

**Comparative Analysis on HDBSCAN and DBSCAN  
Algorithms using Taxi Pickup and Dropoff  
in New York City Dataset**

---

A Thesis Paper Presented to the  
Faculty of Computer Science and Information Technology Department

In Partial Fulfillment of the Requirements  
For the Degree of Bachelor of Science in Computer Science

---

Baylon, Brent V.  
Quirante, Devon Glad L.

2023



## APPROVAL SHEET

In partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science, “**Comparative Analysis on HDBSCAN and DBSCAN Algorithms using Taxi Pickup and Dropoff in New York City Dataset**”, has been examined and is recommended for oral defense.

**ENGR. GUILBERT NICANOR A. ATILLO, DPA**

*Adviser*

This thesis entitled “**Comparative Analysis on HDBSCAN and DBSCAN Algorithms using Taxi Pickup and Dropoff in New York City Dataset**”, prepared and submitted by Brent V. Baylon and Devon Glad L. Quirante has been reviewed and approved by the thesis committee.

**ROCHE L. CABANLIT, MST-CS**

*Chairman, CSIT*

**ENGR. GUILBERT NICANOR A. ATILLO, DPA**

*Adviser*

**ENGR. JOSEPHINE C. MUÑASQUE**

*Examiner*

**Dr. SAMI KHAYAT**

*Examiner*



NEGROS ORIENTAL STATE UNIVERSITY  
College of Arts and Sciences  
Computer Science and Information Technology Department  
Main Campus I and II, Dumaguete City

---



**APPROVED** by the following PANEL OF EXAMINERS ON ORAL DEFENSE on  
June 13, 2023 with a rating of \_\_\_\_\_.

**ENGR. GUILBERT NICANOR A. ATILLO, DPA**  
*Chairman*

Date Signed: \_\_\_\_\_

**ENGR. JOSEPHINE C. MUÑASQUE**  
*Examiner*

**DR. SAMI KHAYAT**  
*Examiner*

Date Signed: \_\_\_\_\_

Date Signed: \_\_\_\_\_

Accepted by the College of Arts and Sciences Dean for the degree of Bachelor of Science  
in Computer Science.

**MICHAEL P. BALDADO JR., Ph.D.**  
*Dean, College of Arts and Sciences*

---



### **Acknowledgement**

The researchers would like to extend their heartfelt gratitude to the following persons who helped them in making their work a successful one.

Engr. Guilbert Nicanor A. Atillo of Negros Oriental State University, their adviser for his encouragement, insight and keen sense of suggestions have enabled the researchers to see this work through to its completion.

Ms. Karima Khayat, the statistician for analyzing and verifying the credibility of the collected data.

The researchers would also like to thank their families who give them guidance, financial and moral support all throughout their studies; Their professors, instructors, friends and classmates for their encouragement.

Above all to the Almighty Father for giving them strength, knowledge, wisdom, guidance, and filling their spiritual needs.

Their sincerest gratitude to all of them.

**The Researchers**



NEGROS ORIENTAL STATE UNIVERSITY  
College of Arts and Sciences  
Computer Science and Information Technology Department  
Main Campus I and II, Dumaguete City

---



*We honor our friends and family who helped and inspired us during our academic path by dedicating this thesis to them. You helped us through the many late hours and times when we questioned our ability, and you joyfully celebrated our accomplishments. This accomplishment is equally yours and ours, and we will always be grateful for having you in our lives.*



### **Abstract**

The pervasiveness of transportation paved way for the taxis to become one of the widely used mode of land transportation. The present study aims to analyze the traffic activity of taxis in one of the most densely populated places in the world which is the New York City and perform data mining techniques such as the Hierarchical Density-Based Spatial Clustering of Application with Noise (HDBSCAN) using the pickup and dropoff variables and then compare it to its predecessor, Density-Based Spatial Clustering of Application with Noise (DBSCAN). Jupyter extension in Visual Studio Code was used in the study to perform the data mining algorithms. The results disclosed Manhattan to be where most of the trips come from. There are also a number of trips that are prevalent in Queens, Bronx, and Brooklyn. Most of the trips carry one passenger with Credit Card as the widely used payment method in trips and usually occurs at 7 PM which reaches its peak time for the greatest number of pickups and dropoffs. When comparing the output of HDBSCAN to DBSCAN, the results favor HDBSCAN as the better algorithm to handle such massive amount of data with 700,000+ rows of data in terms of both efficiency and accuracy. These findings will significantly help the notorious issue in New York City which is traffic congestion by supplying these data to draw useful conclusions as to the traffic activities of yellow taxis which will help mitigate the traffic congestion in one of the most densely populated places in the world.

**Keywords:** HDBSCAN, DBSCAN, Clustering Algorithm, Taxi Pick-up and Drop-off

---



---

## Table of Contents

Title	Page
Approval Sheet .....	i
Acknowledgement.....	iii
Dedication.....	iv
Abstract.....	v
Table of Contents .....	vi
List of Figures.....	vii
List of Table .....	viii
List of Appendices.....	ix
Glossary .....	x
<b>Chapters</b>	
I. Introduction .....	1
II. Statement of the Problem .....	3
III. Review of Technical Literature .....	5
IV. Datasets.....	17
V. Methodology .....	19
VI. Data Analysis and Interpretation of Results .....	25
VII. Summary of Findings, Conclusion and Recommendation.....	48
References .....	54
Appendices .....	60

---



### List of Figures

Figure	Description	Page
1	The phases in processing the research dataset.....	19
2	OpenStreetMap Python import .....	24
3	Passenger Count Distribution .....	25
4	Trip Duration Distribution .....	27
5	Trip Distance Distribution .....	28
6	Pickup Hour Distribution.....	30
7	Dropoff Hour Distribution .....	31
8	Payment Type Distribution .....	32
9	Fare Amount Distribution .....	34
10	Tip Amount Distribution.....	35
11	Total Amount Distribution.....	37
12	Trip Distance vs. Trip Duration Correlation.....	38
13	Trip Distance vs. Fare Amount .....	40
14	Cluster plot for (a) Pickup and (b) Dropoff Locations .....	41
15	Cluster plot for (a) Pickup and (b) Dropoff Locations with map .....	43
16	Efficiency (HDBSCAN vs. DBSCAN) .....	45
17	Accuracy (HDBSCAN vs. DBSCAN).....	46





NEGROS ORIENTAL STATE UNIVERSITY  
College of Arts and Sciences  
Computer Science and Information Technology Department  
Main Campus I and II, Dumaguete City

---



**List of Table**

<b>Table</b>	<b>Description</b>	<b>Page</b>
1	Scale of Pearson's Correlation Coefficient.....	21



NEGROS ORIENTAL STATE UNIVERSITY  
College of Arts and Sciences  
Computer Science and Information Technology Department  
Main Campus I and II, Dumaguete City

---



**List of Appendices**

<b>Appendix</b>	<b>Description</b>	<b>Page</b>
A	Curriculum Vitae .....	60
B	Dataset .....	62
C	Statistician's Certification .....	63
D	Grammarly and Turnitin Results .....	64



## Glossary

<b>Accuracy</b>	How well an algorithm performs based on the desired outcome or result
<b>Clustering</b>	A method of data mining that groups data objects such that the items in a group are comparable or both linked to and different from or not connected to the items in other groupings
<b>Data Mining</b>	A branch of research that integrates methods from statistics, pattern recognition, and machine learning, databases and visualization are used to address the issue information from big databases by way of retrieval
<b>Density-Based Spatial Clustering of Application with Noise (DBSCAN)</b>	An approach to grouping data points by identifying distinct areas of high-density areas of low density are separated to one another



**Efficiency**

How well an algorithm performs with regards to performance and optimization

**Hierarchical Density-Based Spatial Clustering of Application with Noise (HDBSCAN)**

A machine learning algorithm which is the successor of the DBSCAN algorithm with the addition of hierarchy feature

**Latitude**

This term is used to know the location of a vehicle vertically in a geographic scale.

**Longitude**

This term is used to know the location of a vehicle horizontally in a geographic scale.

**Pearson Product – Moment Correlation Coefficient**

A statistical tool that is used in measuring a linear correlation using numbers between  $-1$  and  $1$  that measures the strength and direction of the relationship between two variables.



<b>New York City</b>	The location where the dataset is based and obtained from.
<b>Noise/Outliers</b>	Points that are outside the dense regions
<b>Taxi</b>	A vehicle primarily used as the mode of transportation in the study
<b>Taxi Dropoff</b>	It is where the taxi drops off its passenger
<b>Taxi Pickup</b>	It is where the taxi picks up its passenger



## **Chapter I**

### **Introduction**

Transportation is a crucial part of society as it takes into consideration the development of merchandise and individuals starting from one destination onto another. New York City, considered among the most active cities on the planet, has several taxis roaming around the city to aid the people into their destination. According to Gong et al. (2016), in New York City, individuals use taxis a lot more than some other urban communities of the US.

Baghestani et al. (2020) asserted that New York City's traffic congestion, which affects all five boroughs (Manhattan, The Bronx, Queens, Brooklyn, and Staten Island), has long been a distinguishing characteristic of the city and rates it as the third-worst in the world. Additionally, the 3.94 million people who live in Manhattan during the day (which occupies an area of 22.96 square miles) include about 41% daily commuters, 37% locals, 10% out-of-town visitors, 9% local day-trip visitors, and 3% hospital patients and students (who live in off-campus housing outside Manhattan), which significantly increased traffic in New York City which is why studying the traffic data in New York City is significant especially to the commuters in the area.

By analyzing the taxi data in New York City, one could end up providing useful insights about the transportation status in the city. For instance, processing and providing the necessary and important profiles that are present in the dataset and interpreting them into graphs can be a helpful tool in drawing conclusions. Most especially, this study will





benefit the transportation sector in New York as to the flow of traffic and the different taxi hotspots 24 hours a day.

This study aims to analyze the taxi dataset in New York City from the year 2015 and visualize it through graphs to provide useful insights about the taxi drop-off and pick-up locations in New York City by using the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm and comparing it to Density-Based Spatial Clustering of Applications with Noise (DBSCAN) in terms of performance and efficiency. The study will also be using Pearson's correlation coefficient statistical tool which will measure the relationship between two variables that are available in the dataset such as trip distance versus trip duration and trip distance versus fare amount.





## Chapter II

### Statement of the Problem

This present study aims to compare HDBSCAN and DBSCAN on their efficiency and accuracy using the taxi pick-up and drop-off location dataset.

Specifically, it sought to answer the following questions:

1. What is the frequency of the taxi trips based on:
  - 1.1. Passenger count;
  - 1.2. Trip duration;
  - 1.3. Trip distance;
  - 1.4. Pickup hour
  - 1.5. Dropoff hour;
  - 1.6. Payment type;
  - 1.7. Fare amount;
  - 1.8. Tip amount; and
  - 1.9. Total amount?
2. Is there a correlation between trip distance and:
  - 2.1. Trip duration; and
  - 2.2. Fare amount?
3. What are the different HDBSCAN clusters of taxi trips according to:
  - 3.1. Pickup coordinates; and
  - 3.2. Dropoff coordinates?







4. What is the comparative performance of HDBSCAN and DBSCAN in terms of:
  - 4.1. Efficiency; and
  - 4.2. Accuracy?





### **Chapter III**

#### **Review of Technical Literature**

This section of the paper discusses the related literature and related studies used as a reference to develop, implement and comprehend the comparative analysis on HDBSCAN and DBSCAN algorithms using taxi pick-up and drop-off location in New York City taxi dataset.

#### **Related Literature**

##### **HDBSCAN**

Vijayan & Aziz (2022) defined HDBSCAN as a newly developed algorithm based on DBSCAN. The only difference is that it is capable of identifying clusters of various densities. Moving points are grouped using the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm, that will enforce a flexible technique for creating the least spanning tree. To create the minimal spanning tree, one of the more expensive parts of the HDBSCAN. Moreover, alternating between the Prim and the Boruvka algorithms are recommended. The adaptive HDBSCAN reduces the algorithm's processing duration by 5.31% without sacrificing its accuracy. Furthermore, Campello et al. (2013) asserted that the HDBSCAN library is a collection of tools that use unsupervised learning to identify clusters or dense regions in a dataset.

Moreover, Stewart & Al-Khassaweneh (2022) said that the HDBSCAN clustering algorithm is based on density. In contrast to K-means, which can identify dense clusters, it is not essential to assign each data point to a cluster. Noise or outliers are certain points





that are not assigned to a cluster. An efficient algorithm will be able to detect hastily and identify variety of collections' groups and noises. The HDBSCAN algorithm has a well-known Python implementation that is available as a scikit-learn compatible package. Python is a well-known language that is used frequently for data mining and machine learning applications like clustering. Despite the fact that there are many Java enterprise apps now in use, Java is less common for these kinds of jobs. Furthermore, Stewart & Al-Khassaweneh (2022) claimed that HDBSCAN outperforms DBSCAN in building a hierarchical representation of the clusters. The hierarchy produced by the algorithm's execution can be used pretty successfully for cluster extraction and outlier detection. HDBSCAN overcomes the limitations of DBSCAN by identifying clusters of any density.

Tran et al. (2021) added that recent years have seen a fast growth in data generation due to quick technological advancement. The majority of the data must be processed and categorized in order to become clean, usable data. However, categorizing a sizable amount of data is a difficult operation that necessitates significant human labor. Data clustering is one of the most popular techniques used in the initial stage of data cleaning to aid researchers and speed up this labeling process. Hierarchical Clustering is one of the most effective data clustering methods that can aid in separating noise from data patterns.

## **DBSCAN**

The first density-based algorithm is called Density Based Spatial Clustering of Applications with Noise (DBSCAN). It is a clustering algorithm which enforces unsupervised learning. In addition, Khan et al. (2014) explained how this kind of density-





based clustering algorithm can locate clusters of various sizes and shapes in datasets that contain noise and outliers. Furthermore, Deng (2020) stated that in the event of an unknown data distribution, the density-based clustering technique may cluster data sets of any shape. Due to its simple and efficient features, DBSCAN, a conventional density-based clustering method, is frequently employed for clustering data analysis.

Kumar & Reddy (2016) asserted that DBSCAN, which employs density-based spatial clustering of applications with noise, is efficient at handling outliers and can locate clusters of any shape. DBSCAN can locate clusters by counting the number of different points that are located within a specific radius of a given location. Distances between a given point and every other point in the dataset must be calculated. To expedite neighbor search processes, traditional index-based approaches build a hierarchical structure over the dataset.

### **Data Mining**

The process of looking at large amounts of data to identify patterns, trends, and potential applications is known as data mining. Data miners can then use the findings to make decisions and predict outcomes. According to Vijayan & Aziz (2022), in the mining of static data, clustering techniques are frequently employed. Data segmentation is included into components and data mining for correlations between variables. Clustering algorithms are much less frequently used with real-time data. This is brought on by a number of factors, including the algorithm's high computational demands. As such, the execution of





the calculation continuously or near ongoing may not be plausible. Hyperparameters must also be tuned in clustering algorithms in order to fit the dataset.

Aziz and Robila (2019) also mentioned that big data analysis is sometimes viewed as a difficult method for sifting through enormous amounts of data to find patterns and hidden information. Furthermore, Li (2018) came up with a framework that enables users to respond to inquiries from NYC Street-Hail Services in R since there are countless taxi data in New York City that will be used for data analysis in R.

Hand (2007) defined data mining as the process of looking for interesting, unexpected, or lucrative information in huge databases. Hence, it holds two various characteristics. These include aims to simulate the characteristics of the shapes and the characteristics of distributions which are concerned with large-scale, “global” structures. On the other hand, it focuses on small-scale, “local” structures, while still identifying malfunctions and figuring whether they are intentional or unintentional occurrences.

