CS39000-DM0 Homework 2

Due date: Wednesday October 12, 11:59pm in Blackboard.

Submit a PDF with both your answers to the questions and the R code that you used for analysis. Your homework must be typed.

In this assignment, you will use the R statistical package to explore, transform, and analyze data. Based on your analysis you will formulate hypotheses about the data. To get started, do the following:

- Download and install R from:
  `http://cran.r-project.org/`
  Links to a quick introduction to the R programming language and a short reference card for R are below. A longer tutorial for R is included in the class readings.
  `http://www.stat.cmu.edu/~larry/all-of-statistics/=R/Rintro.pdf`
  `http://cran.r-project.org/doc/contrib/Short-refcard.pdf`
- Download the Yelp dataset from the course page.
  This data set is part of the Yelp academic dataset and consists of data about 45,699 restaurants. The datafile *yelp.csv* contains 28 attributes: 6 numeric and 22 discrete. The first row of the data file is a header row with the names of the attributes where names are separated by a comma (,).

Use R to analyze the Yelp data and complete the questions below.

1 Data import and summarization

Read the data into R using `read.table()` function. Use the argument `sep=","` to specify the column delimiter, the argument `header=TRUE` to read in the column names, the argument `quote="\""` to read in the quoted fields, and the argument `comment.char=""` to treat the # characters as text rather than comments.

(a) Print a summary of the data using the `summary()` function.

2 1D plots

(a) - Plot a histogram of the *'checkins'* attribute. Use the `hist()` function with its default values and make sure to title the plot with the name of the attribute for clarity.
- Compute the logged values for *'checkins'* (you can use `log(d$column_name)` to compute the log of all the values in a column in place). Plot a histogram of the logged values.
- Discuss the differences between the two plots and the information they convey about the distribution of *'checkins'* values in the data.

(b) - Plot a density plot of the logged values of the *'checkins'* attribute using the `density()` function.

- Discuss any similarities and differences between the density plot and the histogram from Q2a with the same logged values (e.g., the information they convey about the distribution of the attribute).

(c) - Plot the logged values of the *'checkins'* attribute using `hist()` with `breaks=50`.
- Plot the logged values of the *'checkins'* attribute using `density()` with `adjust=0.5`.
- Discuss any differences from your previous plots and how the parameter settings change the way the distributions look.
- Plot `hist()` with `breaks=50`, `freq=FALSE` and compare with the density plot from this part.

(d) - Plot a barplot of the *'state'* attribute to show the frequency of each value. Use the `table()` function to get the counts for each value, then use the `barplot()` function with the `names` argument to label the bars with the appropriate value. Again, make sure to title the plot with the name of the attribute for clarity. Note that this will look like a histogram but for nominal values. In small renderings of this plot, you might not see all the state name labels, but if you stretch the window you will be able to see all the labels.

3 Sampling and transforming data

(a) - Transform the attributes: *'alcohol'* and *'noiseLevel'* into ordered, numeric features using the function `factor()` with a set of ordered levels as input parameter. For *'alcohol'* use `levels=c("full_bar","beer_and_wine","none","")`; for *'noiseLevel'* use `levels=c("quiet","average","loud","very_loud","")`. Then transform the resulting levels into integers using `as.integer()`.
- Append the two new columns to the original data frame, using `cbind()` to increase the number of features to 30.

(b) - The attribute *'categories'* and *'mealTimes'* contain comma separated lists of values associated with each restaurant. Compute two new boolean features: 'isAmerican' and 'goodForDinner' with a value of `TRUE` if the list contains *American* (in 'categories'), *dinner* (in 'mealTimes') respectively and `FALSE` otherwise. You can use the function `grepl("str", f$column_name)` to check whether the values in *column_name* contain the string "str".
- Append the two new columns to the original data frame, using `cbind()` to increase the number of features to 32.

(c) - Compute the quantiles (using the function `quantile()`) for the *'reviewCount'* attribute.
- Select a subset of the data with *'reviewCount'* value $\leq$ the 1st quartile (25th percentile). You can use use the `subset()` function or select from the data frame with `[ ]` operations.
- Print a summary of the above subset and compare the results to those from Q1a. Discuss any differences in the distributions of the numerical attributes that you find.

4 2D plots and correlations

(a) - Plot a scatterplot matrix (using `plot()`) for the five attributes:
'stars', *'reviewCount'*, *'checkins'*, *'longitude'*, *'latitude'*.
- Identify which pair of attributes exhibit the most association (as you can determine visually) and discuss whether this is interesting or expected, given your domain knowledge.

(b) - Calculate the pairwise correlation among these five attributes using the `cor()` function.
- Identify the pair of attributes with largest positive correlation and the pair with largest negative correlation. Report the correlations and discuss how it matches with your visual assessment in Q4a.

(c) - Plot a boxplot (using `boxplot()`) for each of the following four attributes vs. the *'goodForGroups'* attribute: *'stars'*, *'reviewCount'*, *'checkins'*, *'priceRange'*.
Make sure to label both axes of the plot with the appropriate attribute names.
- Identify which pair of attributes exhibit the most association (as you can determine visually) and discuss whether this is interesting or expected, given your domain knowledge.
- For the pair of attributes identified, calculate the interquartile range for the continuous attribute for each value of *'goodForGroups'* (i.e., a separate IQR for the "TRUE" instances and the "FALSE" instances). You can do this with the `subset()` and `quantile()` functions. Calculate the overlap between the two IQRs. Discuss whether these results support the conclusion you made based on visual inspection.

5 Identifying potential hypotheses

During your exploration above, investigate other relationships in the data. Establish relationships between variables by assessing plots, computing correlation, or other numerical analysis.

Identify TWO possible relationships in the data (other than the ones specified in earlier questions) and formulate hypotheses about the observed relationship. For each of the two identified relationships:

(a) Include a plot illustrating the observed relationship (between at least two variables).

(b) State whether the variables are discrete or continuous and what type of plot is relevant for comparing these two types of variables.

(c) Formulate a hypothesis about the observed relationship as a function of two random variables (e.g., X is associated with Y).

(d) Write the hypothesis as a claim in English, relating it to the attributes in the data.

(e) Identify the type of hypothesis.