Capstone Project 1: Data Story

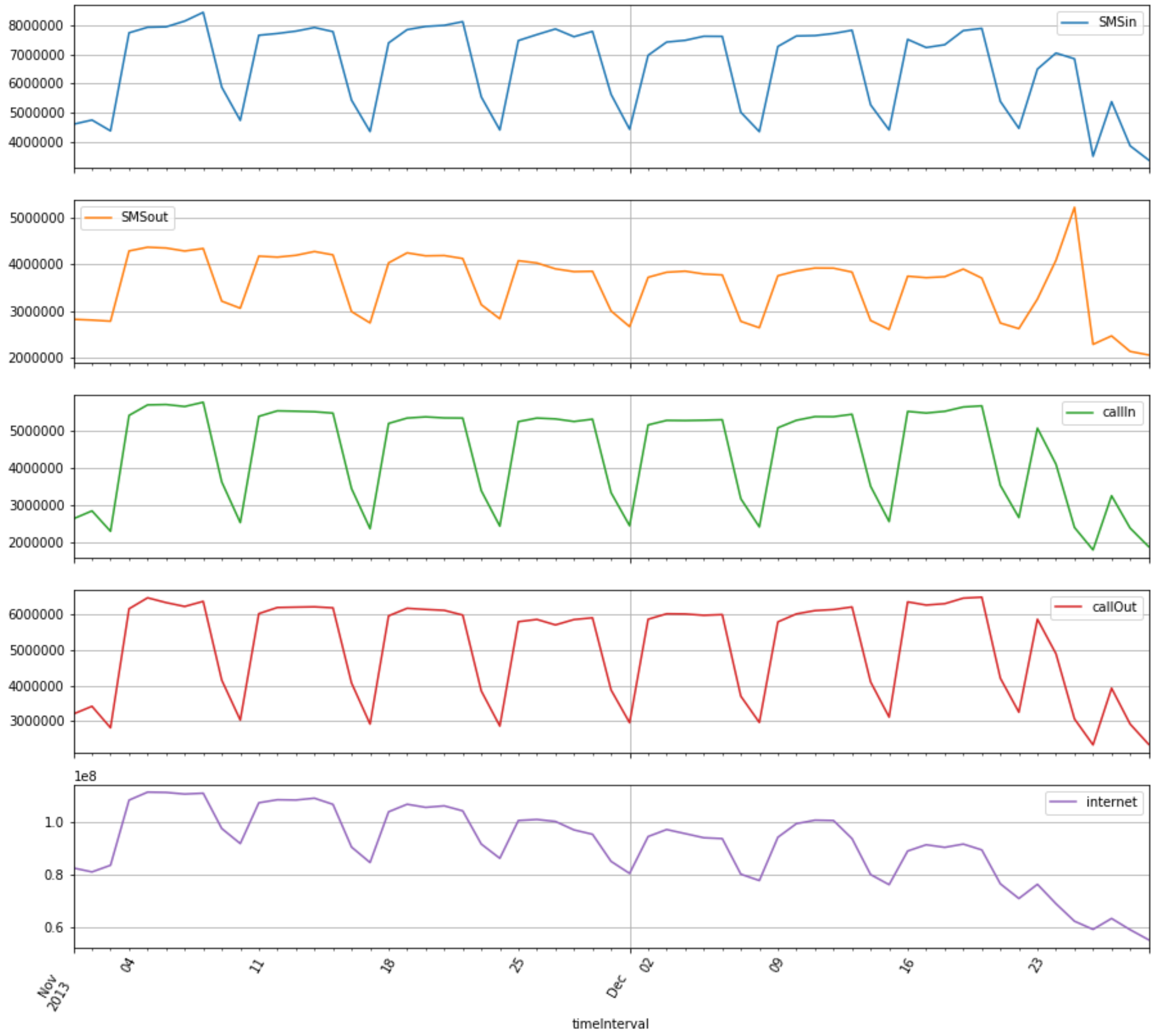
Criteria	Meets Expectations
Completion	❏ A 1-2 page Google doc communicating a data story for the project and data wrangling work completed to date.
Process and understanding	❏ The submission shows that the student used effective questions and exploration of the data. ❏ The submission shows that inferences, correlations, and/other relationships among the data were identified. ❏ The submission shows that a hypothesis was developed.
Presentation	❏ The project presentation demonstrates strong communication skills and presents insights.using text and visuals.

Excellence: The story is not only clear, but extremely well-written! The problem, story and conclusions are crisp and clear even to someone who doesn't know a lot of data science.