Chen et al. (1996) affirmed that several researchers have discovered the crucial database systems and machine learning study topic of mining information and knowledge from enormous databases, and major industrial enterprises have identified this field as one with the potential to make substantial profits. Data mining has been on the rise recently. In order to more easily understand customer behavior, improve the assistance provided, and increase financial potential, information mining procedures are also necessary for a variety of new applications in data-offering types of assistance, including information warehousing and online services over the web.





---

### **Pearson Product - Moment Correlation Coefficient (PMMC)**

The PMMC is one of the statistical analysis method available for determining how strong the relationship is between two variables. In contrast, correlation analysis in research enables us to measure the changes in one variable that are brought about by a change in another variable. According to Taylor (1990), correlation analysis, in which a correlation coefficient is provided to indicate how strong the relationship between two variables, is one of the more commonly discussed statistical approaches. Furthermore, Ratner (2009) stated that the term "correlation coefficient" was coined by Karl Pearson in 1896. Hence, while being almost a century old, this fact is still relevant today. Second only to the mean in terms of usage, it is a widely used statistic nowadays.

Killip et al. (2004) claimed that clustered samples, in which patients are randomly assigned on a group level yet examined individually, are frequently used in primary care research. Without accounting for this grouping, analysis can find significance where none exists. Furthermore, Bartko (1996) discussed a method for determining the intraclass correlation coefficient's reliability of a rating set. The estimation of variance components and variance component estimation was founded.

### **Related Studies**

Wang et al (2022) mentioned that effective and accurate black spot identification is a crucial and difficult undertaking to increase the safety of road traffic. A unique dark spot acknowledgment technique is introduced that consolidates progressive thickness based spatial gathering of uses with clamor with GIS-based handling. To reduce subjectivity in





parameter selection, the density-based clustering validation index, an internal validation indicator, is also utilized. The model is confirmed by collecting 3536 accident data from 1 August to 31 October 2020 in Hangzhou, China, and ultimately identifying 39 "black spots." (1) Regardless of the length of the road network in these areas only accounts for 23.26% from the total length of the road network, accidents that take place in dark spots account for 75% of all accidents. (2) Since many accidents collected per unit of road duration, the model performs better than the usual density-based spatial clustering of applications using an outlier model and also K-means. (3) Survey sample, which included six of the "black spots" that were found, demonstrated that the established framework has seen an increase in identification accuracy recommended that additional research be carried out in these areas.

According to Blanco-Portals et al. (2022), two new cutting-edge clustering analysis and dimensionality reduction algorithms are for the segmentation of core-loss electron energy loss spectroscopy (EELS) spectrum images: hierarchical density-based spatial clustering of applications with noise (HDBSCAN) and uniform manifold approximation and projection (UMAP). Using a well-known synthetic dataset, UMAP and HDBSCAN's performances are carefully compared to those of the other clustering analysis methodologies utilized in EELS in the literature. UMAP and HDBSCAN produce better outcomes. The triple combination nonnegative matrix factorization-UMAP-HDBSCAN, as well as an actual experimental dataset from an iron and manganese oxide core-shell nanoparticle, are utilized to show UMAP and HDBSCAN. The results show how the





complimentary use of various combinations might be advantageous in an actual setting to acquire a complete view, as various algorithms emphasize various features of the dataset under study.

Berg et al (2019) claimed that they offer a fresh approach for DBSCAN, a popular density-based clustering technique. For a group of  $n$ , indices in  $R^2$ , the DBSCAN-clustering is calculated by our method in  $O(n \log n)$  regardless of the scale parameter, time (and assuming that, as is typical in reality, the second parameter *minpts* is set to a fixed constant.) The new method is not only quick in principle, but experiments also demonstrate that a somewhat reduced version is competitive in practice and significantly less susceptible to the selection of compared to the initial DBSCAN algorithm. Also provided is an  $O(n \log n)$ . HDBSCAN is a recently introduced hierarchical variant of DBSCAN, and we demonstrate a randomized approach for computing an approximate version of HDBSCAN in the plane.

On the other hand, Liu et al. (2022) studied a random robot, which requires stability and precise target detection in a messy real-world environment. The robot was assigned a method for recognizing and tracking human legs with a single lidar. The HDBSCAN method, which is enhanced by leg characteristics, is used to cluster the target's leg information and locate it. Based on the density properties of the leg laser radar data, the leg line features are fused. The relative shift angle of the target's motion and the tracking distance lock-in technique enable single target recognition and tracking in an unstructured environment. The feature-based clustering technique improves target leg detection







accuracy by more than 10% over the original algorithm, as shown by the lidar dataset in a variety of scenarios. Conducting additional tests on a solid robot that is self-built is the major plan. The outcome demonstrates that the robot can accurately and steadily follow humans in an area with clutter.

Harrab (2018) declared that the most environmentally friendly method of providing fixed-line clients with very high-speed internet connection is commonly acknowledged as fiber to the home (FttH). Harrab (2018) also added that the method is based on HDBSCAN clustering to estimate the investment cost for these uneven population zones because the deployment of FttH networks is quite expensive and necessitates an accurate estimation of the investment cost. Their study's approach supports variable density clusters, especially for designing the FttH network architecture and calculating the deployment costs.

Ramani et al. (2022) studied that the amount of traffic on the route, late-night driving, which may be a little slower due to limited night vision, and many other factors influence taxi fares and trip length. Ramani et al. (2022) also tried to show how the day of the week, the location of the pickup and drop-off points, and the time of the pickup could all affect how long the trip would take.

AlBatati and Alarabi (2021) asserted that unsupervised learning technique is a type of Clustering Algorithm. The goal of cluster analysis is to categorize incoming data of similar instances so that examples that belong to the same cluster are more similar to one another than to instances that are a part of other clusters.





Liu et al. (2021) also mentioned that the HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) clustering method is utilized in this study to process meteorological data using observation data from the government of Shenzhen's regional meteorological observation stations as the investigated data set. Climate data are categorized using the HDBSCAN clustering method. As analytical indices, variables like temperature and precipitation are selected and dimensioned using PCA (Principal Component Analysis). Specific studies are carried out for some categories based on the obtained categorization results, and pertinent recommendations are given for the pertinent neighborhoods.

Shankar (2016) provided a comprehensive examination of the various approaches to data classification was carried out, including both conventional and novel computer science methods. K-Means, Mean Shift, Hierarchical, and DBSCAN are among the clustering algorithms. For the first time, chemical process data are used with HDBSCAN, a combination of hierarchical and density-based spatial clustering with applications and noise. On a dataset of chemical processes, Shankar (2016) has found that HDBSCAN works better than the other clustering techniques.

Wahyuni et al. (2021) declared that a grouping idea called HDBSCAN uses a *mpts* parameter. The goal of this study is to determine the number of clusters produced using the HDBSCAN method when underdeveloped villages and urban regions are grouped together in Kutai Kartanegara Regency. The study uses *mpts* settings ranging from 2 to 6. According to the analysis's findings, the Kutai Kartanegara Regency's undeveloped villages and urban





areas were grouped into three clusters using the HDBSCAN method. 19 villages and urban areas make up Cluster 0, 4 villages and urban areas make up Cluster 1, and 61 villages and urban areas make up Cluster 2. According to the analysis of Wahyuni et al. (2021), villages and urban areas that are part of cluster 1 may be the government's primary focus when it comes to aid and the construction of regional services and infrastructure.

Zhang et al. (2020) denoted that NYC traffic is a major problem, especially around rush hour. As the primary contributors to NYC traffic, NYC Taxi may be able to provide us with some hints regarding the traffic situation in NYC. Additionally, the researchers can develop a traffic predictor to forecast the journey period given certain conditions. In addition to the extensive data on green and yellow cabs in NYC, processing of traffic data is done using MATLAB and Spark, which is a distributed data processing platform and examines the data using the K-means clustering algorithm of the boarding area.

Faial et al. (2020) also claimed that a method for forecasting taxi demand that makes use of stream machine learning algorithms to deal with concept drift detection on taxi data is presented in the studies. Faial et al. (2020) also added that they utilized the Massive Online Analysis (MOA) device, a structure for information stream mining. Using real data from the New York open platform, it is used to create a stream learning model. The stream model forecasts taxi demand with 78% accuracy, which is encouraging. Despite employing data from a particular location, the methods and findings of the study can help other cities manage demand more proactively.





Wickramasinghe et al. (2019) applied regression-based machine learning approaches to forecast hourly taxi trips for a particular location in a target day of the week and month. Using the trip record data gathered by the Taxi and Limousine Commission (TLC) during 2017 and 2018, Random Forest regression was discovered to be capable of accurately forecasting hourly taxi drop-offs for both a specific taxi zone and the entire city of New York. Stoyanovich et al. (2017) discovered that there are hotspots for taxi usage. For example, JFK, La Guardia or Penn Station as their origin or final destination.

Gunawan (2013) mentioned that in density-based clustering, a density threshold is used to define the clusters. Since it is able to detect clusters of any form of shapes and does not necessarily require to have a knowledge of the number of clusters, DBSCAN is frequently used to group a collection of points. It just uses linear operations and requires two arguments; thus, it runs fairly quickly with a number of data range queries. Moreover, Tran et al. (2013), states that due to its ease of use and capacity to identify clusters of various sizes and forms, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) has been extensively used in numerous scientific fields during the past few years.

Çelik et al. (2011) used the DBSCAN method to find anomalies in monthly temperature data. A density-based clustering technique called DBSCAN is capable of spotting data anomalies. During the experimental evaluation, Çelik et al. (2011) contrasted the DBSCAN algorithm's findings with those of a statistical approach. The investigation revealed that when it comes to finding anomalies, DBSCAN offers a number of advantages over the statistical technique.





Overall, two data mining techniques will be used which are the statistical tool (Pearson's Correlation Coefficient) and clustering algorithms (HDBSCAN and DBSCAN). According to Bose (n.d.), the study of data collection and analysis is the focus of the mathematical discipline known as statistics. Data Mining method is recommended to be used as a statistical method. Moreover, it aids in pattern discovery and the creation of predictive models. On the other hand, clustering is one of the oldest data mining methods. It is the procedure of finding data that are similar to one another. Clustering, also known as segmentation, aids users in comprehending the database's operations. The Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm and the Pearson Correlation Coefficient statistical tool will be applied to the data available. The Pearson Correlation Coefficient and clustering algorithms will significantly help to determine the clustering pick-up and drop-off locations through the use of latitude and longitude variables in which the HDBSCAN algorithm will take care of, correlating two variables such as time and distance, and duration and distance using the Pearson Correlation Coefficient statistical tool. In conjunction, it enables the study to provide and visualize outputs that will be discussed and interpreted later on.

Data Mining techniques such as the HDBSCAN and DBSCAN clustering algorithms aid in converting raw data into useful and understandable visualizations and interpretations. Data mining algorithms paved way to ease in conducting studies that are numerical or quantitative in nature which can pose a positive impact in different areas of studies.





## Chapter IV

### Datasets

#### **Dataset**

The New York City Taxi & Limousine Commission (NYC TLC) provided the dataset that will be used in this study. The Medallion (Yellow) taxi cabs, for-hire vehicles (black cars, community-based liveries, and luxury limousines), commuter vans, and paratransit vehicles of New York City are regulated and licensed by TLC. However, in this particular study, the researchers will only focus on yellow taxi trips around New York City during September of 2015. As mentioned by McCulley (n.d.), in New York City, September is an indication that it is the beginning of Christmas and fall season and is the perfect time of the year to travel to New York City especially for the tourists around the world which will also be the perfect time to conduct this study. The dataset that will be used in the study contains 769,149 rows which will be cleaned up later on in the study. Moreover, on the NYC TLC database of datasets, datasets that are after 2015 no longer have the longitude and latitude variables. These two variables are vital in this study so the researchers will stick to the year 2015 where longitude and latitude variables are available.

The dataset comes from the companies approved by the Taxicab & Livery Passenger Enhancement Programs (TPEP/LPEP) which will then be collected and submitted to the New York City Taxi and Limousine Commission (TLC) database.

The fields in the dataset are the following: vendorID - a code that tells you which TPEP provider gave you the record. tpep\_pickup\_datetime - Time and date when the meter started. tpep\_dropoff\_datetime - Time and date when the meter stopped. passenger\_count





- Total number of passengers inside a taxi. trip\_distance - The measure of the elapsed distance of the trip in miles by the taximeter. pickup\_longitude - Longitude where the meter was started. pickup\_latitude - Latitude where the meter was started. RateCodeID - Final rate code identification in effect right after the trip. store\_and\_fwd\_flag - This variable signifies whether the trip record was held in the taxi database before sending it to the company. dropoff\_longitude - Longitude where the meter was disengaged. dropoff\_latitude - Latitude where the meter was disengaged. payment\_type - the kind of payment the passenger used to pay (in numeric code). fare\_amount - The time-and-distance fare calculated by the meter. MTA\_tax - 0.50 tax amount that will be triggered automatically depending on the metered rate in use. improvement\_surcharge - 0.30 improvement surcharge assessed trips at the flag drop. tip\_amount - this is only applicable for those whose payments are credit cards. tolls\_amount - sum of all tolls paid in a single taxi trip. total\_amount - Total amount to be paid by the passengers in a taxi trip.







## Chapter V

### Methodology

Patel & Patel (2019) asserted that the research challenge can be approached methodically using research methodology. It might be understood as a discipline that investigates the approaches taken in scientific investigation. In it, the researchers look at the many techniques that a researcher typically uses to assess his study topic and the justifications for them. Being knowledgeable about both the approach and the research procedures and methods is a must. Moreover, knowing data analysis, data gathering, different kinds of data techniques, and finally, data interpretation is significant.

#### Design Method

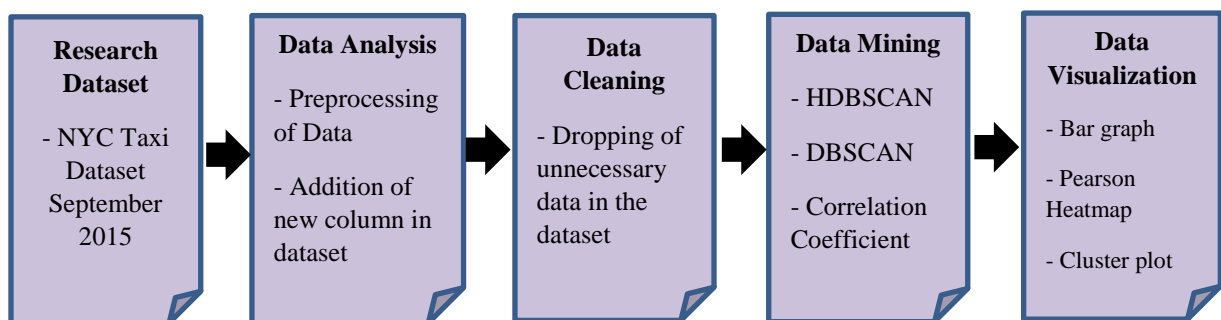


Figure 1. The phases in processing the research dataset

#### Phase 1: Research Dataset

Several datasets have been taken into consideration in this phase. The researchers then establish a goal and eliminate datasets that are of less significance to the goal. There are also other aspects that should be taken into account when meticulously selecting datasets such as the credibility of the source of the dataset and preferring raw dataset over polished ones. Through this, the New York City Taxi dataset for September 2015 is the







compatible dataset from NYC TLC that is feasible in conducting comparative analysis between HDBSCAN and DBSCAN by using the variables pick-up longitude and latitude, and drop-off longitude and latitude.

### **Phase 2: Data Analysis**

After selecting the desired dataset, analysis and planning of the dataset takes place. Analyzing what needs to be done is vital when dealing with research data. According to Selvaraj (2020), programming tools are used in data analysis to decipher relevant information from vast quantities of complex data. It is the process which follows after data collection. Methodizing the processes towards a specific goal tends to produce useful interpretations which can be true in this specific study which is the comparative analysis of HDBSCAN and DBSCAN using the latitude and longitude variables.

