Impact of Web & Social Media on Tourism

Background:

The tourism industry thrives on information. The enormous new big data repositories of information can provide traveller, policymakers and other participants in tourism industry with the data-driven evidence needed to make decisions. Travellers during different stages of travel (prior to, during and post travel) leave different digital traces behind on the Web when using mobile technologies. These web and social media data sources can be used to understand the trends and make predictions in tourism industry.

Problem Statement:

Design a digital support system to

* Analyze selected web & social media data about two most popular tourist attractions of Singapore namely, Universal Studios, Singapore and Gardens by the Bay.
* Use these finding to predict number of tourists visiting to these attractions in next three months.

Business Benefit:

This digital support system will benefit following two major participants in tourism industry.

* Travellers:
  + Get consolidated information about a tourist attraction which will help them in making informed decision.
  + Compare tourist attractions which will help them to plan travel.
* Administrators:
  + Review negative feedback from travellers and take corrective steps to improve travellers overall experience.
  + Prediction of number of tourist visiting a particular attraction will help administrators to implement flexible ticket pricing, promotions, adequate security arrangements and other logistic operations.

Solution Overview

Considering the enormity of activities on Web & Social Media, I propose phased approach for this project. The initial scope for the project(Phase-1) is detailed below. The objective is to have live and usable system at the end of phase-1 which can be upgraded with additional features in subsequent phases.

## Data Sources:

* + Attractions:
    - Universal Studios, Singapore
    - Gardens by the Bay
  + Attraction Review Sites
    - TripAdvisor
  + Web Searches
    - Google Trend data on web searches for above two attractions
  + Social Media data
    - Twitter data for above two attractions.
  + Tourist Data
    - Monthly tourist data
    - Hotel Booking Data

## Data Collection:

* + Web and social media data will be collected either by using API provided or by scraping following website.
    - https://www.tripadvisor.com.sg/
    - https://www.google.com.sg/trends/
    - https://twitter.com/
  + Tourist data will be collected from following Singapore government sites. These agencies provide data in excel format.
    - https://www.stb.gov.sg/
    - http://www.singstat.gov.sg/statistics/browse-by-theme/tourism-tables

## Data Analysis - TripAdvisor Reviews

* + I propose to collect following data for each review recorded on TripAdvisor of selected attraction.
    - Attraction - Name of the attraction
    - Attraction Type - Attraction type e.g. Theme Park, Gardens, Nature & Park
    - Review Month - YearMonth (YYYYMM) when review is recorded
    - Visit Month - YearMonth (YYYYMM) of actual visit
    - Reviewer Rating - On the scale of 1 to 5 where 1 = Terrible and 5 = Excellent
    - Reviewer Level - On the scale of 1 to 10
    - Reason for Poor Rating - Extract reason from Text &/or Long Text. Broadly divided into five categories; expensive, long queue, ongoing maintenance, staff and others.
    - Reviewer Type - Five categories of Reviewers; family, couples, solo, business and friends.
    - Text - High level review description.
    - Long Text - Detailed review text.

Please find attached sample file. This file contains mock data and not the actual data from TripAdvisor.

(link to - <https://github.com/tejascphadnis/visitsingapore/blob/master/Reviews.csv>).

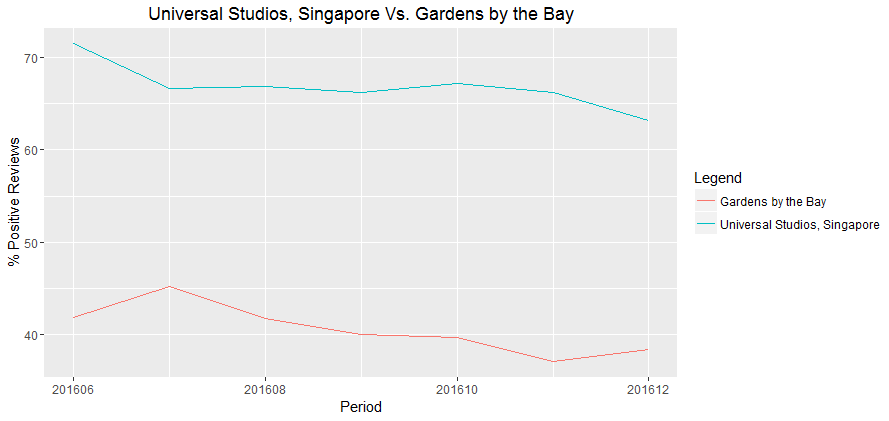
Please find attached R code used for data analysis and plotting graphs.

