**DATA ANALYSIS OF UBER AND OTHER FOR-HIRE-VEHICLES**

**NEW YORK CITY**

**Available Online at:** [Uber and FHV Data Analysis](https://github.com/SHAMINIPUTHOOPPALLIL/UBER-and-other-FHV-Data-Analysis)

Shamini Puthooppallil Baby

[Shaminibaby000@gmail.com](mailto:Shaminibaby000@gmail.com)

Masters In Data Science

[LinkedIn profile](https://www.linkedin.com/in/shamini-puthooppallil-baby-8b00947b)

Contents

[Introduction 4](#_Toc31576273)

[Back ground 5](#_Toc31576274)

[Data Preparation 6](#_Toc31576275)

[Data Source 6](#_Toc31576276)

[Exploratory Analysis 7](#_Toc31576277)

[UBER and Other FHV Taxi 7](#_Toc31576278)

[I. Growth of Uber Business against Other FHV vehicles in NYC City 7](#_Toc31576279)

[II. Monthly Growth of Uber Business against Other FHV vehicles in NYC City 8](#_Toc31576280)

[III. Monthly growth of FHV vehicles Business in NYC City 9](#_Toc31576281)

[IV. Growth of FHV Businesses ( 2014) 10](#_Toc31576282)

[V. Daily Business growth of Taxi Service in a month 10](#_Toc31576283)

[VI. Journeys of Taxi Service by day of week 11](#_Toc31576284)

[VII. Weekday and Weekend Analysis of NYC Taxi services 12](#_Toc31576285)

[Growth of Uber from 2014 to 2015 13](#_Toc31576286)

[I. Total Uber Pickup in 2014 and 2015 13](#_Toc31576287)

[Competitors of Uber 15](#_Toc31576288)

[Model Building – Enhance Uber Business within NYC City 16](#_Toc31576289)

[I. Latitude and longitude of cluster Centroids 16](#_Toc31576290)

[II. Cluster centroids in Google map 17](#_Toc31576291)

[III. Visualization of cluster wise ride request 18](#_Toc31576292)

[Discussion and Future scope 19](#_Toc31576293)

[Conclusion 20](#_Toc31576294)

[References 21](#_Toc31576295)

[List of Figures 22](#_Toc31576296)

[Appendix 23](#_Toc31576297)

[1. Data preparation and Pre validations 23](#_Toc31576298)

[2. Growth of Uber Business against Other FHV vehicles in NYC city 25](#_Toc31576299)

[3. Monthly Growth of Uber Business against Other FHV vehicles in NYC city 26](#_Toc31576300)

[4. Monthly growth of FHV vehicles Business in NYC City 26](#_Toc31576301)

[5. Growth of FHV Businesses (APR-SEP : 2014) 27](#_Toc31576302)

[6. Daily Business growth of Taxi Service in a month 27](#_Toc31576303)

[7. Journeys of Taxi Service by day of week 28](#_Toc31576304)

[8. New York Taxi week day statistics 28](#_Toc31576305)

[9. Individual Growth of Company Businesses in 2014 and 2015 28](#_Toc31576306)

[10. Uber’s Competitors 29](#_Toc31576307)

[11. Data preparation and Elbow method in K means clustering 29](#_Toc31576308)

[12. Cluster centers in a graph 31](#_Toc31576309)

[13. Cluster centers in a Google map 31](#_Toc31576310)

[14. Cluster wise Ride Request 32](#_Toc31576311)

# Introduction

This report describes data analysis of Uber pickups and other for-hire vehicles (FHV) within New York City. It also contains the descriptive and exploratory analysis of the taxi details on comparing Uber pickups with other for-hire vehicles over time.

The main aim is to analyse the data by comparing the productivity of Uber and other taxi services to help Uber, to enhance their business by providing meaningful insights from the results. The analysis is trying to find the best possible geographic zones to obtain maximum rides and thereby increase the overall productivity of Uber Taxi service. It also helps Uber to assign a new service request to the nearby pool of drivers who can arrive in small time frame within New York City

All the data analysis is performed using Python programming language with the help of Jupyter Notebook. The analysis starts with a data cleansing part followed by descriptive and exploratory analysis of Uber and Other taxi pickup data. Finally, the analysis segments the New York City to different geographic zones with the help of statistical modeling techniques .This will help business to make wise design about customer pickups.

# Back ground

Uber Technologies (UBER) [explosive growth and constant controversy](https://www.investopedia.com/terms/d/disruptive-technology.asp) made it one of the most fascinating companies to emerge over the past decade. The firm, founded in 2009, soon grew to become the highest valued private startup company in the world. With Uber's rapid growth came many controversies that knocked down the firm's valuation from a lofty $70 billion to $48 billion in its last funding round in Jan.

During its expansion, Uber has met fierce resistance from the taxi industry and government regulators. There are many competitors raised for Uber taxi from the beginning. Some of the known and strong opponents of Uber within NYC city was ‘Skyline’, ’Dial 7’and ‘Lyft’. Uber is utilizing the advantages of machine learning (ML), artificial intelligence (AI) and global positioning system (GPS) for ride pickup. Uber maintained their high standard in customer pickup and arrival time frame and vehicle facilities. This made Uber one of the most popular and profit earning organizations in the world.

# Data Preparation

Before starting the data analysis and modelling, the main duty of the data analyst is to collect relevant data to solve business problems. Once data collected; they have to make sure that the data provided is in correct format. If the dataset is not cleansed properly, the entire work needs to be repeated.

## Data Source

Data obtained from the [NYC Taxi & Limousine Commission (TLC)](http://www.nyc.gov/html/tlc/html/home/home.shtml) by submitting a Freedom of Information Law request on July 20, 2015. It contains 4.5 million Uber pickups in New York City from April to September 2014, and 14.3 million more Uber pickups from January to June 2015. Trip-level data on 10 other for-hire vehicle (FHV) companies, as well as aggregated data for 329 FHV companies, is also included.

Please find more about the data source and fields at: <https://github.com/fivethirtyeight/uber-tlc-foil-response>

Datasets are loaded to the Jupyter Notebook software and perform required pre validations for the loaded dataset. The analysis performed using the Python 3 and the common statistical tools under the available packages.

The files are validated and combined for the data preparation. Structure of the datasets, column values, null values and impossible values are checked using some basic Python data preprocessing packages. Sample data is viewed and make sure that the data loaded correctly to the R software.

