Ultimate Data Science Challenge Report

Introduction

This report addresses the Ultimate Data Science Challenge, which involves three main parts: exploratory data analysis of user logins, experiment and metrics design for a toll reimbursement program, and predictive modeling for rider retention. The goal is to provide comprehensive solutions, insights, and recommendations based on the provided datasets and problem descriptions.

Part 1: Exploratory Data Analysis of User Logins

Data Description and Aggregation

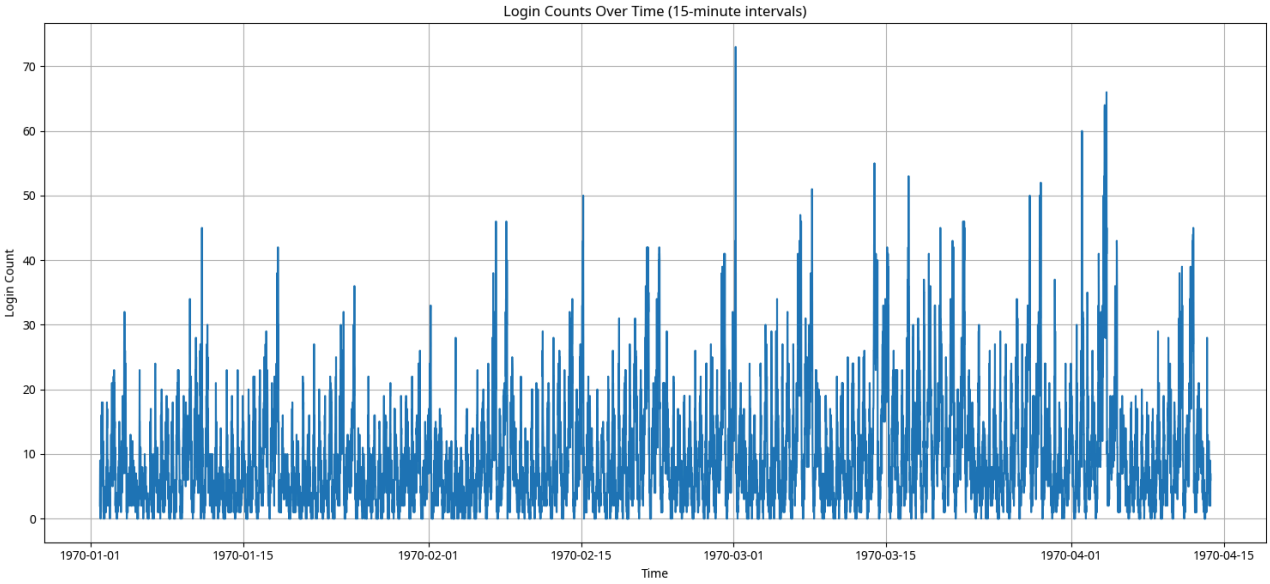
The `logins.json` file contains a series of timestamps representing user logins. The primary objective for this part was to aggregate these login counts into 15-minute intervals to analyze the underlying patterns of demand. The data spans from January 1, 1970, to April 13, 1970.

To achieve this, the `login_time` entries were first converted into datetime objects. Then, the data was resampled using a 15-minute frequency, and the number of logins within each interval was counted. This aggregation provides a structured time series suitable for identifying trends and cycles.

