Lab 6

BSS Stat 20

2022-08-01

NOTICE

If you have any collaborators, please write a sentence before Question 1 which acknowledges them. In addition, make sure that the sentence they therefore are also required to write acknowledges you.

For a template you can follow, see the course syllabus. This notice will be pasted at the beginning of every lab assignment.

All code MUST be shown. No credit will be given for correct answers without supporting code.

Questions

The data-set this week compiles basic information on Berkeley restaurants in the neighborhoods of North Berkeley, Downtown Berkeley and Southside!

Examining the Data

Download the `Restaurants` data set from the course forum thread titled *Lab 6 Materials.*

Question 1

What is the unit of observation in this data set?

Question 2

What are the dimensions of this data set?

Question 3

Based on the variables present in this data set, list a question about Berkeley restaurants that can be answered.

Question 4

Conversely, list a question about Berkeley restaurants that *cannot* be answered with this data set.

Data Analysis - Linear Model EDA

Question 5

Plot the distribution of number of reviews and describe the relationship.

Comment on whether we should consider transforming this data if we use it as an outcome variable in a linear regression model based on our results.

Question 6

Plot the relationship between number of *log* number of photos taken and *log* number of reviews. Describe the relationship in terms of form, strength, and direction.

Comment also on the presence (if any) of outliers.

Question 7

Compute the correlation coefficient r between log number of photos posted and log number of reviews posted.

Question 8

Based on the results of **Question 6** and **Question 7**, does a linear model seem appropriate in this setting?

Data Analysis - Linear Regression

Question 9

Fit a simple linear model to predict log number of reviews by log number of photos taken and save your fit into an object.

Question 10

Write out the equation for the simple linear model and interpret the slope you got.

Question 11

Report the R^2 value from your simple linear model and interpret it in the context of the problem.

Question 12

Superimpose your linear model from **Question 10** onto the plot you made in **Question 7**. Based on this plot and your R^2 value from **Question 12**, does a linear model seem appropriate to you in this setting?

Question 13

Bobby G's Pizzeria has 209 photos posted and 508 reviews posted to Google. Calculate the residual number of reviews (not on the log scale) for *Bobby G's Pizzeria* based on your model.

Question 14

Now, consider a Berkeley restaurant with 350 photos posted on Google Reviews. How many reviews (not on the log scale) does your model predict will be written for that restaurant on Google Reviews?

Question 15

Now, fit a multiple regression model to predict the number of reviews written about a restaurant with number of photos taken and at least one other variable in the data set. Report the adjusted R^2 (R^2_{adj}) and comment on how the R^2 value changes from the previous model as well? Do you think the new model you created predicts number of reviews better?

Data Analysis - Classification

Now, we are going to see if we can use k-Nearest Neighbors to determine whether a given restaurant lies within the North Berkeley or Southside neighborhood.

Filter the original `Restaurants` data-set to include only observations from the North Berkeley or Southside neighborhood.

Question 16

Create a training set of 37 observations and a testing set of 10 observations. *Ensure the training-testing split you make is reproducible.*

Question 17

Using your training set, fit a k-Nearest Neighbors model to predict `Neighborhood` on your 10 testing observations using the two variables `Price_Scale` and `Avg_Rating`. Let $k = 5$. Print out your predictions.

Question 18

Find the misclassification rate (MCR) of your k-Nearest Neighbors model.

Question 19

Identify one (if any) restaurant whose neighborhood was misclassified by your k-Nearest Neighbors model and give reasons as to why the model might have failed to locate the restaurant correctly. If you have no such restaurants, describe the value combinations of `Price_Scale` and `Avg_Rating` that are characteristic to restaurants in North Berkeley.