This code snippet is available in [Appendix 1.](#_Data_preparation_and)

# Exploratory Analysis

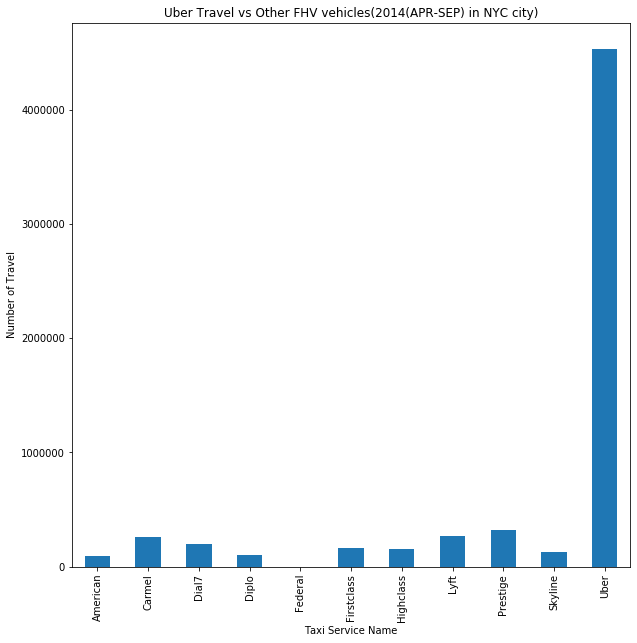
In this stage, the datasets is analyzed using detailed statistical and descriptive tools to get more insights of the time series characteristics.

The data analysis stage considered characteristics of the dataset and figured out the elements of suitable and successful data analysis. The characteristics of the datasets which can be used for the comparison of Uber taxi and other FHV vehicles are analyzed and made visualizations to prove the data if relevant. The findings from the analysis are listing below.

## UBER and Other FHV Taxi

The below observations directly compare the performance of Uber and other FHV vehicles in the New York City by analyzing different characteristics during the period April 2014 to September 2014

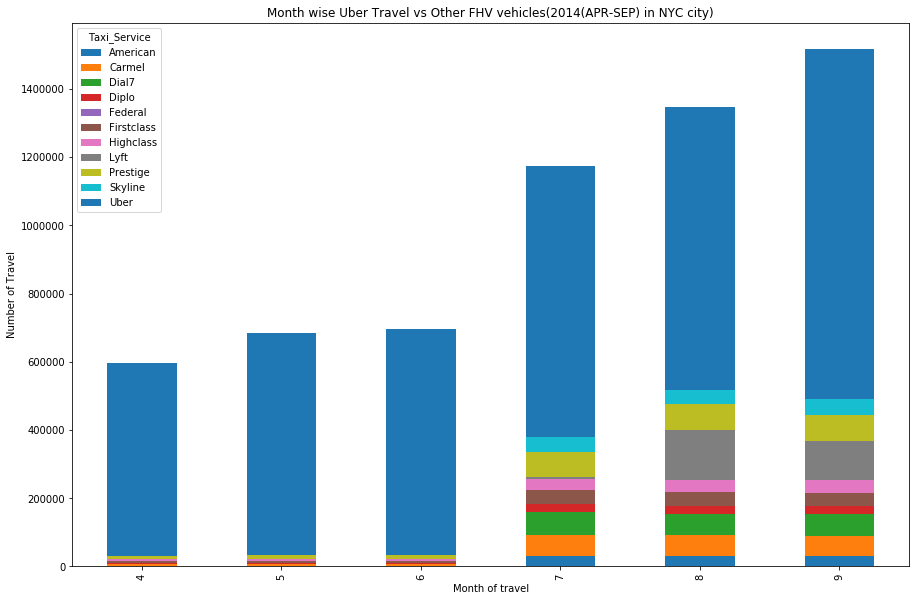
### Growth of Uber Business against Other FHV vehicles in NYC City



We can observe high difference in the number of Uber taxi pickups and other taxi services in the above figure. It is obvious that all other taxi picups contribute less than 50% of the overall taxi services during the period April 2014 to September 2014.While, Uber is processing more than 4 million requests, other taxi picups process less than 1 million request at the same time. ’Carmel’, ‘Prestige’ and ‘Lyft’ are the 3 taxi services which have higher ride request demand as compared to other FHV services.

The code snippet for this visualization is available in [Appendix 2](#_Growth_of_Uber).

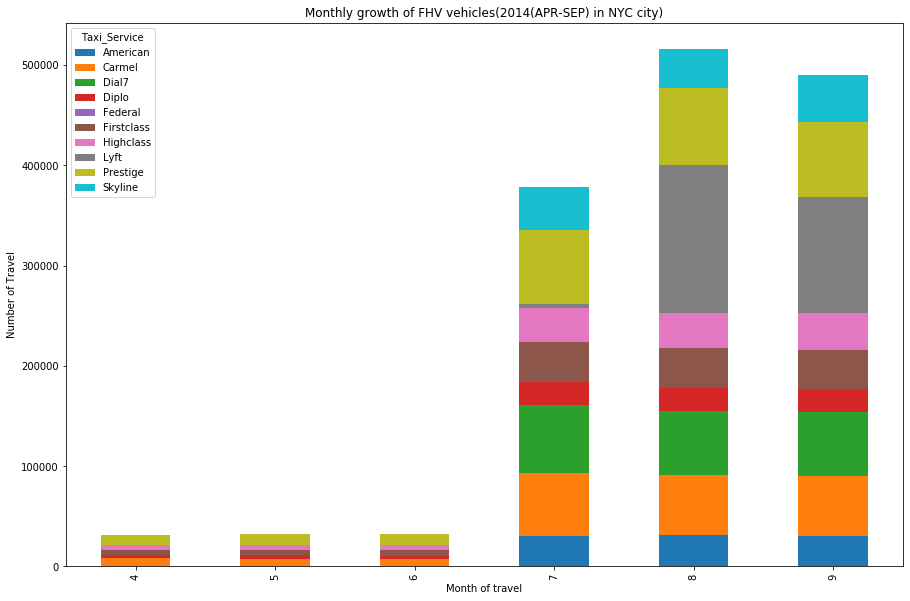
### Monthly Growth of Uber Business against Other FHV vehicles in NYC City



This plot shows a strong upward trend. Overall taxi request in the NYC city gradually increases from April with a value of 60k the growth is almost steady up to June. From July onwards, it shows high increase in the taxi ride requests. The interesting fact is that the highest contribution in every month is from Uber Taxi requests. We can clearly sense that during the initial 3 months, other taxi requests were too low as compared to Uber (<25%) but from July onwards, it is increasing. So from here we can conclude that, even though the Uber demand is high, other taxi services perform a business growth in their own space.

