Insight Report of GPS Data Analysis for Individual Dwells

1. Preprocessing Data

The dataset comprises 1,770,011 rows and covers a broad geographical range, with latitude ranging from 51.482489 to 51.506643 and longitude from -0.121340 to -0.076408. The temporal span extends from January 8, 2018, to January 19, 2018. To manage the large dataset efficiently, I employed batch processing. After cleaning the data, I utilized the 'identify_dwells' function to filter and extract relevant insights. After applying the 'identify_dwells' function to the dataset, I obtained a refined set comprising 43,095 rows of data, which encapsulates the significant locations where users dwelled for extended periods (minimum of duration minutes is 45, and the distance threshold is 30).

	user_id	start_time	end_time	lat	lon
0	0006AACD-AB48-46A6-BC0C-8496A903DCD9	2018-01-12 15:02:32+00:00	2018-01-12 15:32:15+00:00	51.497524	-0.080989
1	0007744c-5c2a-472a-8de5-aabe90ec9fe5	2018-01-12 11:05:28+00:00	2018-01-17 14:01:12+00:00	51.487643	-0.111270
2	0007744c-5c2a-472a-8de5-aabe90ec9fe5	2018-01-17 14:31:21+00:00	2018-01-17 14:46:47+00:00	51.506172	-0.117457
3	0007744c-5c2a-472a-8de5-aabe90ec9fe5	2018-01-17 22:56:58+00:00	2018-01-18 12:40:41+00:00	51.487665	-0.111216
4	0007744c-5c2a-472a-8de5-aabe90ec9fe5	2018-01-18 13:27:01+00:00	2018-01-18 14:28:17+00:00	51.487682	-0.111060
...	...	...	...	...	...
43090	ffed1702-1a50-4760-a6a9-f9f43f7b2b52	2018-01-08 09:56:23+00:00	2018-01-12 10:10:54+00:00	51.485985	-0.120644
43091	ffed1702-1a50-4760-a6a9-f9f43f7b2b52	2018-01-12 10:15:53+00:00	2018-01-12 10:23:53+00:00	51.486268	-0.120735
43092	fffd15f4-bbd5-4bac-9d9f-a9568bb701fc	2018-01-19 07:18:14+00:00	2018-01-19 08:55:34+00:00	51.495754	-0.078319
43093	fffd15f4-bbd5-4bac-9d9f-a9568bb701fc	2018-01-19 09:29:45+00:00	2018-01-19 09:39:52+00:00	51.502507	-0.099963
43094	ff2bc06a-df27-48f8-9177-3a36ef4a8887	2018-01-18 10:11:45+00:00	2018-01-18 10:12:26+00:00	51.504208	-0.114535

2. Analysis

After processing the dataset, I conducted several analyses:

a. Map Plotting and Heatmap: Utilizing the start and end times along with location data, I generated both map plots and heatmaps to visually represent the distribution of dwell locations and activity intensity across the geographical area.

b. Activity Analysis: I examined the frequency and distribution of user activities within the dataset to identify patterns and hotspots of activity.

c. User Activity Analysis: I conducted a detailed analysis of individual user activities to understand their movement patterns, frequent dwell locations, and overall activity behavior.

d. Average Dwells of All Users: By aggregating dwell data across all users, I calculated the average dwell duration to gauge the typical length of stays at various locations.

3. Result

a. Map Plotting and Heatmap

After plotting the dwell locations using a scatter plot and overlaying them onto a heatmap in a real-time map format (accessible via the 'dwell_locations_heatmap2.html'), we observed areas of high concentration. These areas likely represent locations where users spent a significant amount of time, indicating potential points of interest or frequent destinations within the dataset.

