NYC_Final

August 24, 2019

1 Taxi demand prediction in New York City

In [107]:
```
#Importing Libraries
!pip3 install graphviz
!pip3 install dask
!pip3 install toolz
!pip install gpxpy
!pip install folium
!pip install keras
#pip3 install cloudpickle
# https://www.youtube.com/watch?v=ieW3G7ZzRZ0
# https://github.com/dask/dask-tutorial
# please do go through this python notebook: https://github.com/dask/dask-tutorial/b
import dask.dataframe as dd#similar to pandas

import pandas as pd#pandas to create small dataframes

# if this doesnt work refere install_folium.JPG in drive
import folium #open street map

# unix time: https://www.unixtimestamp.com/
import datetime #Convert to unix time

import time #Convert to unix time

# if numpy is not installed already : pip3 install numpy
import numpy as np#Do aritmetic operations on arrays

# matplotlib: used to plot graphs
import matplotlib
%matplotlib inline
# matplotlib.use('nbagg') : matplotlib uses this protocall which makes plots more us
matplotlib.use('nbagg')
import matplotlib.pylab as plt
import seaborn as sns#Plots
from matplotlib import rcParams#Size of plots
```

```
# this lib is used while we calculate the stight line distance between two (lat,lon)
import gpxpy.geo #Get the haversine distance

from sklearn.cluster import MiniBatchKMeans, KMeans#Clustering
import math
import pickle
import os

# download migwin: https://mingw-w64.org/doku.php/download/mingw-builds
# install it in your system and keep the path, migw_path ='installed path'
# mingw_path = 'C:\\Program Files\\mingw-w64\\x86_64-5.3.0-posix-seh-rt_v4-rev0\\min
# os.environ['PATH'] = mingw_path + ';' + os.environ['PATH']

!pip3 install xgboost
# if it didnt happen check install_xgboost.JPG
import xgboost as xgb

# to install sklearn: pip install -U scikit-learn
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error
from sklearn.metrics import mean_absolute_error
import scipy
import pdb
import warnings
from sklearn import linear_model

warnings.filterwarnings("ignore")
```

```
Requirement already satisfied: graphviz in /usr/local/lib/python3.6/dist-packages (0.10.1)
Requirement already satisfied: dask in /usr/local/lib/python3.6/dist-packages (1.1.5)
Requirement already satisfied: toolz in /usr/local/lib/python3.6/dist-packages (0.10.0)
Requirement already satisfied: gpxpy in /usr/local/lib/python3.6/dist-packages (1.3.5)
Requirement already satisfied: folium in /usr/local/lib/python3.6/dist-packages (0.8.3)
Requirement already satisfied: requests in /usr/local/lib/python3.6/dist-packages (from folium)
Requirement already satisfied: numpy in /usr/local/lib/python3.6/dist-packages (from folium) (
Requirement already satisfied: six in /usr/local/lib/python3.6/dist-packages (from folium) (1.
Requirement already satisfied: branca>=0.3.0 in /usr/local/lib/python3.6/dist-packages (from f
Requirement already satisfied: jinja2 in /usr/local/lib/python3.6/dist-packages (from folium)
Requirement already satisfied: urllib3<1.25,>=1.21.1 in /usr/local/lib/python3.6/dist-packages
Requirement already satisfied: idna<2.9,>=2.5 in /usr/local/lib/python3.6/dist-packages (from
Requirement already satisfied: chardet<3.1.0,>=3.0.2 in /usr/local/lib/python3.6/dist-packages
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.6/dist-packages (f
Requirement already satisfied: MarkupSafe>=0.23 in /usr/local/lib/python3.6/dist-packages (fro
Requirement already satisfied: keras in /usr/local/lib/python3.6/dist-packages (2.2.4)
Requirement already satisfied: keras-preprocessing>=1.0.5 in /usr/local/lib/python3.6/dist-pac
Requirement already satisfied: six>=1.9.0 in /usr/local/lib/python3.6/dist-packages (from kera
Requirement already satisfied: numpy>=1.9.1 in /usr/local/lib/python3.6/dist-packages (from ke
Requirement already satisfied: pyyaml in /usr/local/lib/python3.6/dist-packages (from keras) (
```

```
Requirement already satisfied: keras-applications>=1.0.6 in /usr/local/lib/python3.6/dist-packa
Requirement already satisfied: h5py in /usr/local/lib/python3.6/dist-packages (from keras) (2.8
Requirement already satisfied: scipy>=0.14 in /usr/local/lib/python3.6/dist-packages (from kera
Requirement already satisfied: xgboost in /usr/local/lib/python3.6/dist-packages (0.90)
Requirement already satisfied: numpy in /usr/local/lib/python3.6/dist-packages (from xgboost)
Requirement already satisfied: scipy in /usr/local/lib/python3.6/dist-packages (from xgboost)
```

2 Data Information

Ge the data from : http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml (2016 data) The data used in the attached datasets were collected and provided to the NYC Taxi and Limousine Commission (TLC)

```
In [2]: # Load the Drive helper and mount
        from google.colab import drive
        drive.mount('/content/drive')
```

```
Go to this URL in a browser: https://accounts.google.com/o/oauth2/auth?client_id=947318989803-6

Enter your authorization code:
ůůůůůůůůůů
Mounted at /content/drive
```

```
In [0]: import os
        os.chdir("/content/drive/My Drive/Uber")
```

2.1 Information on taxis:

Yellow Taxi: Yellow Medallion Taxicabs

These are the famous NYC yellow taxis that provide transportation exclusively through street-hails. The number of taxicabs is limited by a finite number of medallions issued by the TLC. You access this mode of transportation by standing in the street and hailing an available taxi with your hand. The pickups are not pre-arranged.

For Hire Vehicles (FHVs)

FHV transportation is accessed by a pre-arrangement with a dispatcher or limo company. These FHVs are not permitted to pick up passengers via street hails, as those rides are not considered pre-arranged.

Green Taxi: Street Hail Livery (SHL)

The SHL program will allow livery vehicle owners to license and outfit their vehicles with green borough taxi branding, meters, credit card machines, and ultimately the right to accept street hails in addition to pre-arranged rides.

Credits: Quora

Footnote:

In the given notebook we are considering only the yellow taxis for the time period between Jan - Mar 2015 & Jan - Mar 2016

3 Data Collection

We Have collected all yellow taxi trips data from jan-2015 to dec-2016(Will be using only 2015 data)

file name
file name size
number of records
number of features
yellow_tripdata_2016-01

1. 59G
10906858
19

yellow_tripdata_2016-02

1. 66G
11382049
19
yellow_tripdata_2016-03
   1. 78G
   12210952
   19
   yellow_tripdata_2016-04
      1. 74G
      11934338
      19

yellow_tripdata_2016-05

1. 73G
11836853
19

yellow_tripdata_2016-06

1. 62G
11135470
19

yellow_tripdata_2016-07
884Mb
10294080
17
yellow_tripdata_2016-08
854Mb
9942263
17

yellow_tripdata_2016-09
870Mb
10116018
17
yellow_tripdata_2016-10
933Mb
10854626
17
yellow_tripdata_2016-11
868Mb
10102128
17
yellow_tripdata_2016-12
897Mb
10449408
17
yellow_tripdata_2015-01
1.84Gb
12748986
19
yellow_tripdata_2015-02
1.81Gb
12450521
19
yellow_tripdata_2015-03
1.94Gb
13351609
19
yellow_tripdata_2015-04
1.90Gb
13071789
19
yellow_tripdata_2015-05
1.91Gb
13158262
19
yellow_tripdata_2015-06
1.79Gb
12324935
19
yellow_tripdata_2015-07
1.68Gb
11562783
19
yellow_tripdata_2015-08
1.62Gb
11130304
19

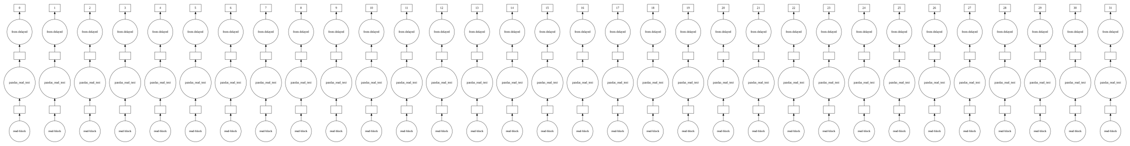
yellow_tripdata_2015-09
1.63Gb
11225063
19
yellow_tripdata_2015-10
1.79Gb
12315488
19
yellow_tripdata_2015-11
1.65Gb
11312676
19
yellow_tripdata_2015-12
1.67Gb
11460573
19

In [4]:
```
#Looking at the features
# dask dataframe  : # https://github.com/dask/dask-tutorial/blob/master/07_dataframe.i
month = dd.read_csv('yellow_tripdata_2015-01.csv')
print(month.columns)
```

```
Index(['VendorID', 'tpep_pickup_datetime', 'tpep_dropoff_datetime',
       'passenger_count', 'trip_distance', 'pickup_longitude',
       'pickup_latitude', 'RateCodeID', 'store_and_fwd_flag',
       'dropoff_longitude', 'dropoff_latitude', 'payment_type', 'fare_amount',
       'extra', 'mta_tax', 'tip_amount', 'tolls_amount',
       'improvement_surcharge', 'total_amount'],
      dtype='object')
```

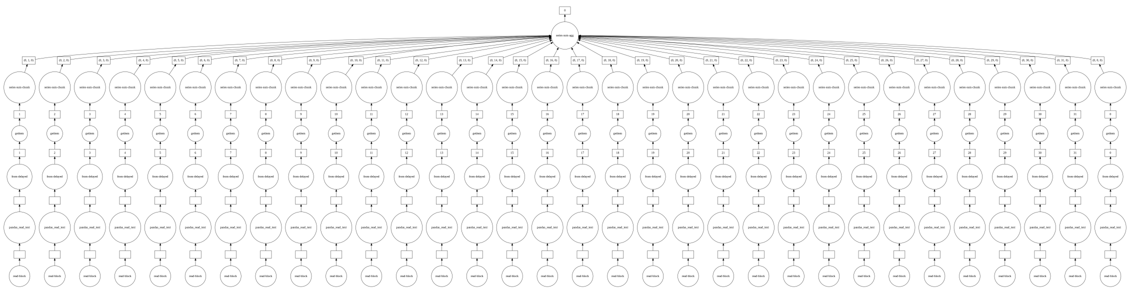
In [5]:
```
# However unlike Pandas, operations on dask.dataframes don't trigger immediate computa
# instead they add key-value pairs to an underlying Dask graph. Recall that in the dia
# circles are operations and rectangles are results.

# to see the visulaization you need to install graphviz
# pip3 install graphviz if this doesnt work please check the install_graphviz.jpg in t
month.visualize()
```

Out[5]:



```python
In [6]: month.fare_amount.sum().visualize()
```

Out[6]:



3.1 Features in the dataset:

```
<tr>
    <th>Field Name</th>
    <th>Description</th>
</tr>
<tr>
    <td>VendorID</td>
    <td>
    A code indicating the TPEP provider that provided the record.
    <ol>
        <li>Creative Mobile Technologies</li>
        <li>VeriFone Inc.</li>
    </ol>
    </td>
</tr>
<tr>
    <td>tpep_pickup_datetime</td>
    <td>The date and time when the meter was engaged.</td>
</tr>
<tr>
    <td>tpep_dropoff_datetime</td>
    <td>The date and time when the meter was disengaged.</td>
</tr>
<tr>
    <td>Passenger_count</td>
    <td>The number of passengers in the vehicle. This is a driver-entered value.</td>
</tr>
<tr>
    <td>Trip_distance</td>
    <td>The elapsed trip distance in miles reported by the taximeter.</td>
</tr>
<tr>
```

```
    <td>Pickup_longitude</td>
    <td>Longitude where the meter was engaged.</td>
</tr>
<tr>
    <td>Pickup_latitude</td>
    <td>Latitude where the meter was engaged.</td>
</tr>
<tr>
    <td>RateCodeID</td>
    <td>The final rate code in effect at the end of the trip.
    <ol>
        <li> Standard rate </li>
        <li> JFK </li>
        <li> Newark </li>
        <li> Nassau or Westchester</li>
        <li> Negotiated fare </li>
        <li> Group ride</li>
    </ol>
    </td>
</tr>
<tr>
    <td>Store_and_fwd_flag</td>
    <td>This flag indicates whether the trip record was held in vehicle memory before sending
    <br\>Y= store and forward trip
    <br\>N= not a store and forward trip
    </td>
</tr>

<tr>
    <td>Dropoff_longitude</td>
    <td>Longitude where the meter was disengaged.</td>
</tr>
<tr>
    <td>Dropoff_ latitude</td>
    <td>Latitude where the meter was disengaged.</td>
</tr>
<tr>
    <td>Payment_type</td>
    <td>A numeric code signifying how the passenger paid for the trip.
    <ol>
        <li> Credit card </li>
        <li> Cash </li>
        <li> No charge </li>
        <li> Dispute</li>
        <li> Unknown </li>
        <li> Voided trip</li>
    </ol>
    </td>
```

```
</tr>
<tr>
    <td>Fare_amount</td>
    <td>The time-and-distance fare calculated by the meter.</td>
</tr>
<tr>
    <td>Extra</td>
    <td>Miscellaneous extras and surcharges. Currently, this only includes. the $0.50 and $1 ru
</tr>
<tr>
    <td>MTA_tax</td>
    <td>0.50 MTA tax that is automatically triggered based on the metered rate in use.</td>
</tr>
<tr>
    <td>Improvement_surcharge</td>
    <td>0.30 improvement surcharge assessed trips at the flag drop. the improvement surcharge b
</tr>
<tr>
    <td>Tip_amount</td>
    <td>Tip amount  This field is automatically populated for credit card tips.Cash tips are no
</tr>
<tr>
    <td>Tolls_amount</td>
    <td>Total amount of all tolls paid in trip.</td>
</tr>
<tr>
    <td>Total_amount</td>
    <td>The total amount charged to passengers. Does not include cash tips.</td>
</tr>
```

4 ML Problem Formulation

Time-series forecasting and Regression

- To find number of pickups, given location cordinates(latitude and longitude) and time, in the query reigion and surrounding regions.

To solve the above we would be using data collected in Jan - Mar 2015 to predict the pickups in Jan - Mar 2016.

5 Performance metrics

1. Mean Absolute percentage error.
2. Mean Squared error.

5.1 Data Cleaning

In this section we will be doing univariate analysis and removing outlier/illegitimate values which may be caused due to some error

```
In [7]: #table below shows few datapoints along with all our features
        month.head(5)

Out[7]:    VendorID tpep_pickup_datetime  ... improvement_surcharge  total_amount
        0         2  2015-01-15 19:05:39  ...                   0.3         17.05
        1         1  2015-01-10 20:33:38  ...                   0.3         17.80
        2         1  2015-01-10 20:33:38  ...                   0.3         10.80
        3         1  2015-01-10 20:33:39  ...                   0.3          4.80
        4         1  2015-01-10 20:33:39  ...                   0.3         16.30

        [5 rows x 19 columns]
```

5.1.1 1. Pickup Latitude and Pickup Longitude

It is inferred from the source https://www.flickr.com/places/info/2459115 that New York is bounded by the location cordinates(lat,long) - (40.5774, -74.15) & (40.9176,-73.7004) so hence any cordinates not within these cordinates are not considered by us as we are only concerned with pickups which originate within New York.

```
In [8]: # Plotting pickup cordinates which are outside the bounding box of New-York
        # we will collect all the points outside the bounding box of newyork city to outlier_l
        outlier_locations = month[((month.pickup_longitude <= -74.15) | (month.pickup_latitude
                           (month.pickup_longitude >= -73.7004) | (month.pickup_latitude >= 40

        # creating a map with the a base location
        # read more about the folium here: http://folium.readthedocs.io/en/latest/quickstart.h

        # note: you dont need to remember any of these, you dont need indeepth knowledge on th

        map_osm = folium.Map(location=[40.734695, -73.990372], tiles='Stamen Toner')

        # we will spot only first 100 outliers on the map, plotting all the outliers will take
        sample_locations = outlier_locations.head(10000)
        for i,j in sample_locations.iterrows():
            if int(j['pickup_latitude']) != 0:
                folium.Marker(list((j['pickup_latitude'],j['pickup_longitude']))).add_to(map_o
        map_osm

Out[8]: <folium.folium.Map at 0x7fe2d8349fd0>
```

Observation:- As you can see above that there are some points just outside the boundary but there are a few that are in either South america, Mexico or Canada

5.1.2 2. Dropoff Latitude & Dropoff Longitude

It is inferred from the source https://www.flickr.com/places/info/2459115 that New York is bounded by the location cordinates(lat,long) - (40.5774, -74.15) & (40.9176,-73.7004) so hence any cordinates not within these cordinates are not considered by us as we are only concerned with dropoffs which are within New York.

```
In [9]: # Plotting dropoff cordinates which are outside the bounding box of New-York
        # we will collect all the points outside the bounding box of newyork city to outlier_l
        outlier_locations = month[((month.dropoff_longitude <= -74.15) | (month.dropoff_latitud
                           (month.dropoff_longitude >= -73.7004) | (month.dropoff_latitude >= 

        # creating a map with the a base location
        # read more about the folium here: http://folium.readthedocs.io/en/latest/quickstart.h

        # note: you dont need to remember any of these, you dont need indeepth knowledge on th

        map_osm = folium.Map(location=[40.734695, -73.990372], tiles='Stamen Toner')

        # we will spot only first 100 outliers on the map, plotting all the outliers will take
        sample_locations = outlier_locations.head(10000)
        for i,j in sample_locations.iterrows():
            if int(j['pickup_latitude']) != 0:
                folium.Marker(list((j['dropoff_latitude'],j['dropoff_longitude']))).add_to(map_
        map_osm

Out[9]: <folium.folium.Map at 0x7fe2d8349748>
```

Observation:- The observations here are similar to those obtained while analysing pickup latitude and longitude

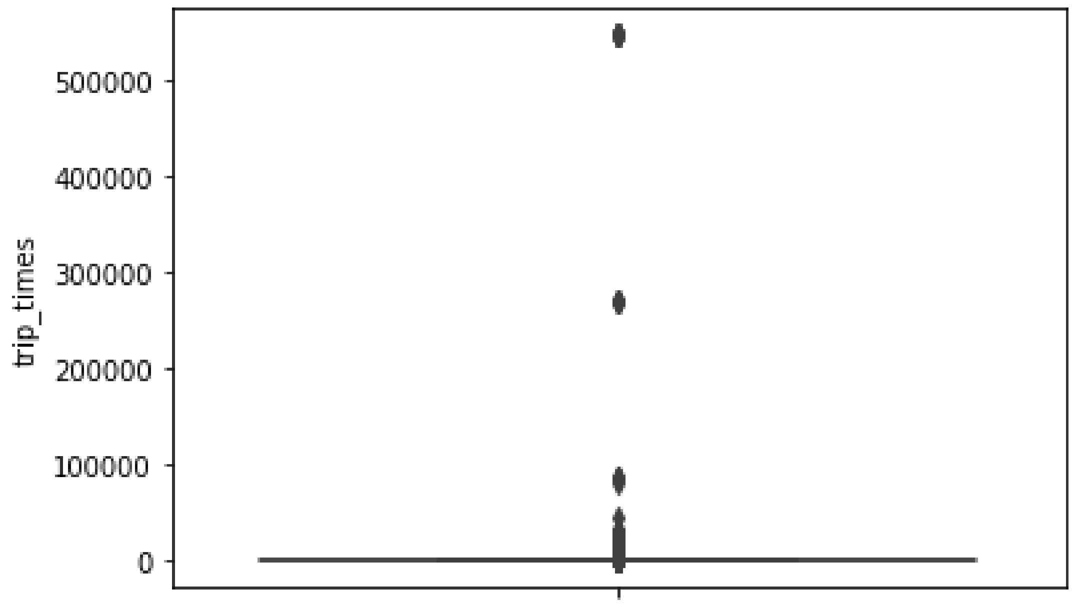
5.1.3 3. Trip Durations:

