Project 2: Analysis Spark Application

Group C: Yash Dhayal, Grace Alberts, Hyung Ro Yoon, Leo Chen, & Cameron Lim
Lead: Yash Dhayal

Focus: Yelp Dataset
https://www.yelp.com/dataset/

Utilization: Scala, Spark, Hadoop, Zeppelin
Version Control: Github
Project Management Tool: Trello

Business.json

- Split into 2 tables
  - details on business
  - review score and count for each business
  - Find what are restaurants by the restaurant take out property in attributes

Checkin.json

- Need for date range comparison

Purpose:
Accomplish a series of queries to analyze restaurants, scores, and locations to assist with identifying key points for new restaurateurs or customers.

- City/Town scoring based on businesses in that location
  - To determine the value of cities/towns based on business success

- Average scorings by cuisine type
  - To determine the success rate of type of cuisine

- Popularity of business based on scores by date range
  - Produces a trendline graph on the popularity of a store based on the check-in

- Rating to review count comparison
  - To gauge validity of reviews to the avg review score

- Best restaurants & cities by avg review score and review counts
  - Like a top 10 restaurants listing

- Cuisine to city-level comparison
  - Best type of cuisine in a city based on ratings