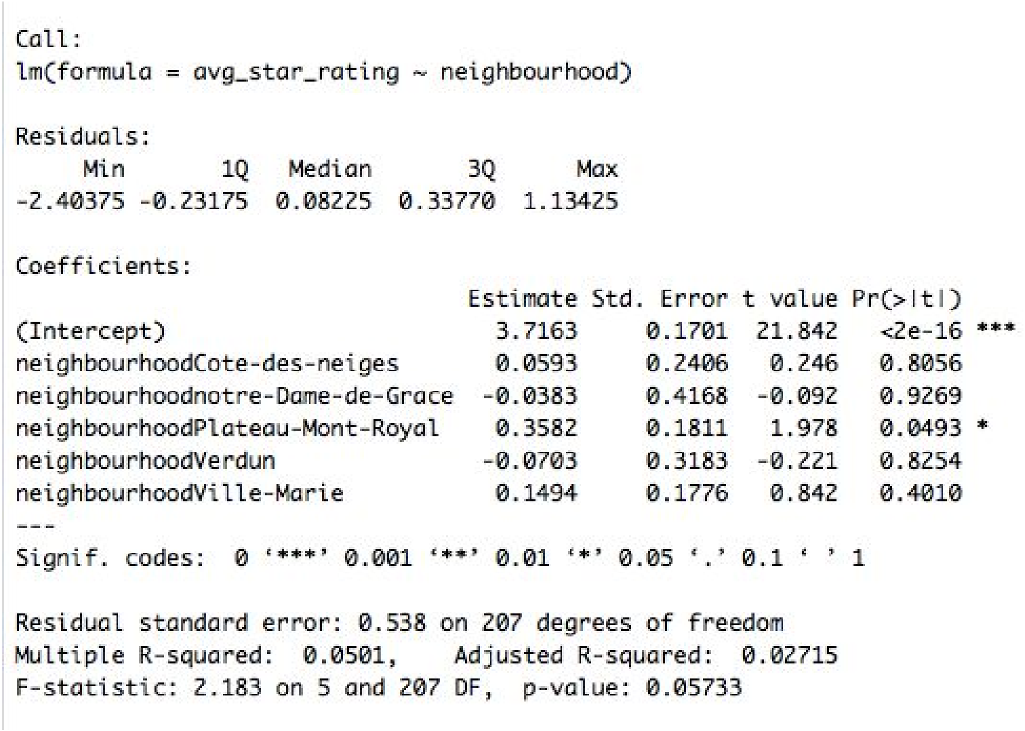
Hint: Check the data to see if all categories are properly named. You might have to rename some categories if you find inconsistencies.

Write the regression equation below:



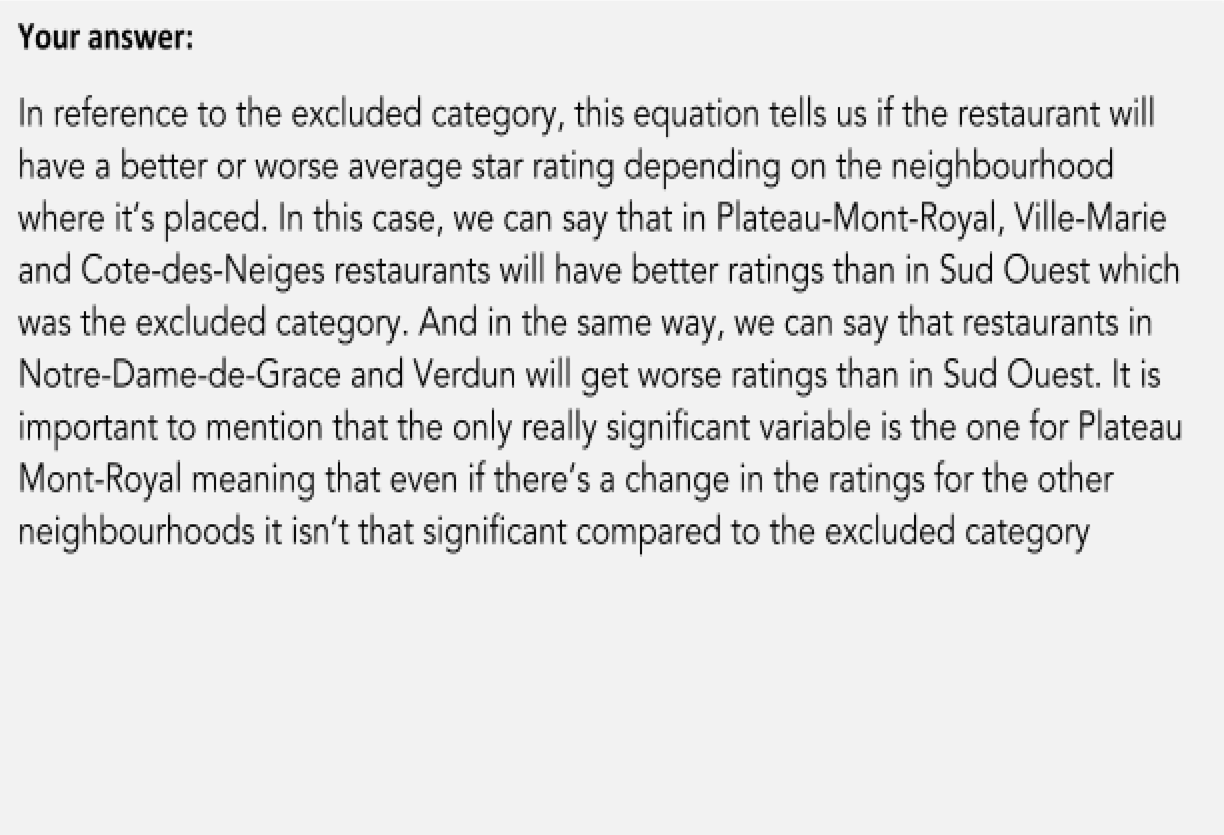
```
Yelp_lab2$neighbourhood=as.factor(Yelp_lab2$neighbourhood)

levels(neighbourhood)[levels(neighbourhood)=="Cote des neiges"] <- "Cote-des-neiges"

neighbourhood=relevel(neighbourhood,ref="Sud-Ouest")

nreg=lm(avg_star_rating~neighbourhood)

summary(nreg)
```

