ITCS 5102 SURVEY OF PROGRAMMING LANGUAGES

A PROJECT REPORT

ON

UBER/LYFT DATA ANALYSIS USING PYTHON

TEAM MEMBERS

Geeta Priyanka Janapareddy	800988736
Harsh Hundiwala	800986810
Heena Khan	800340986
Parul Thakral	800986821
Suvajit Chakrabarty	800976080

ABSTRACT:

Uber Technologies, Inc. is a transportation network company and a pioneer in the shared economy business. The Uber software application, better known as the “Uber App”, was developed using GPS technologies for Uber drivers and their clients to use during a transaction. Pertinent details regarding a transaction include the date, time, and geographical location of the Uber request. Data sets containing this information for New York City were found online at Kaggle.com, and were used to create visual depictions such as charts, graphs, and maps to highlight the busiest locations for Uber drivers.

The problem addressed in this project is determining where Uber pick-ups and Lyft pick-ups are popular. Using Uber datasets and Lyft datasets that document transactions from April 2014 to September 2014, a clustering algorithm specifically K-means was applied to predict the geographical location of Uber/Lyft pick-ups in New York City, in September 2014. After a detailed process of refining, training, and testing the data, we were able to make predictions of Uber pick-up locations, by longitude and latitude coordinates, within high percent accuracy.

INTRODUCTION:

Currently, Uber/Lyft drivers have no way of knowing when and where their next ride pick-up will be. Drivers typically go to a specific location that gets busy at a certain time of the day, and will wait for an Uber request. Records and statistics of past Uber transactions may show correlations between certain areas, such as particular boroughs in New York City, and certain times (morning, afternoon, or evening). However, there is no way of predicting where future pick-ups may be. Using relevant data sets and a range of machine learning methods, our project aims to make predictions of the longitude and latitude coordinates of Uber pickups for September 2014 in New York City.

THE PYTHON LANGUAGE:

Python language was created by the developer Guido van Rossum in 1991 and since then it is one of the fastest growing language. Python is an interpreted language i.e. it can execute instructions directly, without previously compiling a program into machine-language instructions. The design philosophy of python emphasizes on code readability and it uses white space indentation to delimit code blocks rather than curly braces. Syntax in Python allows a programmer to express concepts in few lines of code and constructs that enable clear programming. It is a widely used high-level programming language due to its features and multiple programming paradigms. The Python is widely used in bigger organizations because of its multiple programming paradigms. They usually involve imperative and object-oriented functional programming. It has a comprehensive and large standard library that has automatic

memory management and dynamic features. It supports libraries (numpy, scipy and matplotlib) and functions for almost any statistical operation and model creation. Since introduction of pandas one of its library, it has become very strong in operations on structured data.
.

WHY PYTHON?

- Python is Open Source and free to install
- Flexible language focuses more on Readability.
- Python testing framework is a built-in, low-barrier-to-entry testing framework that encourages good test coverage. This guarantees code reusability.
- Easy to understand language that is known by programmers and that can easily be learnt by statisticians, you can build a single tool that integrates with every part of your workflow.
- Python has some nice visualization libraries, such as Seaborn, Bokeh and Pygal, there are maybe too many options to choose from.
- Can become a common language for data science and production of web based analytics products.

DESCRIPTION OF LANGUAGE SYNTAX:

Syntax in python is the set of rules that defines how a Python program will be written and interpreted. It is an ideal language for scripting and rapid application development in many areas on most platforms due to its elegant syntax and interpreted nature. We can write Python programs using any text editor and the file should have extension .py. The easiest way is to use the IDE such as jupyter notebook, spyder, all are available in Anaconda navigator. The .py files can also be executed with command prompt by simply typing python filename.py. Python language mainly focuses on readability so instead of braces, blocks are identified by having same indentation.

ELEMENTS OF THE LANGUAGE:

Keywords:

KEYWORD	DESCRIPTION
If,else,while,for,next,repeat.	The basic control-flow constructs of Python language.
and, or , not	and, or, not are the logical operators in Python. and will result into True only if both the operands are True.
TRUE & FALSE	True and False are truth values in Python. They are the results of comparison operations or logical (Boolean) operations in Python.

