Bike-Sharing Operations Analysis: Insights and Recommendations for Optimization

Assignment 4: Creating Reports and Dashboards for Predictive and Prescriptive Analysis

Dayanara Torres Macalino (NF1001047),

Miko L. Tan (NF1008647),

Dev D. Rabadia (NF1005560),

Stephy Marvin Christi (NF1003839)

Master of Data Analytics, University of Niagara Falls Canada

CPSC-510-5: Winter 2025 Data Warehousing/Visualization

Professor Mehdi (Matt) Mostofi

March 23, 2025

Overview of Workflow and Methodology

Data Loading and Initial Exploration

The Bike Sharing Dataset (Bike Sharing Dataset.csv) was loaded using pandas (pd.read_csv), containing 731 daily entries from January 1, 2011, to December 31, 2012. The dataset's structure was examined with bike_share_data.info(), revealing 16 columns with no missing values (Non-Null Count: 731 for all columns). Key features included numerical columns like temp, hum, windspeed, and cnt (total rentals), and categorical columns such as season, holiday, and weathersit. The nunique() method confirmed that instant and dteday had 731 unique values, aligning with the dataset size, while categorical variables like season (4 unique values) and numerical variables like cnt (696 unique values) showed expected variability.

Data Cleaning and Preprocessing

The dataset was checked for missing values using bike_share_data.isnull().sum(), which returned 0 for all columns, indicating no missing data. Duplicate rows were assessed with bike_share_data.duplicated().sum(), confirming no duplicates (result: 0).

Outliers in the hum (humidity) column were identified using the Interquartile Range (IQR) method. The IQR was calculated (Q1 - 1.5*IQR for the lower bound and Q3 + 1.5*IQR for the upper bound), and 2 rows were removed, reducing the dataset from 731 to 729 entries. Boxplots and histograms visualized the distribution of hum before and after outlier removal, confirming improved data quality.

The uniqueness of columns was verified using nunique(), ensuring instant and dteday were unique identifiers, while other columns like weathersit (3 unique values) and casual (606 unique values) aligned with their expected ranges.

Exploratory Data Analysis (EDA)

Histograms of numerical features (cnt, temp, hum, windspeed, etc.) were plotted using bike_share_data.hist() and custom plt.hist() loops to assess distributions. For instance, cnt showed a slight left skew (-0.047), while windspeed was right-skewed (0.677). Bar plots for categorical features like season were created using sns.countplot(), revealing balanced seasonal distribution.

Scatter plots, such as temp vs. cnt, were generated using sns.scatterplot(), showing a positive relationship between temperature and rentals. A correlation matrix of numerical features was computed and visualized as a heatmap (sns.heatmap(corr)), highlighting a strong positive correlation between temp and cnt (0.63) and between registered and cnt (0.95).

Boxplots for numerical features were created to confirm outliers, particularly in hum, which were addressed during preprocessing.

Feature Engineering and Transformation

The dteday column was converted to datetime format (pd.to_datetime), and new features were extracted: hour, day_of_week (e.g., Monday), month_name (e.g., January), and season_name (mapped from season, e.g., 1 to 'Spring').

The weathersit column was mapped to descriptive categories (e.g., 1 to 'Clear'), and numerical features like temp, hum, and windspeed were binned into categories (e.g., temp_category: 'Cold', 'Mild', 'Hot').

A user_type column was created to classify rentals as 'Casual' or 'Registered' based on the dominant user type, and rentals_per_user was computed as cnt / (casual + registered). A day_type column categorized days as 'Holiday', 'Working Day', or 'Weekend', and an is_weekend binary indicator was added for weekends. A comfort_index was calculated as temp - hum + (1 - windspeed), and a peak_hour indicator flagged rush hours (7–9 AM, 5–7 PM).

The windspeed column, initially right-skewed (0.677), underwent a Box-Cox transformation, reducing skewness to -0.003. The weathersit column, also right-skewed (0.957), was transformed using a logit function, achieving a skewness of 0.0. Visualizations (histograms with KDE) confirmed improved normality post-transformation.

Statistical Analysis

Skewness was computed for numerical features, identifying holiday (5.65) and casual (1.27) as extremely right-skewed, indicating rare but significant spikes in casual usage on holidays. workingday was left-skewed (-0.79), while most features like cnt (-0.047) and temp (-0.055) showed minimal skewness.

