ASSIGNMENT 1

Understanding the Data Set-

The dataset used in the report is a global-scale check-in data collected from Foursquare. Data is collected for 18 months from April 2012 to September 2013. Below is the link of data.

https://drive.google.com/file/d/0BwrgZ-IdrTotZ0U0ZER2ejI3VVk/view

The dataset contains three .tsv files.

The first file is named 'dataset_TIST2015_Checkins' which contains all the 33M check-ins occurred in the world for the aforementioned timeline. This file has the data of 266,909 users checking into the 3.6M venues spread across the 415 cities of 77 countries.

The second file is named 'dataset_TIST2015_POIs' which contains the information like Venue-ID, Latitude, Longitude, Venue Category, and Country Code about the 3.6M venues. Country Code is a two-letter code given to the countries as per ISO 3166-1 alpha-2 standard. There are total 429 venue categories which has sub-categorize the 3.6M venues into groups like Bar, Home, Office, Train Stations, University etc.

The third file is named 'dataset_TIST2015_Cities' which contains the information of 415 cities of the 77 countries. This file contains City Name, Latitude and Longitude of City Centre, Country Name, City Type, and Country Code. The cities are sub-categorize by City Types like National Capital, Provincial Capital etc.

Technical Overview of the Report-

This report is designed by keeping three major data modelling aspects as foundation and using those models in depicting the most famous venue of the world. Decision of a venue being famous has been based on the number of check-ins occurred at that venue.

First model used is the Relational Model. This modelling technique helps in performing operations like Join, Sort etc. on tables. Linking the 3 files given using relational modelling to frame an entity relationship tends to display key information of the dataset. Using relational database constraints pointed out the most checked-in venue of the world and in which country that venue is.

Second model used is the Spatial Model. After finding the key information about the venues from our first model, spatial modelling is used to scrutinize the given Latitude and Longitude of cities' centres and Latitude and Longitude of Venues. So, this modelling unveils the venue's location with respect to the city.

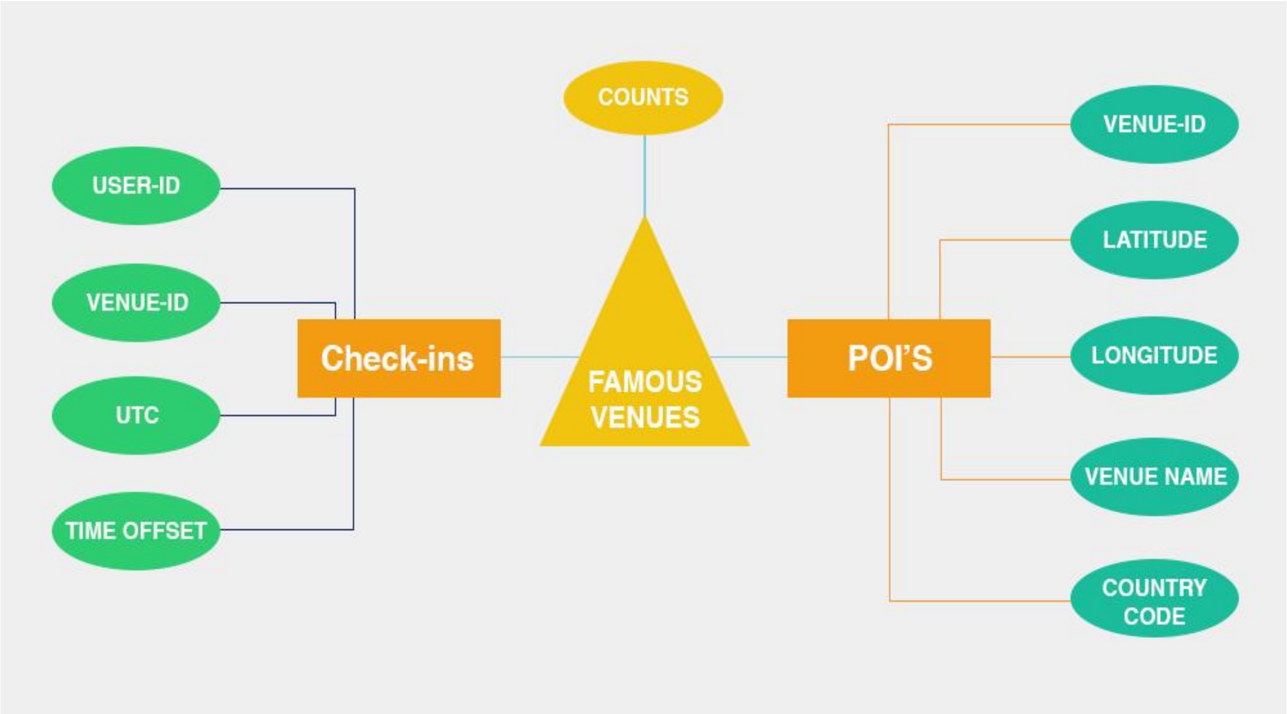
Third model used is the Temporal Model. Since, given dataset contains UTC time and Time Offset of each check-in that has occurred, this modelling is advantageous in figuring out the pattern of check-ins over the course of 18 months in the most famous venues of the world.