The code snippet for this visualization is available in [Appendix 3](#_Monthly_Growth_of).

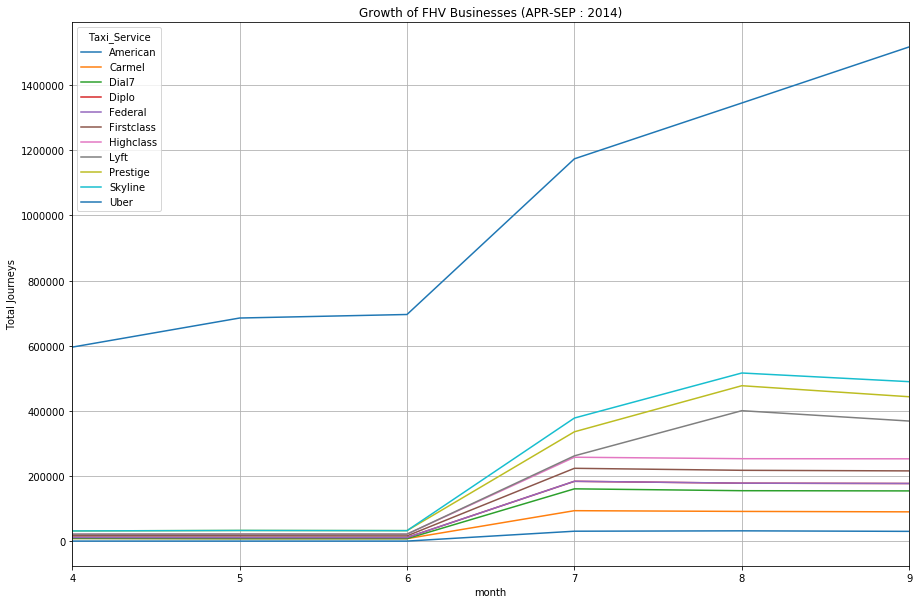
### Monthly growth of FHV vehicles Business in NYC City



As mentioned in the previous sessions, the growth of other vehicles shows a dramatic increase from April to September 2014.A detailed analysis done by plotting only the other taxi details without Uber. It is very clear that all of the FHV services are growing their business and they shows a very high upward trend in the number of taxi requests especially in the last 3 months. New York City is a great platform for FHV vehicles to enhance their business. There might be some healthy competitions between these taxi services, but these competitions not affecting the growth of Uber business as per the above data.

The code snippet for this visualization is available in [Appendix 4](#_Linear_Trend).

### Growth of FHV Businesses ( 2014)



This plot gives a detailed description of the business growth of Uber and other FHV vehicles in a single frame .As we can see; there is a huge difference in the growth curve of Uber and other taxi services. But surprisingly, all taxi services depict an upward growth curve. But, Uber shows a huge increase in the number of pickups from April to September. The growth of Uber is not even comparable with other services and the FHV services growth rate is very low as compared to Uber. Also, ‘Skyline, ‘Prestige’ and ‘Lyft’ are the services which are growing fast in other FHV category.

The code snippet for this visualization is available in [Appendix 5.](#_Growth_of_FHV)

### Daily Business growth of Taxi Service in a month

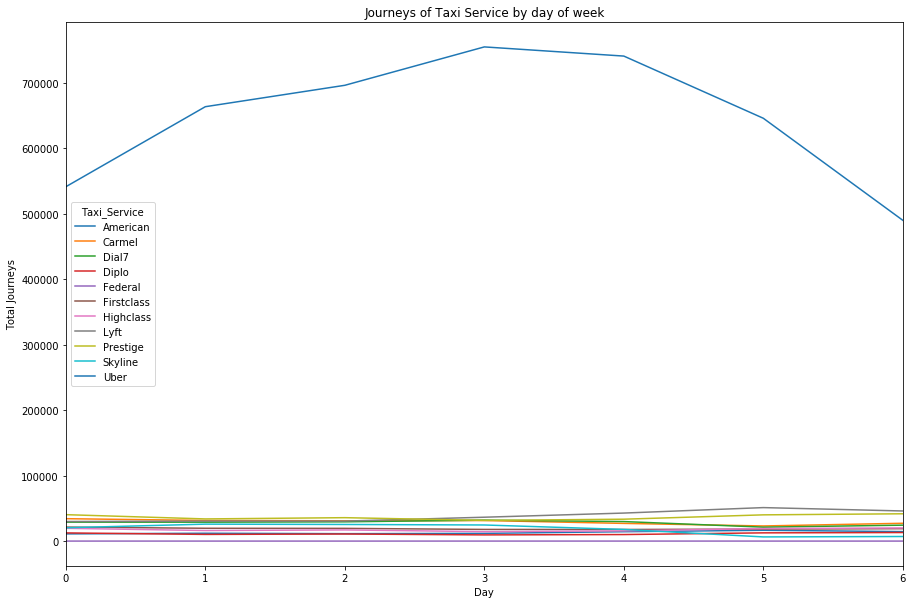
The below plots show the overall business growth of Uber Vehicles and Other FHV vehicles according to day of every month. This will help us to provide details of the taxi pickups from Uber and other taxi picups separately in a month.

|  |  |
| --- | --- |
|  |  |

From the above figure, we can see that Uber shows steady business all over the month and there are no much high fluctuations other than the month end. But other FHV vehicles shows fluctuations in business in every month. Their business is almost dull during first half of the month other than 2,3 days. But from the second half of the month, it shows a steady performance but which is too small as compared to Uber pickup counts.

The code snippet for this visualization is available in [Appendix 6](#_Daily_Business_growth).

