PANIMALAR ENGINEERING COLLEGE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Project Review III , 14-06-21

DATA ANALYSIS : TAXI PASSENGER TRAVEL SPATIAL AND TEMPORAL CHARACTERISTICS ANALYSIS AND APPLICATION BASED ON RIDE SOURCING DATA

Project Guide:
Prof. Mohan

M Eharrsha narayanaa-211417104057
N Jagadish kumar-211417104088
N Joshua dhanraj-2114117104100

Batch Number: E21

Abstract

- Nowadays, with the development of humanistic consciousness, the residents' travel behavior is becoming more and more important to be considered in urban planning, and has become an important reference for urban traffic construction . The ride sourcing software like Uber and Didi have been widely accepted and used. As a model of travel, the ride sourcing has many features, like convenience and flexibility, and the origin and destination of every trip are completely determined by passengers, the running track of vehicles can directly reflect the travel behavior of urban residents. So this paper will study the passengers' travel behavior by data mining based on the Didi order data, try to reveal the travel characteristics of urban residents from a macro perspective. This paper focuses on these issues: to analyze the order data from ride sourcing company Didi; to explore the temporal characteristics of passengers' travel from different angles like the temporal distribution of travel, the distribution of peak hour, time consumption; to analyze the spatial characteristics from the different aspects including the distribution of travel distance, travel hot spots; to propose an optimization model of the taxi stand location by analyzing travel characteristics, which can improve user's experience of residents by setting up reasonable site layout. Finally we found that the ride sourcing track data can well reveal travel characteristics of urban residents, which could be helpful for the urban planning and road network optimization.

Literature Survey

❖ Real-Time Uber Data Analysis of Popular Uber Locations in Kubernetes Environment

- There is a growing demand for Big Data applications to extract and evaluate information, which will provide the necessary knowledge that will help us make important rational decisions

- Uber is using real-time Big Data to perfect its processes, from calculating Uber's pricing to finding the optimal positioning of taxis to maximize profits. Real-time data analysis is very challenging for the implementation because we need to process data in real-time, if we use Big Data, it is more complex than before.

- Implementation of real-time data analysis by Uber to identify their popular pickups would be advantageous in various ways. So far no research has been done on real-time analysis for identifying popular Uber locations within Big Data in a distributed environment, particularly on the Kubernetes environment.

- To address these issues, we have created a machine learning model with a Spark framework to identify the popular Uber locations.

- The future development will consist of visualizing the real-time popular Uber locations on Google map.

An approach to predict taxi-passenger demand using quantitative histogram on Uber data

- The precise prediction of the day to day and monthly transactions is of great value for companies. This information can be beneficial for the companies in analyzing their ups and downs and draw other plans.

- This paper presents the use of data analytics in analyzing the transaction dataset provided by Uber to predict the possible outcomes and the changes to be made.

- The histograms and heat maps drawn provide us a clear visualization of the dataset and we must predict the rest out of it.

❖ A Preliminary Exploration of Uber Data as an Indicator of Urban Liveability

- Urban liveability is a key concept in the New Urban Agenda (NUA) adopted by the United Nations (UN) in 2016. The UN has recognized that effective benchmarks and monitoring mechanisms are essential for the successful implementation of the NUA. However, the timely and cost effective collection of objective international quality of life urban data remains a significant challenge.

- This paper explores the use of Uber data as a simple real-time indicator of urban liveability.

- Using data from the Uber Ride Request (URR) API for the Brazilian city of Natal, our preliminary findings suggest that Uber Estimated Time to Arrive (ETA) data is strongly correlated with selected quality of life indicators at a neighborhood and region level.

- This preliminary study finds strong evidence that Uber data can provide a simple, comparable, low cost, international urban liveability indicator at both city and neighborhood level for urban policy setting and planning.

❖ Green Cabs vs. Uber in New York City

- This paper reports on the process and outcomes of big data analytics of ride records for Green cabs and Uber in the outer boroughs of New York City (NYC), USA.

- The problem investigated revolves around where exactly Green cabs are losing market share to Uber outside Manhattan and what, if any, measures can be taken to preserve market share?

- Two datasets were included in the analysis including all rides of Green cabs and Uber respectively from April-September 2014 in New York excluding Manhattan and NYC's two airports.

- Tableau was used as the visual analytics tool, and PostgreSQL in combination with PostGIS was used as the data processing engine.

- Our findings show that the performance of Green cabs in isolated zip codes differ significantly, and that Uber is growing faster than Green cabs in general and especially in the areas close to Manhattan.

Intelligent Analysis of City Residents Mobility Data for Transport Simulation

- Data analysis is one of the key elements in the process of investigating of the transport processes in cities. Despite the large number of scientific papers in this area, the problem of obtaining of the urban population mobility data is not well studied.

- This paper describes a method for obtaining and processing population mobility open data taken from the Uber Movement source, as well as creating a theoretical model of population mobility based on people density distribution data.

- The developed method was tested on the example of the city of Amsterdam. In addition, a comparative analysis was carried out using open source data and data computed by developed method.

- As a result, we get a working method for obtaining population mobility data and analyzing these data.

❖ Technology Stack

NumPy

- NumPy is the fundamental package for scientific computing with Python. It contains among other things: a powerful N-dimensional array object sophisticated (broadcasting) functions tools for integrating C/C++ and Fortran code useful linear algebra, Fourier transform, and random number capabilities Besides its obvious scientific uses, NumPy can also be used as an efficient multi- dimensional container of generic data. Arbitrary data-types can be defined. This allows NumPy to seamlessly and speedily integrate with a wide variety of databases.

