Hello, in my project I have worked on my usage logs in Google Maps between December 2015-December 2023. There were a few reasons for me to select Google Maps as the source of data. Firstly, the interval of records is wide, which is 8 years of time. Also, I believe I use Google Maps more actively than health apps or social media apps since I enjoy walking around new places. It might not be used more actively than WhatsApp or Spotify, but I did not want to use Spotify as it is used by many in CS210 project or WhatsApp because of personal privacy issues.

Obtaining data from Google Maps was fairly easy, I just had to go to the section for my data from the main page of Google Maps and select the data that I wanted to download. I pushed the data within .zip file as Records.json file, which has about 615000 logs inside of it. I already explained my workflow in the Jupyter file, but it would be nice to make a brief summary for it.

Firstly, I needed to tidy up the data in .json file as it would not be efficient to process it as a single piece of string, so I created a dataframe out of it for preprocessing and I altered that dataframe with a few new columns.

	latitudeE7	longitudeE7	accuracy	activity	source	deviceTag	timestamp	deviceDesignation	velocity	heading	...	serverTimestamp	deviceTimestamp	batteryCharging	formFactor	placeId	locationMetadata	inferredLocation	activeWifiScan	predicted_activity	activity_confidence
0	410108451	288561243	830	[{'type': 'UNKNOWN', 'confidence': 42}, {'type...	CELL	1442429258	2015-12-16T15:47:39.481Z	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	IN_VEHICLE	21
1	410096099	288579703	29	NaN	WIFI	1442429258	2015-12-16T15:48:44.961Z	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	No log held	0
2	410095633	288578685	37	[{'type': 'TILTING', 'confidence': 100}]	WIFI	1442429258	2015-12-16T15:49:45.015Z	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	TILTING	100
3	410095633	288578685	37	NaN	WIFI	1442429258	2015-12-16T15:50:45.357Z	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	No log held	0
4	410094077	288576223	58	[{'type': 'STILL', 'confidence': 95}, {'type'...	WIFI	1442429258	2015-12-16T15:51:11.995Z	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	STILL	95
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
615740	408912312	293814795	15	NaN	WIFI	-1798405216	2023-12-05T06:37:14.548Z	NaN	NaN	NaN	...	2023-12-05T06:42:27.849Z	2023-12-05T06:42:26.990Z	False	PHONE	NaN	NaN	NaN	{'accessPoints': [{'mac': '185816304111777', '...	No log held	0
615741	408905369	293807440	300	NaN	CELL	-1798405216	2023-12-05T06:38:38.207Z	NaN	NaN	NaN	...	2023-12-05T06:42:27.849Z	2023-12-05T06:42:26.990Z	False	PHONE	NaN	NaN	NaN	NaN	No log held	0
615742	408912265	293814728	16	NaN	WIFI	-1798405216	2023-12-05T06:41:41.229Z	NaN	NaN	NaN	...	2023-12-05T06:42:27.849Z	2023-12-05T06:42:26.990Z	False	PHONE	NaN	NaN	NaN	{'accessPoints': [{'mac': '185816304111776', '...	No log held	0
615743	408912394	293814858	63	NaN	WIFI	-1798405216	2023-12-05T06:42:10.931Z	NaN	NaN	NaN	...	2023-12-05T06:42:27.849Z	2023-12-05T06:42:26.990Z	False	PHONE	NaN	NaN	NaN	NaN	No log held	0
615744	408912255	293814892	15	NaN	WIFI	-1798405216	2023-12-05T06:46:25.647Z	NaN	NaN	NaN	...	2023-12-05T06:46:25.985Z	2023-12-05T06:46:25.733Z	False	PHONE	NaN	{'wifiScan': {'accessPoints': [{'mac': '18581...	{'timestamp': '2023-12-05T06:42:28.316Z', 'la...	{'accessPoints': [{'mac': '185816304111776', '...	No log held	0