According to NYC Taxi & Limousine Commision Regulations the maximum allowed trip duration in a 24 hour interval is 12 hours.

```
In [0]: #The timestamps are converted to unix so as to get duration(trip-time) & speed also pi

        # in out data we have time in the formate "YYYY-MM-DD HH:MM:SS" we convert thiss sting
        # https://stackoverflow.com/a/27914405
        def convert_to_unix(s):
            return time.mktime(datetime.datetime.strptime(s, "%Y-%m-%d %H:%M:%S").timetuple())



        # we return a data frame which contains the columns
        # 1.'passenger_count' : self explanatory
        # 2.'trip_distance' : self explanatory
        # 3.'pickup_longitude' : self explanatory
        # 4.'pickup_latitude' : self explanatory
        # 5.'dropoff_longitude' : self explanatory
        # 6.'dropoff_latitude' : self explanatory
        # 7.'total_amount' : total fair that was paid
        # 8.'trip_times' : duration of each trip
        # 9.'pickup_times : pickup time converted into unix time
        # 10.'Speed' : velocity of each trip
```

```
def return_with_trip_times(month):
    duration = month[['tpep_pickup_datetime','tpep_dropoff_datetime']].compute()
    #pickups and dropoffs to unix time
    duration_pickup = [convert_to_unix(x) for x in duration['tpep_pickup_datetime'].val
    duration_drop = [convert_to_unix(x) for x in duration['tpep_dropoff_datetime'].valu
    #calculate duration of trips
    durations = (np.array(duration_drop) - np.array(duration_pickup))/float(60)

    #append durations of trips and speed in miles/hr to a new dataframe
    new_frame = month[['passenger_count','trip_distance','pickup_longitude','pickup_la

    new_frame['trip_times'] = durations
    new_frame['pickup_times'] = duration_pickup
    new_frame['Speed'] = 60*(new_frame['trip_distance']/new_frame['trip_times'])

    return new_frame

# print(frame_with_durations.head())
#  passenger_count        trip_distance        pickup_longitude        pickup_latitude
#   1                  1.59              -73.993896              40.750111
#   1                       3.30              -74.001648              40.724243
#   1                       1.80               -73.963341              40.802788
#   1                       0.50              -74.009087              40.713818
#   1                       3.00              -73.971176              40.762428
frame_with_durations = return_with_trip_times(month)
```

In [11]:
```
%matplotlib inline


# the skewed box plot shows us the presence of outliers
sns.boxplot(y="trip_times", data =frame_with_durations)
plt.show()
```




```
In [12]: #calculating 0-100th percentile to find a the correct percentile value for removal of
         for i in range(0,100,10):
             var =frame_with_durations["trip_times"].values
             var = np.sort(var,axis = None)
             print("{} percentile value is {}".format(i,var[int(len(var)*(float(i)/100))]))
         print ("100 percentile value is ",var[-1])
```

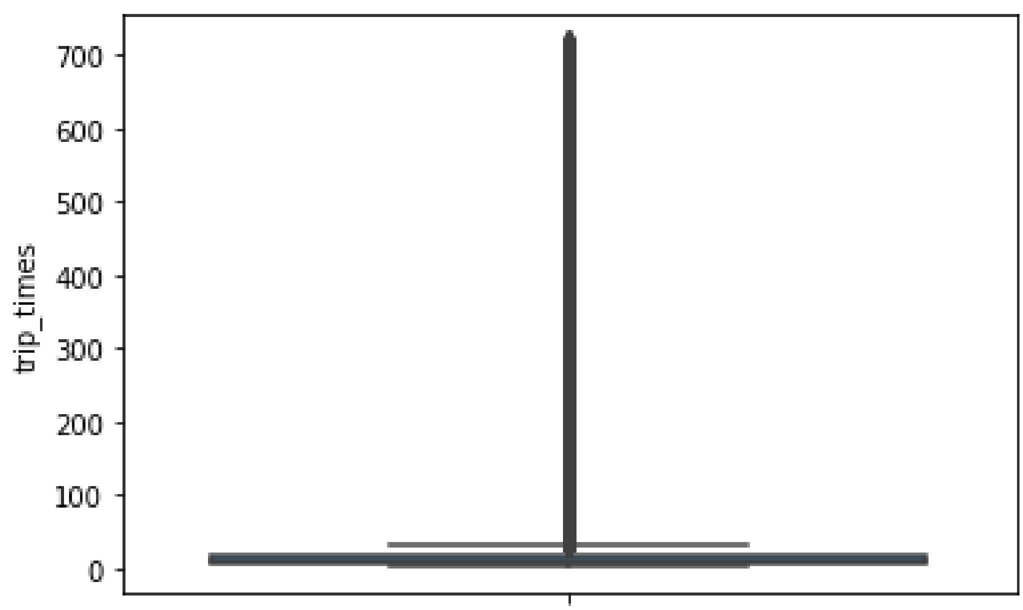
```
0 percentile value is -1211.0166666666667
10 percentile value is 3.8333333333333335
20 percentile value is 5.383333333333334
30 percentile value is 6.816666666666666
40 percentile value is 8.3
50 percentile value is 9.95
60 percentile value is 11.866666666666667
70 percentile value is 14.283333333333333
80 percentile value is 17.633333333333333
90 percentile value is 23.45
100 percentile value is  548555.6333333333
```

```
In [13]: #looking further from the 99th percecntile
         for i in range(90,100):
             var =frame_with_durations["trip_times"].values
             var = np.sort(var,axis = None)
             print("{} percentile value is {}".format(i,var[int(len(var)*(float(i)/100))]))
         print ("100 percentile value is ",var[-1])
```

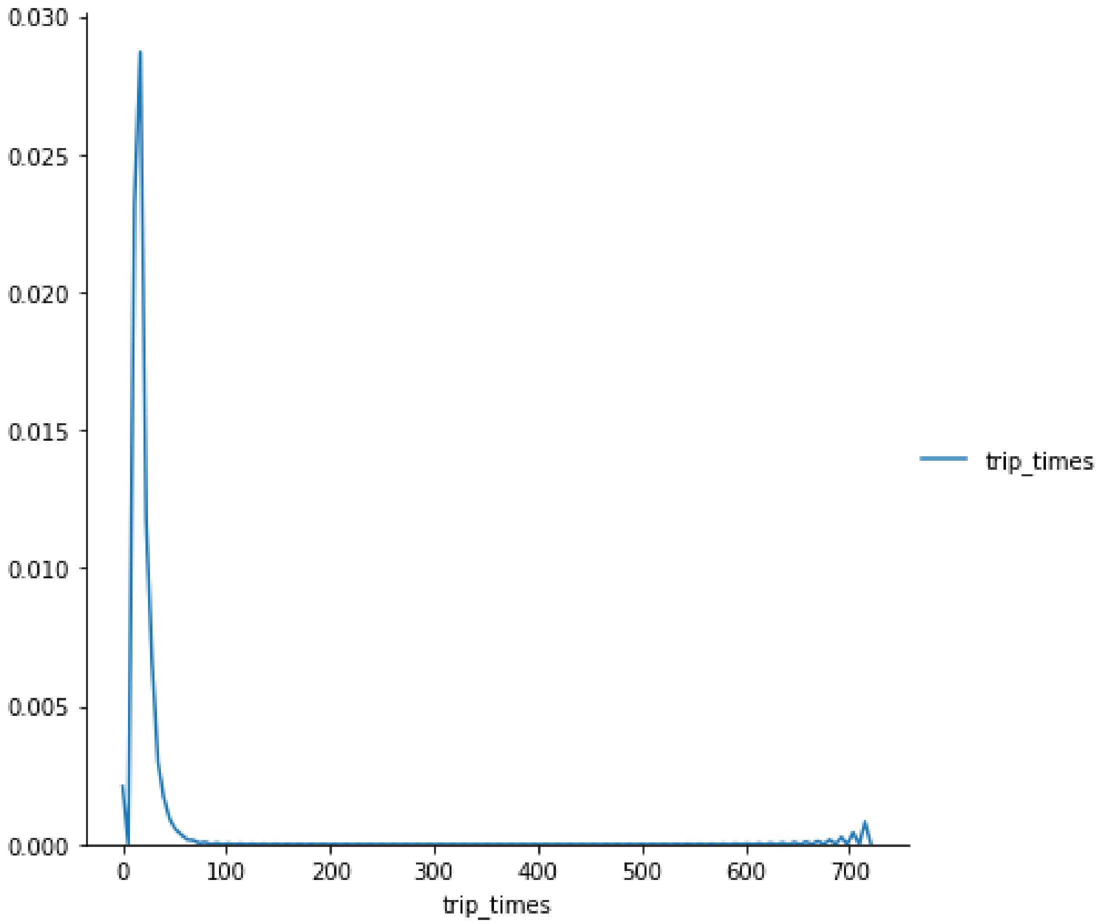
```
90 percentile value is 23.45
91 percentile value is 24.35
92 percentile value is 25.383333333333333
93 percentile value is 26.55
94 percentile value is 27.933333333333334
95 percentile value is 29.583333333333332
96 percentile value is 31.683333333333334
97 percentile value is 34.46666666666667
98 percentile value is 38.71666666666667
99 percentile value is 46.75
100 percentile value is  548555.6333333333
```

In [0]:
```
#removing data based on our analysis and TLC regulations
frame_with_durations_modified=frame_with_durations[(frame_with_durations.trip_times>1)
```

In [15]:
```
%matplotlib inline


#box-plot after removal of outliers
sns.boxplot(y="trip_times", data =frame_with_durations_modified)
plt.show()
```

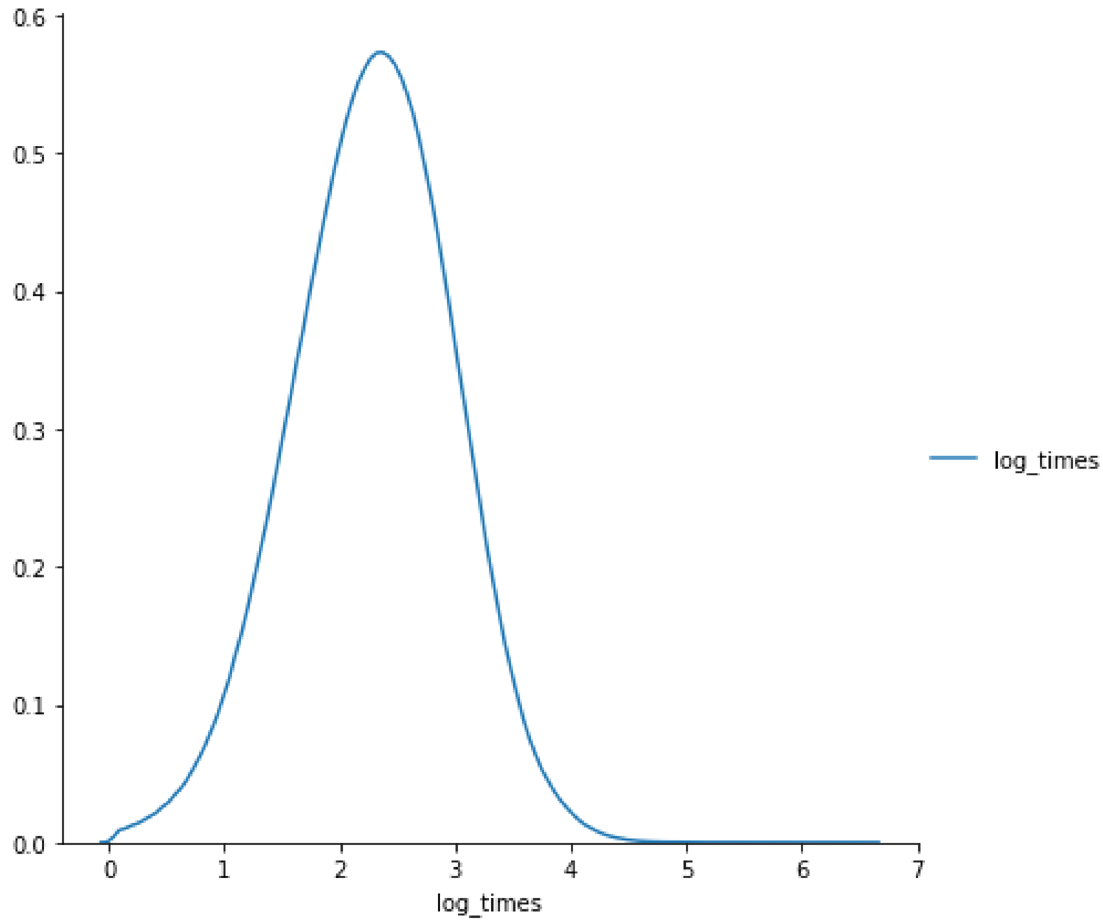


In [16]:
```
#pdf of trip-times after removing the outliers
sns.FacetGrid(frame_with_durations_modified,size=6) \
      .map(sns.kdeplot,"trip_times") \
      .add_legend();
plt.show();
```

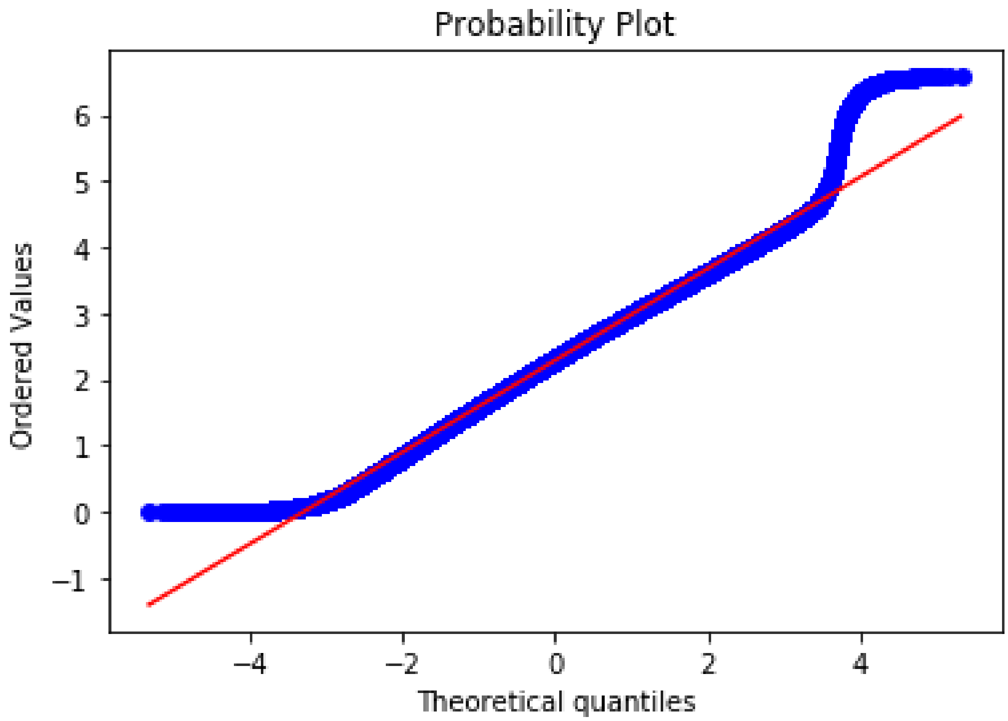


In [0]:
```
#converting the values to log-values to chec for log-normal
import math
frame_with_durations_modified['log_times']=[math.log(i) for i in frame_with_durations_m
```

In [18]:
```
#pdf of log-values
sns.FacetGrid(frame_with_durations_modified,size=6) \
      .map(sns.kdeplot,"log_times") \
      .add_legend();
plt.show();
```

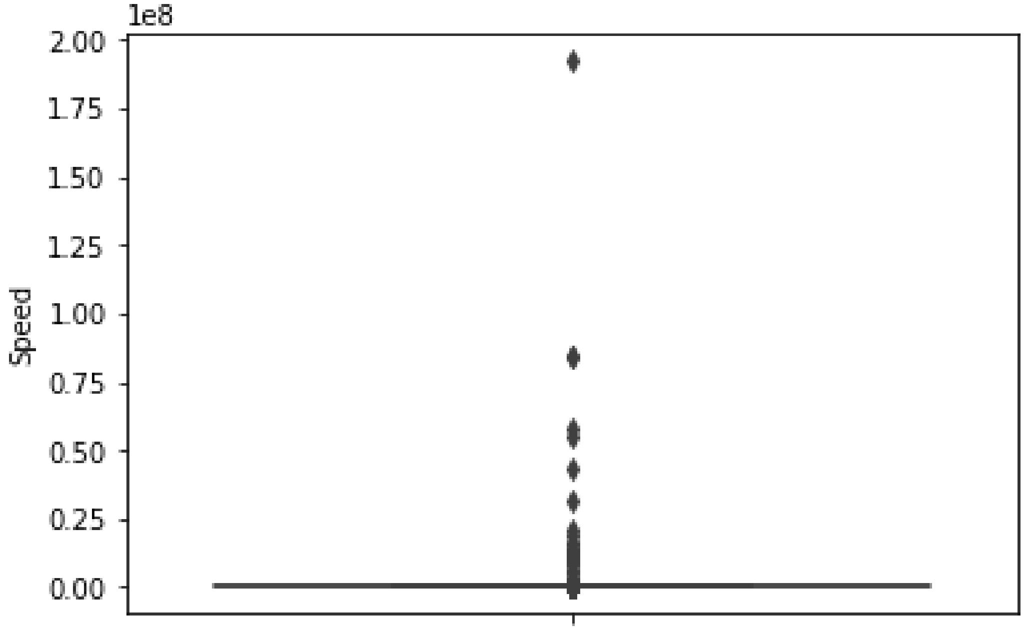



```
In [19]: #Q-Q plot for checking if trip-times is log-normal
         scipy.stats.probplot(frame_with_durations_modified['log_times'].values, plot=plt)
         plt.show()
```



5.1.4 4. Speed

```
In [20]: # check for any outliers in the data after trip duration outliers removed
         # box-plot for speeds with outliers
         frame_with_durations_modified['Speed'] = 60*(frame_with_durations_modified['trip_dista
         sns.boxplot(y="Speed", data =frame_with_durations_modified)
         plt.show()
```



In [21]:
```
#calculating speed values at each percntile 0,10,20,30,40,50,60,70,80,90,100
for i in range(0,100,10):
    var =frame_with_durations_modified["Speed"].values
    var = np.sort(var,axis = None)
    print("{} percentile value is {}".format(i,var[int(len(var)*(float(i)/100))]))
print("100 percentile value is ",var[-1])
```

```
0 percentile value is 0.0
10 percentile value is 6.409495548961425
20 percentile value is 7.80952380952381
30 percentile value is 8.929133858267717
40 percentile value is 9.98019801980198
50 percentile value is 11.06865671641791
60 percentile value is 12.286689419795222
70 percentile value is 13.796407185628745
80 percentile value is 15.963224893917962
90 percentile value is 20.186915887850468
100 percentile value is  192857142.85714284
```