Figure 1 : Scatter Plot of all dwells

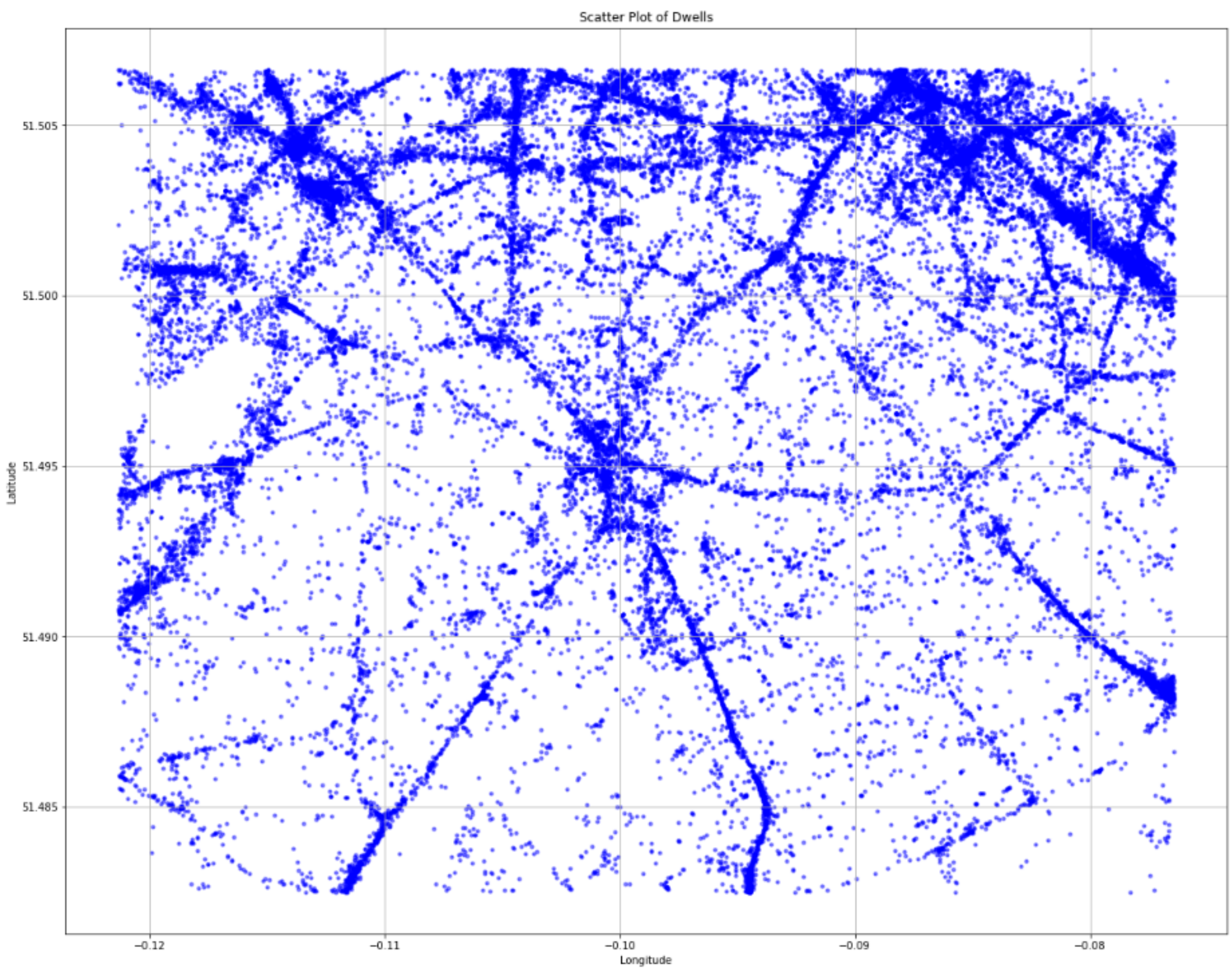