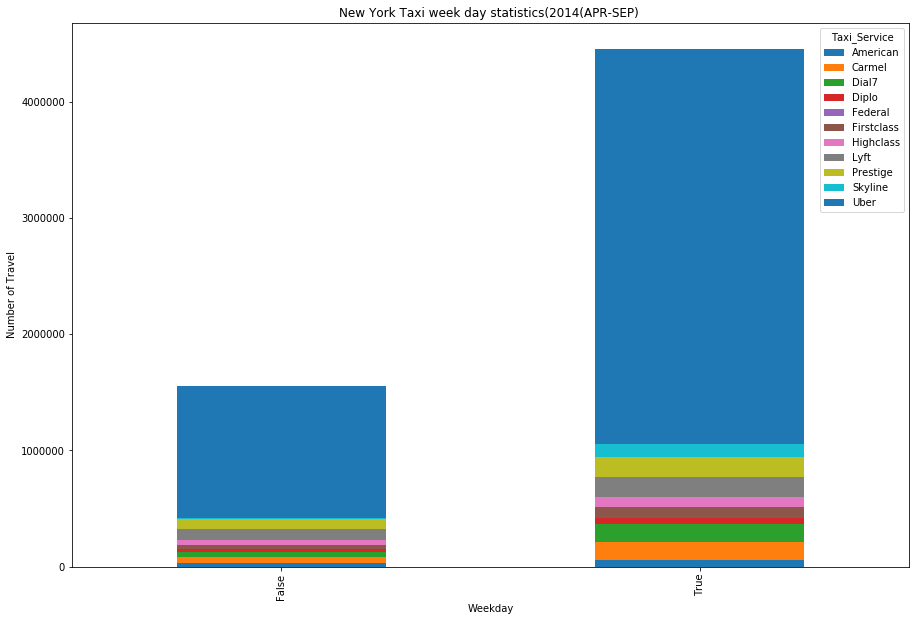
### Journeys of Taxi Service by day of week



The code snippet for this visualization is available in [Appendix 7](#_Journeys_of_Taxi).

Here we are doing a weekly analysis of Uber and Other FHV vehicle services. It clearly shows that the Uber services show a curved trend that is the business is high during the weekdays. It’s increasing up from Sunday to Wednesday and then decreasing back to Saturday. So in the case of Uber, the peak days are Tuesday, Wednesday and Thursday. But other taxi services act differently from this and shows a steady trend during weeks.

### Weekday and Weekend Analysis of NYC Taxi services



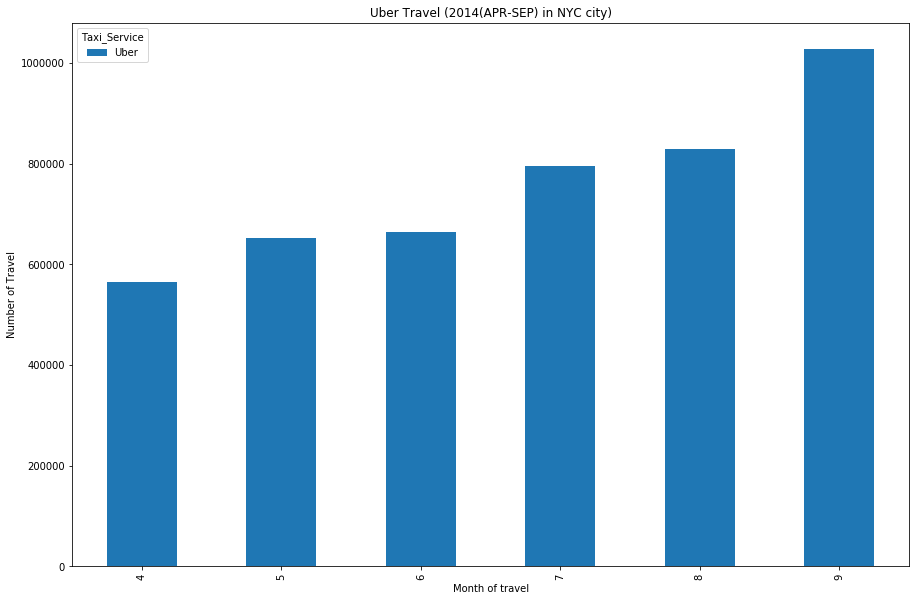
The above bar chart describes a comparison between week day and weekend statistics of Uber and other FHV vehicles pick up services. The Taxi business is more up in weekdays as compared to weekends. Overall performance and rides are really high in weekdays and it’s comparatively low in weekends. This may be due to the holidays and very few people are travelling for work those days. When we compare Uber and other FHV we can see that, Uber has the maximum number of rides in weekdays as well as weekends.

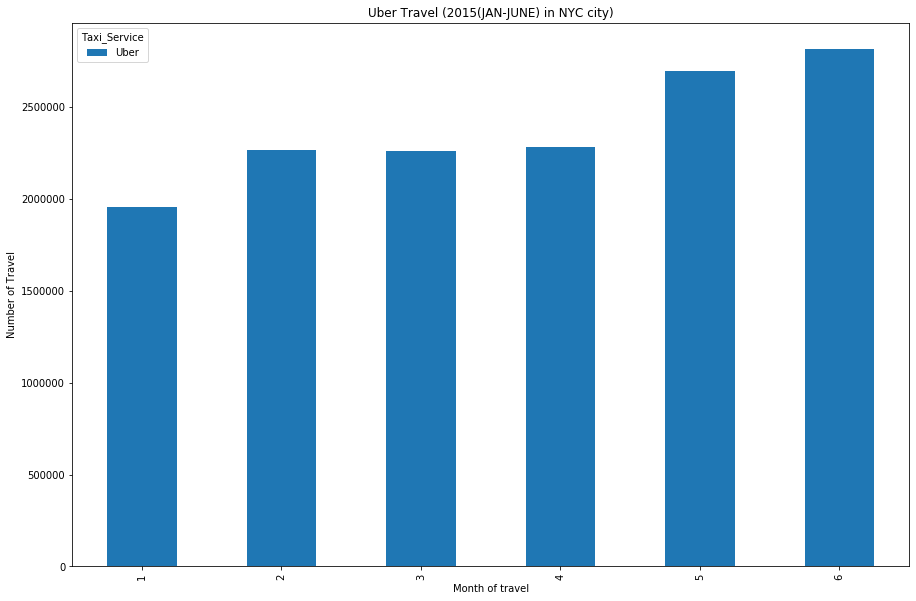
The code snippet for this visualization is available in [Appendix 8.](#_New_York_Taxi)