### **Phase 3: Data Cleaning**

Since the researchers use a raw dataset, it is important to tidy the things that are found in the dataset. This includes missing rows, missing columns, and dropping columns that are unnecessary for performing data mining later on. In this specific study, adding and dropping empty fields and unnecessary columns is necessary since it can affect the efficiency of the algorithm later on. After cleaning the dataset, the researchers were left from 769,149 rows of data to 738,656. For example, latitudes and longitudes that are outside New York City's geographical latitude and longitude are removed. Modifying the dataset can be done mostly using Jupyter Notebook. But in the researchers' case, Visual Studio Code is used by using the Jupyter extension and installing Python version 3.9.13.





#### Phase 4: Data Mining

In this phase, the researchers decide what algorithms and statistical tools that suit well in the dataset that they have selected.

#### Pearson's Product-Moment Correlation Coefficient

The Pearson Product - Moment Correlation Coefficient will be performed in this particular study. This statistical tool will help the researchers as to how two variables correlate to each other. Pearson's Product Moment Correlation Coefficient (PPMCC) is a measurement of correlation between variables. This particular study includes 3 variables that will be used to measure the correlation against each other. The variables include: trip distance versus trip duration and trip distance versus fare amount.

The Pearson's  $r$  should be a number between -1 and 1. In this study, most of the values of  $r$  are  $r > 0$ , hence, the scale of correlation coefficient should be followed.

Table 1. Scale of Pearson's Correlation Coefficient

Scale of correlation coefficient	Value
$0 < r \leq 0.19$	Very Low Correlation
$0.2 \leq r \leq 0.39$	Low Correlation
$0.4 \leq r \leq 0.59$	Moderate Correlation
$0.6 \leq r \leq 0.79$	High Correlation
$0.8 \leq r \leq 1.0$	Very High Correlation

The values of Pearson's  $r$  have corresponding values ranging from different categories. Values that are greater than zero but less than or equal 0.19 are considered as Very Low Correlation between two variables. Values that are greater than 0.19 but less





than or equal to 0.39 are considered as Low Correlation. On the other hand, values that are greater than 0.39 but less than or equal to 0.59 are considered as Moderate Correlation. Values that are greater than 0.59 but less than or equal to 0.79 are considered High Correlation. Lastly, values that are greater than 0.79 but less than or equal to 1.0 are considered Very High Correlation. This means there is a direct proportion between two variables as they have a strong correlation with each other.

Kenton (2022) defined Pearson Coefficient as a type of correlation coefficient that shows how two variables that are measured on the same ratio or interval scale relate to one another. Pearson Coefficient statistical tool will be applied for the different variables that are in the dataset such as comparing trip distance and trip duration, and trip distance and fare amount. Moreover, it is also important to visualize certain profiles of the taxi dataset such as number of passengers, trip duration, trip distance, pick up hour, drop off hour, payment type, fare amount, tips amount, and total amount into graphs and will then be interpreted later on.

### **Hierarchical Density-Based Spatial Clustering of Applications with Noise**

Stewart & Al-Khassaweneh (2022) mentioned that HDBSCAN is a hierarchical density-based algorithm proposed by Campello, Moulavi, and Sander in the early 2010s. Campello, Moulavi, and Sander introduced their work by implementing a complete framework for clustering and outlier or noise detection which includes a more in-depth explanation of HDBSCAN algorithm. HDBSCAN is the successor to DBSCAN which introduces a hierarchical analysis of the clusters. HDBSCAN is a more recently developed





algorithm built upon DBSCAN, which, unlike its predecessor, is capable of identifying clusters of varying density. According to Dorfer (2022), HDBSCAN, while not perfect, is typically more prudent with the assignment of noisy data points to clusters. Moreover, DBSCAN tends to underperform while identifying clusters with non-uniform density. Unlike HDBSCAN, DBSCAN is quite susceptible to noise, which can result in incorrect clustering. This problem was the main motivation behind the development of HDBSCAN and, as a result, it handles clusters of varying density much better than DBSCAN.

Fuchs (2020) added that HDBSCAN is a hierarchical density-based clustering algorithm. In order to apply the HDBSCAN algorithm to the dataset, the following things are done simultaneously. Firstly, the researchers look for the longitude and latitude variables for both pick-up and drop-off longitude and latitude in the dataset. Secondly, the researchers then used these variables: pickup\_latitude, pickup\_longitude, dropoff\_latitude, and dropoff\_longitude to provide a cluster in the New York City map. Then, modifying the min\_samples and min\_cluster\_size takes place as to what suits the performance of the algorithm when used in the dataset. Values 1 and 5 are used respectively in min\_samples and min\_cluster\_size. Next, using the OpenStreetMap (OSM) and HDBSCAN python import to provide a map background in the HDBSCAN cluster to provide a more comprehensible visualization of the clusters in the New York City setting. Lastly, adjusting the coordinates (longitude and latitude) in accordance to the geographical location of New York City is a must to put a limit as to the specifics of the study.





Moreover, the researchers chose OpenStreetMap (OSM) import as the background image of the map in providing HDBSCAN and DBSCAN cluster points.

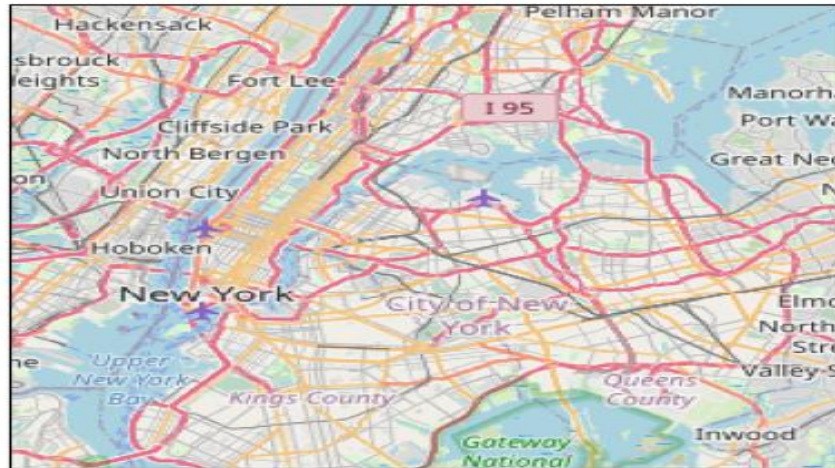


Figure 2: OpenStreetMap Python import

### Phase 5: Data Interpretation

Schoen (n.d.) affirmed that data interpretation includes a sequence of inquiries regarding your data that are relevant to your research question(s). Schoen (n.d.) also stated that responses to these queries are arranged in the form of results and conclusions. After analyzing and performing certain algorithms to the dataset, interpretation should take place as it is vital throughout the process. Using the graph outputs of performing profiling, correlation, and HDBSCAN, interpretation of data follows in order to better understand the results and produce useful insights and draw impactful conclusions onto the study.

The tools that are used in visualization of data are: bar graphs for frequency count, Pearson's Correlation Matrix for comparison of two variables, and lastly, HDBSCAN and DBSCAN map clusters for finding the dense regions of the taxi hotspots and also comparing the results of HDBSCAN and DBSCAN output in terms of efficiency and accuracy.





## Chapter VI

### Data Analysis and Interpretation of Results

This part of the chapter presents the data and interpretation of the results.

#### I. Frequency of taxi trips in terms of:

##### I.1. Passenger Count Distribution

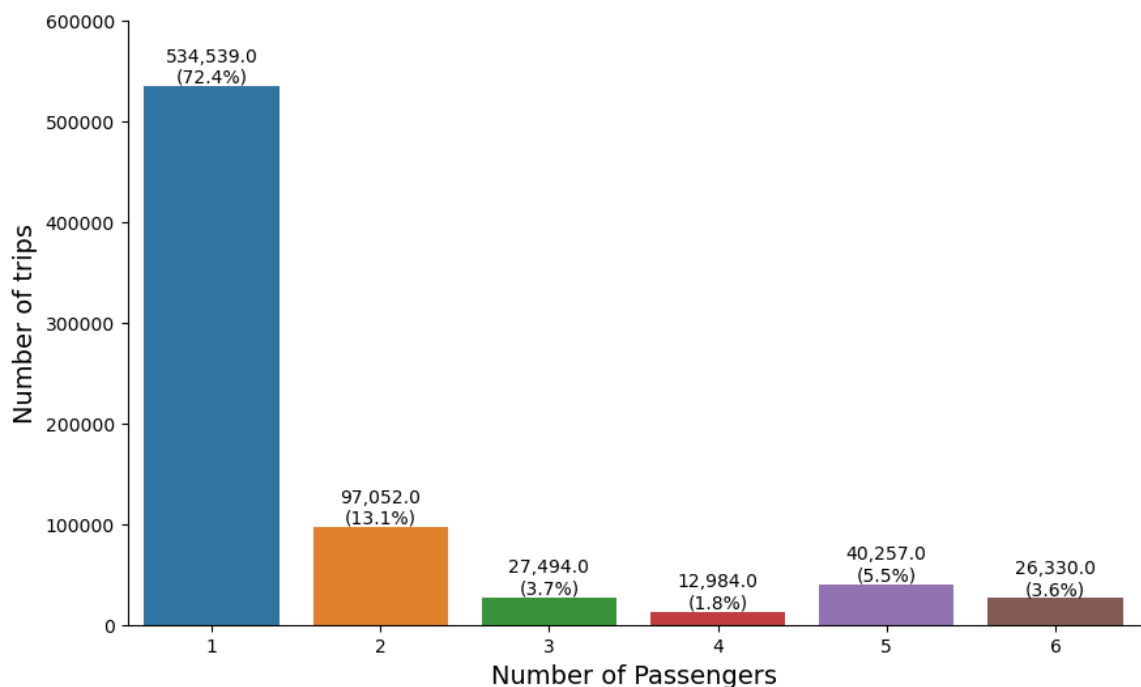


Figure 3: Passenger Count Distribution

Figure 3 shows that there are 534,539 taxi trips (72.4%) that had only one passenger, 97,052 taxi trips (13.1%) that had two passengers, 27,494 taxi trips (3.7%) that had three passengers, 12,984 taxi trips (1.8%) that had four passengers, 40,257 taxi trips (5.5%) that had five passengers, and lastly, 26,330 taxi trips (3.6%) that had six passengers.







New York City is still considered among the most densely populated cities in the planet. According to Bass (2021), New York City is still dense. New York County remained number one in the recent Census, with 74,781 individuals per square mile, then Kings County with 39,438 individuals, Bronx County with 34,920 individuals, and lastly, Queens County with 22,124 individuals. By being densely populated, it can equate to more demands in transportation, which means that taxis are often used for individual trips rather than shared rides. Taxis are a common choice for personal or single person transportation in NYC since they are practical and dependable.

Many passengers prefer to travel alone in a taxi for privacy and comfort reasons. They may not want to share the ride with strangers, or they may simply prefer to have the vehicle to themselves. Also, New York City has a proper and adequate public transportation system, including subways and buses, that provides affordable and convenient alternatives for group travel which leaves individual tourists to take taxis for their individual transportation needs which answers as to why there are more trips that carries one passenger in New York City.





## I.2. Trip Duration

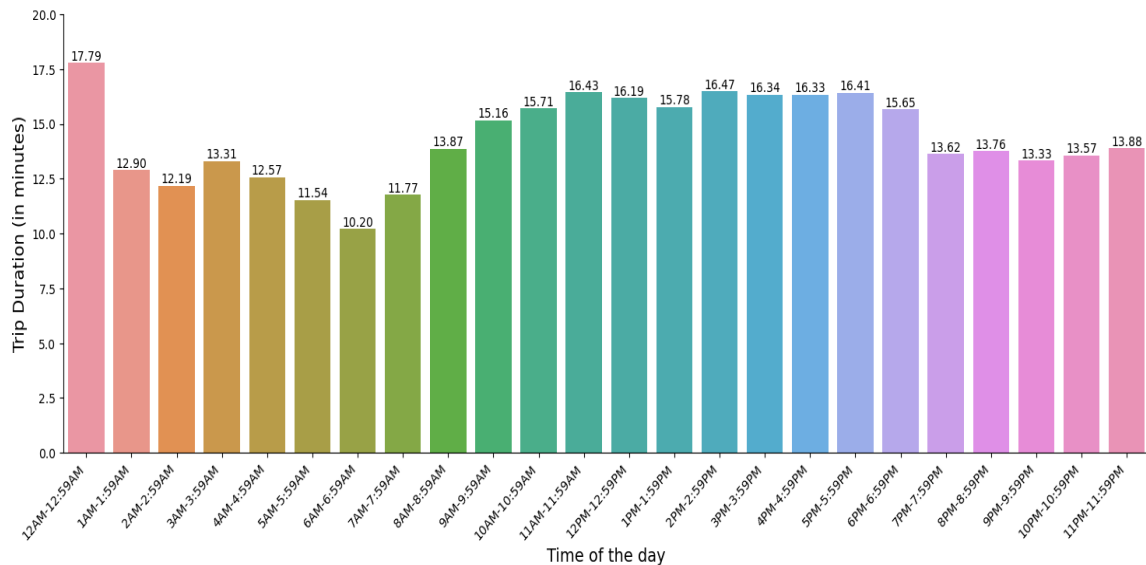


Figure 4: Trip Duration Distribution

In Figure 4, it shows the trip duration distribution in minutes. From 12 midnight until 5:59AM (early morning), the average trip duration is 13.38 minutes, then in the morning around 6AM to 11:59AM, the average trip duration is 13.86 minutes. On the afternoon, from 1PM to 5:59PM, the average trip duration is 16.27 minutes. At evening, from 6PM to 11:59PM, the average trip duration is 13.97 minutes. In addition, it is shown in the figure that 12AM-12:59AM has the highest trip duration, with 17.79, while 6AM has the shortest, having an average trip duration of 10.20 minutes. One of the contributing factors for this is the traffic in New York City.

Poongodi et al. (2022) described that the majority of traffic in New York City is made up of taxi rides. The numerous rides that New Yorkers take throughout the hectic metropolis each day might be an excellent indicator of traffic volumes, roadblocks, etc.







Additionally, during the early afternoon, many people are on the road, either going to work, running errands, or heading home, which can result in heavy traffic and longer travel times. Moreover, weather conditions, such as rain or snow, can also impact travel times and make taxi trips longer in the early afternoon. It is also crucial to estimate how long a taxi ride will take because a user always wants to know exactly how long it will take him to get from one area to another.

### I.3. Trip Distance

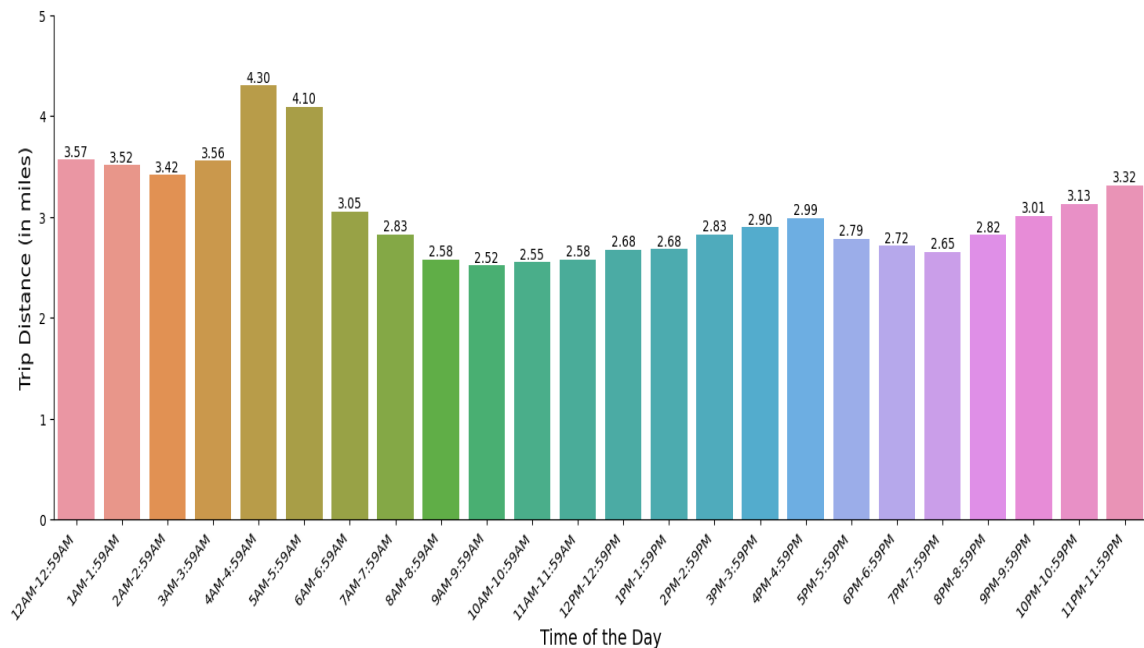


Figure 5: Trip Distance Distribution

Figure 5 shows the longest average trip distance which is around 4AM-4:59AM, having an average of 4.30 miles. The shortest average trip distance is around 9AM-9:59AM with an average of 2.52 miles. One of the contributing factors for this is the traffic in New





York. This is justifiable due to the fact that there are a number of passengers that are very distant to their workplaces which sometime might take more time to reach their workplaces.