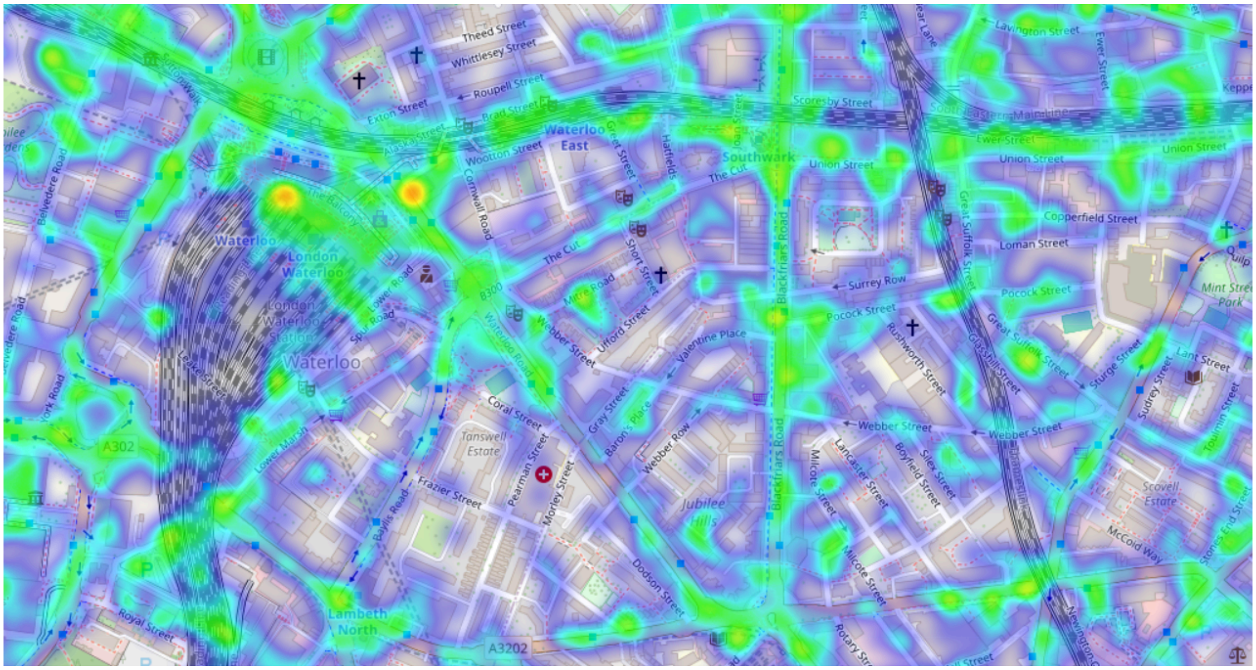