Major module used in this report is Pandas of Python. It is used for data import, data export, and performing operations on dataset. For Data Visualizations, Tableau software is used.

Most Famous Venues of each Country-

To extract the information of venues, entity-relationship model is used to join the 2 files, i.e. check-ins and POIs. Using ER model, a relation was framed depicting the number of check-ins that has occurred in that particular venue. A column named 'Counts' was formed to show the number. Also, using country code and pandas Data-Frame concatenation operation, a column of country name was also added to the 'Famous Venues' relation.

Below is the basic ER diagram that was used to form the 'Famous Venues' relation.



(Fig. 1)

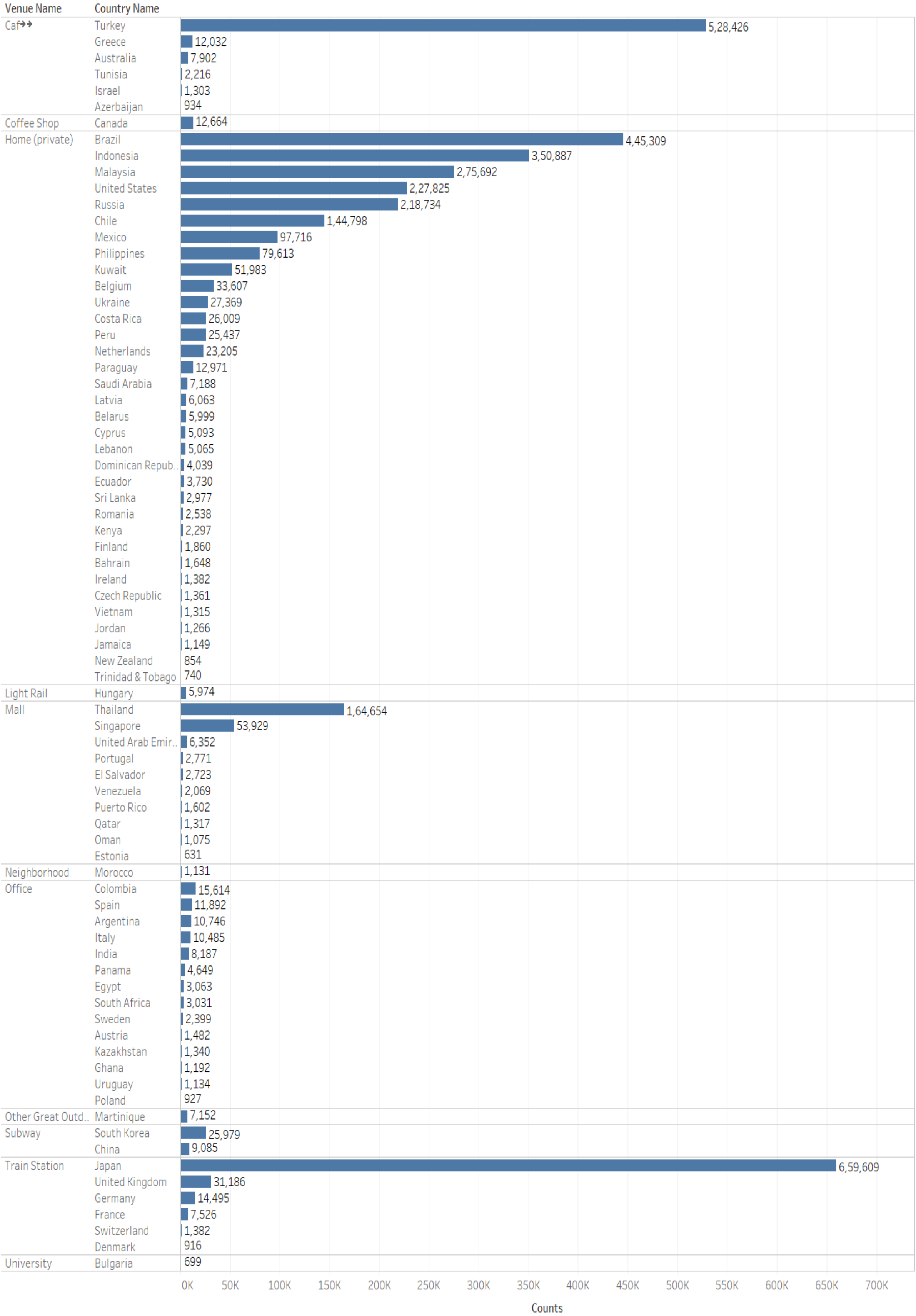
The dimensional relations are joined using a left join operation of Pandas tool in python to form a fact relation.

Using, the above Famous Venues Relation, a data visualization is created to pictorially represent the most famous venues of each country and the number of times they were visited.

Below, visual rendition clearly displays that the Train Stations of Japan are the most famous venue of the world. These train stations have been visited 659,609 times over the course of 18 months. This portrays that Japanese people have a genuine tendency to use public trains for commuting as compared to the other countries. This habit is very helpful in the preservation of environment and in reduction of pollution.

Also, from below bar diagram, it's clear that nearly 50% of the countries' maximum check-ins happened at the users' private property, their homes.