Deng et al. (2020) claimed that most early route choice behavior analysis studies used data from expressed preference surveys or from small-scale tests. If many passengers are traveling to or from outlying areas such as suburbs, remote workplaces, or residential neighborhoods located outside of the city center, the trips may be longer. This is because these areas are often located farther away from the city center, and reaching them may require longer routes. In addition, these areas may also have less efficient road networks, which can also add distance to the trip. In general, passengers who live or work in outlying areas may have longer commutes and require more time and distance to reach their destinations. These longer trips may contribute to the overall increase in taxi trip distance in the early morning.





#### I.4. Pickup Hour Distribution

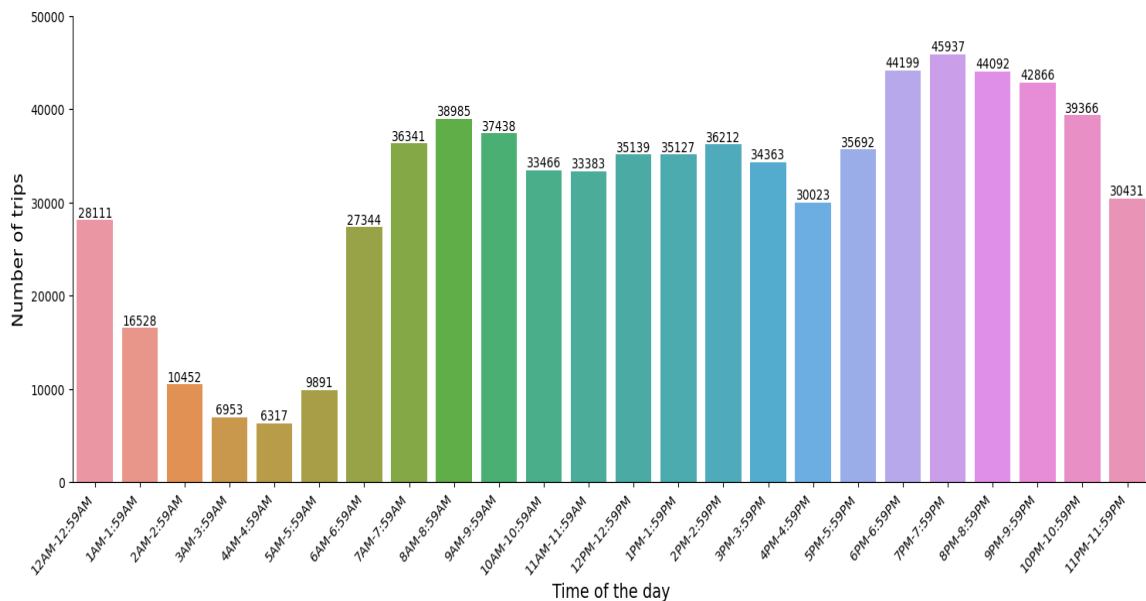


Figure 6: Pickup Hour Distribution

The figure 6 above represents the pickup hour distribution. It shows that 7:00 PM has the highest number of pickup trips with 45,937 trips. In contrast, the hour that had the least number of pickup trips is 4:00 AM which only had 6,317 trips. In addition, there are usually more pickup trips in the early evening hours which usually starts at 6:00PM until 10:00PM while lesser pickup trips are happening during dawn starting at 1:00AM until 5:00 AM.

Bischoff et al. (2015) described that there is often a morning peak in demand followed by an afternoon peak that lasts for a lengthier period of time. Weekends see a shift in the demand peaks from day to night. 6:00AM starts to get busier since it is when people go to their respective work. And during early evening, people start to commute on their way home which explains why 7:00PM is the busiest hour in taxi trips.





### I.5. Dropoff Hour Distribution

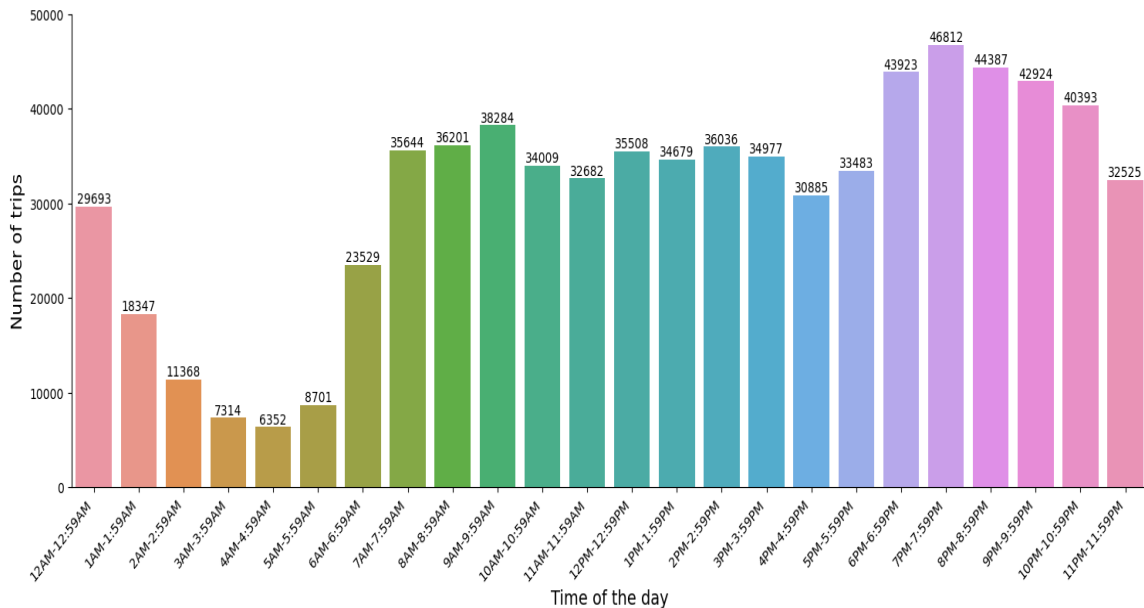


Figure 7: Dropoff Hour Distribution

Figure 7 shows that 7:00PM has the highest number of drop off trips with 46,812 trips. On the other hand, the hour that had the least number of drop off trips is 4:00AM which only had 6,352 trips. In addition, there are usually more drop off trips in the early evening hours which usually starts at 6:00PM until 10:00PM while lesser drop off trips happen during dawn starting at 1:00AM until 5:00AM and then more drop off trips happened 6:00AM onwards.

There is an increased demand during the late evening as people may be finishing work, going out for dinner or evening activities, or traveling to and from the airport. Bischoff et al. (2015) asserted that there is often a morning peak in demand followed by an afternoon peak that lasts for a lengthier period of time. The rise in the number of people using services like Uber and Lyft may also contribute to the increase in taxi drop-offs





during this time. Additionally, factors such as traffic, road conditions, and weather can also play a role in determining the number of taxi drop-offs during early evening.

### I.6. Payment Type Distribution

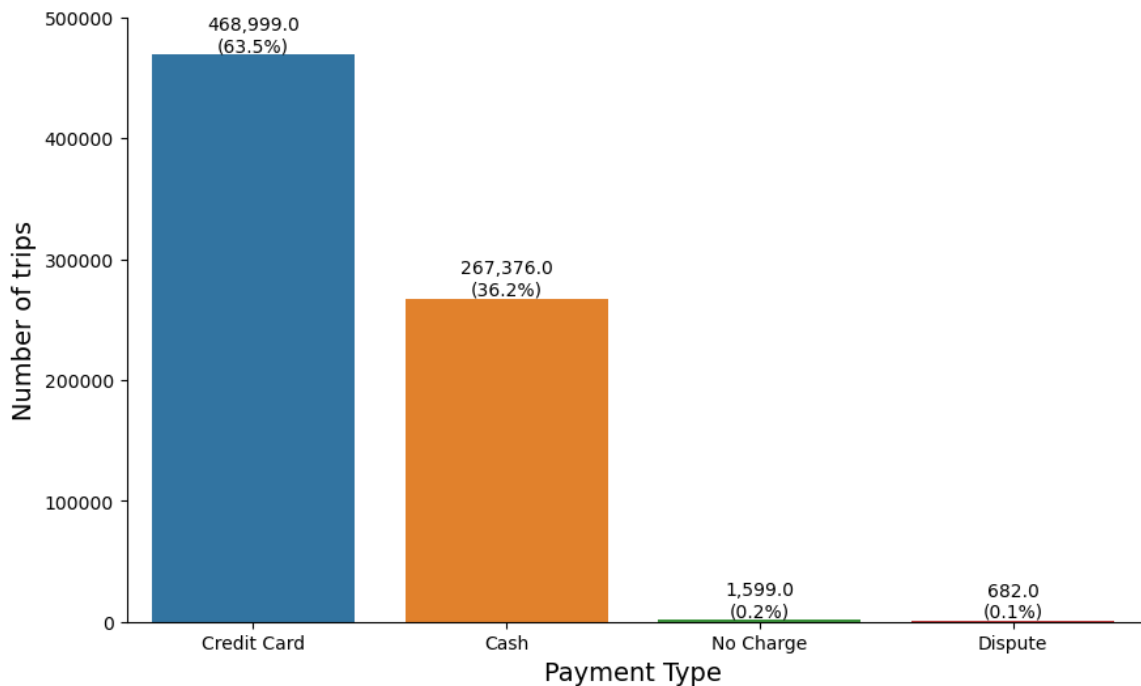


Figure 8: Payment Type Distribution

Figure 8 shows the frequency count of the taxi trips in terms of payment type. The figure shows that the Credit Card is the most used payment type in taxi trips with 468,999 trips (63.5%), followed by Cash with 267,376 trips (36.2%), No Charge with 1,599 trips (0.2%), and Dispute with 682 trips (0.1%).

Grynbaum (2009) purported that cab drivers in New York City were shocked when government officials required them to accept credit cards as payment methods. Most of the cab drivers went on strike, accusing the regulations of being burdensome for drivers and a





kowtow to visitors. However, two years later, even as fleets in other cities were having difficulty in obtaining payments in a severe recession, the back-of-the-cab swipe emerged as an unforeseen savior for New York's taxi sector. Both ridership and income have grown generally. Even for shorter rides, credit card payments for fares are becoming more and more common. And tips for drivers, who are typically the first to suffer in hard times, have increased significantly since the advent of plastic. Grynbaum (2009) also stated that credit cards are beneficial for business, even according to cab drivers.

King & Saldarriaga (2017) also asserted that the payment of taxi fares by neighborhood and looked into the relationship between paying for taxi fares and using traditional banking services. Riders' tendency to pay in cash has distinct spatial dimensions. Moreover, both immigration status and being "unbanked" are significant determinants of payment transactions. Through these findings, there have been consequences on regional laws governing the for-hire vehicle market, specifically in light of explosive expansion of services requiring credit card payments.





### I.7. Fare Amount Distribution

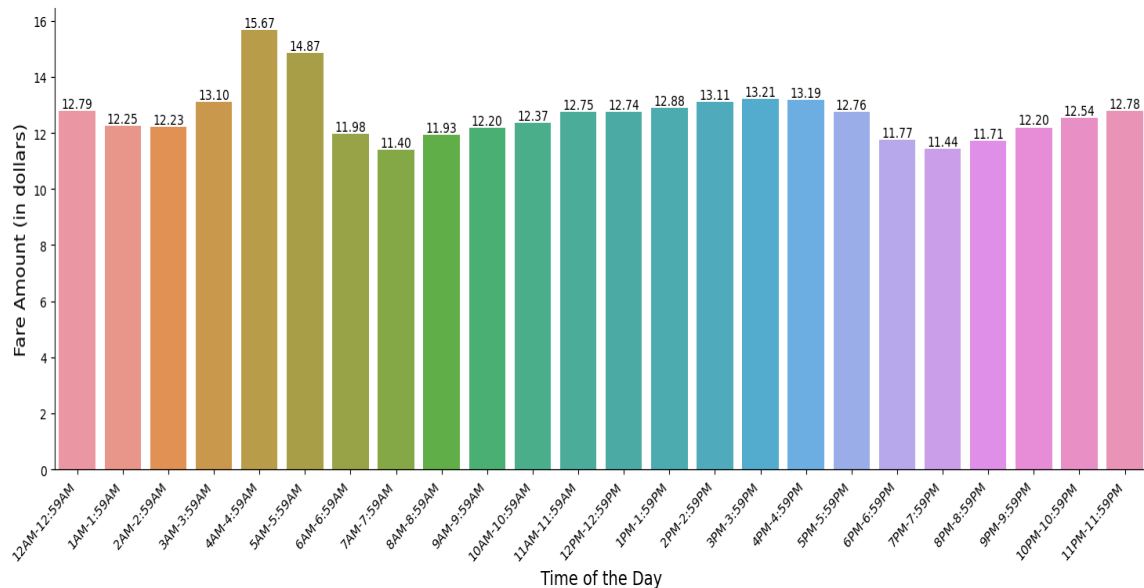


Figure 9: Fare Amount Distribution

As seen in Figure 9, the hour with the highest average fare amount, that is \$15.67, was during 4 AM while the hour with the least average fare amount, that is \$11.40, was during 7 AM. Overall, the fare amounts are averaging between \$11.40 to \$15.67.

Schaller (1999) declared the impact of rising taxi fares on demand for trips and the accessibility of taxi services. The elasticity of trip demand about prices is 0.22; the elasticity of service availability about taxi rates is 0.28; and the elasticity of service availability about the entire supply of services is close to 1.0. For judgments about taxi regulations, these findings have significant ramifications. Schaller (1999) also stated that fare hikes significantly boost industry income at a slower rate than the percentage increase in the fare. The policy-making conclusion is that fare elasticities must be considered appropriate to achieve desired increases in drivers' incomes. Moreover, Schaller (1999)





states that given the significant influence of prices on availability, service availability—a crucial component of service quality frequently disregarded during discussions on fare policy—should be a factor in fare setting. Furthermore, when there is a need to increase the supply of taxis, it is possible to do so without negatively impacting the income of current operators.