In [22]:
```
#calculating speed values at each percntile 90,91,92,93,94,95,96,97,98,99,100
for i in range(90,100):
    var =frame_with_durations_modified["Speed"].values
    var = np.sort(var,axis = None)
    print("{} percentile value is {}".format(i,var[int(len(var)*(float(i)/100))]))
print("100 percentile value is ",var[-1])
```

```
90 percentile value is 20.186915887850468
91 percentile value is 20.91645569620253
92 percentile value is 21.752988047808763
93 percentile value is 22.721893491124263
94 percentile value is 23.844155844155843
95 percentile value is 25.182552504038775
96 percentile value is 26.80851063829787
97 percentile value is 28.84304932735426
98 percentile value is 31.591128254580514
99 percentile value is 35.7513566847558
100 percentile value is  192857142.85714284
```

```
In [23]: #calculating speed values at each percntile 99.0,99.1,99.2,99.3,99.4,99.5,99.6,99.7,9
         for i in np.arange(0.0, 1.0, 0.1):
             var =frame_with_durations_modified["Speed"].values
             var = np.sort(var,axis = None)
             print("{} percentile value is {}".format(99+i,var[int(len(var)*(float(99+i)/100))]
         print("100 percentile value is ",var[-1])
```

```
99.0 percentile value is 35.7513566847558
99.1 percentile value is 36.31084727468969
99.2 percentile value is 36.91470054446461
99.3 percentile value is 37.588235294117645
99.4 percentile value is 38.33035714285714
99.5 percentile value is 39.17580340264651
99.6 percentile value is 40.15384615384615
99.7 percentile value is 41.338301043219076
99.8 percentile value is 42.86631016042781
99.9 percentile value is 45.3107822410148
100 percentile value is  192857142.85714284
```

```
In [0]: #removing further outliers based on the 99.9th percentile value
        frame_with_durations_modified=frame_with_durations[(frame_with_durations.Speed>0) & (f
```

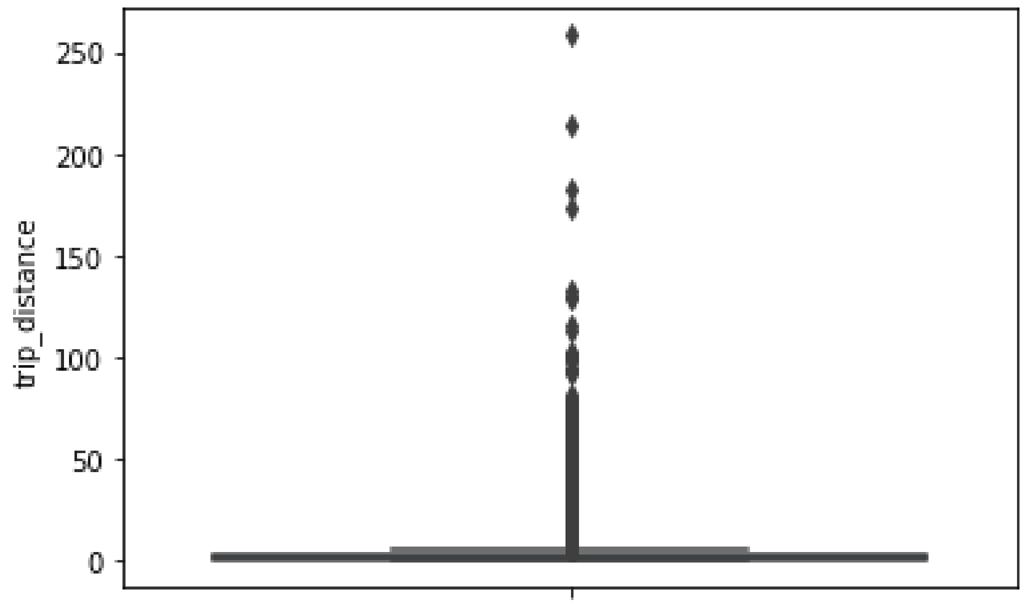
```
In [25]: #avg.speed of cabs in New-York
         sum(frame_with_durations_modified["Speed"]) / float(len(frame_with_durations_modified
```

```
Out[25]: 12.450173996027528
```

The avg speed in Newyork speed is 12.45miles/hr, so a cab driver can travel 2 miles per 10min on avg.

5.1.5 4. Trip Distance

```
In [26]: # up to now we have removed the outliers based on trip durations and cab speeds
         # lets try if there are any outliers in trip distances
         # box-plot showing outliers in trip-distance values
         sns.boxplot(y="trip_distance", data =frame_with_durations_modified)
         plt.show()
```



In [27]:
```
#calculating trip distance values at each percntile 0,10,20,30,40,50,60,70,80,90,100
for i in range(0,100,10):
    var =frame_with_durations_modified["trip_distance"].values
    var = np.sort(var,axis = None)
    print("{} percentile value is {}".format(i,var[int(len(var)*(float(i)/100))]))
print("100 percentile value is ",var[-1])
```

```
0 percentile value is 0.01
10 percentile value is 0.66
20 percentile value is 0.9
30 percentile value is 1.1
40 percentile value is 1.39
50 percentile value is 1.69
60 percentile value is 2.07
70 percentile value is 2.6
80 percentile value is 3.6
90 percentile value is 5.97
100 percentile value is  258.9
```

In [28]:
```
#calculating trip distance values at each percntile 90,91,92,93,94,95,96,97,98,99,100
for i in range(90,100):
    var =frame_with_durations_modified["trip_distance"].values
    var = np.sort(var,axis = None)
    print("{} percentile value is {}".format(i,var[int(len(var)*(float(i)/100))]))
print("100 percentile value is ",var[-1])
```

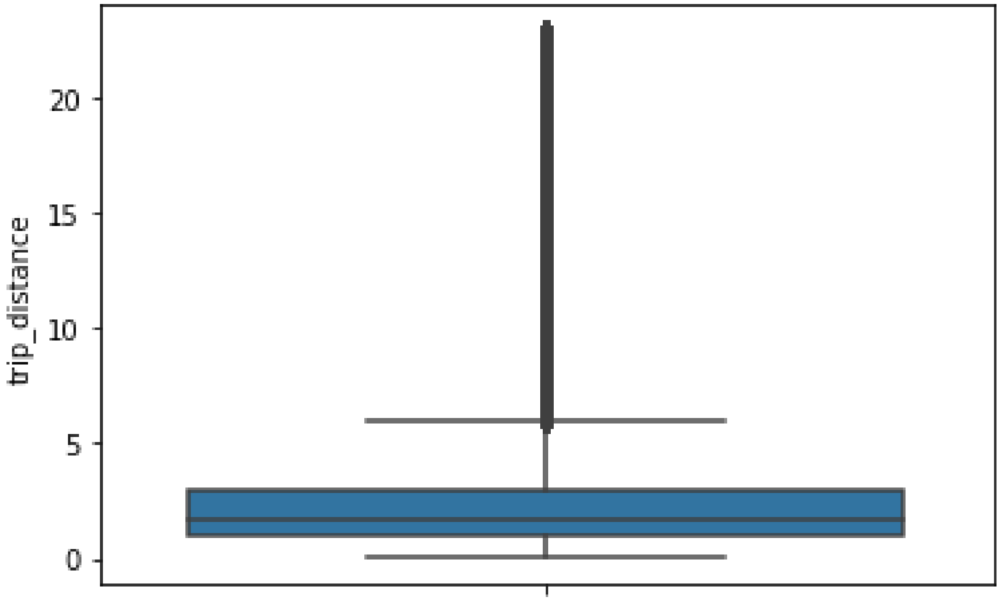
```
90 percentile value is 5.97
91 percentile value is 6.45
92 percentile value is 7.07
93 percentile value is 7.85
94 percentile value is 8.72
95 percentile value is 9.6
96 percentile value is 10.6
97 percentile value is 12.1
98 percentile value is 16.03
99 percentile value is 18.17
100 percentile value is  258.9
```

In [29]:
```
#calculating trip distance values at each percntile 99.0,99.1,99.2,99.3,99.4,99.5,99.
for i in np.arange(0.0, 1.0, 0.1):
    var =frame_with_durations_modified["trip_distance"].values
    var = np.sort(var,axis = None)
    print("{} percentile value is {}".format(99+i,var[int(len(var)*(float(99+i)/100))]
print("100 percentile value is ",var[-1])
```

```
99.0 percentile value is 18.17
99.1 percentile value is 18.37
99.2 percentile value is 18.6
99.3 percentile value is 18.83
99.4 percentile value is 19.13
99.5 percentile value is 19.5
99.6 percentile value is 19.96
99.7 percentile value is 20.5
99.8 percentile value is 21.22
99.9 percentile value is 22.57
100 percentile value is  258.9
```

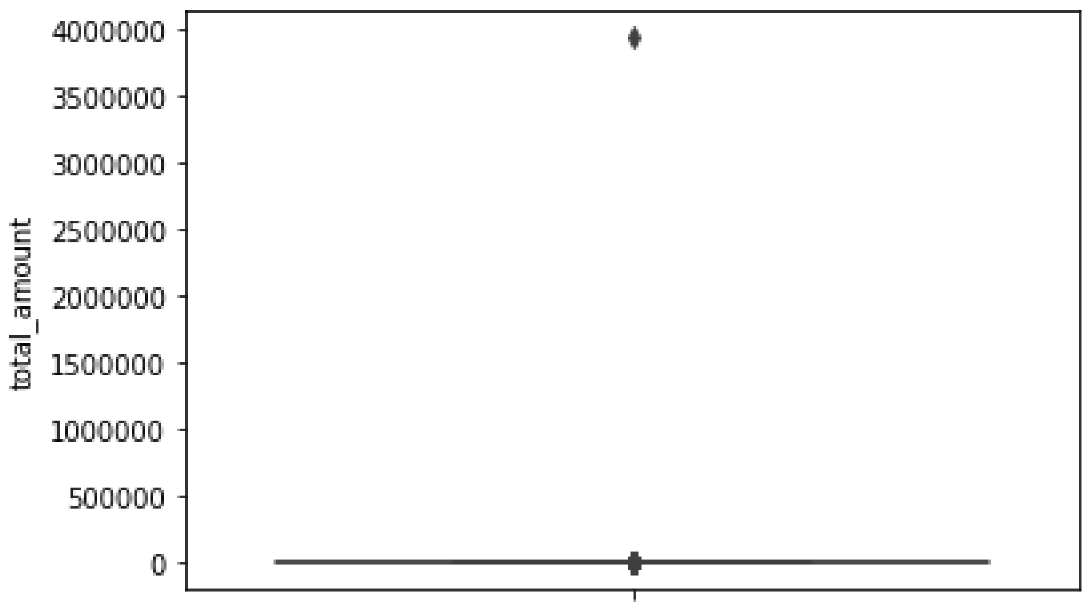
In [0]:
```
#removing further outliers based on the 99.9th percentile value
frame_with_durations_modified=frame_with_durations[(frame_with_durations.trip_distance
```

In [31]:
```
#box-plot after removal of outliers
sns.boxplot(y="trip_distance", data = frame_with_durations_modified)
plt.show()
```



5.1.6 5. Total Fare

In [32]:
```
# up to now we have removed the outliers based on trip durations, cab speeds, and tri
# lets try if there are any outliers in based on the total_amount
# box-plot showing outliers in fare
sns.boxplot(y="total_amount", data =frame_with_durations_modified)
plt.show()
```



In [33]:
```
#calculating total fare amount values at each percntile 0,10,20,30,40,50,60,70,80,90,
for i in range(0,100,10):
    var = frame_with_durations_modified["total_amount"].values
    var = np.sort(var,axis = None)
    print("{} percentile value is {}".format(i,var[int(len(var)*(float(i)/100))]))
print("100 percentile value is ",var[-1])
```

```
0 percentile value is -242.55
10 percentile value is 6.3
20 percentile value is 7.8
30 percentile value is 8.8
40 percentile value is 9.8
50 percentile value is 11.16
60 percentile value is 12.8
70 percentile value is 14.8
80 percentile value is 18.3
90 percentile value is 25.8
100 percentile value is  3950611.6
```

In [34]:
```
#calculating total fare amount values at each percntile 90,91,92,93,94,95,96,97,98,99
for i in range(90,100):
    var = frame_with_durations_modified["total_amount"].values
    var = np.sort(var,axis = None)
    print("{} percentile value is {}".format(i,var[int(len(var)*(float(i)/100))]))
print("100 percentile value is ",var[-1])
```

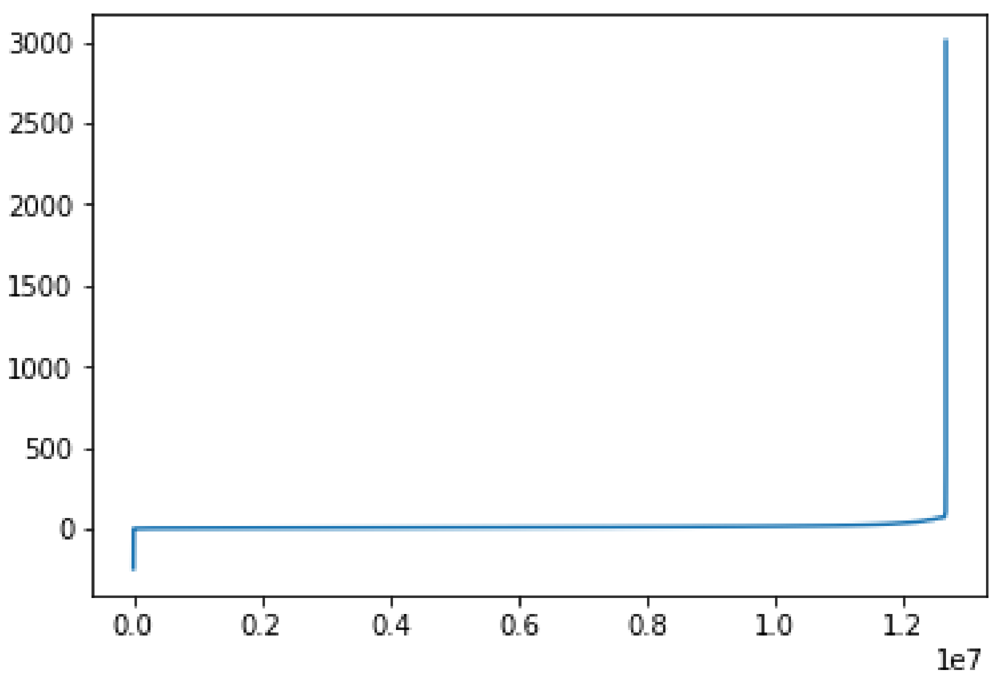
```
90 percentile value is 25.8
91 percentile value is 27.3
92 percentile value is 29.3
93 percentile value is 31.8
94 percentile value is 34.8
95 percentile value is 38.53
96 percentile value is 42.6
97 percentile value is 48.13
98 percentile value is 58.13
99 percentile value is 66.13
100 percentile value is  3950611.6
```

In [35]:
```
#calculating total fare amount values at each percntile 99.0,99.1,99.2,99.3,99.4,99.5
for i in np.arange(0.0, 1.0, 0.1):
    var = frame_with_durations_modified["total_amount"].values
    var = np.sort(var,axis = None)
    print("{} percentile value is {}".format(99+i,var[int(len(var)*(float(99+i)/100))]
print("100 percentile value is ",var[-1])
```

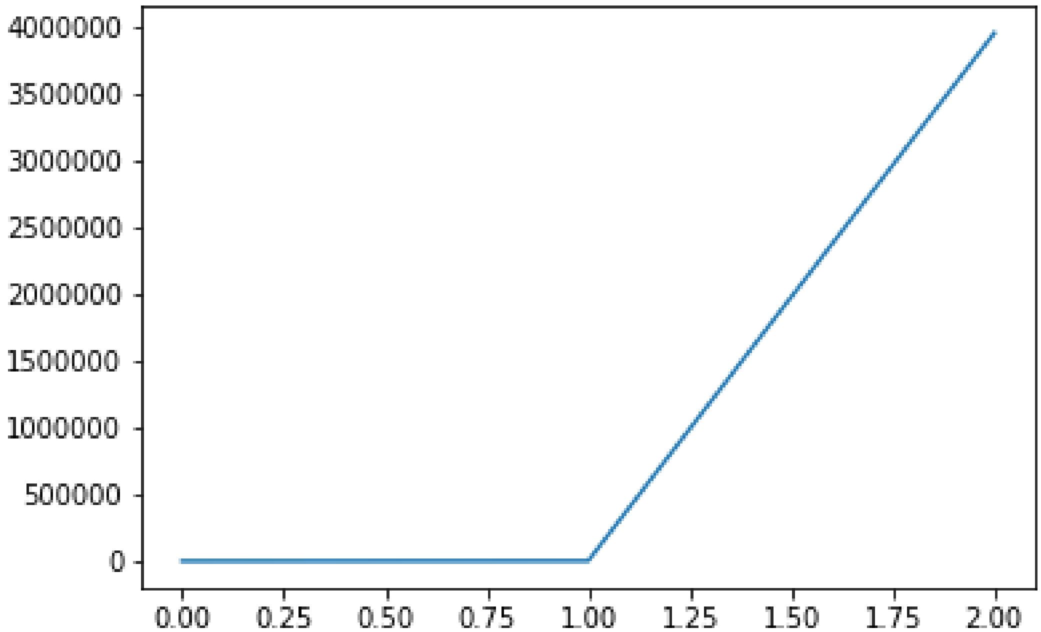
```
99.0 percentile value is 66.13
99.1 percentile value is 68.13
99.2 percentile value is 69.6
99.3 percentile value is 69.6
99.4 percentile value is 69.73
99.5 percentile value is 69.75
99.6 percentile value is 69.76
99.7 percentile value is 72.58
99.8 percentile value is 75.35
99.9 percentile value is 88.28
100 percentile value is  3950611.6
```

Observation:- As even the 99.9th percentile value doesnt look like an outlier,as there is not much difference between the 99.8th percentile and 99.9th percentile, we move on to do graphical analyis

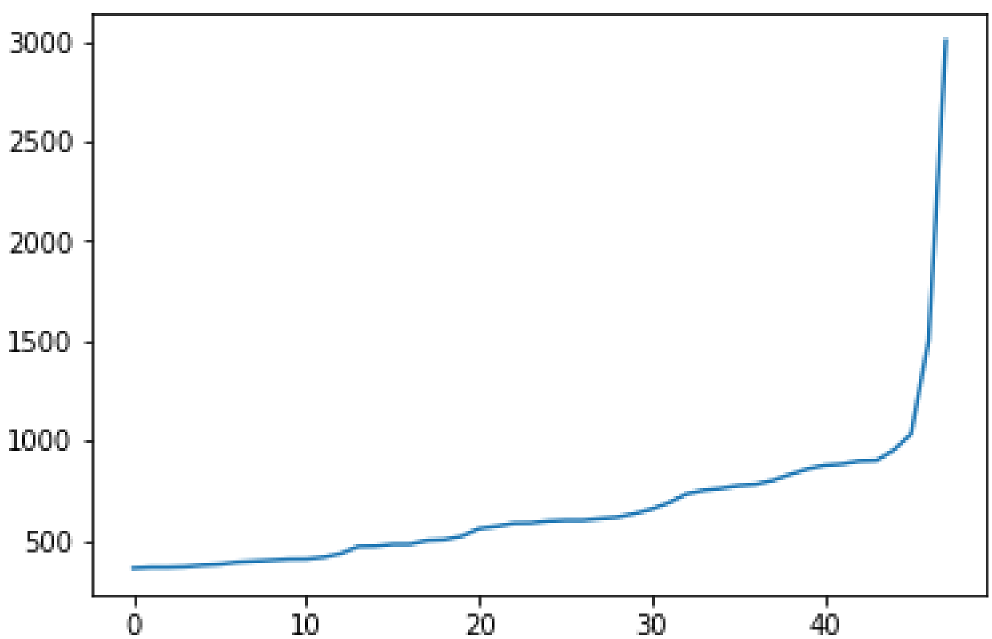
In [36]:
```
#below plot shows us the fare values(sorted) to find a sharp increase to remove those
# plot the fare amount excluding last two values in sorted data
plt.plot(var[:-2])
plt.show()
```



In [37]:
```
# a very sharp increase in fare values can be seen
# plotting last three total fare values, and we can observe there is share increase i
plt.plot(var[-3:])
plt.show()
```




```
In [38]: #now looking at values not including the last two points we again find a drastic incr
         # we plot last 50 values excluding last two values
         plt.plot(var[-50:-2])
         plt.show()
```



