Project Report- CS528 Spring 2024

PRIVACY-PRESERVING ANALYSIS OF BIKE-SHARING NETWORKS FOR COMMUNITY DETECTION

DEPARTMENT OF COMPUTER SCIENCE

Data Privacy and Security - CS528

BY

URVA SURTI (A20505142)

ATHARVA NIRALI (A20517247)

SUMIT SAKARKAR (A20516976)

FROM

ILLINOIS INSTITUTE OF TECHNOLOGY, CHICAGO

Table of Contents

Introduction

In today's bustling cities, bike-sharing networks like Chicago's Divvy system have become vital for offering convenient and eco-friendly transportation options. Yet, while these systems help countless people get around, they also raise concerns about privacy. The data collected by Divvy, including where and when people ride, could reveal a lot about individuals' habits and whereabouts.

Our project tackles this challenge head-on. We want to dig into the wealth of data from Divvy while ensuring people's privacy remains protected. By using smart techniques to keep personal details safe, we aim to learn valuable insights about how different groups use the bike-sharing system.

Our main goal is simple: to understand how communities interact with Divvy bikes, like where they ride most and when. This knowledge can help city planners and others make biking even better for everyone.

This report outlines our journey, explaining how we did our research, what data we used, and how we plan to apply our findings. Through our work, we hope to not only understand bike-sharing better but also set a good example for keeping people's privacy intact in similar studies.

Team Description

Our project team, composed of three members from DPS Spring 2024, each pursuing their master's degrees in computer science, brings a diverse set of skills and expertise to the table:

Urva Surti: With a strong focus on software development, Urva brings expertise in data analysis, particularly in unraveling temporal and spatial patterns within datasets.

Atharva Nirali: Specializing in data privacy, Atharva applies cryptographic schemes to protect sensitive information, ensuring the security of deployed software and safeguarding proprietary data.

Sumit Sakarkar: Contributing knowledge in network analysis and graph-based algorithms, Sumit adds depth to our team's analytical capabilities, allowing for comprehensive data analysis and insights generation.

Together, our team is well-equipped to tackle the challenges of our project, leveraging our combined expertise to drive meaningful analysis while prioritizing privacy and security concerns.

Application

Our project aims to develop a versatile tool tailored for analyzing Divvy's dataset, with a particular focus on practical applications and utility. By offering user-friendly features and actionable insights, our tool can be deployed in various contexts to enhance the efficiency and accessibility of bike-sharing systems and urban transportation infrastructure.

1. Urban Planning and Infrastructure Development: Urban planners and city officials can leverage our tool to identify areas within the Divvy network that experience high demand or congestion. By pinpointing these hotspots, decision-makers can prioritize infrastructure development, such as the installation of additional bike stations or bike lanes, to alleviate congestion and improve accessibility for users.

2. Resource Allocation and Optimization: Transportation authorities can use our tool to optimize resource allocation within the Divvy network. By analyzing usage patterns and user demographics, decision-makers can allocate bikes and other resources more effectively to meet fluctuating demand throughout the day and across different locations. This ensures that resources are distributed efficiently, enhancing the overall user experience.

3. Marketing and User Engagement: Divvy operators can utilize our tool to gain insights into user behavior and preferences, allowing them to tailor marketing campaigns and promotional activities more effectively. By understanding where and when different user groups are most active, operators can target their marketing efforts to specific locations and times, increasing user engagement and participation in the bike-sharing system.

4. Community Engagement and Stakeholder Collaboration: Our tool can facilitate collaboration and engagement between stakeholders, including community groups, businesses, and advocacy organizations. By providing transparent and accessible data on bike-sharing usage patterns, decision-makers can foster dialogue and collaboration with stakeholders to address community needs and priorities related to urban transportation.

5. Research and Policy Development: Researchers and policymakers can leverage our tool to conduct in-depth analyses of bike-sharing usage patterns and trends. By examining spatial and temporal data, researchers can identify correlations and trends that inform the development of evidence-based policies and interventions aimed at promoting sustainable urban transportation and reducing carbon emissions.