d. Paste the R-results of the regression below (as an image):



```
Call:
lm(formula = avg_star_rating ~ neighbourhood)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.40375 -0.23175  0.08225  0.33770  1.13425 

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)                         3.7163     0.1701  21.842   <2e-16 ***
neighbourhoodCote-des-neiges        0.0593     0.2406   0.246   0.8056    
neighbourhoodnotre-Dame-de-Grace   -0.0383     0.4168  -0.092   0.9269    
neighbourhoodPlateau-Mont-Royal     0.3582     0.1811   1.978   0.0493 *  
neighbourhoodVerdun                -0.0703     0.3183  -0.221   0.8254    
neighbourhoodVille-Marie            0.1494     0.1776   0.842   0.4010    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.538 on 207 degrees of freedom
Multiple R-squared:  0.0501,    Adjusted R-squared:  0.02715 
F-statistic: 2.183 on 5 and 207 DF,  p-value: 0.05733
```

e. In one paragraph, what does this equation tell you, in reference to the excluded category? Which restaurants locations are more popular? Which are less popular? Comment on statistical significance.



Your answer:

In reference to the excluded category, this equation tells us if the restaurant will have a better or worse average star rating depending on the neighbourhood where it's placed. In this case, we can say that in Plateau-Mont-Royal, Ville-Marie and Cote-des-Neiges restaurants will have better ratings than in Sud Ouest which was the excluded category. And in the same way, we can say that restaurants in Notre-Dame-de-Grace and Verdun will get worse ratings than in Sud Ouest. It is important to mention that the only really significant variable is the one for Plateau Mont-Royal meaning that even if there's a change in the ratings for the other neighbourhoods it isn't that significant compared to the excluded category

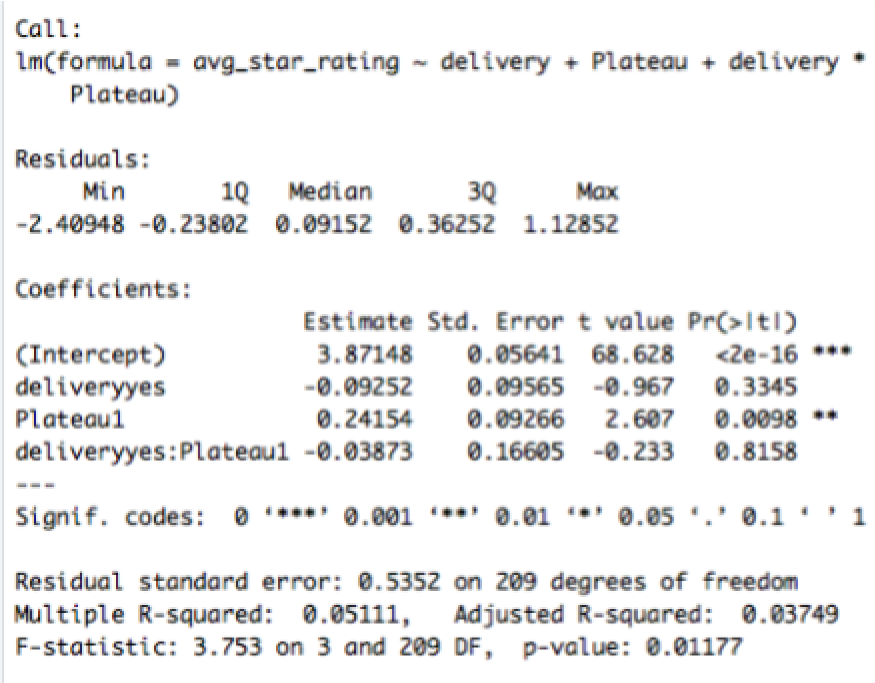
5. Interaction terms (5 points)