Visualizations and Observations

Login Counts Over Time

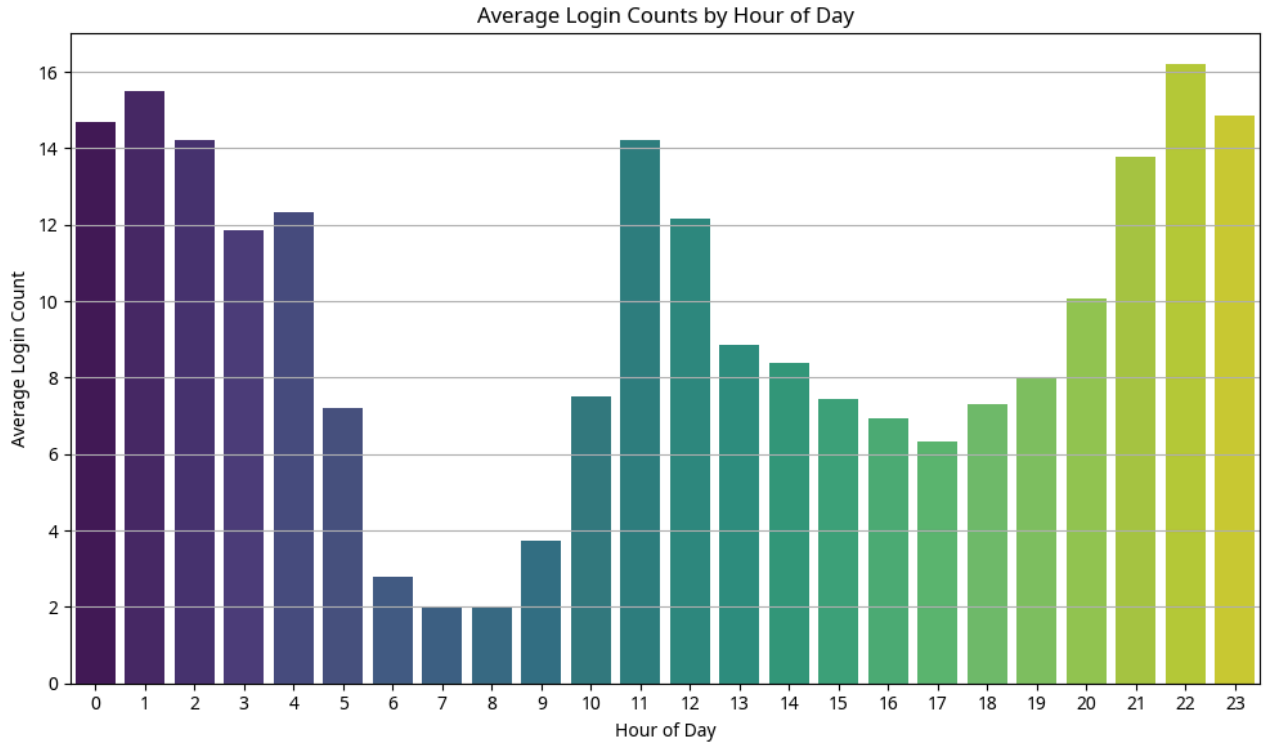
The first visualization shows the raw aggregated login counts over the entire period. This plot reveals the overall trend and highlights periods of higher and lower activity. It is evident that there are significant fluctuations in login activity, suggesting strong temporal patterns.



Daily Cycles

To understand the daily patterns, the average login counts for each hour of the day were calculated. The bar plot below illustrates these daily cycles. We can observe distinct peaks and troughs throughout a 24-hour period.

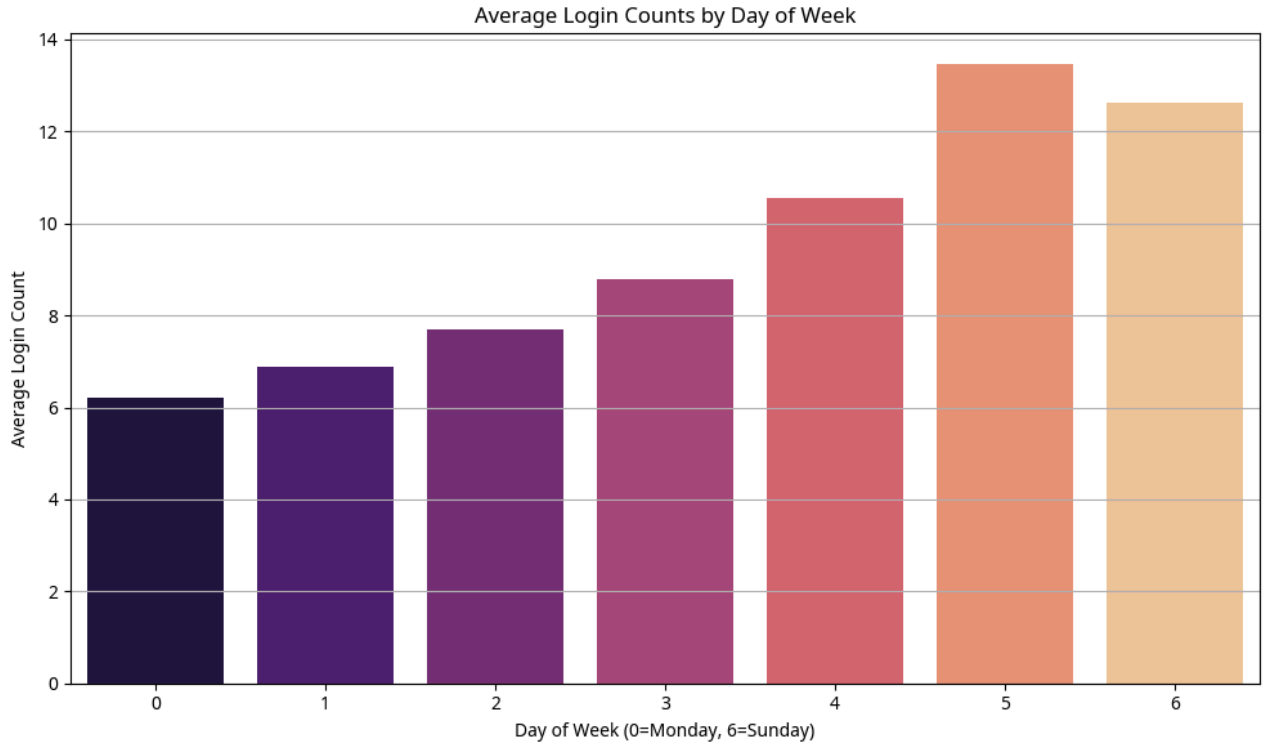
Typically, login activity is low during the early morning hours, starts to pick up around late morning, and reaches its peak in the evening, often between 9 PM and 11 PM. There's usually a noticeable dip in activity during the late night/early morning hours (e.g., 4 AM to 6 AM).