Overall, our tool offers a versatile solution for analyzing Divvy's dataset, providing valuable insights that can inform decision-making processes and improve the efficiency, accessibility, and sustainability of bike-sharing systems and urban transportation infrastructure. With its practical applications and user-friendly interface, our tool has the potential to make a significant impact in shaping the future of urban mobility.

Data Sets

The dataset provides comprehensive records of Divvy bike-sharing trips, detailing key aspects such as the starting and ending stations, trip durations, and corresponding timestamps. Additionally, it classifies users into two main groups: subscribers, who are regular users with a membership, and casual riders, who use the service occasionally.

With a rich collection of geospatial and temporal data, the dataset enables thorough analysis of mobility patterns and station usage across various times and locations within the Divvy network. This information is instrumental in understanding how people navigate the city and utilize bike-sharing services.

Overall, the dataset serves as a valuable resource for uncovering insights into urban transportation dynamics and optimizing the efficiency of bike-sharing systems. By harnessing the wealth of information contained within the dataset, stakeholders can make informed decisions to enhance the accessibility and sustainability of urban mobility while respecting user privacy.

Link to Data Sets:

Chicago Data Portal

Divvy - Lyft Data

Data Source

Privacy and Security Technique

In analyzing the Divvy bike-sharing system, our foremost priority is to ensure the privacy and security of user data. This section outlines the advanced privacy-preserving techniques we've implemented to protect user information while enabling us to extract meaningful insights from the data. These techniques are designed to address the dual challenges of achieving high utility from data analytics and adhering to stringent privacy standards.

Differential Privacy

To guarantee that individual users cannot be identified from our datasets, we have employed differential privacy techniques. This involves adding controlled random noise to the results of queries on the dataset, which helps mask the contributions of individual data points. Specifically, we utilized the Laplace mechanism for its simplicity and effectiveness in providing epsilon-differential privacy, where epsilon determines the level of privacy and accuracy of the data:

Laplace Noise Addition: For each aggregate statistic calculated, such as average trip duration or frequency of station usage, Laplace noise calibrated to the sensitivity of the query is added. This ensures that no single individual's data significantly influences the output, thereby protecting their privacy.

Secure Multi-Party Computation (SMPC)

We considered the use of Secure Multi-Party Computation for scenarios where data might be analyzed collaboratively by multiple stakeholders without revealing their individual datasets to one another. While not implemented directly in this project, SMPC represents a forward-thinking approach to privacy-preserving data analysis, particularly useful in multi-organizational studies of transportation data:

Theoretical Application: For community detection and usage pattern analysis, SMPC could allow us to securely aggregate data across different regions without any party having access to the others' raw data.

Homomorphic Encryption

Homomorphic encryption was evaluated for its potential to allow computations to be performed on encrypted data. This method would enable the Divvy system to derive analytical insights without ever accessing plaintext data:

Potential Use Case: Encrypting user trip data while still allowing for operations like counting trips or calculating average speeds, which could be beneficial for real-time traffic management and predictive analytics in smart city applications.

Data Anonymization and Masking

We have implemented data anonymization techniques to remove or mask identifiers that are linked to individual users. This ensures that the data can be used for analysis without risking personal privacy:

K-Anonymity: Ensuring that each person included in the release of data cannot be distinguished from at least k-1 individuals whose information also appears in the release.

Pseudonymization: Replacing private identifiers with artificial identifiers (pseudonyms) to protect the individual's identity while allowing their behavior within the dataset to be analyzed.

Analysis

For our project, we undertake a detailed analysis of various data points within the Divvy dataset to glean insights into bike-sharing usage patterns. Here's a breakdown of our analysis:

1. Bike Type Analysis: We scrutinize the type of bikes used in rides, distinguishing between electric and classic Divvy bikes. By examining the frequency of each bike type's usage, we aim to understand user preferences and trends regarding bike selection.