a. Run the following interaction model:

$$\textit{avg\_star\_rating} = b_o + b_1(\textit{delivery}) + b_2(\textit{Plateau}) + b3(\text{delivery}*\textit{Plateau})$$

Where *Plateau* Is a variable you will need to create (*Plateau*=1 if restaurant is in le Plateau neighbourhood, *Plateau*=0 if it is not).

Write the regression output below:



```
Call:
lm(formula = avg_star_rating ~ delivery + Plateau + delivery *
    Plateau)

Residuals:
     Min       1Q   Median       3Q      Max
-2.40948 -0.23802  0.09152  0.36252  1.12852

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            3.87148    0.05641  68.628   <2e-16 ***
deliveryyes           -0.09252    0.09565  -0.967   0.3345
Plateau1               0.24154    0.09266   2.607   0.0098 **
deliveryyes:Plateau1 -0.03873    0.16605  -0.233   0.8158
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5352 on 209 degrees of freedom
Multiple R-squared:  0.05111,   Adjusted R-squared:  0.03749
F-statistic: 3.753 on 3 and 209 DF,  p-value: 0.01177
```

b. In two sentences, what does the coefficient b_3 tell you?

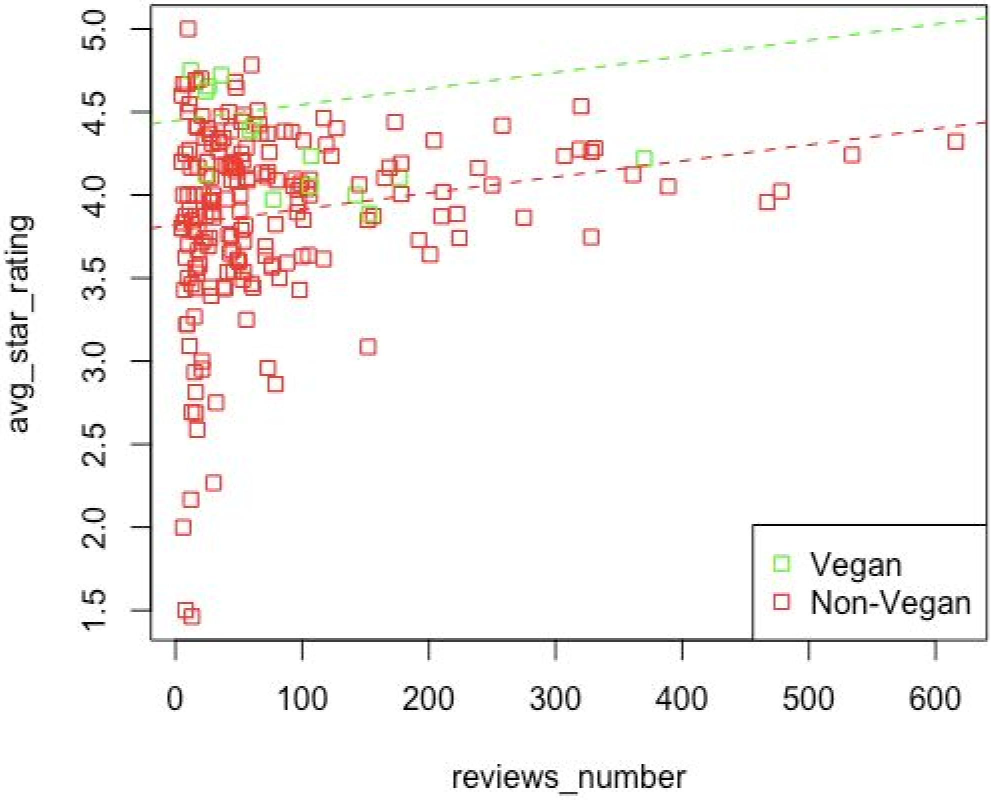
Your answer: If the restaurant can provide delivery and is in Plateau-Mont-Royal, it will bring less 0.03873 stars to the average star ratings. If there's no delivery provided or the restaurant is not in Plateau-Mont-Royal, there's no additional effect to the average star ratings.