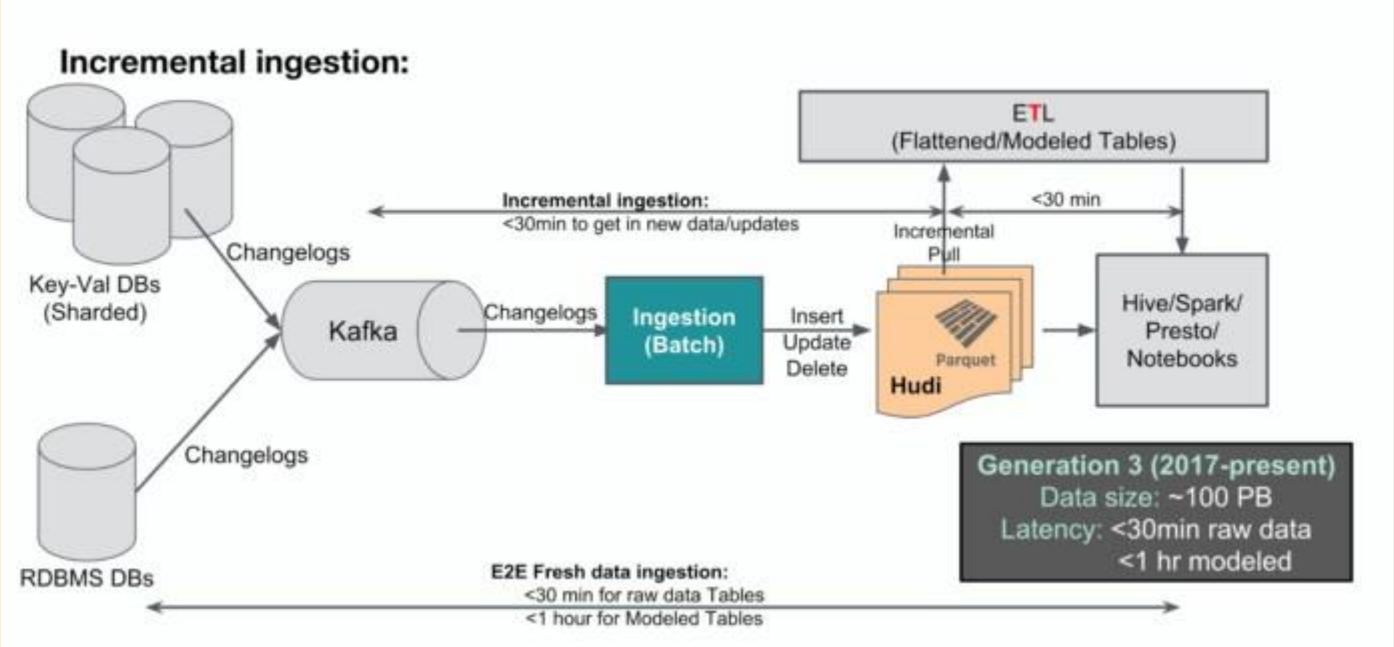
Pandas

- Pandas is an open source, BSD-licensed library providing high-performance, easy- to-use data structures and data analysis tools for the Python programming language. pandas is a NumFOCUS sponsored project. This will help ensure the success of development of pandas as a world-class open-source project, and makes it possible to donate to the project.

MATPLOTLIB

- Matplotlib is a plotting library for the Python programming language and its numericalmathematics extension NumPy. it provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt or GTK+.There is also a procedural "pylab" interface based on a state machine (like OpenGL),designed to closely resemble thatof MATLAB, though its use is discouraged. SciPy makes use of matplotlib.

❖ SYSTEM ARCHITECTURE



- By early 2017, our Big Data platform was used by engineering and operations teams across the company, enabling them to access new and historical data all in one place.

- Users could easily access data in Hive, Presto, Spark, Vertica, Notebook, and more warehouse options all through a single UI portal tailored to their needs. With over 100 petabytes of data in HDFS, 100,000 vcores in our compute cluster, 100,000 Presto queries per day, 10,000 Spark jobs per day, and 20,000 Hive queries per day, our Hadoop analytics architecture was hitting scalability limitations and many services were affected by high data latency.

- Fortunately, since our underlying infrastructure was horizontally scalable to address the immediate business needs, we had enough time to study our data content, data access patterns, and user-specific requirements to identify the most pressing concerns before building the next generation. Our research revealed four main pain points:

❖ Implementation of algorithms

EXISTING SYSTEM

- Customers are often dissatisfied with traditional cab companies because of their high prices and long waiting time and hence can exploit new and big markets in countries like India. Regular Taxi service regulations are not applicable for uber.

- One documented means of reducing these overheads is to focus the routing optimizations on the paths to popular destinations, and thus be able to shift a large volume of traffic with a small number of path switches, rather than shifting the traffic for all the prefixes.

- we track traffic with three distinct predictors which vary in degrees of complexity: a very simple predictor, the Last Value (LV), the classical Moving Average (MA) , and an adaptive but more complex predictor, the LpEMA (Low pass Exponential Moving Average),s.

- This implies that it is enough to take account of only a small fraction of the total number of destinations to control the routing of the majority of the traffic.
- Lastly, we describe the problem of selecting the popular destinations, and present our proposal.

Disadvantage

- However, there is a lack of simple and pragmatic methods for selecting popular destinations

- Low-profit margins causes dissatisfaction among the drivers. This might lead to bad publicity, which can in turn discourage the new drivers from joining Uber.