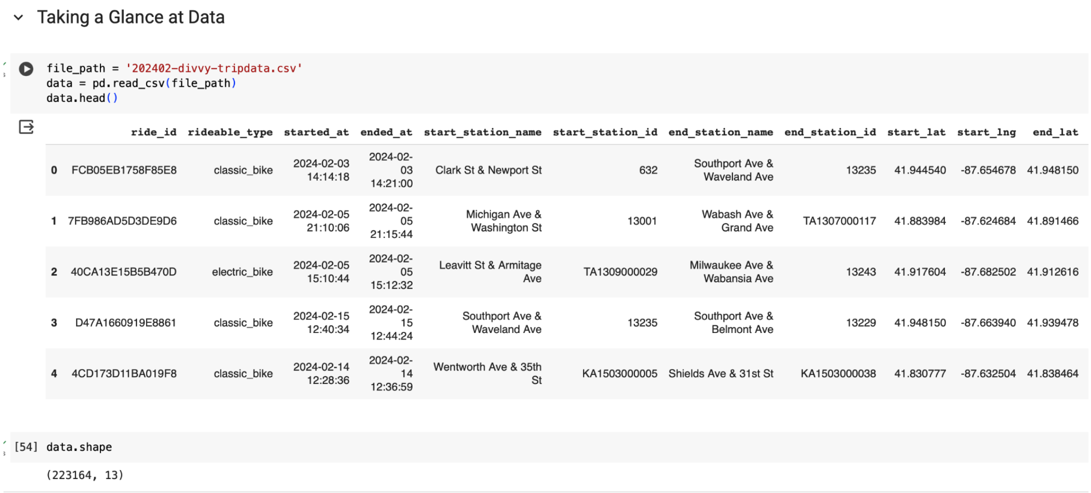
2. User Categorization: We categorize riders based on their membership status, distinguishing between members who subscribe to the service and casual riders who use it on an ad-hoc basis. This segmentation allows us to analyze usage patterns and behaviors among different user groups.

3. Peak Usage Times: We analyze the times of the day when Divvy bikes are most heavily utilized. By identifying peak usage periods, we can discern trends in user behavior and demand, which can inform decisions related to resource allocation and service provision.

4. Station Utilization: We examine the popularity of different stations within the Divvy network, identifying which stations are most frequently used by riders. This analysis sheds light on high-traffic areas and can guide decisions regarding station placement and infrastructure investment.

5. Average Trip Duration: We calculate the average duration of trips taken with Divvy bikes. Understanding typical ride lengths allows us to gauge user behavior and preferences, as well as the overall efficiency of the bike-sharing system.

By conducting a thorough analysis of these key metrics, we aim to provide valuable insights into urban mobility patterns and the effectiveness of the Divvy bike-sharing system. These insights can inform decision-making processes aimed at optimizing service delivery, enhancing user experience, and promoting sustainable transportation options in urban environments.

