Data Science Project

Traffic Incidents in the UAE – An Analysis

IBM Data Science Professional – Applied Data Science Capstone

Aparna Vadakedathu Thulasiprasad
2/16/2020

Contents

Introduction -Business Problem

According to the UN-WHO Road Accidents is the second major cause of deaths in the Emirates and this rate is increasing at a rapid speed with every passing year. A survey by the WHO conducted on UAE Road Accidents shows that 63% of children deaths occurred in road accidents. UAE has a modern and state of the art roads and traffic infrastructure, but there are many other factors involved in the rapidly increasing number of road accidents. Dubai government takes immense efforts to ensure the happiness of its residents and the quality of travel is one of the key indicators.

Based on information that is available on past accidents, is it possible to prevent accidents or response from emergency responders once an accident has occurred

1. What are the major causes of accidents in Dubai?
2. Are there locations that are more accident prone than others?
3. Can emergency responder hubs be established at strategic locations to improve response time once a distress call is placed after an accident?

Data

Middlesex InsightsX Lab: Applied Data Analytics (https://www.mdx.ac.ae/insightsx-lab) is the Data Science hub to connect the analytics community and academic fraternity with organizations willing to derive insights from their data. The lab enables academic researchers and the data science community in the UAE to help companies understand how to better monetize their data through the development and application of new predictive models and analytical approaches. Several datasets have been put together and curated for use by the data science researches in UAE.

Description of Dataset1: Traffic_accidents.csv

This particular data set has been sourced from an organisation called Bayanat (https://data.bayanat.ae/) which hosts several data sets related to governament or public entities in Dubai. A traffic incidents data set for the year 2017 could be obtained and has been used in this analysis

Each row is an incident of the 2017 accident case and columns are the attributes related to the accident. Some fields are in Arabic, but the English one could be downloaded. Very important details like accident location, weather, seatbelt, intoxication, and others are shared for each accident.

Link: www.bit.ly/mdxtrafficdata1 (Data File) Related
(Source): http://data.bayanat.ae/en_GB/dataset/traffic-accidents

Description of dataset2: Traffic_incidents.csv

This particular data set has been sourced from an organisation called Dubai Pulse (https://www.dubaipulse.gov.ae/) which hosts several data sets related to governament or public entities in Dubai.

It is a data set related to the traffic accidents since August 2018 in Dubai. The best part about this data set is that the data is getting constantly updated by Dubai Police. Which can give us an opportunity to verify the preventive actions taken based on our findings.

For each accident: Unique Accident ID, accident Time, name of the person involved, accident place coordinates are provided.

Link: http://bit.ly/2qjZGmw (Data) Related: https://www.dubaipulse.gov.ae/data/dp-traffic/dp_traffic_incidents-open?organisation=dubai-police&service=dp-traffic

Foursquare data

Foursquare data can be used to identify the top venues around which accidents occur. This can be strategically used by law enforcement for strengthening patrolling, surveillance and establishing emergency responder hubs

