Group Members: Dae Hyun Kim(dk58), Yong Hyun Kwon(yk3), Sang Ha Park(sp86), Junsu Choi (jchoi21)

Vision:

- **Idea :** Use the dataset provided by Yelp Dataset Challenge to examine attributes / characteristics that determine the general trend of reviews, visits, etc for a restaurant (e.g. How much does the location factor into a venue's success? How does the venue receive its first "jump" - are there social trendsetters?) Ideally, we would like to submit the project into the Yelp Dataset Challenge.

Data:

- **Dataset (size, parsing/cleaning, etc)**
  - **Size:** ~1.8G for Yelp dataset (+ for additional data - Uber, foursquare, tripadvisor)
  - **Content:** 4 million reviews, 900,000 tips, 1 million visit attributes, aggregated check-ins, etc
  - **Parsing / Cleansing**: JSON format

Methodology:

- **Techniques:** Pandas, Numpy, matplotlib, scikit-learn, machine learning algorithms, regression analysis, IPython Notebook
- **Visualization:** Venues vs location (heatmaps), if possible a correlation between traffic data vs venue check-in data, visualization of social trendsetters (location, activity)

Getting Started:

- **Deliverable by first TA check-in:**
  - Data parsing / cleansing (preprocessing), exploratory data analysis, request Uber movement data
- **By midterm report:**
  - Obtain clear grasp of final deliverable, figure out which additional data sources to use along with Yelp and complete data preprocessing / data aggregation