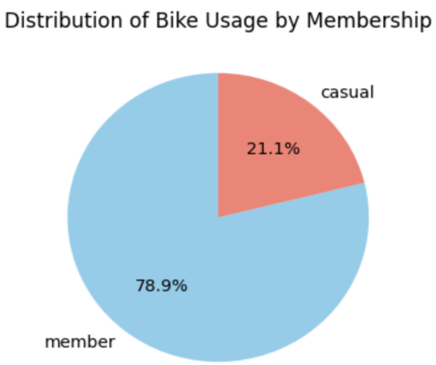


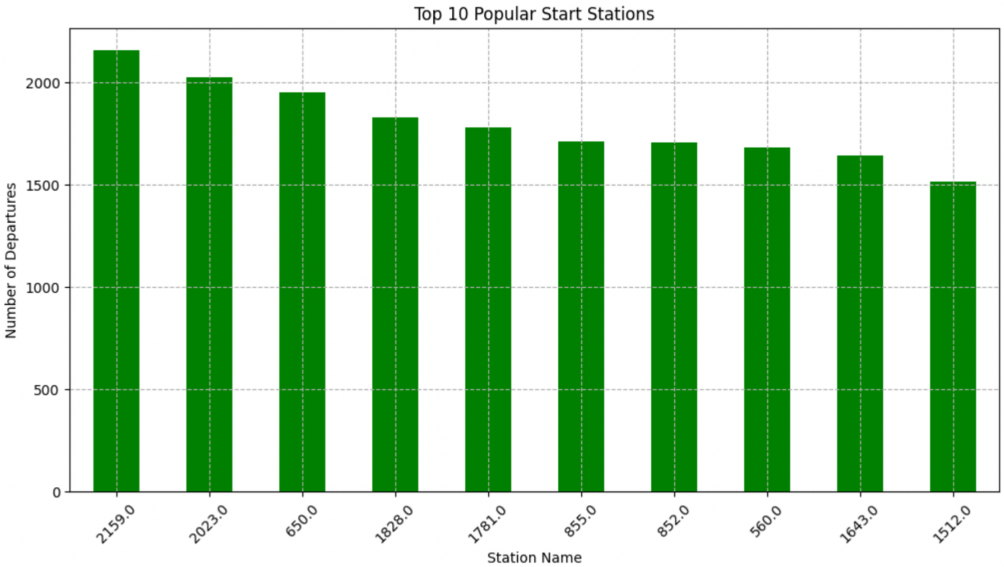
Taking a Glance at Data

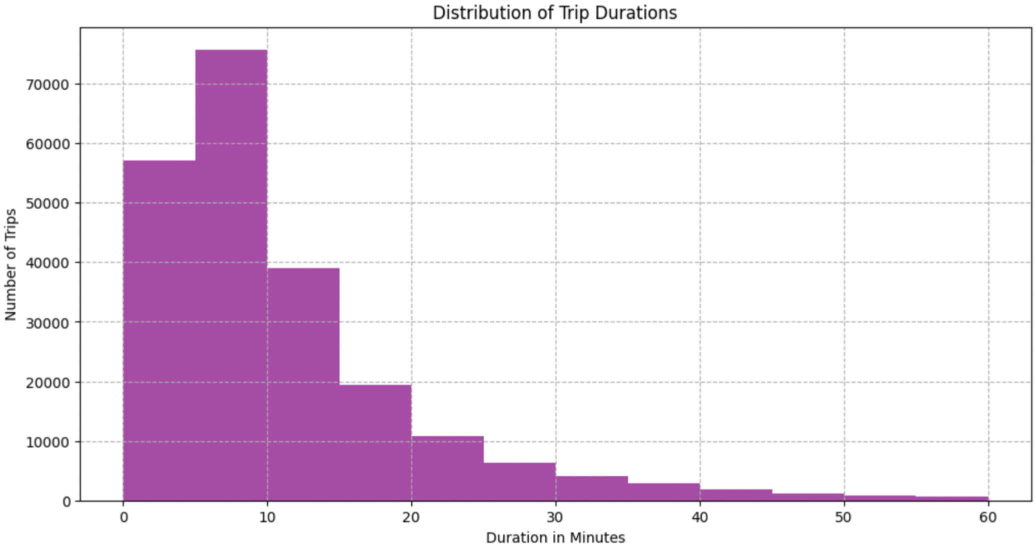
```
file_path = '202402-divvy-tripdata.csv'
data = pd.read_csv(file_path)
data.head()
```

	ride_id	rideable_type	started_at	ended_at	start_station_name	start_station_id	end_station_name	end_station_id	start_lat	start_lng	end_lat
0	FCB05EB1758F85E8	classic_bike	2024-02-03 14:14:18	2024-02-03 14:21:00	Clark St & Newport St	632	Southport Ave & Waveland Ave	13235	41.944540	-87.654678	41.948150
1	7FB986AD5D3DE9D6	classic_bike	2024-02-05 21:10:06	2024-02-05 21:15:44	Michigan Ave & Washington St	13001	Wabash Ave & Grand Ave	TA1307000117	41.883984	-87.624684	41.891466
2	40CA13E15B5B470D	electric_bike	2024-02-05 15:10:44	2024-02-05 15:12:32	Leavitt St & Armitage Ave	TA1309000029	Milwaukee Ave & Wabansia Ave	13243	41.917604	-87.682502	41.912616
3	D47A1660919E8861	classic_bike	2024-02-15 12:40:34	2024-02-15 12:44:24	Southport Ave & Waveland Ave	13235	Southport Ave & Belmont Ave	13229	41.948150	-87.663940	41.939478
4	4CD173D11BA019F8	classic_bike	2024-02-14 12:28:36	2024-02-14 12:36:59	Wentworth Ave & 35th St	KA1503000005	Shields Ave & 31st St	KA1503000038	41.830777	-87.632504	41.838464

```
[54] data.shape

(223164, 13)
```




