Data Science: Coursework Assignment

Introduction

This is the coursework assignment associated with the Data Science module. The hand-out date for these instructions is 11th February 2021. A report response is to be submitted by **1700hrs on 25th March 2021**. A hard copy of the report should be submitted to the course management team. The report should be written using Google colab.

Background

The dataset is a modified version of the VAST 2015[1] challenge. Some information taken from the VAST 2015 website is repeated here, to ensure that these instructions are self-contained.

DinoFun World

DinoFun World is a typical (but simulated, totally hypothetical) modest-sized amusement park, sitting on about 215 hectares and hosting thousands of visitors each day. It has a small-town feel, but it is well known for its exciting rides and events. You are supplied with two datasets covering a three-day period: movement data and communications data.

Movement data

The first dataset (movement data) contains positional information of all visitors to the theme park covering three days (Friday, Saturday and Sunday). The simulated park covers a large geographic space (approx. 500x500m^2) and is populated with ride attractions, restaurants and food stops, souvenir and game stores, an arcade, a show hall, and a performance stage. The attractions are categorised into Thrill Rides, Kiddie Rides, Rides for Everyone, Food, Restrooms, Shopping, and Shows & Entertainment. All visitors to the park (except for very young children) use a park app to check in to the park and attractions and to communicate with fellow visitors. If visitors do not have compatible phones, they are provided with loan devices. Visitors are assigned unique IDs and must use the app to check into rides and some other attractions. The park is equipped with sensor beacons that record movements within the park. Sensors are sensitive within a 5m x 5m grid cell. All pathways in the park are covered by these sensors, as are the ride check in locations. Locations are not recorded while people are on rides or inside attractions (including restaurants, stores, and rest rooms).

The park area is gridded to assist in specifying locations. Each movement data file contains movement information around the grid, with coordinate locations. This data appears as follow: 2014-6-06 08:00:08,5231584,check-in,63,99

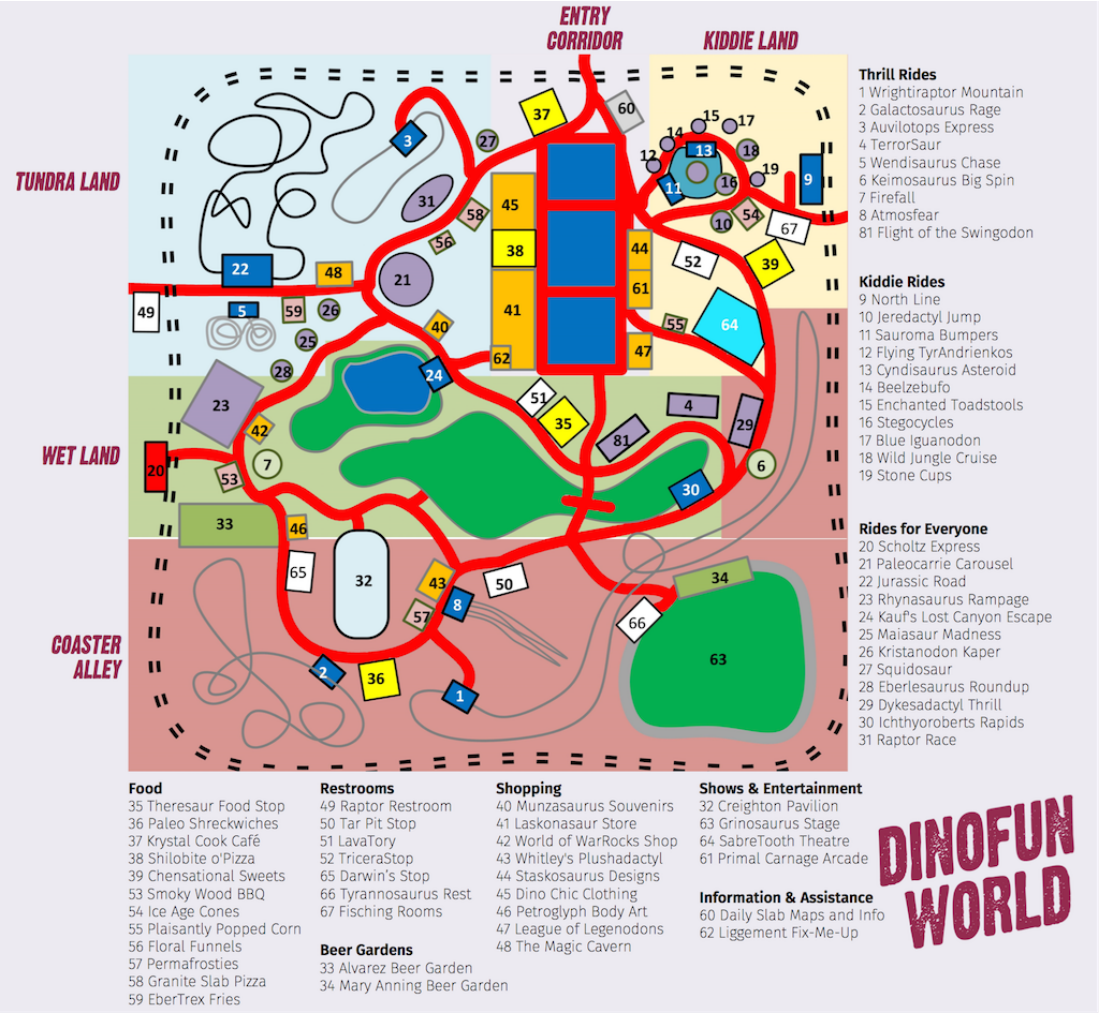
2014-6-06 08:00:42,93275931,movement,64,98

The values are: timestamp, person-id, type of activity (either “check-in” or “movement”), X-coordinate, and Y-coordinate.

