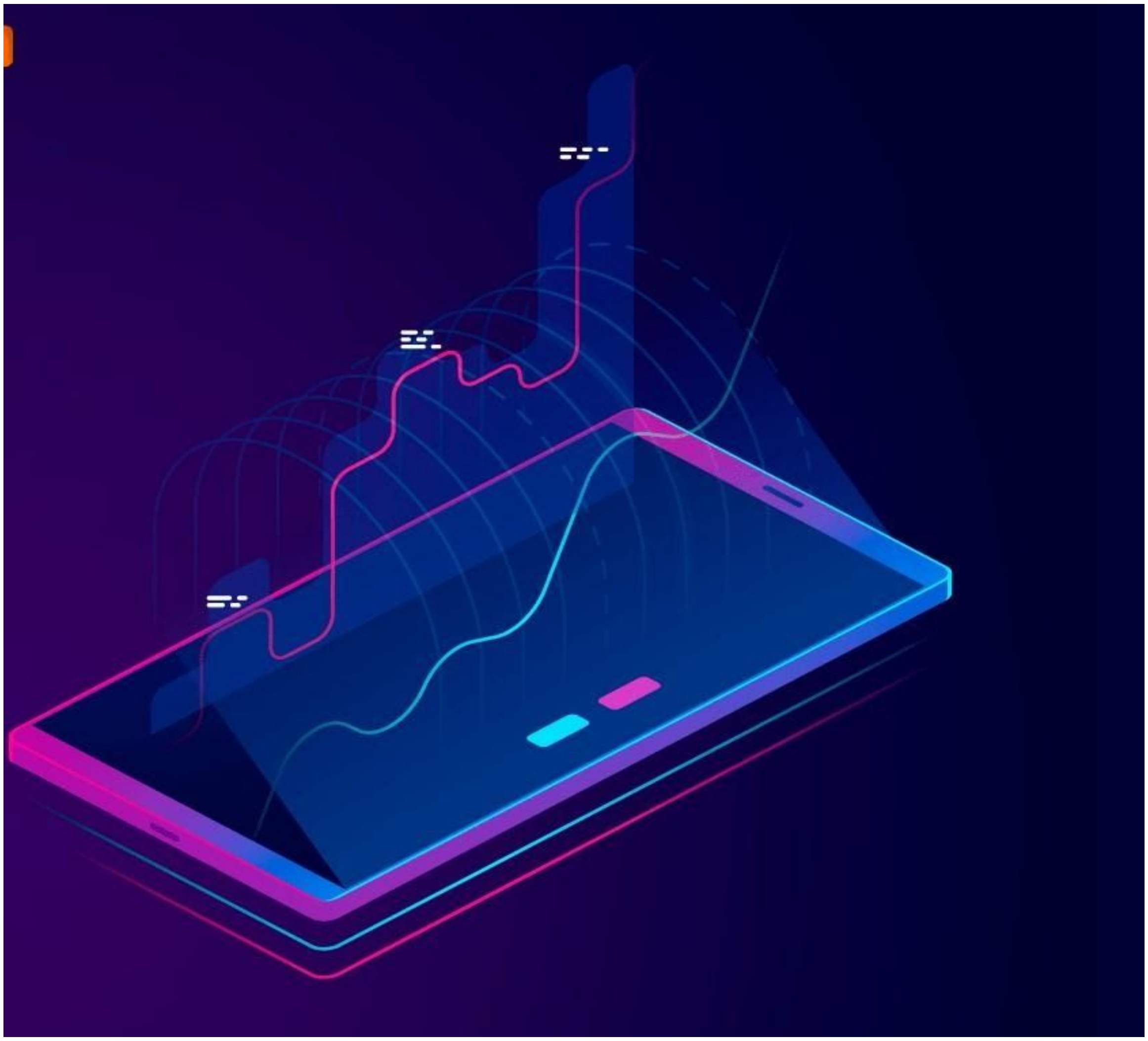
Forecasting Mobile Network Traffic Assignment-IMDEA Networks Institute

Aujasvi, Graduate Student in Data Science, University of Nantes (Polytech Nantes)
Email: aujasvimoudgil@gmail.com



Submitted To:
Marco Fiore
Research Associate Professor
IMDEA Networks Institute

Task I

(A) A brief description of the hardware and software setup used to process the data.

Solution: The hardware setup: Local Windows10 machine with the following configuration:

1. Processor: Intel(R) Core(TM) i7-7500U CPU @2.70GHz with Turbo Boost @2.90GHz
2. RAM: 16GB
3. System Type: 64-bit Operating System
4. GPU: NVIDIA GeForce 940MX (2GB)

The software setup used to preprocess the data, analysing and modelling is Jupyter Ipython Notebook. All the libraries and packages were installed using pip and conda, depending on the task.

(B) Figures illustrating the distribution (eg. probability density function) of the total traffic recorded in a geographical area during the two month, with an accompanying text where the applicant comments on the result (eg. discussing the homogeneity or diversity of the total traffic per area, and providing possible explanations for the observed behaviors)

Solution:

About Dataset The dataset is taken from a multi-source dataset of urban life in the city of Milan. The reference dataset was released by Telecom Italia Mobile (TIM) for a data analysis challenge. Although, there were many dataset available for the challenge, only two two are relevant for this assignment namely, Telecommunications activity and Grid dataset.

The telecommunication activity dataset gives a measure of level of interaction between the users and the mobile phone network. Following variables / features are present :

1. Squareid
2. Time Interval
3. SMS-in-activity
4. SMS-out-activity
5. Call-in-activity

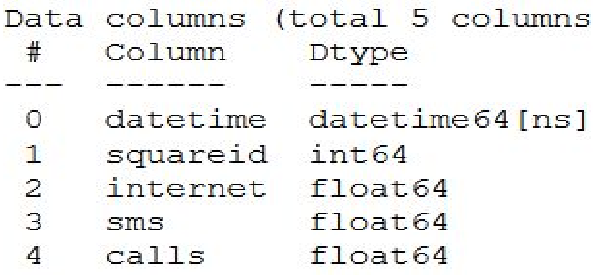
6. Call-out-activity
7. Internet traffic activity
8. Country code

For the assignment tasks, only some features are relevant such as Square id, Time interval and Internet traffic activity. We will consider these and ignore the rest while doing the analysis.

Analysis: First, we read all the datafiles separately for November 2013 and December 2013 Call Detail Records (CDRs). For ease, we are doing the analysis separately as it is done on a local machine. Total size of two months data is 9.5 GB (Approx.). Each datafile is in 24 hrs format with start-time 04:00:00.

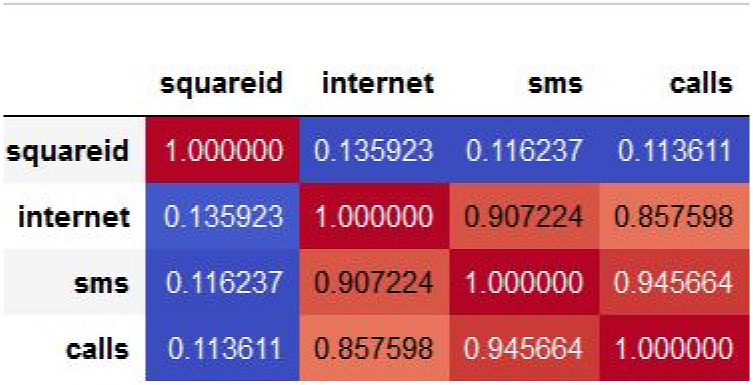
Basic summary and statistics of Telecommunication activity data set:

Table 1. Columns / Features:



```
Data columns (total 5 columns
 #   Column    Dtype
---  ------    -----
 0   datetime  datetime64[ns]
 1   squareid  int64
 2   internet  float64
 3   sms       float64
 4   calls     float64
```

Table 2. Correlations between different variables:



	squareid	internet	sms	calls
squareid	1.000000	0.135923	0.116237	0.113611
internet	0.135923	1.000000	0.907224	0.857598
sms	0.116237	0.907224	1.000000	0.945664
calls	0.113611	0.857598	0.945664	1.000000

Scatterplot Matrix

A Scatterplot can be formed between different features of the Telecommunication activity dataset where each feature is plotted against each other.