615745 rows × 24 columns

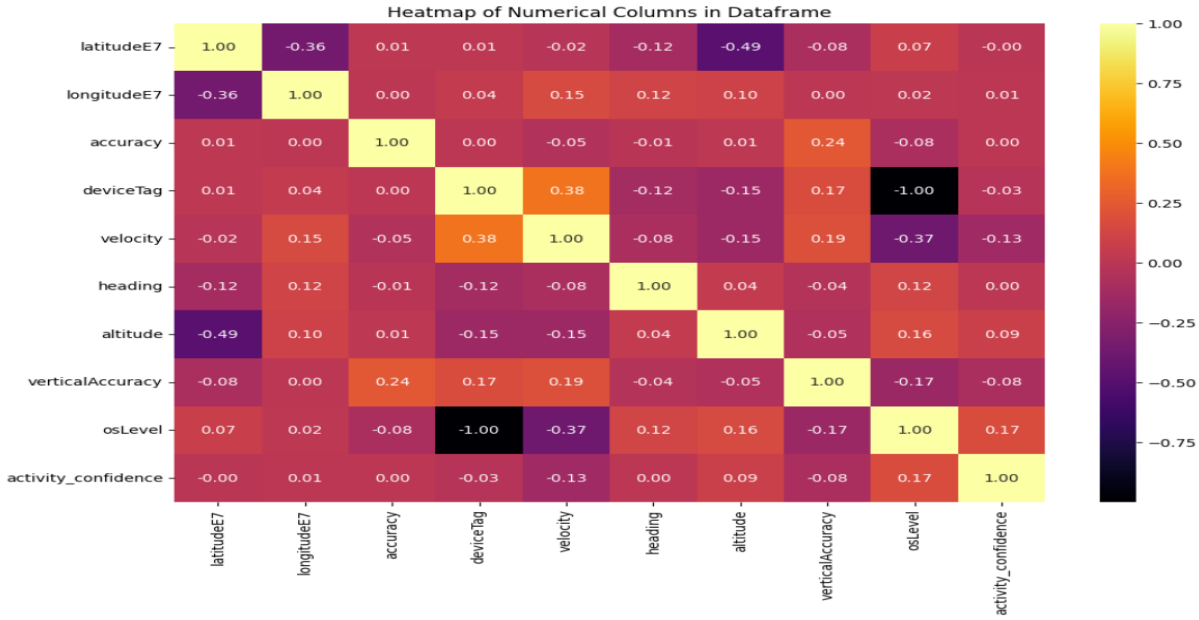
Then I inspected some columns to reach some conclusions on my usage trends and on the missing parts in dataframe as exploratory data analysis. While doing all these things, I noted my null hypotheses inside the notebook file so that I could change them when needed.

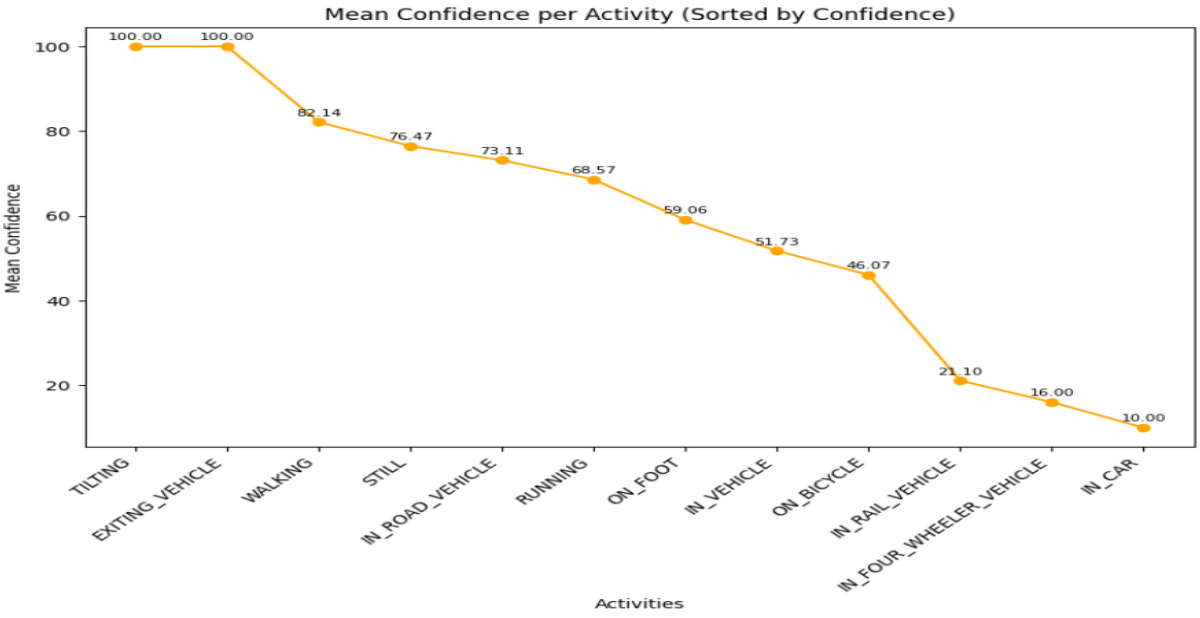
```
Usage Values by Years:

2015 :  656

2016 :  4912

2017 :  55699

2018 :  185519

2019 :  318834

2020 :  41133

2021 :  1648

2022 :  4077

2023 :  3267
```