Methodology

Data Wrangling

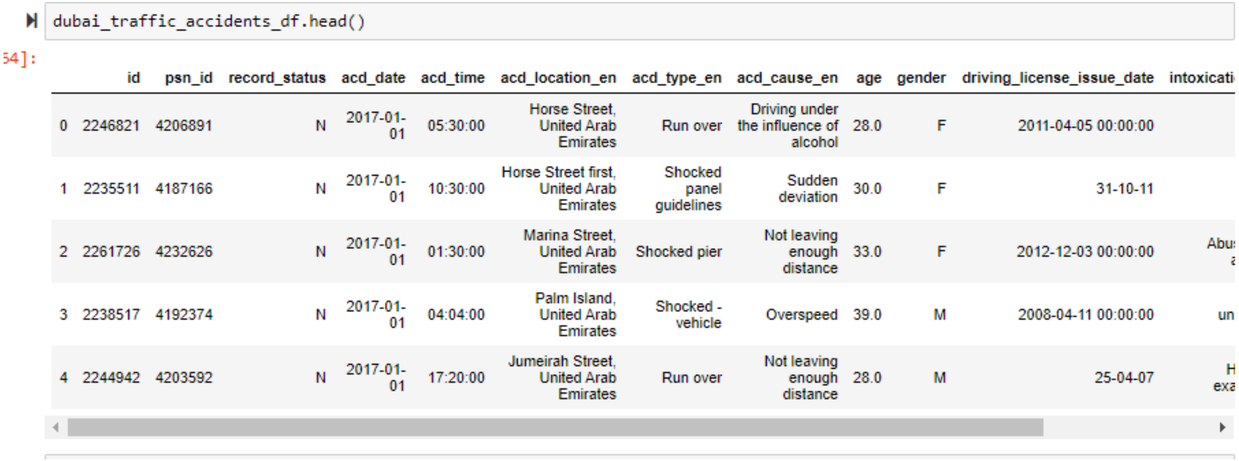
Traffic Accidents Data Set

A look at the data set provides insight into the information contained. There are several important attributes like Location, cause of accident, Age of driver, year of obtaining driver's license and gender of driver.

In the original data set, many fields were in Arabic. This could easily be translated by uploading in Google Sheets and using the GoogleTranslate() function. Alternately I have also explored the possibility of using googletrans API. However this proved to be more cumbersome due to the timeouts that frequently occur even with throttled requests.

Below Data Cleansing steps were carried out to clean up the data frame

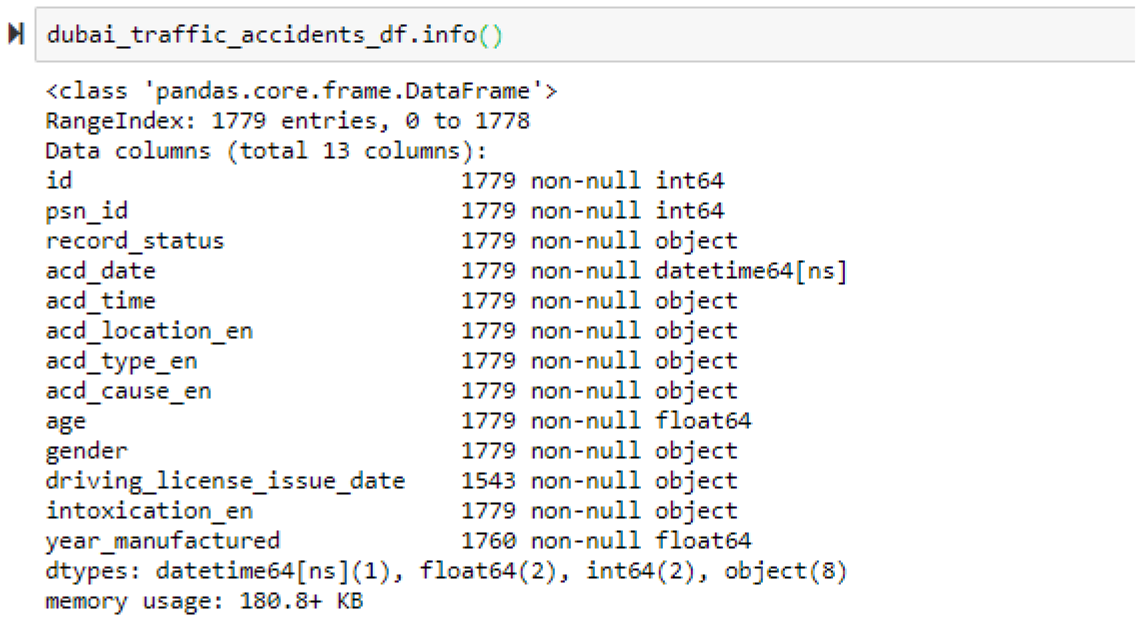
1. Drop all rows containing arabic text or attributes that are not sufficient or significant for analysis
2. Drop all rows that does not have an accident
3. Replacing missing age values with average age
4. Drop rows for which gender data is missing.
5. Convert accident time to date time format



```
dubai_traffic_accidents_df.head()
```