Def	def is used to define a user-defined function
As	as is used to create an alias while importing a module. It means giving a different name (user-defined) to a module while importing it.
Assert	assert is used for debugging purposes.
None	None is a special constant in Python that represents the absence of a value or a null value.
Break, continue	break will end the smallest loop it is in and control flows to the statement immediately below the loop. continue causes to end the current iteration of the loop, but not the whole loop.
Class	class is used to define a new user-defined class in Python. Class is a collection of related attributes and methods that try to represent a real world situation.
Del	del is used to delete the reference to an object. Everything is object in Python. We can delete a variable reference using del
Except	except, raise, try are used with exceptions in Python.
Finally	finally is used with try…except block to close up resources or file streams.
from, import	import keyword is used to import modules in the current namespace. from…import is used to import specific attributes or functions into the current namespace.
Global	global is used to declare that a variable inside the function is global (outside the function).
In	in is used to test if a sequence (list, tuple, string etc.) contains a value. It returns True if the value is present, else it returns False.
Is	is used in Python for testing object identity. While the == operator is used to test if two variables are equal or not, is used to test if the two variables refer to the same object.
Lambda	lambda is used to create an anonymous function (function with no name). It is an inline function that does not contain

	a return statement. It consists of an expression that is evaluated and returned
Nonlocal	The use of nonlocal keyword is very much similar to the global keyword. nonlocal is used to declare that a variable inside a nested function (function inside a function) is not local to it, meaning it lies in the outer enclosing function.
Pass	pass is a null statement in Python. Nothing happens when it is executed. It is used as a placeholder.
Return	return statement is used inside a function to exit it and return a value.
With	with statement is used to wrap the execution of a block of code within methods defined by the context manager.
Yield	yield is used inside a function like a return statement. But yield returns a generator.

Datatypes:

DATA TYPE	DESCRIPTION
Integer numbers Floating point numbers	Any non decimal number is called an integer number Any decimal number is called a floating point number.
String	String is a sequence of characters that are enclosed within single quotes or double quotes.
List	A list is a container that holds many objects under a single name.
Tuple	A tuple is a container that holds many objects under a single name. A tuple is immutable which means, a tuple once defined cannot be modified.
Dictionary	A dictionary is a set of key-value pairs referenced by a single name

Importing data in Python

Using Pandas Libraries

- read_csv()
  Reads in data from csv file

CONTROL STRUCTURE:

Conditional Execution:

Condition statements are a block of statements whose execution depends on a certain condition.

1. If:
- A "simple if" condition is one where a block of statements get executed if the condition mentioned in the "if" statement evaluates to true.

  Syntax: if (condition) expr_1

2. If-Else
- An "If-Else" statement is one where a block of statements under "if" condition gets executed if the condition evaluates to true. If the condition evaluates to false, the block of statements under "else" is executed.

  Syntax: if (condition) expr_1 else expr_2

3. If-Elif-Else
- An "If-Elif-Else" statement is one where multiple "if" conditions are evaluated one after another if an "if" statement evaluates to false. "elif" stands for else-if. If all the if conditions evaluates to false, the block of statements under "else" gets executed.

  Syntax: if (condition) expr_1 elif expr2 elif expr3 else expr4

Repetitive Execution:

Looping is used to repeatedly perform a block of statements over and over again.

1.While Loop:
- "While loop" is used to repeatedly execute a block of statements as long as the condition mentioned in the "while loop" holds true.

  Syntax:
  while condition:
  statement 1
  statement 2
  - - -
  - - -
  statement 3

2.For Loop:

- For loop is used to iterate over a sequence, starting from the first value to the last. The number of iterations to be performed depends upon the length of the list.

  Syntax:
  for iteratingVariable in sequence:
  statement 1
  statement 2
  - - -
  - - -
  statement n

3. Nested Loop:

A loop within another loop is called a nested loop. The concept of nested loop could be a little bit of trouble understanding at first, but can be simplified with the help of an example.

4.The **break** statement can be used to terminate any loop, possibly abnormally. This is the only way to terminate repeat loops.