- Uber and its customers have no bonding. Incentive remaining with Uber is low.

❖ PROPOSED SYSTEM

- We proposed that we will find the days on which each basement has more trips. Find the days on which each basement has more number of active vehicles. Can tap growing markets in suburban areas where taxi services are not available.

- Based on the data, we will find the top 20 destination people travel the most, top 20 locations from where people travel the most, top 20 cities that generate high airline revenues for travel, based on booked trip count. Top 20 destination people travel the most: Based on the given data, we can find the most popular destination that people travel frequently. We are creating an RDD by loading a new dataset which is in HDFS.

- We are creating the key-value pair, where key is the destination that is in 3rd column and the value is 1. Since we need to count the cities which are popular, we are using the reduceByKey method to count them.

- After counting the destinations, we are swapping the key-value pairs. The sortByKey method sorts the data with keys and false stands for descending order.

- Once the sorting is complete, we are considering the top 20 destinations. We can find the places from where most of the trips are undertaken, based on the booked trip count.

Advantage

- Estimated Time of Arrival can be reduced with rise in the number of Uber drivers which in turn will make Uber more liked by the customers and hence, the startup will get more revenue and drivers will also be profited.
- Convenient system for the drivers. They can work for flexible hours and can even choose to be a part-time employee. Drivers can also reject unwanted clients.
- There are many destinations out of which we will find only first 20, based on trips booked for particular destinations.
- We are using the sortByKey method which sorts the data with keys where false stands for descending order.

FEASIBILITY STUDY

- Feasibility study is the initial design stage of any project, which brings together the elements of knowledge that indicate if a project is possible or not. All projects are feasible if they have unlimited resources and infinite time. But the development of the software is plagued by the scarcity of resources and difficult delivery rate. It is necessary and prudent to evaluate the feasibility of the project at the earlier times.

TECHNICAL FEASIBILITY

- This is concerned with specifying the software will successfully satisfy the user requirement. Open source and business-friendly and it is truly cross platform, easily deployed and highly extensible.

ECONOMIC FEASIBILITY

- Economic analysis is the most frequently used technique for evaluating the effectiveness of a proposed system. The enhancement of the existing system doesn't incur any kind of drastic increase in the expenses. Python is open source and ready available for all users. Since the project is runned in python and jupyter notebook hence is cost efficient.

HARDWARE REQUIREMENTS

- Processor : Intel Pentium Dual Core 2.00GHz Hard disk : 40 GB
- RAM : 2 GB (minimum)

SOFTWARE REQUIREMENTS

- Jupyter Notebook
- Python 3.6.4 Version

Data Analysis

- **Data analysis** is defined as a process of cleaning, transforming, and modeling data to discover useful information for business decision-making.
- The purpose of Data Analysis is to extract useful information from data and taking the decision based upon the data analysis. So that is nothing but data analysis. Now same thing analyst does for business purposes, is called Data Analysis.

Types of Data Analysis: Techniques and Methods

- There are several **types of Data Analysis** techniques that exist based on business and technology. However, the major Data Analysis methods are:
  - Text Analysis
  - Statistical Analysis
  - Diagnostic Analysis
  - Predictive Analysis
  - Prescriptive Analysis

Text Analysis

- Text Analysis is also referred to as Data Mining. It is one of the methods of data analysis to discover a pattern in large data sets using databases or data mining tools. It used to transform raw data into business information. Business Intelligence tools are present in the market which is used to take strategic business decisions. Overall it offers a way to extract and examine data and deriving patterns and finally interpretation of the data.

Statistical Analysis

- Statistical Analysis shows "What happen?" by using past data in the form of dashboards. Statistical Analysis includes collection, Analysis, interpretation, presentation, and modelling of data. It analyses a set of data or a sample of data. There are two categories of this type of Analysis - Descriptive Analysis and Inferential Analysis.

Descriptive Analysis

- Analyses complete data or a sample of summarized numerical data. It shows mean and deviation for continuous data whereas percentage and frequency for categorical data.

Inferential Analysis

- Analyses sample from complete data. In this type of Analysis, you can find different conclusions from the same data by selecting different samples.

Diagnostic Analysis

- Diagnostic Analysis shows "Why did it happen?" by finding the cause from the insight found in Statistical Analysis. This Analysis is useful to identify behavior patterns of data. If a new problem arrives in your business process, then you can look into this Analysis to find similar patterns of that problem. And it may have chances to use similar prescriptions for the new problems.
-