c. Consider the following model:

Avg_star_rating =b_o+b_1(*review_number*)+ b_2(*Vegan*)+b_3(review_number*Vegan)

I want you to run the regression for the above model. Then draw the regression lines for (i) Vegan restaurants and (ii) non-Vegan restaurants.

- The regression line for Vegan restaurants should be in green, the line for non-Vegan restaurants should be red.
- Make sure you create a legend.
- Instead of circles, I want squares in the dots of the scatter plot (you will need to figure how to do this on your own).
- I want the regression to be dashed lines (you will need to figure how to do this on your own).



6. Interpretation of a full regression model

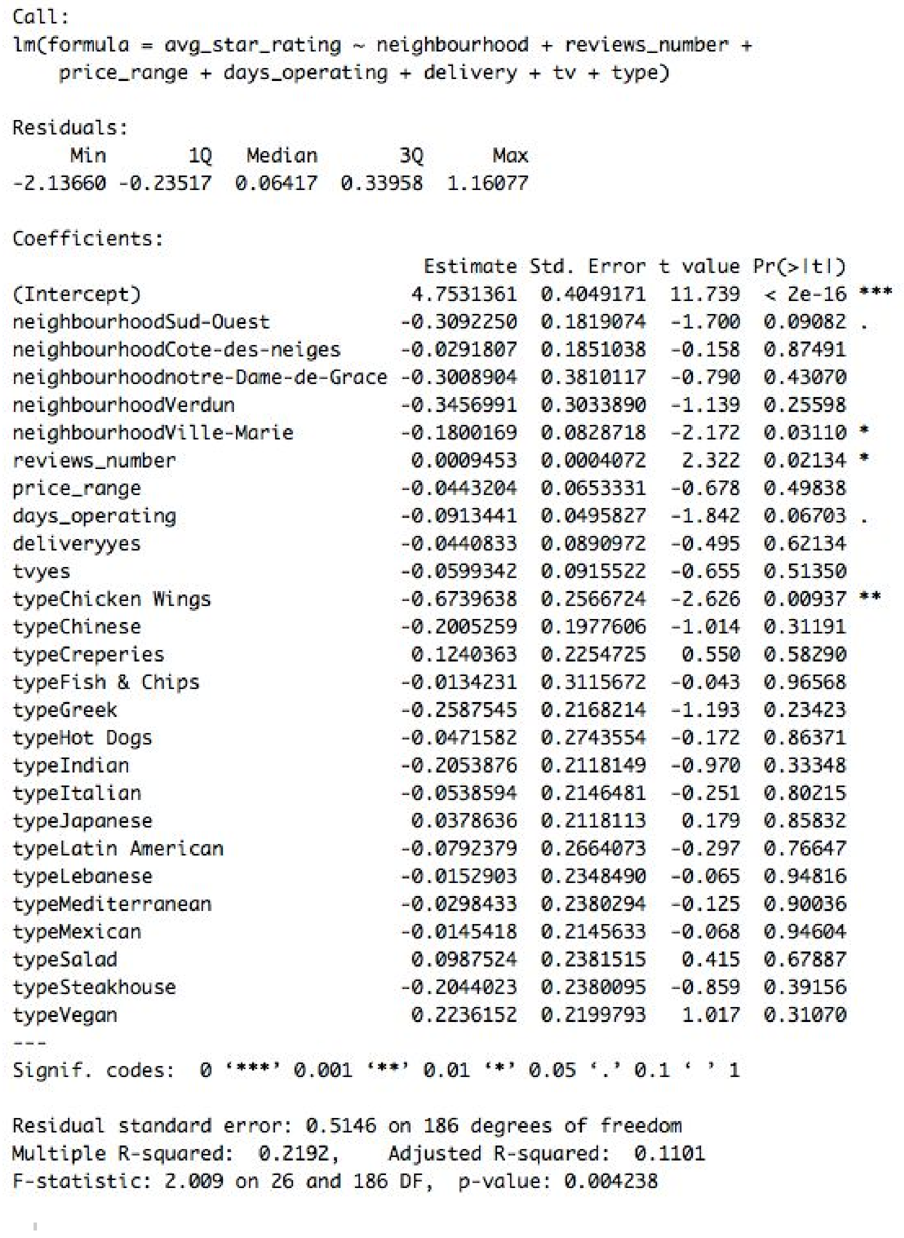
(5 points)

