Literature Review

1. Introduction

In the ever-evolving landscape of the modern sharing economy, ride-sharing platforms have emerged as prominent players in the urban transportation paradigm. Services like Uber and Lyft have not only transformed the way people move from one point to another but also have created unique income-generating opportunities for countless individuals across the globe. At the heart of this transformation lies the intricate dance of supply and demand, where drivers aim to be in the right place at the right time to maximize their earnings while passengers seek timely, convenient, and cost-effective transportation. In this dynamic ecosystem, an emerging field of research and practice has arisen, one that focuses on utilizing data analytics techniques to optimize the positioning of ride-sharing drivers during periods of low demand. The objective of this project is to contribute to this growing body of knowledge by exploring the strategic positioning of drivers during off-peak hours. By identifying optimal locations for drivers during these intervals, we aim to provide a data-driven solution that enhances the probability of drivers receiving passenger requests within the shortest amount of time possible.

The implications of this endeavor are multifaceted, touching upon the livelihood of drivers, the efficiency and competitiveness of ride-sharing platforms, and the overall experience of passengers. Through our literature review, we embark on an exploratory journey into existing research and practices that have sought to address this fundamental challenge. This review serves as a critical foundation for our project, as it not only highlights the current state of knowledge in this field but also informs our research methodology, strategies, and potential avenues for innovation. In the following sections, we delve into the rich tapestry of work reports that have explored the optimization of ride-sharing driver positioning, encompassing a wide spectrum of data analytics, algorithms, approaches and designs. By synthesizing the insights from previous works, we aim to discern the prevailing trends, challenges, and gaps in the papers, thus laying the groundwork for our own contributions to the field. In doing so, we aspire to enhance the earnings and service efficiency of ride-sharing drivers while enriching the experience of passengers, thereby advancing the broader sharing economy discourse and offering practical recommendations for drivers in this dynamic industry.

2. Approach

The paper on "**Real-Time Uber Data Analysis of Popular Uber Locations in Kubernetes Environment**" presents a wealth of insights that can greatly benefit our project aimed at identifying optimal locations for ride-sharing drivers during periods of low demand. By delving into Uber's data analytics techniques, we can gain a deeper understanding of how to harness data effectively for decision-making. Uber's emphasis on real-time data analysis aligns with our project's goal of making timely decisions about driver positioning, especially during periods of low demand. Their use of technologies like Kubernetes and distributed systems provides valuable lessons in optimizing the performance of data processing, which can be directly applicable to our project. Additionally, their implementation of machine learning models, such as K-means clustering, offers a potential avenue to enhance the accuracy of identifying optimal driver locations. The paper's insights into resource optimization, performance metrics, and scalability can guide our project in efficient resource allocation and preparation for future growth.

As for **'An Empirical Data Analytics and Visualization for UBER Services: A Data Analysis Based Web Search Engine'**, it offers valuable insights and methodologies that can significantly benefit our project, which aims to identify optimal locations for ride-sharing drivers during periods of low demand. Leveraging real data from ride-sharing platforms like Uber or Lyft, we can explore the patterns in supply and demand, thus enabling us to identify strategic locations where drivers can position themselves during low-demand periods to maximize their earnings and enhance service efficiency. Additionally, the paper's visualization techniques, including heat maps, bar

graphs, and tables, can be a valuable reference for creating visually compelling representations of demand patterns. This visualization can be a powerful tool for ride-sharing drivers, allowing them to make informed decisions about their positioning in real time, which is crucial for our project's success. Furthermore, the paper introduces the use of R programming and machine learning [Sudar22], providing a solid foundation for statistical analysis and data modeling. These techniques can be instrumental in developing models that predict demand patterns and assist drivers in making data-driven decisions.