### I.8. Tip Amount Distribution

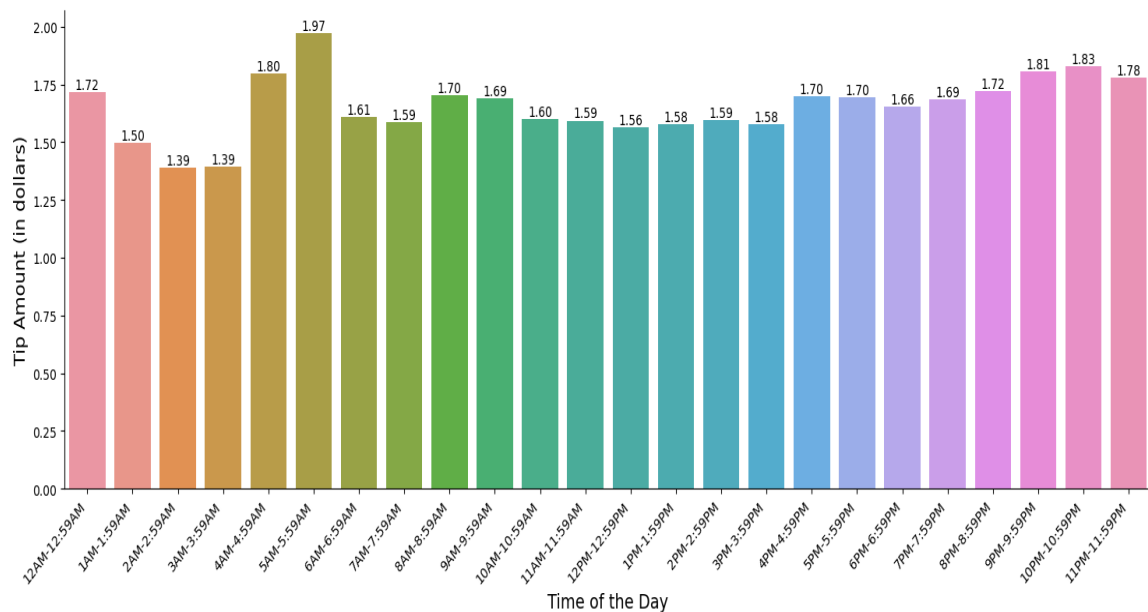


Figure 10: Tip Amount Distribution

Figure 10 shows the distribution of average tip amounts of taxi trips per hour. The hour with the highest average tip amount, that is \$1.97, was during 5AM while the hours with the least average fare amounts, that is \$1.39, were during 2AM and 3AM. Overall, the tips are averaging between \$1.39 to \$1.97.







Azar (2020) described that millions of workers depend heavily on tips, which generates lots of money each year in the US alone. Beyond its economic significance and ramifications, tipping is considered a distinctive economic phenomenon because it is not forced which means that it is voluntary and unconstrained by law. Tipping shows how psychological and social incentives can be a significant factor in economic conduct and how economic models should account for more than just a self-interested, emotionless economic agent in order to encompass the whole spectrum of economic activities. Furthermore, Lynn et al. (1993) stated that it is one's decision as to whether to tip. The causes of people's tipping decisions have been the subject of academic research on this unusual but common consumer behavior.





### I.9. Total Amount Distribution

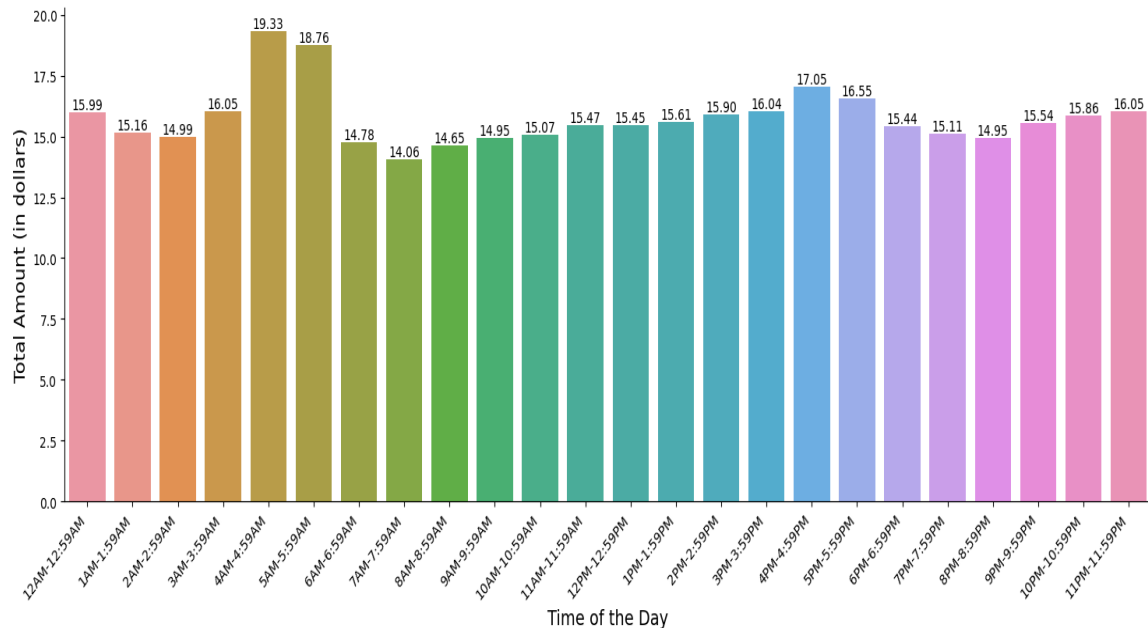


Figure 11: Total Amount Distribution

Figure 11 above shows the hour with the highest average total amount, that is \$19.33, was during 4AM while the hour with the least average total amount, that is \$14.06, was during 7AM. Overall, the tips are averaging between \$14.06 to \$19.33. This can be due to various reasons such as the distance of the trip being far enough.

Land (2022) denoted that New York City's TLC voted to increase the metered fares by 23% which is the really new for how many decades. The increase in price includes a 10-cent increase in fares for taxis including green and yellow, from \$2.50 to \$3. Depending on the distance, a trip may cost between \$8 and \$10. The taxi or a limousine are two options. One-way costs for the trip might range from \$50 to \$170.





## 2. Pearson's Correlation Coefficient

### 2.1. Correlation Between Trip Distance and Trip Duration

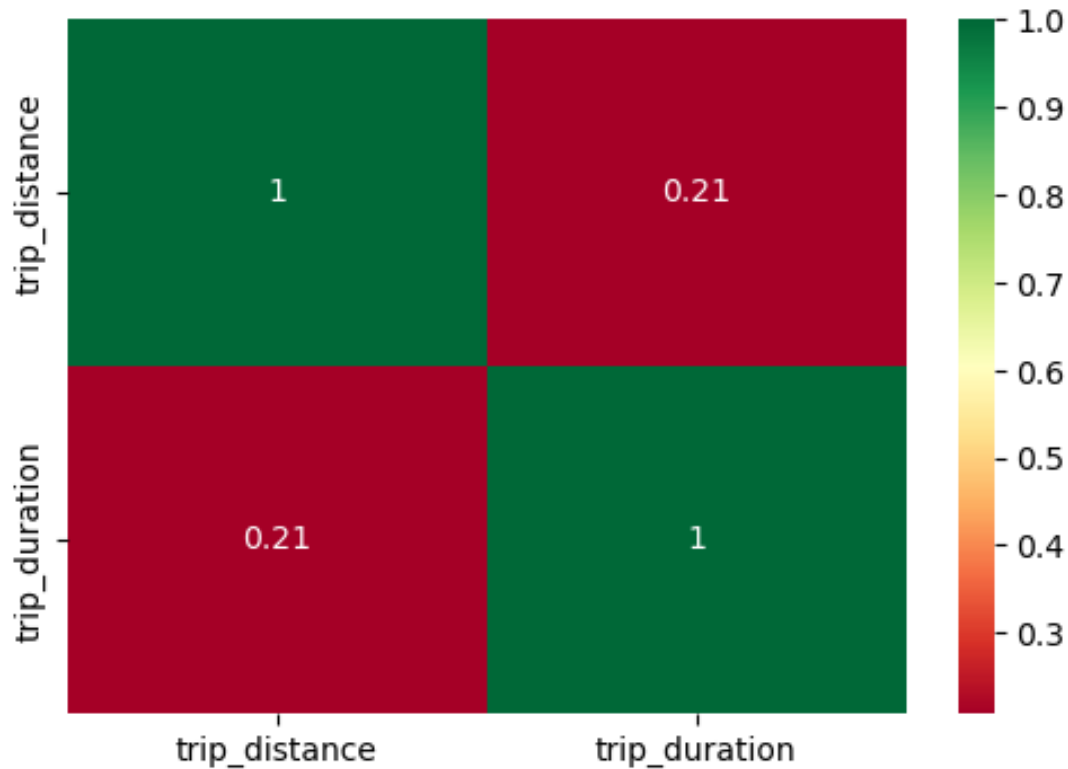


Figure 12: Trip Distance vs. Trip Duration Correlation

Figure 12 shows the value of Pearson's  $r$  between trip distance and trip duration using heatmap matrix. The values in red-colored squares,  $r = 0.21$ , indicate the value of how strongly the two variables correlate with each other. When based on the scale of correlation coefficient from Table 1, the value of  $r$  which is 0.21 falls under the Low Correlation category meaning the two variables have a weak correlation with each other.

According to Pu (2011), there are numerous ways of estimating travel time dependability. In transportation engineering, on-time arrival, the frequency of congestion,





the standard deviation, the coefficient of variation, the percent of variation, the buffer index, the planning time index, the travel time index, the misery index, and the travel time index are all used. Moreover, trip duration and trip distance in taxis may have a low correlation for several reasons such as traffic conditions, speed limits, driving style, and stopovers.

According to Morris (2015), trip duration affects mood and emotions significantly but weakly. It can also be heavily influenced by traffic conditions, such as congestion, accidents, and road closures, which may not be directly related to the distance of the trip. In addition, the correlation between distance and duration also depends on the speed at which the taxi driver is driving, and whether they are following the speed limits or not. Lastly, trip duration can also include stopovers, such as for restroom breaks, food, or to drop off additional passengers, which can add time to the trip without increasing the distance traveled all of these probable reasons provide support as to why there is a low correlation between the two variables, namely, trip distance and trip duration.





## 2.2. Correlation Between Trip Distance and Fare Amount

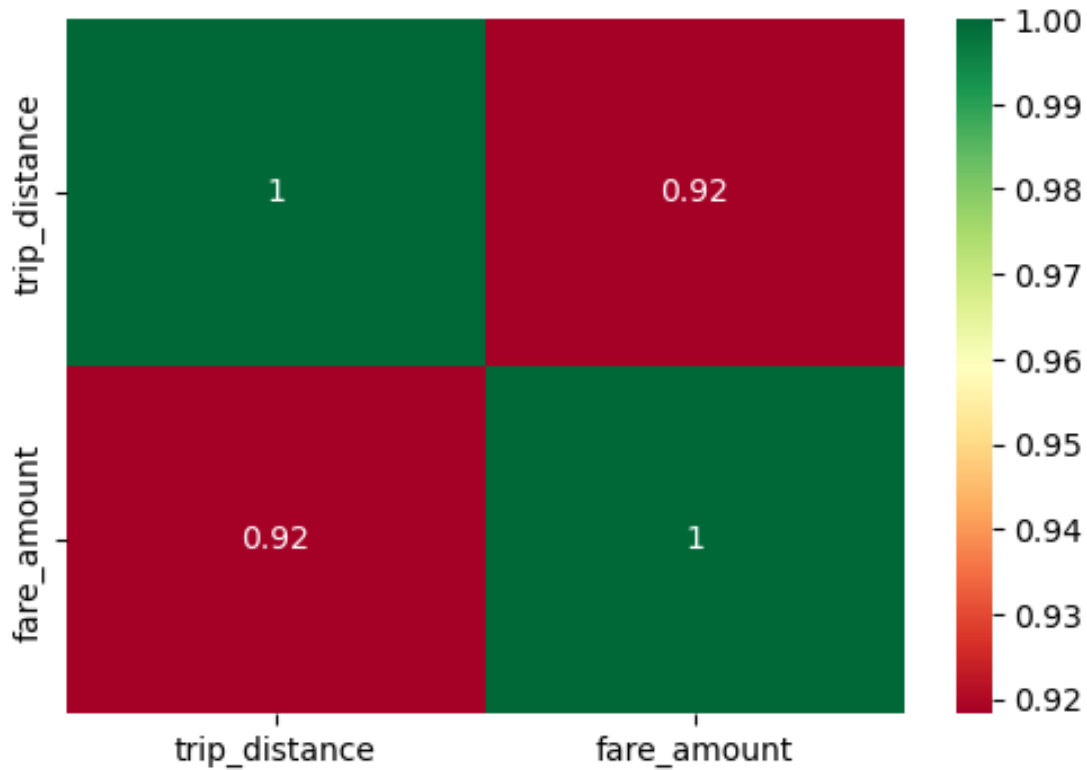


Figure 13: Trip Distance vs. Fare Amount

Figure 13 shows the value of Pearson's  $r$  between trip distance and fare amount using heatmap matrix. The values in red-colored squares,  $r = 0.92$ , indicate the value of how strongly the two variables correlate with each other. When based on the scale of correlation coefficient from Table 1, the value of  $r$  which is 0.92 falls under the Very High Correlation category meaning the two variables have a very strong relation with each other.

According to Tedesco (2023), it's a myth that taxis in NYC are absurdly expensive. Trip distance and fare amount in taxis are likely to have a very high correlation because the fare amount is typically calculated based on the distance traveled by the taxis. Most





taxi companies use a combination of a base fare and a per-mile or per-kilometer rate to calculate the fare, so as the distance traveled increases, the fare amount will also increase.

When applying in the New York City setting, it is still the same as to the fare calculations. According to Taylor (2022), having and driving a car in New York City can be a difficult and expensive option. There is limited parking, and car owners frequently rent parking spots. In contrast to popular rumors, taxi cab prices in New York are not excessively high. A typical trip in Manhattan costs about \$10 on average. Additionally, some taxi companies may also charge extra for certain services such as tolls, waiting time, or traffic congestion, which can also affect the fare amount but these charges are not related to the distance traveled. Hence, there is a direct relationship between the trip distance and the fare amount.

### 3. HDBSCAN clustering for all trips for pickup and dropoff

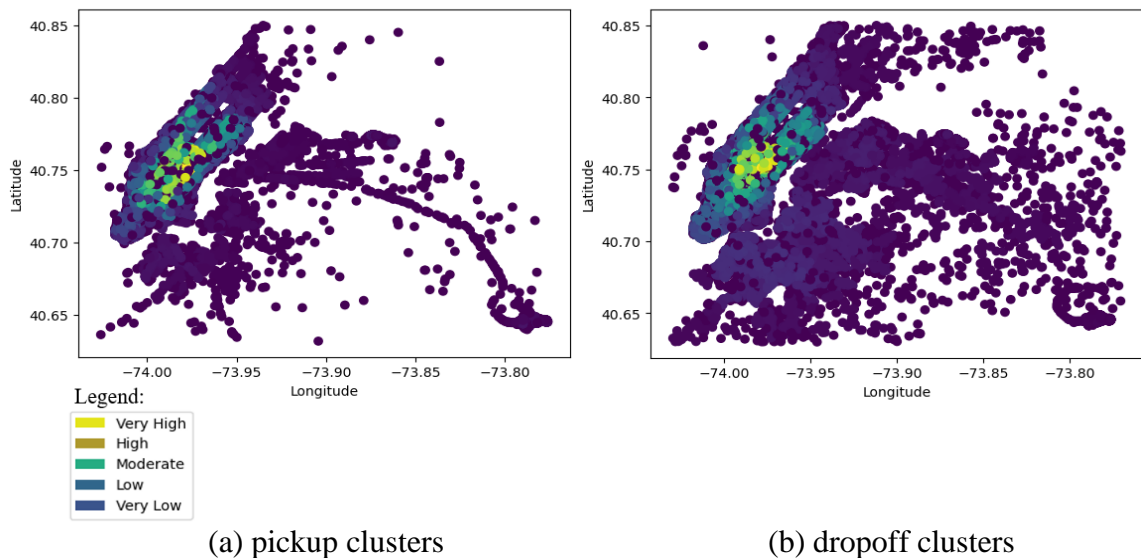


Figure 14: Cluster plot for (a) pick-up and (b) drop-off locations





Figure 14 illustrates the HDBSCAN pick up cluster points and drop off cluster points. The clusters have a property of minimum cluster size which is equal to 5 which means that there are 5 different strokes of colors in the graph, and a minimum sample of 1. The x-axis are longitude values and the y-axis are latitude values. Dark violet colors which are present among the graphs, are considered noises while the other colors have their corresponding values each. The yellow colors indicate that there is a very high accumulation of taxi pick-ups and drop-offs in that region, the brown colors indicate a high accumulation of taxi pick-ups and drop-offs in that region, the green colors indicate a moderate accumulation of taxi pick-ups and drop-offs in that region, the blue colors indicate a low accumulation of taxi pick-ups and drop-offs in that region, and the dark blue colors indicate a very low accumulation of taxi pick-ups and drop-offs in that region which can be clearly seen approximately in  $(-74.00, 40.75)$  coordinates is considered to be where frequent trips have happened.