In this section, various insights produced through descriptive statistics and data visualization is presented

Consolidated usage patterns across all the grids

A time series line graph is provided below to illustrate the variation in the dependent variables over time.

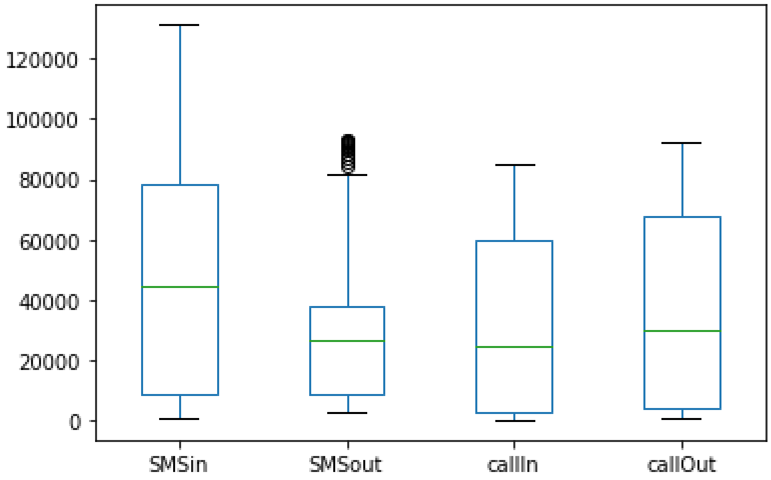


Inferences:

A. All telecom activities (SMS, calls, Internet) follow a similar pattern with usage being relatively the same across the weekdays and a sharp fall during the weekends

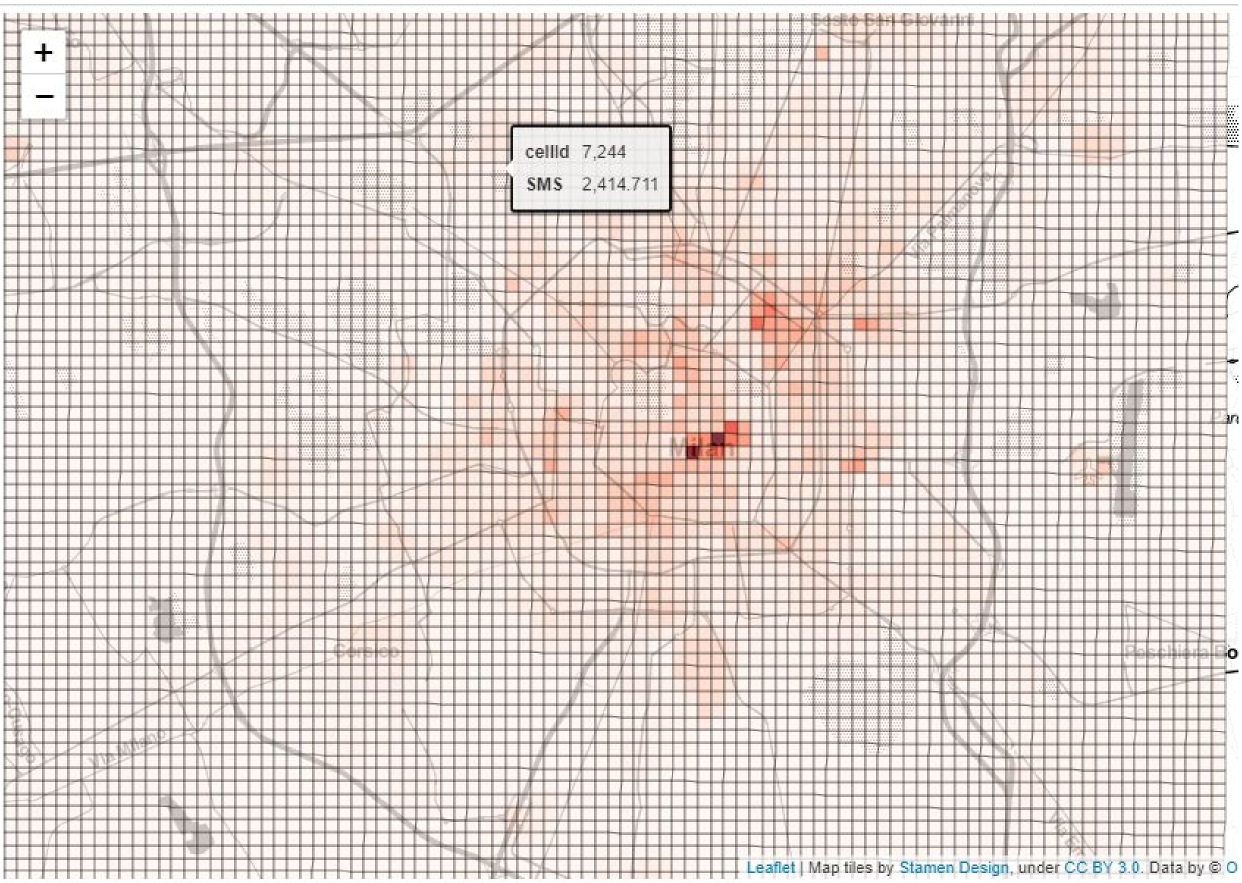
B. This pattern seems to be disrupted at the end of the timeline, this Christmas period and normal usage cannot be expected due to furlough

