Twitter and Circadian Rhythms

December 3, 2021

1 Abstract

Circadian rhythm helps to guide our body to let it know when to sleep and when to wake up. It's crucial in helping us regain energy lost from being awake and performing daily activities. Hence better understanding of circadian cycle would help us to maintain energy levels. We are trying to predict the sleep patterns and which features play a significant role in affecting the circadian cycles based mainly on twitter data.

2 Introduction

Our goal is to perform analysis on the sleep patterns of the different categories of people. We will use different features like Job Title, Timezone, Seasons and perform analysis on why the sleep patterns are differing w.r.t to those features. At the end, we will build a classification model to predict which users falls under which category of the sleep pattern.

3 Data Collection

For this progress report, we analyzed the 80 Million Rows data set given by the professor. For extracting latitude, longitude, we used an external API called Geopy. For extracting weather information of a particular place at a particular time, we used the World Weather Online API.

4 Data Preprocessing

4.1 Limitations

1. As the data-set was very huge, i.e. 80 million rows, it was very daunting task to process the whole data set in a sequential manner.
2. 70% of the all users have less than 1000 tweets over a year in the given data-set. So, this would cause an unnecessary fluctuations in our analysis.
3. Almost all the API's available for extracting location and weather information are paid and very expensive API's.
4. The address column which was given was very bad and it was difficult to parse.

4.2 Steps Taken

1. We have done multiprocessing in order to make the data-processing faster as it was taking around 20 hours to process around 5 million rows. This technique has helped us to reduce the time by 15 hours where we were able to focus on the analysis more. Below are the features which are present in the data set.
2. We have extracted only those users who have more than 1000 tweets over the entire given data set. This has reduced the data set drastically which resulted us to focus more time on analysis.
3. All the API's had free tire something like 15,000 calls. So, we ran multiple iterations in a batches of 15,000 and by using parallel processing techniques, we were able to utilize this free tier in an effecient way.

4. We have to find latitude and longitude from the given abstract data. The location column given in the dataset has a lot of bad values. We tried to extract proper location using Scapy library, but it only gives out entities such as if its a country or not, and hence it doesn't satisfy our requirement. So, we used another library called Geopy [GPY]. Using Geopy we processed the given location string into a proper address and extracted the exact latitude and longitude of the determined location address.

So, below are the original columns present in our initial dataset.

1. **User Hash** - Hash of the Twitter User.
2. **Characteristic** - We found 18 unique characteristics of people namely - attorney, bartender, designer, developer, doctor, engineer, firefighter, gamer, lawyer, manager, nurse, officer, professor, retired, streamer, teacher.
3. **UTC Timestamp** - UTC EPOCH
4. **Location** - Location from which a tweet has been tweeted.

From the above columns, We have done some initial data pre-processing and below are the additional columns which we extracted.

1. Latitude and Longitude - Extracted using Geopy. Using Geopy we processed the given location string into a proper string and we extracted the exact latitude and longitude of the given location string.

2. Finding Local Time - We extracted the local time using the below steps:

1. Finding UTC time from the EPOCH time.
2. From latitude and longitude, we identified the timezone.
3. Using the timezone information, we added the UTC offset value to the UTC time.

This resulted in the exact local time of a given location.

3. Weather - We extracted the weather of the particular place from where the user is tweeting using the World Weather Online API [WWO]. Weather information consists the below colums

1. Temperature at a particular time at a particular place.
2. Min & Max Temperature of that particular day.
3. Some other important features such as Total snow in cms, sunrise time, sunset time, moon-rise, moonset time, dew point in C, Feels like In C, Sun Index, UV Index, Heat Index, Wind Chill, Visibility, Humidity, Windspeed.