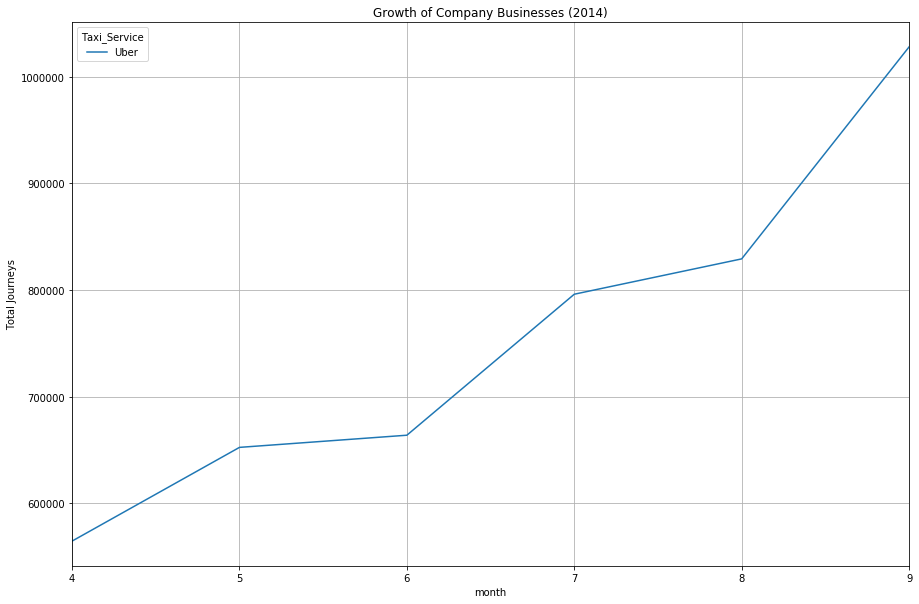
## Growth of Uber from 2014 to 2015

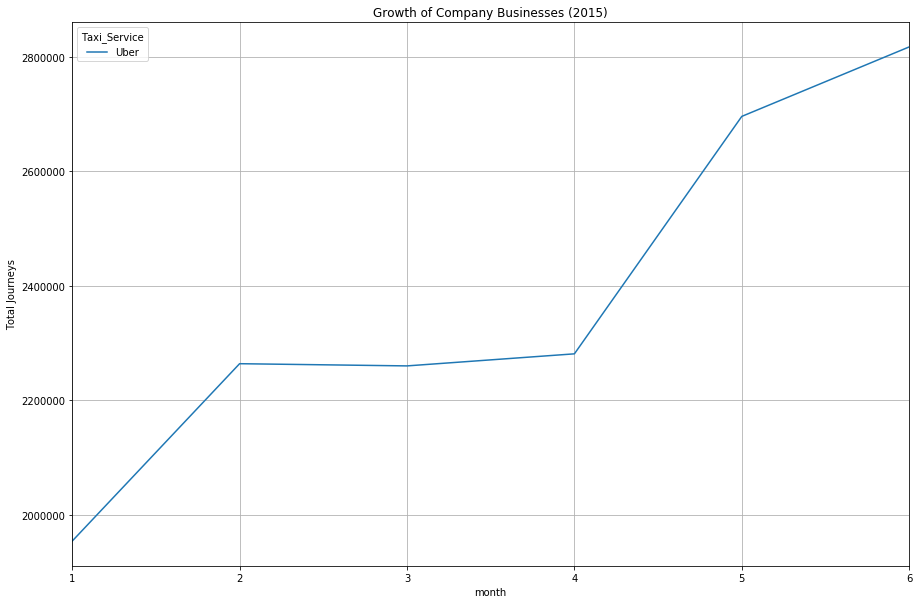
A comparison of the growth of Uber in the second half of 2014 and first half of 2015 are done with the help of statistical visualizations.

## Total Uber Pickup in 2014 and 2015









The above figures explain the growth of Uber taxies within New York from April 2014 to June 2015.We can see an exponential growth of business in case of Uber. At the end of September, the total number of rides in NYC was around 1 million. But, when we analyze thoroughly, we can see that at the beginning of January the ride count was starting from 2 million. This shows a huge hike in Uber taxies in NYC city.

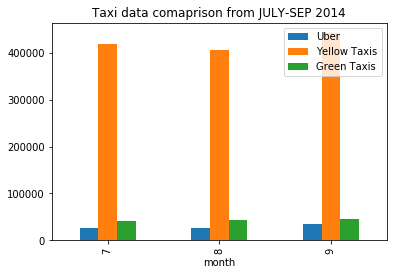
In the first 6 months of 2015, Uber doesn’t show a very good performance as expected. Increase in the number of Uber rides was around 1 million from September 2014 to Jan 2014,. But, at the end of June 2015, the ride count was just 2.8 million. This is 0.8 million increases from the beginning of 2015.This indicates that the overall Uber growth rate is decreased from 2014 to 2015

From this point onwards, this analysis is concentrating on the root cause analysis behind this slow growth and provides recommendations to increase the Uber business. Also, provide evidence from the available data.

The code snippet for this visualization is available in [Appendix 9](#_Individual_Growth_of).

## Competitors of Uber

Uber didn’t perform a steady growth as expected in the first half of 2015. So it is assume that some competitors rise of or grown up in the market at the same time. Below figure proves that that assumption is correct.



The above figure shows that Yellow Taxies and Green Taxies shows higher number of pickup rides than Uber. So, this may be one of the reasons behind Uber’s unexpected fall in growth. The data is available only for 3 months in 2014 and we can clearly see that Yellow taxi and green taxi growth is increasing over time. This steady trend might have followed to 2015 and this could affect the growth of Uber rides.

The code snippet for this visualization is available in [Appendix 10](#_Uber’s__Competitors).

# Model Building – Enhance Uber Business within NYC City

In this approach I’m trying to fit our Uber data to a clustering technique. The main objective of this process is to find a suitable model which is appropriate and help to enhance the Uber business within New York City.

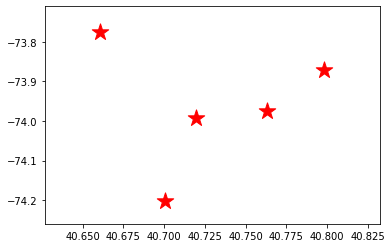
Clustering is one of the most common exploratory data analysis technique used to get an intuition about the structure of the data. It can be defined as the task of identifying subgroups in the data such that data points in the same subgroup (cluster) are very similar while data points in different clusters are very different. In other words, we try to find homogeneous subgroups within the data such that data points in each cluster are as similar as possible according to a similarity .The decision of which similarity measure to use is application-specific.