On further analysis based on the above pattern disruption, It was observed that there were some outliers in the SMSout category. These outliers were on one day 25-12-2013 between 10:00 AM and 12:20 PM. This peak SMSout is on **Christmas day** which is an expected behaviour during this time and these outliers does not seem like an data collection error and hence no further action is taken on these outliers



Consolidated Gridwise usage patterns

The SMS, calls and internet usage is aggregated on each grid ignoring the country code, this will give us the volume of usage per grid. A GeoPandas plot is created using this consolidated dataframe.



AllGrids.html

Inferences:

A. Grids concentrated in and around the city center have highest usage, whereas the usage is minimal in the outskirts. Usage per grid seems to gradually decrease as the distance of the grid from the city center increases.

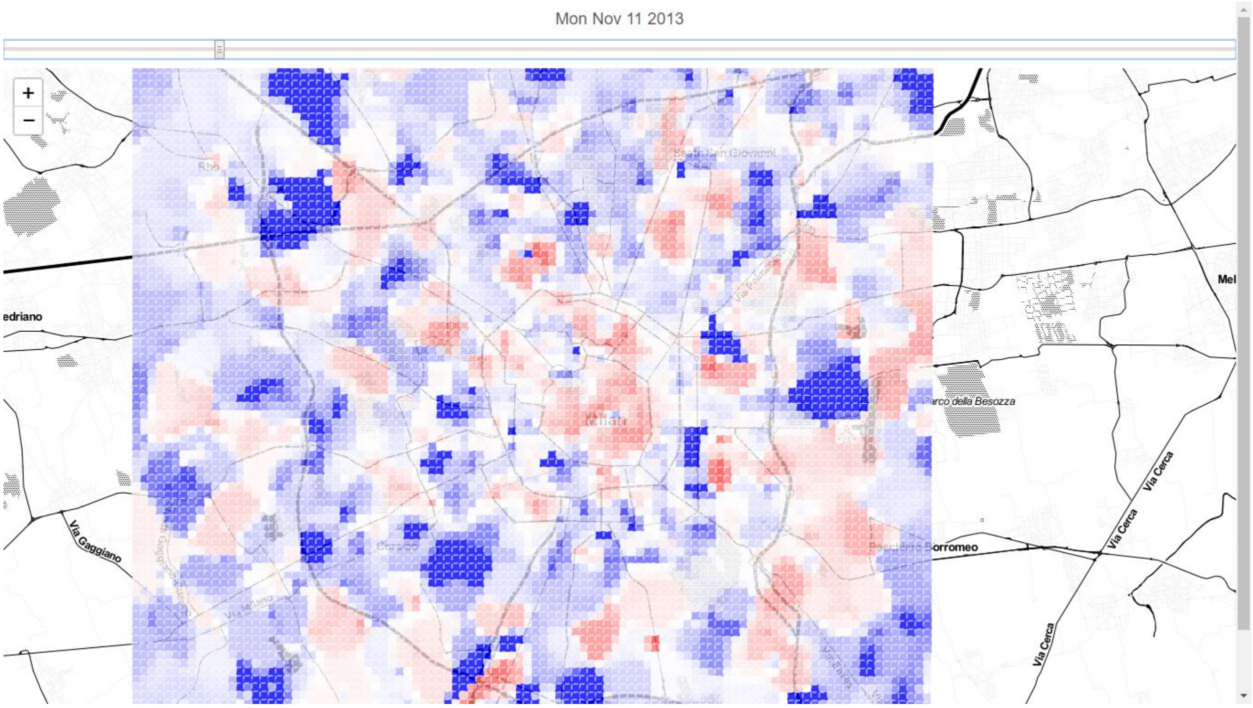
B. Couple of grids right in the center of Milan seems to have the highest usage.