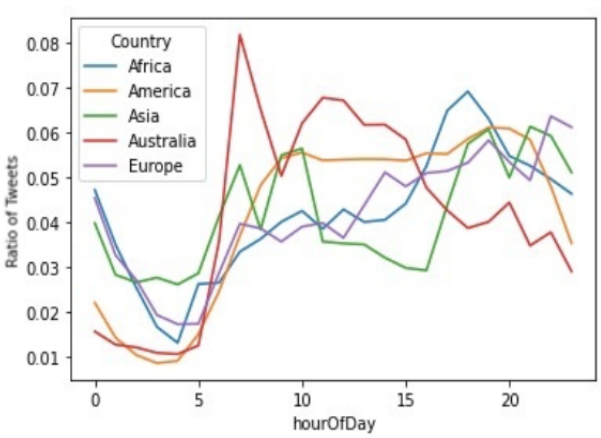
We used database merging techniques like Left-Join to merge the given data-set with the above extracted values by using Date, Time and Location as primary keys.

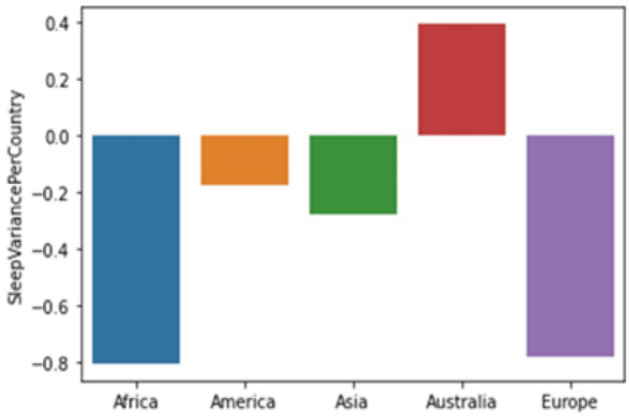
5 Areas of Analysis

5.1 Time Zone Effects

In the timezone effects column given in the document, a hypothesis was given which says that people on the east end of a time zone wake up earlier than the west. We explored this question in the below way: After mapping local time to each and every location, we mapped the ratio of tweets (number of tweets at that particular hour divided by total tweet on that day).

As we can see from the below graph in the left, the people living in east end countries like Asia, Australia have a jump in tweet percentage from 5 to 8 AM, where as the increase in tweet percentage for America and Europe started something around 9:00 AM. This supports the given hypothesis that people living in the eastern end wake up earlier than the people in the west.