Most Famous Venues of Each Country



(Fig. 2)

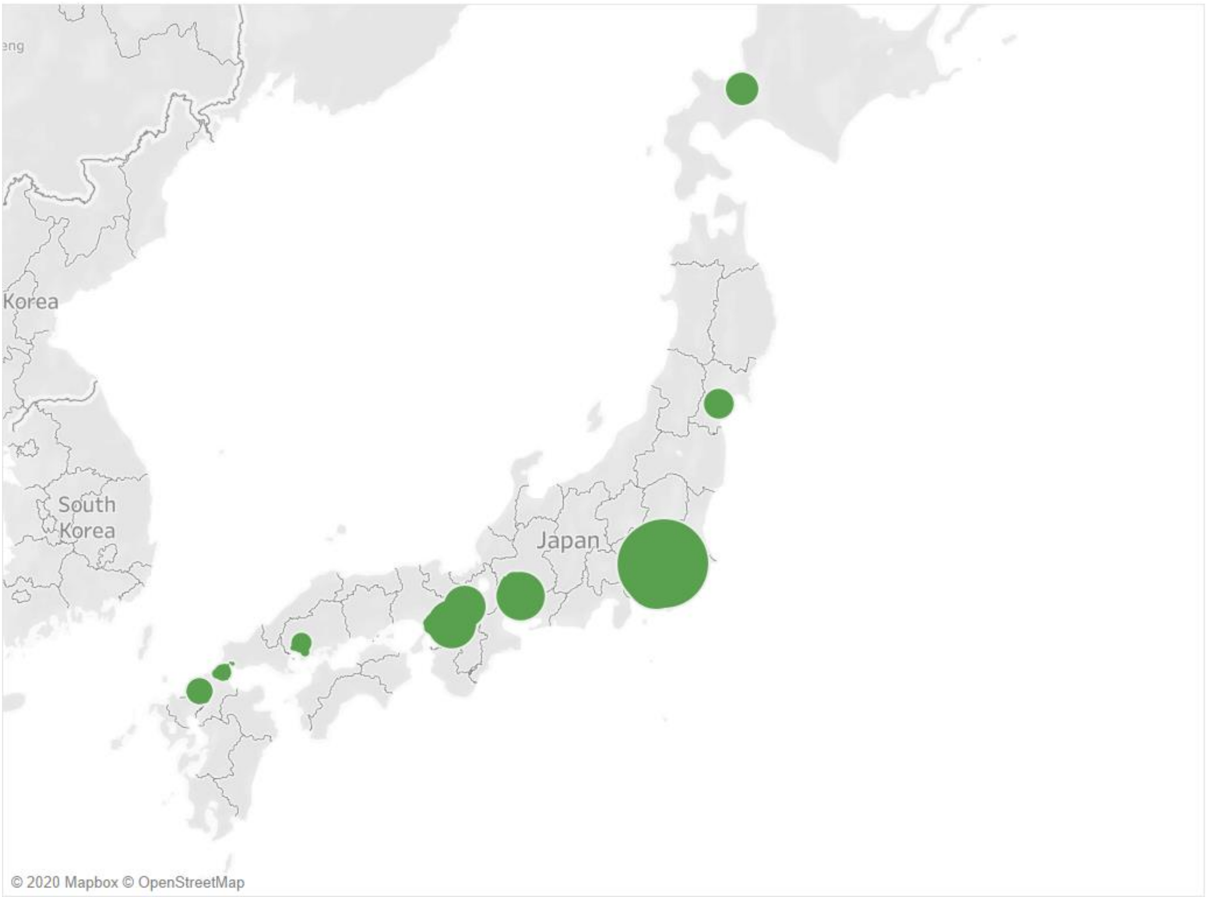
Spatial Representation of Train Stations of Japan-

After depicting that the most famous venues of the world, out of all the countries, are Japan's Train Stations, now it's time to display the provinces in which those check-ins have occurred in Japan.

So, below is the map of Japan showcasing the provinces in which Train Stations were checked-in.

The diameter of the marker indicates that prominence of the train stations in that province.

Train Stations of Japan



(Fig. 3)