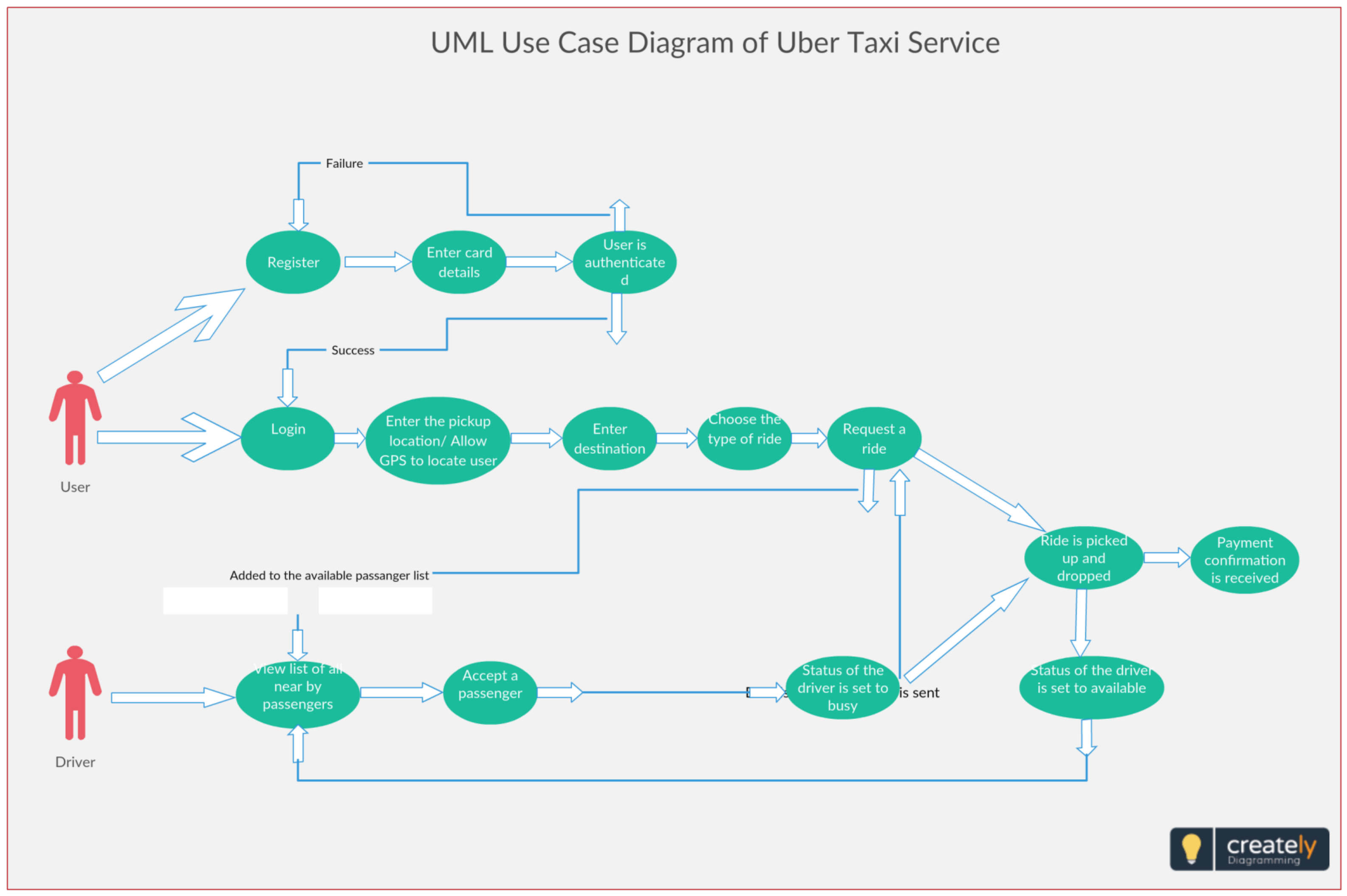
Predictive Analysis

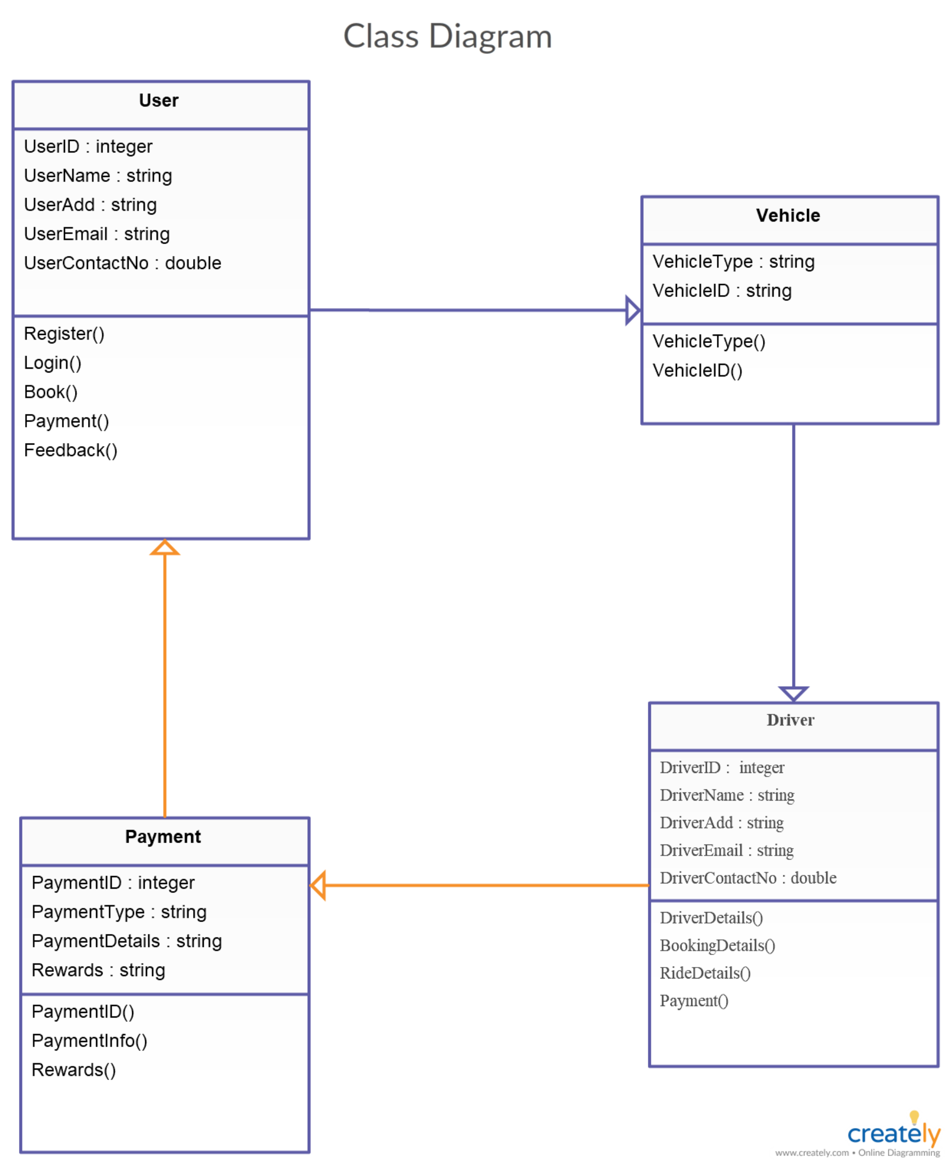
- Predictive Analysis shows "what is likely to happen" by using previous data. The simplest data analysis example is like if last year I bought two dresses based on my savings and if this year my salary is increasing double then I can buy four dresses.
- circumstances like chances of prices of clothes is increased this year or maybe instead of dresses you want to buy a new bike! So here, this Analysis makes predictions about future outcomes based on current or past data. Forecasting is just an estimate. Its accuracy is based on how much detailed information you have and how much you dig in it.