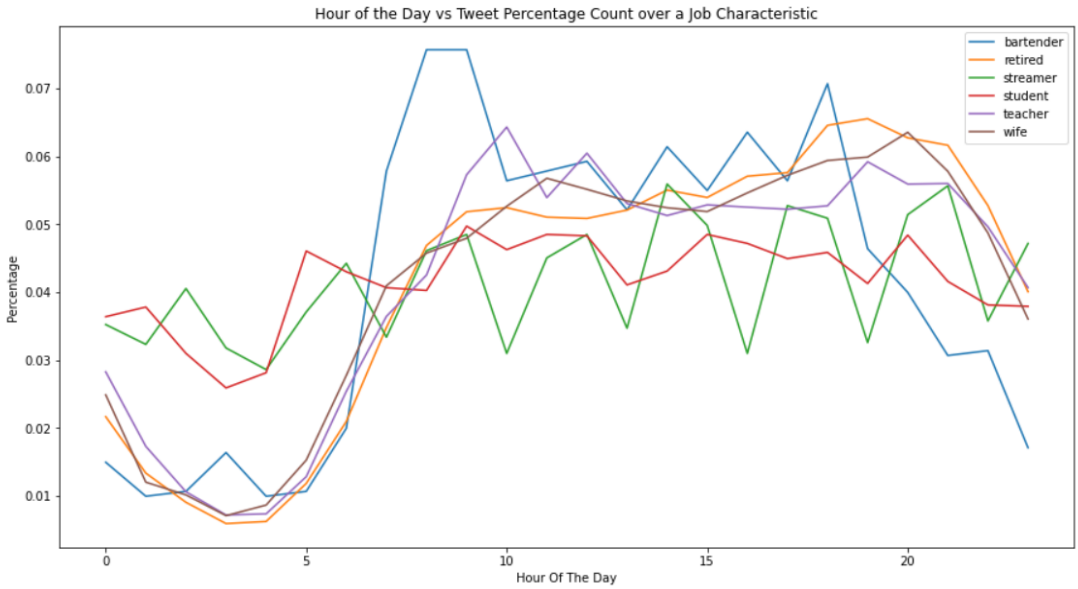




We have identified twitter inactive timings of every user on the days that he tweeted by calculating the difference between the time in first tweet and last tweet on the previous day. As we know that the inactive time in twitter is proportional to the sleep time. We considered 8 hours as the average sleep time, we calculated the Sleep timing variance from 8 hours standard sleep time for different time zones. Interestingly, (in the right graph) we can see that Africa and European countries are lacking a healthy sleep timing by a huge deficit; where as Australia is maintaining a healthy sleep timing pattern (which is evident from the left graph that there is a decrease in tweets after 15:00 hours). So, from the above observations, we conclude that time zones are having a huge impact on the people's circadian cycles.

5.2 Job Title Effects

There are around 26 different job characteristics in our dataset: 'artist', 'attorney', 'bartender', 'consultant', 'dad', 'designer', 'developer', 'doctor', 'emt', 'engineer', 'father', 'firefighter', 'gamer', 'husband', 'lawyer', 'manager', 'mom', 'mother', 'nurse', 'officer', 'professor', 'retired', 'streamer', 'student', 'teacher', 'wife'. We conducted KS pairwise two sample tests with each and every characteristic and found some distributions which are same. For demonstration, we choose 6 categories - bartender, retired, streamer, student, teacher, wife and plotted percentage of their tweet distribution over hour of the day.



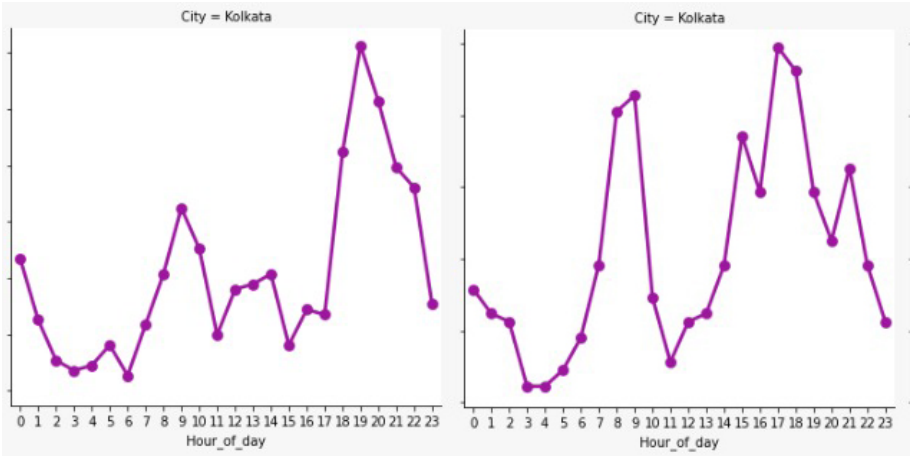
From the above graph, we can see that distribution of tweets for bartender and streamer comes under different distributions. We performed KS test for that and we got the below values:

Retired and Wife - KSsampResult(statistic=0.125, pvalue=0.994161229482218) Since the statistic value is closer to 0, the retired and wife doesn't have significant difference and have same distribution. The tweet activity of retired people is decreasing after 20:00 hours. It is also obvious that retired people sleep early due to their old age. So as the house wifes, their circadian cycles match with those of retired people.
Bartender and Streamer - KSsampResult(statistic=0.41666, pvalue=0.029913567122680163) - Since the statistic value is a little far from zero, the bartender and streamer come from different distribution. The tweet activity of bartender is decreased from 17:00 hours because their usual working times is 18:00 hours to 2:00 hours.