54]:

	id	psn_id	record_status	acd_date	acd_time	acd_location_en	acd_type_en	acd_cause_en	age	gender	driving_license_issue_date	intoxicati
0	2246821	4206891	N	2017-01-01	05:30:00	Horse Street, United Arab Emirates	Run over	Driving under the influence of alcohol	28.0	F	2011-04-05 00:00:00	
1	2235511	4187166	N	2017-01-01	10:30:00	Horse Street first, United Arab Emirates	Shocked panel guidelines	Sudden deviation	30.0	F	31-10-11	
2	2261726	4232626	N	2017-01-01	01:30:00	Marina Street, United Arab Emirates	Shocked pier	Not leaving enough distance	33.0	F	2012-12-03 00:00:00	Abu a
3	2238517	4192374	N	2017-01-01	04:04:00	Palm Island, United Arab Emirates	Shocked - vehicle	Overspeed	39.0	M	2008-04-11 00:00:00	un
4	2244942	4203592	N	2017-01-01	17:20:00	Jumeirah Street, United Arab Emirates	Run over	Not leaving enough distance	28.0	M	25-04-07	H exa



```
dubai_traffic_accidents_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1779 entries, 0 to 1778
Data columns (total 13 columns):
id                             1779 non-null int64
psn_id                         1779 non-null int64
record_status                  1779 non-null object
acd_date                       1779 non-null datetime64[ns]
acd_time                       1779 non-null object
acd_location_en                1779 non-null object
acd_type_en                    1779 non-null object
acd_cause_en                   1779 non-null object
age                            1779 non-null float64
gender                         1779 non-null object
driving_license_issue_date     1543 non-null object
intoxication_en                1779 non-null object
year_manufactured              1760 non-null float64
dtypes: datetime64[ns](1), float64(2), int64(2), object(8)
memory usage: 180.8+ KB
```

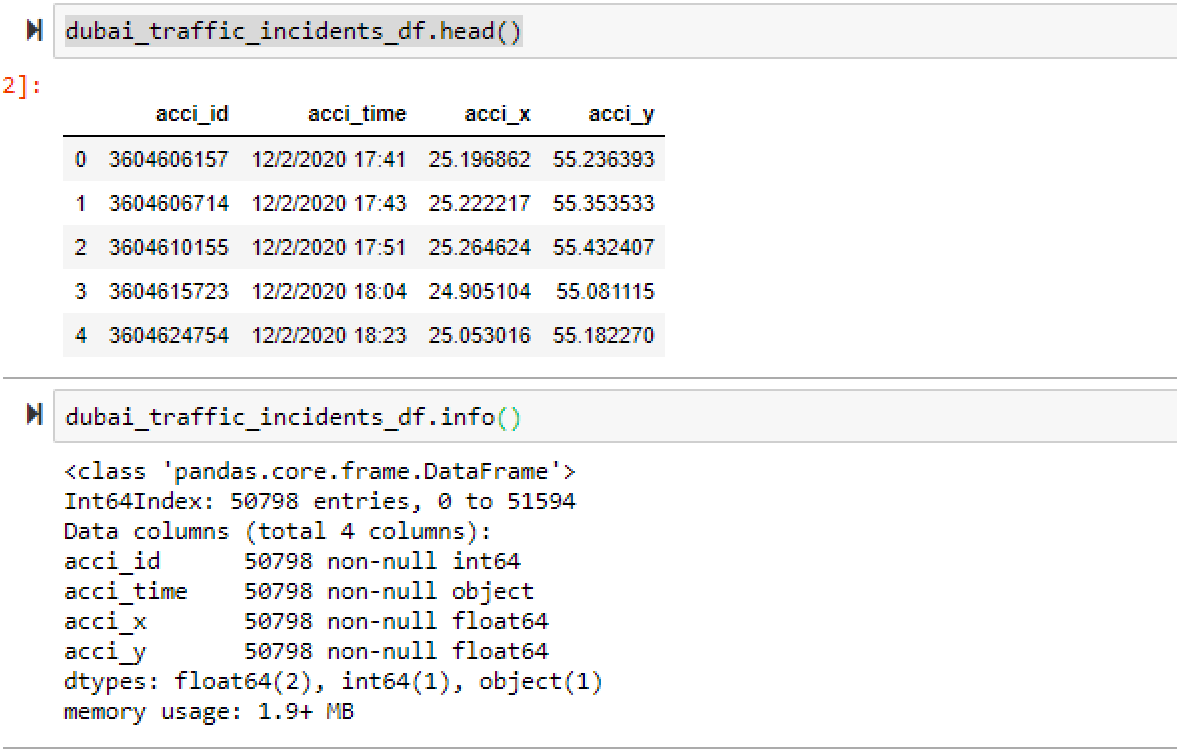
Traffic Incidents Data Set

This data set has information on accident ID, latitude and longitude of incident accident time.