Prescriptive Analysis

- Prescriptive Analysis combines the insight from all previous Analysis to determine which action to take in a current problem or decision. Most data-driven companies are utilizing Prescriptive Analysis because predictive and descriptive Analysis is not enough to improve data performance. Based on current situations and problems, they analyse the data and make decisions.

❖ UML DIAGRAM


UML Use Case Diagram of Uber Taxi Service
Failure
Register
Enter card details
User is authenticated
Success
User
Login
Enter the pickup location/ Allow GPS to locate user
Enter destination
Choose the type of ride
Request a ride
Added to the available passanger list
Ride is picked up and dropped
Payment confirmation is received
Driver
View list of all near by passengers
Accept a passenger
Status of the driver is set to busy
is sent
Status of the driver is set to available
creately
Diagramming



Class Diagram

User

UserID : integer
UserName : string
UserAdd : string
UserEmail : string
UserContactNo : double

Register()
Login()
Book()
Payment()
Feedback()

Vehicle

VehicleType : string
VehicleID : string

VehicleType()
VehicleID()

Driver

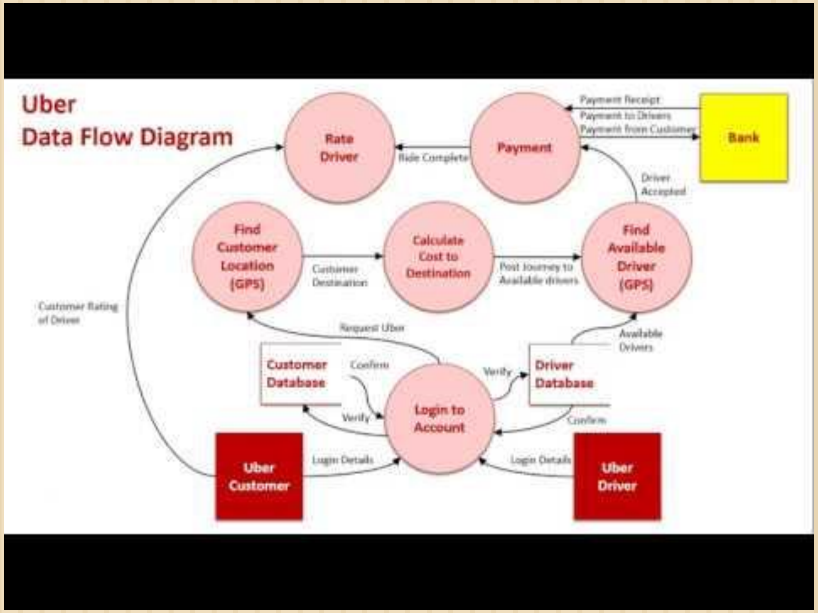
DriverID : integer
DriverName : string
DriverAdd : string
DriverEmail : string
DriverContactNo : double

DriverDetails()
BookingDetails()
RideDetails()
Payment()

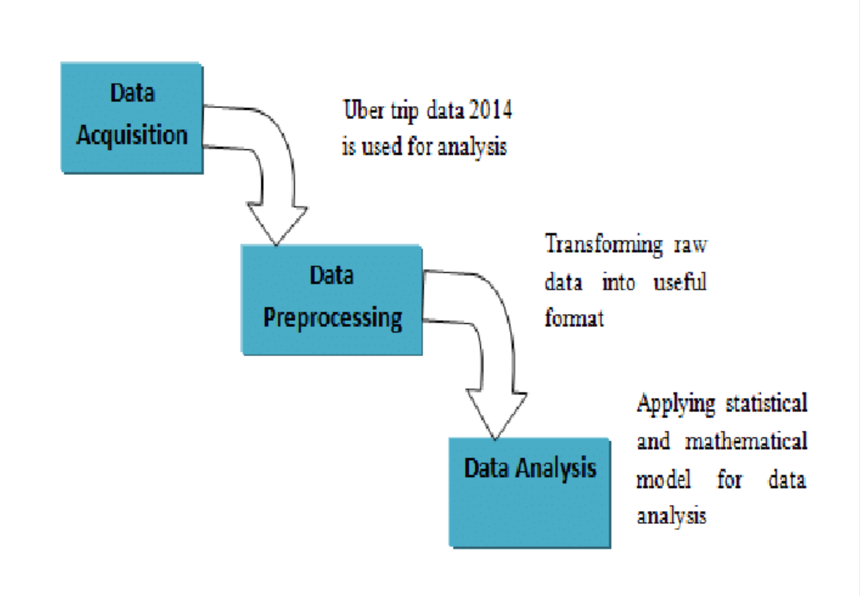
Payment

PaymentID : integer
PaymentType : string
PaymentDetails : string
Rewards : string