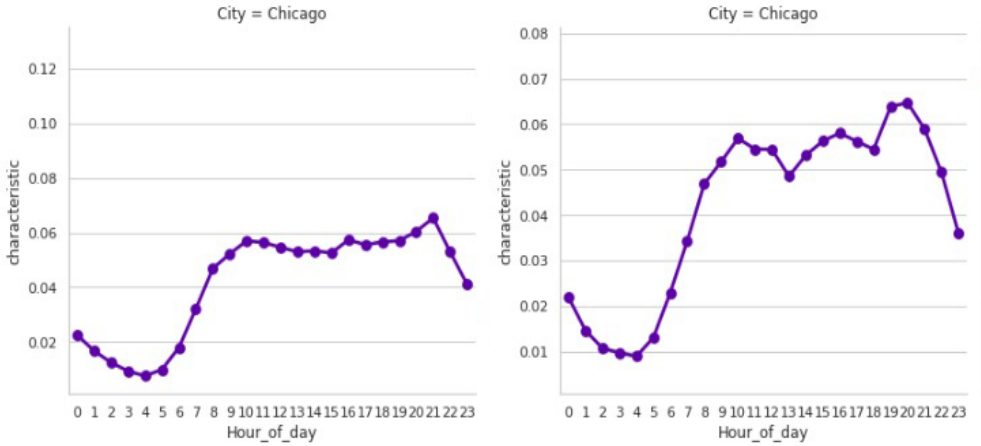
Conclusion: It is true that a persons professions affects their circadian cycle and it is also very evident from our analysis.Student and Streamer are active constantly through out the day irrespective of the time and its obvious because students tend to use more social media and have irregular sleep patterns. Also same case with the streamers as compared to those of professions like retired, wife, and teacher.

5.3 Seasonal Effects

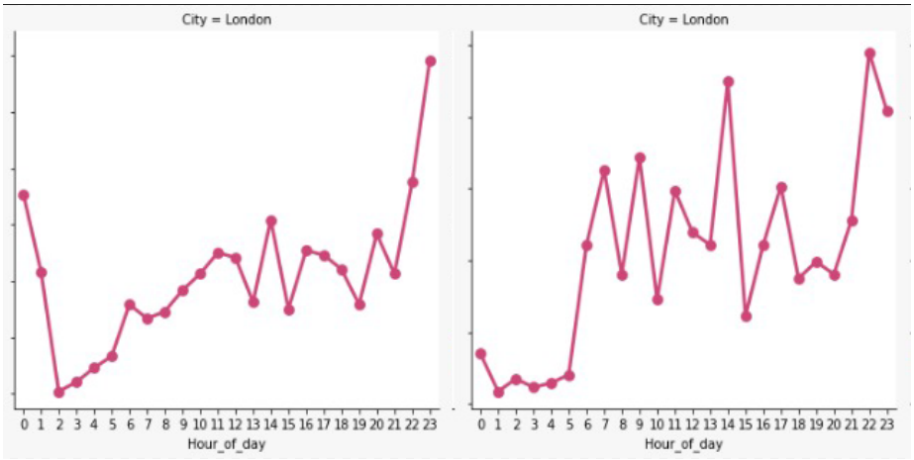
Circadian cycles change with a gradual change in temperature. The countries which are closer to the equator have very less change in their circadian cycle with respect to change in seasons as compared to the countries which are located farther from equator. They have a conspicuous change in cycles with respect to the seasons. To demonstrate the analysis, we choose 3 different places namely Kolkata(closer to equator), Chicago (a little farther from equator) and London (farthest from equator). We took June-Aug as Summer season and December - February as Winter season and plotted their Tweet ratio per hour. The left graph is of summer season and right graph is of winter season.



As expected, we can see that there isn't much fluctuation in the twitter activity during Summer(left) and Winter(right) season for Kolkata. This is true because India is a tropical country and the change in seasons isn't that harsh as compared to other countries which are far away from equator.



From the above graphs left(Summer) and right(Winter) for Chicago, it is evident that people sleep early during winter as the decrease in twitter activity started from 20:00 hours in winter where as the decrease in tweet activity during Summer started from 21:00 hours.