Here, I have applied a K-Means clustering algorithm whose main goal is to group similar elements or data points of Uber ridesharing data. “K” in K-means represents the number of clusters. The optimum value of K is found as 5 and proceeded the clustering with to find 5 meaningful segments.

The code snippet for this Data preparation and finding optimum value of K is available at [Appendix 11](#_Data_preparation_and_1).

## Latitude and longitude of cluster Centroids

Once the cluster number is fixed performed the K-means clustering with the help of Python libraries and plotting the cluster centroids or the geographical centers for Uber rides below.

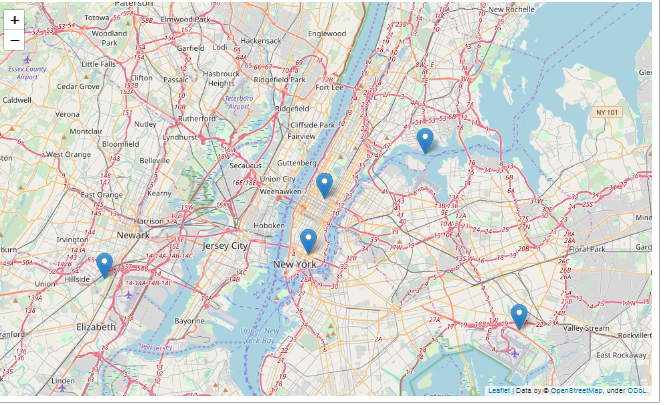
****

This doesn’t show any meaningful information. The above plot just tells us the latitude and longitude of 5 geographic centroids and these will act as the Uber taxi driver pool location for easy access to customers.

The code snippet for this visualization is available in [Appendix 12](#_Cluster_centers_in).

## Cluster centroids in Google map

Let’s plot the same in Google map (latitude & longitude) and visualize it in more detail. This will give us a general idea of the points we selected as the geographic centers.



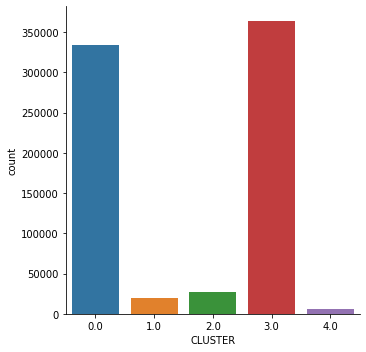
Here we can see that we got 5 geographic centers which can act as Uber hubs. Two of them are from Manhattan and others are from Bronx, Brooklyn and Newark respectively.

When Uber receive a new ride request, they can check the closeness with each of these centroids. Whichever particular centroid is closer .Then the Uber can direct the vehicle from that particular location to the customer location. If they are getting a lot of ride request then, they can place their driver’s in good location so that probabilities of getting a ride request are huge.

The code snippet for this visualization is available in [Appendix 13.](#_Cluster_centers_in_1)

## Visualization of cluster wise ride request

As a data analyst, once the segmentation is done, next step is mapping the clusters back to original data and find out which cluster has the maximum ride request. This step is crucial because as a business perspective we always eager to know the spots which have the maximum ride request. If we can provide more services to that areas, the net growth of the company can be increase with little effort.



Cluster 3 received maximum ride request followed by cluster 0. Cluster 4 received the least request. Uber can place more vehicles in Cluster 3 to meet higher demands

The code snippet for this visualization is available in [Appendix 14](#_Cluster_wise_Ride).

# Discussion and Future scope

The main challenges of this analysis were the size of the dataset and the time frame allotted to process the data. The details of Uber and other taxies were scattered in different files. As a first step, data merging and feature extraction was performed. Since the analysis performed in local system, considerable amount of time was utilized for these data cleansing and feature extraction steps. This report contains data analysis of Uber taxi and other FHV data in a high level perspective and tried to elaborate the areas which I found interesting. A clustering algorithm was developed to find more interesting facts about dataset.

Geographical location details calculation from other taxi pickup details was a time consuming task. As a future goal, geographical aspects of other FHV vehicles can be explored and a detailed analysis of other FHV vehicles can be done according to the geographical location. In addition to that, Base Number of Uber taxi and other FHV vehicles are not considered in this analysis, a study of these details can provide more insights to the business. The entire analysis can be performed in a better cloud environment having high performance processors .This can save the time frame needed for huge data processing.

# Conclusion

The main goal of this data analysis was to compare the Uber taxi ride and other FHV vehicles ride within New York City. A set of data for Uber and other vehicles were available for analysis. During the initial analysis data cleansing, basic exploration analysis was performed.

A detailed comparison between Uber taxi details were done with the help of strong python packages and visualizations. Growth of Uber was really appreciable and high as compared to the other taxi pickups. During the second set of analysis, it’s noticed that the growth rate of Uber in the first half of 2015 was not as expected. A root cause analysis performed to find the reason behind this unexpected behavior. Yellow taxi and Green taxi was more in demand during the period July to September 2014.

To enhance the performance of Uber taxi demand within New York City, a clustering algorithm was proposed and found 5 meaningful geometric clusters. A detailed analysis of these cluster centers was performed to find out the maximum taxi request geometric zones within NYC city. This will help the business to put more vehicles on these particular zones to increase the business growth.

# References

1. <https://www.uber.com/global/en/cities/new-york/>
2. <http://benalexkeen.com/k-means-clustering-in-python/>
3. <https://www.geeksforgeeks.org/python-plotting-google-map-using-folium-package/>
4. <https://data.cityofnewyork.us/Transportation/uber-Data/3jeu-mn7j>
5. [https://towardsdatascience.com](https://towardsdatascience.com/how-does-uber-use-clustering-43b21e3e6b7d)

# List of Figures

1. [Growth of Uber Business against Other FHV vehicles in NYC City](#_Growth_of_Uber_1)
2. [Monthly Growth of Uber Business against Other FHV vehicles in NYC City](#_Growth_of_Uber)
3. [Monthly growth of FHV vehicles Business in NYC City](#_Linear_Trend)
4. [Growth of FHV Businesses ( 2014)](#_Quadratic_Trend)
5. [Daily Business growth of Taxi Service in a month](#_Daily_Business_growth)
6. [Journeys of Taxi Service by day of week](#_Splitting__Data)
7. [Weekday and Weekend Analysis of NYC Taxi services](#_New_York_Taxi)
8. [Total Uber Pickup in 2014 and 2015](#_Residual_Analysis_-)
9. [Competitors of Uber](#_Uber’s__Competitors)
10. [Elbow Curve](#_Data_preparation_and_1)
11. [Latitude and longitude of cluster Centroids](#_Cluster_centers_in)
12. [Cluster centroids in Google map](#_Cluster_centers_in_1)
13. [Visualization of cluster wise ride request](#_Cluster_wise_Ride)

# Appendix

Please find the attached python notebook for this entire analysis.