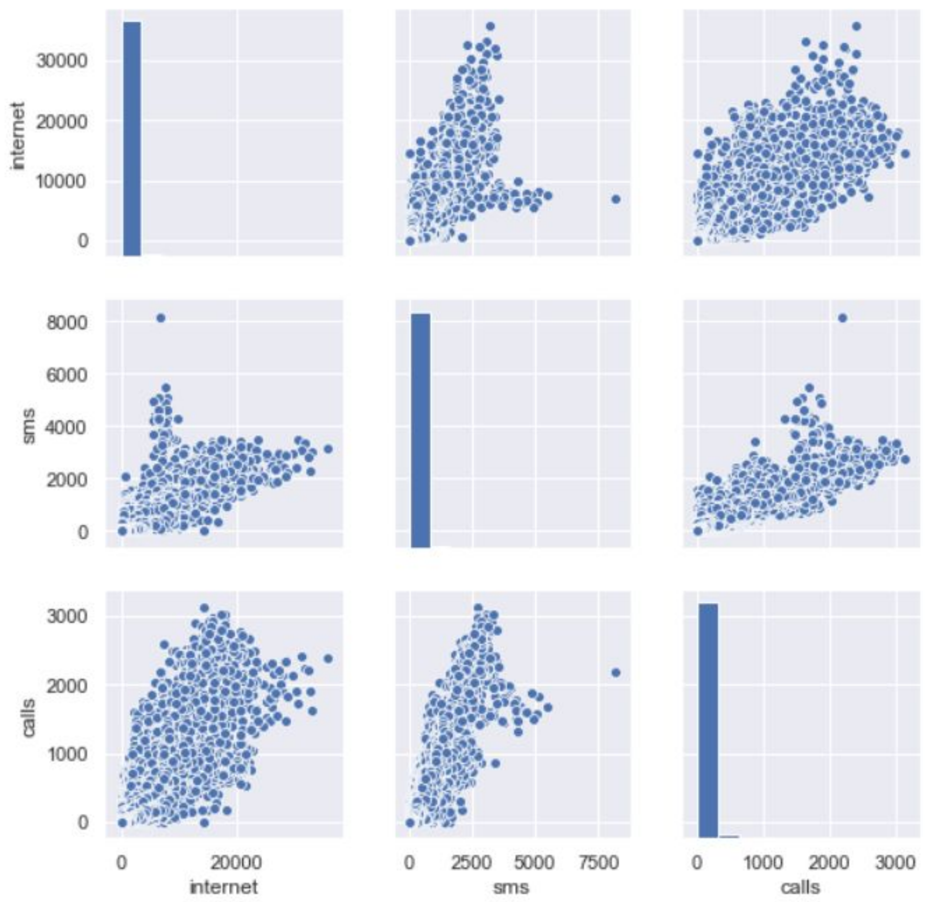
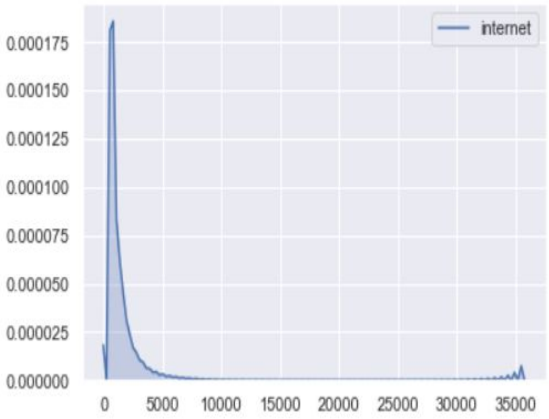


Fig. 1 Scatterplot Matrix

We can see there is a high positive correlation between sms, internet and calls from the above table and chart.

But, only three fields **Squareid**, **Time interval** and **Internet traffic** activity are relevant to the assignment from telecommunication activity dataset. So, we will consider only those.

Distribution (Probability Density Function) of total traffic recorded in the geographical area in two months



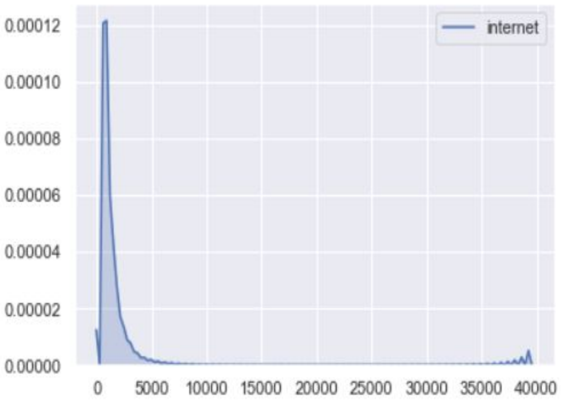


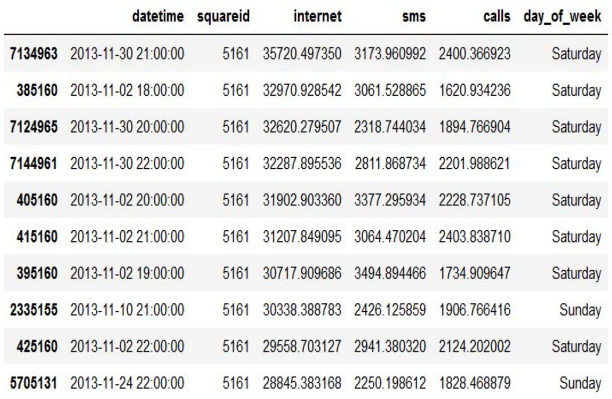
Fig. 2 PDF for November 2013

Fig. 3 PDF for December 2013

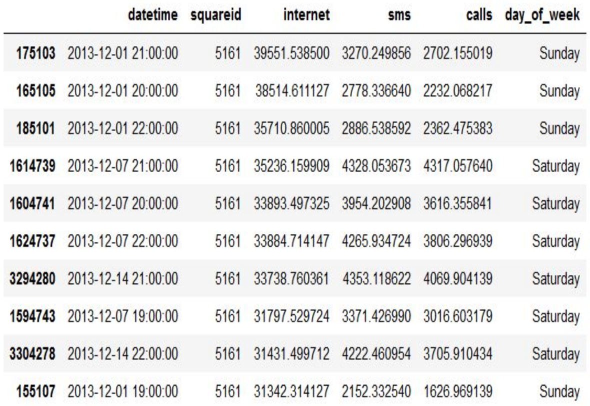
Above Fig. shows the Probability density Function (PDF) for total traffic recorded in the whole geographical area during the period of two months November and December in the city of Milan. Here, we can see both the PDF are heavily tailed distributed. Probably, this is due to network and data usage at different heterogeneous locations in the entire city of Milan. Popular locations have high network and data usage as compared to other locations in the whole Geographical area.

(C) A figure of the time series of network traffic during the first two weeks at four areas, namely (i) the area with the highest total traffic during the two-month period, (ii) the area with square id 4159, (iii) the area with squareid 4556, and (iv) the area with squareid 5160, with accompanying text where the applicant comments on the result (eg., discussing the similarities or differences observed in the temporal dynamics of each area, and speculating on their causes).

Solution: Check the area with Highest Total Traffic:



	datetime	squareid	internet	sms	calls	day_of_week
7134963	2013-11-30 21:00:00	5161	35720.497350	3173.960992	2400.366923	Saturday
385160	2013-11-02 18:00:00	5161	32970.928542	3061.528865	1620.934236	Saturday
7124965	2013-11-30 20:00:00	5161	32620.279507	2318.744034	1894.766904	Saturday
7144961	2013-11-30 22:00:00	5161	32287.895536	2811.868734	2201.988621	Saturday
405160	2013-11-02 20:00:00	5161	31902.903360	3377.295934	2228.737105	Saturday
415160	2013-11-02 21:00:00	5161	31207.849095	3064.470204	2403.838710	Saturday
395160	2013-11-02 19:00:00	5161	30717.909686	3494.894466	1734.909647	Saturday
2335155	2013-11-10 21:00:00	5161	30338.388783	2426.125859	1906.766416	Sunday
425160	2013-11-02 22:00:00	5161	29558.703127	2941.380320	2124.202002	Saturday
5705131	2013-11-24 22:00:00	5161	28845.383168	2250.198612	1828.468879	Sunday



	datetime	squareid	internet	sms	calls	day_of_week
175103	2013-12-01 21:00:00	5161	39551.538500	3270.249856	2702.155019	Sunday
165105	2013-12-01 20:00:00	5161	38514.611127	2778.336640	2232.068217	Sunday
185101	2013-12-01 22:00:00	5161	35710.860005	2886.538592	2362.475383	Sunday
1614739	2013-12-07 21:00:00	5161	35236.159909	4328.053673	4317.057640	Saturday
1604741	2013-12-07 20:00:00	5161	33893.497325	3954.202908	3616.355841	Saturday
1624737	2013-12-07 22:00:00	5161	33884.714147	4265.934724	3806.296939	Saturday
3294280	2013-12-14 21:00:00	5161	33738.760361	4353.118622	4069.904139	Saturday
1594743	2013-12-07 19:00:00	5161	31797.529724	3371.426990	3016.603179	Saturday
3304278	2013-12-14 22:00:00	5161	31431.499712	4222.460954	3705.910434	Saturday
155107	2013-12-01 19:00:00	5161	31342.314127	2152.332540	1626.969139	Sunday

Table 3. Highest Traffic Area in Nov 2013 **Table 4. Highest Traffic Area in Dec. 2013**

From the above tables, we can see **squareid 5161** has the Highest Total Traffic in two months.