Top 10 Popular Start Stations
Number of Departures
0
500
1000
1500
2000
2159.0
2023.0
650.0
1828.0
1781.0
855.0
852.0
560.0
1643.0
1512.0
Station Name


Distribution of Trip Durations
Number of Trips
0
10000
20000
30000
40000
50000
60000
70000
0
10
20
30
40
50
60
Duration in Minutes

Code Implementation

For code implementation, we utilize Python as our primary programming language, leveraging the powerful libraries and frameworks available within its ecosystem. Additionally, we make use of Google Colab, a cloud-based platform that provides a convenient environment for running Python code, especially for data analysis and machine learning tasks.

Our code relies on various dependencies, including but not limited to:

- Pandas for data manipulation and analysis

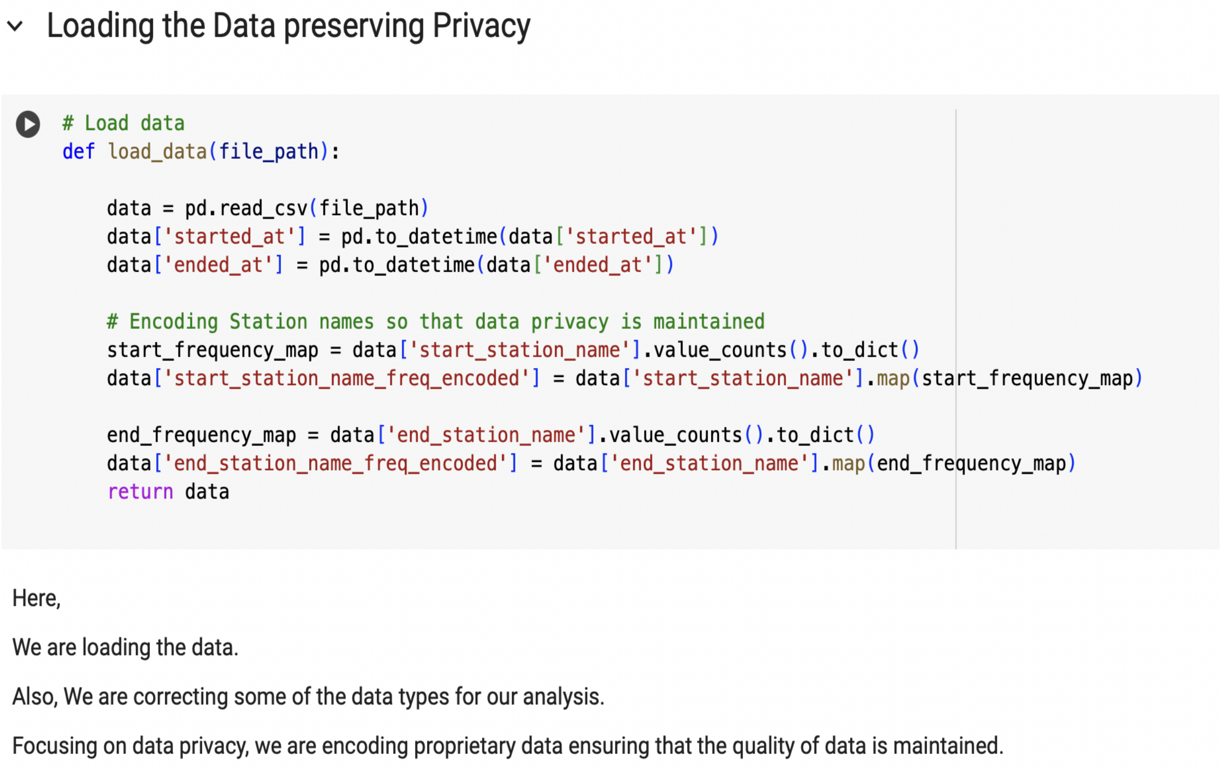
- NumPy for numerical computing

- Matplotlib and Seaborn for data visualization

- Scikit-learn for machine learning algorithms (if applicable)

- TensorFlow or PyTorch for deep learning models (if applicable)

These dependencies enable us to perform a wide range of tasks, from data preprocessing and exploratory analysis to model development and evaluation. By leveraging Python and Google Colab, along with these essential libraries, we ensure a seamless and efficient code implementation process for our project.



