Analyzing Driver Performance and Ride Metrics

Fang-Wen, Hsiao
25th June, 2024

Fang-Wen, Hsiao
Data Analyst
fangoaish@gmail.com / LinkedIn

1.0 Summary of Analysis

Data Source: The analysis used data from three tables:

1. driver_ids_df: Contains driver IDs and onboard dates. Total 937 driver users, no duplicates, no missing values.
2. ride_ids_df: Contains ride details for 193,502 rides, including distance and duration.
3. ride_timestamps_df: Contains ride event timestamps for 970,405 events, with one null value in the timestamp column, no duplicates.

Analytics Approach:

- Data Exploration: Overview and information on datasets, identifying duplicates and anomalies.
- Data Cleaning: Handling anomalies, null values, converting data types, and converting units.
- Data Analytics: Calculated metrics such as average lifetime value (LTV), projected driver lifetime, churn rate, driver segments and analyzed factors affecting LTV and driver churn.

Insights & Takeaways:

- **Average Lifetime Value (LTV):** The average LTV is $1531.
- **Main Factors Affecting LTV:** Total earnings positively impact LTV by approximately 0.25 units per unit increase. Longer wait times before arrival significantly increase churn risk.
- **Average Projected Lifetime of a Driver:** On average, a driver continues driving with Kiwi for 55 days.
- **Driver Churn Indicators:** High churn in early stages due to longer wait times and lower earnings. Strategies to optimize dispatch and improve compensation can reduce turnover.
- **Driver Segments by LTV:**
    - **High-LTV, High-Frequency:** 17% with the highest average revenue ($1157).
    - **Low-LTV, Low-Frequency:** 15% with the second-lowest average revenue ($75.95).
    - **Mid-High-LTV, Mid-High-Frequency:** 13% with significant average revenue ($655.93).
    - **Mid-Low-LTV, Mid-Low-Frequency:** 11% with moderate average revenue ($322.97).
- **Business Uses of Driver Segments Metric Metric:** Offer exclusive loyalty rewards to "High-LTV, High-Frequency" drivers, target mid-frequency, mid-LTV segments with upsell and cross-sell campaigns, and use tailored promotions to increase usage and value in low-frequency, low-LTV segments.

Agenda

- Analytics approach
  - Data exploration
  - Data Cleaning
- Analytics insights
  - Average lifetime value of a driver?
  - Main factors affect a driver’s LTV?
  - Avg. projected lifetime of a driver?
  - What is driver churn rate? Any predictive indicators for driver churn?
  - Do all drivers act alike? specific segments of drivers generating more value?
- Q&A
- Reference: Python code for KPIs

1.1 Analytics approach: Data Exploration

	driver_ids	ride_ids	ride_timestamps
Data Overview	• Driver ids and onboard dates • Total 937 driver users	• Ride details • Total 193,502 rides • 75% ride is < 6 miles *(avg. 4 miles)* • 96% ride is < 33 mins *(avg. 14 mins)* • 60% rids has 0 prime time *(avg.17)*	• Ride events and timestamps • Total 970,405 events
Data Quality	• No duplicates, no missing values	• No duplicates, no missing values	• 1 null value in the timestamp column, no duplicates

1.2 Analytics approach: Data Cleaning

Suspicious data points		Actions
ride_distance: ● Negative ride distance (-2 meters) ● 38 extremely high ride distance (> 62 miles) ride_duration: ● 165 extremely high ride duration (> 83 mins) ride_prime_time: ● 27 extremely high ride prime time (> 300%) timestamp: ● 1 null value	▶	Drop values: These outliers are less than 5%*
● driver_onboard_date column: object data type ● timestamp column: object data type	▶	Convert the data type from object to datetime
● ride_distance: ride distance in meters ● ride_duration: ride duration in seconds	▶	Convert the units ● seconds to mins ● meters to miles

5% references from datacamp: https://campus.datacamp.com/courses/feature-engineering-for-machine-learning-in-python/conforming-to-statistical-assumptions?ex=10

Agenda

- Analytics approach
  - Data exploration
  - Data Cleaning
- Analytics insights
  - Average Lifetime Value of a Driver?
  - Main factors affect a driver’s LTV?
  - Avg. Projected Lifetime of a driver?
  - What is driver churn rate? Any predictive indicators for driver churn?
  - Do All Drivers Act Alike? Specific Segments of Drivers Generating More Value?
- Q&A
- Reference: Python code for KPIs

2.1 Analytics Insights: Average Lifetime Value of a Driver?

Lifetime Value (LTV)

How much revenue a customer represents a business over the life of that relationship.