Figure 2 : A heatmap representation indicating areas of high concentration



Following the analysis of the heatmap, the areas exhibiting the highest concentration of dwell activity are identified as follows:

- Waterloo Station
- Capital Tower of London
- Tooley Street and London Bridge Hospital
- Guy's Hospital
- Sainsbury New Kent Road
- Astral Street
- Perspective Building
- Tower Bridge Road
- Potter Field Park and Tower Bridge Plaza
- Elephant Park

b. Activity Analysis

For Activity Analysis, data visualization techniques were employed to discern usage patterns among dwell users.

Figure 3 : Hourly Distribution of Dwells

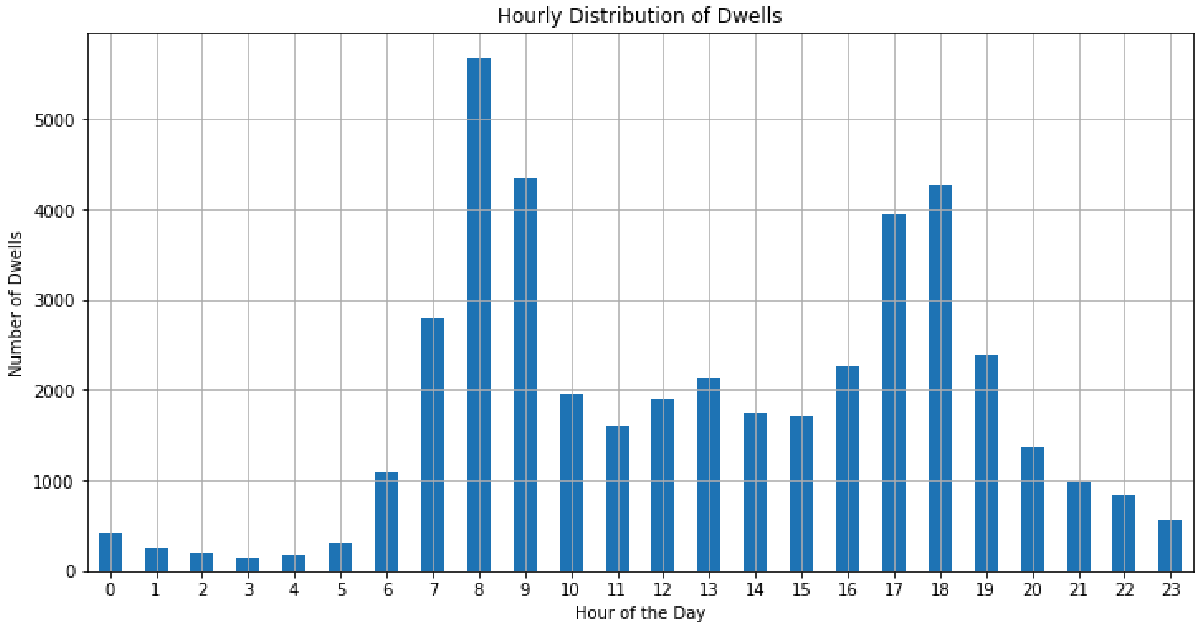
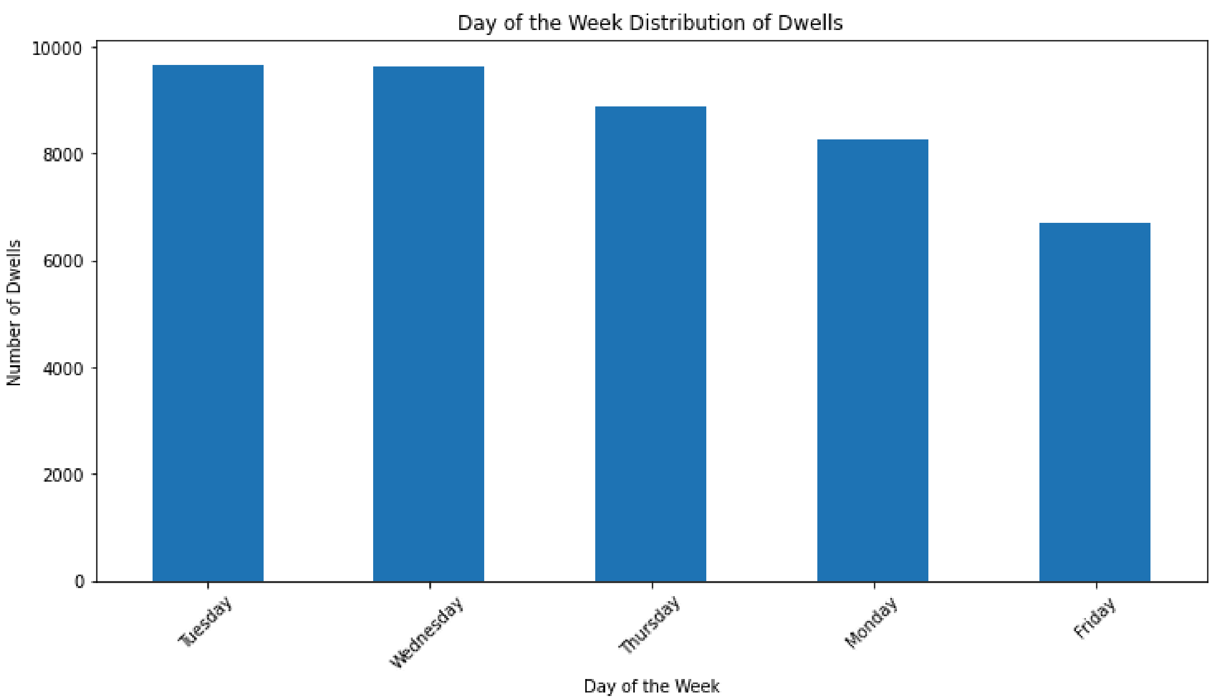