5.2 Remove all outliers/erronous points.

```
In [0]: #removing all outliers based on our univariate analysis above
        def remove_outliers(new_frame):


            a = new_frame.shape[0]
            print ("Number of pickup records = ",a)
            temp_frame = new_frame[((new_frame.dropoff_longitude >= -74.15) & (new_frame.dropo
                                   (new_frame.dropoff_latitude >= 40.5774) & (new_frame.dropoff_la
                                   ((new_frame.pickup_longitude >= -74.15) & (new_frame.pickup_lat
                                   (new_frame.pickup_longitude <= -73.7004) & (new_frame.pickup_la
            b = temp_frame.shape[0]
            print ("Number of outlier coordinates lying outside NY boundaries:",(a-b))


            temp_frame = new_frame[(new_frame.trip_times > 0) & (new_frame.trip_times < 720)]
            c = temp_frame.shape[0]
            print ("Number of outliers from trip times analysis:",(a-c))


            temp_frame = new_frame[(new_frame.trip_distance > 0) & (new_frame.trip_distance <
            d = temp_frame.shape[0]
            print ("Number of outliers from trip distance analysis:",(a-d))

            temp_frame = new_frame[(new_frame.Speed <= 65) & (new_frame.Speed >= 0)]
            e = temp_frame.shape[0]
            print ("Number of outliers from speed analysis:",(a-e))

            temp_frame = new_frame[(new_frame.total_amount <1000) & (new_frame.total_amount >0
            f = temp_frame.shape[0]
            print ("Number of outliers from fare analysis:",(a-f))


            new_frame = new_frame[((new_frame.dropoff_longitude >= -74.15) & (new_frame.dropof
                                   (new_frame.dropoff_latitude >= 40.5774) & (new_frame.dropoff_la
                                   ((new_frame.pickup_longitude >= -74.15) & (new_frame.pickup_lat
                                   (new_frame.pickup_longitude <= -73.7004) & (new_frame.pickup_la

            new_frame = new_frame[(new_frame.trip_times > 0) & (new_frame.trip_times < 720)]
            new_frame = new_frame[(new_frame.trip_distance > 0) & (new_frame.trip_distance < 2
            new_frame = new_frame[(new_frame.Speed < 45.31) & (new_frame.Speed > 0)]
            new_frame = new_frame[(new_frame.total_amount <1000) & (new_frame.total_amount >0)

            print ("Total outliers removed",a - new_frame.shape[0])
            print ("---")
            return new_frame

In [40]: print ("Removing outliers in the month of Jan-2015")
```

```
print ("----")
frame_with_durations_outliers_removed = remove_outliers(frame_with_durations)
print("fraction of data points that remain after removing outliers", float(len(frame_w
```

```
Removing outliers in the month of Jan-2015
----
Number of pickup records =  12748986
Number of outlier coordinates lying outside NY boundaries: 293919
Number of outliers from trip times analysis: 23889
Number of outliers from trip distance analysis: 92597
Number of outliers from speed analysis: 24473
Number of outliers from fare analysis: 5275
Total outliers removed 377910
---
fraction of data points that remain after removing outliers 0.9703576425607495
```

6 Data-preperation

6.1 Clustering/Segmentation

In [41]:
```
#trying different cluster sizes to choose the right K in K-means
coords = frame_with_durations_outliers_removed[['pickup_latitude', 'pickup_longitude']
neighbours=[]

def find_min_distance(cluster_centers, cluster_len):
    nice_points = 0
    wrong_points = 0
    less2 = []
    more2 = []
    min_dist=1000
    for i in range(0, cluster_len):
        nice_points = 0
        wrong_points = 0
        for j in range(0, cluster_len):
            if j!=i:
                distance = gpxpy.geo.haversine_distance(cluster_centers[i][0], cluste
                min_dist = min(min_dist,distance/(1.60934*1000))
                if (distance/(1.60934*1000)) <= 2:
                    nice_points +=1
                else:
                    wrong_points += 1
        less2.append(nice_points)
        more2.append(wrong_points)
    neighbours.append(less2)
    print ("On choosing a cluster size of ",cluster_len,"\nAvg. Number of Clusters wit
           np.ceil(sum(less2)/len(less2)), "\nAvg. Number of Clusters outside the vic
           "\nMin inter-cluster distance = ",min_dist,"\n---")
```

```python
def find_clusters(increment):
    kmeans = MiniBatchKMeans(n_clusters=increment, batch_size=10000,random_state=42).
    frame_with_durations_outliers_removed['pickup_cluster'] = kmeans.predict(frame_wi
    cluster_centers = kmeans.cluster_centers_
    cluster_len = len(cluster_centers)
    return cluster_centers, cluster_len

# we need to choose number of clusters so that, there are more number of cluster regi
#that are close to any cluster center
# and make sure that the minimum inter cluster should not be very less
for increment in range(10, 100, 10):
    cluster_centers, cluster_len = find_clusters(increment)
    find_min_distance(cluster_centers, cluster_len)
```

```
On choosing a cluster size of  10
Avg. Number of Clusters within the vicinity (i.e. intercluster-distance < 2): 2.0
Avg. Number of Clusters outside the vicinity (i.e. intercluster-distance > 2): 8.0
Min inter-cluster distance =  1.0945442325142543
---
On choosing a cluster size of  20
Avg. Number of Clusters within the vicinity (i.e. intercluster-distance < 2): 4.0
Avg. Number of Clusters outside the vicinity (i.e. intercluster-distance > 2): 16.0
Min inter-cluster distance =  0.7131298007387813
---
On choosing a cluster size of  30
Avg. Number of Clusters within the vicinity (i.e. intercluster-distance < 2): 8.0
Avg. Number of Clusters outside the vicinity (i.e. intercluster-distance > 2): 22.0
Min inter-cluster distance =  0.5185088176172206
---
On choosing a cluster size of  40
Avg. Number of Clusters within the vicinity (i.e. intercluster-distance < 2): 8.0
Avg. Number of Clusters outside the vicinity (i.e. intercluster-distance > 2): 32.0
Min inter-cluster distance =  0.5069768450363973
---
On choosing a cluster size of  50
Avg. Number of Clusters within the vicinity (i.e. intercluster-distance < 2): 12.0
Avg. Number of Clusters outside the vicinity (i.e. intercluster-distance > 2): 38.0
Min inter-cluster distance =  0.365363025983595
---
On choosing a cluster size of  60
Avg. Number of Clusters within the vicinity (i.e. intercluster-distance < 2): 14.0
Avg. Number of Clusters outside the vicinity (i.e. intercluster-distance > 2): 46.0
Min inter-cluster distance =  0.34704283494187155
---
On choosing a cluster size of  70
Avg. Number of Clusters within the vicinity (i.e. intercluster-distance < 2): 16.0
Avg. Number of Clusters outside the vicinity (i.e. intercluster-distance > 2): 54.0
```

```
Min inter-cluster distance =  0.30502203163244707
---
On choosing a cluster size of  80
Avg. Number of Clusters within the vicinity (i.e. intercluster-distance < 2): 18.0
Avg. Number of Clusters outside the vicinity (i.e. intercluster-distance > 2): 62.0
Min inter-cluster distance =  0.29220324531738534
---
On choosing a cluster size of  90
Avg. Number of Clusters within the vicinity (i.e. intercluster-distance < 2): 21.0
Avg. Number of Clusters outside the vicinity (i.e. intercluster-distance > 2): 69.0
Min inter-cluster distance =  0.18257992857034985
---
```

6.1.1 Inference:

- The main objective was to find a optimal min. distance(Which roughly estimates to the radius of a cluster) between the clusters which we got was 40

```
In [0]: # if check for the 50 clusters you can observe that there are two clusters with only 0
        # so we choose 40 clusters for solve the further problem

        # Getting 40 clusters using the kmeans
        kmeans = MiniBatchKMeans(n_clusters=40, batch_size=10000,random_state=0).fit(coords)
        frame_with_durations_outliers_removed['pickup_cluster'] = kmeans.predict(frame_with_du
```

6.1.2 Plotting the cluster centers:

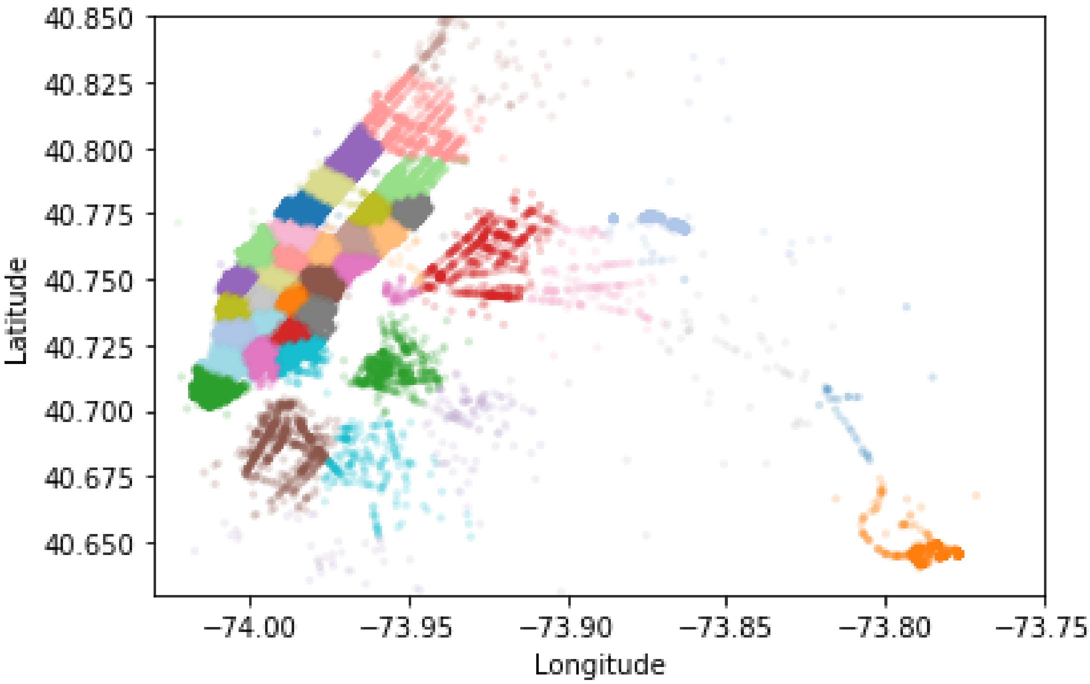
```
In [43]: # Plotting the cluster centers on OSM
         cluster_centers = kmeans.cluster_centers_
         cluster_len = len(cluster_centers)
         map_osm = folium.Map(location=[40.734695, -73.990372], tiles='Stamen Toner')
         for i in range(cluster_len):
             folium.Marker(list((cluster_centers[i][0],cluster_centers[i][1])), popup=(str(clus
         map_osm

Out[43]: <folium.folium.Map at 0x7fe2d7016cf8>
```

6.1.3 Plotting the clusters:

```
In [44]: # visualise clusters on map


         def plot_clusters(frame):
             city_long_border = (-74.03, -73.75)
             city_lat_border = (40.63, 40.85)
             fig, ax = plt.subplots(ncols=1,nrows=1)
             ax.scatter(frame.pickup_longitude.values[:100000], frame.pickup_latitude.values[:1
                        s=10,lw=0, c= frame.pickup_cluster.values[:100000],cmap='tab20',alpha=0
             ax.set_xlim(city_long_border)
```

```
    ax.set_ylim(city_lat_border)
    ax.set_xlabel('Longitude')
    ax.set_ylabel('Latitude')
    plt.show()
```

```
plot_clusters(frame_with_durations_outliers_removed)
```



6.2 Time-binning

```
In [0]: def add_pickup_bins(frame,month,year):
          unix_pickup_times = [i for i in frame['pickup_times'].values]
          unix_times = [[1420070400,1422748800,1425168000,1427846400,1430438400,1433116800],\
                        [1451606400,1454284800,1456790400,1459468800,1462060800,1464739200]
          start_pickup_unix = unix_times[year-2015][month-1]
          tenminutewise_binned_unix_pickup_times=[(int((i-start_pickup_unix)/600)+33) for i in
          frame['pickup_bins'] = np.array(tenminutewise_binned_unix_pickup_times)

          return frame
```

```
In [0]: # clustering, making pickup bins and grouping by pickup cluster and pickup bins
        frame_with_durations_outliers_removed['pickup_cluster'] = kmeans.predict(frame_with_du
        jan_2015_frame = add_pickup_bins(frame_with_durations_outliers_removed,1,2015)
```

```
In [0]:   jan_2015_groupby = jan_2015_frame[['pickup_cluster','pickup_bins','trip_distance']].
```

In [48]:
```
# we add two more columns 'pickup_cluster'(to which cluster it belogns to)
# and 'pickup_bins' (to which 10min intravel the trip belongs to)
jan_2015_frame.head()
```

Out[48]:
```
   passenger_count  trip_distance  ...  pickup_cluster  pickup_bins
0                1           1.59  ...              34         2163
1                1           3.30  ...               2         1452
2                1           1.80  ...              16         1452
3                1           0.50  ...              38         1452
4                1           3.00  ...              22         1452

[5 rows x 12 columns]
```

In [49]:
```
# hear the trip_distance represents the number of pickups that are happend in that pa
# this data frame has two indices
# primary index: pickup_cluster (cluster number)
# secondary index : pickup_bins (we devid whole months time into 10min intravels 24*3
jan_2015_groupby.head()
```

Out[49]:
```
                             trip_distance
pickup_cluster pickup_bins
0              33                      104
               34                      200
               35                      208
               36                      141
               37                      155
```

In [50]:
```
# upto now we cleaned data and prepared data for the month 2015,

# now do the same operations for months Jan, Feb, March of 2016
# 1. get the dataframe which inlcudes only required colums
# 2. adding trip times, speed, unix time stamp of pickup_time
# 4. remove the outliers based on trip_times, speed, trip_duration, total_amount
# 5. add pickup_cluster to each data point
# 6. add pickup_bin (index of 10min intravel to which that trip belongs to)
# 7. group by data, based on 'pickup_cluster' and 'pickuo_bin'

# Data Preparation for the months of Jan,Feb and March 2016
def datapreparation(month,kmeans,month_no,year_no):

    print ("Return with trip times..")

    frame_with_durations = return_with_trip_times(month)

    print ("Remove outliers..")
    frame_with_durations_outliers_removed = remove_outliers(frame_with_durations)

    print ("Estimating clusters..")
    frame_with_durations_outliers_removed['pickup_cluster'] = kmeans.predict(frame_wi
```

```
    #frame_with_durations_outliers_removed_2016['pickup_cluster'] = kmeans.predict(fr

    print ("Final groupbying..")
    final_updated_frame = add_pickup_bins(frame_with_durations_outliers_removed,month_
    final_groupby_frame = final_updated_frame[['pickup_cluster','pickup_bins','trip_di

    return final_updated_frame,final_groupby_frame

month_jan_2016 = dd.read_csv('yellow_tripdata_2016-01.csv')
month_feb_2016 = dd.read_csv('yellow_tripdata_2016-02.csv')
month_mar_2016 = dd.read_csv('yellow_tripdata_2016-03.csv')

jan_2016_frame,jan_2016_groupby = datapreparation(month_jan_2016,kmeans,1,2016)
feb_2016_frame,feb_2016_groupby = datapreparation(month_feb_2016,kmeans,2,2016)
mar_2016_frame,mar_2016_groupby = datapreparation(month_mar_2016,kmeans,3,2016)
```

```
Return with trip times..
Remove outliers..
Number of pickup records =  10906858
Number of outlier coordinates lying outside NY boundaries: 214677
Number of outliers from trip times analysis: 27190
Number of outliers from trip distance analysis: 79742
Number of outliers from speed analysis: 21047
Number of outliers from fare analysis: 4991
Total outliers removed 297784
---
Estimating clusters..
Final groupbying..
Return with trip times..
Remove outliers..
Number of pickup records =  11382049
Number of outlier coordinates lying outside NY boundaries: 223161
Number of outliers from trip times analysis: 27670
Number of outliers from trip distance analysis: 81902
Number of outliers from speed analysis: 22437
Number of outliers from fare analysis: 5476
Total outliers removed 308177
---
Estimating clusters..
Final groupbying..
Return with trip times..
Remove outliers..
Number of pickup records =  12210952
Number of outlier coordinates lying outside NY boundaries: 232444
Number of outliers from trip times analysis: 30868
Number of outliers from trip distance analysis: 87318
Number of outliers from speed analysis: 23889
Number of outliers from fare analysis: 5859
```

```
Total outliers removed 324635
---
Estimating clusters..
Final groupbying..
```

6.3 Smoothing

```
In [0]: # Gets the unique bins where pickup values are present for each each reigion

        # for each cluster region we will collect all the indices of 10min intravels in which
        # we got an observation that there are some pickpbins that doesnt have any pickups
        def return_unq_pickup_bins(frame):
            values = []
            for i in range(0,40):
                new = frame[frame['pickup_cluster'] == i]
                list_unq = list(set(new['pickup_bins']))
                list_unq.sort()
                values.append(list_unq)
            return values
```

```
In [0]: # for every month we get all indices of 10min intravels in which atleast one pickup go

        #jan
        jan_2015_unique = return_unq_pickup_bins(jan_2015_frame)
        jan_2016_unique = return_unq_pickup_bins(jan_2016_frame)

        #feb
        feb_2016_unique = return_unq_pickup_bins(feb_2016_frame)

        #march
        mar_2016_unique = return_unq_pickup_bins(mar_2016_frame)
```

```
In [53]: # for each cluster number of 10min intravels with 0 pickups
         for i in range(40):
             print("for the ",i,"th cluster number of 10min intavels with zero pickups: ",4464
             print('-'*60)
```

```
for the  0 th cluster number of 10min intavels with zero pickups:  40
------------------------------------------------------------
for the  1 th cluster number of 10min intavels with zero pickups:  1985
------------------------------------------------------------
for the  2 th cluster number of 10min intavels with zero pickups:  29
------------------------------------------------------------
for the  3 th cluster number of 10min intavels with zero pickups:  354
------------------------------------------------------------
for the  4 th cluster number of 10min intavels with zero pickups:  37
------------------------------------------------------------
for the  5 th cluster number of 10min intavels with zero pickups:  153
```

```
------------------------------------------------------------
for the  6 th cluster number of 10min intavels with zero pickups:  34
------------------------------------------------------------
for the  7 th cluster number of 10min intavels with zero pickups:  34
------------------------------------------------------------
for the  8 th cluster number of 10min intavels with zero pickups:  117
------------------------------------------------------------
for the  9 th cluster number of 10min intavels with zero pickups:  40
------------------------------------------------------------
for the  10 th cluster number of 10min intavels with zero pickups:  25
------------------------------------------------------------
for the  11 th cluster number of 10min intavels with zero pickups:  44
------------------------------------------------------------
for the  12 th cluster number of 10min intavels with zero pickups:  42
------------------------------------------------------------
for the  13 th cluster number of 10min intavels with zero pickups:  28
------------------------------------------------------------
for the  14 th cluster number of 10min intavels with zero pickups:  26
------------------------------------------------------------
for the  15 th cluster number of 10min intavels with zero pickups:  31
------------------------------------------------------------
for the  16 th cluster number of 10min intavels with zero pickups:  40
------------------------------------------------------------
for the  17 th cluster number of 10min intavels with zero pickups:  58
------------------------------------------------------------
for the  18 th cluster number of 10min intavels with zero pickups:  1190
------------------------------------------------------------
for the  19 th cluster number of 10min intavels with zero pickups:  1357
------------------------------------------------------------
for the  20 th cluster number of 10min intavels with zero pickups:  53
------------------------------------------------------------
for the  21 th cluster number of 10min intavels with zero pickups:  29
------------------------------------------------------------
for the  22 th cluster number of 10min intavels with zero pickups:  29
------------------------------------------------------------
for the  23 th cluster number of 10min intavels with zero pickups:  163
------------------------------------------------------------
for the  24 th cluster number of 10min intavels with zero pickups:  35
------------------------------------------------------------
for the  25 th cluster number of 10min intavels with zero pickups:  41
------------------------------------------------------------
for the  26 th cluster number of 10min intavels with zero pickups:  31
------------------------------------------------------------
for the  27 th cluster number of 10min intavels with zero pickups:  214
------------------------------------------------------------
for the  28 th cluster number of 10min intavels with zero pickups:  36
------------------------------------------------------------
for the  29 th cluster number of 10min intavels with zero pickups:  41
```

```
------------------------------------------------------------
for the  30 th cluster number of 10min intavels with zero pickups:  1180
------------------------------------------------------------
for the  31 th cluster number of 10min intavels with zero pickups:  42
------------------------------------------------------------
for the  32 th cluster number of 10min intavels with zero pickups:  44
------------------------------------------------------------
for the  33 th cluster number of 10min intavels with zero pickups:  43
------------------------------------------------------------
for the  34 th cluster number of 10min intavels with zero pickups:  39
------------------------------------------------------------
for the  35 th cluster number of 10min intavels with zero pickups:  42
------------------------------------------------------------
for the  36 th cluster number of 10min intavels with zero pickups:  36
------------------------------------------------------------
for the  37 th cluster number of 10min intavels with zero pickups:  321
------------------------------------------------------------
for the  38 th cluster number of 10min intavels with zero pickups:  36
------------------------------------------------------------
for the  39 th cluster number of 10min intavels with zero pickups:  43
------------------------------------------------------------
```