Post-transformation distributions of windspeed_boxcox and weathersit_logit were visualized, confirming improved normality. Boxplots further validated the removal of outliers in hum, enhancing the dataset's suitability for analysis.

Exporting the Dataset

The cleaned and transformed dataset, now with 19 columns (including engineered features), was saved as cleaned_bike_sharing_dataset.csv using bike_share_data.to_csv for further visualization and analysis in Power BI.

Power BI Visualization

The cleaned dataset (cleaned_bike_sharing_dataset.csv) was imported into Power BI Desktop using the "Get Data > Text/CSV" option, loading all 729 entries and 19 columns. Data types were verified, ensuring dteday was set as a Date type, and numerical columns like cnt and temp were correctly formatted. Visualizations were created to reflect key insights: A Line Chart for "Bike Rental Performance" was built using dteday on the X-axis and cnt on the Y-axis to show rental trends over time, highlighting peaks in warmer months**.** A Clustered Column Chart for "Bike Rental by Weather" used weathersit on the X-axis and cnt on the Y-axis, showing higher rentals in clear weather (2.3M) compared to misty (1.0M) and light rain (0.0M) conditions. A Stacked Column Chart for "Rental Counts by Season and Weather Category" was created with season_name on the X-axis, cnt on the Y-axis, and weathersit as the Legend, illustrating higher rentals in Summer and Fall on clear days. A Donut Chart for "Customer Breakdown" displayed the proportion of user_type, showing 728K registered users (99.59%) and 3K casual users (0.41%). A Histogram for "Monthly Bike Rentals" was generated using month_name on the X-axis and cnt on the Y-axis, confirming peak rentals from June to August. A Time-Series Forecast (SARIMA) plot for "Rental Count Forecast" was added using Power BI's forecasting feature on the cnt time-series data, showing historical and predicted trends.

Total Bike Rentals (3M), Daily Bike Rentals (4.50K), and Daily Registered Users (3.66K). Slicers for user_type, season_name, and weathersit were included to enable interactive filtering. The dashboard was saved as BikeSharingAnalysis.pbix and published to the Power BI Service for sharing and scheduled refreshes.

This methodology ensured a robust foundation for deriving insights, combining data cleaning, EDA, feature engineering, statistical analysis, and Power BI visualization to prepare the dataset for prescriptive analysis in the bike-sharing operations report.

Summary Report

Overview of the Datasets and Their Attributes

The bike-sharing dataset encompasses 731 daily entries from January 1, 2011, to December 31, 2012, capturing bike rental activities over two years. After cleaning, the dataset (cleaned_bike_sharing_dataset.csv) contains 729 entries following the removal of outliers in the hum column, with no duplicates identified. Key attributes include categorical features such as season (1: Spring, 2: Summer, 3: Fall, 4: Winter; mean: 2.5, std: 1.11), holiday (binary: 0/1; mean: 0.029, std: 0.17), weekday (0 to 6; mean: 3.0, std: 2.0), workingday (binary: 0/1; mean: 0.68, std: 0.47), and weathersit (1 to 4; mean: 1.4, std: 0.54). Numerical features include temp (normalized temperature; mean: 0.50, std: 0.18), hum (normalized humidity; mean: 0.63, std: 0.14), windspeed (normalized wind speed; mean: 0.19, std: 0.08), and cnt (total rentals; mean: 4,504, std: 1,937, min: 22, max: 8,714). The dataset's overall rental statistics are summarized in the dashboard, showing 3M total bike rentals, 4.50K daily bike rentals, and 3.66K daily registered users (Dashboard: Left Panel). The dataset provides a robust foundation for analyzing rental patterns influenced by temporal, weather, and user-related factors.