= **ARPU** (average revenue per user) x **%Gross Margin*** / **% Churn Rate**

Key Metrics

- Average Lifetime Value (LTV): $1531
- ARPU (Average Revenue Per User): $487
- Churn Rate: 15.9%

Distribution Analysis

- Positive skewness, indicating the majority has relatively lower LTV(high churn or low engagement) and a long tail towards higher values.
- Majority LTV values between $200 - $800; a smaller segment has LTV values above $1000 - $2000.

Recommendations

- Enhance Retention:
  - Reducing the (early)churn rate by implementing improvements based on feedback, provide support or incentives.

- Leverage Middle/High-Value Drivers:
  - Offer a bonus/retention program to increase engagement, and replicate the success of drivers with LTV above $1000.

**Gross margin = 0.5 is used in this data analysis - example value, adjust based on actual data*
LTV formula reference: https://www.mosaic.tech/financial-metrics/ltv-lifetime-value

2.1 Analytics Insights: Main Factors Affect a Driver's LTV?

Drivers' Total Earnings

Coefficient: 0.2500

- For each unit increase in total earnings, the LTV increases by approximately 0.25 units.
- Higher earnings are associated with higher LTV, reflecting the financial motivation for drivers to continue working.

#of Completed Rides

Coefficient: -0.4375

- For each additional ride, ***the LTV decreases by approximately 0.44 units.***
- Might suggest diminishing returns for additional rides or potential burnout effects on drivers.

Avg. Wait Time before Arrived

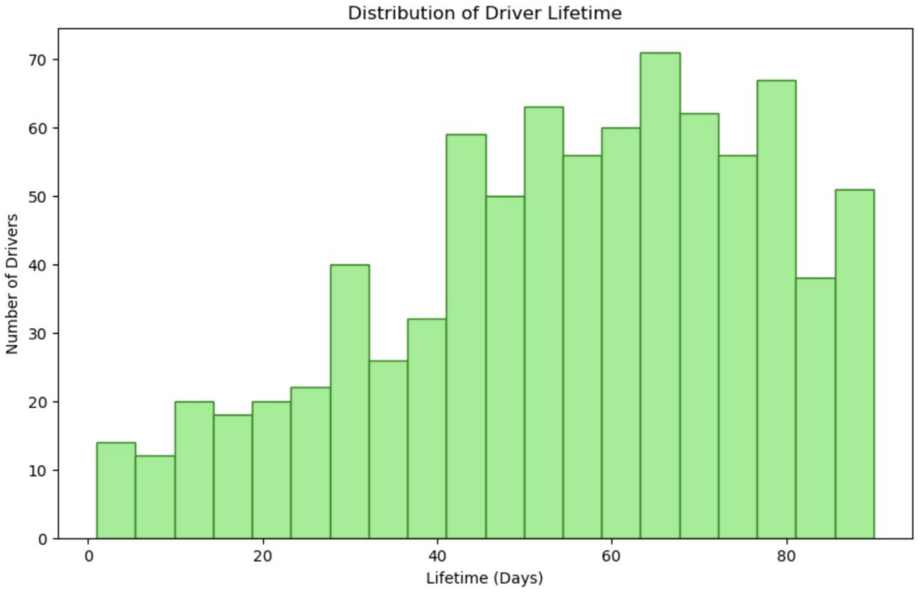
Coefficient: -8.891e-14

- The average wait time before arrived up has a very small negative impact on LTV.
- Indicating that wait time before arrived does not significantly affect LTV.

 # Linear Regression Analysis - the relationship between input & output variables

2.2 Analytics Insights: Avg. Projected Lifetime of a Driver?

Avg. Projected Lifetime: 55 days



Findings:

- Mid-range Engagement:
  - Steady driver activity for 20-80 days (1-3 months).
- Initial Drop-off:
  - Significant churn within the first 0-10 days.
- Long-term Drivers:
  - Decline in drivers after 80 days.