Observations and comments regarding squareid 5161 (Temporal and Spatial aspects):

1. The squareid 5161 is near to Duomo (squareid 5060) which is the city centre of Milan and the most important tourist attraction.
2. Increase in network traffic during the weekends at the city centre and nearby area.
3. Increase in network traffic during the evening (Between 6pm - 10pm) shows people are out in the city centre and nearby area for different activities (e.g., eating, watching favourite football match, jogging, shopping) during weekends.
4. Dt. 30 Nov 2013, which had the highest traffic during the November month had a Animal rights activity protest where a large number of people were on streets. Dt. 2nd Nov 2013, which has also the second highest total traffic had a very important football match event between AC Milan and Fiorentina team.

5. Also, similar football match events also happened in the month of December during the first two weeks.

Time Series of Network traffic during the first two weeks at four Areas:

1. Area with the highest total traffic during the two month period (ie. squareid 5161)

2. Area with squareid 4159

3. Area with squareid 4556

4. Area with squareid 5160

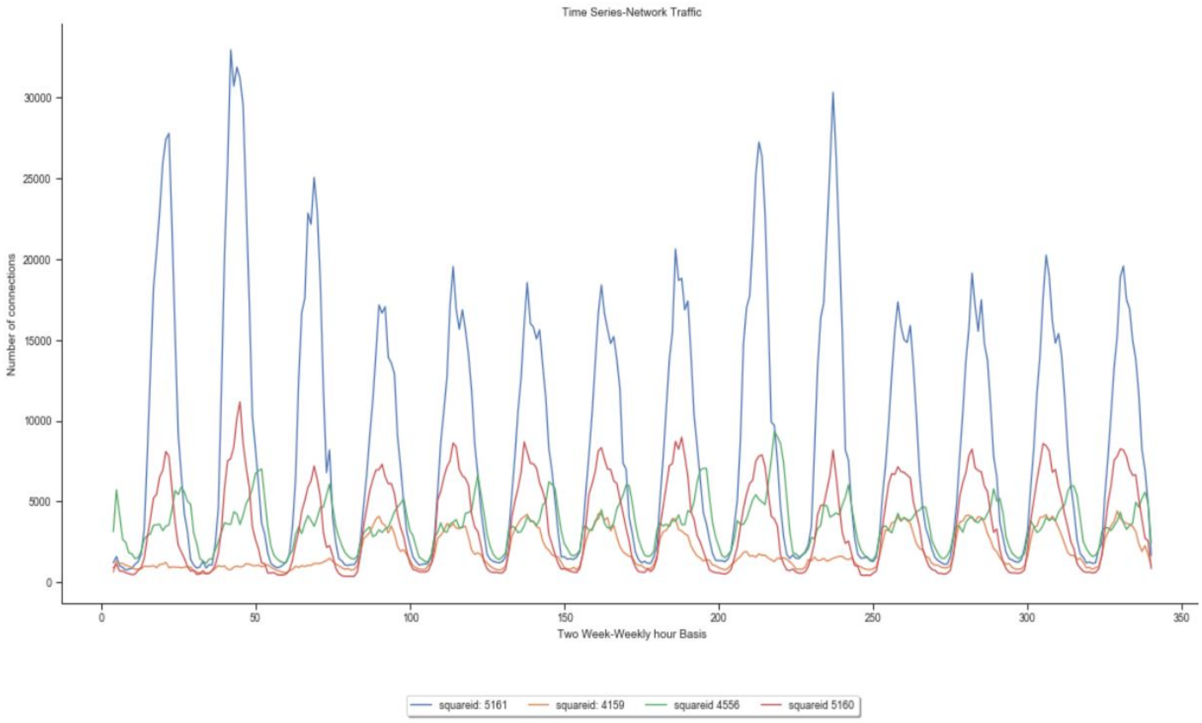


Fig 4. Time Series-Network Traffic (Four Areas)

Observations and Comments:

1. In the time series of network traffic in four geographical areas, the squareid 5161 and squareid 5160 are having high network traffic during the first two weeks. These two squareids are near to the city centre Duomo (squareid 5060).

2. squareid 4159 has overall low totaltraffic and number of connections got dropped during the weekends. This squareid is located near to Bocconi (squareid 4259), one of the famous Universities in Milan. (Weekdays Schedule)

3. squareid 4556 has less network traffic and number of connections as compared to the city centre locations/sites. But, there is a slight increase in the network traffic and number of connections as weekend approaches. This squareid is located near to the Navigli district (squareid 4456), one of the most famous nightlife places in Milan.

Representing Geographical Data using Grid GeoJSON dataset

About GeoJSON:

A GeoJSON file format is an open standard format that contains both geospatial data and attribute data. It is an extension from the JSON (JavaScript Object Notation) standard format.

It basically contains three different parts:

1. Geometry Object: This is either the point,line or polygon
2. Feature Object: This is the geometry object and the associated random ad hoc data
3. FeatureCollection: List of feature objects

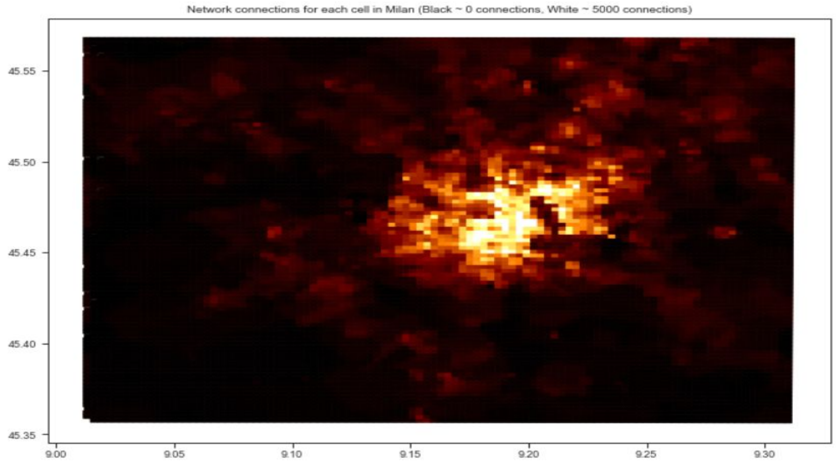


Fig 5. Hotspots for internet activity in Milan City (Highest at City Center)

Task II

In the second task, the applicant is asked to code an algorithm for one-step prediction of future traffic in a single area. Formally, let us denote as $x_a(t)$ the traffic observed at area a during the time interval t. At each time t, the algorithm receives as input a history x_t, ie., a vector of traffic values in past time intervals, up to t included. The algorithm shall then produce as output an estimate $x_a(t+1)$ of future traffic t+1 in area a. The applicant shall run the to forecast traffic in the four geographical areas identified at the second item of Task I, during the week from December 16 to 22. The following results should be produced as a minimum:

- **a self-contained description of the proposed algorithm;**
- **four plots reporting the superposed time series of (i) the original traffic, and (ii) the predicted traffic in the week from December 16 to 22;**
- **a table reporting the mean absolute error (MSE) and the mean absolute percentage error (MAPE) computed for time series at the second item above;**
- **exact statistics on the training and/or execution time of the algorithm, with details on the process used to compute such statistics and on the hardware on which they are recorded;**
- **a short text where the applicant provides personal considerations on the design and performance of the algorithm**

Solution:

Stationarity, Trend and Seasonality of Time Series

A time series is a collection of data points at constant time intervals and most of the time series models work on the assumption that the time series is stationary.

Stationary Series

The three basic criteria for Time Series to be classified as Stationary :

1. The mean of the series should not be a function of time rather should be constant.
2. The variance of the series should not be a function of time rather should be constant. This property is also called Homoscedasticity.
3. The covariance of the ith term and (i+m)th term should not be a function of time or the autocovariance that does not depend on time.

We can identify Stationarity of our time series using one of the statistical tests which is called the **Dickey-Fuller Test**.

