CSP 571 Data Preparation and Analysis

Project Proposal and Outline

Analyzing Divvy Bike Usage Patterns in Chicago

Mohammad Hamza Piracha
mpiracha@hawk.iit.edu
A20554741

Poshan Pandey
Ppandey6@hawk.iit.edu
A20519852

Usman Matheen Hameed
uhameed@hawk.iit.edu
A20564338

Project Proposal

Objective

The goal of this project is to analyze Divvy bike-sharing data in Chicago to identify the busier and less busy areas and times, provide recommendations for optimizing the distribution of docks and cycles, and compare the usage patterns of electric versus classic bicycles and scooters, as well as regular guest users versus Chicago resident users.

Research Questions

- Which areas and times are the busiest for Divvy bike usage in Chicago?
- Which areas need more docks and cycles based on usage patterns?
- How does the usage of electric bicycles compare to classic bicycles and scooters?
- What are the differences in usage patterns between regular guest users and Chicago resident users?

Proposed Methodology

Data Collection: Gather all necessary datasets from the Divvy bike data repository.

Data Preparation: Clean the datasets by removing unnecessary data and combining them into a single comprehensive dataset.

Data Analysis: Use descriptive and inferential statistical methods to identify trends and patterns in the data.

Visualization: Create visualizations to highlight insights and findings from the data.

Recommendations: Provide constructive suggestions based on the analysis for optimizing dock and cycle distribution.

Metrics for Analysis

Number of Rides: Total number of rides per area, user type, bike type and time period.

Dock Utilization: Frequency of dock usage in different areas.

Ride Duration: Average duration of rides for different bike types.

User Patterns: Distribution of rides between guest users and resident users.

Project Outline

Data Sources and Reference Data

Main Datasets:

- *Index of bucket "divvy-tripdata"*

Dataset Features:

- ride_id: Unique identifier for the ride
- rideable_type: Type of the bike (classic/electric)
- started_at: Start date and time of the ride
- ended_at: End date and time of the ride
- start_station_name: Name of the station where the ride started
- start_station_id: Unique identifier for the station where the ride started
- end_station_name: Name of the station where the ride ended
- end_station_id: Unique identifier for the station where the ride ended
- start_lat: Latitude of the location where the trip started
- start_lng: Longitude of the location where the trip started
- end_lat: Latitude of the location where the trip ended
- end_lng: Longitude of the location where the trip ended
- member_casual: Type of rider (member/casual)

Data Processing

- **Data Cleaning:** Removing duplicate entries, handling missing values, and correcting any data inconsistencies.
- **Data Imputing:** Addressing missing values where applicable.
- **Data Transformation:** Converting date and time fields into appropriate formats, aggregating data by different time intervals (hourly, daily, monthly).
- **Outlier Detection:** Identifying and handling outliers to ensure data quality.
- **Distributional Analysis:** Analyzing the distribution of ride counts, durations, and distances.
- **Clustering:** Identifying clusters of high and low activity areas.
- **Dimensionality Reduction:** Using techniques like PCA to reduce the dataset's dimensionality for visualization purposes.

Model Selection

- **Feature Selection:** Identifying the most relevant features for the analysis.
- **Classification/Regression Approaches:** Using regression models to predict dock and cycle needs based on historical usage patterns(hopefully).
- **Reference/Baseline Model:** Establishing a baseline model for comparison.

Software Packages and Tools

- **R Language:** The primary programming language for analysis.
- **R Studio:** Integrated development environment for R.

Libraries:

- dplyr for data manipulation
- ggplot2 for data visualization
- tidyr for data tidying
- lubridate for date and time handling
- caret for machine learning model training and evaluation (hopefully)
- more on implementation

This project will leverage these tools and methodologies to provide actionable insights into the Divvy bike-sharing system in Chicago, ultimately aiming to enhance its efficiency and user satisfaction.

Future Scope

- **Integration with Public Transit:** Analyze the relationship between Divvy bike docks and CTA bus stops to optimize the multimodal transportation network in Chicago.
- **Population Analysis:** Correlate Divvy usage data with population density data to identify areas with unmet demand for bike docks and cycles.
- **Expansion to Other Cities:** Apply the analysis methodology to other cities with bike-sharing programs to compare and improve overall urban mobility.
- **Real-Time Data Integration:** Incorporate real-time data feeds to provide dynamic recommendations for bike redistribution and dock availability.
- **User Experience Enhancement:** Analyze user feedback and usage patterns to improve overall user experience.
- **Sustainability Impact:** Evaluate the environmental benefits of the bike-sharing program and suggest improvements for increasing its positive impact on urban sustainability.

References:

L. Czarlnski, "Exploratory Data Analysis (EDA) of the Chicago Divvy Bikes Dataset," Medium. [Online]. Available: Exploratory Data Analysis (EDA) of the Chicago Divvy Bikes Dataset | by Leon Czarlnski | Medium.

"Exploring variations in Divvy bike station usage volume: from historical trip records to Google Street view images," MACS 37000 (Spring 2021) Thinking with Deep Learning for Complex Social & Cultural Data Analysis, uchicago.edu. [Online]. Available: Exploring variations in Divvy bike station usage volume: from historical trip records to Google street view images | MACS 37000 (Spring 2021) Thinking with Deep Learning for Complex Social & Cultural Data Analysis (uchicago.edu).

"Divvy Trips," City of Chicago, Data Portal. [Online]. Available: Divvy Trips | City of Chicago | Data Portal.

Shivaniwac, "Quarterly Success: Divvy Bike's 2024 Growth Analysis," Medium, May 2024. [Online]. Available: Quarterly Success: Divvy Bike's 2024 Growth Analysis | by Shivaniwac | May, 2024 | Medium.

"Index of bucket 'divvy-tripdata'," [Online]. Available: Index of bucket "divvy-tripdata".