Data Analytics Project

Modugula, Jagadeep; McGilvry-James, Tyler J; Okafor, Enyinaya S; Marikanti, Santhosh; Merugu, Naga Prudhvi S

Requirements to run the code

- Programs required
    - Java
    - Python
- Packages required
    - Pyspark
    - Pandas
    - Numpy
    - Pyarrow
    - Matplotlib

The packages can be installed with pip and added to the right environment variables using the PATH. The code can be run on the python terminal on a Windows machine.

- Python “name_of_file.py”

The code should be ran in an IDE capable of visual (plot) outputs and the cleaned.xlsx file should be within the same directory as the python script. It is also recommended you have all the resources in the same folder for easy loading.

Dataset

The dataset is a compilation of restaurants from a review platform. The dataset has several columns such as the name, location, and the ratings. The original dataset had all columns labelled with the schema of strings. After applying the infer schema option in pyspark, we were still left with several mismatches. So, we had to enforce schema using several methods such as cast and struct types in the background. After the preprocessing, the dataset includes datatypes of Boolean, strings and integers and floats.

While working with the original dataset, we discovered that pandas loaded heavier strings better than pyspark. So, we did some cleaning with pandas before loading into pyspark, this is evident in the importation of pandas in the pyspark code.

Columns: name, online_order, book_table, rate, votes, location, rest_type, cost_for_two_people(cost).

The online_order defines restaurants that have an option for customers to order online while the book_table defines restaurants with the option to book tables online. The rest_type defines the type of the restaurant.

Analysis

The restaurants are rated by customers on an online platform. We began the analysis by describing the basic statistics that could be applied on the dataset. The analysis continued by drawing random samples to compare costs per location. We looked at other manipulations to the location and how prices are affected depending on the restaurant type. We further manipulated the dataset to discover restaurants based on voting numbers. Through analysis, we could observe the type of restaurants preferred in different locations.

There is usually a stereotype that expensive restaurants are better, so we analyzed the correlation between price and ratings using Pearson coefficient method on pandas and pyspark. There was a positive correlation but not high enough to conclude on that hypothesis. Also, to see the trend between restaurants that offer online order capabilities and restaurants that offer reservations, we used covariance to measure the variability of the two variables. The cov function was required to be imported in pyspark while importation wasn't required in pandas. The analysis also included a classification of restaurants according to their locations and count, we used an inner join to specify the exact columns required for the analysis.

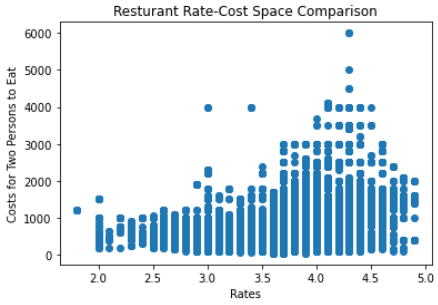
For better analysis, we utilized matplotlib to visualize some of our results. We analyzed the frequency distribution of costs by restaurants. The data is converted using pandas to the appropriate type of dataframe. Queries are submitted to form frequency-domain histograms about resturant ratings and costs. These queries apply two filters in sequential order over each bin determined by the numBin variable. A representation of the cost-rating space is also given from the same accumulated dataset. To properly run this portion of the code, matplotlib with associated dependencies must be installed in conjunction with pandas and pyspark. It is uncertain currently why there is a discrepancy between the histogram of restaurant costs, but it is thought to be due to a mismatch between the excel and csv formats when converting for the pyspark implementation. The final runtime is output at the end of each component of the script.

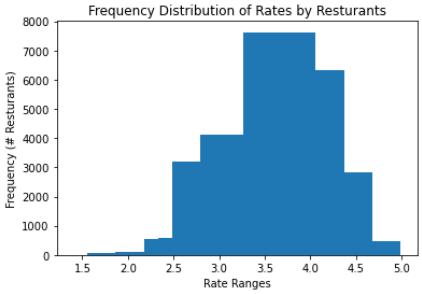
Results

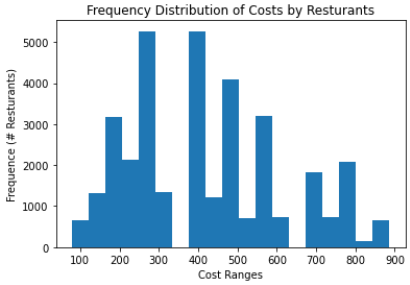
The frequency distribution of rates and frequencies differ, as represented by the somewhat skewed normal distribution in the rate-cost representation. Many restaurants are of above average quality, but most of any quality are at average prices. Only a few outliers at 4.5-5.0 skew the set. It is uncertain why there is a discrepancy between the histogram of restaurant costs, but it is thought to be due to a mismatch between the excel and csv formats when converting for the pyspark implementation.

The final runtime is output at the end of each component of the script.

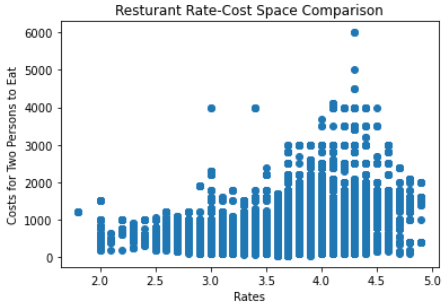
Pyspark Results

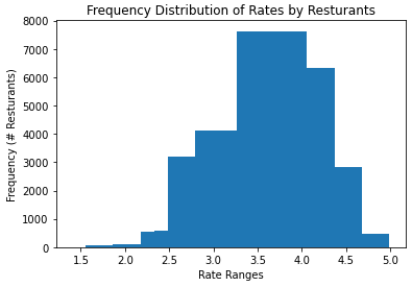


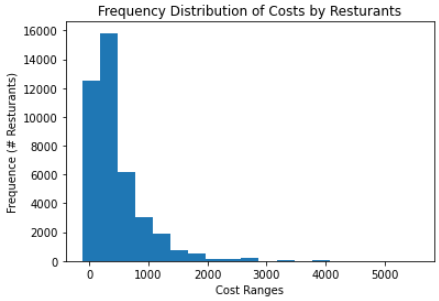




Pandas Result


Resturant Rate-Cost Space Comparison
Costs for Two Persons to Eat
Rates


Frequency Distribution of Rates by Resturants
Frequency (# Resturants)
Rate Ranges


Frequency Distribution of Costs by Resturants
Frequence (# Resturants)
Cost Ranges