It is important to note that there are certain latitude, longitude values that have been assigned as 0.0. Also there are a few latitude, longitude values that point to locations outside UAE. These are outliers and can distort the results, hence to be removed from the dataset.

Below Data Cleansing steps were carried out to clean up the data frame.

1. Convert the latitude and longitude values to float data type
2. Remove rows containing latitide/longitude values that are outside the bounding coordinates for UAE - 51.5795186705, 3. .4969475367, 56.3968473651, 26.055464179
3. Drop all rows that does not have a latitude/longitude detail.



```
dubai_traffic_incidents_df.head()
```

2]:

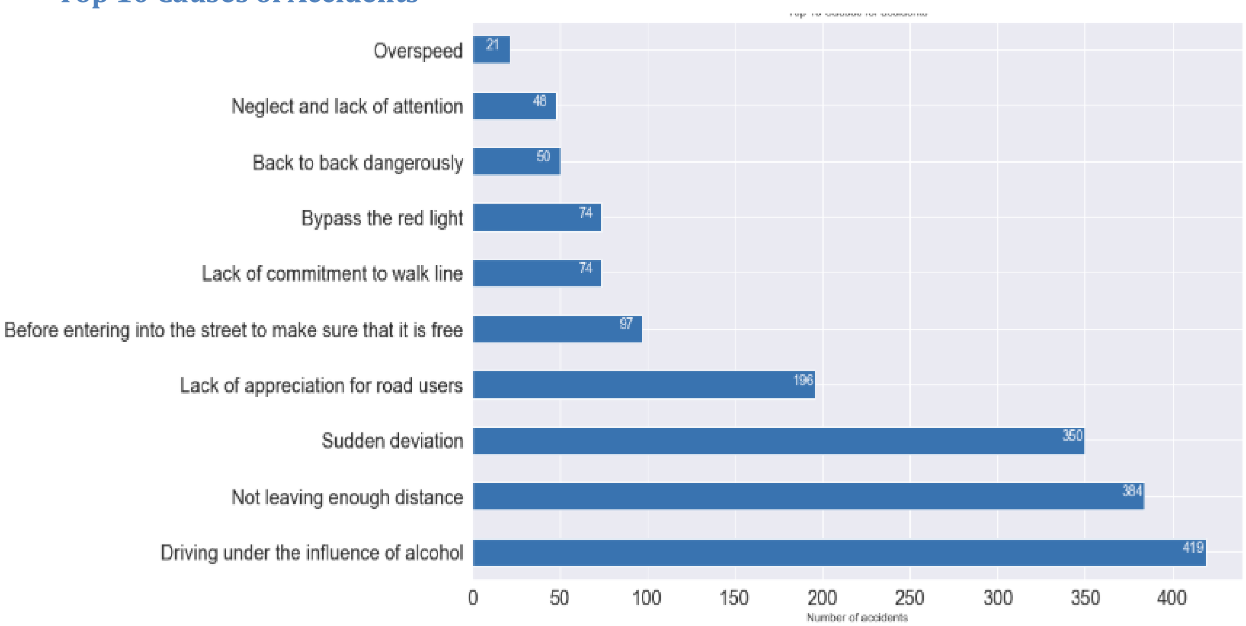
	acci_id	acci_time	acci_x	acci_y
0	3604606157	12/2/2020 17:41	25.196862	55.236393
1	3604606714	12/2/2020 17:43	25.222217	55.353533
2	3604610155	12/2/2020 17:51	25.264624	55.432407
3	3604615723	12/2/2020 18:04	24.905104	55.081115
4	3604624754	12/2/2020 18:23	25.053016	55.182270

```
dubai_traffic_incidents_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 50798 entries, 0 to 51594
Data columns (total 4 columns):
acci_id      50798 non-null int64
acci_time    50798 non-null object
acci_x       50798 non-null float64
acci_y       50798 non-null float64
dtypes: float64(2), int64(1), object(1)
memory usage: 1.9+ MB
```