Key Findings from Optimization and Decision Analysis

Analysis of the bike-sharing dataset reveals significant insights into rental patterns and operational optimization. Rentals (cnt) peak in Summer and Fall, with a strong positive correlation between temp and cnt (0.63), indicating higher usage in warmer conditions and on clear days (weathersit = 1). This seasonal trend is evident in the "Rental Counts by Season and Weather Category" visualization, where Summer and Fall show the highest rental counts, particularly on clear days (Dashboard: Bottom Right), and in the "Monthly Bike Rentals" histogram, which peaks from June to August (Dashboard: Top Center). Working days exhibit higher demand (mean: 4,504 rentals) compared to weekends or holidays, reflecting commuting trends, while casual users dominate on weekends and registered users on working days. The user breakdown is illustrated in the "Customer Breakdown" donut chart, showing 728K registered users (99.59%) and 3K casual users (0.41%) (Dashboard: Top Right). Weather impacts are notable, with temp and lower hum driving increased rentals, as captured by the comfort_index, while windspeed shows a mild negative effect. The "Bike Rental by Weather" bar chart highlights that clear weather (2.3M rentals) significantly outperforms misty (1.0M) and light rain (0.0M) conditions (Report Tab: Bike Rental by Weather), and the "Bike Rental by Temperature" chart shows higher rentals in mild (1.53M) and hot (1.44M) conditions compared to cold (0.32M) (Report Tab: Bike Rental by Temperature). Statistically, holiday (skewness: 5.65) and casual (1.27) are extremely right-skewed, indicating rare but significant spikes in casual usage on holidays, whereas cnt is slightly left-skewed (-0.047), suggesting occasional low-rental days. The "Bike Rental Performance" time-series plot shows a downward trend in rentals during colder months, with peaks in warmer months (Report Tab: Bike Rental Performance). After outlier removal, hum distribution improved, enhancing the reliability of weather-based insights.

Real-World Applications and Business Impact

The insights derived offer practical applications for bike-sharing operations. Optimizing bike distribution involves allocating more bikes during Summer and Fall, particularly on clear days, to meet heightened demand driven by favorable weather, while reducing the fleet on holidays and rainy days (weathersit = 3 or 4) to avoid underutilization. This strategy aligns with the "Rental Counts by Season and Weather Category" visualization, which shows higher rentals in Summer and Fall on clear days (Dashboard: Bottom Right), and the "Monthly Bike Rentals" histogram, which peaks in warmer months (Dashboard: Top Center). Dynamic pricing strategies can include surge pricing during peak hours (7–9 AM, 5–7 PM) and warmer months to maximize revenue, alongside offering discounts on rainy days or in winter to boost ridership among casual users, who spike on holidays. The "Key Insights and Recommendations" section of the dashboard suggests surge pricing in Summer and on peak hours, as well as discounts in Winter (Report Tab: Key Insights and Recommendations). Weather-based adjustments involve promoting indoor alternatives on adverse weather days, reducing the active fleet to lower maintenance costs. The "Bike Rental by Weather" chart supports this by showing minimal rentals during light rain (Report Tab: Bike Rental by Weather). Infrastructure planning recommends expanding stations in areas with high registered usage, likely business districts, to cater to consistent weekday demand, and adding temporary stations during summer events or holidays to accommodate casual users. The "Customer Breakdown" chart indicates a dominance of registered users, supporting the focus on business districts (Dashboard: Top Right), and the "Friday Peak Bike Rental Day" insight highlights the need for increased capacity on peak days (Report Tab: Friday Peak Bike Rental Day).

Recommendations for Further Exploration or Applications of the Insights

Further exploration can enhance operational efficiency and user experience. Enhanced demand forecasting should incorporate external factors like local events or traffic patterns to improve predictions, especially for casual users, and employ predictive models like regression or time-series analysis to forecast cnt based on temp, weathersit, and workingday. The "Rental Count Forecast (SARIMA)" plot provides a foundation for such forecasting, showing historical and predicted rental trends (Dashboard: Bottom Center). Customer segmentation could analyze user demographics to distinguish commuters from leisure riders, tailoring pricing or promotions, and investigate holiday impacts on specific groups to refine strategies. The "Customer Breakdown" chart can guide segmentation by highlighting the small but significant casual user base (Dashboard: Top Right). Operational efficiency can be improved by using predictive maintenance models to schedule repairs based on usage patterns, minimizing downtime, and implementing real-time bike rebalancing to address intraday demand shifts. The "Friday Peak Bike Rental Day" insight suggests focusing on peak days for rebalancing efforts (Report Tab: Friday Peak Bike Rental Day). Infrastructure and sustainability efforts should explore partnerships with local governments to integrate bike stations with public transit, assess environmental impacts, and consider electric bikes to increase ridership in residential or hilly areas.