(link - <https://github.com/tejascphadnis/visitsingapore/blob/master/Reviews.R>)

## Inferences from Data Analysis

Graph1 - % Positive Reviews for Universal Studios, Singapore and Gardens by the Bay.

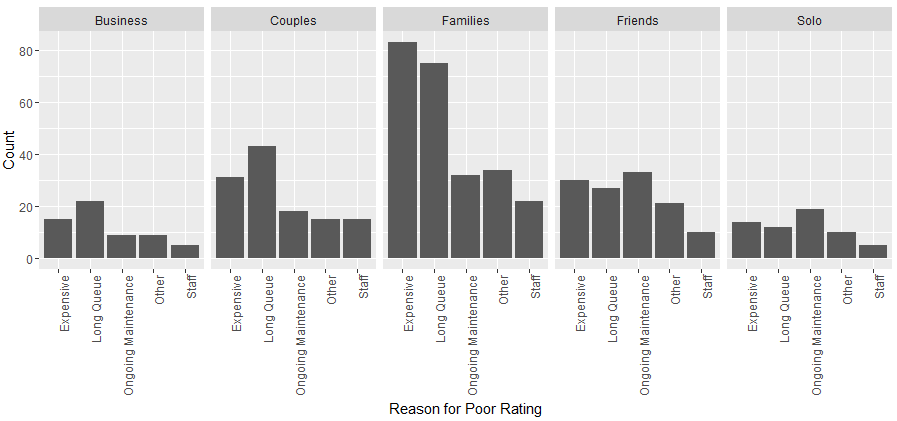
(link - <https://github.com/tejascphadnis/visitsingapore/blob/master/Graphs_Latest.docx> )



* Universal Studios, Singapore receives significantly more positive reviews than Gardens by the Bay.
* % of Positive Reviews for both the attractions is reducing over the period of six months.

Graph2 - Reviewer Type analysis.

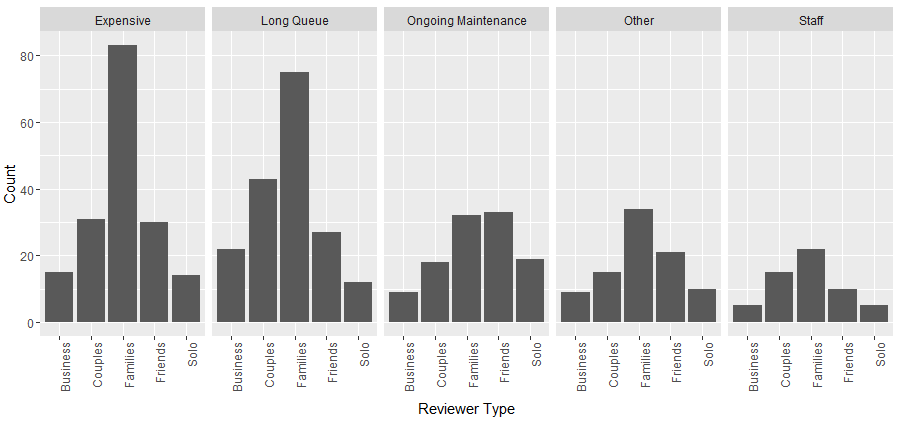
(link - <https://github.com/tejascphadnis/visitsingapore/blob/master/Graphs_Latest.docx>)



* Tourists travelling with "Families" post high number of negative reviews followed by "Couples".

Graph - 3: Reason for Poor rating analysis.

(link - <https://github.com/tejascphadnis/visitsingapore/blob/master/Graphs_Latest.docx>)



* "Expensive" is the main reason for "Families" to post negative reviews.
* "Couples" and "Business" tourist mainly post negative reviews due to "Long Queue".
* "Friends" and "Solo" travellers find non-working rides (Ongoing Maintenance) as the main reason for providing negative reviews.

Sub-sequent phases:

I propose following data sources to be added in subsequent phases.

Data Sources:

* + Attractions:
    - Singapore Zoo
    - Night Safari
    - Jurong Bird Park
  + Web Searches
    - Google popular times info about eateries (like McDonalds) in the vicinity of attraction might indicate number of travellers visiting attraction.
  + Social Media data
    - Facebook
    - Flickr / Instagram
  + Blogging Sites
  + Transaction Data
    - Airline Booking
    - Tour online booking