there are two ways to fill up these values
Fill the missing value with 0's
Fill the missing values with the avg values
Case 1:(values missing at the start) Ex1: _ _ _ x =>ceil(x/4), ceil(x/4), ceil(x/4), ceil(x/4) Ex2: _ _ x => ceil(x/3), ceil(x/3), ceil(x/3)
Case 2:(values missing in middle) Ex1: x _ _ y => ceil((x+y)/4), ceil((x+y)/4), ceil((x+y)/4), ceil((x+y)/4) Ex2: x _ _ _ y => ceil((x+y)/5), ceil((x+y)/5), ceil((x+y)/5), ceil((x+y)/5), ceil((x+y)/5)
Case 3:(values missing at the end) Ex1: x _ _ _ => ceil(x/4), ceil(x/4), ceil(x/4), ceil(x/4) Ex2: x _ => ceil(x/2), ceil(x/2)

```
In [0]: # Fills a value of zero for every bin where no pickup data is present
        # the count_values: number pickps that are happened in each region for each 10min intr
        # there wont be any value if there are no picksups.
        # values: number of unique bins

        # for every 10min intravel(pickup_bin) we will check it is there in our unique bin,
        # if it is there we will add the count_values[index] to smoothed data
        # if not we add 0 to the smoothed data
        # we finally return smoothed data
        def fill_missing(count_values,values):
            smoothed_regions=[]
            ind=0
            for r in range(0,40):
                smoothed_bins=[]
                for i in range(4464):
```

```
            if i in values[r]:
                smoothed_bins.append(count_values[ind])
                ind+=1
            else:
                smoothed_bins.append(0)
        smoothed_regions.extend(smoothed_bins)
    return smoothed_regions
```

```
In [0]: # Fills a value of zero for every bin where no pickup data is present
        # the count_values: number pickps that are happened in each region for each 10min intr
        # there wont be any value if there are no picksups.
        # values: number of unique bins

        # for every 10min intravel(pickup_bin) we will check it is there in our unique bin,
        # if it is there we will add the count_values[index] to smoothed data
        # if not we add smoothed data (which is calculated based on the methods that are discu
        # we finally return smoothed data
        def smoothing(count_values,values):
            smoothed_regions=[] # stores list of final smoothed values of each reigion
            ind=0
            repeat=0
            smoothed_value=0
            for r in range(0,40):
                smoothed_bins=[] #stores the final smoothed values
                repeat=0
                for i in range(4464):
                    if repeat!=0: # prevents iteration for a value which is already visited/re
                        repeat-=1
                        continue
                    if i in values[r]: #checks if the pickup-bin exists
                        smoothed_bins.append(count_values[ind]) # appends the value of the pic
                    else:
                        if i!=0:
                            right_hand_limit=0
                            for j in range(i,4464):
                                if  j not in values[r]: #searches for the left-limit or the pi
                                    continue
                                else:
                                    right_hand_limit=j
                                    break
                            if right_hand_limit==0:
                            #Case 1: When we have the last/last few values are found to be mis
                                smoothed_value=count_values[ind-1]*1.0/((4463-i)+2)*1.0
                                for j in range(i,4464):
                                    smoothed_bins.append(math.ceil(smoothed_value))
                                smoothed_bins[i-1] = math.ceil(smoothed_value)
                                repeat=(4463-i)
                                ind-=1
```

```
                        else:
                        #Case 2: When we have the missing values between two known values
                            smoothed_value=(count_values[ind-1]+count_values[ind])*1.0/((ri
                            for j in range(i,right_hand_limit+1):
                                smoothed_bins.append(math.ceil(smoothed_value))
                            smoothed_bins[i-1] = math.ceil(smoothed_value)
                            repeat=(right_hand_limit-i)
                    else:
                        #Case 3: When we have the first/first few values are found to be m
                        right_hand_limit=0
                        for j in range(i,4464):
                            if  j not in values[r]:
                                continue
                            else:
                                right_hand_limit=j
                                break
                        smoothed_value=count_values[ind]*1.0/((right_hand_limit-i)+1)*1.0
                        for j in range(i,right_hand_limit+1):
                                smoothed_bins.append(math.ceil(smoothed_value))
                        repeat=(right_hand_limit-i)
                ind+=1
            smoothed_regions.extend(smoothed_bins)
        return smoothed_regions
```

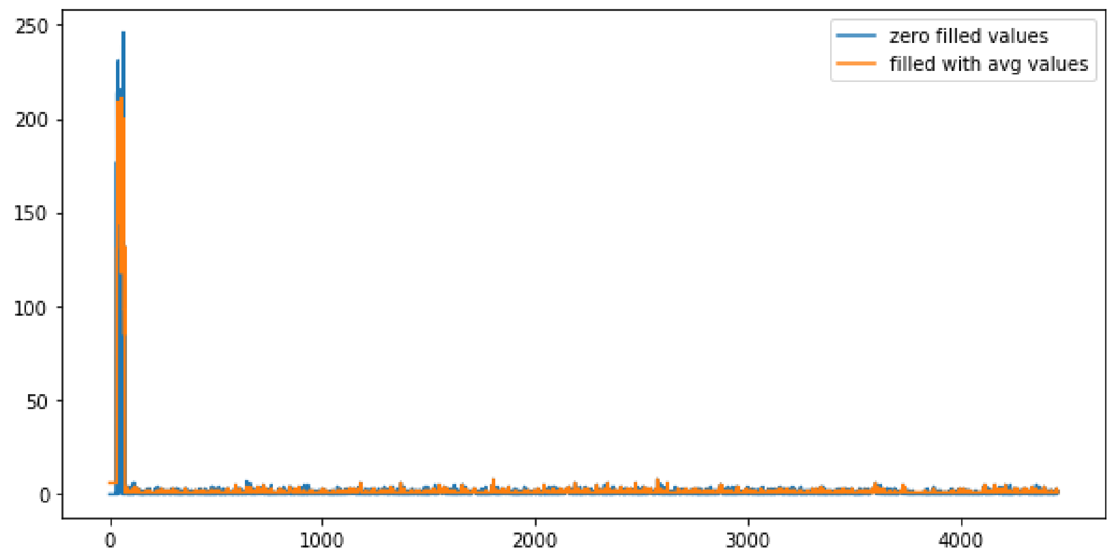
In [0]:
```
#Filling Missing values of Jan-2015 with 0
# here in jan_2015_groupby dataframe the trip_distance represents the number of pickup
jan_2015_fill = fill_missing(jan_2015_groupby['trip_distance'].values,jan_2015_unique)
```

In [0]:
```
#Smoothing Missing values of Jan-2015
jan_2015_smooth = smoothing(jan_2015_groupby['trip_distance'].values,jan_2015_unique)
```

In [58]:
```
# number of 10min indices for jan 2015= 24*31*60/10 = 4464
# number of 10min indices for jan 2016 = 24*31*60/10 = 4464
# number of 10min indices for feb 2016 = 24*29*60/10 = 4176
# number of 10min indices for march 2016 = 24*30*60/10 = 4320
# for each cluster we will have 4464 values, therefore 40*4464 = 178560 (length of th
print("number of 10min intravels among all the clusters ",len(jan_2015_fill))
```

```
number of 10min intravels among all the clusters  178560
```

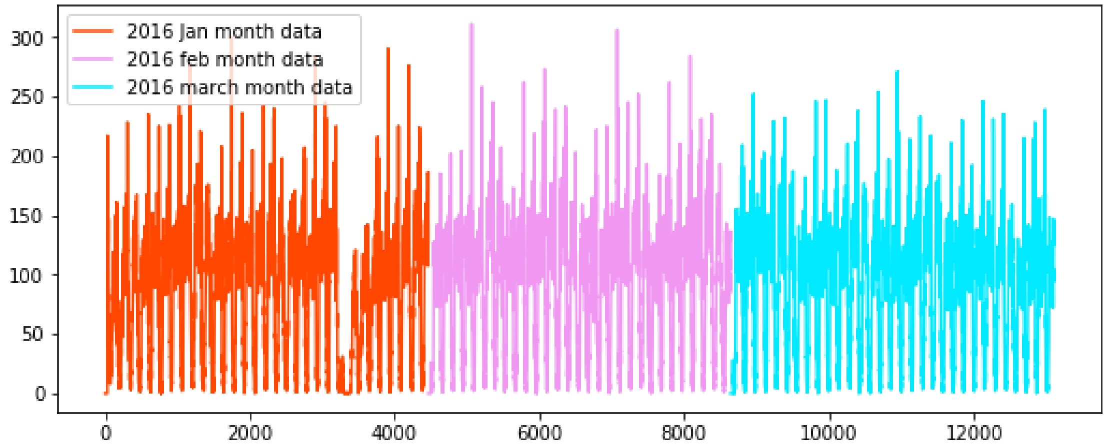
In [59]:
```
# Smoothing vs Filling
# sample plot that shows two variations of filling missing values
# we have taken the number of pickups for cluster region 2
plt.figure(figsize=(10,5))
plt.plot(jan_2015_fill[4464:8920], label="zero filled values")
plt.plot(jan_2015_smooth[4464:8920], label="filled with avg values")
plt.legend()
plt.show()
```

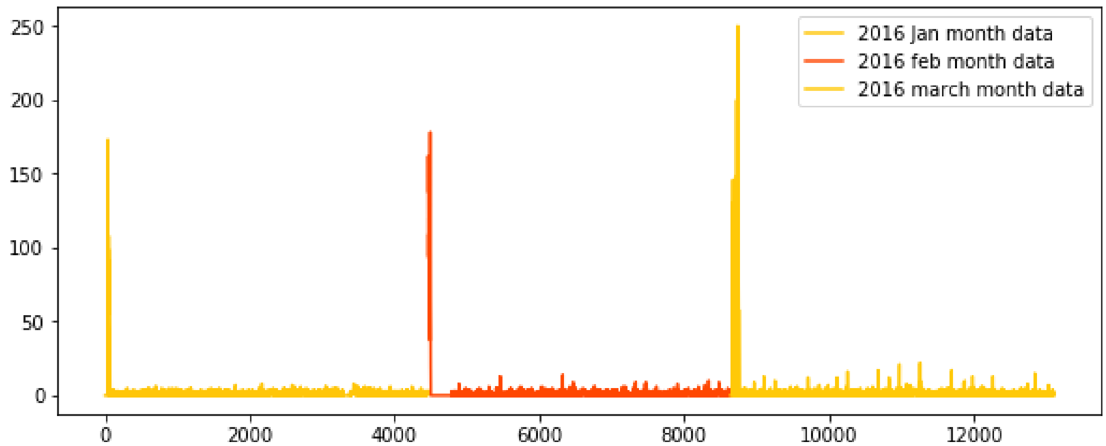


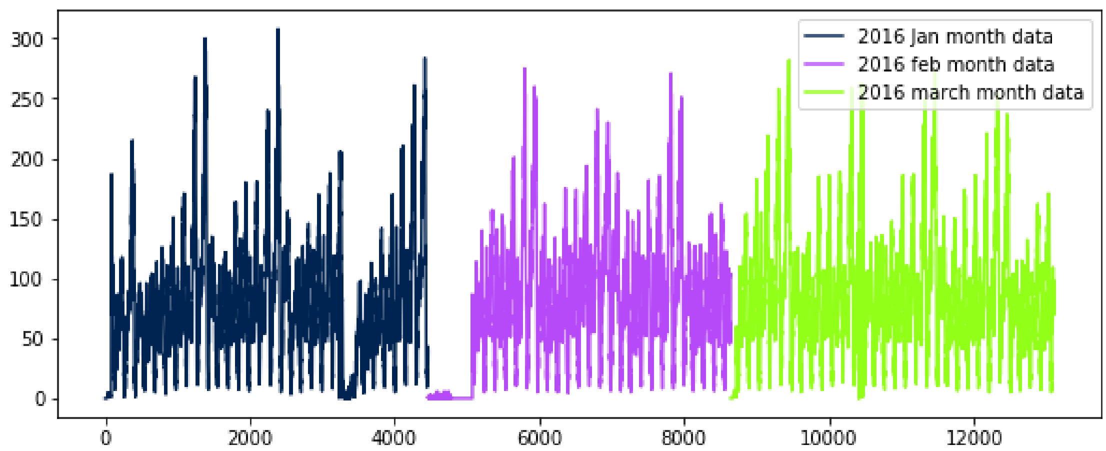
In [0]:
```
# Jan-2015 data is smoothed, Jan,Feb & March 2016 data missing values are filled with
jan_2015_smooth = smoothing(jan_2015_groupby['trip_distance'].values,jan_2015_unique)
jan_2016_smooth = fill_missing(jan_2016_groupby['trip_distance'].values,jan_2016_unique
feb_2016_smooth = fill_missing(feb_2016_groupby['trip_distance'].values,feb_2016_unique
mar_2016_smooth = fill_missing(mar_2016_groupby['trip_distance'].values,mar_2016_unique

# Making list of all the values of pickup data in every bin for a period of 3 months a
regions_cum = []

# a =[1,2,3]
# b = [2,3,4]
# a+b = [1, 2, 3, 2, 3, 4]

# number of 10min indices for jan 2015= 24*31*60/10 = 4464
# number of 10min indices for jan 2016 = 24*31*60/10 = 4464
# number of 10min indices for feb 2016 = 24*29*60/10 = 4176
# number of 10min indices for march 2016 = 24*31*60/10 = 4464
# regions_cum: it will contain 40 lists, each list will contain 4464+4176+4464 values
# that are happened for three months in 2016 data

for i in range(0,40):
    regions_cum.append(jan_2016_smooth[4464*i:4464*(i+1)]+feb_2016_smooth[4176*i:4176*

# print(len(regions_cum))
# 40
# print(len(regions_cum[0]))
# 13104
```

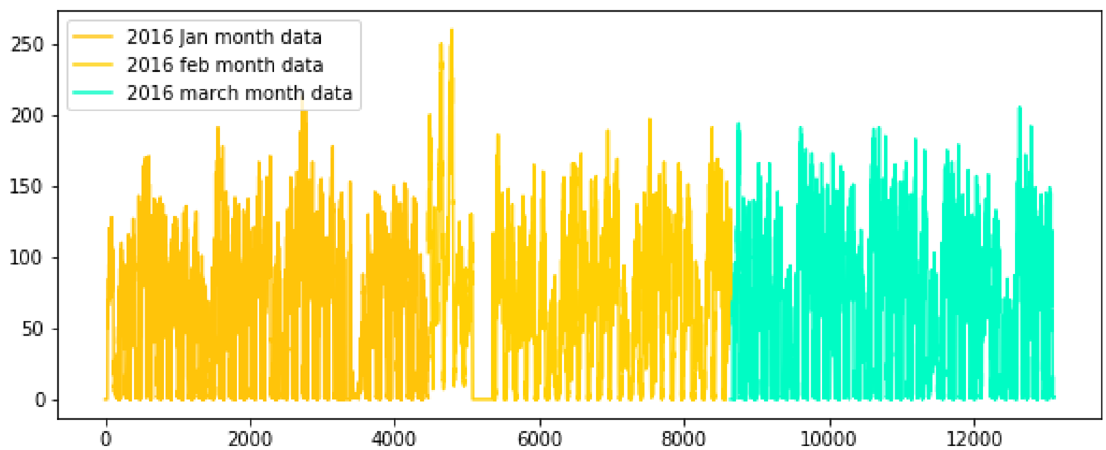
6.4 Time series and Fourier Transforms

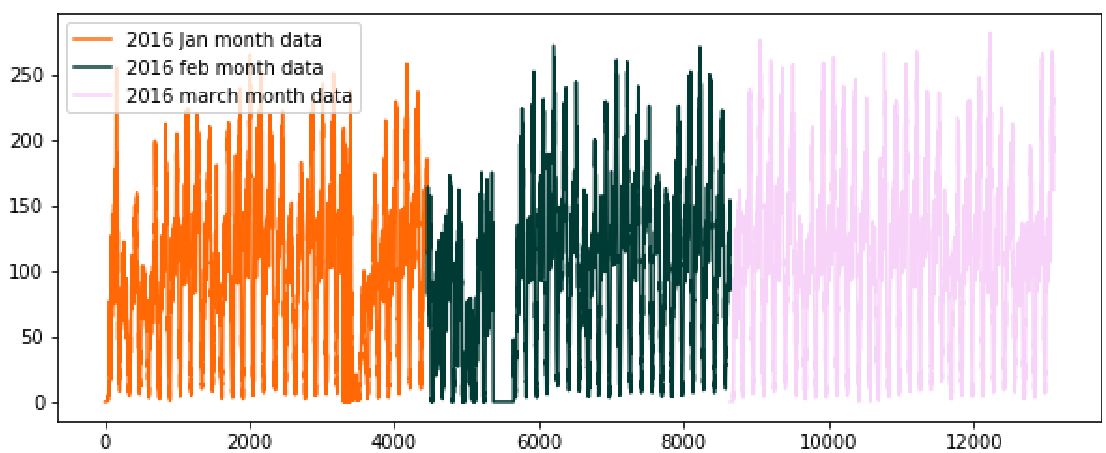
```
In [61]: def uniqueish_color():
             """There're better ways to generate unique colors, but this isn't awful."""
             return plt.cm.gist_ncar(np.random.random())
         first_x = list(range(0,4464))
         second_x = list(range(4464,8640))
         third_x = list(range(8640,13104))
         for i in range(40):
             plt.figure(figsize=(10,4))
             plt.plot(first_x,regions_cum[i][:4464], color=uniqueish_color(), label='2016 Jan m
             plt.plot(second_x,regions_cum[i][4464:8640], color=uniqueish_color(), label='2016 
             plt.plot(third_x,regions_cum[i][8640:], color=uniqueish_color(), label='2016 marcl
             plt.legend()
             plt.show()
```