Loading the Data preserving Privacy

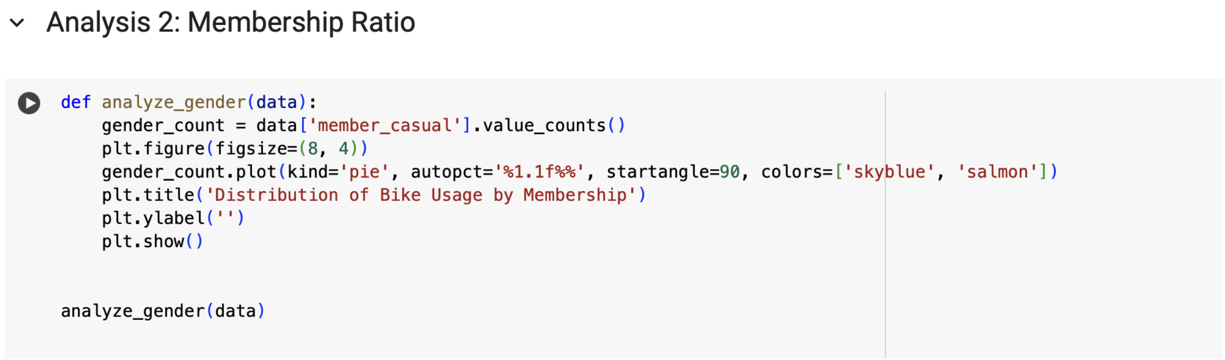
```python
# Load data
def load_data(file_path):

    data = pd.read_csv(file_path)
    data['started_at'] = pd.to_datetime(data['started_at'])
    data['ended_at'] = pd.to_datetime(data['ended_at'])

    # Encoding Station names so that data privacy is maintained
    start_frequency_map = data['start_station_name'].value_counts().to_dict()
    data['start_station_name_freq_encoded'] = data['start_station_name'].map(start_frequency_map)

    end_frequency_map = data['end_station_name'].value_counts().to_dict()
    data['end_station_name_freq_encoded'] = data['end_station_name'].map(end_frequency_map)
    return data
```

Here,

We are loading the data.

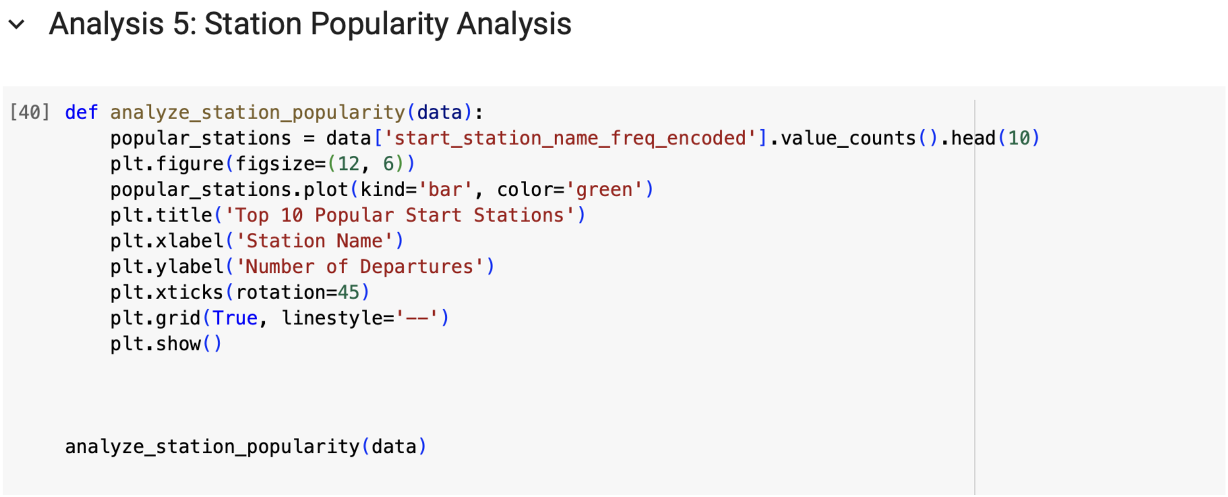
Also, We are correcting some of the data types for our analysis.