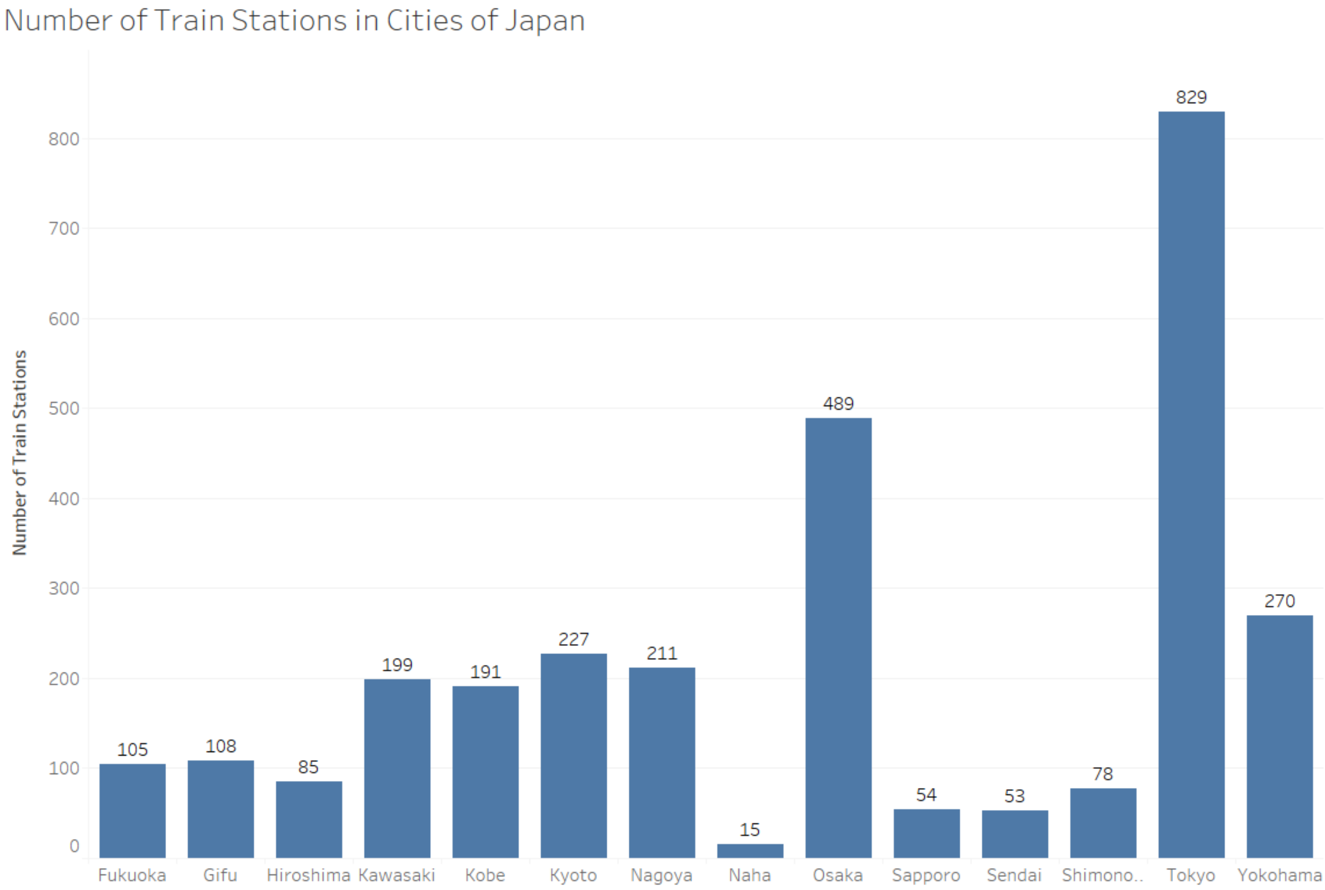
It's been observed that the Train stations of Japan are most checked-in venue of the world but still the number of stations in the cities mapped above is unknown. So, it's time to find out in or near which cities of Japan have these train stations are located.

Using the K-Nearest Neighbours case of Spatial Modelling, the number of train stations in or near those cities are depicted.

Since, the location of city centre is given in data set and we have locations of all the train stations in Japan. We fixed the location of city centre of all the cities in Japan and performed spatial operation of finding the distance between the centre and the train stations. So, whichever train station has a minimum distance to the city centre, that station is mapped to that city's name concluding that the aforementioned station belonged to the city with which it's distance to the centre is minimum.

To calculate the distances between the two coordinates, we used Haversine's formula. We first converted the latitude and longitude into radians using Python library 'NumPy'. Then we calculated the distance using the formula and stored the values in Pandas series. After that we imported that series into a DataFrame columns corresponding to the City Name which gave us a 2D representation of the distances from the city centres to the train stations.

Then we traversed through the 2D array and framed a series which gave us all the minimum distances for all the train stations corresponding to the city centres. Then we used data visualization tool to display the below bar graph which represents the Number of Train Stations which are either in that city or closest to the mentioned cities.



(Fig. 4)

We can clearly see that Tokyo has the maximum number of Train stations and hence the maximum number of check-ins as seen in the map of Japan above.

Now, since the number of stations is known in each city of Japan. It's time to figure out the pattern in which these check-ins were made.