Here, we consider the Null Hypothesis that the Time series is not stationary. The test results consist of a Test Statistic and some Critical Values for different confidence intervals. If the 'Test Statistic' is less than 'Critical Value', we can reject the null hypothesis and can say that our Time Series is stationary.

Once we identify that the time series is stationary, we can check **Trend, Seasonality** and **Noise** :

Trend : The increasing or decreasing value in the Time Series

Seasonality : The repeating cycles in the Time Series

Noise: The random variations in the Time Series

Firstly, we consider the area with highest traffic highest traffic ie. **squareid 5161**

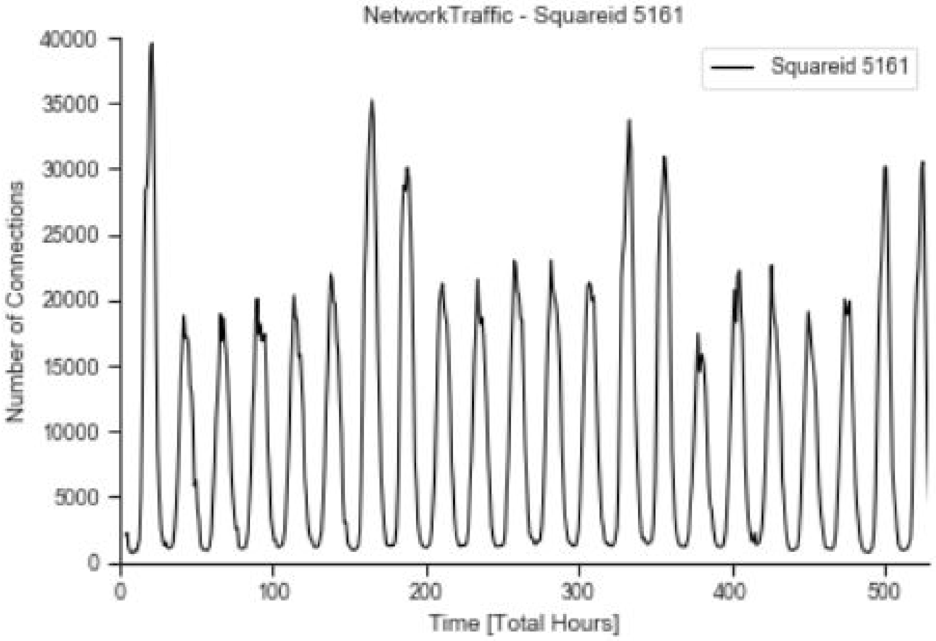
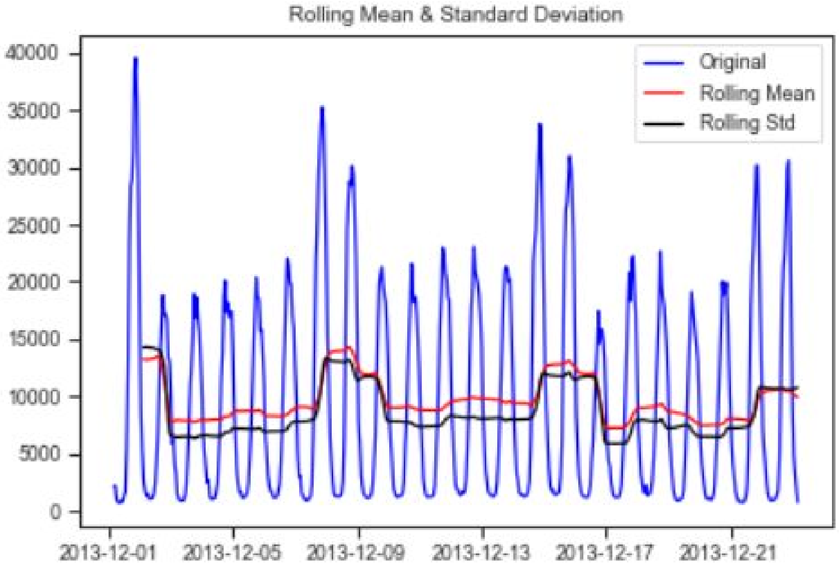
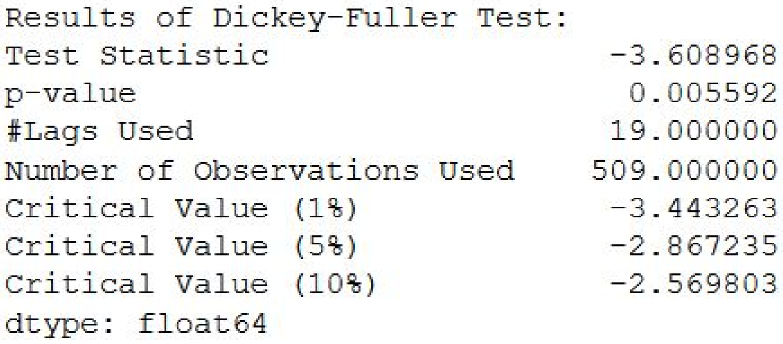


Fig 6. Time Series-Internet Traffic (squareid 5161)

From the above figure, we can see the periodic nature of the time series with total number of connections and we can see the **daily seasonality** (with no specific trend) which is consistent at every 24 hr period in the internet traffic activity.

Check Stationarity using Dickey-Fuller test:





```
Results of Dickey-Fuller Test:
Test Statistic                  -3.608968
p-value                          0.005592
#Lags Used                      19.000000
Number of Observations Used    509.000000
Critical Value (1%)             -3.443263
Critical Value (5%)             -2.867235
Critical Value (10%)            -2.569803
dtype: float64
```

Fig 7. Dickey-Fuller Test Results (squareid 5161)

Comments and Observations:

From the above fig, we can see there is small variation in the rolling values of mean and standard deviation and there is no specific continuously increasing or decreasing trend. This is more of a visual technique to check whether a time series is stationary or not from a graph.

To verify, we can use the Dickey-Fuller test results, Test statistic & Critical values for different confidence intervals. Here, the Test static (-3.60) is smaller than 1% Critical values. Therefore, we can say with 99% confidence interval that this time series is stationary. Also, with the p-value is less than 0.05 so we reject the null hypothesis & confirm that the series is stationary. Similarly, we can see the Dickey-Fuller test results for other squareids to check the stationarity of time series.

Decomposition:

In the method of STL decomposition (Seasonal and Trend decomposition by Loess), both trend and seasonality are modeled separately and the remaining part of the series is returned.

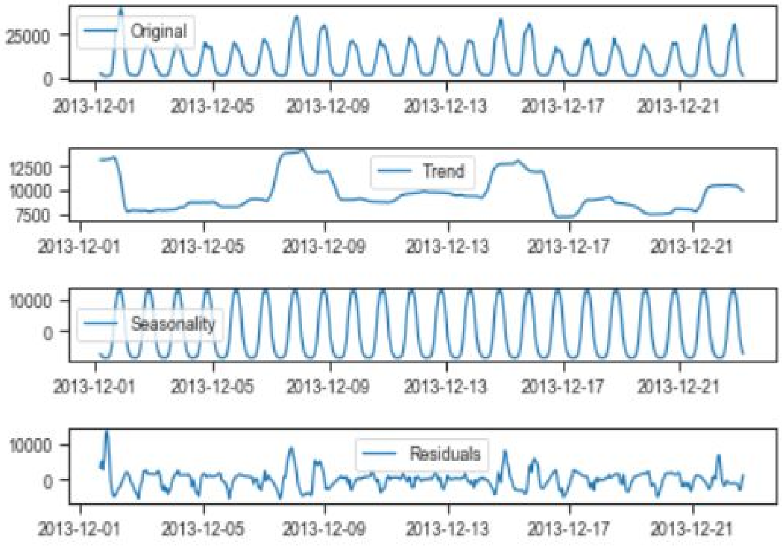
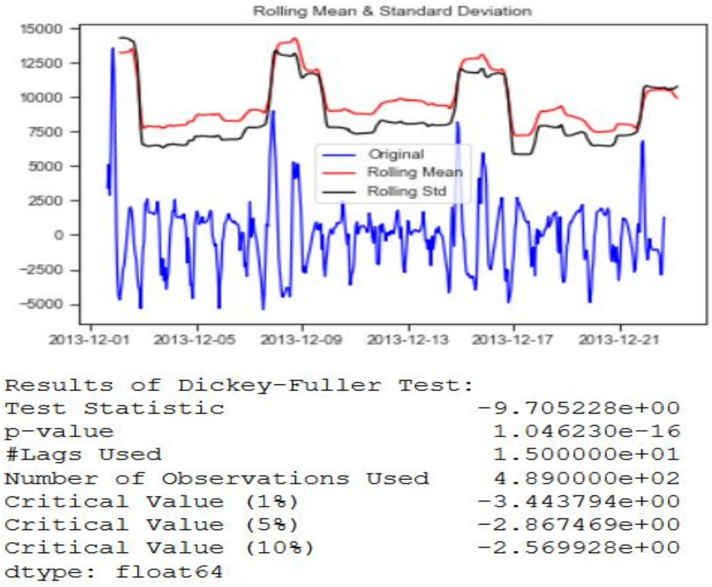


Fig 8. STL Decomposition (squareid 5161)

Here, we can see Trend and Seasonality are separated out of data. Now, we can model our residuals.



```
Results of Dickey-Fuller Test:
Test Statistic                 -9.705228e+00
p-value                         1.046230e-16
#Lags Used                      1.500000e+01
Number of Observations Used     4.890000e+02
Critical Value (1%)            -3.443794e+00
Critical Value (5%)            -2.867469e+00
Critical Value (10%)           -2.569928e+00
dtype: float64
```

Fig 9. Dickey-Fuller Test Results (Residuals)

The Dickey-Fuller test also shows that the Test-Static value for residuals is also smaller than 1% Critical values. So, this time series is very close to stationary.