Stewart & Al-Khassaweneh (2022) asserted that in contrast to K-means, the HDBSCAN algorithm does not require every data point to be assigned to a cluster in order to identify dense clusters. Moreover, outliers or noises are certain points that don't belong to a cluster. It is a useful algorithm that can quickly locate distinct groups in a collection and spot outliers.

In addition, Dorfer (2022) explained that, unlike HDBSCAN, DBSCAN is quite susceptible to noise, which can result in incorrect clustering. HDBSCAN, while not perfect, is typically more prudent with the assignment of noisy data points to clusters. Stewart &

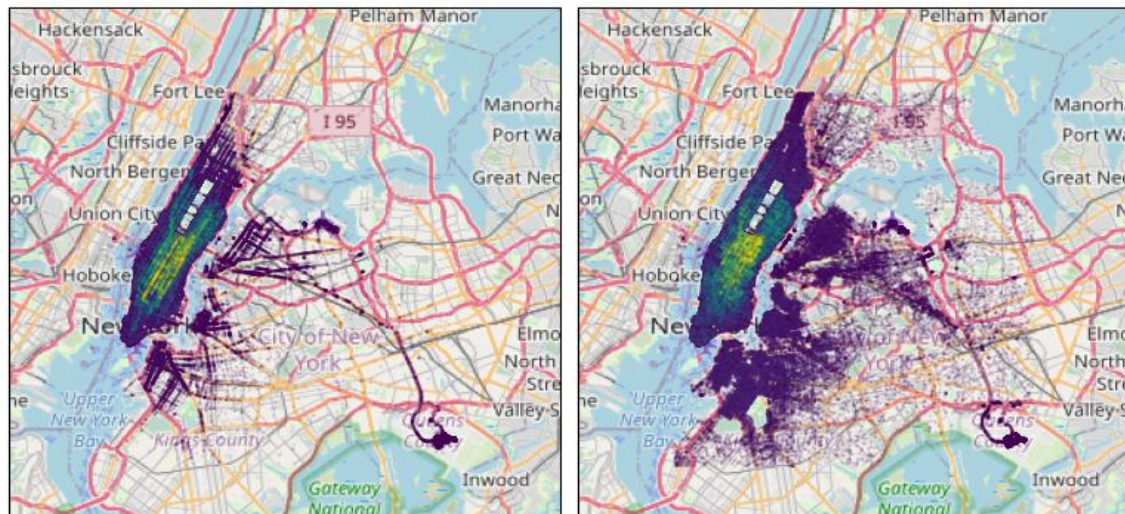






Al-Khassaweneh (2022) also added that by making a hierarchical cluster of the groups, HDBSCAN beats DBSCAN. The hierarchy created by the algorithm's execution can be used effectively for outlier detection and cluster extraction. Through the identification of clusters of any density, HDBSCAN gets beyond DBSCAN's drawbacks.

It is worth mentioning that in drop off cluster points, more noise or outlier points are detected. This can be justified since taxi drivers drop off their passengers onto their respective houses. When dealing with New York City setting, residential areas aren't just located in Manhattan which justifies as to why there are lots of noises outside the denser regions near the  $(-74.00, 40.75)$  coordinates.



Legend:



(a) pickup clusters

(b) dropoff clusters

Figure 15: Cluster plot for (a) pick up, (b) drop off locations







Figure 15 illustrates the HDBSCAN pick up cluster points and drop off cluster points in a New York City map.

The largest hotspot region in the graph is Manhattan. The majority of the total trips are within Manhattan. After that, trips between LaGuardia Airport and Manhattan, and also, between JFK Airport and Manhattan are in large numbers for any day. Moreover, as seen in the graphs above, there are dense clusters for pick up and drop off routes from Manhattan, Brooklyn, Queens, and John F. Kennedy International Airport.

According to Rosenthal (2020), there are far fewer people in New York City than in any other major city in the United States. New York City has 28,000 residents per square mile, while San Francisco has 17,000, which is the next most crowded city. Besides, on a normal business day, in excess of 5 million individuals jar onto the city's metro trains and cabs — however many outings as Los Angeles finds in a portion of a month. New York City has 400,000 more people living in cramped public housing units than any other city. Additionally, Times Square is one of the world's busiest tourist attractions, drawing nearly 40 million visitors annually.

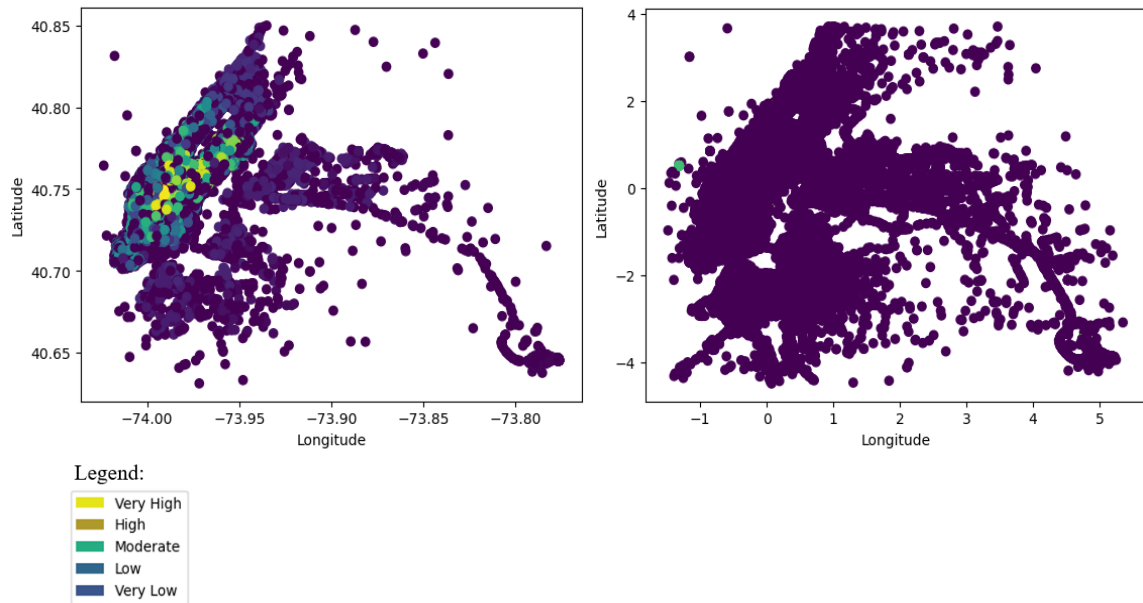
Manhattan, being one of the hotspot regions for taxis as seen in the graphs above, is considered the busiest city in New York since it is famous for its tourist attractions such as the famous Times Square. This explains as to why there are numerous trips in Manhattan over any towns in New York. In addition, yellow taxicabs frequently drop off street-hailing passengers in the Manhattan area and pick them up anywhere in New York City.





#### 4. Comparison against DBSCAN for both pick-up and drop-off locations

##### 4.1. Efficiency (HDBSCAN vs. DBSCAN)



(a) HDBSCAN

(b) DBSCAN

Figure 16: Cluster plot of pick-up and drop-off locations

Figure 16 shows comparison between HDBSCAN and DBSCAN in terms of efficiency. Based on the graphs, it is evident that the DBSCAN algorithm is having a hard time dealing with large datasets such as in this study compared to HDBSCAN which produces a more desirable output. Moreover, the DBSCAN algorithm is also having a hard time finding the dense region in the graph and it also produces more outliers and noise compared to HDBSCAN.

Dorfer (2022) asserted that DBSCAN tends to fall short of identifying clusters with non-uniform density. This problem was the main motivation behind the development of





HDBSCAN and, as a result, it handles clusters of varying density much better. In terms of performance, scalability studies also demonstrate that HDBSCAN outperforms DBSCAN in computational performance as the data increases in size. HDBSCAN is about twice as fast when compared to its predecessor, DBSCAN. In terms of efficiency, it is recommended to use HDBSCAN instead of its predecessor, DBSCAN, to produce a more desirable output.

#### 4.2. Accuracy (HDBSCAN vs. DBSCAN)

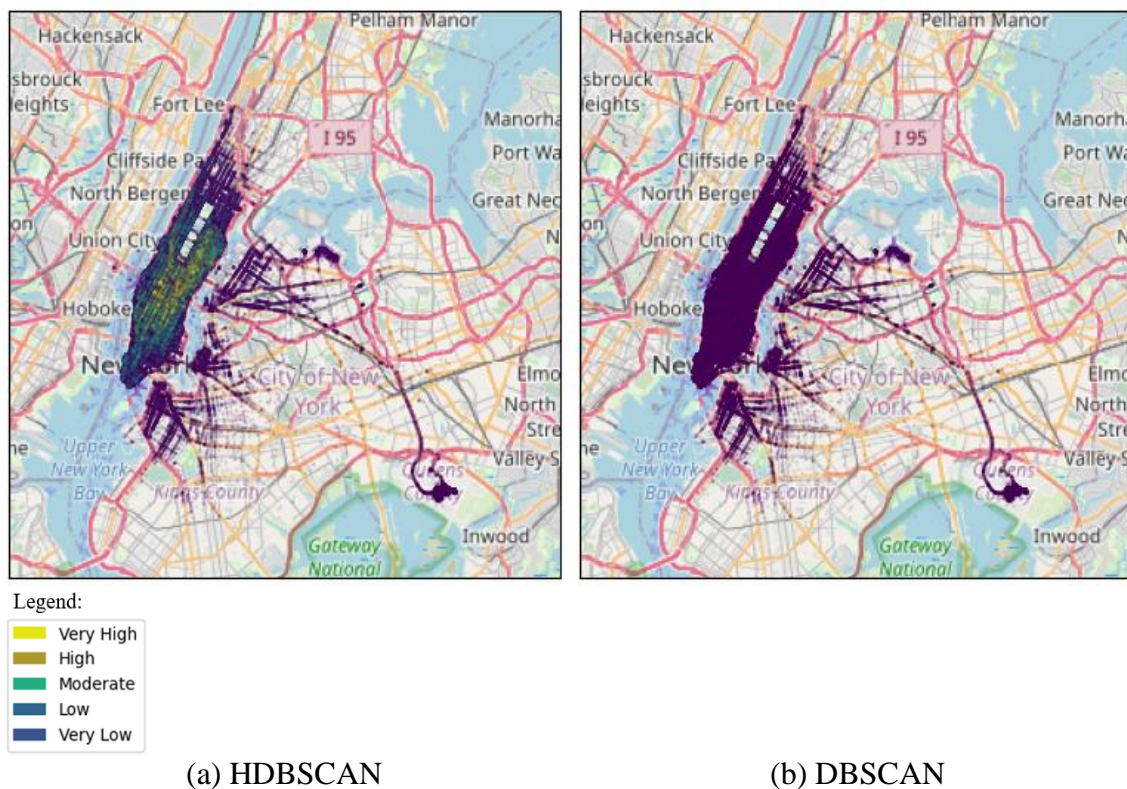


Figure 17: Cluster plot of pick-up and drop-off locations with map

Figure 17 shows the comparison between HDBSCAN and DBSCAN in terms of accuracy. Based on the graphs, the DBSCAN algorithm is having a hard time finding the





dense region in the graph and it also produces more outliers and noise compared to HDBSCAN. DBSCAN's clusters aren't accurate compared to HDBSCAN's as it is having difficulties finding clusters in the data.

Dorfer (2022) explained that DBSCAN is quite susceptible to noise, which can result in incorrect clustering. HDBSCAN, while not perfect, is typically more prudent with the assignment of noisy data points to clusters. Even though both DBSCAN and HDBSCAN perform well on data with noise and clusters of any size, there are some subtle differences between the two. Although the additional epsilon (eps) parameter of DBSCAN can be helpful if you have domain expertise regarding the data, it is frequently regarded as a very difficult parameter to optimize. In contrast, setting HDBSCAN's `min_cluster_size` is much easier to understand. HDBSCAN also wins when it comes to classifying noise and clusters of varying densities.

Although both DBSCAN and HDBSCAN work well for data containing noise and clusters of arbitrary shapes and sizes, they do have some intricate differences. Dang (2015) asserted that DBSCAN is not recommended with datasets that have altering densities. In addition, as the comparison in this study demonstrates, it will occasionally fail to identify clusters when the dataset is too sparse or has a variable density. On the other hand, Dorfer (2022) noted that HDBSCAN handles clusters of varying density much better. By contrast, `min_cluster_size` for HDBSCAN feels a lot more intuitive to set. HDBSCAN also has the upper hand when it comes to classifying noise and clusters of varying densities. All in all, HDBSCAN outperforms DBSCAN in both efficiency and accuracy categories.





## **Chapter VII**

### **Summary of Findings, Conclusion and Recommendation**

#### **Findings**

In this part of the study, the results are shown and analyzed in relation to the problems that have been identified. Based on the problem stated, the researchers have found the following:

##### **1. Frequency of Taxi Trips:**

###### **1.1. Passenger count**

In 738,656 total taxi trips, 534,539 taxi trips have one passenger. On the other hand, taxi trips that have four passengers accounts to 12,984 trips which is the least among all.

###### **1.2. Trip duration**

The trips that transpired during the 6AM-6:59AM time frame have the least trip duration throughout the entire day with an average trip duration of 10.20 minutes while the trips that occurred during the 12AM-12:59AM time frame have the most trip duration with an average trip duration of 17.79 minutes.

###### **1.3. Trip distance**

The shortest average trip distance is 2.52 miles which happened during the 9AM-9:59AM time frame. In contrast, the longest average trip distance is 4.30 miles which occurred during the 4AM-4:59AM time frame.





#### 1.4. Pickup hour

The hour with the least number of taxi pickups occurred during 4 AM with 6,317 trips recorded. On the other hand, the hour with the greatest number of taxi pickups occurred during 7 PM with 45,937 trips recorded.

#### 1.5. Dropoff hour

The hour with the least number of taxi dropoffs happened during 4 AM with 6,352 trips while the hour with the most number of taxi dropoffs happened during 7 PM with 46,812 trips.

#### 1.6. Payment type

Credit Card is the most used payment type accounting to 468,999 taxi trips. On the other hand, Dispute is the least used payment type accounting to only 682 taxi trips.

#### 1.7. Fare amount

The hour with the highest average fare amount was on 4 AM with an average fare of \$15.67 while the hour with the lowest average fare amount was on 7 AM with an average fare of \$11.40.

#### 1.8. Tip amount

The hour with the highest average tip amount was on 5 AM with \$1.97. On the other hand, the hours with the lowest average tip amount were on 2 AM and 3 AM with \$1.39 average tips.







### 1.9. Total amount

The hour with the highest average total amount for trips was on 4 AM with \$19.33 average amounts while the hour with the lowest average total amount for trips was on 7 AM with \$14.06 average total amount.

## 2. Correlation between trip distance and:

### 2.1. Trip duration

Pearson's Correlation Coefficient resulted to the value of  $r$  which is  $r = 0.21$  which can be based on the Pearson's scale of correlation as Low Correlation which means that the variables trip distance and trip duration have low correlation to each other.