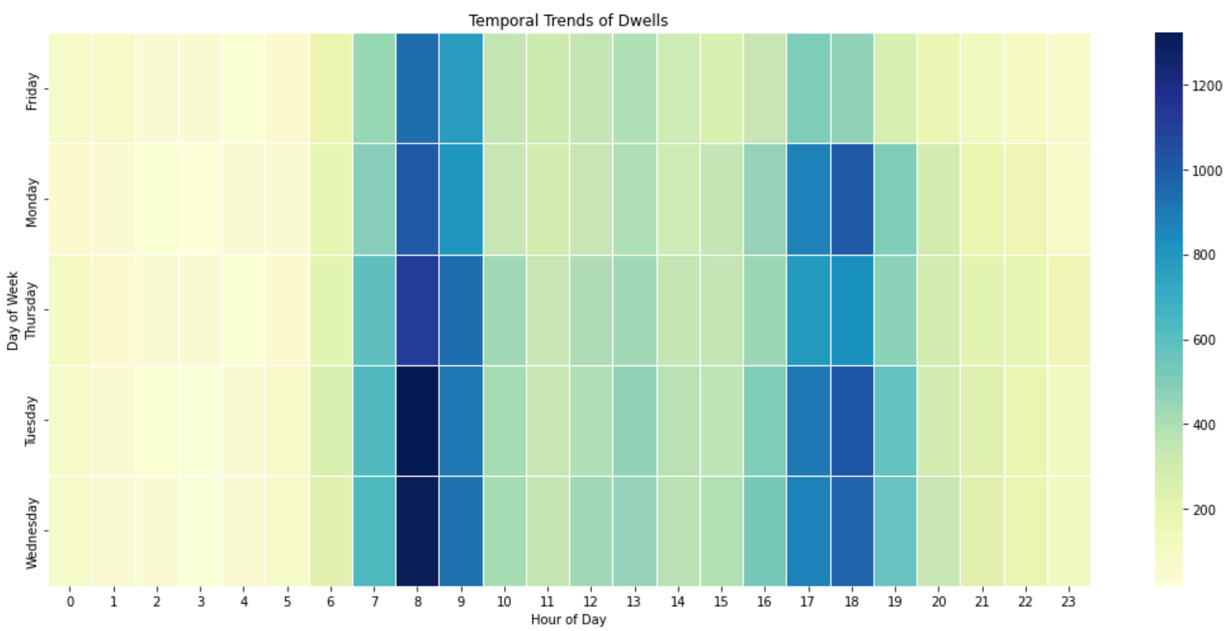
Figure 4 : Day of the Week Distribution



The lowest frequency of dwells occurs during the early morning hours from midnight to around 5 AM. This is typical as these hours generally correspond with the time most people are at home and inactive in terms of traveling. A significant rise starts at 7 AM and 8 AM, suggesting the beginning of the daily commute as people start their day heading to work, school, or other activities. There is a noticeable peak at 12 PM, representing the lunch hour. There is also a peak at 6 PM, which likely corresponds to the end of the standard workday, resulting in a high volume of people dwelling as they leave work, visit shops, attend social events, or engage in evening activities.