## 

## Data preparation and Pre validations

#Import available libraries

import numpy as np

import pandas as pd # data processing, CSV file I/O (e.g. pd.read\_csv)

import os

import matplotlib.pyplot as plt

#Read Uber Data

UBER\_apr14 = pd.read\_csv('uber-raw-data-apr14.csv')

UBER\_may14 = pd.read\_csv('uber-raw-data-may14.csv')

UBER\_jun14 = pd.read\_csv('uber-raw-data-jun14.csv')

UBER\_jul14 = pd.read\_csv('uber-raw-data-jul14.csv')

UBER\_aug14 = pd.read\_csv('uber-raw-data-aug14.csv')

UBER\_sep14 = pd.read\_csv('uber-raw-data-sep14.csv')

UBER\_2015\_jan\_june = pd.read\_csv('uber-raw-data-janjune-15.csv')

#Read Other taxi Data

#read csv files

American\_B01362=pd.read\_csv('American\_B01362.csv')

Dial7\_B00887=pd.read\_csv('Dial7\_B00887.csv')

Federal\_02216=pd.read\_csv('Federal\_02216.csv')

Lyft\_B02510=pd.read\_csv('Lyft\_B02510.csv')

Skyline\_B00111=pd.read\_csv('Skyline\_B00111.csv')

#read xlsx files

Carmel\_B00256=pd.read\_excel('Carmel\_B00256.xlsx')

Diplo\_B01196=pd.read\_excel('Diplo\_B01196.xlsx')

Firstclass\_B01536=pd.read\_excel('Firstclass\_B01536.xlsx')

Highclass\_B01717=pd.read\_excel('Highclass\_B01717.xlsx')

Prestige\_B01338=pd.read\_excel('Prestige\_B01338.xlsx')

#Rename the Date column -Pickup for symmetric fields

Dial7\_B00887.rename(columns={'Date': 'PickupDatetime'}, inplace=True)

Lyft\_B02510.rename(columns={'time\_of\_trip': 'PickupDatetime'}, inplace=True)

Skyline\_B00111.rename(columns={'Date': 'PickupDatetime'}, inplace=True)

American\_B01362.rename(columns={'DATE': 'PickupDatetime'}, inplace=True)

Federal\_02216.rename(columns={'Date': 'PickupDatetime'}, inplace=True)

Carmel\_B00256.rename(columns={'Date': 'PickupDatetime'}, inplace=True)

Diplo\_B01196.rename(columns={'Date': 'PickupDatetime'}, inplace=True)

Firstclass\_B01536.rename(columns={'DATE': 'PickupDatetime'}, inplace=True)

Highclass\_B01717.rename(columns={'DATE': 'PickupDatetime'}, inplace=True)

Prestige\_B01338.rename(columns={'DATE': 'PickupDatetime'}, inplace=True)

UBER\_2015\_jan\_june.rename(columns={'Pickup\_date ': 'PickupDatetime'}, inplace=True)

#Bind 4 months data and make one data frame

uber\_data = pd.concat([UBER\_apr14, UBER\_may14, UBER\_jun14, UBER\_jul14, UBER\_aug14,UBER\_sep14])

#check the shape to make sure that all the rows added

uber\_data.shape

#combine all the data from other FHV

other\_taxi\_data = pd.concat([Dial7\_B00887\_1, Lyft\_B02510\_1, Skyline\_B00111\_1, American\_B01362\_1, Federal\_02216\_1,Carmel\_B00256\_1,Diplo\_B01196\_1,Firstclass\_B01536\_1,Highclass\_B01717\_1,Prestige\_B01338\_1])

#merge all taxi data and do primary checks

NYC\_TAXI\_data = pd.concat([other\_taxi\_data,uber\_data])

#check the shape and datatypes

print(NYC\_TAXI\_data.shape)

print(NYC\_TAXI\_data.dtypes)

### 

## Growth of Uber Business against Other FHV vehicles in NYC city

# Uber Business and other taxi analysis in new york city

NYC\_TAXI\_data.groupby(['Taxi\_Service']).count().unstack('Taxi\_Service')['PickupDatetime'].plot(kind='bar', figsize = (10,10),stacked=True)

plt.ylabel('Number of Travel')

plt.xlabel('Taxi Service Name')

plt.title('Uber Travel vs Other FHV vehicles(2014(APR-SEP) in NYC city)');

## Monthly Growth of Uber Business against Other FHV vehicles in NYC city

#convert pickup date to date time format

NYC\_TAXI\_data['PickupDatetime'] = pd.to\_datetime(NYC\_TAXI\_data['PickupDatetime'])

#create month field and filter unwanted data(data is between apriland september 2014)

NYC\_TAXI\_data['month'] = NYC\_TAXI\_data['PickupDatetime'].dt.month

NYC\_TAXI\_data = NYC\_TAXI\_data[(NYC\_TAXI\_data['month']>3) & (NYC\_TAXI\_data['month']<10)]

NYC\_TAXI\_data.groupby(['month','Taxi\_Service']).count().unstack('Taxi\_Service')['PickupDatetime'].plot(kind='bar', figsize = (15,10),stacked=True)

plt.ylabel('Number of Travel')

plt.xlabel('Month of travel')

plt.title('Uber Travel vs Other FHV vehicles(2014(APR-SEP) in NYC city)');

## Monthly growth of FHV vehicles Business in NYC City

#convert pickup date to date time format

other\_taxi\_data['PickupDatetime'] = pd.to\_datetime(other\_taxi\_data['PickupDatetime'])

