Group No: 5.5

ANALYSIS OF YELP BUSINESS & REVIEWS DATASET
To Help New & Old Businesses Gain Leverage Over Competitors

Project Description

In a bid to stay ahead, companies use data analytics to gauge their standing against competitors and gain leverage over them.

Yelp is a business directory service and crowd-sourced review forum and collects data such as user ratings on a particular business, along with their locations, reviews (text), ratings, timings, likes, dislikes, etc.
A business has many choices when it comes to investing money. Should a restaurant get well-known chefs or excellent lighting? Should a salon invest in high tech appliances or comfortable seats? If many customers talk about "ambience" while giving a 5-star review, it would mean customers really care about great ambience and hence the business should focus on it.

Our aim is to explore this mass of data and answer the following and come up with interesting questions and analyze how or what a business can use to plan their next step to position their product/brand.

- How can we use this dataset to make critical decisions to capture and/or expand market stance?
- Would operating hours impact our businesses' success? Does location affect the business profits?
- What is the impact of business type on its success – e.g. should we invest in a hair salon, or a restaurant?
- Can we gauge if a review posted by a user is a positive or negative, as opposed to reading through millions of records manually?
- If so, can we capture keywords from this text to get an understanding of where the business is going right or wrong?

Dataset Summary

Source – www.Kaggle.com

Dataset (.csv Files **)**	**Shape (** Rows x Columns **)**	**Fields (** Column Headers **)**
Yelp_Business	(174,567 x 13)	business_id, name, neighborhood, address, city, state, postal_code, latitude, longitude, stars, review_count, is_open, categories
Yelp_Business.Hours	(174,567 x 8)	business_id, monday, tuesday, wednesday, thursday, friday, saturday, sunday
Yelp_Review	(5,261,668 x 9)	review_id, user_id, business_id, stars, date, text, useful, funny, cool
Yelp_User	(1,326,100 x 11)	user_id, name, review_count, yelping_since, friends, useful, funny, cool, fans, elite, average_stars

The dataset is a subset of Yelp's businesses, their ratings and reviews, and the user's data across the whole of USA.

- **Yelp_Business.csv**: Includes information about the 174,567 business in the USA - their name, type, location and rating
- **Yelp_Business.Hours.csv**: Includes information about working hours on each day of the week for the business
- **Yelp_Review.csv**: Includes 5 million reviews from 1.3 million users and info. such as - user ID, review ID, ratings, text reviews
- **Yelp _User.csv**: Includes information about 1,326,100 users - user ID, name, no. of reviews posted, friends, status

Teammates & Work Division

- Mahesh. M. Iyer (mmi170000) – Data Manipulation, Analysis, NLP modelling, Report formatting
- Sujith Nair (sxn180033) – Data manipulation, Analysis, Regression modelling, Report formatting
- My Pham (mtp 180000) – Data manipulation, Analysis, Visualization, Regression modelling, Report formatting

Potential Methodologies

- ***Python*** (Numpy, Pandas, SciKit learn, MatplotLib, Plotly, Seaborn, Word-Cloud)
- ***Methodologies*** – Machine learning (Regression modelling, Natural language processing)