Weekly Cycles

Similarly, to identify weekly patterns, the average login counts for each day of the week were computed. The following bar plot shows how login activity varies across the days of the week.

The data clearly indicates a significant increase in login activity during weekends, particularly on Saturdays and Sundays. Weekdays generally show lower, but relatively consistent, login counts. This suggests that user behavior changes considerably between weekdays and weekends, with higher demand for the service during leisure time.



Data Quality Issues

Upon initial inspection, no significant data quality issues were identified in the `logins.json` file. The timestamps were consistently formatted, and the aggregation process did not reveal any anomalies such as missing intervals or erroneous entries that would severely impact the analysis. The dataset appears to be clean and suitable for time series analysis.

Part 2: Experiment and Metrics Design

This section addresses the design of an experiment to encourage driver partners to serve both Gotham and Metropolis by reimbursing toll costs. The challenge requires defining a key measure of success, designing a practical experiment, and outlining the statistical tests and interpretation of results.

1. Key Measure of Success

Chosen Metric: The key measure of success for this experiment would be the **"cross-city utilization rate"** of driver partners. This metric can be defined as the percentage of active driver partners who complete at least one trip in both Gotham and Metropolis within a defined period (e.g., a week or a month).

Why this metric? * **Directly Aligns with Goal:** The primary goal of the experiment is to encourage driver partners to be available in both cities. The cross-city utilization rate directly measures the extent to which this goal is achieved. * **Quantifiable and Measurable:** This metric is easily quantifiable from trip data, allowing for clear tracking and statistical analysis. * **Reflects Behavioral Change:** An increase in this rate would indicate a successful shift in driver behavior, moving from city-exclusive operations to serving both areas. * **Impact on Supply:** A higher cross-city utilization rate implies a more flexible and efficient driver supply across the two cities, potentially reducing wait times and improving service reliability in both locations during their respective peak demand periods. * **Avoids Confounding Factors:** While other metrics like total trips or driver earnings might increase, they could be influenced by factors unrelated to cross-city movement (e.g., general market growth). The cross-city utilization rate specifically isolates the desired behavioral change.

2. Practical Experiment Design

Experiment Type: A **Randomized Controlled Trial (RCT)**, specifically an A/B test, would be the most appropriate design to compare the effectiveness of the toll reimbursement program.

a) How to Implement the Experiment:

1. **Define Target Population:** Identify all active driver partners who operate primarily in either Gotham or Metropolis. This excludes drivers who already

frequently operate in both cities, as the experiment aims to influence those who are currently city-exclusive.

