HW2 – Statistics

This is an individual assignment. It consists of 4 tasks on statistics. All of them are based on the Yelp datasets. See the problem statements for the detailed dataset to be used. You do not need to clean the data for this homework. Use functions from existing libraries as much as possible. Don't reinvent the wheels! For example, use:

- **pandas** to manipulate and recognize data (https://pandas.pydata.org/docs/)
- **scipy** to calculate certain statistics (https://docs.scipy.org/doc/scipy/)
- **matplotlib** for plotting (https://matplotlib.org/stable/users/index.html)

Dataset:

The dataset we are using is the Yelp Dataset (https://www.yelp.com/dataset). We will focus on the user data and the business data. They are in JSON format. Due to the large volume, a subset of the data is used for this assignment. The data file can be download from the Canvas assignment page.
Please read the documentation (https://www.yelp.com/dataset/documentation/main) to understand the meaning of the attributes.

Q0. Format (10pt)

Please read the instructions carefully and follow the name and format on your submission files. In particular, please use existing libraries as much as possible.

Q1. Correlation coefficient. (30pt)

Dataset: yelp_academic_dataset_user_no_friend.json. Download the provided dataset from the assignment page.

Task: We would like to study the correlation of different user attributes.

- Q1a: Among "funny", "cool", "useful", "fans", and "review_count", which two of them have the strongest linearly correlation?
- Q1b: Investigate the correlation coefficients between the average star of all reviewers and the number of years since this user joined Yelp. What's your conclusion?
- Q1c: Compare the Pearson correlation coefficient and Spearman correlation coefficient between the number of reviews that a user gives and the number of years a user was an elite. What did you observe? Why do you think this is the case? Use some visualization to explain what you observe from the coefficient calculation.

Hint: scipy.stats (https://docs.scipy.org/doc/scipy/reference/stats.html) contains functions to calculate coefficients directly. Use dataframe.apply to extract indirect attributes.

Q2. Chi-Square test. (10 pt)

Dataset: yelp_academic_dataset_business.json. Download the provided dataset from the assignment page.

Task: Use the chi-square test to test if the following two events are independent:

-If the business is open or closed

-If the *stars* are greater than 2.5 or not

Please provide your observations if the two events are independent or not (assume a significance level of 0.05).

Q3. Association Analysis. (20pt)

Dataset: yelp_academic_dataset_business.json. Every business comes with a list of categories. For example, the business with id "`0bPLkL0QhhPO5kt1_EXmNQ`" has five categories (separated by commas): Food, Delis, Italian, Bakeries, Restaurants.

Task: Perform association rule analysis to find the common association between keywords in business categories. For example, one common rule is 'Fitness & Instruction' -> 'Active Life'.

- Q3a: Among all rules with a support of at least 0.05 and confidence of at least 90%, which rule has the maximum lift? Write down the rule as well as its (support, confidence, lift) values.
- Q3b: Among all rules with support at least 0.01, confidence at least 0.9, which of them are relevant to 'Auto Repair'? Among all of them, which one(s) has the highest lift? Explain why this association happens in practice.

Hint: 1) when processing the category strings, make sure you remove spaces before/after each category; 2) most data structures can be cast to a list.

Q4. ANOVA (30 pt)

Dataset: yelp_academic_dataset_business.json

Task: We want to study if the average star rating of a restaurant differs among different cuisine types. We will focus on Chinese, American, Mexican, and Italian.

- Q4a: Make a side-by-side boxplot to compare them qualitatively using matplotlib.
- Q4b: Formulate a hypothesis that you can use ANOVA to test and perform an ANOVA analysis. What's your conclusion from ANOVA test?

- Q4c: Repeat the above two steps, but this time on review count. What's your conclusion from the plot and from ANOVA test? Do you think ANOVA is applicable to test review count? Why?

Hint: 1) Remember that you need to check two ANOVA applicability conditions; 2) the stand ANOVA works when the sample sizes from different groups are the same. You can subsample each group to 1000-entry sets.

Submission Instructions

Put your code in the template notebook as provided on the Canvas assignment page. Please submit two files:

1) The ipynb file. Make sure every question is executed with the printed results. Make sure all explanations are printed.

2) Export your code and results in a separate html document. After you are done with the ipynb file, go to File --> Download/Export As --> HTML in Jupyter notebook.

Submit **two documents** using the following naming convention. Please submit them directly **without zipping them.**

```
HW2.ipynb
HW2.html
```