Recommendations:

- Retention:
  - Reduce early churn with better onboarding support and incentives.
- Mid-term Support:
  - Maintain motivation with training, recognition, and check-ins.
- Long-term Incentives:
  - Investigate and implement long-term engagement bonuses.

2.3 Analytics Insights: Driver Churn Rate? Any predictive indicators for driver churn?

1

Key Metrics from Logistic Regression

Classification Report:

- Precision: 0.89 for non-churned, 0.62 for churned.
- Recall: 0.96 for non-churned, 0.37 for churned.
    - Overall accuracy: 86.31%.

2

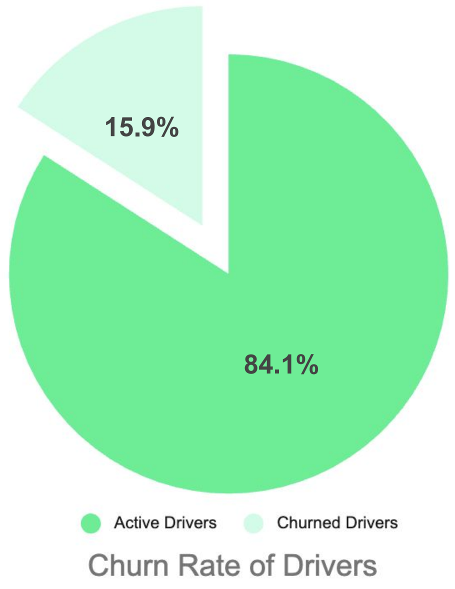
Findings

- Total Earnings: Small positive impact on churn.
- Average Wait Before Accepted: Slight positive impact, not significant.
- ***Average Wait Before Arrived: Strong positive impact, indicating longer waits increase churn risk.***

3

Recommendations

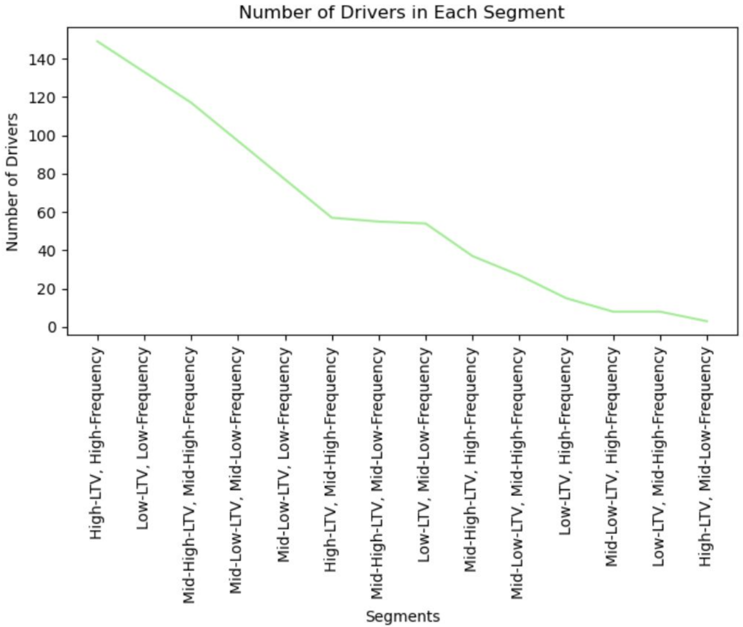
- Reduce Wait Times: Optimize real-time traffic updates and offer efficient dispatch system to reduce the wait time
- Enhance Data Quality: Explore additional factors like demographics and behavior to better predict churn.