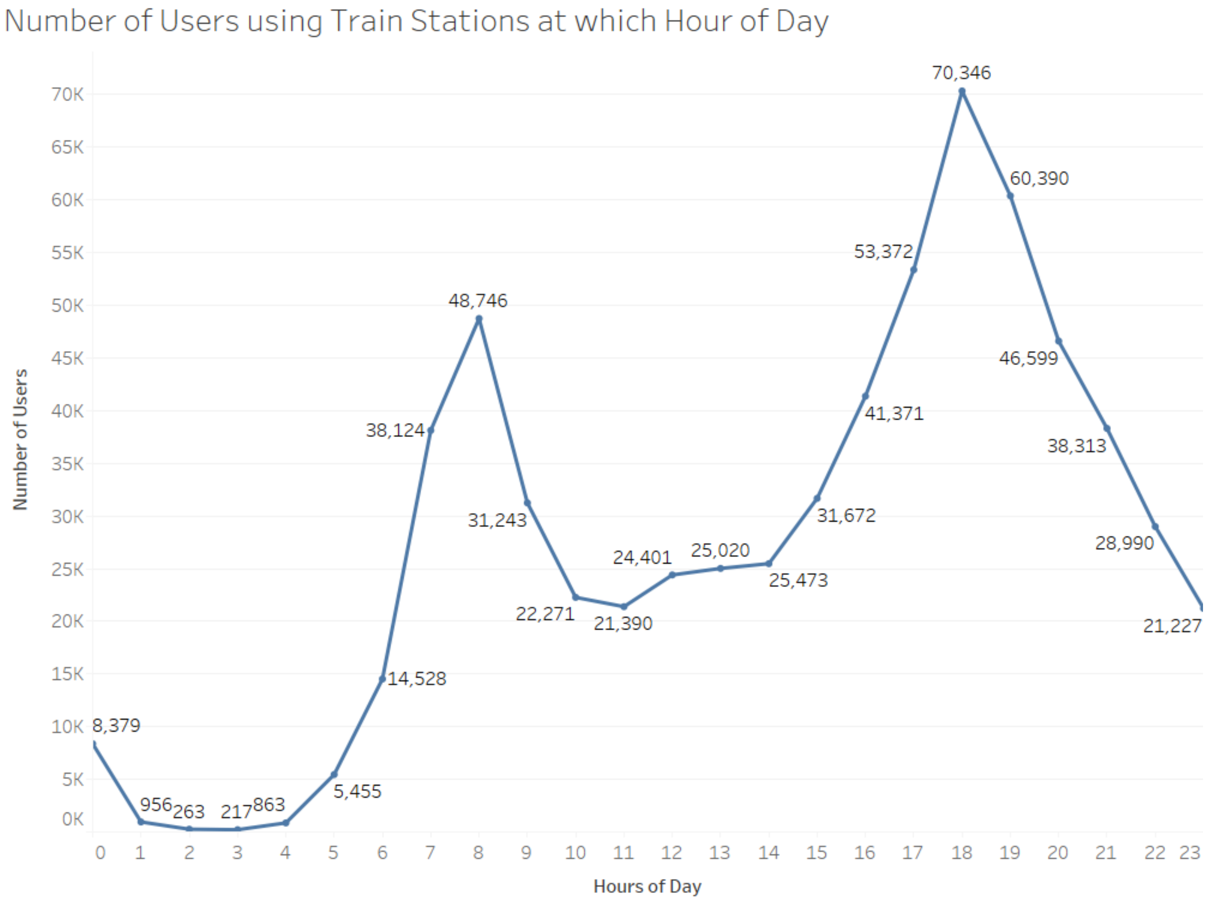
Hence, let's investigate and see at what time these train stations of Japan were used and at what time they were not in much use.

Temporal Pattern of Train Stations of Japan-

In order to see the pattern of check-ins in train stations, we have mapped the Number of Users with the Time of the Day in the below graph.

Here, we have displayed that at each hour of the day, how many users are using the train stations in Japan. We have calculated the number of users checked in at each hour in the station and then mapped it onto the pictorial graph representation.

We calculated the local time of Japan from the given UTC time and Time Offset. We used python's library 'DateTime' to convert UTC time in a desired format and then add the Time offset to the UTC time which gives us Local Time.



(Fig. 5)

From above chart, we can conclude that between 00:00 and 06:00 hours, these train stations are least utilized as it's obvious and unsafe to travel in night time. Also, most of these train stations might be closed for this particular time interval which supports the argument of least utilization at this time.

As we all know, that a normal working person starts their day in the morning, goes to work and comes back home in the night after finishing the office hours. This general theory is also supported by two major peaks in the graph.

There's a drastic increase in the check-ins from 06:00 to 08:00 in the morning which indicates the start of the day for people who like to travel to their work using trains. Also, second peak at 18:00 in the evening shows that people likes to travel back from work in train as well. So, we can depict a time related pattern here in the usage of train stations by the people of Japan.

Since, we are done observing a pattern in the most famous venue of the world per country. Now, as we know from Bar graph 1, most famous venue category in the world is their Home. Fig. 2 clearly shows that nearly 50% countries have maximum checked-in venue as their Home. So, let's try to find a spatial or temporal pattern in the check-ins done by people at their Home.