Eliminating Trend and Seasonality:

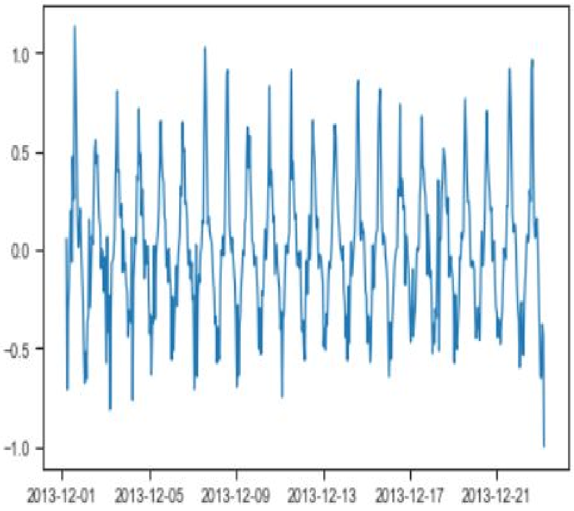
Trend can be removed by transformation and we can use Log Transformation to eliminate the trend.

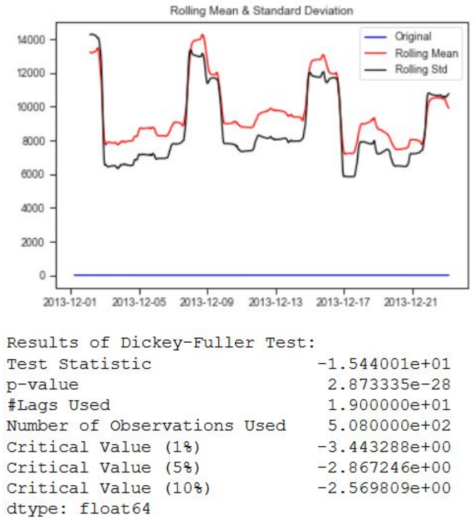
Our time series has some daily Seasonal component, which can be removed. A time series where the seasonal component has been removed is called seasonal stationary.

There are different methods to correct seasonal component and eliminate seasonality :

1. Difference: Taking the difference at a particular time lag
2. Decomposition: Modelling both trend and seasonality and removing them from the model

In our case, we have seasonal components at a lag of 24 hrs. We can remove this seasonal component by subtracting the value from the previous day.





```
Results of Dickey-Fuller Test:
Test Statistic                -1.544001e+01
p-value                        2.873335e-28
#Lags Used                     1.900000e+01
Number of Observations Used    5.080000e+02
Critical Value (1%)           -3.443288e+00
Critical Value (5%)           -2.867246e+00
Critical Value (10%)          -2.569809e+00
dtype: float64
```

Fig 10. Log-Diff Time Series **Fig 11. Dickey-Fuller Test Results (Log-Diff)**

Dickey-Fuller test statistic is less than the 1% Critical value.

Forecasting the Time Series:

Forecasting the Time Series can be divided into **two** types:

1. If we use only the previous values of the time series to predict its future values it is called Univariate Time Series Forecasting.

2. If we use predictors values other than series values (Eg. Exogenous Variables) to forecast it is called Multivariate Time Series Forecasting.

In our case, we are considering only internet network traffic therefore it is Univariate Time Series Forecasting.

Here, we will be using the **ARIMA** model as a forecasting method.

ARIMA Model which stands for **Autoregressive Integrated Moving Average** is a generalization of ARMA (Autoregressive Moving Average) model. It works on the idea that the information in the past values of the time series can alone be used to predict the future values.

An ARIMA model is characterized by three terms: **p, d, q**

where,

p is the order of the AR term

q is the order of the MA term

d is the number of differencing required to make the time series stationary

Both Seasonal and Non Seasonal time series can be modeled with ARIMA. If a time series, has seasonal patterns, then we need to add seasonal terms and it becomes SARIMA ie. 'Seasonal ARIMA'.

p and q terms:

'p' is the order of the 'Auto Regressive' (AR) term. It refers to lags of dependent variables. And 'q' is the order of the 'Moving Average' (MA) term. It refers to the number of lagged forecast errors that should go into the ARIMA Model.

ACF and PACF plots:

These plots are basically used to determine the value of 'p' and 'q'.

AutoCorrelation Function (ACF):

A plot of auto-correlation of different lags is called ACF.

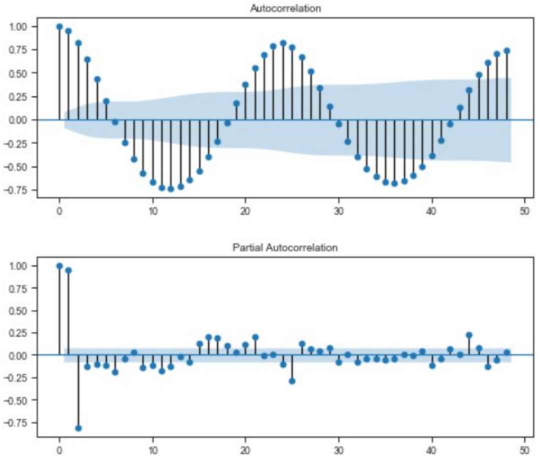
The plot summarizes the correlation of an observation with lag values. The x-axis shows the lag and the y-axis shows the correlation coefficient between -1 and 1 for negative and positive correlation.

Partial AutoCorrelation Function (PACF):

A plot of partial auto-correlation for different values of lags is called PACF.

The plot summarizes the correlations for an observation with lag values that is not accounted for by prior lagged observations.

Plot ACF and PACF



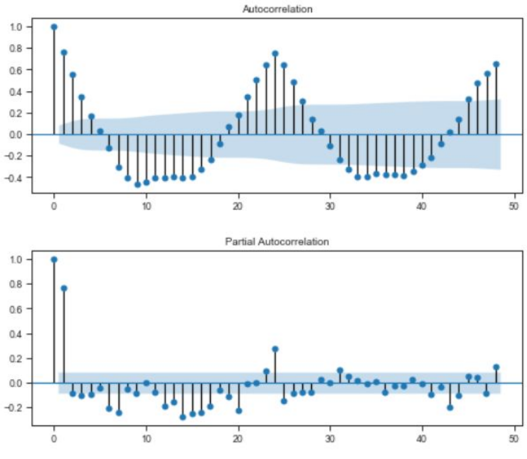


Fig 12. ACF and PACF (No Diff.)

Fig 13. ACF and PACF (Ist order Diff.)

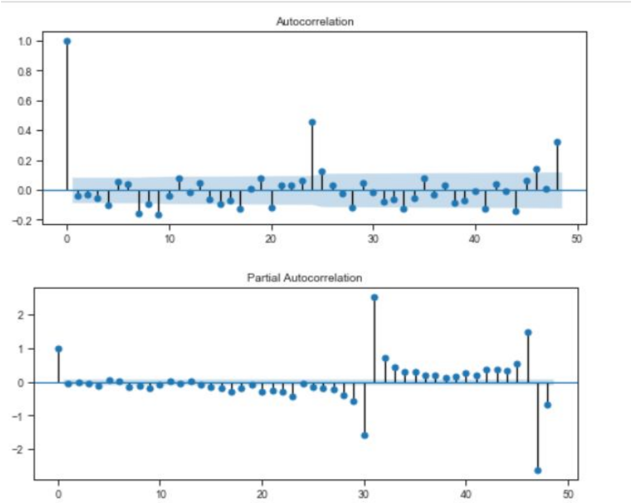


Fig 14. ACF and PACF (IInd order Diff.)