### 2.2. Fare amount

Pearson's Correlation Coefficient resulted to the value of  $r$  which is  $r = 0.92$  which can be based on the Pearson's scale of correlation as Very High Correlation which means that the variables trip distance and fare amount have very high correlation to each other.

## 3. HDBSCAN cluster of taxi trips according to:

### 3.1. Pickup coordinates

The largest hotspot region for pickup is Manhattan. The majority of the total trips are within Manhattan. There are also numerous trips in Brooklyn and Queens, specifically in John F. Kennedy International Airport.





### 3.2. Dropoff coordinates

There is a high concentration of dropoff in Manhattan. Moreover, there are some accumulations of trips in JFK airport in Queens, Brooklyn and most of the places surrounding Manhattan.

## 4. Comparative performance of HDBSCAN and DBSCAN in terms of:

### 4.1. Efficiency

HDBSCAN is superior than DBSCAN in terms of efficiency. DBSCAN falls short in identifying clusters with non-uniform density. Moreover, scalability is also a huge improvement of HDBSCAN from DBSCAN as it can handle large datasets better than DBSCAN.

### 4.2. Accuracy

HDBSCAN is more favorable versus DBSCAN when speaking of accuracy. DBSCAN is having a hard time finding dense regions in large datasets when compared to HDBSCAN. Moreover, DBSCAN tends to produce more outliers than HDBSCAN which results to incorrect clustering of points.

## Conclusion

This study primarily uses the Hierarchical Density-Based Spatial Clustering Application with Noise (HDBSCAN) for clustering the NYC taxi dataset. Then it will be compared to its predecessor, Density-Based Spatial Clustering Application with Noise (DBSCAN) in terms of their efficiency and accuracy. According to the study's findings and evaluation, the following conclusions are hereby drawn:







1. Most trips have one passenger with trip durations ranges between 9 minutes to 19 minutes which frequently happen during 1 PM to 5 PM with distances ranging from 2.52 miles to 4.53 miles. Most of the pickups and dropoffs that have been recorded transpired mostly during 7 PM with Credit Card as the most common payment method. Additionally, tips are averaging between \$1.39 and \$1.97 while the total amount is ranging between \$14.06 up to \$19.33.
2. The value of  $r$  in the Pearson's Correlation Coefficient for the variables trip distance and trip duration resulted to  $r = 0.21$  which can be classified in the Pearson's scale of correlation as Low Correlation which means that there is a low correlation between trip distance and trip duration. On the other hand, the value of  $r$  for the variables trip distance and fare amount resulted to  $r = 0.92$  which can be interpreted as Very High Correlation when basing from the Pearson's scale of correlation which means that there is a very high correlation between trip distance and fare amount.
3. HDBSCAN's output concluded that Manhattan, LaGuardia Airport and John F. Kennedy International Airport are the hotspot regions with the highest number of pickups and dropoffs. Furthermore, most of the trips are between Manhattan and the two Airports, John F. Kennedy International Airport and LaGuardia Airport and then there are few concentrations of trips surrounding Manhattan such as Brooklyn and Queens.
4. HDBSCAN outperforms DBSCAN in two major aspects, namely, efficiency and accuracy. DBSCAN incorrectly clusters points when dealing with large datasets





whereas HDBSCAN can correctly cluster points regardless of how large the dataset is. Moreover, DBSCAN is having a hard time finding dense regions in large datasets when compared to HDBSCAN.

### **Recommendation**

The recommendations given as a result of the in-depth study and comparative analysis of taxi pick-up and drop-off in New York City using the HDBSCAN and DBSCAN algorithms are the following for future research:

1. It is recommended to explore and compare various machine learning algorithms, including FDBSCAN, in terms of their speed and efficiency compared to DBSCAN.
2. It is advisable to use diverse dates other than just September 2015 to enhance the variety of taxi trip data analyzed.
3. Furthermore, incorporating additional variables from the dataset such as tolls amount and surcharges could provide valuable insights for future studies.
4. It is suggested to extend the analysis beyond yellow taxis in New York City and include other taxi types such as green taxis.
5. Investigating different locations apart from New York City would enable a comparison of the most traffic-congested areas worldwide, thereby contributing to a more comprehensive analysis of taxi operations.





## References

- Ahmad, H. P., & Dang, S. (2015). Performance Evaluation of Clustering Algorithm Using different dataset. *International Journal of Advance Research in Computer Science and Management Studies*, 8.
- Azar, O. H. (2020). The economics of tipping. *Journal of Economic Perspectives*, 34(2), 215-236.
- Aziz, Z., & Robila, S. (2019, May). Interface for Querying and Data Mining for NYC Yellow and Green Taxi Trip Data. In *2019 IEEE Long Island Systems, Applications and Technology Conference (LISAT)* (pp. 1-7). IEEE.
- Baghestani, A., Tayarani, M., Allahviranloo, M., Gao, H. (2020, May 1). Evaluating the Traffic and Emissions Impacts of Congestion Pricing in New York City. *Advanced Travel Demand Modelling for Sustainable Transportation*. IEEE.
- Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological reports*, 19(1), 3-11.
- Bass, F. (2021, October 22). *Population Density, 2020*. Social Explorer. <https://www.socialexplorer.com/blog/post/population-density-2020-11910>
- Berechman, J., & Paaswell, R. E. (2005). Evaluation, prioritization and selection of transportation investment projects in New York City. *Transportation*, 32, 223-249.
- Berg, M., Gunawan, A., & Roeloffzen, M. (2019). Faster dbscan and hdbscan in low-dimensional euclidean spaces. *International Journal of Computational Geometry & Applications*, 29(01), 21-47.
- Bischoff, J., Maciejewski, M., & Sohr, A. (2015, June). Analysis of Berlin's taxi services by exploring GPS traces. In *2015 International conference on models and technologies for intelligent transportation systems (MT-ITS)* (pp. 209-215). IEEE.
- Blanco-Portals, J., Peiró, F., & Estradé, S. (2022). Strategies for EELS data analysis. Introducing UMAP and HDBSCAN for dimensionality reduction and clustering. *Microscopy and Microanalysis*, 28(1), 109-122.
- Bose, B. (n.d.). *What Is Data Mining: Definition, Purpose, And Techniques*. Digital Vidya. <https://www.digitalvidya.com/blog/what-is-data-mining/>





- Campello, R. J., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining: 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14-17, 2013, Proceedings, Part II* 17 (pp. 160-172). Springer Berlin Heidelberg.
- Çelik, M., Dadaşer-Çelik, F., & Dokuz, A. Ş. (2011, June). Anomaly detection in temperature data using DBSCAN algorithm. In *2011 international symposium on innovations in intelligent systems and applications* (pp. 91-95). IEEE.
- Chen, M. S., Han, J., & Yu, P. S. (1996). Data mining: an overview from a database perspective. *IEEE Transactions on Knowledge and data Engineering*, 8(6), 866-883.
- Deng, D. (2020, September). DBSCAN clustering algorithm based on density. In *2020 7th international forum on electrical engineering and automation (IFEAA)* (pp. 949-953). IEEE.
- Deng, Y., Li, M., Tang, Q., He, R., & Hu, X. (2020). Heterogenous trip distance-based route choice behavior analysis using real-world large-scale taxi trajectory data. *Journal of Advanced Transportation*, 2020, 1-16.
- Dorfer, T. A. (2022, December 6). *Density-Based Clustering: DBSCAN vs. HDBSCAN*. Towards Data Science. <https://towardsdatascience.com/density-based-clustering-dbscan-vs-hdbscan-39e02af990c7>
- Faial, D., Bernardini, F., Meza, E. M., Miranda, L., & Viterbo, J. (2020, July). A methodology for taxi demand prediction using stream learning. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)* (pp. 417-422). IEEE.
- Fuchs, M. (2020, June 20). *HDBSCAN*. <https://michael-fuchs-python.netlify.app/2020/06/20/hdbscan/>
- Gong, Y., Fang, B., Zhang, S., Zhang, J., Chugani, V., Zhang, V., Lee, P., Berke, A., Meier, B., & Villar, I. (2016, September 21). *Data Study to Predict New York City Taxi Demand*. <https://nycdatascience.com/blog/student-works/data-study-to-predict-new-york-city-taxi-demand/>
- Grynbaum, M. M. (2009). *New York's Cabbies Like Credit Cards? Go Figure*. The New York Times. <https://www.nytimes.com/2009/11/08/nyregion/08taxi.html>





- Gunawan, A., & de Berg, M. (2013). A faster algorithm for DBSCAN. *Master's thesis*.
- Hand, D. J. (2007). Principles of data mining. *Drug safety*, 30, 621-622.
- El Harrab, M. S. (2018). *Cost Estimation of FttH Deployment with HDBSCAN Clustering* (No. hal-03746138).
- Hong, F., Zhou, H., Zhu, X., Li, H., & Liu, Z. (2021). Lidar-based panoptic segmentation via dynamic shifting network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 13090-13099).
- Ji, Y., Mishalani, R. G., & McCord, M. R. (2015). Transit passenger origin–destination flow estimation: Efficiently combining onboard survey and large automatic passenger count datasets. *Transportation Research Part C: Emerging Technologies*, 58, 178-192.
- Kamga, C., Yazici, M. A., & Singhal, A. (2013, January). Hailing in the rain: Temporal and weather-related variations in taxi ridership and taxi demand-supply equilibrium. In *Transportation research board 92nd annual meeting* (Vol. 1).
- Kenton, W. (2022, May 06). *What Is the Pearson Coefficient? Definition, Benefits, and History*. Investopedia.  
<https://www.investopedia.com/terms/p/pearsoncoefficient.asp>
- Khan, K., Rehman, S. U., Aziz, K., Fong, S., & Sarasvady, S. (2014, February). DBSCAN: Past, present and future. In *The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014)* (pp. 232-238). IEEE.
- Killip, S., Mahfoud, Z., & Pearce, K. (2004). What is an intracluster correlation coefficient? Crucial concepts for primary care researchers. *The Annals of Family Medicine*, 2(3), 204-208.
- King, D. A., & Saldarriaga, J. F. (2017). Access to taxicabs for unbanked households: An exploratory analysis in New York City. *Journal of Public Transportation*, 20(1), 1-19.
- Kumar, K. M., & Reddy, A. R. M. (2016). A fast DBSCAN clustering algorithm by accelerating neighbor searching using Groups method. *Pattern Recognition*, 58, 39-48.





- Land, O. (2022, November 16). *Here's how much taxi fares are going to go up in NYC*. New York Post. [https://nypost.com/2022/11/16/heres-how-much-taxi-fares-are-going-to-go-up-in-nyc/?fbclid=IwAR0XvHcblAvfZAGWBuWvCyytgLnK-9XFF\\_h0Qv7Xmewd59S\\_Ji-Pn7z-1gY](https://nypost.com/2022/11/16/heres-how-much-taxi-fares-are-going-to-go-up-in-nyc/?fbclid=IwAR0XvHcblAvfZAGWBuWvCyytgLnK-9XFF_h0Qv7Xmewd59S_Ji-Pn7z-1gY)
- Li, W. (2018, May 14). *Tools for understanding taxicab and e-hail services use in New York City*. Smith College. <https://scholarworks.smith.edu/theses/2032/>
- Liu, J., Sun, S., & Chen, C. (2021, October). Big data Analysis of Regional Meteorological Observation Based: On Hierarchical Density Clustering Algorithm HDBSCAN. In *2021 2nd International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT)* (pp. 111-116). IEEE.
- Liu, P., Yao, H., Dai, H., & Fu, W. (2022, March). The Detection and Following of Human Legs Based on Feature Optimized HDBSCAN for Mobile Robot. In *Journal of Physics: Conference Series* (Vol. 2216, No. 1, p. 012009). IOP Publishing.
- Liu, Z., Xia, X., Zhang, H., & Xie, Z. (2021, May). Analyze the impact of the epidemic on New York taxis by machine learning algorithms and recommendations for optimal prediction algorithms. In *2021 the 3rd International Conference on Robotics Systems and Automation Engineering (RSAE)* (pp. 46-52).
- Lynn, M., Zinkhan, G. M., & Harris, J. (1993). Consumer tipping: A cross-country study. *Journal of Consumer Research*, 20(3), 478-488.
- McCulley, K. (n.d.). *The Very Best Time to Visit New York (by a local!)*. Adventurous Kate. <https://www.adventurouskate.com/best-time-to-visit-new-york/>
- Morris, E. A., & Guerra, E. (2015). Are we there yet? Trip duration and mood during travel. *Transportation research part F: traffic psychology and behaviour*, 33, 38-47.
- Patel, M., & Patel, N. (2019). Exploring Research Methodology. *International Journal of Research and Review*, 6(3), 48-55.
- Patel, U., & Chandan, A. (2015, October). *NYC Taxi Trip and Fare Data Analytics using BigData*. ResearchGate. [https://www.researchgate.net/profile/Umang-Patel/publication/287205718\\_NYC\\_Taxi\\_Trip\\_and\\_Fare\\_Data\\_Analytics\\_using\\_BigData/links/567318ab08ae1557cf49472f/NYC-Taxi-Trip-and-Fare-Data-Analytics-using-BigData.pdf](https://www.researchgate.net/profile/Umang-Patel/publication/287205718_NYC_Taxi_Trip_and_Fare_Data_Analytics_using_BigData/links/567318ab08ae1557cf49472f/NYC-Taxi-Trip-and-Fare-Data-Analytics-using-BigData.pdf)







- Poongodi, M., Malviya, M., Kumar, C., Hamdi, M., Vijayakumar, V., Nebhen, J., & Alyamani, H. (2022). New York City taxi trip duration prediction using MLP and XGBoost. *International Journal of System Assurance Engineering and Management*, 1-12.
- Pu, W. (2011). Analytic relationships between travel time reliability measures. *Transportation Research Record*, 2254(1), 122-130.
- Ramani, S., Ghiya, A., Aravind, P. S., Karuppiah, M., & Pelusi, D. (2022). Predicting New York Taxi Trip Duration Based on Regression Analysis Using ML and Time Series Forecasting Using DL. In *Intelligent Sustainable Systems: Proceedings of ICISS 2022* (pp. 15-28). Singapore: Springer Nature Singapore.
- Ratner, B. (2009). The correlation coefficient: Its values range between+ 1/- 1, or do they?. *Journal of targeting, measurement and analysis for marketing*, 17(2), 139-142.
- Schaller, B. (1999). Elasticities for taxicab fares and service availability. *Transportation*, 26(3), 283-297.
- Schoen, J. (n.d.). *Data Interpretation*. University of Massachusetts. <https://www.umass.edu/mwwp/pdf/intmanl.pdf>
- Selvaraj, N. (2020, October 15). *A Beginner's Guide to Data Analysis in Python*. Towards Data Science. <https://towardsdatascience.com/a-beginners-guide-to-data-analysis-in-python-188706df5447>
- Gowri Shankar, V. (2016). Chemical Process Data Classification and Visualization for Process Monitoring. Digital Commons. 10.31390/gradschool\_theses.4472
- Stewart, G., & Al-Khassaweneh, M. (2022). An implementation of the HDBSCAN\* clustering algorithm. *Applied Sciences*, 12(5), 2405. <http://dx.doi.org/10.3390/app12052405>
- Stoyanovich, J., Gilbride, M., & Moffitt, V. Z. (2017). Zooming in on NYC taxi data with Portal. *arXiv preprint arXiv:1709.06176*.
- Taylor, M. (2022, July 5). *Cost of Living in New York City*. [https://www.bankrate.com/real-estate/cost-of-living/nyc/?fbclid=IwAR3ALd-xs4tw7Ni4mIuKfWcuoGiqpQPSJU\\_FYs9HjChnsxUNVF1gpCfIMR0](https://www.bankrate.com/real-estate/cost-of-living/nyc/?fbclid=IwAR3ALd-xs4tw7Ni4mIuKfWcuoGiqpQPSJU_FYs9HjChnsxUNVF1gpCfIMR0)