PaymentID()
PaymentInfo()
Rewards()



Uber
Data Flow Diagram



Data Acquisition

Uber trip data 2014 is used for analysis

Data Preprocessing

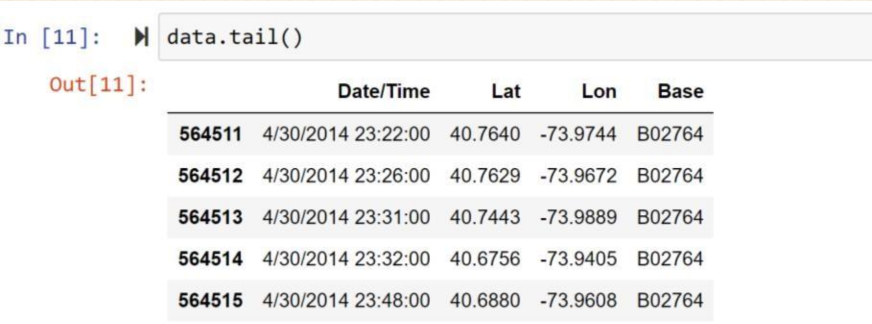
Transforming raw data into useful format

Data Analysis

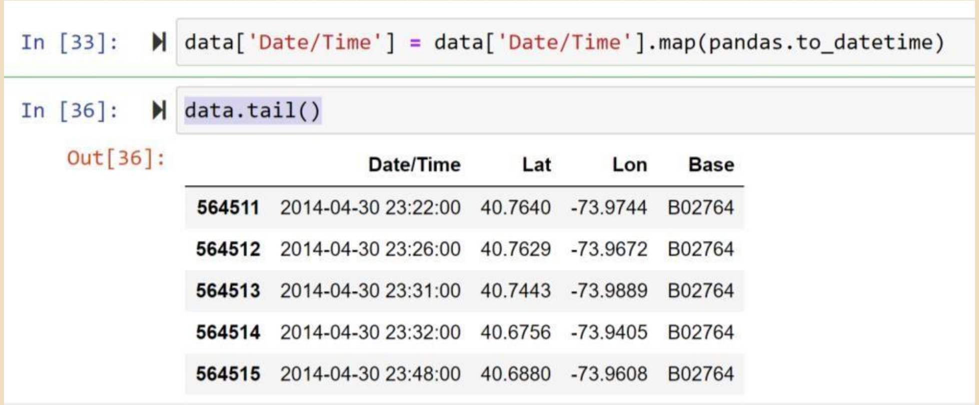
Applying statistical and mathematical model for data analysis

❖ Testing and performance analysis

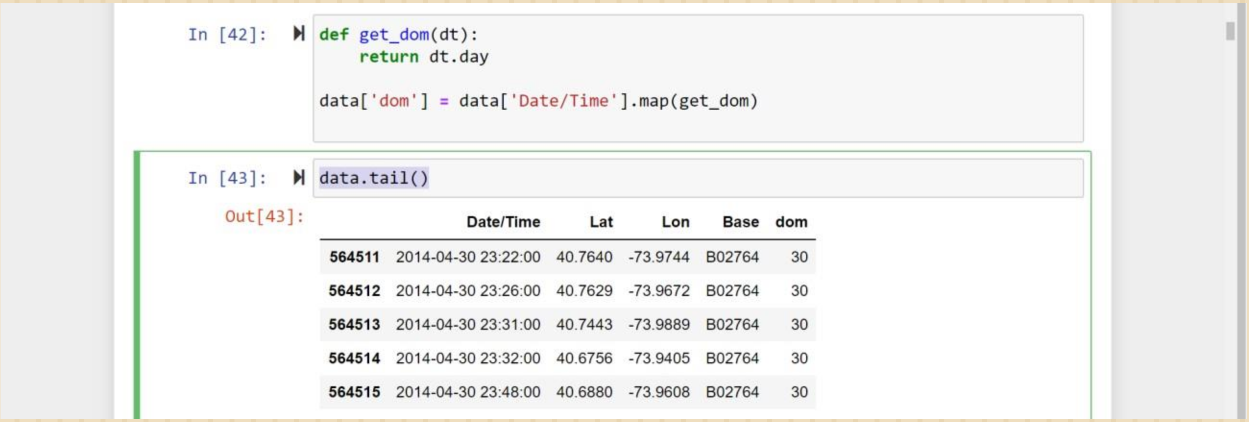
SAMPLE OUTPUT



```
In [11]:  data.tail()

Out[11]:
                   Date/Time      Lat       Lon     Base
564511  4/30/2014 23:22:00  40.7640  -73.9744  B02764
564512  4/30/2014 23:26:00  40.7629  -73.9672  B02764
564513  4/30/2014 23:31:00  40.7443  -73.9889  B02764
564514  4/30/2014 23:32:00  40.6756  -73.9405  B02764
564515  4/30/2014 23:48:00  40.6880  -73.9608  B02764
```



```
In [33]:  data['Date/Time'] = data['Date/Time'].map(pandas.to_datetime)

In [36]:  data.tail()

Out[36]:
                    Date/Time      Lat       Lon     Base
564511  2014-04-30 23:22:00  40.7640  -73.9744  B02764
564512  2014-04-30 23:26:00  40.7629  -73.9672  B02764
564513  2014-04-30 23:31:00  40.7443  -73.9889  B02764
564514  2014-04-30 23:32:00  40.6756  -73.9405  B02764
564515  2014-04-30 23:48:00  40.6880  -73.9608  B02764
```