The right order of differencing is the minimum differencing required to get a near stationary series which roams around a defined mean and the ACF plot reaches to zero fairly quickly.

If the autocorrelations are positive for many numbers of lags (10 or more), then the series needs further differencing. On the other hand, if the lag 1 autocorrelation itself is too negative, then the series is probably over-differenced.

For the above time series, looking at the autocorrelation plot for the IInd order differencing the lag goes into the far negative zone more quickly, which indicates, the series might be over-differenced. Therefore, we should consider ACF and PACF plots with Ist order of differencing / Without differencing.

AR term can be inspected by the PACF plot. We can observe, the PACF lag 1 and lag 2 is quite significant and is well above the significance line.

MA term can be inspected by ACF plot.

PACF & ACF suggested that AR(2) and MA(2).

Model Building:

As we know, we have a daily seasonality (with no specific trend) at every 24 hr period in the internet traffic activity. Therefore, we can either use **Auto-Arima or SARIMA (Seasonal Arima)**.

The advantage of using Auto-Arima over ARIMA is that after data preprocessing we can skip ACF and PACF steps for manually determining / inspecting AR and MA values. It uses the AIC (Akaike Information Criteria) and BIC (Bayesian Information Criteria) values generated by trying different combinations of p,d & q values to fit the model.

Auto-Arima

Squareid 5161 (Area with the highest traffic in the four geographical areas)

Superposed time series of (i) the original traffic (ii) the predicted traffic in the week from December 16 to 22

Plot

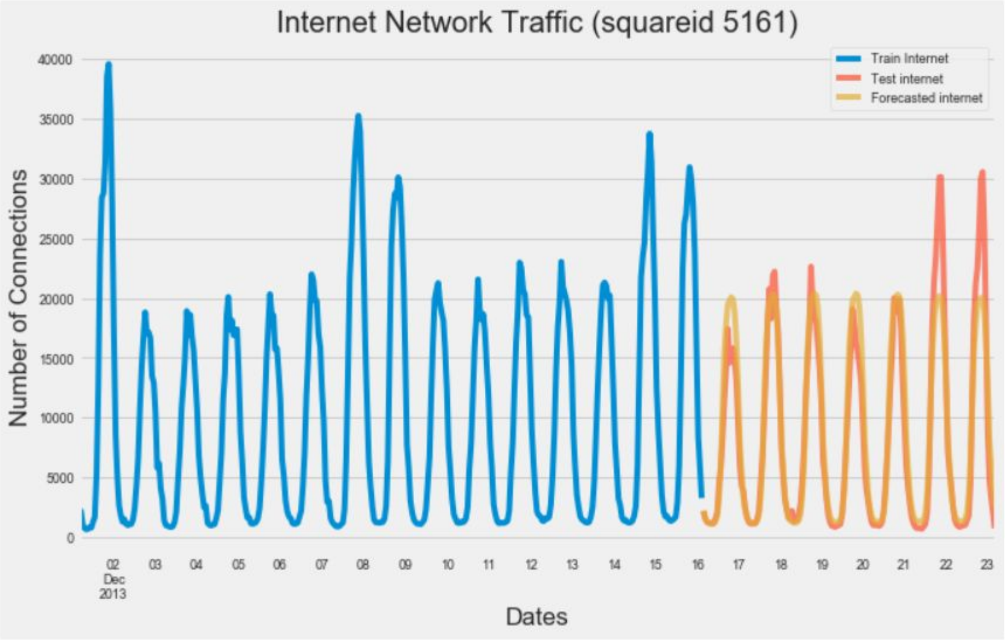


Fig 15. Superposed Time Series (Original and Forecasted)

Comments and Observations:

1. From the graph, we can see Dates on x-axis and Number of internet connections made during the month of December 2013.
2. The BLUE line belongs to network traffic for the first 15 days used to train the model and Red Line belongs to network traffic used as a test set for the week 16-22, December 2013. Yellow line belongs to the forecasted value for that week.
3. Here, it is clearly visible that the Auto-ARIMA model forecast is much closer to the actual pattern during the weekdays than weekends.

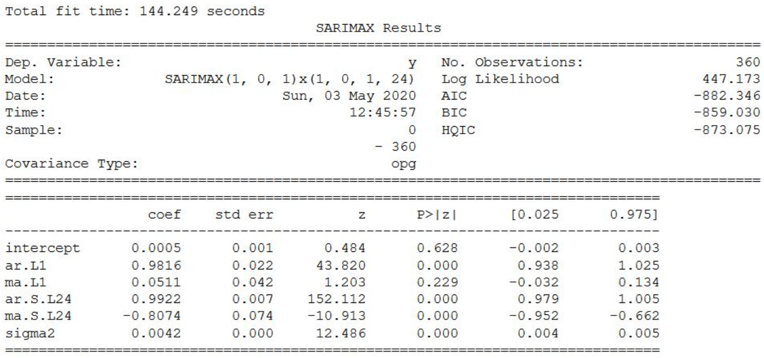
Table 5. Accuracy Metrics

MSE	1582.24
MAPE	24.59

Comments and Observations:

1. Around 24.59% MAPE implies that the model is about 75.41 % accurate in predicting the 16-22 week observations (squareid 5161). Although, this accuracy is not so good.
2. Probably, this is because we have an hourly dataset which exhibits multiple-seasonalities. For example, On a particular day, we can have intra-daily seasonalities as well as seasonality at every 24 hrs between previous and next day. Also, these can be different on weekdays and weekends.
3. Generally, ARIMA does not handle multiple seasonalities quite well (https://stats.stackexchange.com/questions/360167/arima-model-configuration-for-hourly-forecasting-problem)

Table 6. Statistics on the training and/or execution time



```
Total fit time: 144.249 seconds
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  360
Model:             SARIMAX(1, 0, 1)x(1, 0, 1, 24)   Log Likelihood                 447.173
Date:                            Sun, 03 May 2020   AIC                           -882.346
Time:                                    12:45:57   BIC                           -859.030
Sample:                                         0   HQIC                          -873.075
                                            - 360
Covariance Type:                              opg
==========================================================================================
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept      0.0005      0.001      0.484      0.628      -0.002       0.003
ar.L1          0.9816      0.022     43.820      0.000       0.938       1.025
ma.L1          0.0511      0.042      1.203      0.229      -0.032       0.134
ar.S.L24       0.9922      0.007    152.112      0.000       0.979       1.005
ma.S.L24      -0.8074      0.074    -10.913      0.000      -0.952      -0.662
sigma2         0.0042      0.000     12.486      0.000       0.004       0.005
==============================================================================
```

Squareid 4159

Superposed time series of (i) the original traffic (ii) the predicted traffic in the week from December 16 to 22

Plot

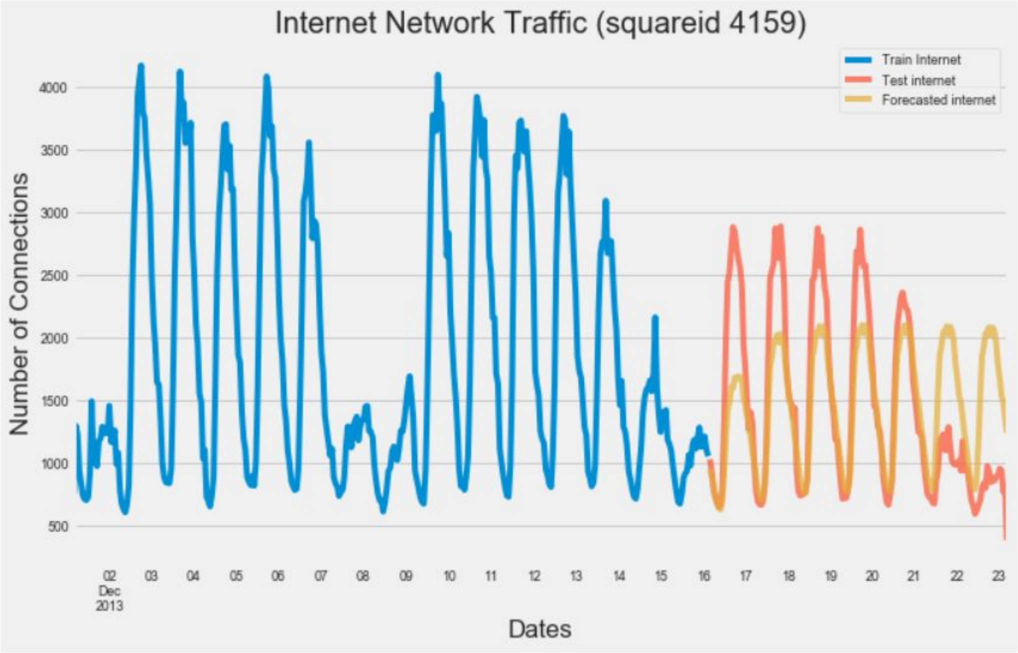


Fig 16. Superposed Time Series (Original and Forecasted)

Comments and Observations:

1. From the graph, we can see Dates on x-axis and Number of internet connections made during the month of December 2013.
2. The BLUE line belongs to network traffic for the first 15 days used to train the model and Red Line belongs to network traffic used as test set for the week 16-22, December 2013. Yellow line belongs to the forecasted value for that week.
3. Here, it is clearly visible that there is a lot of variability in the network traffic. The traffic during the weekdays is much higher than weekends because of the Universities schedule. There is abrupt change in network traffic on weekdays following the weekends. The Auto-ARIMA model is not able to capture those changes / differences.

