Tasks one so far:

* Converted Yelp json file into a csv file using spark; (pyspark\_data\_load.py)
* Generate logstash file
* Generated final\_dataset.csv

data\_cleaned .py

* Extract categories related to “restaurants”; & added them as attributes.
* Removed those attributes/categories from final\_dataset.csv which contains null values greater than 0.95%.

Those are : 'AcceptsInsurance', 'AgesAllowed', 'BYOB', 'BYOBCorkage', 'ByAppointmentOnly', 'CoatCheck', 'Corkage', 'DietaryRestrictions', 'HairSpecializesIn', 'Open24Hours', 'RestaurantsCounterService', 'Smoking', 'friday', 'monday', 'saturday', 'sunday', 'thursday', 'tuesday', 'wednesday', 'no\_music'

* one -hot encoding to split each category into attributes.
* Created csv file containing restaurant related words (restaurant\_related\_words.csv)
* Retain the categories that are common to df\_categories\_sum.columns and restaurant\_valid\_words
* Delete df\_categories\_sum.
* Delete unnamed columns too if any,
* Generate a csv. **Yelp\_cleaned.csv**
* **From yelp\_cleaned, removed features having >60% null values,**
* **It removed happy hour; an important feature**
* **Hence we filled happy hour values based on : if alcohol = “full bar”, happy hour=true; else false.**
* **Then we removed >60% null values features.**
* **Now, we are checking which city contains more rows, i.e. top 3 are toronto, las vegas, phoenix.**
* **In US, the city containing more number of rows is vegas.**
* **We extracted a csv file containing only vegas and are working on it.**
* **Filename: yelp\_vegas.csv**

**To do (yelp\_cleaned)**

* **Histogram for city vs review\_count**
* **Histogram for city vs stars**
* **gmaps based clustering based on lat-lon**
* **Develop word cloud for restaurant names, pick popular ones.**
* **Develop a relative graph for top 5-6 popular names vs their rating. (V)**
* **For each city, develop city vs price range visualizations.**
* **Visualise which type of food is most selling in each stateFor types of food(pizza,burger,...) ; vs state;**
* **Visualise which type of food is most selling in each stateFor types of food(pizza,burger,...) ; vs state;(v)**
* **Visualise which type of cosine is most selling in each stateFor types of food(pizza,burger,...) ; vs state;(v)**

**For yelp\_vegas.csv**

* **Gmaps for lat lon**

## Most Frequent Restaurant Categories

## Wordcloud of Top 20 Most Frequent Categories

## Bar Chart of Top 6 Frequent Cuisines

* For a given cuisine, what are the top review counts(american, mexican,indian)

## 

To do :

Perform visualisations on yelp\_cleaned.csv

//json file to csv file: raw file: final\_dataset.csv

//dropped columns