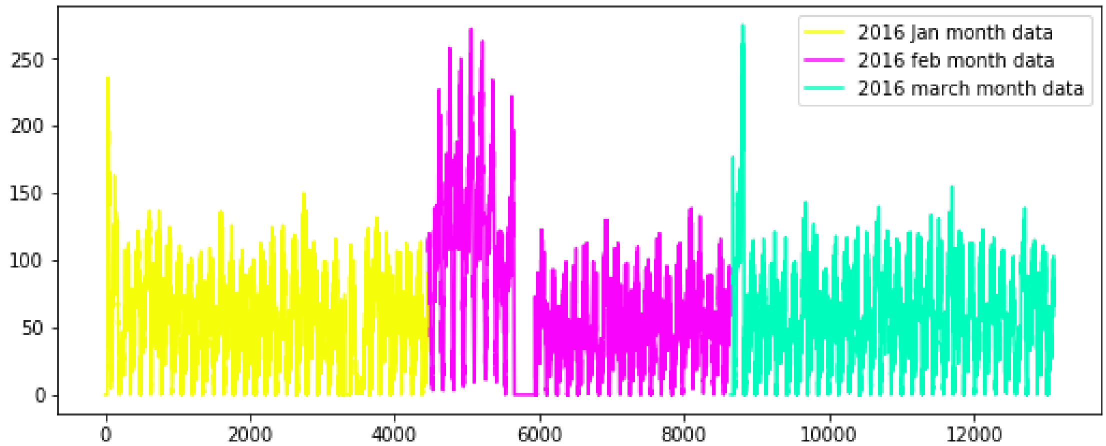


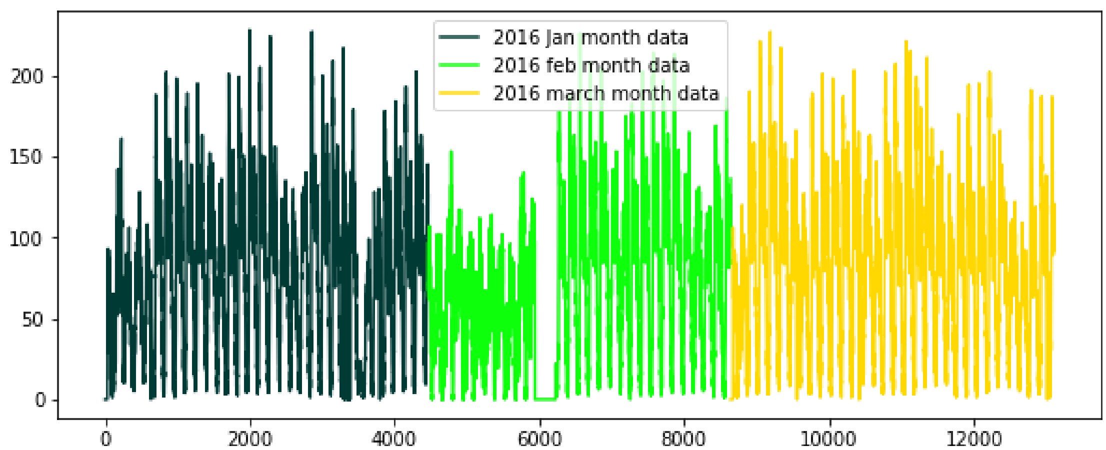


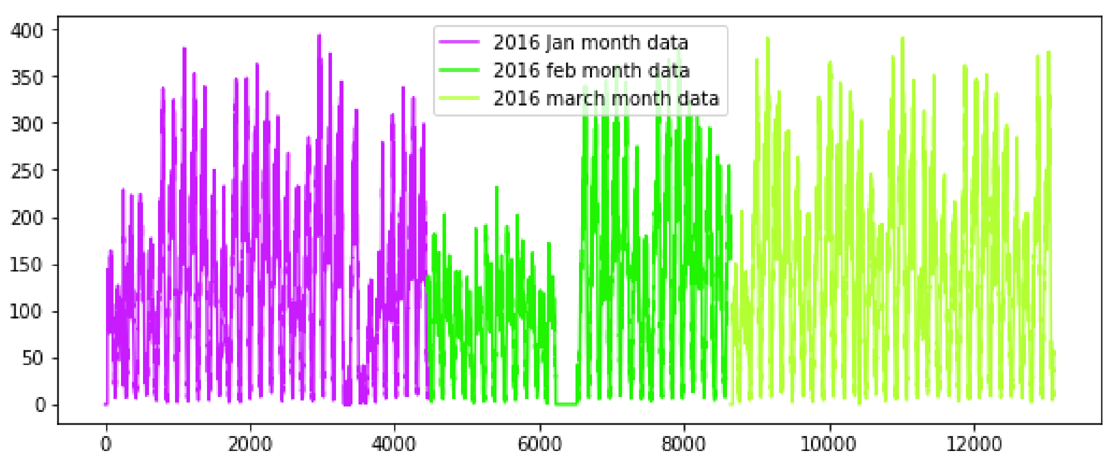

2016 Jan month data
2016 feb month data
2016 march month data

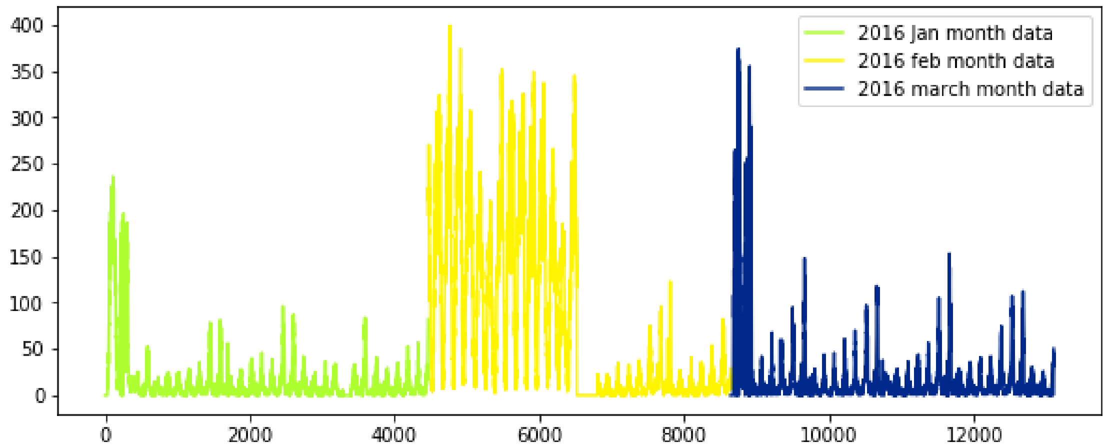

2016 Jan month data
2016 feb month data
2016 march month data

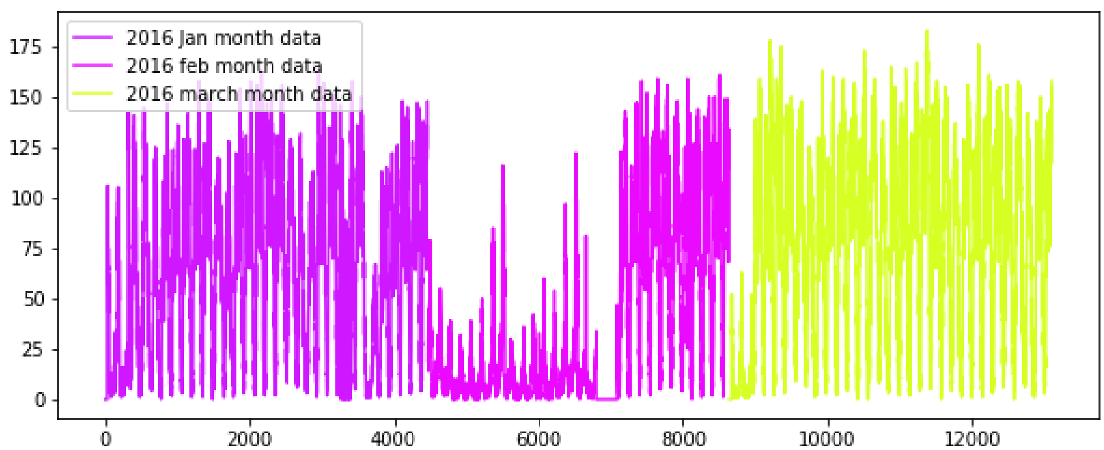

2016 Jan month data
2016 feb month data
2016 march month data

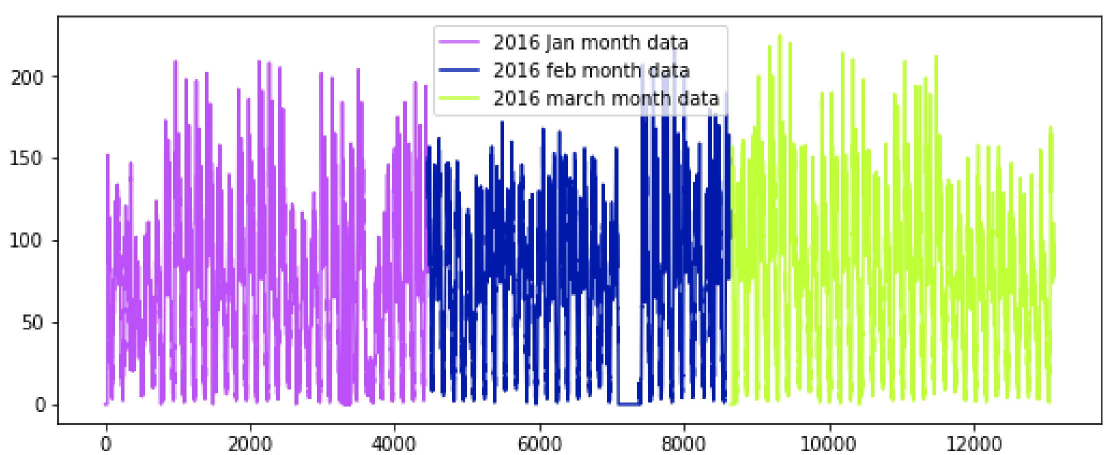

2016 Jan month data
2016 feb month data
2016 march month data

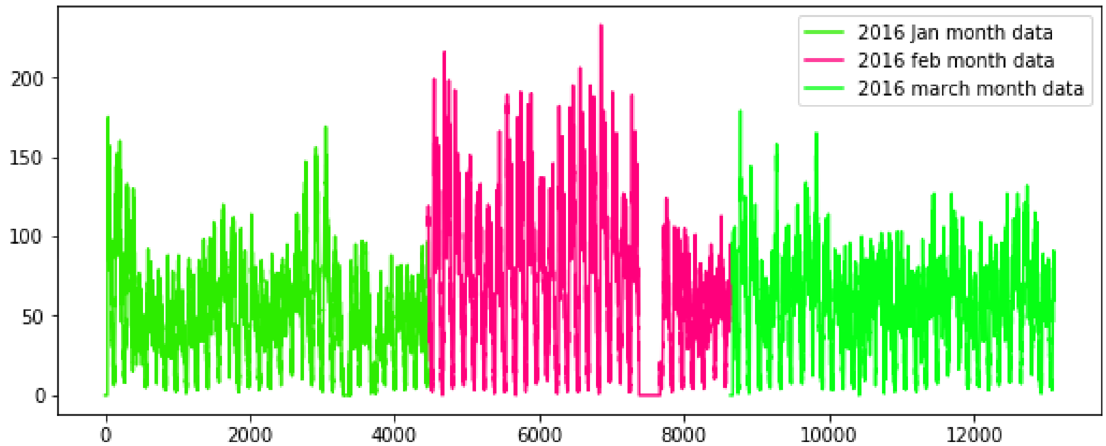

2016 Jan month data
2016 feb month data
2016 march month data

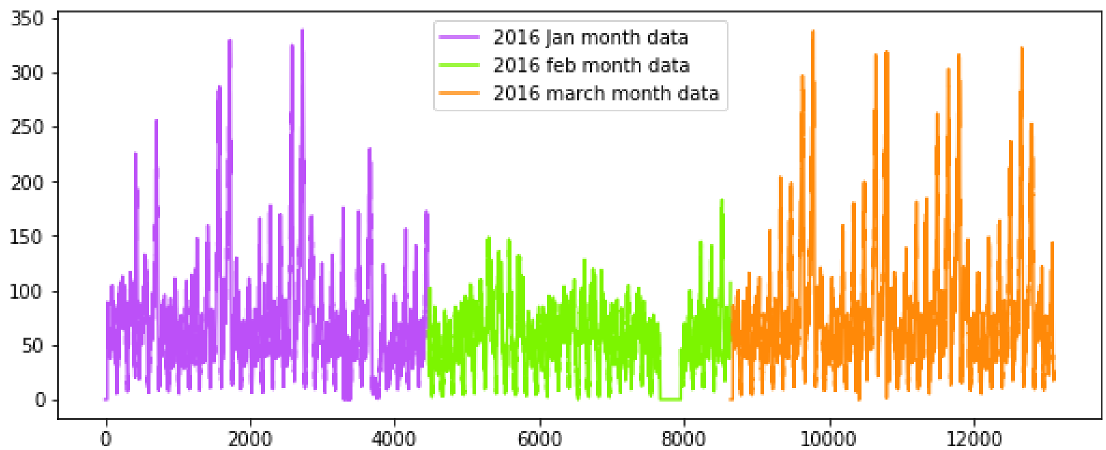

2016 Jan month data
2016 feb month data
2016 march month data

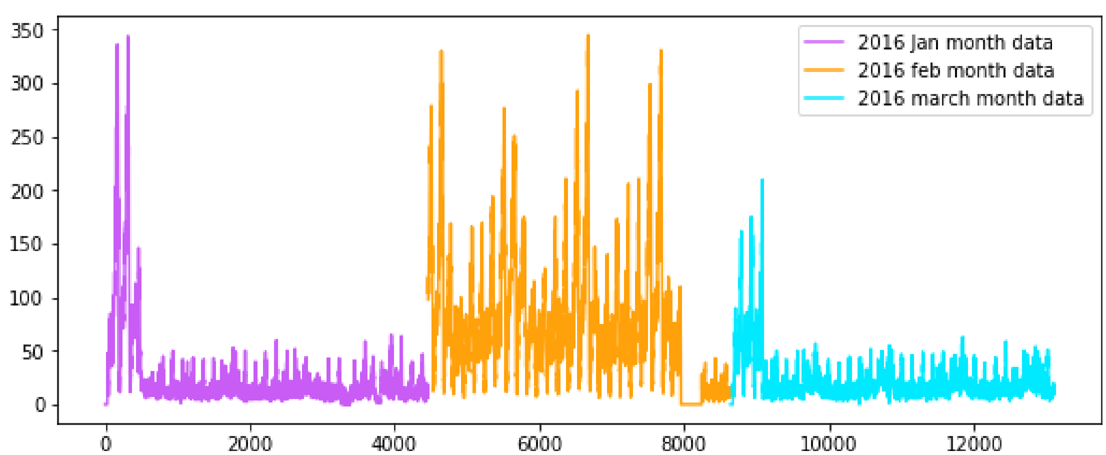

2016 Jan month data
2016 feb month data
2016 march month data

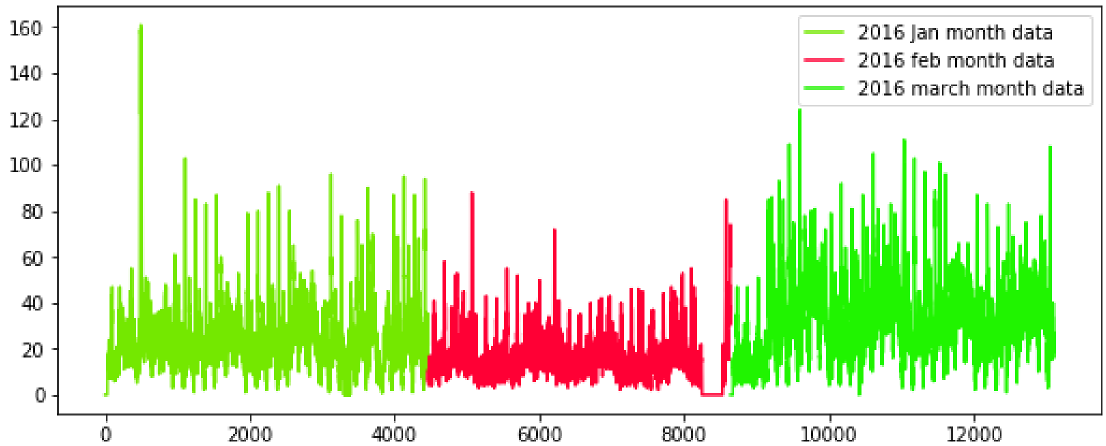

2016 Jan month data
2016 feb month data
2016 march month data

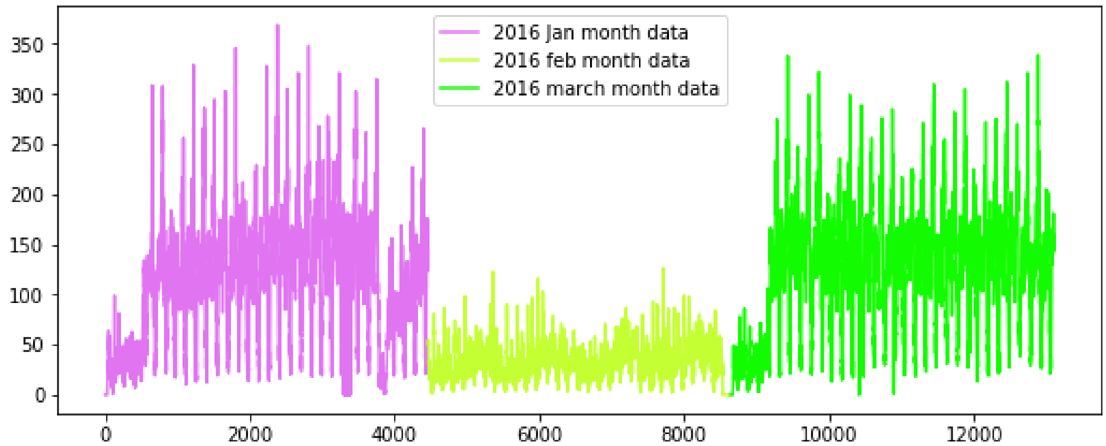

2016 Jan month data
2016 feb month data
2016 march month data

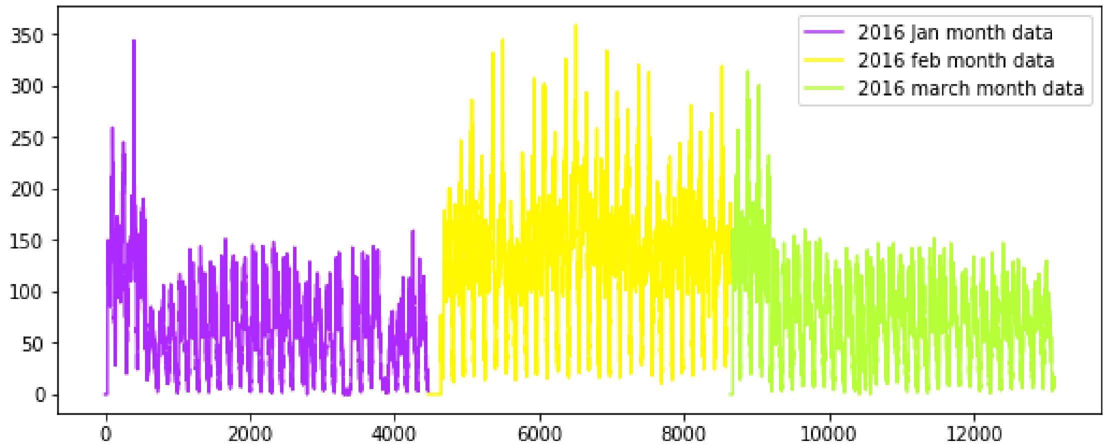

2016 Jan month data
2016 feb month data
2016 march month data

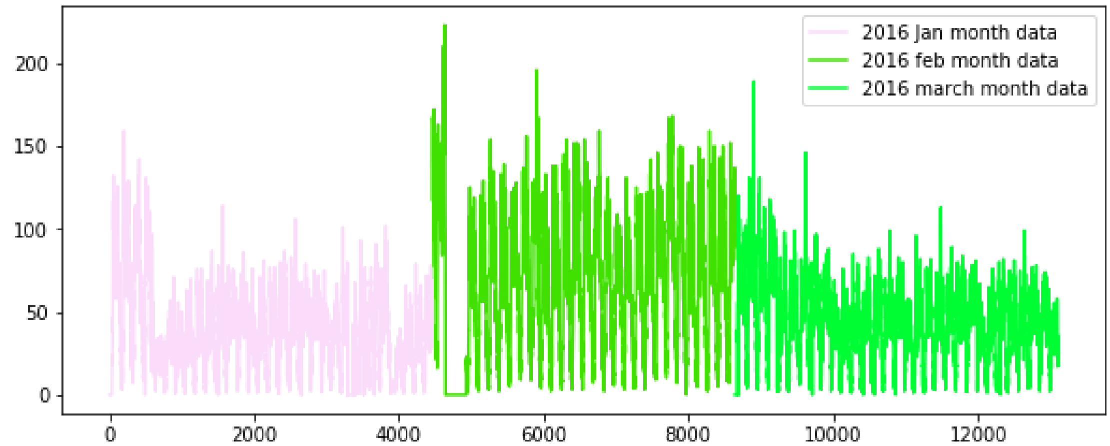

2016 Jan month data
2016 feb month data
2016 march month data

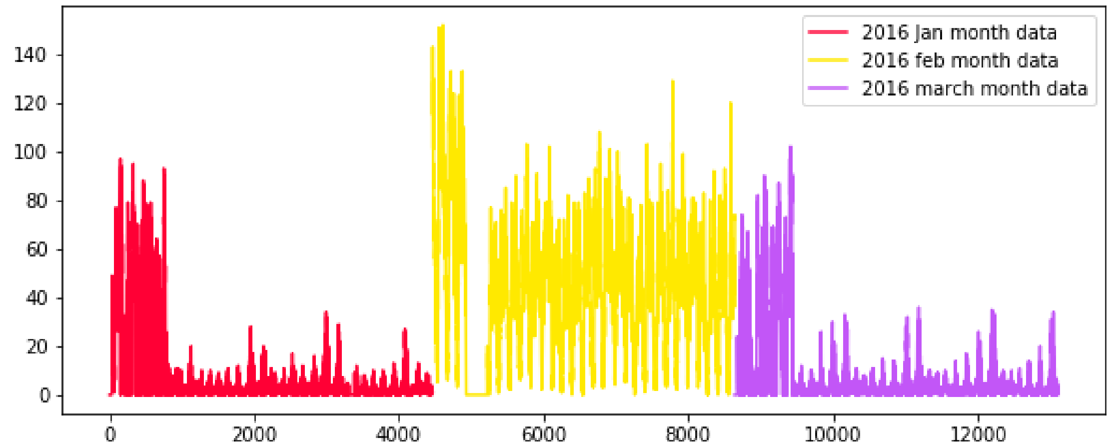

2016 Jan month data
2016 feb month data