C. Few groups of grids a little distance away from the grid also have relatively higher usage. These grids seem to be located around business areas and commercial hubs

D. North eastern outskirts has more usage than other outskirts of Milan

Consolidated Gridwise Interactive pattern per day

Further building on the previous plot a time series interactive map is generated. This indicates variation in dependent variable usage across all the grids. The map is generated using Geopandas and Folium. timeSeriesChoropleth is used to add a time slider to the map



TimeSliderChoropleth-SMSpctchange.html

This is a daily time series plot for percentage change in the SMS usage per grid on a folium map.

Inferences:

A. The usage in grids near the city center are almost constant (with slight changes, highlighted by white) during the weekdays

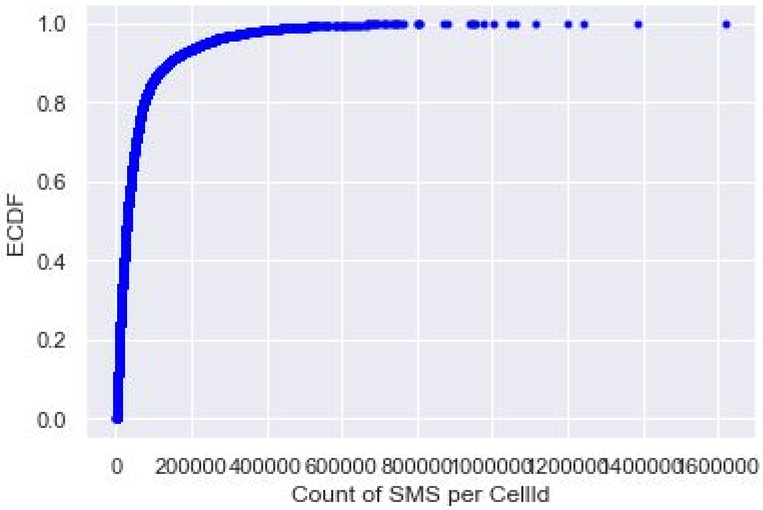
B. Some pockets of grids away from the city get highlighted during the weekends. Rest of the grids don't see any change. Based on this we can infer that these are residential non-commercial localities where people more time during weekends

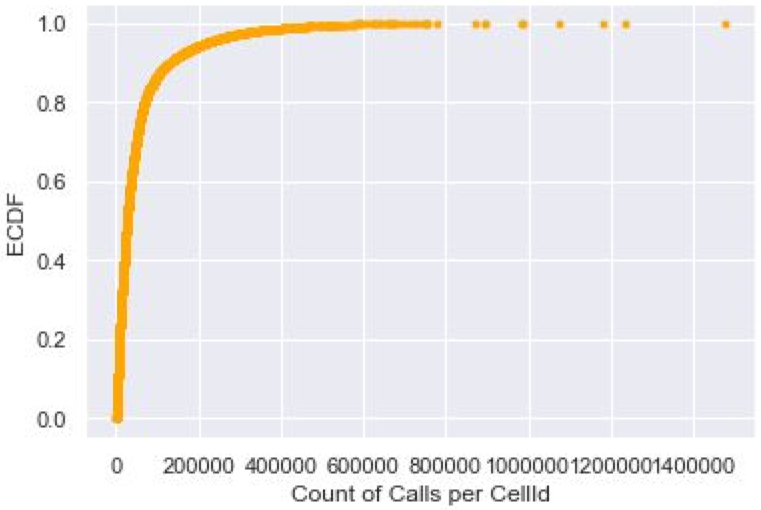
C. On Mondays, almost all the grids see an increase in usage irrespective of the location, this proves that during the weekend there is least usage across all the grids and the usage increases suddenly on Mondays.

D. This pattern of weekday and weekend is disrupted during the last week of the year when people are at home enjoying the holidays.

ECDF Gridwise

Using the same consolidated gridwise dataframe (consolidated dataframe for each cell irrespective of the date), aggregated based on the cellID is used to generate CDF to identify the frequency of usage for SMS and Calls.





Inferences:

A. 80% of the cells are producing less than 100k SMS and Calls over a period of 2 months
B. Only a few cells are utilizing more than 800k SMS and Calls
C. The grid utilization seems to conform to pareto principle (80% of the SMS and calls are generated on 20% of the Grids)