The paper on "**Uber Related Data Analysis using Machine Learning"** is aiming to predict cab pickups from a coordinated cluster of points so that we can bridge the supply demand gap of cab services, deduce ETA and optimize the route selection. The k-mean cluster method is a possible solution/tool to use for our project [Srinivas21]. The authors employ k-means to divide the trajectory dataset into clusters based on the longitude, latitude and the frequency of trips. A machine learning model is also trained with this data to predict pickup locations of cabs based on the cluster. Since the goal of this paper is to predict pickup locations, this method will be useful for our project since we are also trying to find a good spot for the driver to wander.

In the paper on "**Identifying the Hotspots in Urban Areas Using Taxi GPS Trajectories"**, Bai Feng and Xu are trying to identify hotspots areas in Lanzhou city using taxi data [Bai18]. Urban hotspots are areas with high levels of human activities, where pick-up cab service demands tend to be relatively high. So arranging more cab services in these areas will increase the working efficiency. Similar to our project, finding hotspots in urban areas is the goal. They divided the area into grid cells of the same size and made a directed graph with each grid cell being a node; the weight of each weighted edge between each node is determined by the amount of trips between them[Bai18]. After that, to find out the hotspots in the city, a method of DWNodeRank is used. By inputting an adjacency matrix based on the directed weighted graph, the output will be a vector that ranks all the nodes[Bai18].

In the paper "**Hunting or waiting? Discovering passenger-finding strategies from a large-scale real-world taxi dataset"** , they designed a set of taxi-patterns, and obtained a rich collection of features. If the distance covered by a taxi between the last dropoff and the next pickup is short, the taxi driver is very likely to have found passengers locally in the region where the last passenger got off. The authors used L1-Norm SVM to select the most discriminative features, a predictor based on the selected patterns, and a classifier [B. Li11] . In addition, the SVM is used to select a small subset of the most salient features of good and ordinary taxis from a collection of taxi-patterns, and finally get the result of predicting the performance of a set of test taxis based on the selected patterns: Time, Location, Strategy [B.Li11].

In the paper "**On recommending urban hotspots to find our next passenger"**, they are trying to find their passengers by using The Use Sliding Window Ensemble Framework, which harnesses the strengths of multiple predictive models by combining them in an integrated framework [Zhao11]. This approach aims to leverage the strengths of each model, potentially leading to more accurate and robust forecasts. This includes the Prediction models, Poisson Model, Weighted Time Varying Poisson Model, and Autoregressive Integrated Moving Average (ARIMA) Model. Afterwards, they designed a recommendation model. The decision is framed as a minimization problem, aiming to rank the stands based on the minimum waiting time to pick up another passenger. It includes the following 3 Steps, Service Deficit Calculation, Normalized Distance Calculation, and finally a recommendation score, calculated with the first two steps [Zhao11]. This is done to get the highest score which will be shown visibly. This paper provides a good example of combining multiple predictive models and also a good algorithm to get recommendation scores.

"A Survey on Trajectory Data Mining: Techniques and Applications" provided various strategies and tools that all pertain to handling and analyzing trajectory data as a whole. In terms of data types, semantic trajectory was of note. Semantic trajectory is trajectory data accompanied with semantic entities, such as location information [Feng16]. In terms of storing data, Feng et al. has introduced us to quite a few useful tools. For example, SharkDB and TrajStore are two storage tools that specialize in trajectory data [Feng16]. TrajStore, more specifically, is a dynamic storage system for specific geo-spatial locations [Feng16]. As opposed to simple index trajectories, TrajStore slices trajectories into sub trajectories based on their region and stores them all together [Feng16]. SharkDB uses column-oriented storage, where trajectories are partitioned into frames and further compressed [Feng16]. Lastly, various queries were mentioned but aggregate queries fit our needs. Aggregate queries search for measurements of datapoint groups, such as average number of members within a specific region.