2016 march month data

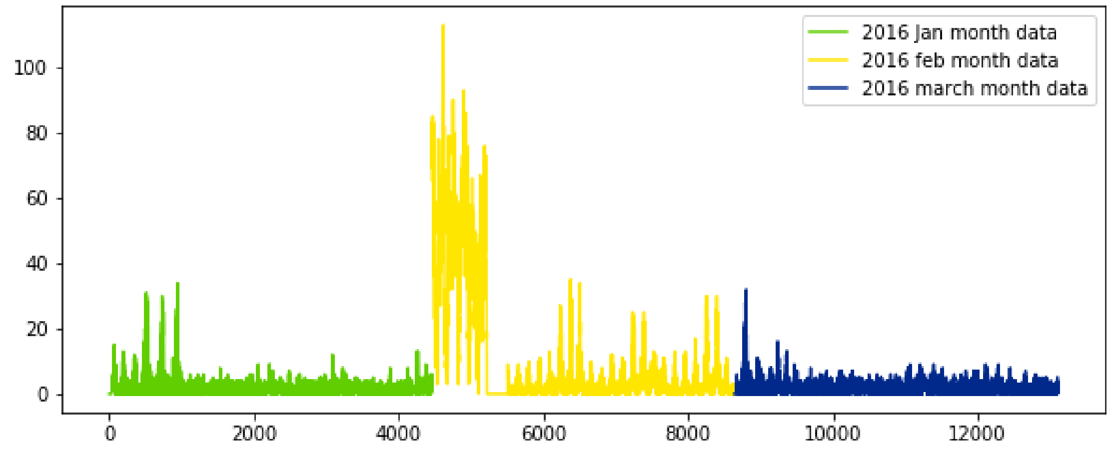

2016 Jan month data
2016 feb month data
2016 march month data

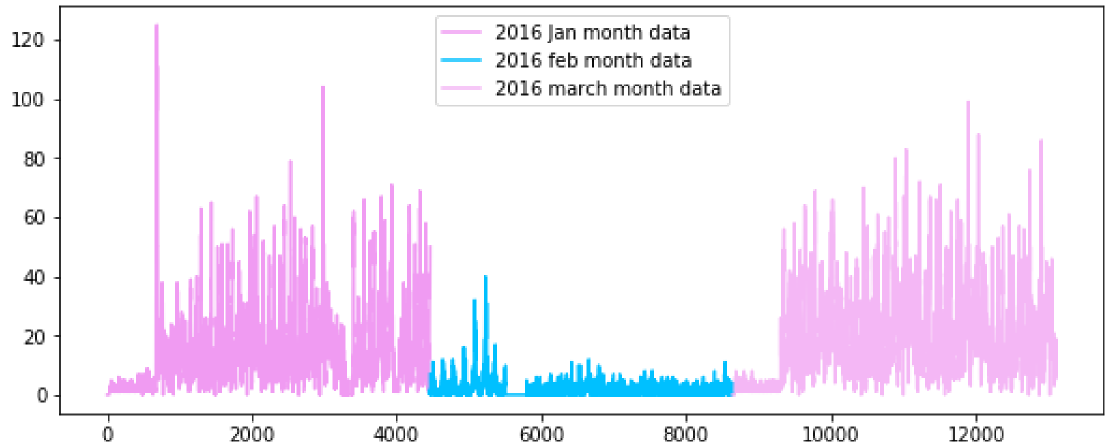

2016 Jan month data
2016 feb month data
2016 march month data

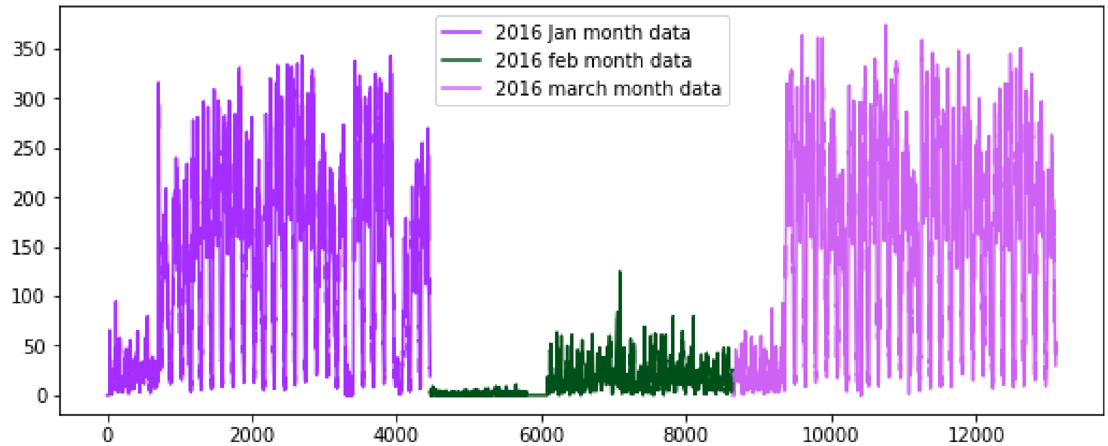

2016 Jan month data
2016 feb month data
2016 march month data

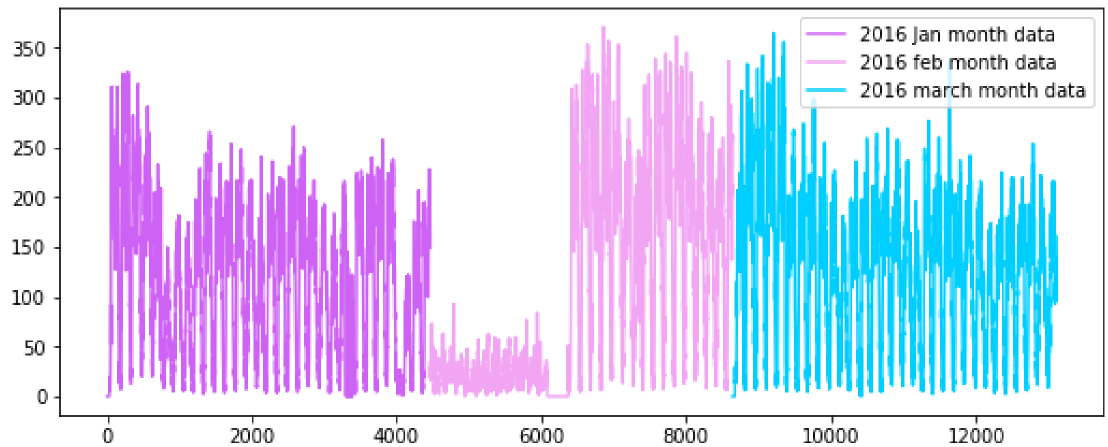

2016 Jan month data
2016 feb month data
2016 march month data

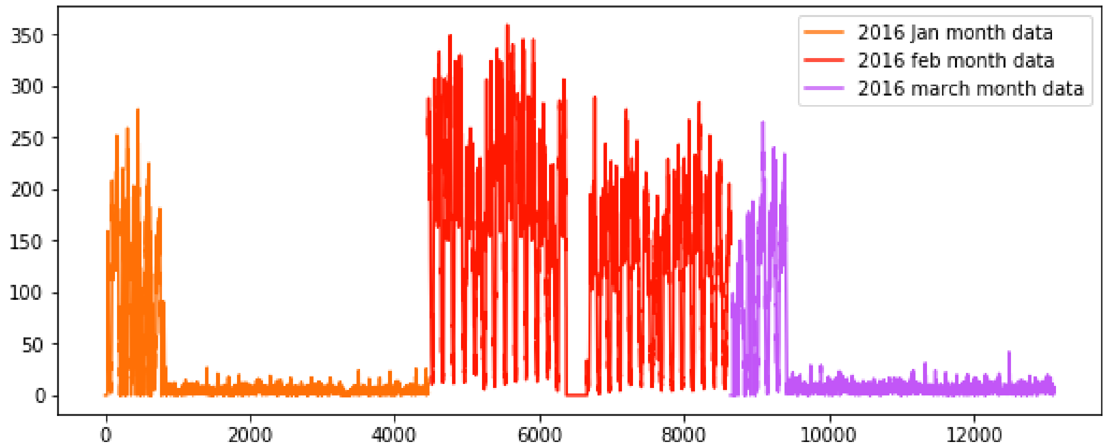

2016 Jan month data
2016 feb month data
2016 march month data

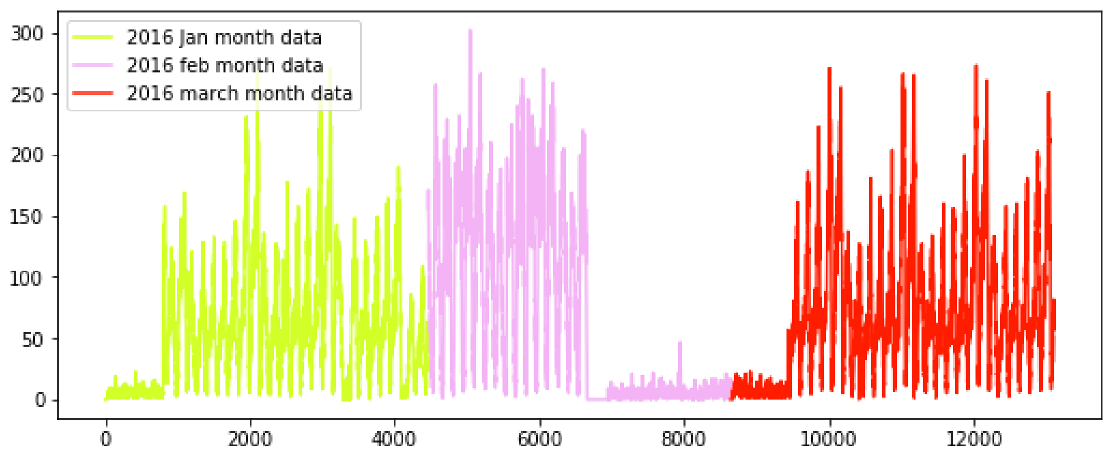

2016 Jan month data
2016 feb month data
2016 march month data

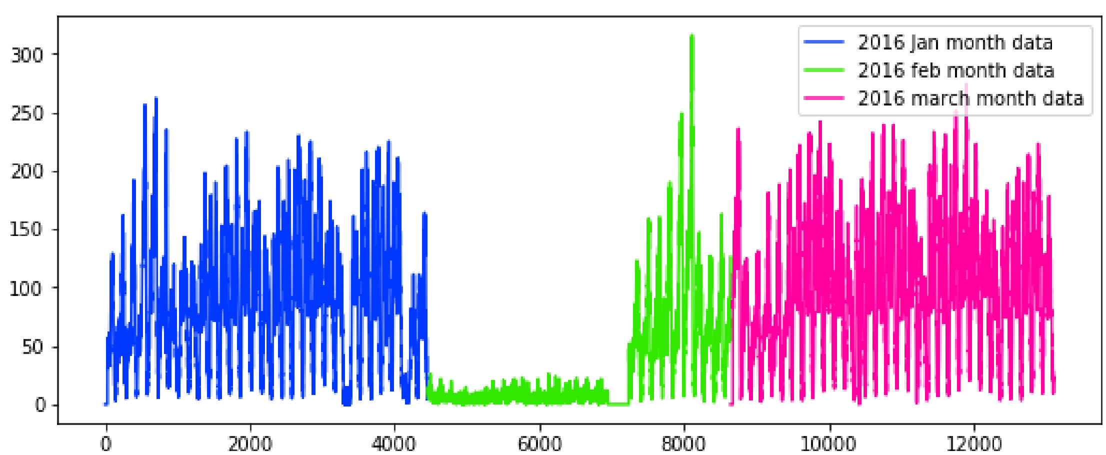

2016 Jan month data
2016 feb month data
2016 march month data
120
100
80
60
40
20
0
0
2000
4000
6000
8000
10000
12000

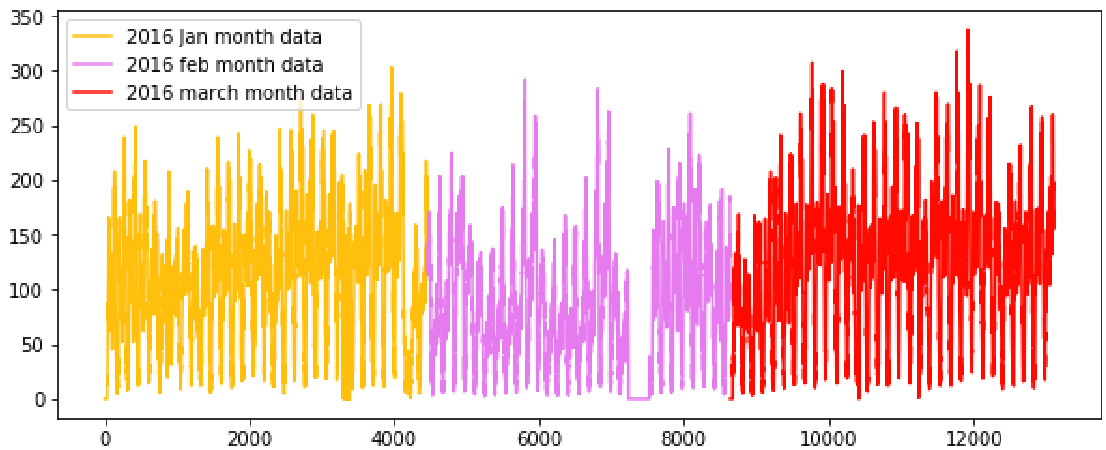

2016 Jan month data
2016 feb month data
2016 march month data
350
300
250
200
150
100
50
0
0
2000
4000
6000
8000
10000
12000

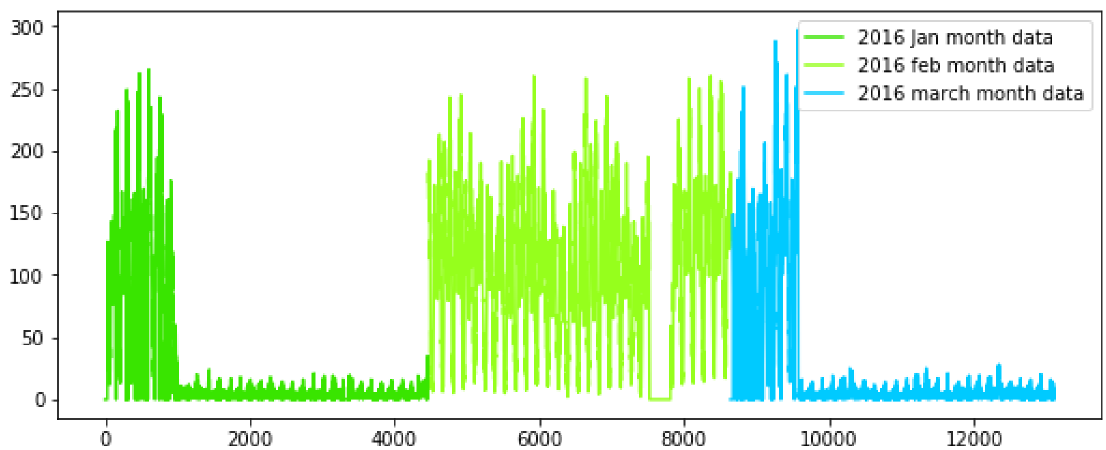

2016 Jan month data
2016 feb month data
2016 march month data
350
300
250
200
150
100
50
0
0
2000
4000
6000
8000
10000
12000

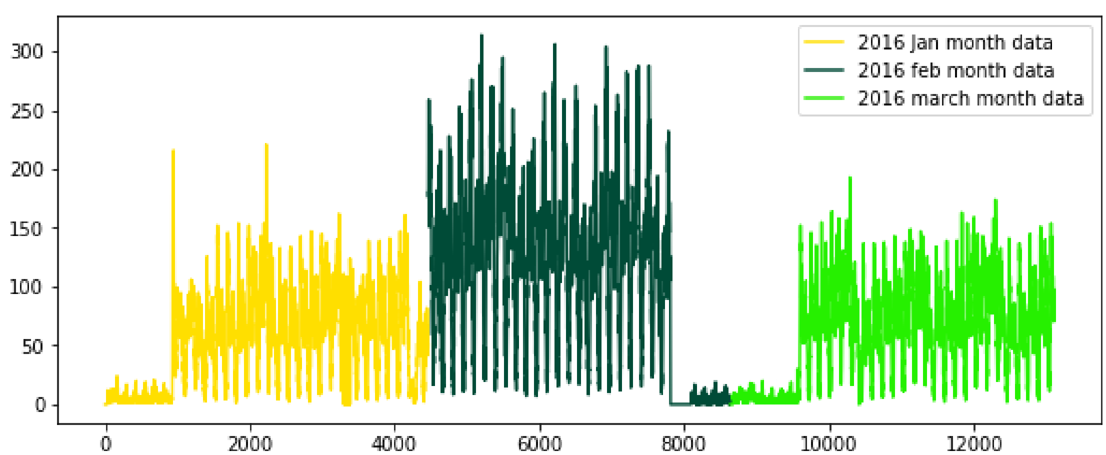

2016 Jan month data
2016 feb month data
2016 march month data
350
300
250
200
150
100
50
0
0
2000
4000
6000
8000
10000
12000