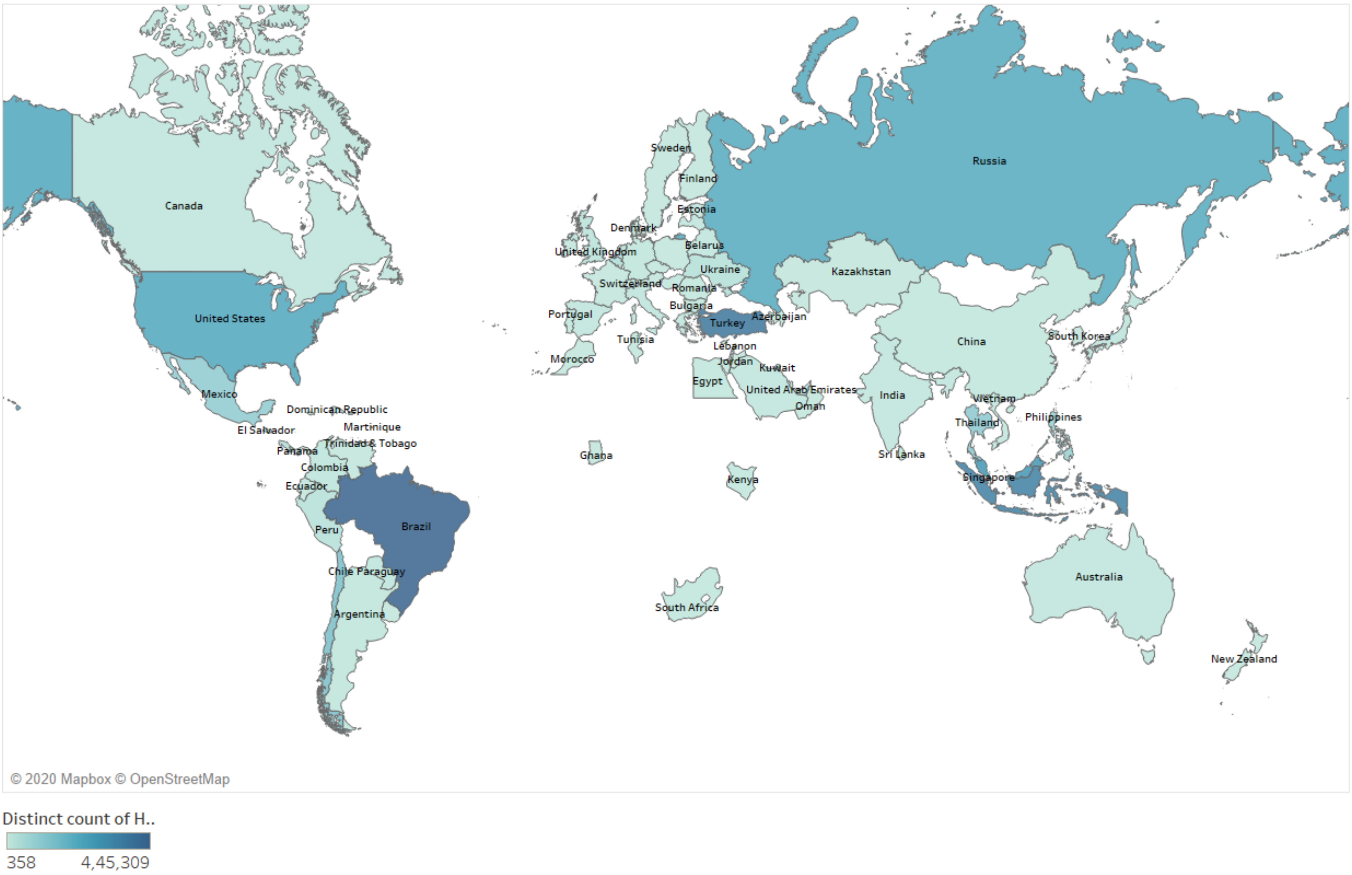
Representation of Check-ins done at Home (Private) –

Since, we have established from bar graph 1 that Home is the most checked-in venue category of the world, we will now use our three modelling techniques to represent the database relations for the users and countries with check-ins at home.

Below is the visualization of count of check-ins occurred in each country present in the database.

It shows the increasing number of check-ins as the shade of blue gets darken in the visualization as shown in the scale below.

Count of Home Check-ins colouring across 77 countries



© 2020 Mapbox © OpenStreetMap

(Fig. 6)