People either move from grid square to grid square, or “check in” at attractions, meaning they either get in line or onto the ride. So, above, a person with ID 5231584, checked into a ride located at (63,99) at 8:00:08 AM. Person 93275931 just moved to location (64,98) at 8:00:42. As a person travels through the park, their locations are recorded in this file at a one-second resolution. If there is no record during a particular second of time, then the individual has not moved out of their previous grid square. People

[1] http://vacommunity.org/VAST+Challenge+2015

are not tracked after they check into a ride – they will eventually appear back on the grid when the ride is over. The park map (and associated rides, attractions, etc) are displayed below.



The attractions are numbered and coded according to type. The red line indicates the pathway through the park, although dark green areas are also areas where people can move. (Attraction 30 in the middle of the map is a water rapids ride, so people can watch from the "inside" of the ride boundaries. For other rides, people are not allowed to wander inside of the ride footprint. Attraction 63 is a show stage area, so people populate this area during performances).

Communications data

The second dataset relates to communication data; you have access to the in-app communication data over the three days, covering the same period as the movement data. This includes communications between the paying park visitors, as well as communications between the visitors and park services. In addition, the data also contains records indicating if and when the user sent a text to an external party. The data appears as follows:

2014-6-06 08:03:19,439105,1053224,Kiddie Land

2014-6-06 08:03:47,1836139,1593258,Entry Corridor

The fields for this file are: timestamp, from (the sender ID), to (the recipient ID), and location (area where communications occurred). Location can be: Entry Corridor, Kiddie Land, Tundra Land, Wet Land, or Coaster Alley.

Acquiring the data

The data are available in the following folder:
https://drive.google.com/drive/folders/1HjjenE-JFRE0n-0BtFnvTqwHr--Ed13-?usp=sharing

The following three files contain movement information, one for each day:

movement-Fri.csv
https://drive.google.com/open?id=1-PtQsQK_CFW5p_kMyMPEYpfVsdOanWGR

movement-Sat.csv
https://drive.google.com/open?id=1lNVQSMi68hgCPd4ZGPE-fFjW8FR-_BPr

movement-Sun.csv
https://drive.google.com/open?id=1FmC7-XxYPrRk3M0nHfefAict41UnKamy

The following three files contain communications information, one for each day:

comm-data-Fri.csv
https://drive.google.com/open?id=1cRCZxuCMRgi1w5Ni2H3flkUsAXMSQ0o1

comm-data-Sat.csv
https://drive.google.com/open?id=1k4iExgtVW6YgRYg_wGQLcIGodLcihz5t

comm-data-Sun.csv
https://drive.google.com/open?id=1zd9DnPxdVzNgbxofgIxq5_FfhGtbSCFA

Assignment

During the three days of the data supplied, DinoFun World held a weekend tribute to Scott Jones, internationally renowned football star. Scott Jones is from a town near to DinoFun World. He was a classic hometown hero, with thousands of fans who cheered his success as if he were a beloved family member. To celebrate his years of stardom in international play, DinoFun World declared “Scott Jones Weekend”, where Scott was scheduled to appear in two stage shows each on Friday, Saturday, and Sunday to talk about his life and career. In addition, a show of memorabilia related to his illustrious career would be displayed in the park’s Pavilion.

However, the event did not go as planned. Scott’s weekend was marred by crime and mayhem perpetrated by a poor, misguided and disgruntled figure from Scott’s past.

While the crimes were rapidly solved, park officials are interested in understanding just what happened during that weekend to better prepare themselves for future events. They are interested in understanding how people move and communicate in the park, as well as how patterns change and evolve over time, and what can be understood about motivations for changing patterns.

Write a report in Google colab summarising your exploration of the data, any modelling that is performed, and any conclusions that you may derive. The recipients of your report are park officials, some (but not all) of whom are data science literate. The report should document your exploration of the data, mathematical details (where appropriate), python code, and visualisations.

In order to assist you in this assignment, the following questions are provided to guide your thinking. You may answer zero, one, or more of the questions, and/or derive your own questions and answers.

Movement data

- Characterise the attendance at the park on this weekend. Describe up to five different types of groups at the park on this weekend.
    a. How big is the group type?
    b. Where does this type of group like to go in the park?
    c. How common is this type of group?
    d. What are your other observations about this type of group?
    e. What can you infer about the group?
    f. If you were to make one improvement to the park to better meet this group's needs, what would it be?
- Are there notable differences in the patterns of activity on in the park across the three days? Please describe the notable difference you see.
- What anomalies or unusual patterns do you see? Describe no more than five anomalies and prioritise those unusual patterns that you think are most likely to be relevant to the crime.

Communication data

- Identify those IDs that stand out for their large volumes of communication. For each of these IDs:
    a. Characterise the communication patterns you see.
    b. Based on these patterns, what do you hypothesize about these IDs?
- Describe up to five communications patterns in the data. Characterise who is communicating, with whom, when and where. If you have more than five patterns to report, please prioritise those patterns that are most likely to relate to the crime.
- From this data, can you hypothesise when the crime was discovered? Describe your rationale.