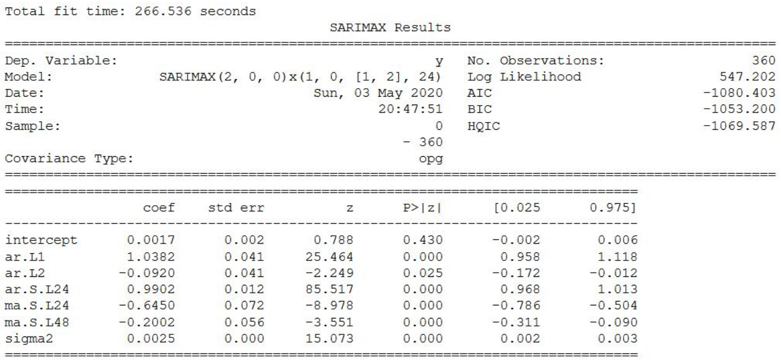
Table 6. Accuracy Metrics

MSE	390.99
MAPE	30.66

Comments and Observations:

1. Around 30.66% MAPE implies that the model is about 69.39 % accurate in predicting the 16-22 week observations (squareid 4159).
2. As we saw earlier this squareid is near to squareid 4259 where famous Universities of Milan are present. There is higher network traffic during weekdays than weekends and there is a lot of variability in network traffic. Also, as discussed earlier, we have an hourly dataset which exhibits multiple-seasonalities.
3. Generally, ARIMA does not handle multiple seasonalities quite well (https://stats.stackexchange.com/questions/360167/arima-model-configuration-for-hourly-forecasting-problem)

Table 7. Statistics on the training and/or execution time



```
Total fit time: 266.536 seconds
                                        SARIMAX Results
===============================================================================================
Dep. Variable:                                       y   No. Observations:                  360
Model:             SARIMAX(2, 0, 0)x(1, 0, [1, 2], 24)   Log Likelihood                 547.202
Date:                                 Sun, 03 May 2020   AIC                          -1080.403
Time:                                         20:47:51   BIC                          -1053.200
Sample:                                              0   HQIC                         -1069.587
                                                 - 360
Covariance Type:                                   opg
===============================================================================================
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept      0.0017      0.002      0.788      0.430      -0.002       0.006
ar.L1          1.0382      0.041     25.464      0.000       0.958       1.118
ar.L2         -0.0920      0.041     -2.249      0.025      -0.172      -0.012
ar.S.L24       0.9902      0.012     85.517      0.000       0.968       1.013
ma.S.L24      -0.6450      0.072     -8.978      0.000      -0.786      -0.504
ma.S.L48      -0.2002      0.056     -3.551      0.000      -0.311      -0.090
sigma2         0.0025      0.000     15.073      0.000       0.002       0.003
==============================================================================
```

Squareid 4556

Superposed time series of (i) the original traffic (ii) the predicted traffic in the week from December 16 to 22

Plot

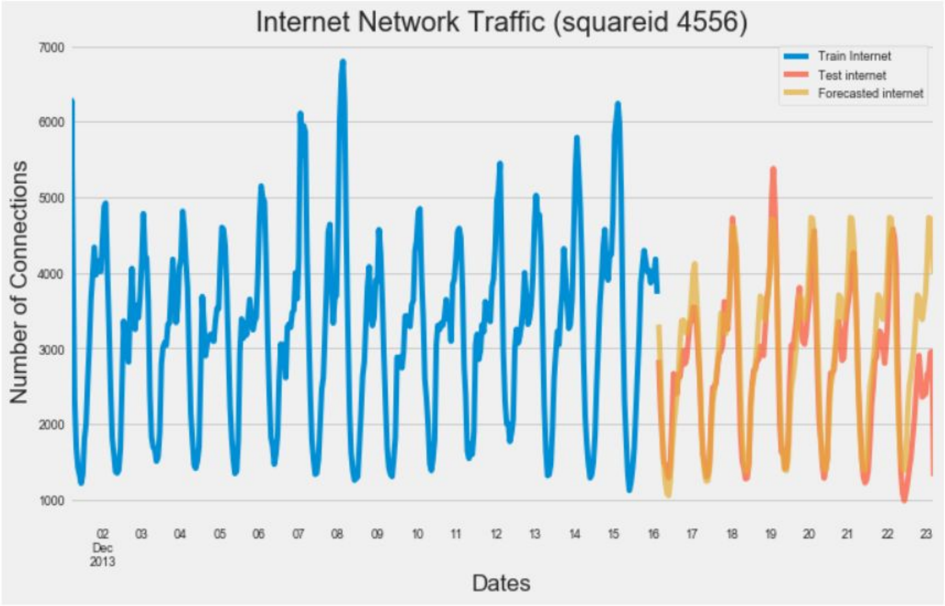


Fig 17. Superposed Time Series (Original and Forecasted)

Comments and Observations:

1. From the graph, we can see Dates on x-axis and Number of internet connections made during the month of December 2013.
2. The **BLUE line** belongs to network traffic for the first 15 days used to train the model and **Red Line** belongs to network traffic used as a test set for the week 16-22, December 2013. **Yellow line** belongs to the forecasted value for that week.
3. Here, We see some sharp peaks in the graph on each day repeatedly. This increase in traffic is during evening hours. The Auto-ARIMA model is able to capture this pattern quite well.

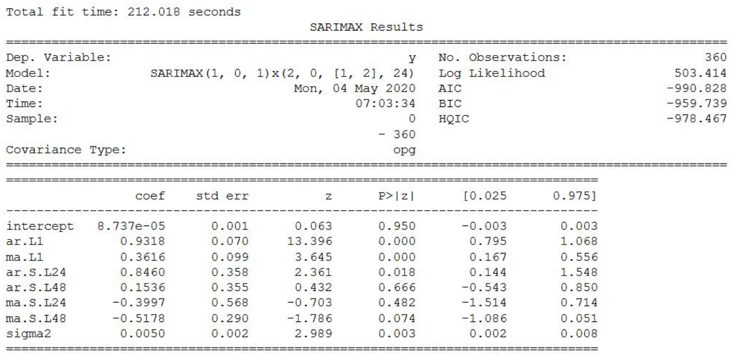
Table 8. Accuracy Metrics

MSE	394.127
MAPE	16.51

Comments and Observations:

1. Around 16.51% MAPE implies that the model is about 83.49 % accurate in predicting the 16-22 week observations (squareid 4159). This accuracy is more as compared to other squareids calculated above.
2. As we saw earlier the squareid is near to squareid 4456 which is present in the city centre and one of the most famous nightlife places in Milan city. There is high network traffic during evening hours.