Focusing on data privacy, we are encoding proprietary data ensuring that the quality of data is maintained.



Analysis 2: Membership Ratio

```python
def analyze_gender(data):
    gender_count = data['member_casual'].value_counts()
    plt.figure(figsize=(8, 4))
    gender_count.plot(kind='pie', autopct='%1.1f%%', startangle=90, colors=['skyblue', 'salmon'])
    plt.title('Distribution of Bike Usage by Membership')
    plt.ylabel('')
    plt.show()


analyze_gender(data)
```



Analysis 5: Station Popularity Analysis

```python
[40] def analyze_station_popularity(data):
         popular_stations = data['start_station_name_freq_encoded'].value_counts().head(10)
         plt.figure(figsize=(12, 6))
         popular_stations.plot(kind='bar', color='green')
         plt.title('Top 10 Popular Start Stations')
         plt.xlabel('Station Name')
         plt.ylabel('Number of Departures')
         plt.xticks(rotation=45)
         plt.grid(True, linestyle='--')
         plt.show()



     analyze_station_popularity(data)
```

Result

Our analysis of the Divvy dataset unveiled significant insights: Electric bikes are gaining popularity, while members tend to make more frequent and longer trips compared to casual riders. Peak usage occurs during morning and evening commuting hours, with certain stations experiencing heavy traffic. Additionally, members take longer trips on average, highlighting variations in travel behavior.

Privacy Preservation:

Privacy preservation is a critical aspect of our analysis process. To safeguard user information, we employed several techniques, including data encoding and differential privacy:

- Data Encoding: We encoded proprietary information to prevent the disclosure of personally identifiable details during analysis. This encoding ensured that sensitive information remained protected while still allowing for meaningful insights to be derived from the dataset.
- Differential Privacy: Differential privacy techniques were applied to further enhance privacy protection. By adding noise to the analysis results, we ensured that individual user data could not be inferred from the aggregated outputs, thus preserving the privacy of Divvy users.

Through these privacy-preserving measures, we maintained the confidentiality of user information while still deriving valuable insights to inform decision-making processes and enhance the efficiency and accessibility of the Divvy bike-sharing system.

Summary

Our project embarked on a comprehensive analysis of the Divvy bike-sharing system with a twofold objective: to uncover insights into how communities engage with the service and to ensure the stringent protection of user privacy throughout our analytical processes. This report has detailed our journey from conceptualization through to the execution of advanced data analysis techniques underpinned by robust privacy-preserving protocols.

Key Findings

Our analyses yielded several important insights into the usage patterns of the Divvy bike-sharing system:

- **Bike Type Preference**: We observed distinct preferences in bike types between different user groups, with electric bikes growing in popularity among casual riders.
- **User Demographics**: Members showed different usage patterns compared to casual riders, including more frequent and longer trips, indicating their reliance on the service for daily commutes.
- **Peak Usage Times**: The analysis identified peak hours which correspond with morning and evening rush hours, suggesting the critical role of the bike-sharing system in supporting urban commuting.
- **Station Utilization**: Certain stations demonstrated significantly higher traffic, highlighting potential areas for infrastructural improvement.

Privacy and Security Achievements

We implemented several cutting-edge techniques to protect the privacy of Divvy users:

- **Differential Privacy**: This was integral to our approach, ensuring that our data outputs could not be used to identify individual users.
- **Data Anonymization**: Through techniques like k-anonymity and pseudonymization, we safeguarded personal identifiers, ensuring the anonymization of the dataset before analysis.

Impact and Future Work

The findings from this project have significant implications for urban planning and transportation management. They provide evidence-based insights that can help optimize the distribution of bike stations, improve the allocation of resources, and tailor marketing strategies to enhance user engagement and system efficiency. Looking forward, we plan to explore the integration of real-time data analysis and predictive modeling to further enhance the responsiveness of the Divvy system to changing urban dynamics. Additionally, the adoption of Secure Multi-Party Computation (SMPC) and Homomorphic Encryption in future studies could open new avenues for collaborative, privacy-preserving analyses across multiple stakeholders.