Now, run a multiple linear regression, where you include all the following predictors:

- □ Response Variable (Y): *avg_star_rating*
- □ Predictors (X): *neighbourhood*, *reviews_number, price_range, Days_operating, Delivery, TV, Type.*

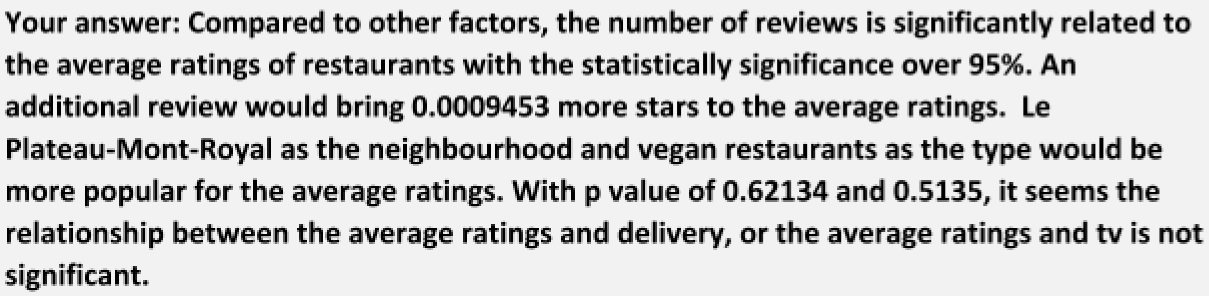
  - o Note: Don't forget that *type* and *neighbourhood* have multiple categories. Make *Pizza* and "Le Plateau" the excluded categories in each case. Please paste the results of this regression below:
  - o **Hint**: Check the data to see if all categories are properly named. You might have to rename some categories.

Regression Output:



```
Call:
lm(formula = avg_star_rating ~ neighbourhood + reviews_number +
    price_range + days_operating + delivery + tv + type)

Residuals:
     Min       1Q   Median       3Q      Max
-2.13660 -0.23517  0.06417  0.33958  1.16077

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                       4.7531361  0.4049171  11.739  < 2e-16 ***
neighbourhoodSud-Ouest           -0.3092250  0.1819074  -1.700  0.09082 .
neighbourhoodCote-des-neiges     -0.0291807  0.1851038  -0.158  0.87491
neighbourhoodnotre-Dame-de-Grace -0.3008904  0.3810117  -0.790  0.43070
neighbourhoodVerdun              -0.3456991  0.3033890  -1.139  0.25598
neighbourhoodVille-Marie         -0.1800169  0.0828718  -2.172  0.03110 *
reviews_number                    0.0009453  0.0004072   2.322  0.02134 *
price_range                      -0.0443204  0.0653331  -0.678  0.49838
days_operating                   -0.0913441  0.0495827  -1.842  0.06703 .
deliveryyes                      -0.0440833  0.0890972  -0.495  0.62134
tvyes                            -0.0599342  0.0915522  -0.655  0.51350
typeChicken Wings                -0.6739638  0.2566724  -2.626  0.00937 **
typeChinese                      -0.2005259  0.1977606  -1.014  0.31191
typeCreperies                     0.1240363  0.2254725   0.550  0.58290
typeFish & Chips                 -0.0134231  0.3115672  -0.043  0.96568
typeGreek                        -0.2587545  0.2168214  -1.193  0.23423
typeHot Dogs                     -0.0471582  0.2743554  -0.172  0.86371
typeIndian                       -0.2053876  0.2118149  -0.970  0.33348
typeItalian                      -0.0538594  0.2146481  -0.251  0.80215
typeJapanese                      0.0378636  0.2118113   0.179  0.85832
typeLatin American               -0.0792379  0.2664073  -0.297  0.76647
typeLebanese                     -0.0152903  0.2348490  -0.065  0.94816
typeMediterranean                -0.0298433  0.2380294  -0.125  0.90036
typeMexican                      -0.0145418  0.2145633  -0.068  0.94604
typeSalad                         0.0987524  0.2381515   0.415  0.67887
typeSteakhouse                   -0.2044023  0.2380095  -0.859  0.39156
typeVegan                         0.2236152  0.2199793   1.017  0.31070
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5146 on 186 degrees of freedom
Multiple R-squared:  0.2192,    Adjusted R-squared:  0.1101
F-statistic: 2.009 on 26 and 186 DF,  p-value: 0.004238
```

Based on this result, what have you learned about restaurant popularity in Montreal (type and neighborhood wise)? Please limit your answer to the space provided, and discuss about direction of predictors, and their statistical significance. I want you to draft a quick explanation to a manager who does not know statistics. Therefore, you will need to develop this explanation using no jargon, leveling it down to non-technical managers.