#create month field and filter unwanted data(data is between apriland september 2014)

other\_taxi\_data['month'] = other\_taxi\_data['PickupDatetime'].dt.month

other\_taxi\_data = other\_taxi\_data[(other\_taxi\_data['month']>3) & (other\_taxi\_data['month']<10)]

other\_taxi\_data.groupby(['month','Taxi\_Service']).count().unstack('Taxi\_Service')['PickupDatetime'].plot(kind='bar', figsize = (15,10),stacked=True)

plt.ylabel('Number of Travel')

plt.xlabel('Month of travel')

plt.title('Monthly growth of FHV vehicles(2014(APR-SEP) in NYC city)');

The trend line is plotted Using the below R code:

plot(ozone\_thickness,type='o',ylab='y')  
plot.new

abline(model.ozone\_thickness.ln)

## Growth of FHV Businesses (APR-SEP : 2014)

#Individual Growth of Company Businesses in 2014

NYC\_TAXI\_data.groupby(['month','Taxi\_Service']).count().unstack('Taxi\_Service')['PickupDatetime'].plot(figsize = (15,10),stacked=True)

plt.ylabel('Total Journeys')

plt.title('Growth of FHV Businesses (APR-SEP : 2014)');

plt.grid()

# Uber did not seem to hurt the business of other companies in 2014

# as all other companies experienced a growth in their business along with Uber

## Daily Business growth of Taxi Service in a month

The below R chunk will fit the data into a harmonic model.

# UBER JOURNEY BY MONTH

uber\_data.groupby('MonthDayNum').count()['Taxi\_Service'].plot(kind='bar', figsize = (8,6))

plt.ylabel('Total Journeys')

plt.Xlabel('Day Of Month')

plt.title('Uber Journeys by Month Day');

# OTHER FHV JOURNEY BY MONTH

other\_taxi\_data.groupby('MonthDayNum').count()['Taxi\_Service'].plot(kind='bar', figsize = (8,6))

plt.ylabel('Total Journeys')

plt.Xlabel('Day Of Month')

plt.title('Other FHV Journeys by Month Day');

lines(as.vector(ozone\_thickness),type="o")

## Journeys of Taxi Service by day of week

#*Taxi service growth by week day*

*NYC\_TAXI\_data.groupby(['DayOfWeekNum','Taxi\_Service']).count().unstack('Taxi\_Service')['PickupDatetime'].plot(figsize = (15,10))*

*plt.ylabel('Total Journeys')*

*plt.xlabel('Day')*

*plt.title('Journeys of Taxi Service by day of week');*

## New York Taxi week day statistics

#weekday/weekend statistics

NYC\_TAXI\_data.groupby(['week/Weekend','Taxi\_Service']).count().unstack('Taxi\_Service')['PickupDatetime'].plot(kind='bar', figsize = (15,10),stacked=True)

plt.ylabel('Number of Travel')

plt.xlabel('Month of travel')

plt.title('New York Taxi week day statistics(2014(APR-SEP) ');

## Individual Growth of Company Businesses in 2014 and 2015

#Individual Growth of Company Businesses in 2015

UBER\_2015\_jan\_june.groupby(['month','Taxi\_Service']).count().unstack('Taxi\_Service')['PickupDatetime'].plot(figsize = (15,10),stacked=True)

plt.ylabel('Total Journeys')

plt.title('Growth of Company Businesses (2015)');

plt.grid()

# Uber did not seem to hurt the business of other companies in 2014

#Individual Growth of Company Businesses in 2015

UBER\_2015\_jan\_june.groupby(['month','Taxi\_Service']).count().unstack('Taxi\_Service')['PickupDatetime'].plot(figsize = (15,10),stacked=True)

plt.ylabel('Total Journeys')

plt.title('Growth of Company Businesses (2015)');

plt.grid()

## Uber’s Competitors

#Competitors of Uber

y= AggregateFHVData\_1.set\_index('month')

z=y.groupby('month').mean()

z.plot.bar(stacked=False,title="Taxi data comaprison from JULY-SEP 2014")

## Data preparation and Elbow method in K means clustering

#take the relevant data fields for clustering algorithm

cluster=uber\_data\_backup[['Lat','Lon']]

#import necessary libraries

# Import required packages

import pandas as pd

from sklearn.preprocessing import MinMaxScaler

from sklearn.cluster import KMeans

import matplotlib.pyplot as plt

#transform data

mms = MinMaxScaler()

mms.fit(cluster)

data\_transformed = mms.transform(cluster)

#Optimum value of K

Sum\_of\_squared\_distances = []

K = range(1,15)

for k in K:

km = KMeans(n\_clusters=k)

km = km.fit(data\_transformed)

Sum\_of\_squared\_distances.append(km.inertia\_)

#Elbow method to find optimal value of K

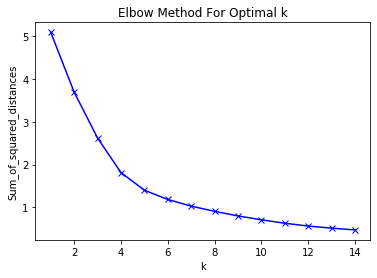
plt.plot(K, Sum\_of\_squared\_distances, 'bx-')

plt.xlabel('k')

plt.ylabel('Sum\_of\_squared\_distances')

plt.title('Elbow Method For Optimal k')

plt.show()

****

## Cluster centers in a graph

**#**plot the centers

plt.scatter(km.cluster\_centers\_[:, 0], km.cluster\_centers\_[:, 1], s=300, c='red',marker='\*')

plt.show()

## Cluster centers in a Google map

import folium

# Plotting the centroids on google map using Folium library.

map = folium.Map(location=[ 40.76317395 ,-73.97565024], zoom\_start = 10)

folium.Marker(location=[ 40.76317395 ,-73.97565024], popup = [ 40.76317395 ,-73.97565024]).add\_to(map)

folium.Marker(location=[40.66077836 ,-73.77532178], popup = [40.66077836 ,-73.77532178]).add\_to(map)

folium.Marker(location=[40.79822769 ,-73.87235133], popup = [40.79822769 ,-73.87235133]).add\_to(map)

folium.Marker(location=[40.71972119 ,-73.99233214], popup = [40.71972119 ,-73.99233214]).add\_to(map)

folium.Marker(location=[40.70076096 ,-74.2026252], popup = [40.70076096 ,-74.2026252]).add\_to(map)

map

## Cluster wise Ride Request

#plot the clusters back to data

import seaborn as sb

sb.factorplot(data=uber\_data\_backup[0:750000],x='CLUSTER',kind='count')