The second graph shows the number of dwells for different days of the week, from Tuesday to Friday. The number of dwells is relatively consistent from Tuesday to Thursday, with a slight decrease on Thursday. However, there is a notable drop in the number of dwells on Friday. This trend might suggest a change in routine towards the end of the week, possibly as people prepare for the weekend or have different schedules.

Figure 5 : Temporal Trends of Dwells



The 'Temporal Trends of Dwells' reveals the frequency of dwells across various times and days. Weekday mornings, especially on Tuesday and Wednesday, exhibit a high concentration of dwells around 8 AM, indicative of rush hour patterns. A consistent midday peak suggests common lunchtime activities, while Friday evenings show extended activity, likely reflecting leisure or social engagements post-work

c. User Activity Analysis

For user activity analysis, I count the occurrences of dwell instances for each user. Extracted the frequency of dwellings per user and put it in the table to track user activity patterns efficiently.

Figure 6 : High-activity users

	user_id	dwells count
0	DEF1AC85-350F-4C8A-872F-8BE67ED30D73	53
1	BB03FFB0-B6BA-44CE-B382-B4CDF944ACEC	39
2	61D56F4C-5F18-473E-A0A8-160C6FDFBD3E	33
3	308B68F1-137B-4A47-9EC2-9F3BEBC94D7F	31
4	CA08BCB6-DA9C-4C47-A24B-54C9E1562455	31
5	AEEF074C-E974-4918-8724-F1059153FC97	30
6	57e5b56c-18bc-4f50-a66b-f9f625ea3b64	29
7	21b6858f-3438-4052-8cb5-131fd7833442	29
8	E112CB96-32DC-4A86-89E3-E17D7B708344	28
9	784616B6-9F69-422C-8221-51E6D3698FCD	28
10	435771D7-6329-4664-8C4A-C1366A1CDDB0	27

These values represent the number of times they were recorded as dwelling in a location. This frequency can be used to identify users with high activity within the areas. These users might be of particular interest for network usage studies, targeted marketing campaigns, or other customer engagement strategies. The user's characteristic can also be used as a benchmark about how the potential candidate will be a target market.

d. Average Dwells of All Users

Figure 6 : Dwells Duration Calculation

```
# Calculate the duration of each dwell in minutes
dwells['dwell_duration'] = (dwells['end_time'] - dwells['start_time']).dt.total_seconds() / 60

# Calculate the average dwell time
average_dwell_time = dwells['dwell_duration'].mean()

print("Average Dwell Time:", average_dwell_time, "minutes")
print("Average Dwell Time:", average_dwell_time / 60 , "hour")
```

```
Average Dwell Time: 1502.0837266139206 minutes
Average Dwell Time: 25.034728776898678 hour
```

```
# Calculate the median dwell time
mode_dwell_time = dwells['dwell_duration'].mode()
mode_dwell_time_value = mode_dwell_time.iloc[0]  # Extract the mode value
print("Mode Dwell Time:", mode_dwell_time_value, "minutes")
```

```
Mode Dwell Time: 51.93333333333333 minutes
```

In analyzing the dwell times of all users, I opted to utilize the mode instead of the mean to capture the most frequently occurring dwell duration within the dataset. This is due to mode dwell time representing the most common duration observed, which is around 51.93 minutes. This choice offers insight into the prevalent patterns of activity rather than the overall average, which might be skewed by outliers or less common durations.

4. Shortcomings of the Data

This data provides valuable insights into user behavior during the available time periods. The existing data offers a foundation for understanding spatial and temporal patterns.

As for shortcomings of the data, the dataset exhibits incompleteness, primarily due to its failure to cover a full 24-hour period. This limitation results in distorted metrics. For instance, if the last recorded data point is at 15:00 on a Tuesday, any subsequent data not available until 06:00 on Wednesday would erroneously register as a single prolonged dwell.

On the other hand, adding the dataset with supplementary data sources could significantly enhance the analysis. For example, If demographic data is available, we can correlate dwell patterns with demographic segments to understand which population groups are spending time in certain areas.