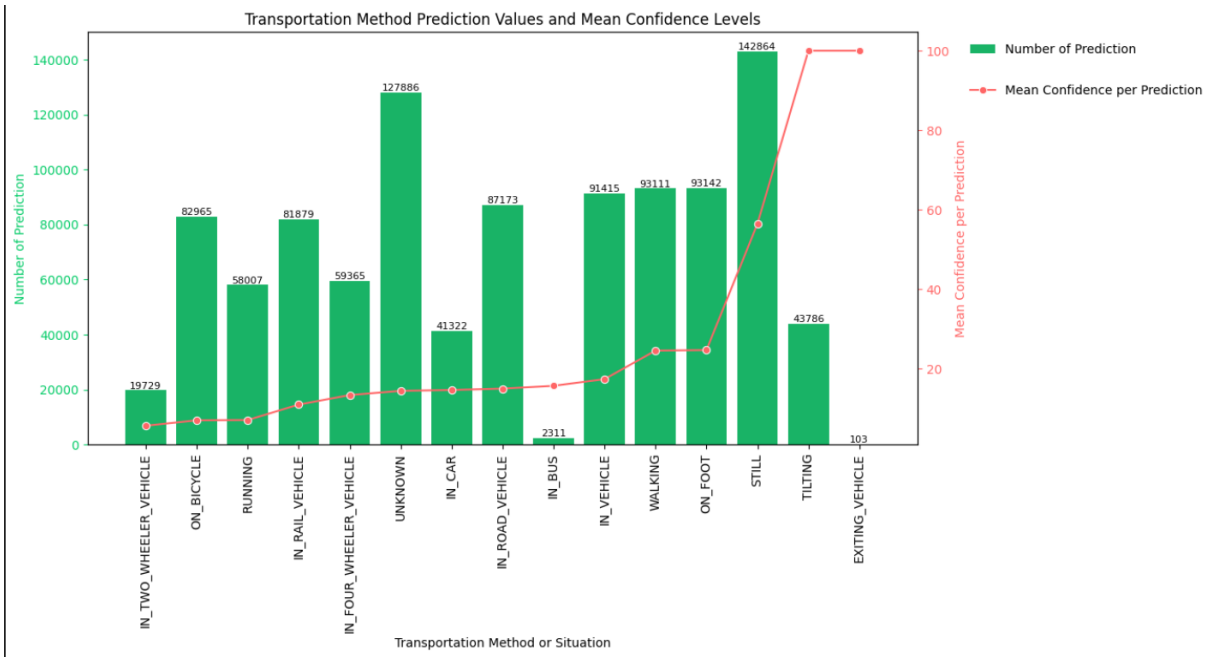
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 615745 entries, 0 to 615744
Data columns (total 24 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   latitudeE7            615745 non-null  int64
 1   longitudeE7           615745 non-null  int64
 2   accuracy              615745 non-null  int64
 3   activity              196467 non-null  object
 4   source                615745 non-null  object
 5   deviceTag             615745 non-null  int64
 6   timestamp             615745 non-null  object
 7   deviceDesignation     257407 non-null  object
 8   velocity              1663 non-null    float64
 9   heading               1294 non-null    float64
 10  altitude              3708 non-null    float64
 11  verticalAccuracy      3708 non-null    float64
 12  platformType          3913 non-null    object
 13  osLevel               3913 non-null    float64
 14  serverTimestamp       3913 non-null    object
 15  deviceTimestamp       3913 non-null    object
 16  batteryCharging       3205 non-null    object
 17  formFactor            3913 non-null    object
 18  placeId               191 non-null     object
 19  locationMetadata      1830 non-null    object
...
 22  predicted_activity    615745 non-null  object
 23  activity_confidence   615745 non-null  int64
```

```
Number of Usages in Istanbul: 549962

Number of Usages in other cities of Turkey: 64037

Number of Usages outside of Turkey: 1746

Ratio of Number of Usage in Istanbul over Number of Usage in whole Turkey: 0.8957050418648891

You have been using Google Maps in Istanbul more often than you do outside of it.
```

At the next part, I used data visualization to discover and show patterns in my data, such as regression, correlation, and causation. Some of the charts can be seen below.







Finally, I trained a model to fill the missing values for a column with predictions based on the known values in that column. Please check the notebook file to see more on these steps.

```
The accuracy of the model is 0.5153837849508799
The value counts of the predictions for the unknown sources are:
WIFI    6478
CELL    2263
GPS      522
Name: count, dtype: int64
```

In my project, I learned that I spent so much time outside in 2018 and 2019. Indeed, when I think about it, those 2 years might be the most relaxed times of my life. The pressure of university exam did not kick in yet and there was no pandemic in Turkey. Additionally, I went abroad for the first time in

my life in 2018 to Western Europe countries, and the results of data showed themselves accordingly. Another finding was on my monthly usage trends. I was surprised when I found out that I used Google Maps mostly in November, December, and October, since my claim was that those months would be the months of summer or close to summer season. And a different finding was that about 90% of my usages were in Istanbul. It was not shocking to see the usages in Istanbul to have the majority, but the ratio is far greater than what I expected. These are some of my findings and if you want to know more about my conclusions, please see the notebook file.

In my opinion, I could improve the accuracy of prediction model that I trained. Its accuracy is slightly higher than 50% and it would be nice to work on that in the future with more logs held in some of the metrics. Moreover, I plan to increase the level of optimization for my methods as I struggled at some parts in terms of memory and time. I believe these will be easier as I move on in my courses.

Thanks for reading. Hope you enjoyed it!

Arda Barış Tonbil 31130