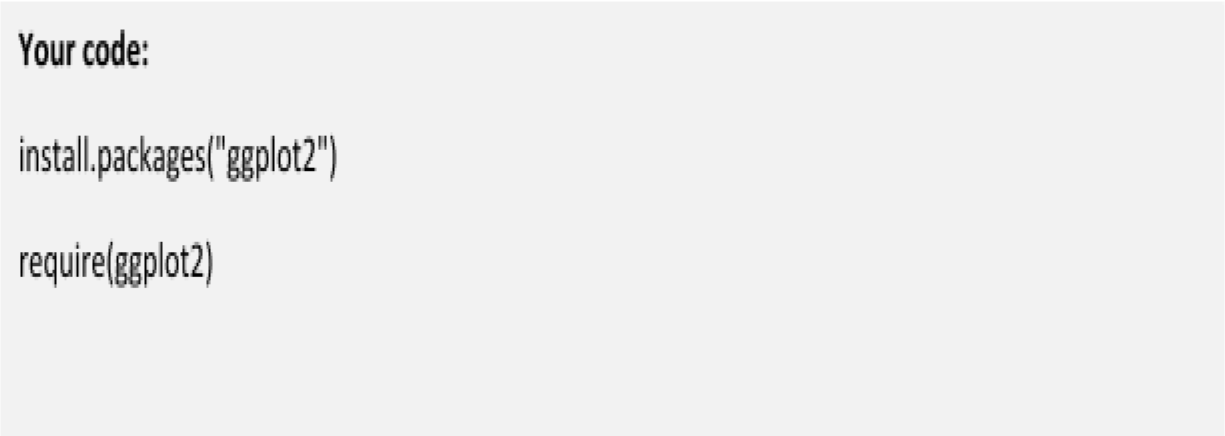


Your answer: Compared to other factors, the number of reviews is significantly related to the average ratings of restaurants with the statistically significance over 95%. An additional review would bring 0.0009453 more stars to the average ratings. Le Plateau-Mont-Royal as the neighbourhood and vegan restaurants as the type would be more popular for the average ratings. With p value of 0.62134 and 0.5135, it seems the relationship between the average ratings and delivery, or the average ratings and tv is not significant.

7. GgPlot2 (3 points)

Now, we will be learning a way of creating visually appealing and highly customizable graphs. Specifically, in *R,* there is a package called *ggplot2. The Economist* magazine (and other important outlets) use ggPlot2 to make their graphs. For the following variable, you need to create the same plots, but using ggplot2.

1. You will need to first install the ggplot2 package (we learned in class how to install a package). You also need to tell *R* that you will be using ggplot for this exercise. Paste the two code lines below required to perform this task:



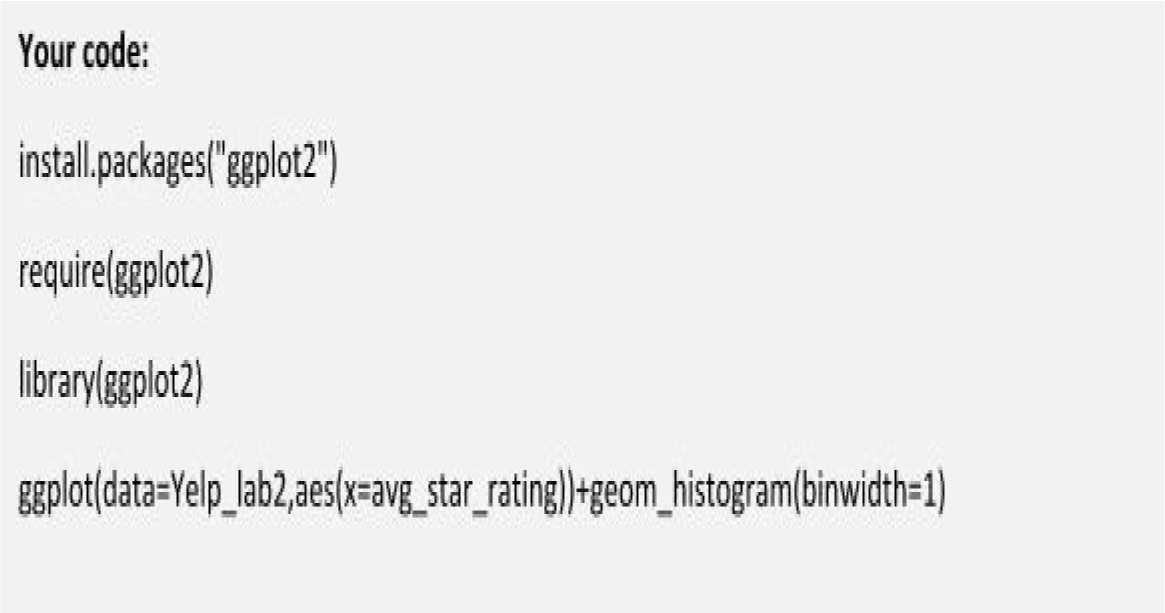
Your code:

```
install.packages("ggplot2")

require(ggplot2)
```