APPROACH:

- The data pre-processing part involved cleaning empty data, converting the locations from address to latitude and longitude.
- We are trying to predict based on latitude, longitude and hour of the day for the based on August month data who is most likely to get the ride that is Uber or Lyft(as competitor).
- Using K-means clustering algorithm, we clustered the datasets based on longitude and latitude. Once the data is clustered, we were able to plot the data using Scatter plot to visualise it.
- The visualisation of clustered data depicted that it is easy for Lyft drivers to find rides quickly in Downtown of New York. Whereas, Uber had ubiquitous popularity around the New York. Still the popularity of Uber peaked in central part of New York.

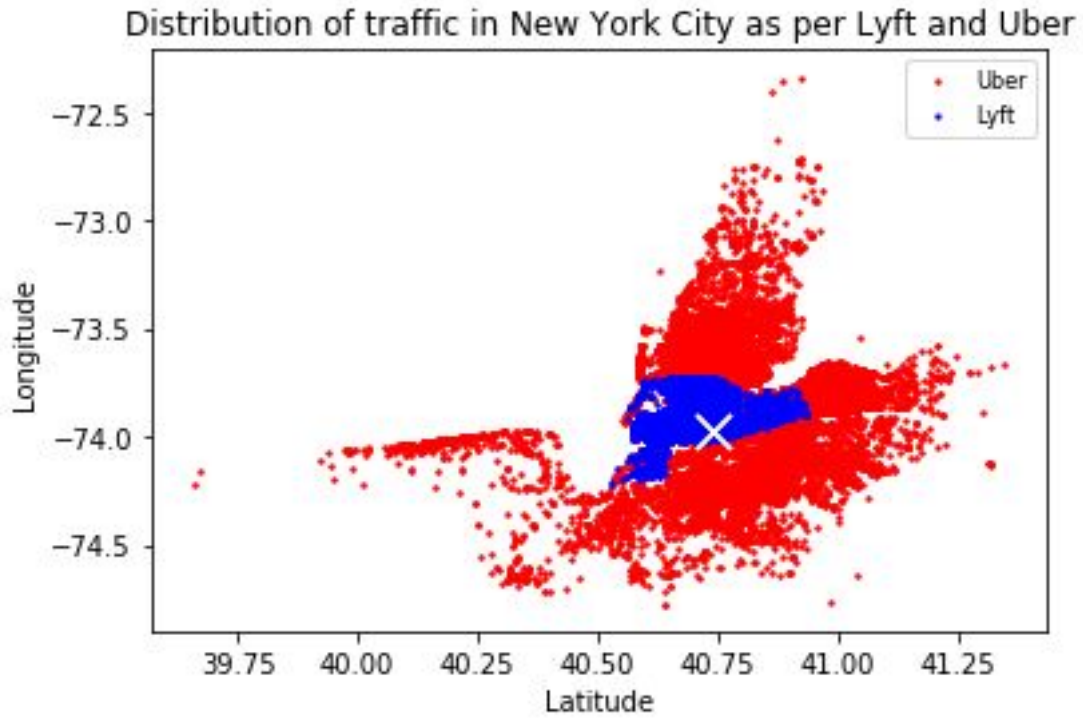
Materials and Methods:

The primary language used in this project is in Python. The libraries used are from Sci-kit Learn, an open-source library archive that implements a range of machine learning algorithms. The machine learning algorithms used in this project to predict Uber/Lyft pick-up locations are K Means Clustering.

RESULTS:

The table below shows comparison of accuracy obtained from various approaches:

Visualization



STRENGTHS AND WEAKNESS OF PYTHON:

Strengths:

- **Open Language:** Python is free and open source software, allowing anyone to use and modify it.
- **Extensive Support Libraries:** It provides large standard libraries that include the areas like string operations, Internet, web service tools, operating system interfaces and protocols.
- **Integration Feature:** Python integrates the Enterprise Application Integration that makes it easy to develop Web services by invoking COM or COBRA components.

Weakness:

- **Run-time Errors:** The Python language is dynamically typed so it has many design and it requires more testing time, and the errors show up when the applications are finally run.
- **Difficulty in Using Other Languages:** Due to its libraries and features it causes difficulty for one to learn other languages as there are no semicolons or curly braces required in Python