Visualization, Modeling and Analysis

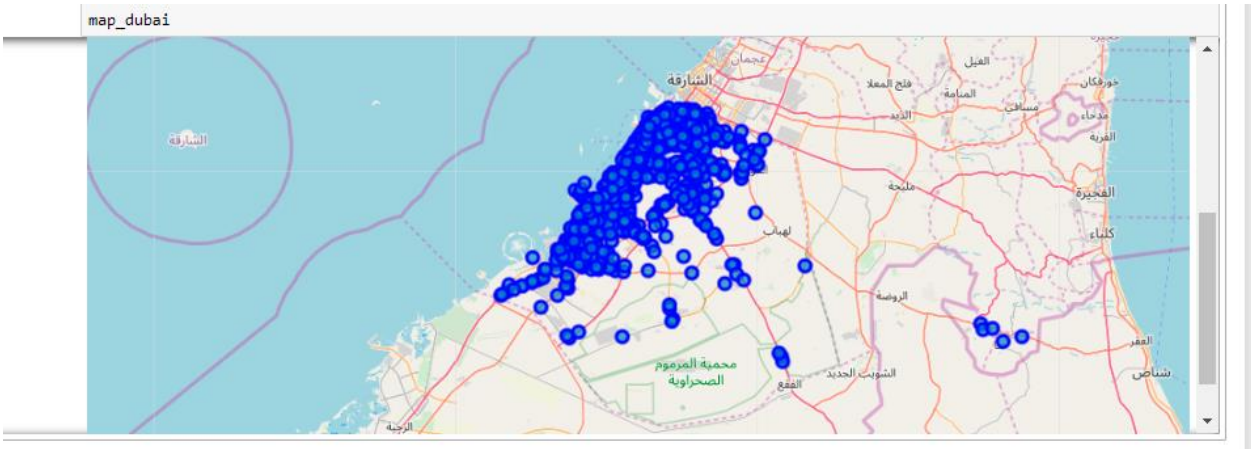
Top 10 Causes of Accidents



It can be observed that most accidents were due to driving under the influence of alcohol. The second most common reason is not leaving enough distance from the vehicle in front.

Top Locations where accidents occur

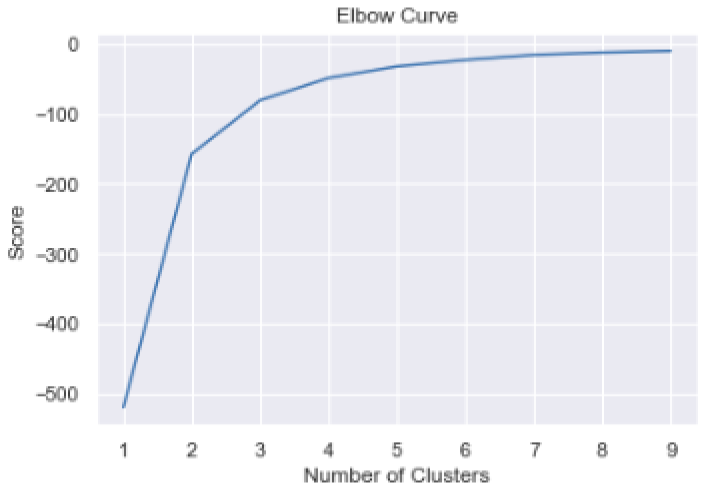
To visualize the pattern of locations where accidents occuu, the first 1000 accident locations from Traffic_incidents data set has been plotted on the map



The next step was to understand if these locations are clustered around any location. To identify this two clustering algorithms have been applies on the latitude/longitude values: K means Clustering and DBSCAN