"Mining Interesting Locations and Travel Sequences from GPS Trajectories" primarily focus on providing a travel recommendation tool. Zheng et al implemented a tree-based hierarchy graph that we believe can be useful [Zheng09]. According to Zeng et al, it is basically two structures, a tree-based hierarchy and tree-based graph [Zheng09]. The hierarchy is a density-based clustering of regions based on proximity. The child clusters are sub-clusters of larger clustered regions. This approach could allow us to find high traffic locations and calculate average prices for specific regions quite quickly. In this paper, the tree-based hierarchical graph is used to connect data points that multiple tourists traveled to [Zheng09]. For us, we could use the tree-based hierarchical graph to make connections between neighbor clusters based on drop off points. By connecting clusters on the same level with edges, we can see instances where the drop off point of one cluster is connected to a pickup point in a close neighbor cluster. This approach will allow us to determine regions where drivers will have the opportunity to continue working in high paying regions throughout their work path with little downtime.

3. Challenges

"Real-Time Uber Data Analysis of Popular Uber Locations in Kubernetes Environment" highlights several challenges that Uber faces in its operations. First and foremost, Uber grapples with the overwhelming volume of data generated daily, which poses significant hurdles in terms of data storage, processing, and analysis. Managing and making sense of this massive data influx is a continuous challenge. Secondly, real-time data analysis presents a critical challenge for Uber. Given the time-sensitive nature of their service, quick decision-making is essential, and efficient data processing in real-time is a demanding task [Gunawardena20]. Moreover, the diverse variety of data sources and formats that Uber encounters presents yet another challenge. Integrating and analyzing data from numerous sources require comprehensive data integration strategies, which can be complex and resource-intensive.

In addition to that, Addressing the concerns of passenger safety and trust is essential but challenging. Sudar et al. highlights some of these issues, especially in the context of ride-sharing services in India [Sudar20]. Ensuring passengers' safety and gaining their trust in the ride-sharing platform is vital to the success of our project. Addressing this challenge involves not only offering data-driven solutions for drivers' positioning but also incorporating features that enhance passengers' sense of security. Successfully navigating these challenges is essential for realizing the full potential of our project and for making ride-sharing services a safer and more attractive mode of transportation for all passengers.

In the paper "**Uber Related Data Analysis using Machine Learning"**, the challenges are how to predict cab pickup location efficiently, handle large datasets and reduce forecasting errors and overfitting when dealing with

extensive trip data [Srinivas21]. For our project, having data visualization before we actually start analyzing the dataset can be helpful to understand the data and try our best to reduce the noise.

In the paper "**Identifying the Hotspots in Urban Areas Using Taxi GPS Trajectories"**, the main challenges are about data cleaning, outlier detection and matching GPS data to road network topology [Bai18]. Our project will rely on a map api, so a question remains if we are able to divide the area into correct pieces that match up to the road network.

In the research paper titled "**Hunting or waiting? Discovering passenger-finding strategies from a large-scale real-world taxi dataset**" the authors adopt a sophisticated supervised feature selection method known as L1-Norm SVM [Li11]. While this method offers robustness and efficiency in the context of feature selection, a significant challenge arises in determining the optimal weight settings for each parameter. Setting these weights appropriately is crucial as they can significantly impact the accuracy and generalizability of the model. Striking the right balance to ensure both model performance and interpretability becomes an intricate task that demands experimentation and validation.

In the paper, "**On recommending urban hotspots to find our next passenger",** the authors grapple with a significant computational challenge stemming from their methodological choice: the Sliding Window Ensemble Framework [Zhao11]. This framework, while potentially powerful in analyzing urban hotspots and passenger behaviors, necessitates substantial computational resources [Zhao11]. The dynamic nature of sliding windows, combined with ensemble techniques, leads to intricate calculations that demand high processing capability. Deploying such a framework in real-time scenarios or on large datasets might pose scalability issues, making it imperative to ensure that the underlying infrastructure is robust enough to handle the computational load.