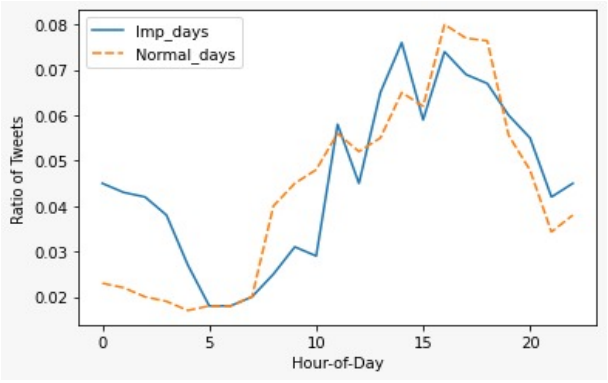


As we can see from the left(Summer) graph and right(Winter) graph for London, people tend to sleep early in winter which is evidently visible with the quick decrease in tweet activity after 22:00 hours; also people wake up early in winter as there is less daylight. Where as, tweet activity in summer gradually decreased after 00:00 hours and the jump in wake up time is also something after 7:00 hours.

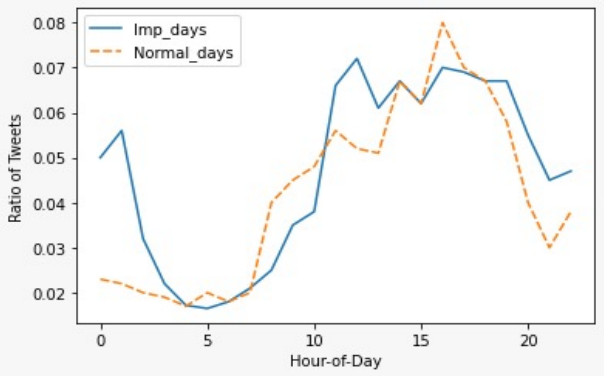
Conclusion: This analysis proves our hypothesis that Distance of a place from equator and Seasons have a correlation with the change in circadian cycles. This analysis gives us two important features(latitude-longitude, Weather) to predict circadian cycles.

5.4 Important Days Effects

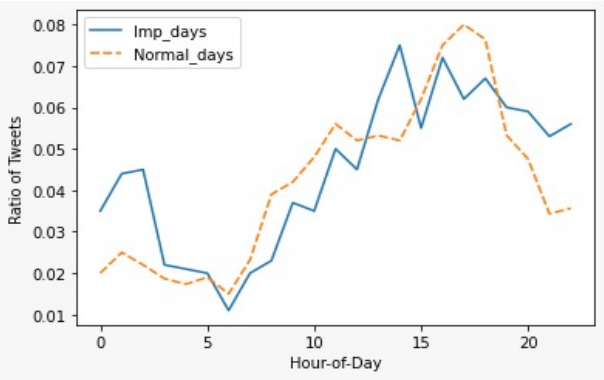
These days we get to know about the news first in social media and then in news channels, that is the rate at which important events travel through the media, and twitter has been one of such platform where such events spread like wild fire. Not only that, twitter explodes when we get to see a nail biting games between major sporting teams across all the sports, For example, on Oct. 11 The Lakers defeat the Miami Heat to capture a record-tying 17th NBA title. This lead to the explosion of tweets. So, in our analysis we observed that the user activity has changed abruptly when important events occur and it also changes with countries. There is always a twitter burst whenever a good movie is released in India, or when ever there is an IPL cricket match. Similarly every country has their own influential dates. We used credible news articles to identify such dates where twitter exploded for every country. [NyP]



Few Important dates for USA: Oct. 11: The Lakers defeat the Miami Heat to capture a record-tying 17th NBA title; Nov. 12: Coronavirus infections in California surpass 1 million; Dec. 14: The first COVID-19 vaccinations start in the United States; March 13: President Trump declares the coronavirus a national emergency; Oct. 2: Trump tests positive for COVID-19; Sept 12 : West coast wild fires Home Games of Baseball, BasketBall, Football for respective cities; Election days



Few important dates for INDIA: 19th Sept - 3rd Nov: Indian Premier league start date; 1/10/2020, Tanhaji Movie release; 24 January 2020 - Street dancer 3d movie release; March 6, 2020 - Bhaagi 3 movie release; 21 February 2020 - Subh Mangal movie release; 7 February 2020 - Malang movie release; 3 September 2020 - Pubg and other Chinese apps banned; Election days [IID]



Few Important days for Rest of world: Covid lockdown; Kobe Bryant death; Vaccine release announcement; Travel bans between respective countries.