2. Now, create a histogram of the variable avg_star_rating using ggplot, with binwidth=100. The idea here is to get you to figure out how to code. *R* has tons of help in the internet (e.g. stackoverflow). You will need to research how to use this package.

For instance, here’s a tutorial on ggplot2:
https://tutorials.iq.harvard.edu/R/Rgraphics/Rgraphics.html

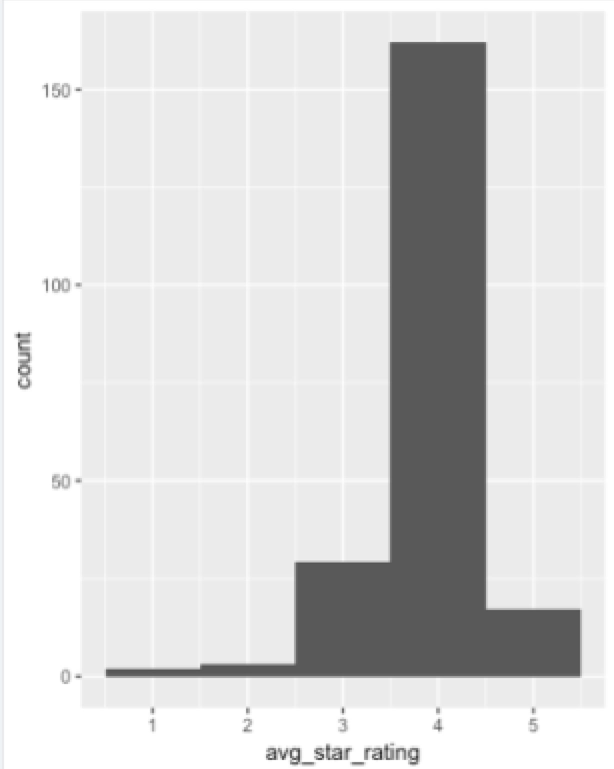
Hint: “aes” , “binwidth”, and “geom_histogram” should appear in your code.



Your code:

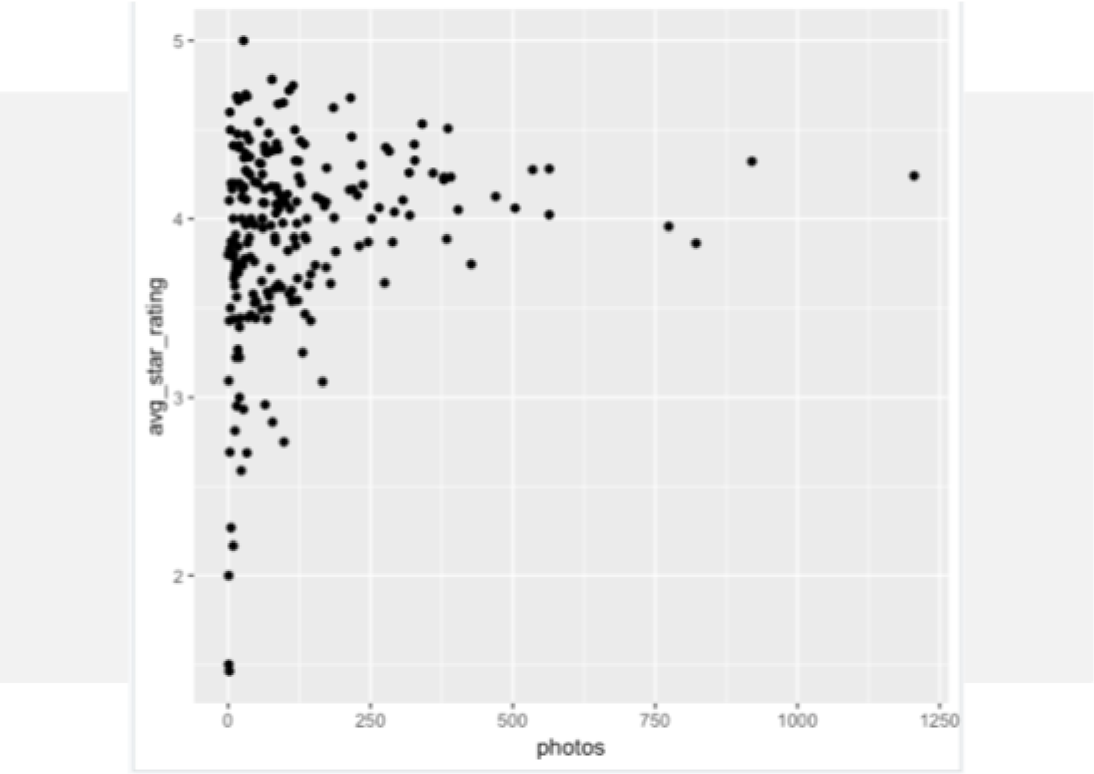
```
install.packages("ggplot2")

require(ggplot2)

library(ggplot2)

ggplot(data=Yelp_lab2,aes(x=avg_star_rating))+geom_histogram(binwidth=1)
```

In



3. Create a scatter plot, where y=avg_star_rating and x=photos. Hint: “geom_point” should be in your code:

Your code:ggplot(data=Yelp_lab2, aes(y=avg_star_rating, x=photos))+geom_point()



CODE FOR LAB 2

```
attach(Yelp_lab2)
sreg1=lm(avg_star_rating~reviews_number)
summary(sreg1)

sreg2=lm(avg_star_rating~price_range)
summary(sreg2)

sreg3=lm(avg_star_rating~days_operating)
summary(sreg3)

sreg4=lm(avg_star_rating~photos)
summary(sreg4)

pred1=coef(sreg1)[1]+100*coef(sreg1)[2]
```

```
pred1
pred2=coef(sreg2)[1]+3*coef(sreg2)[2]
pred2
pred3=coef(sreg3)[1]+5*coef(sreg3)[2]
pred3
pred4=coef(sreg4)[1]+300*coef(sreg4)[2]
pred4

pred5=coef(mreg1)[1]+100*coef(mreg1)[2]+3*coef(mreg1)[3]+5*coef(mreg1)[4]+300*coef(mreg1)[5]
pred5
```

3. categorical

```
Yelp_lab2$reservations=as.factor(Yelp_lab2$reservations)
creg=lm(avg_star_rating~reviews_number+reservations)
summary(creg)
```

5. Interaction

```
Yelp_lab2$neighbourhood=as.factor(Yelp_lab2$neighbourhood)
Yelp_lab2$delivery=as.factor(Yelp_lab2$delivery)
Plateau=ifelse(neighbourhood=="Plateau-Mont-Royal","1","0")
preg=lm(avg_star_rating~delivery+Plateau+delivery*Plateau)
summary(preg)

Yelp_lab2$type=as.factor(Yelp_lab2$type)
Vegan=ifelse(type=="Vegan","1","0")
vreg=lm(avg_star_rating~reviews_number+Vegan+reviews_number*Vegan)
plot(reviews_number,avg_star_rating,col=ifelse(type=="Vegan","Green","Red"),pch=0)
```

```
abline(coef(vreg)[1]+coef(vreg)[3],coef(vreg)[2],col="green",lty=2)

abline(coef(vreg)[1],coef(vreg)[2],col="red",lty=2)
```

6. interpretation of a full regression model

```
Yelp_lab2$neighbourhood=as.factor(Yelp_lab2$neighbourhood)

Yelp_lab2$delivery=as.factor(Yelp_lab2$delivery)

Yelp_lab2$tv=as.factor(Yelp_lab2$tv)

Yelp_lab2$type=as.factor(Yelp_lab2$type)

levels(neighbourhood)[levels(neighbourhood)=="Cote des neiges"] <- "Cote-des-neiges"

type=relevel(type, ref="Pizza")

neighbourhood=relevel(neighbourhood, ref="Plateau-Mont-Royal")

final_reg=lm(avg_star_rating~neighbourhood+reviews_number+price_range+days_operating+delivery+tv+type)

summary(final_reg)
```