COLUMBIA DATA ANALYTICS BOOTCAMP

RESTAURANTS: A TALE OF FOUR CITIES

Amanda Luna, Caiti Donovan, Duvan Diaz, Eline Van Eldere

TODAY'S OUTLINE

OUR MAIN POINTS

1. Selecting a dataset
2. Methodology & Analysis
3. Visualizations
4. Q&A

DATASET

A BRIEF INTRODUCTION

yelp Dataset

VERY LARGE YELP ACADEMIC DATASET WITH DATA ON "10 METROPOLITAN AREAS"

LIMITATIONS OF THE DATASET

After initial scrubbing we realized our preferred metropolitan area, New York City, is badly represented in the dataset

- Just 1 record on New York City
- Other lage metropolitan areas scarcely represented
- We chose 4 larger cities to evaluate: Las Vegas, Phoenix, Charlotte and Toronto (CA)

DATA SCRUBBING

CLEANING APPROACH:

- Downloaded the JSON files of the dataset
- Used Pandas in Jupyter Notebook, extracted JSONs into dataframes
- Reduce data to only include cuisine types
- Reduce data population to fit our choice of 4 cities
- Many actions to correct a LOT of spelling errors
- Slimmed down to top 40 overall for our cities

NONSQL APPROACH TO DATABASE CONNECTION

SET UP MONGO CONNECTION:

- Created final csv files to be connected to Mongo
- Our Yelp_FinalScrub.csv uploaded to use in Mongo Compass
- Created app.py file for flask API connection
- Set up 3 collections for city cuisine, city count and city facts which correspond with our asset files and eventual visualization setup

FRONT END VISUALIZATIONS: TOWARDS AN INTERACTIVE DASHBOARD AND MAP

DASHBOARD AND MAP:

- We chose an interactive dashboard as our visual canvas, as we felt our goal would be best reflected in that way
- Furthermore we added a clickable map to give the user an opportunity to geographically track quality restaurants by rating (1 up to 5 stars)
    - colored indicators
    - hoverable map

AAAAAAANY QUESTIONS?

BIBLIOGRAPHY

Dataset sourced from Yelp:

https://www.yelp.com/dataset/download

Github Repository:

https://github.com/aluna0806/Group_Project