Conclusion: From the above graphs on considering around 72 important days, it is clearly visible that people are active for a longer duration's and have a change in their circadian cycles as they are more active on twitter. Hence, there is an effect of important days on the circadian cycles of people.

6 Sleep Pattern Prediction Model

6.1 Dataset and Features

Using the above dataset described in the data preprocessing step, we used features that intuitively felt important which were supported by our analysis. Features include Total snow in cms, sunrise, sunset, moon-rise, moon set time, dew point in C, Feels like In C, Sun Index, UV Index, Heat Index, Wind Chill, Visibility, Humidity, Windspeed, Latitude, Longitude, TimeZone, Profession, Counntry, City, Month, Hour of the Day as training features. Now we had to have a relevant target variable that would exactly justify the features and the circadian cycles.

6.2 Target-Value

We categorized the sleep patterns in 4 categories based upon the common sleep cycles.

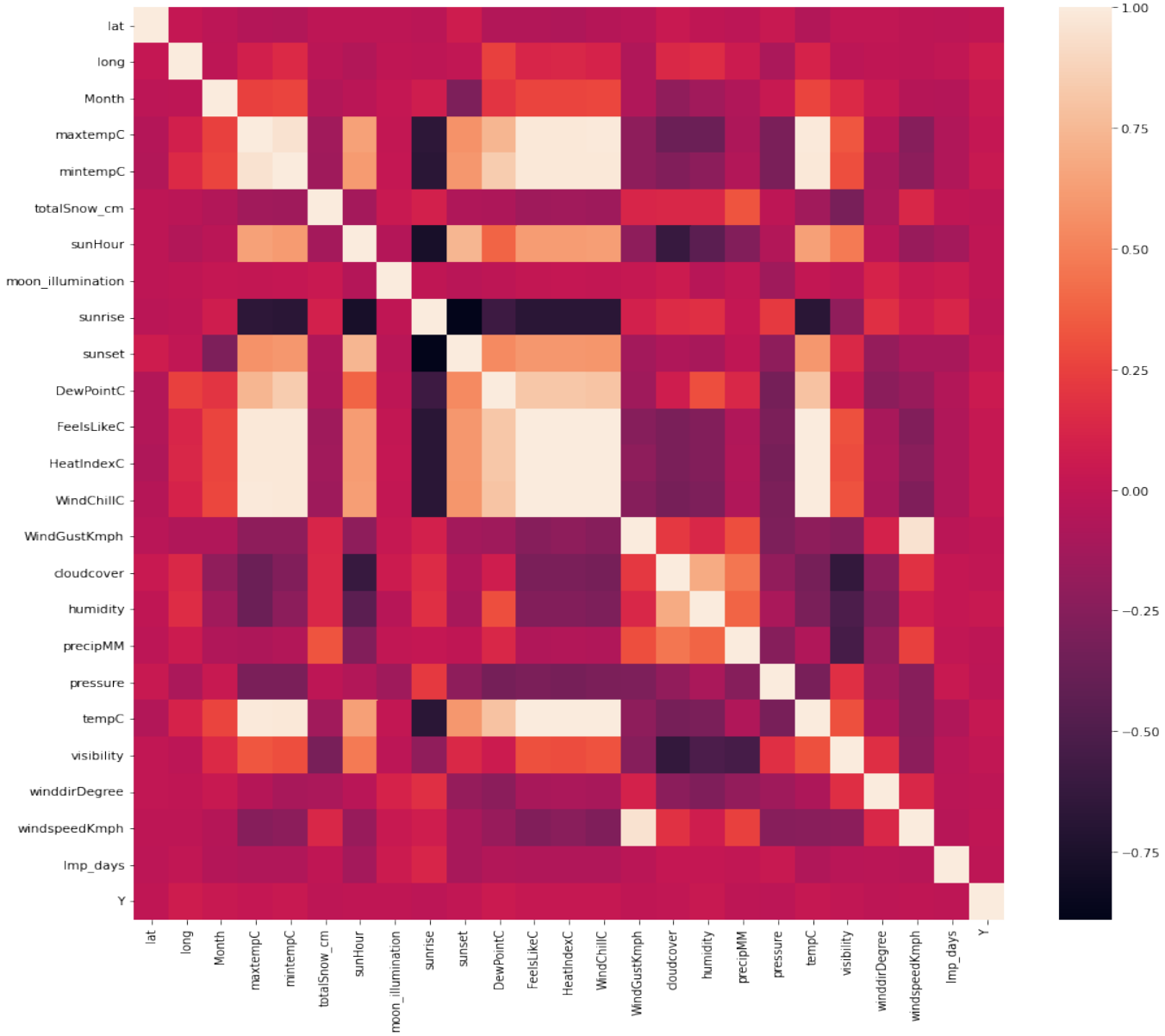
1. Early to Bed - Early to Rise : Sleep at 21:00 hours and wake up at 6:00 hours.
2. Early to Bed - Late to Rise: Sleep at 21:00 hours and wake up at 10:00 hours.
3. Late to Bed - Early to Rise: Sleep at 00:00 hours and wake up at 6:00 hours.