- Taylor, R. (1990). Interpretation of the correlation coefficient: a basic review. *Journal of diagnostic medical sonography*, 6(1), 35-39.
- Tedesco, L. (2023, January 20). *Everything You Need To Know About NYC's Taxis*. <https://www.thetravel.com/what-to-know-about-taxi-cabs-in-nyc-costs/>
- Tran, T. H., Cao, T. D., & Tran, T. T. H. (2021). HDBSCAN: Evaluating the Performance of Hierarchical Clustering for Big Data. In *Soft Computing: Biomedical and Related Applications* (pp. 273-283). Cham: Springer International Publishing.
- Tran, T. N., Drab, K., & Daszykowski, M. (2013). Revised DBSCAN algorithm to cluster data with dense adjacent clusters. *Chemometrics and Intelligent Laboratory Systems*, 120, 92-96.
- Vijayan, D., & Aziz, I. (2022). Adaptive Hierarchical Density-Based Spatial Clustering Algorithm for Streaming Applications. *Telecom*, 4(1), 1–14. MDPI AG. <http://dx.doi.org/10.3390/telecom4010001>
- Wahyuni, N. A., Hayati, M. N., & Rizki, N. A. (2021). Metode Hierarchical Density-Based Spatial Clustering of Application with Noise (HDBSCAN) Pada Wilayah Desa/Kelurahan Tertinggal di Kabupaten Kutai Kartanegara. *EKSPONENSIAL*, 12(1), 47-52.
- Wang, D., Huang, Y., & Cai, Z. (2022). A two-phase clustering approach for traffic accident black spots identification: integrated GIS-based processing and HDBSCAN model. *International journal of injury control and safety promotion*, 1-12.
- Wickramasinghe, C. S., Marino, D., Yucel, F., Bulut, E., & Manic, M. (2019, June). Data driven hourly taxi drop-offs prediction using tlc trip record data. In *2019 12th International Conference on Human System Interaction (HSI)* (pp. 168-173). IEEE.
- Yang, C., & Gonzales, E. J. (2017). Modeling taxi demand and supply in New York city using large-scale taxi GPS data. *Seeing Cities Through Big Data: Research, Methods and Applications in Urban Informatics*, 405-425.
- Zhang, X., Gong, L., & Ge, O. (2020). *Deep Analysis of NYC/Limo System*. <http://csis.pace.edu/~scha/MID2020/Abstract/a17.pdf>







## Appendix A

### Curriculum Vitae



**Age:** 22

**Address:** North Poblacion, Bacong, Negros Oriental

**Contact Number:** 09776778182

**Email Address:** brent.baylon29@gmail.com

---

### EDUCATIONAL ATTAINMENT

<b>College (2019 – Present)</b>	Bachelor of Science in Computer Science Negros Oriental State University Kagawasan Avenue, Dumaguete City, Negros Oriental
<b>Senior High School (2017-2019)</b>	Dauin Science High School Dauin, Negros Oriental
<b>Junior High School (2013-2017)</b>	Saint Louis College Cebu Maguikay, Mandaue City, Cebu
<b>Elementary (2007-2013)</b>	Tabok II Elementary School Tabok II, Mandaue City, Cebu





**Age:** 22

**Address:** Purok Sampaguita, Napolan, Pagadian City, Zamboanga del Sur

**Contact Number:** 09760544453

**Email Address:** devongladquirante17@gmail.com

---

## EDUCATIONAL ATTAINMENT

<b>College (2019-Present)</b>	Bachelor of Science in Computer Science Negros Oriental State University Kagawasan Avenue, Dumaguete City, Negros Oriental
<b>Senior High School (2017-2019)</b>	Saint Columban College Pagadian City, Zamboanga del Sur
<b>Junior High School (2013-2017)</b>	Josefina Herera Cerilles State College Caridad, Dumingag, Zamboanga del Sur
<b>Elementary (2007-2013)</b>	Kabasalan Special Education School Poblacion, Kabasalan, Zamboanga Sibugay





**NEGROS ORIENTAL STATE UNIVERSITY**  
**College of Arts and Sciences**  
**Computer Science and Information Technology Department**  
**Main Campus I and II, Dumaguete City**



## Appendix B

### Dataset

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	VendorID	trip_pickup_datetime	trip_dropoff_datetime	passenger	trip_distar	pickup_lat	pickup_lat	Ratecode	store_and	dropoff_lat	dropoff_lat	payment_fare_amo	extra	mta_tax	tip_amo	tolls_amo	improvement	total_amount	
2	2	01/09/2015 0:05	01/09/2015 0:31	1	17.45	-73.791	40.6467	1 N	-73.857	40.8483	1	47.5	0.5	0.5	5	5.54	0.3	59.34	
3	1	01/09/2015 0:05	01/09/2015 0:07	1	0.4	-73.979	40.7529	1 N	-73.986	40.7554	2	3.5	0.5	0.5	0	0	0.3	4.8	
4	1	01/09/2015 0:05	01/09/2015 0:16	1	1.5	-73.991	40.724	1 N	-74.01	40.7289	2	9	0.5	0.5	0	0	0.3	10.3	
5	1	01/09/2015 0:05	01/09/2015 0:05	1	0	-73.933	40.8038	1 N	0	0	2	4	0.5	0.5	0	0	0.3	5.3	
6	1	01/09/2015 0:05	01/09/2015 0:30	1	7.5	-73.988	40.7382	1 N	-73.945	40.8282	1	23.5	0.5	0.5	4.95	0	0.3	29.75	
7	2	01/09/2015 0:05	01/09/2015 0:12	3	1.61	-73.99	40.7373	1 N	-73.995	40.7449	2	7	0.5	0.5	0	0	0.3	8.3	
8	2	01/09/2015 0:05	01/09/2015 0:08	2	0.69	-73.944	40.8206	1 N	-73.938	40.8287	2	4.5	0.5	0.5	0	0	0.3	5.8	
9	1	01/09/2015 0:05	01/09/2015 0:10	1	0.8	-74.004	40.7518	1 N	-74.004	40.742	1	5.5	0.5	0.5	1	0	0.3	7.8	
10	1	01/09/2015 0:05	01/09/2015 0:32	3	5.8	-73.987	40.7665	1 N	-73.911	40.7711	1	22	0.5	0.5	4.65	0	0.3	27.95	
11	1	01/09/2015 0:05	01/09/2015 0:19	1	3.9	-73.997	40.7253	1 N	-73.952	40.6928	1	14	0.5	0.5	3.06	0	0.3	18.36	
12	2	01/09/2015 0:05	01/09/2015 0:22	1	4.21	-73.987	40.7301	1 N	-73.981	40.6755	2	15.5	0.5	0.5	0	0	0.3	16.8	
13	1	01/09/2015 0:05	01/09/2015 0:22	3	3.7	-73.994	40.7323	1 N	-73.981	40.7686	2	15	0.5	0.5	0	0	0.3	16.3	
14	2	01/09/2015 0:06	01/09/2015 0:08	1	0.84	-73.971	40.7553	1 N	-73.963	40.7659	1	4.5	0.5	0.5	1.2	0	0.3	7	
15	2	01/09/2015 0:06	01/09/2015 0:15	1	1.21	-73.986	40.7306	1 N	-74.003	40.7309	1	7.5	0.5	0.5	1.76	0	0.3	10.56	
16	1	01/09/2015 0:06	01/09/2015 0:10	2	2	-73.982	40.7746	1 N	-73.968	40.8005	2	7.5	0.5	0.5	0	0	0.3	8.8	
17	1	01/09/2015 0:06	01/09/2015 0:15	1	4.3	-73.863	40.7691	1 N	-73.925	40.7594	1	14.5	0.5	0.5	2	0	0.3	17.8	
18	2	01/09/2015 0:06	01/09/2015 0:23	1	5.14	-74.008	40.7229	1 N	-73.983	40.782	1	17.5	0.5	0.5	1.8	0	0.3	20.6	
19	2	01/09/2015 0:06	01/09/2015 0:15	1	2.44	-73.982	40.7782	1 N	-73.967	40.7527	1	10	0.5	0.5	0.7	0	0.3	12	
20	2	01/09/2015 0:06	01/09/2015 0:51	1	11.15	-73.994	40.7615	1 N	-73.966	40.6346	1	38.5	0.5	0.5	7.96	0	0.3	47.76	
21	2	01/09/2015 0:06	01/09/2015 0:11	1	1.47	-73.981	40.7823	1 N	-73.967	40.7985	1	7	0.5	0.5	0.7	0	0.3	9	
22	1	01/09/2015 0:06	01/09/2015 0:19	1	3.8	-73.975	40.7643	1 N	-73.935	40.7984	1	13	0.5	0.5	2.85	0	0.3	17.15	
23	2	01/09/2015 0:06	01/09/2015 0:37	1	11.58	-73.875	40.7742	1 N	-73.998	40.7319	1	35	0.5	0.5	8.37	5.54	0.3	50.21	
24	1	01/09/2015 0:06	01/09/2015 0:16	1	1.9	-73.987	40.7486	1 N	-73.984	40.7672	2	9.5	0.5	0.5	0	0	0.3	10.8	
25	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
25	1	01/09/2015 0:06	01/09/2015 0:22	1	9.2	-73.982	40.7653	1 N	-73.919	40.8673	2	27	0.5	0.5	0	0	0.3	28.3	
26	2	01/09/2015 0:06	01/09/2015 0:11	1	1.32	-73.998	40.7218	1 N	-74.004	40.7072	1	6.5	0.5	0.5	1.56	0	0.3	9.36	
27	2	01/09/2015 0:06	01/09/2015 0:13	5	1.92	-73.978	40.7731	1 N	-73.976	40.7938	1	8.5	0.5	0.5	1	0	0.3	10.8	
28	2	01/09/2015 0:06	01/09/2015 0:15	1	1.25	-73.973	40.794	1 N	-73.985	40.7825	2	8	0.5	0.5	0	0	0.3	9.3	
29	2	01/09/2015 0:06	01/09/2015 0:21	1	7.54	-73.973	40.7535	1 N	-73.998	40.69	2	22.5	0.5	0.5	0	0	0.3	23.8	
30	2	01/09/2015 0:06	01/09/2015 0:23	1	4.45	-74.01	40.7123	1 N	-73.992	40.7683	1	16.5	0.5	0.5	0	0	0.3	17.8	
31	1	01/09/2015 0:06	01/09/2015 0:07	1	0.3	-73.995	40.7422	1 N	-73.999	40.7395	1	3.5	0.5	0.5	7.2	0	0.3	12	
32	1	01/09/2015 0:06	01/09/2015 0:11	1	1	-73.998	40.7208	1 N	-73.99	40.7347	2	6	0.5	0.5	0	0	0.3	7.3	
33	2	01/09/2015 0:06	01/09/2015 0:12	6	0.98	-73.992	40.7599	1 N	-73.978	40.7518	2	6.5	0.5	0.5	0	0	0.3	7.8	
34	2	01/09/2015 0:06	01/09/2015 0:19	2	4.94	-73.874	40.7741	1 N	-73.924	40.7423	1	16.5	0.5	0.5	5.34	0	0.3	23.14	
35	1	01/09/2015 0:06	01/09/2015 0:16	1	2.5	-73.979	40.7533	1 N	-73.976	40.7287	1	10	0.5	0.5	2.8	0	0.3	14.1	
36	2	01/09/2015 0:06	01/09/2015 0:12	1	1.28	0	0	1 N	0	0	2	6	0.5	0.5	0	0	0.3	7.3	
37	2	01/09/2015 0:06	01/09/2015 0:25	1	5.69	-73.988	40.738	1 N	-73.956	40.6851	1	19	0.5	0.5	3.5	0	0.3	23.8	
38	1	01/09/2015 0:06	01/09/2015 0:16	2	3.8	-73.988	40.7203	1 N	-73.962	40.6825	1	13	0.5	0.5	2.85	0	0.3	17.15	
39	2	01/09/2015 0:06	01/09/2015 0:18	1	2.86	-74.001	40.731	1 N	-74	40.7619	2	11.5	0.5	0.5	0	0	0.3	12.8	
40	2	01/09/2015 0:06	01/09/2015 0:16	1	2.94	-74.012	40.7025	1 N	-73.984	40.6938	1	11.5	0.5	0.5	0	0	0.3	12.8	
41	1	01/09/2015 0:06	01/09/2015 0:09	1	0.9	-73.969	40.7967	1 N	-73.966	40.8056	1	4.5	0.5	0.5	1.15	0	0.3	6.95	
42	1	01/09/2015 0:06	01/09/2015 0:17	2	2.6	-73.961	40.7142	1 N	-73.927	40.699	1	11	0.5	0.5	2	0	0.3	14.3	
43	1	01/09/2015 0:06	01/09/2015 0:32	1	9.6	-73.864	40.7696	1 N	-73.907	40.6971	1	30	0.5	0.5	6.25	0	0.3	37.55	
44	2	01/09/2015 0:06	01/09/2015 0:38	1	7.64	-73.977	40.7849	1 N	-73.927	40.867	1	29	0.5	0.5	6.06	0	0.3	36.36	
45	2	01/09/2015 0:06	01/09/2015 0:12	4	1.25	-73.984	40.7554	1 N	-73.97	40.7631	1	6.5	0.5	0.5	0	0	0.3	7.8	
46	2	01/09/2015 0:06	01/09/2015 0:12	2	1.51	-74.01	40.7203	1 N	-74.006	40.7376	1	7	0.5	0.5	1.66	0	0.3	9.96	
47	2	01/09/2015 0:06	01/09/2015 0:12	6	1.07	-73.995	40.7252	1 N	-74.007	40.7329	1	6.5	0.5	0.5	1.56	0	0.3	9.36	
48	2	01/09/2015 0:06	01/09/2015 0:14	3	2.36	-73.974	40.7837	1 N	-73.959	40.8095	1	9	0.5	0.5	2.06	0	0.3	12.36	

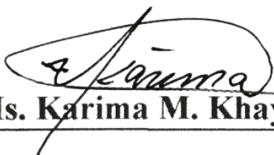




## Appendix C

### STATISTICIAN'S CERTIFICATION

This is to certify that this research study entitled, “**Comparative Analysis on HDBSCAN and DBSCAN Algorithms using Taxi Pickup and Dropoff in New York City Dataset**” prepared and submitted by Brent V. Baylon and Devon Glad L. Quirante in partial fulfillment for the Degree of Bachelor of Science in Computer Science, has been statistically reviewed by the undersigned.



Ms. Karima M. Khayat

*Statistician*

Date Signed: May 26, 2023





## Appendix D

### Grammarly and Turnitin Results

#### E1. Turnitin Result

The screenshot shows the Turnitin Match Overview interface. The main text area displays a document titled "in New York City Datasets" with several highlighted matches. The right sidebar shows a list of matches with their respective similarity percentages. The overall match percentage is 15%.

Match Number	Source	Similarity Percentage
1	Submitted to Negros O... Student Paper	9%
2	www.mdpi.com Internet Source	1%
3	academic.oup.com Internet Source	1%
4	www.researchgate.net Internet Source	<1%
5	Submitted to Internatio... Student Paper	<1%
6	Penghua Liu, Hanchen ... Publication	<1%

#### E2. Grammarly Result

The screenshot shows the Grammarly interface for a document titled "YAP-thesis-the-last-of-us". The document text is displayed on the left, and the right sidebar shows the Grammarly score and various metrics.

Document Text:

NEGROS ORIENTAL STATE UNIVERSITY  
College of Arts and Sciences  
Computer Science and Information Technology Department  
Main Campus I and II, Dumaguete City  
|  
Bachelor of Science in Computer Science  
2

NEGROS ORIENTAL STATE UNIVERSITY  
College of Arts and Sciences  
Computer Science and Information Technology Department  
Main Campus I and II, Dumaguete City

Grammarly Score: 96 (Overall score)

Correctness: 7 alerts

Clarity: Clear

Engagement: Engaging

Plagiarism: 0%