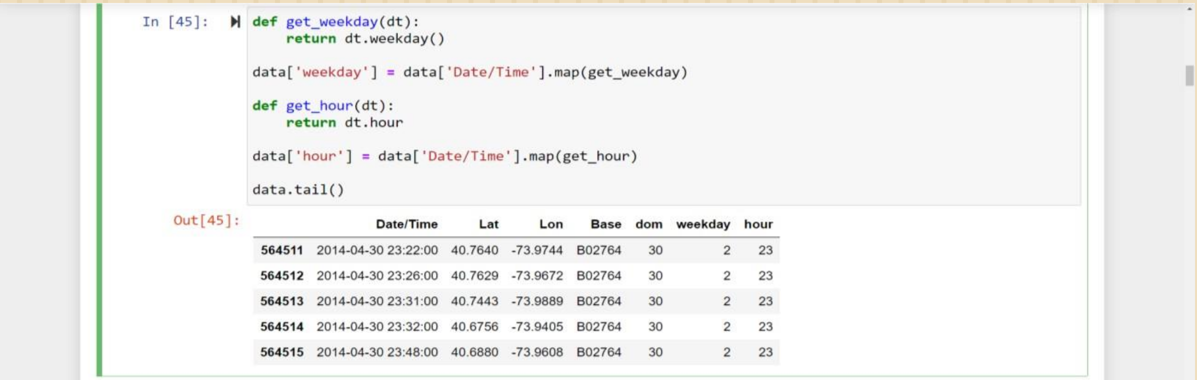
```
In [42]:  def get_dom(dt):
              return dt.day

          data['dom'] = data['Date/Time'].map(get_dom)

In [43]:  data.tail()

Out[43]:
                    Date/Time      Lat       Lon     Base  dom
564511  2014-04-30 23:22:00  40.7640  -73.9744  B02764   30
564512  2014-04-30 23:26:00  40.7629  -73.9672  B02764   30
564513  2014-04-30 23:31:00  40.7443  -73.9889  B02764   30
564514  2014-04-30 23:32:00  40.6756  -73.9405  B02764   30
564515  2014-04-30 23:48:00  40.6880  -73.9608  B02764   30
```

IN[] : def get_weekday(dt): return dt.weekday()

data['weekday'] = data['Date/Time'].map(get_weekday) def get_hour(dt):

return dt.hour

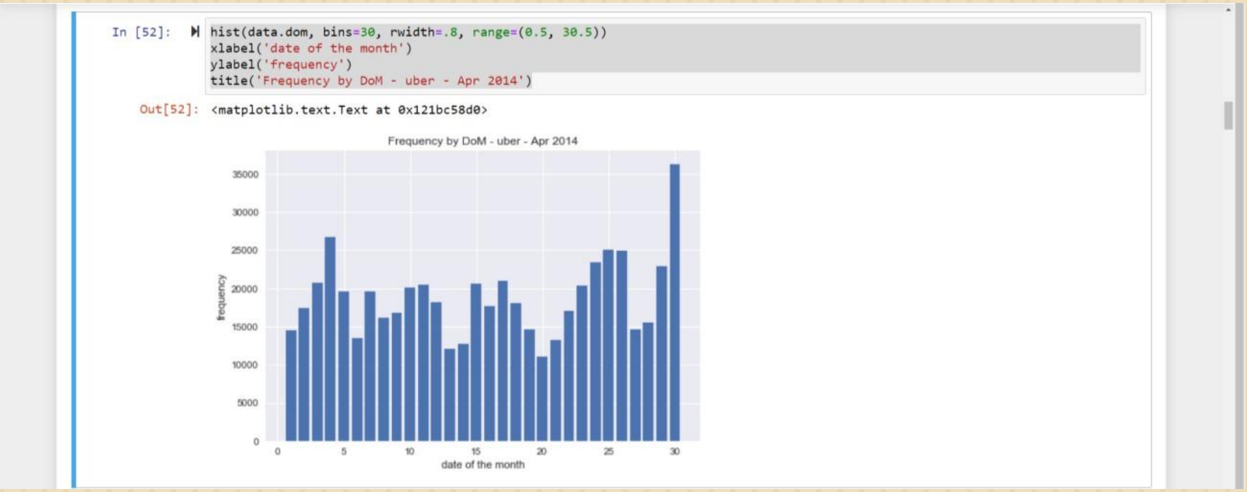
data['hour'] = data['Date/Time'].map(get_hour) data.tail()



```
In [45]:  def get_weekday(dt):
              return dt.weekday()

          data['weekday'] = data['Date/Time'].map(get_weekday)

          def get_hour(dt):
              return dt.hour

          data['hour'] = data['Date/Time'].map(get_hour)

          data.tail()
```

Out[45]:

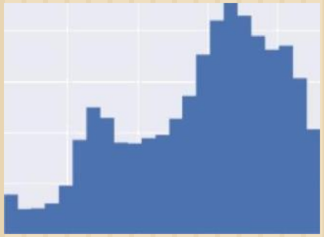
	Date/Time	Lat	Lon	Base	dom	weekday	hour
564511	2014-04-30 23:22:00	40.7640	-73.9744	B02764	30	2	23
564512	2014-04-30 23:26:00	40.7629	-73.9672	B02764	30	2	23
564513	2014-04-30 23:31:00	40.7443	-73.9889	B02764	30	2	23
564514	2014-04-30 23:32:00	40.6756	-73.9405	B02764	30	2	23
564515	2014-04-30 23:48:00	40.6880	-73.9608	B02764	30	2	23



```
In [52]:  hist(data.dom, bins=30, rwidth=.8, range=(0.5, 30.5))
          xlabel('date of the month')
          ylabel('frequency')
          title('Frequency by DoM - uber - Apr 2014')

Out[52]: <matplotlib.text.Text at 0x121bc58d0>
```

- □ analyze the hour

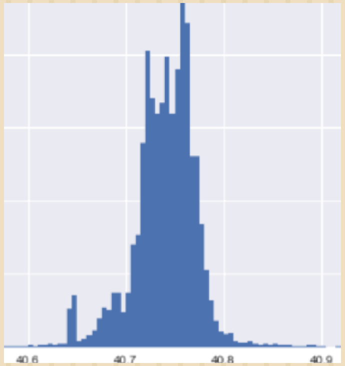
- □ analyze the weekday



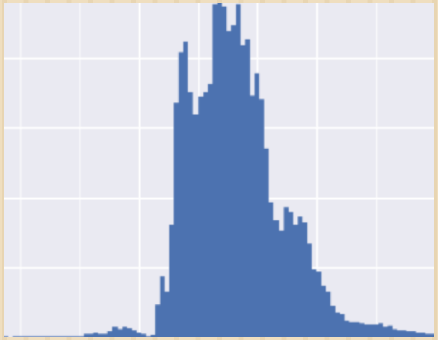
- □ cross analysis (hour, dow)

N b cross = data rou b ' eekd ' n ' . s l it a 1 count rows . unstack

- □ by lat and lon
- □ n [ ] : H hi st (data [’ Let ’ , bi ns = ra nge = (80.
  41)°› fi- '



- □ [",8 ! N his l (data [’ La n ’] , bi ns=1ee, ra nge = (-za . 1, - . 9))
  ;

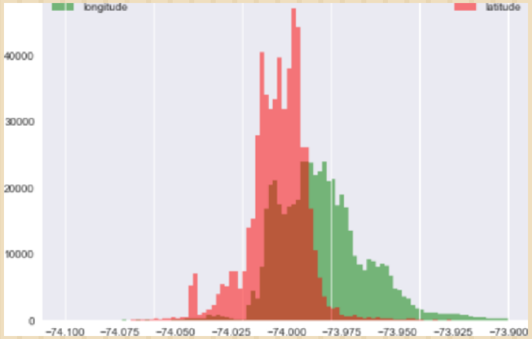


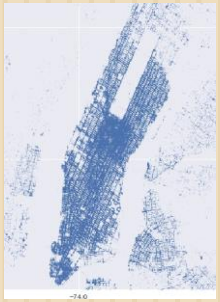
- In [108] : N hist (data[' Lan'] , bins=100, = (-74.l, -73. 9), color= ' g' , alpha= . 5, label = ’ longitude ')

legend (1oc= ' uppe r left ')

hist(data['Lat'], bius=100, = 41), color='r', alpha=.5, label =

legend (loc= ' best ')





❖ REFERENCES

- M. Lowe, C. Whitzman, H. Badland, M. Davern, L. Aye, D. Hes, I. Butterworth, and
- B. Giles-Corti, “Planning healthy, liveable and sustainable cities: how can indicators inform policy?” Urban Policy and Research, vol. 33, no. 2, pp. 131–144, 2015.
- The Economist Intelligence Unit, “Global liveability index 2018,” 2018.
- United Nations, “New urban agenda,” 2016.
- C. Barnett and S. Parnell, “Ideas, implementation and indicators: epistemologies of the post-2015 urban agenda,” Environment and Urbanization, vol. 28, no. 1, pp. 87–98, 2016.
- F. Caprotti, R. Cowley, A. Datta, V. C. Broto, E. Gao, L. Georgeson, C. Herrick, N. Odendal, and S. Joss, “The new urban agenda: key opportunities and challenges for policy and practice,” Urban research & practice, vol. 10, no. 3, pp. 367–378, 2017.
- United Nations, “Glossary of the habitat iii,” 2016.
- P. W. Newton, “Livable and sustainable? socio-technical challenges for twenty-first- century cities,” Journal of Urban Technology, vol. 19, no. 1, pp. 81–102, 2012.
- M. Ruth and R. S. Franklin, “Livability for all? conceptual limits and practical implications,” Applied Geography, vol. 49, pp. 18–23, 2014.
- H. J. Miller, F. Witlox, and C. P. Tribby, “Developing context-sensitive livability indicators for transportation planning: a measurement framework,” Journal of Transport Geography, vol. 26, pp. 51–64, 2013.
- A. Ley and P. Newton, “Creating and sustaining livable cities,” in Developing living cities: From analysis to action. World Scientific, 2010, pp. 191–229.
- R. Cervero, Transit-oriented development in the United States: Experiences, challenges, and prospects. Transportation Research Board, 2004, vol. 102.