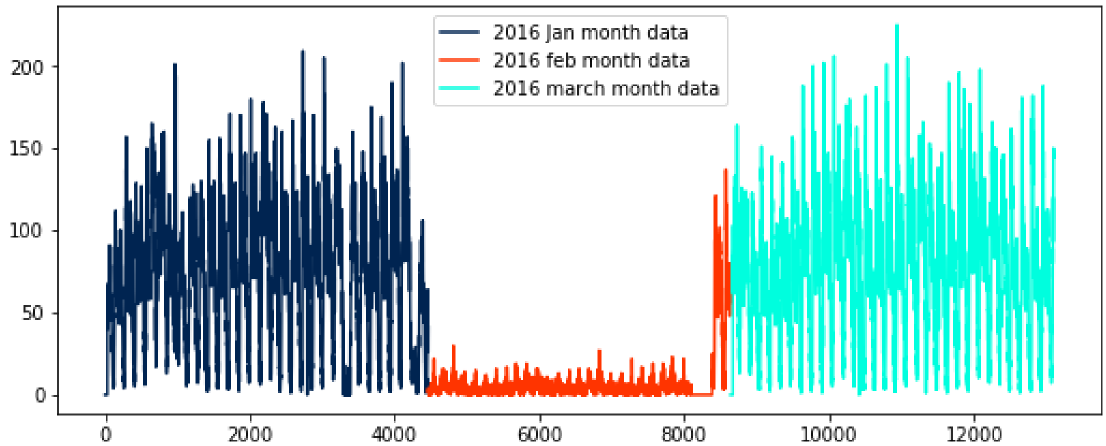

2016 Jan month data
2016 feb month data
2016 march month data
300
250
200
150
100
50
0
0
2000
4000
6000
8000
10000
12000

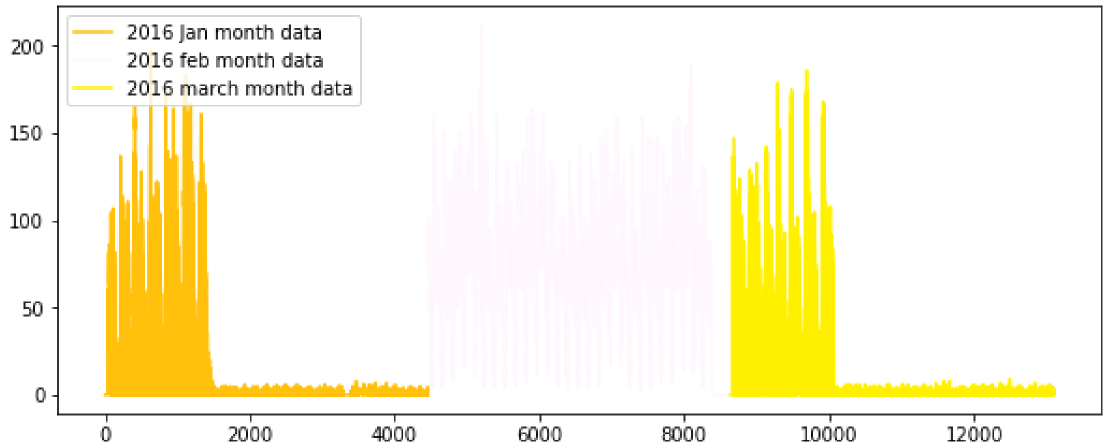

2016 Jan month data
2016 feb month data
2016 march month data
300
250
200
150
100
50
0
0
2000
4000
6000
8000
10000
12000

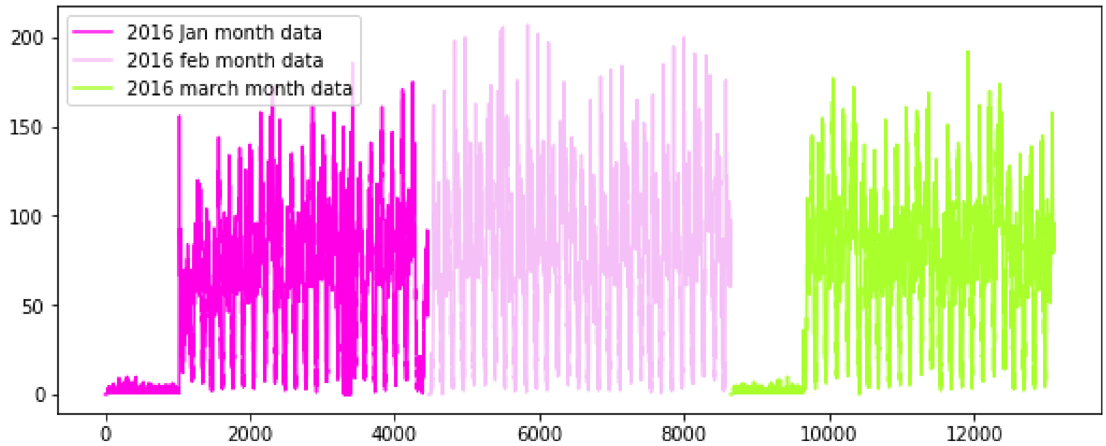

2016 Jan month data
2016 feb month data
2016 march month data

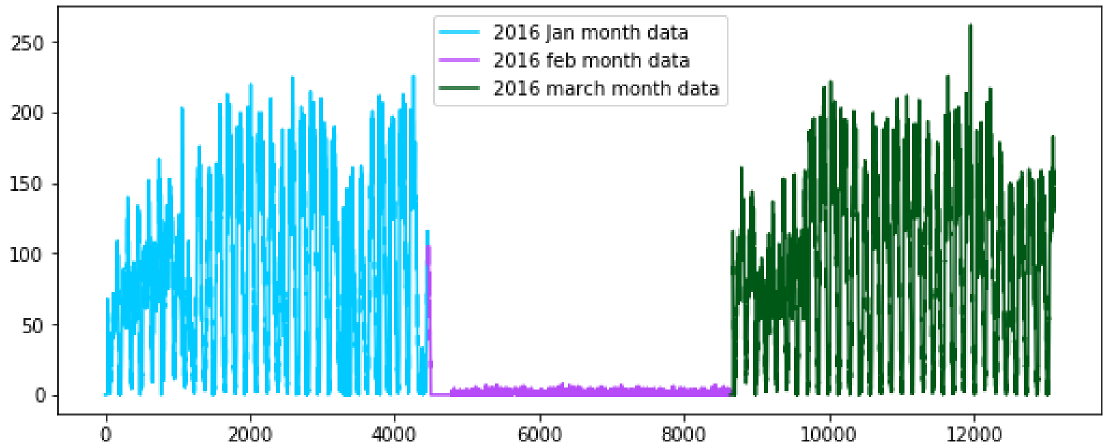

2016 Jan month data
2016 feb month data
2016 march month data

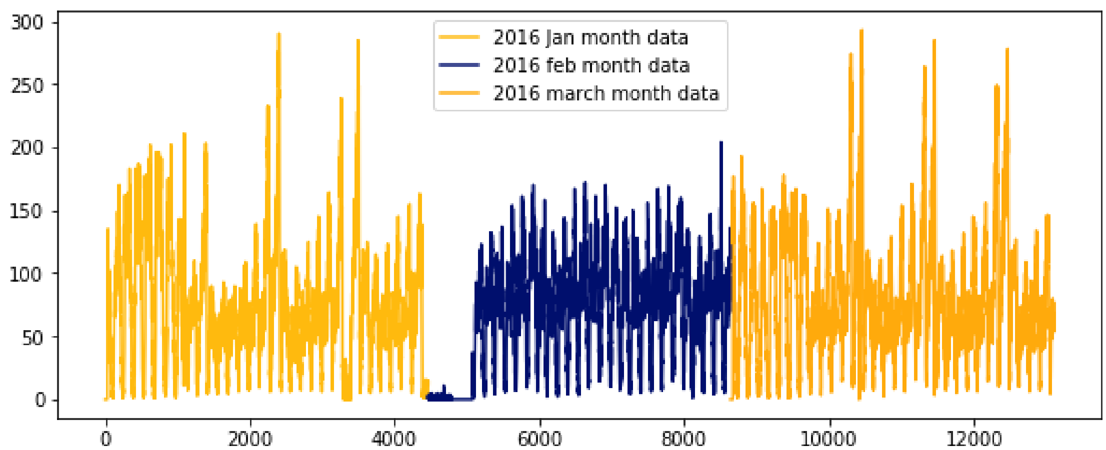

2016 Jan month data
2016 feb month data
2016 march month data

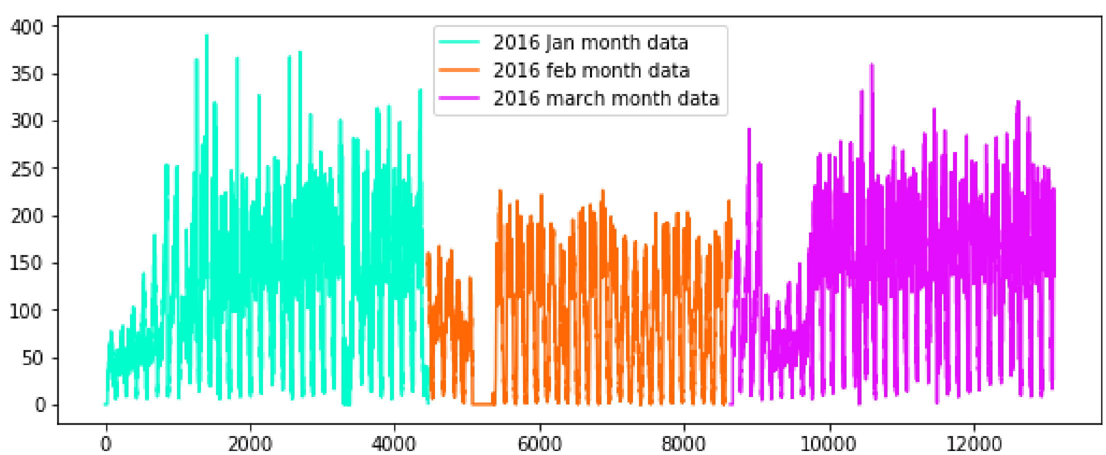

2016 Jan month data
2016 feb month data
2016 march month data
0
50
100
150
200
0
2000
4000
6000
8000
10000
12000

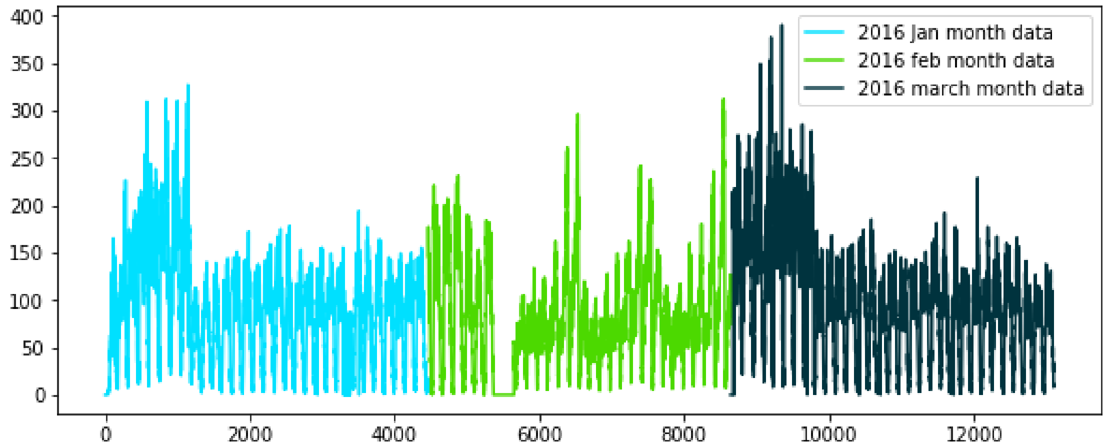

2016 Jan month data
2016 feb month data
2016 march month data
0
50
100
150
200
0
2000
4000
6000
8000
10000
12000

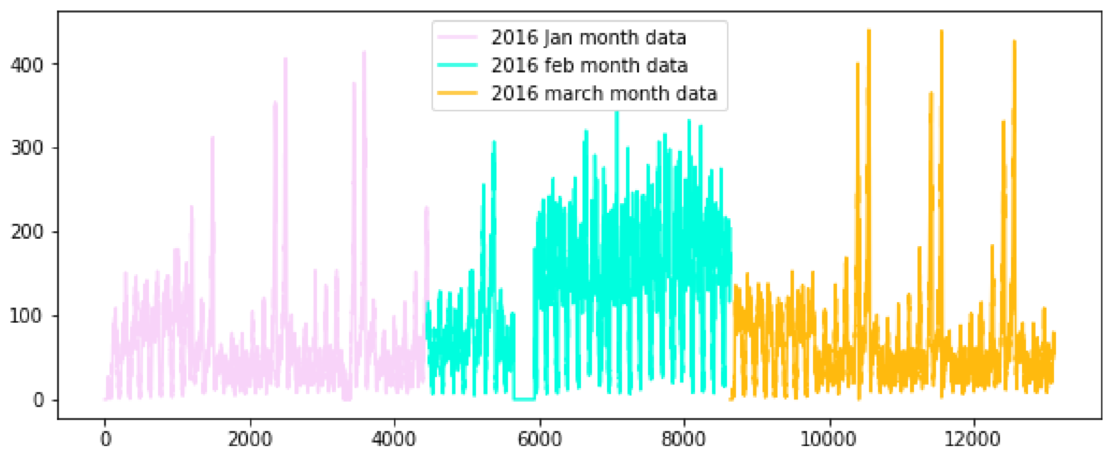

2016 Jan month data
2016 feb month data
2016 march month data
0
50
100
150
200
0
2000
4000
6000
8000
10000
12000

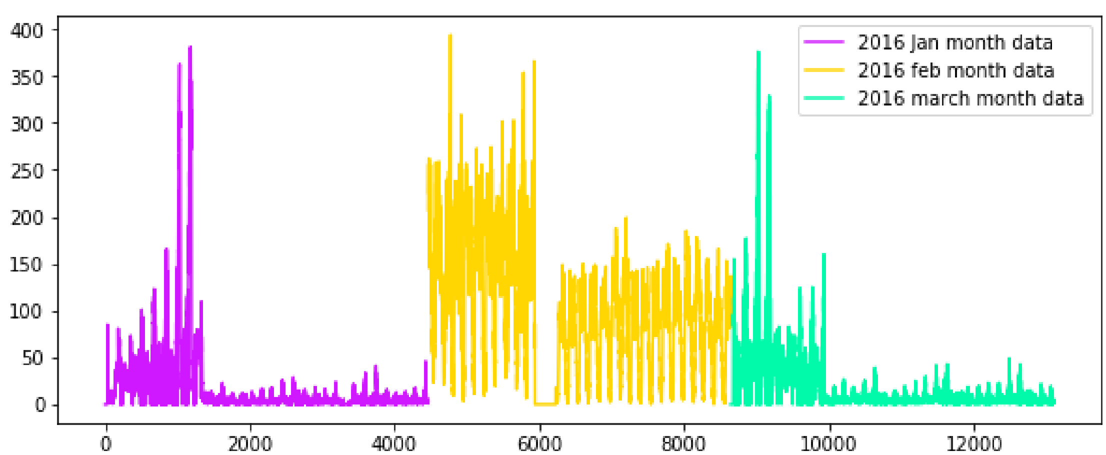

2016 Jan month data
2016 feb month data
2016 march month data
0
50
100
150
200
250
2000
4000
6000
8000
10000
12000

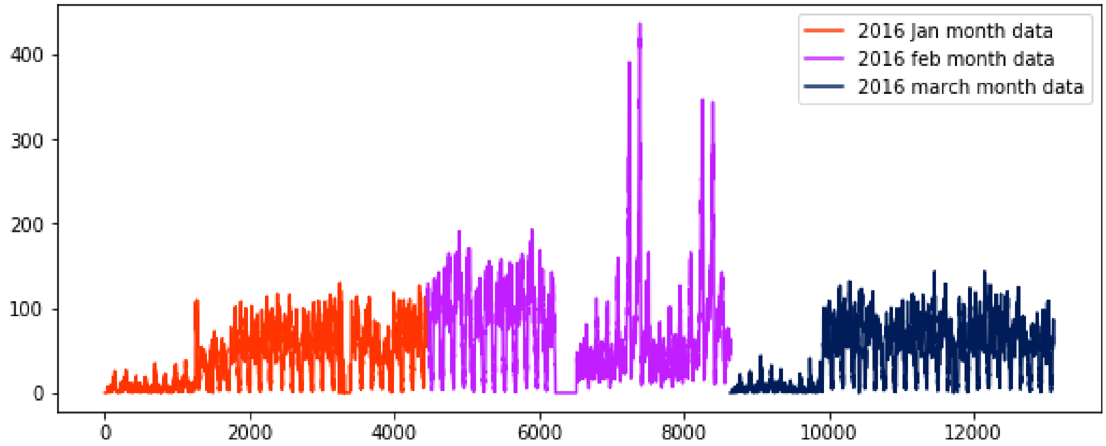

2016 Jan month data
2016 feb month data
2016 march month data
0
50
100
150
200
250
300
2000
4000
6000
8000
10000
12000

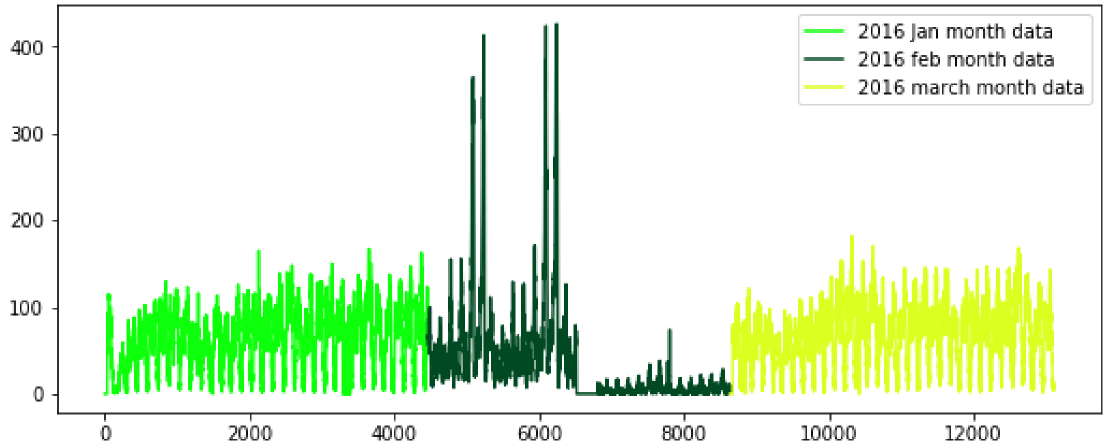

2016 Jan month data
2016 feb month data
2016 march month data
0
50
100
150
200
250
300
350
400
2000
4000
6000
8000
10000
12000


2016 Jan month data
2016 feb month data
2016 march month data


2016 Jan month data
2016 feb month data
2016 march month data


2016 Jan month data
2016 feb month data
2016 march month data





In [62]:
```
# getting peaks: https://blog.ytotech.com/2015/11/01/findpeaks-in-python/
# read more about fft function : https://docs.scipy.org/doc/numpy/reference/generated/
Y    = np.fft.fft(np.array(jan_2016_smooth)[0:4460])
# read more about the fftfreq: https://docs.scipy.org/doc/numpy/reference/generated/n
freq = np.fft.fftfreq(4460, 1)
n = len(freq)
plt.figure()
plt.plot( freq[:int(n/2)], np.abs(Y)[:int(n/2)] )
plt.xlabel("Frequency")
plt.ylabel("Amplitude")
plt.show()
```