4. Late to Bed - Late to Rise: Sleep at 00:00 hours and wake up at 10:00 hours.

We have considered tweets of every user on a particular day in all the above time periods and considered the time period that has very less amount of tweets as the time period that he is sleeping. We have considered this as the ground truth or Target values to predict given the features. We got a equal distribution of users between all the sleep patterns.

Category	No. of Data Points
21:00 - 10:00	114099
21:00 - 06:00	99013
00:00 - 10:00	93786
00:00 - 06:00	86947

6.3 Correlation Map of the Features



We plotted the heatmap of our correlation matrix and found out that many features are important to predict our target value. Out of all the extracted features HeatIndexC ,FeelsLikeC ,DewPointC

,Month ,long ,tempC are highly co-related to that of the sleep cycles. Out of those many were proved in the above made analysis.

6.4 Model Training

After preparing the data set, Label encode the characteristic features such as profession, TimeZone, and Country etc. We have divided the total data set into training and testing set maintaining the distribution. We trained with different models to predict the target-value described above. Below are the models and accuracy rates:

Model Name	Accuracy
Catboost	56.63%
KNN Clustering	42.83%
XGBoost	83.18%

The best model that gave good accuracy is XGBoost with 500 estimators and 12 depth. These changes were made keeping in mind the complexity of the data set as well as the high number of feature columns. We also calculated T-test values among all our models.

Model Names	T-test Score
KNN and XGboost3	-12.45270929
Catboost and XGboost3	-7.12202067
XGboost1 and XGboost2	1.44398598
XGboost1 and XGboost3	5.1223969
XGboost2 and XGboost3	3.59998728

Since KNN and catboost have very low accuracy in predicting sleep cycles, their t-test values with XGBoost came up in negative. Where as XGBoost with different iterations performed better. Hence the positive value of XGBoost models.

6.5 Precision, Recall, F-1 Score

We calculated the below statistics values for our best model - XGBoost.

Statistics	Value
Precision	0.83
Recall	0.828
F1-Score	0.829

We can see that F1 score is around 0.829 which indicates the fact that the model is performing very well in predicting the sleep patterns based upon the conditions of the day along with some historical information of the user.

7 Conclusion

Hence, we were able to make proper analysis and a good prediction model using extra features like local time, plethora of weather data and important dates.

References

[GPY] GPY. Geopy - https://geopy.readthedocs.io/en/stable/.

[IID] IID. India important dates - https://en.wikipedia.org/wiki/2020$_i n_i ndia$.

[NyP] NyPost. Nypost - https://nypost.com/list/major-2020-events/.

[WWO] WWOA. Getting historical weather data - https://www.worldweatheronline.com/developer/.