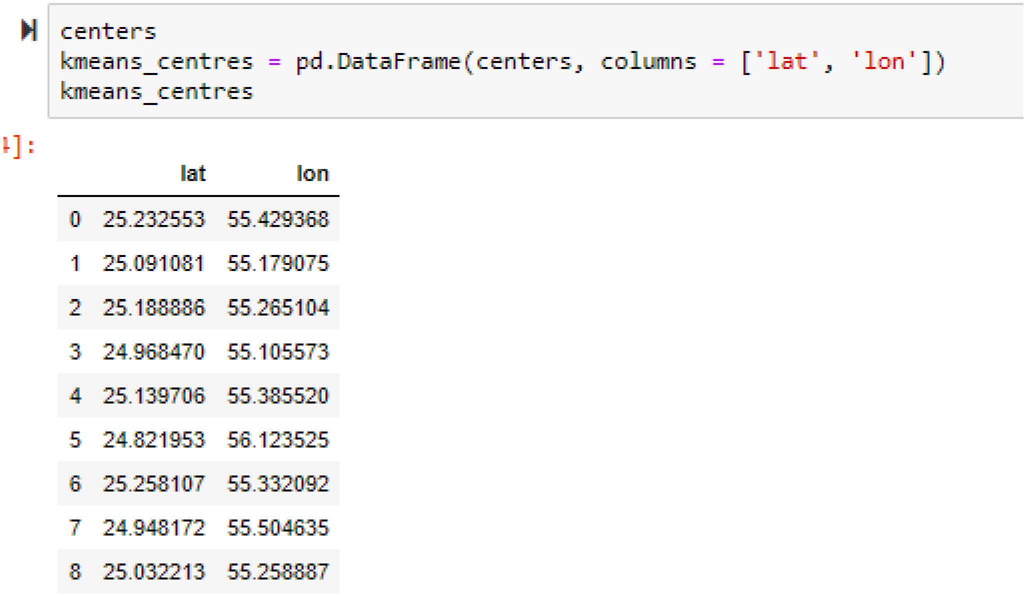
K Means Clustering on Latitude, Longitude Coordinates

To determine an optimum value for K, an elbow curve was plotted



The elbow curve flattens for values above K=9. Hence K=9 was chosen as the optimum number of clusters.

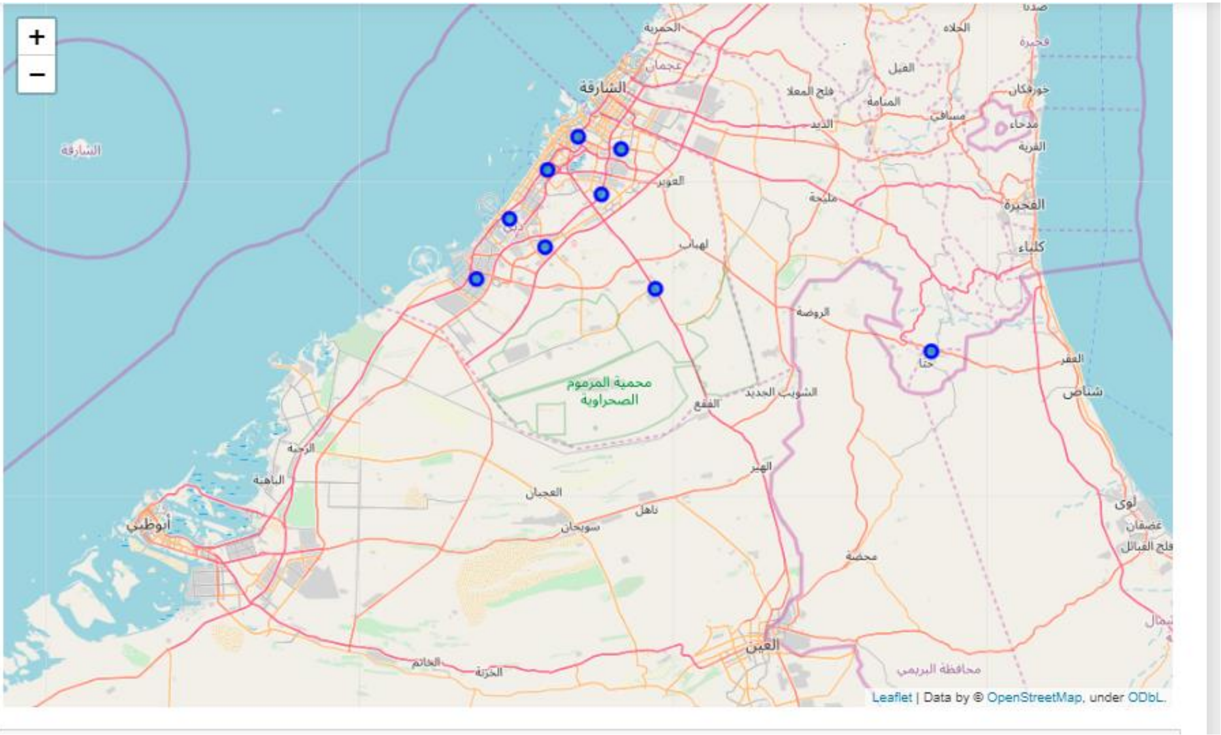
K means predicted below 9 coordinates as cluster centres. These locations are the focal points for accidents in UA as per this model.