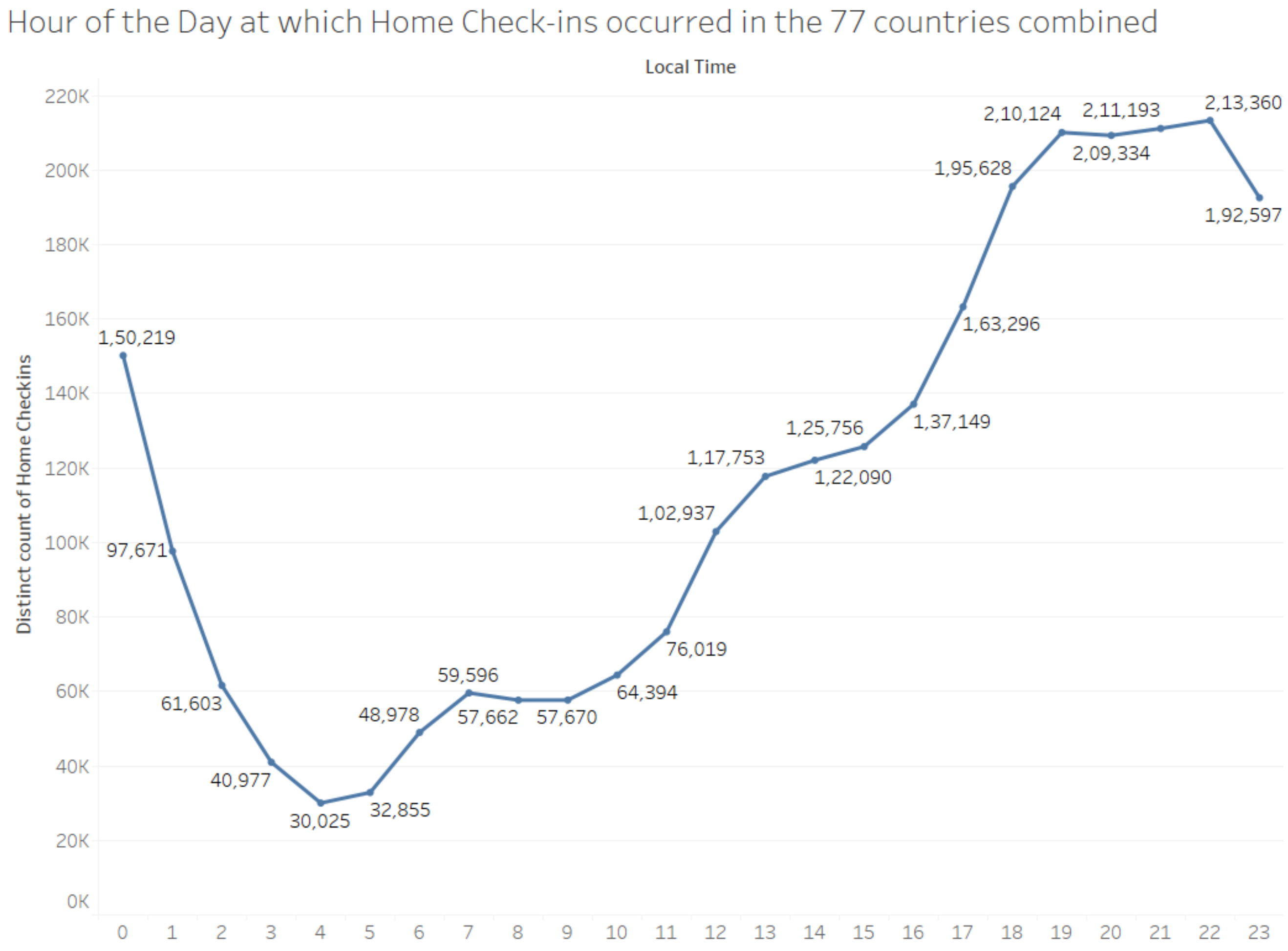
So, as we can see that Brazil has the darkest shade of Blue in the map, it indicates that Brazil has the most home check-ins out of the 77 countries of the world. And from the scale we can see that total check-ins in Brazil matches the Bar graph 1's depiction in Fig. 2 (Home Check-ins in Brazil = 445,309).

To get this constrained database, we used join operations on the dataset files and applied condition 'VenueName' = 'Home (private)'. This gave us a relation with all the Home check-ins done by all the users in all the countries of the dataset.

Now, with the relation we have, let's see if there's a pattern of check-ins happening at Homes with respect to time.

So, we used the same formula and modules to convert UTC time into Local Time for each country. We saved those times into a column replacing UTC time in the relation we have. Using local time helps us to investigate a day-night habit, if any, shown by the users just like there was in Train Station case of Japan.