Table 9. Statistics on the training and/or execution time



```
Total fit time: 212.018 seconds
                                      SARIMAX Results
============================================================================================
Dep. Variable:                                   y   No. Observations:                   360
Model:             SARIMAX(1, 0, 1)x(2, 0, [1, 2], 24)   Log Likelihood                 503.414
Date:                             Mon, 04 May 2020   AIC                            -990.828
Time:                                     07:03:34   BIC                            -959.739
Sample:                                          0   HQIC                           -978.467
                                             - 360
Covariance Type:                               opg
============================================================================================
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept   8.737e-05      0.001      0.063      0.950      -0.003       0.003
ar.L1          0.9318      0.070     13.396      0.000       0.795       1.068
ma.L1          0.3616      0.099      3.645      0.000       0.167       0.556
ar.S.L24       0.8460      0.358      2.361      0.018       0.144       1.548
ar.S.L48       0.1536      0.355      0.432      0.666      -0.543       0.850
ma.S.L24      -0.3997      0.568     -0.703      0.482      -1.514       0.714
ma.S.L48      -0.5178      0.290     -1.786      0.074      -1.086       0.051
sigma2         0.0050      0.002      2.989      0.003       0.002       0.008
==============================================================================
```

Squareid 5160

Superposed time series of (i) the original traffic (ii) the predicted traffic in the week from December 16 to 22

Plot

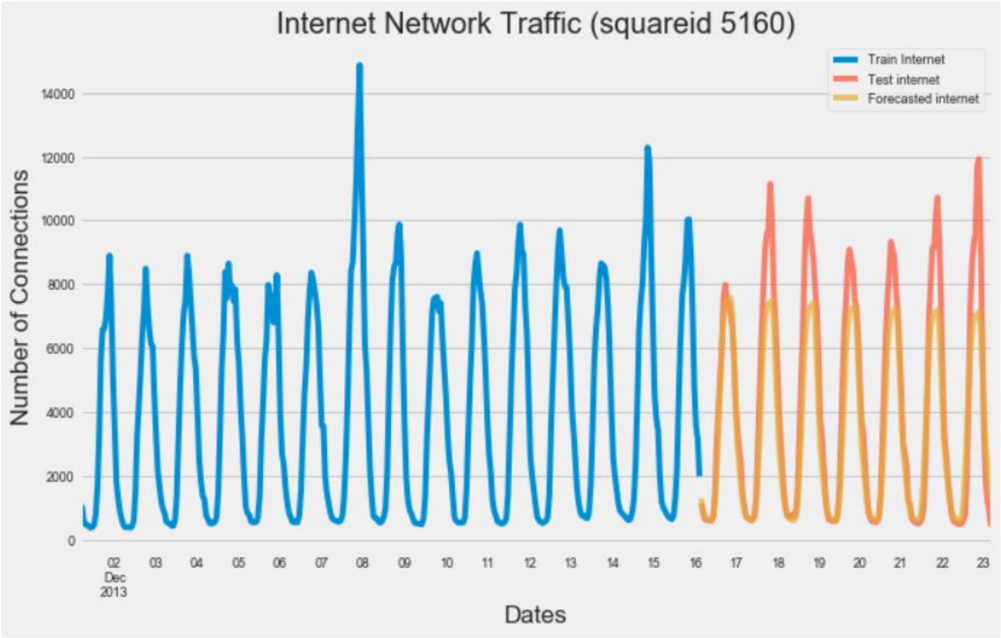


Fig 18. Superposed Time Series (Original and Forecasted)

Comments and Observations:

1. From the graph, we can see Dates on x-axis and Number of internet connections made during the month of December 2013.
2. The **BLUE line** belongs to network traffic for the first 15 days used to train the model and **Red Line** belongs to network traffic used as test set for the week 16-22, December 2013. **Yellow line** belongs to the forecasted value for that week.
3. Here, the ARIMA model is able to capture internet network traffic where there are not many sharp peaks. The original internet traffic and number of connections made during the first 15 days of December 2013 is less as compared to the internet traffic during 16-22 week of december 2013. The Auto-ARIMA model is not able to spot these surge in the network traffic and number of connections.

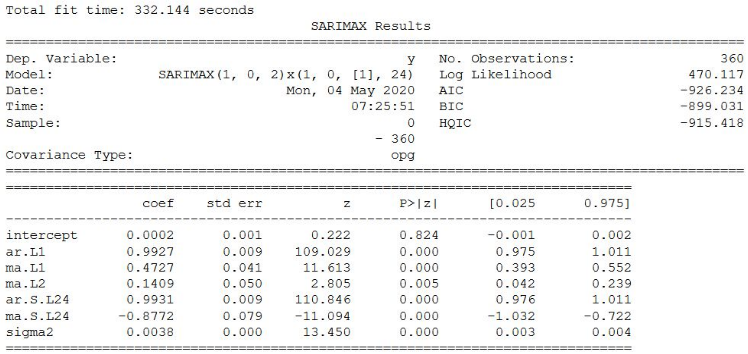
Table 10. Accuracy Metrics

MSE	888.31
MAPE	22.41

Comments and Observations:

1. Around 22.41% MAPE implies that the model is about 77.59 % accurate in predicting the 16-22 week observations (squareid 5160).
2. This squareid is near to the area(squareid 5161) with highest traffic during a two month period. The model performed better for this squareid as compared to squareid 5161 because there is no surge in the Internet Network connections. Also, it is in the city centre of Milan.

Table 11. Statistics on the training and/or execution time



```
Total fit time: 332.144 seconds
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  360
Model:             SARIMAX(1, 0, 2)x(1, 0, [1], 24)   Log Likelihood                 470.117
Date:                            Mon, 04 May 2020   AIC                           -926.234
Time:                                    07:25:51   BIC                           -899.031
Sample:                                         0   HQIC                          -915.418
                                            - 360
Covariance Type:                              opg
==========================================================================================
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept      0.0002      0.001      0.222      0.824      -0.001       0.002
ar.L1          0.9927      0.009    109.029      0.000       0.975       1.011
ma.L1          0.4727      0.041     11.613      0.000       0.393       0.552
ma.L2          0.1409      0.050      2.805      0.005       0.042       0.239
ar.S.L24       0.9931      0.009    110.846      0.000       0.976       1.011
ma.S.L24      -0.8772      0.079    -11.094      0.000      -1.032      -0.722
sigma2         0.0038      0.000     13.450      0.000       0.003       0.004
==============================================================================
```

Table 12. MSE and MAPE (Four Areas)

Squareid	**MSE**	**MAPE**
5161	1582.84	24.59
4159	390.99	30.66
4556	394.12	16.51
5160	888.31	22.41

Conclusions:

1. The Auto-ARIMA Model gives better accuracy and is able to capture the patterns in the areas / squareid where there is less sudden increase/decrease in the Internet network traffic.
2. Squareid 4556: Model performs best and has better accuracy score on Internet network traffic.
3. Squareid 4159: Model performs worst in this case because there is sudden drop in the Internet network traffic and number of connections during the weekends.

Some interesting follow up to explore / Personal recommendations:

1. **Enrich prediction by adding some more features**: Explore more possibilities to incorporate more features/variables such as national holidays, holidays periods (Ex. Christmas & New Year holidays), events etc.
2. **Make model self-learning**: Let the model learn from it's past mistakes in prediction and update it with new observations without rerunning the entire model.
3. **Develop a separate model for weekdays and weekends**: It can be valuable to develop a separate prediction model for the weekdays and weekends in particular, as we know there is a major difference in internet network traffic during the weekdays and weekends in some squareids.
4. **Prediction with other Time Series, Machine Learning and Deep Learning Models**: Finally, it can be interesting to run some state of the art time series, machine learning and deep learning models such as facebook/prophet, XGBoost, ADABoost, GluonTS (Introduced by AWS recently) on this dataset.

Libraries Used:

1. **Numpy :** NumPy
2. **Pandas :** Pandas
3. **Matplotlib :** Matplotlib
4. **Seaborn :** Seaborn
5. **geojson :** geojson
6. **descartes :** descartes
7. **Sklearn :** scikit-learn
8. **Scipy :** SciPy.org
9. **Statsmodels :** statsmodels
10. **pmdarima :** pmdarima

References:

[1] Barlacchi, G., De Nadai, M., Larcher, R. *et al.* A multi-source dataset of urban life in the city of Milan and the Province of Trentino. *Sci Data* 2, 150055 (2015).
url: **https://rdcu.be/b3WmL**

[2] Trinh, Hoang Duy & Bui, Nicola & Widmer, Joerg & Giupponi, L. & Dini, Paolo. (2017). Analysis and Modeling of Mobile Traffic Using Real Traces. 10.1109/PIMRC.2017.8292200.
url: PDF

[3] Aarshay Jain, (2016, Feb 6), A comprehensive beginner's guide to create Time Series Forecast [Blog Post]
url: https://www.analyticsvidhya.com/blog/2016/02/time-series-forecasting-codes-python/

[4] Selva Prabhakaran, (2019, Feb 18), ARIMA Model - Time Series Forecasting [Blog Post]
url: ARIMA

[5] Kaggle: Mobile Activity in a city
url: https://www.kaggle.com/marcodena/mobile-phone-activity/kernels

[6] Raymond Camden, (2019, Sep 4), An Introduction to GeoJSON [Blog Post]
url: https://developer.here.com/blog/an-introduction-to-geojson