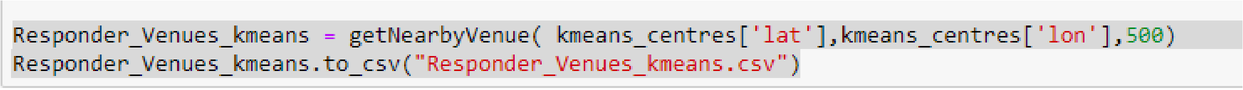
```
centers
kmeans_centres = pd.DataFrame(centers, columns = ['lat', 'lon'])
kmeans_centres
```

]:

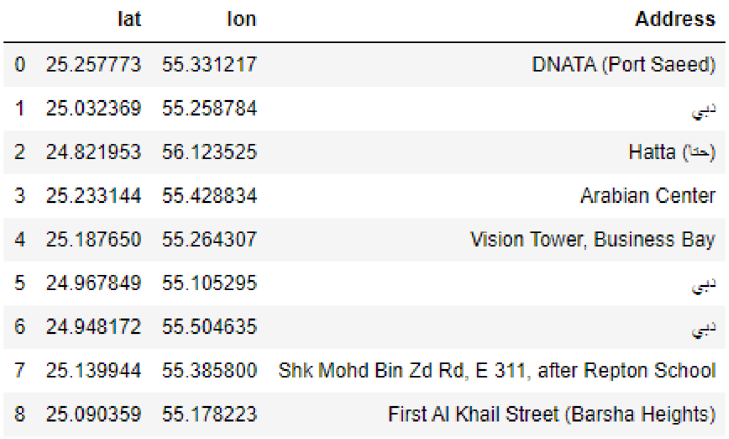
	lat	lon
0	25.232553	55.429368
1	25.091081	55.179075
2	25.188886	55.265104
3	24.968470	55.105573
4	25.139706	55.385520
5	24.821953	56.123525
6	25.258107	55.332092
7	24.948172	55.504635
8	25.032213	55.258887



Next step was to identify the nearest venue near these coordinates using Foursquare API. This information could be used by law enforcement authorities or emergency responders to centralize their operations.