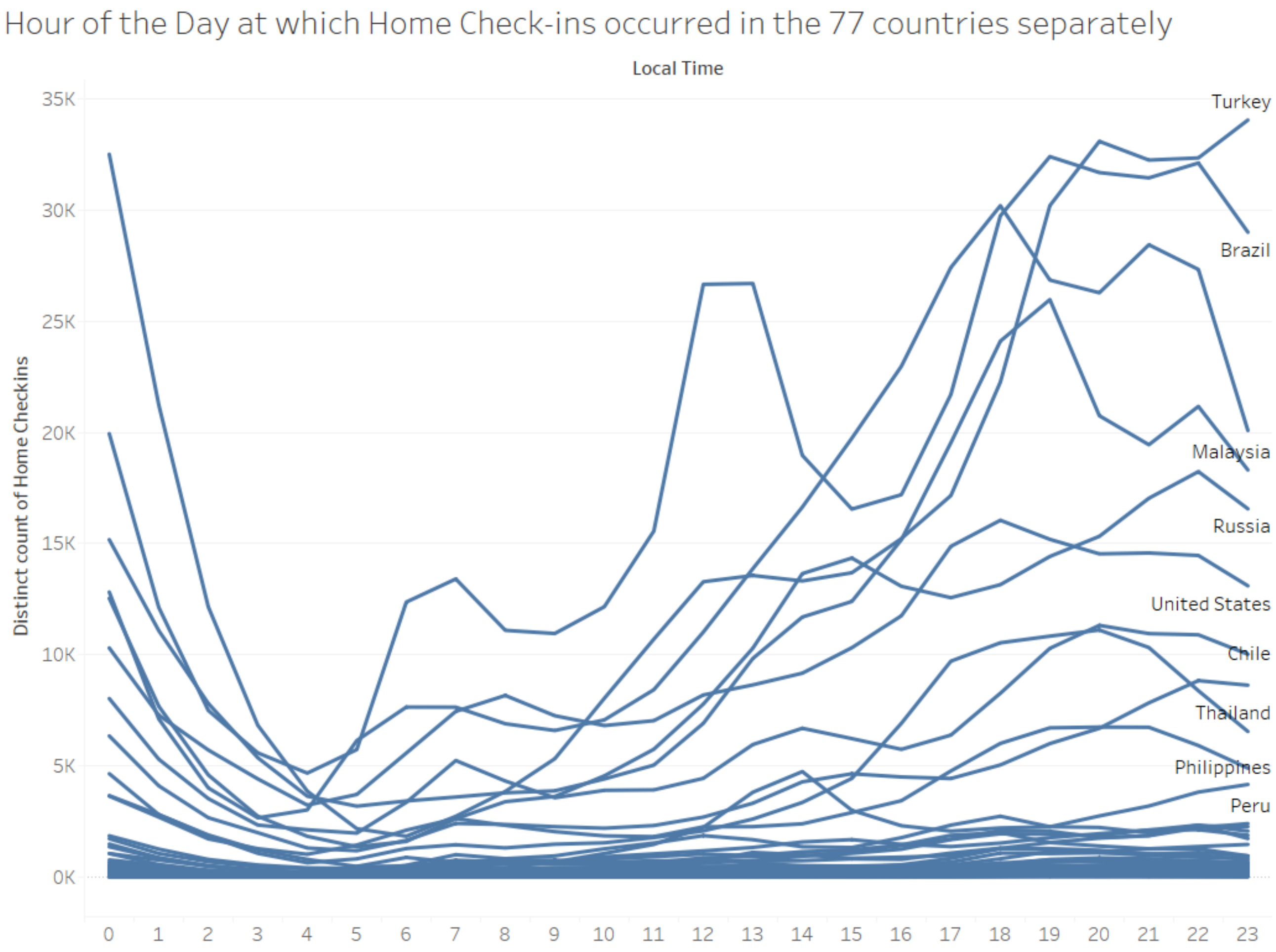
Hence, we build a temporal model spread around the 24 hours of the day. We calculated all the check-ins occurring in that particular 60mins of the day and plotted a pathway showcasing the maximum number of check-ins happening at Home in each hour.



(Fig. 7)

In above graph, x-axis represents the Hours of the day and Y-axis represents the count of check-ins happening at that particular hour.

Let's plot the above graph separately for each country and see if the pattern holds true for all the countries of the world seen distinctly.



(Fig. 8)

As we can clearly see a similarity in the pattern of Home check-ins if we compare each country separately with the count of all the countries taken together.

Above graph displays a dip in the Home Check-ins from 00:00 to 04:30. Which is quite obvious to state that at that time of the night most of the people must have already been checked into their home in the preceding hours and might be long asleep in this timeline.

Also, we can clearly observe that the maximum Home check-ins happened from 18:30 to 23:00 hours. This proves our previous point and also tells us that people who have gone out of their house for work, party, or any leisurely activity, have returned to their home in the night.

So, we can conclude a definite time-based pattern in the home check-ins all across the world.

Conclusion of the Report-

We have analysed the given dataset and have seen some interesting insights which relates to the real-world habits or attributes people have in daily life.

Firstly, we did operations on our database to find out the most checked-in venues of each country in our dataset. We performed relational database modelling on our files and created a data visualization to show the values of maximum check-ins for each country with the help of a horizontal bar graph (Fig. 2).

Secondly, after figuring out that the most checked-in venues of the world per country are Train Stations of Japan, we tried to display the count of check-ins occurring onto the map of Japan using the Latitude and Longitude of the venues we had in our dataset (Fig. 3). This showed us the particular areas and provinces of Japan where the train stations were actually located and were checked-in by the users.

Thirdly, we try to pin point the location of each train station of Japan to the cities present in our dataset using a spatial modelling concept known as K-Nearest Neighbour. Using the Latitude and longitude of the city centres of Japan's major cities, we divided our train stations data and mapped each train station to that particular city with which it's distance from the centre was minimum. This is displayed in Fig. 4 with the help of a vertical bar graph.

Now, after mapping each train station to its city, we were tempted to see any time-based pattern that must have occurred in these check-ins of train stations. So, we converted the given UTC time to Local time and used our temporal concepts to define the occurrence of check-ins of train stations (Fig. 5) which showcased a real-life scenario of most of the people in this world.

Lastly, we depicted from the Fig. 1 that the most famous venue category across the whole world was Home. So, we modelled a relation for check-ins occurring at the home of users in all the 77 countries and we tried to analyse a pattern in those check-ins. We mapped the count of Home check-ins for each country to a world map which displays the count as a shade of blue colour (Fig. 6). Then, to see a time-based pattern in Home check-ins we applied temporal model concepts and visualized 2 graphs showing total count of check-ins per Hour for whole world collectively (Fig. 7) and showing total Home check-ins per Hour for each country separately (Fig. 8).

These analyzations display a definite pattern with a real-world daily life attribute of people.