We believe that heterogeneity won't be the biggest problem as there is a good chance that the data will be from the same source, Uber. We not only need to make sure that users are in a quality location but also one where, in the event they must travel a great distance, that quality doesn't drastically change when they do arrive. Feng et al. touches on this issue with the use of pattern matching[Feng16]. However, there is the issue of how often we would do these operations, impacting performance and memory. Lastly, we need to consider not only factors like surge charge and distance, but also common drop off locations. We want drivers to continue being placed near locations that have quality opportunities to make money and that depends on where trips end. Tree-based Hierarchical Graphs prove useful in solving this problem[Zheng09].

4. Project Proposal

Our project isn't entirely unique as Uber data has been analyzed to find hotspots in the past. Fortunately, their work provides us with a plethora of tools and methods at our disposal. Our papers today showed us methods to not only break up location data, but also how to visualize it as well. For example k-mean data clustering can break down maps into hot spots while heat maps can help visually represent it. However, what makes our project unique is that we are looking at methods to have drivers get quality pickups throughout a shift, not just finding singular hotspot locations. We want drivers to experience few wait times between calls and to get the most money for each consecutive call. Fortunately, there are also methods that have tackled this problem as well, but they arose from works outside of Uber directly. In fact, these methods arose from analyzing taxis and tourist travel data. We hope to use these same methods to further enhance the experience of Uber drivers.

Cited Work

[1] T. M. Gunawardena and K. P. N. Jayasena, "Real-Time Uber Data Analysis of Popular Uber Locations in Kubernetes Environment," 2020 5th International Conference on Information Technology Research (ICITR), Moratuwa, Sri Lanka, 2020, pp. 1-6, doi: 10.1109/ICITR51448.2020.9310851.

[2] K. M. Sudar, P. Nagaraj, V. Muneeswaran, S. K. Jeevana Swetha, K. M. Nikhila and R. Venkatesh, "An Empirical Data Analytics and Visualization for UBER Services: A Data Analysis Based Web Search Engine," 2022 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2022, pp. 1-6, doi: 10.1109/ICCCI54379.2022.9741016.

[3] R. Srinivas, B. Ankayarkanni and R. S. B. Krishna, "Uber Related Data Analysis using Machine Learning," 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2021, pp. 1148-1153, doi: 10.1109/ICICCS51141.2021.9432347.

[4] F. Bai, H. Feng and Y. Xu, "Identifying the Hotspots in Urban Areas Using Taxi GPS Trajectories," 2018 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), Huangshan, China, 2018, pp. 900-904, doi: 10.1109/FSKD.2018.8686932.

[5] B. Li et al., "Hunting or waiting? Discovering passenger-finding strategies from a large-scale real-world taxi dataset," 2011 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops), Seattle, WA, USA, 2011, pp. 63-68, doi: 10.1109/PERCOMW.2011.5766967.

[6] K. Zhao, D. Khryashchev and H. Vo, "Predicting Taxi and Uber Demand in Cities: Approaching the Limit of Predictability," in IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 6, pp. 2723-2736, 1 June 2021, doi: 10.1109/TKDE.2019.2955686.

[7] Z. Feng and Y. Zhu, "A Survey on Trajectory Data Mining: Techniques and Applications," in IEEE Access, vol. 4, pp. 2056-2067, 2016, doi: 10.1109/ACCESS.2016.2553681.

[8] Yu Zheng, Lizhu Zhang, Xing Xie, and Wei-Ying Ma. 2009. Mining interesting locations and travel sequences from GPS trajectories. In Proceedings of the 18th international conference on World wide web (WWW '09). Association for Computing Machinery, New York, NY, USA, 791–800. https://doi.org/10.1145/1526709.1526816

[9] Yu Zheng, Lizhu Zhang, Xing Xie, and Wei-Ying Ma. 2009. Mining interesting locations and travel sequences from GPS trajectories. In Proceedings of the 18th international conference on World wide web (WWW '09). Association for Computing Machinery, New York, NY, USA, 791–800. https://doi.org/10.1145/1526709.1526816