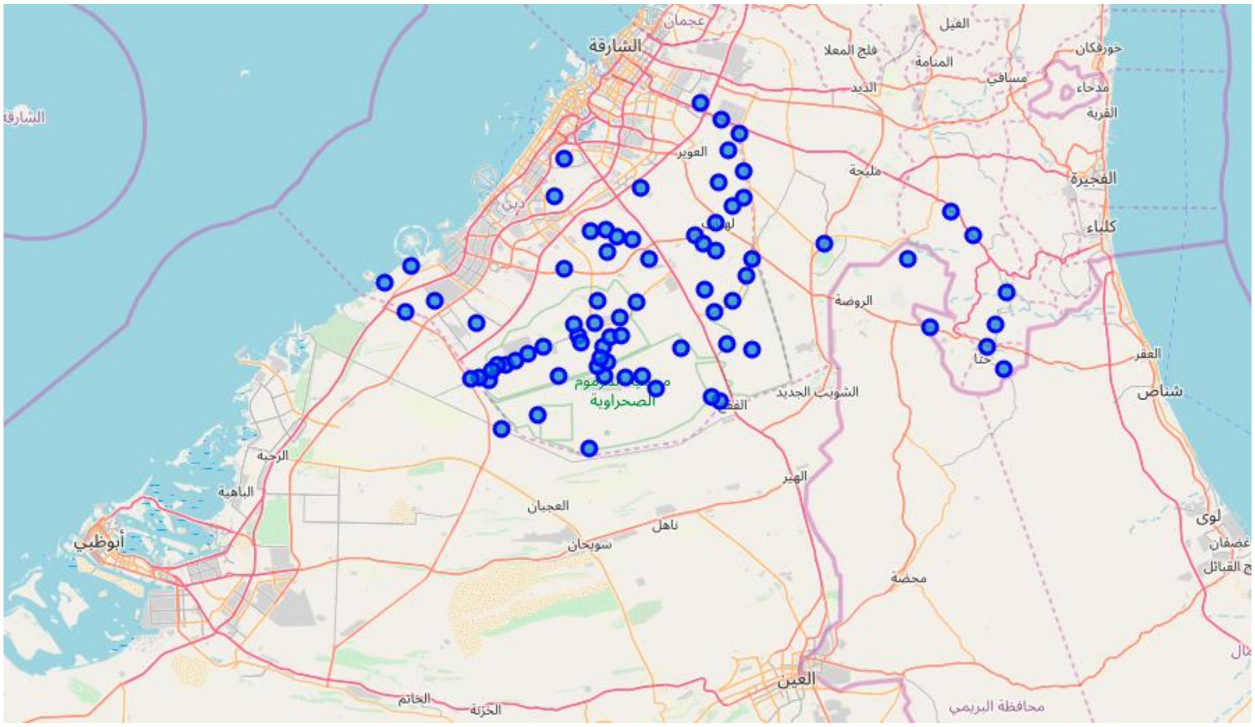
```
Responder_Venues_kmeans = getNearbyVenue( kmeans_centres['lat'],kmeans_centres['lon'],500)
Responder_Venues_kmeans.to_csv("Responder_Venues_kmeans.csv")
```



	lat	lon	Address
0	25.257773	55.331217	DNATA (Port Saeed)
1	25.032369	55.258784	دبي
2	24.821953	56.123525	Hatta (حتا)
3	25.233144	55.428834	Arabian Center
4	25.187650	55.264307	Vision Tower, Business Bay
5	24.967849	55.105295	دبي
6	24.948172	55.504635	دبي
7	25.139944	55.385800	Shk Mohd Bin Zd Rd, E 311, after Repton School
8	25.090359	55.178223	First Al Khail Street (Barsha Heights)

DBSCAN

DBSCAN is an alternate clustering algorithm and is supposed to be better performing that K means when it comes to GPS coordinates. If GPS coordinates spill over multiple zones, K means clustering can give erroneous results. DBSCAN on the other hand converts coordinated to radian and uses haversine distances to calculate centroids and can give better results. In the analysis below, DBSCAN is used to produce 77 centroid locations



Results

From this Data Analysis exercise, below observations could be made

1. A significant number of accidents are caused due to driving under the influence of alcohol. Law enforcement authorities need to perform occasional inspections for driving under the influence
2. The second major cause of accidents is tailgating or not leaving enough distance with the vehicle in front. Improving surveillance regarding this and effecting fines can be measure to curb this habit of drivers.
3. There are certain locations around which accidents are more frequent. Measures for improved monitoring and surveillance, quicker emergency response maybe set up around these locations to effect overall improvement

Discussion

One of the challenges encountered during this project was that the location data for UAE often is in Arabic and translation to English is not straightforward. Googletrans API was experimented with but often results in timeouts even with throttled traffic. A paid account is required.

The analysis can be further expanded to include the influence of gender on causes of accidents, peak times during which accidents occur and what has been the overall trend over the years.

Conclusion

This project helps provide insights to the driver behavior in UAE and patterns of accidents/incidents. Locations that are more dangerous and which require improved surveillance/monitoring has been identified and adopting measures can improve the overall experience on UAE roads