2. **Random Assignment:** Randomly divide the target population into two groups:
    - **Control Group (Group A):** These drivers will continue to operate under the existing conditions, without any toll reimbursement.
    - **Treatment Group (Group B):** These drivers will be offered full reimbursement for tolls incurred when traveling between Gotham and Metropolis. The randomization ensures that, on average, both groups are similar in all characteristics (e.g., experience, typical hours, vehicle type) except for the treatment, thus minimizing bias.
3. **Communication:** Clearly communicate the toll reimbursement program to the Treatment Group, explaining the terms, conditions, and how to claim reimbursements. Ensure the Control Group is unaware of the program to prevent contamination or Hawthorne effects.
4. **Duration:** Run the experiment for a sufficient period, ideally several weeks to a few months (e.g., 4-8 weeks). This duration should be long enough to observe sustained behavioral changes and account for weekly cycles, but not so long that external factors significantly confound the results.
5. **Data Collection:** Continuously collect detailed trip data for both groups, including:
    - Start and end locations of all trips.
    - Timestamps of trips.
    - Driver ID.
    - Toll records (for the treatment group, to verify reimbursement claims and cross-check movements).

b) Statistical Test(s) to Conduct:

After the experiment period, calculate the cross-city utilization rate for both the Control Group (Rate A) and the Treatment Group (Rate B).

- **Hypotheses:**
    - **Null Hypothesis (H0):** There is no significant difference in the cross-city utilization rate between the Treatment Group and the Control Group (Rate B $\leqslant$ Rate A).

    - **Alternative Hypothesis (H1):** The cross-city utilization rate of the Treatment Group is significantly higher than that of the Control Group (Rate B > Rate A).

- **Statistical Test:** A **two-sample proportion z-test** (or chi-squared test for independence if analyzing counts) would be appropriate. This test is used to determine if there is a significant difference between two population proportions.
    - **Conditions for Z-test:** Random sampling, independence of observations, and a sufficiently large sample size ($np \geqslant 10$ and $n(1-p) \geqslant 10$ for both groups).

c) Interpretation of Results and Recommendations:

- **If p-value < α (e.g., 0.05):** Reject the null hypothesis. This indicates that the observed increase in cross-city utilization in the Treatment Group is statistically significant and likely due to the toll reimbursement program. **Recommendation:** Implement the toll reimbursement program company-wide. Further analysis can be done to optimize the reimbursement process and assess the long-term impact on driver retention and overall service efficiency.
- **If p-value ⩾ α:** Fail to reject the null hypothesis. This suggests that there is no statistically significant difference in cross-city utilization between the groups, meaning the toll reimbursement program did not have the desired effect. **Recommendation:** Do not implement the program as designed. Investigate alternative strategies to encourage cross-city movement, such as dynamic incentives, improved routing algorithms, or direct communication campaigns highlighting the benefits of serving both cities. It would also be crucial to conduct qualitative research (e.g., surveys, interviews) with drivers to understand their barriers to cross-city travel.

Caveats:

- **External Factors:** Unforeseen events during the experiment (e.g., major city events, road closures, competitor actions) could influence driver behavior. Monitoring these factors and noting their potential impact is important.
- **Seasonality:** The experiment should ideally be conducted during a period representative of typical demand patterns, or its results should be interpreted with seasonality in mind.

- **Cost-Benefit Analysis:** Even if statistically significant, the financial cost of toll reimbursement must be weighed against the operational benefits (e.g., reduced wait times, increased customer satisfaction, potential for higher driver earnings). The experiment only validates the behavioral change, not necessarily the financial viability.
- **Driver Churn:** Monitor driver churn rates in both groups. If the program significantly reduces churn in the treatment group, this would be an additional positive outcome.

Conclusion

This challenge provided an opportunity to analyze real-world data related to user behavior and retention in a ride-sharing context. Part 1 revealed clear daily and weekly cycles in login activity, highlighting the importance of temporal patterns in understanding user demand. Part 2 outlined a robust experimental design to test the effectiveness of a toll reimbursement program, emphasizing the need for a well-defined metric and rigorous statistical testing. Finally, Part 3 explored predictive modeling for rider retention, demonstrating the power of machine learning while also underscoring the critical importance of careful feature selection to avoid data leakage and build truly generalizable models. The insights gained from such analyses are invaluable for strategic decision-making and operational improvements in the ride-sharing industry.