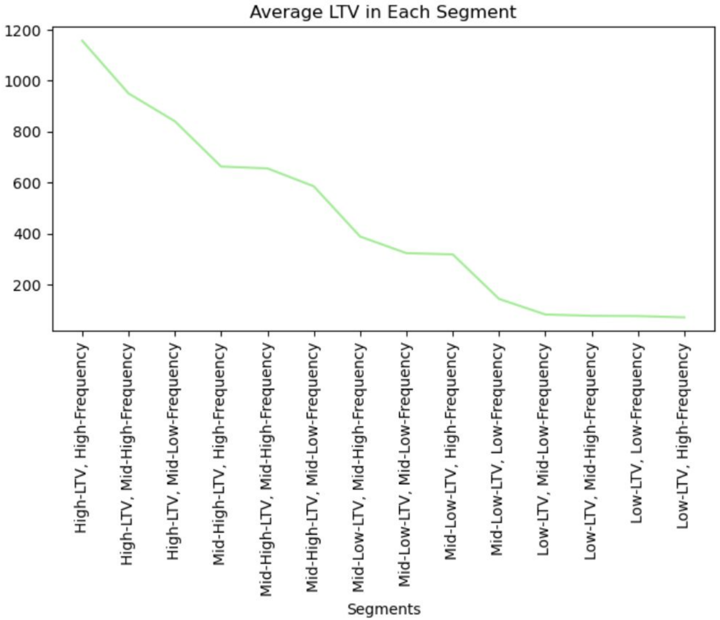


 # Use logistic regression to predict the internal factors: total_earnings, avg_wait_before_accepted,avg_wait_before_arrived

2.4 Do All Drivers Act Alike? Specific Segments of Drivers Generating More Value?

The charts show that drivers do not act alike, with the "High-LTV, High-Frequency" segment generating the most value. This indicates that specific driver segments, defined by higher LTV and frequency, contribute more significantly to overall revenue.





2.4 Do All Drivers Act Alike? Specific Segments of Drivers Generating More Value?

Segment	Findings
High-LTV, High-Frequency	• Largest segment (17%) with highest revenue ($1157).
Low-LTV, Low-Frequency	• Second largest segment (15%) with low revenue ($75.95).
Mid-High-LTV, Mid-High-Frequency	• 13% drivers with significant revenue ($655.93)
Mid-Low-LTV, Mid-Low-Frequency	• 11% drivers with moderate revenue ($322.97)

Recommend business actions

Strategic Planning:

- Use segment insights for business strategies and financial planning.
- Predict trends and allocate budgets effectively.

Targeted Retention and Incentives:

- Personalize loyalty programs for high-LTV & mid-LTV drivers to to maintain their loyalty and maximize their LTV.
- Address low-LTV drivers' challenges with tailored promotions increasing their usage frequency and overall value.

Resource Allocation:

- Focus resources on high-LTV segments.
- Optimize onboarding to reduce churn.

 # Use Quantile-based discretization function to divide & label the segments

2.4 Backup Code

Kiwi Analytics Assignment

Fang-Wen, Hsiao (Data Analyst)

June 10th, 2024

Analysis Conclusion

- Key Insights: Most rides are under 10 km and 33 minutes. The average LTV is $1531 and the driver lifetime is 55 days, with a churn rate of 15.9%.
- Factors Affecting LTV: For each unit increase in total earnings, the LTV increases by approximately 0.25 units.
- Churn Indicators: Longer wait times and lower earnings predict driver churn; optimizing dispatch and improving compensation can reduce turnover.
- Segment-Specific Approaches: Retain top performers (average revenue $1157) with personalized rewards, and boost mid-value driver engagement with targeted

In [37]:

```
# Rate card
base_fare = 2
cost_per_mile = 1.15
cost_per_min = 0.22
min_fare = 5
max_fare = 400
service_fee = 1.75
kiwi_revenue_percentage = 0.20
gross_margin = 0.5  # Example value, adjus


# Fare calculation formula
## 1. $2.00 + Duration(min) * 0.22 + Dista
## 2. Add prime-time if applicable
## 3. Enforce $5 minimum and $400 maximum
## 4. Add 1.75 Service Fee

# Add a new column to calculate the fare
# 1. Initial fare calculation
ride_ids_df['fare'] = base_fare + (ride_ic
# 2. Prime time calculation: fare = fare +
ride_ids_df['fare'] += ride_ids_df['fare']
# 3. Check the minimum and maximum limit,
# 3.1 Use lambda function to make sure the
# 3.1-1 min(x, max_fare): Catch the fare a
# 3.1-1 max(min_fare, min(x, max_fare)): (
ride_ids_df['fare'] = ride_ids_df['fare'].


# Revenue calculation formula: value for K
# Add a new column to calculate the kiwi_r
ride_ids_df['kiwi_revenue'] = (ride_ids_df


# Calculate total revenue & totla_drivers
total_revenue = ride_ids_df['kiwi_revenue'
total_drivers = driver_ids_df['driver_id']
```

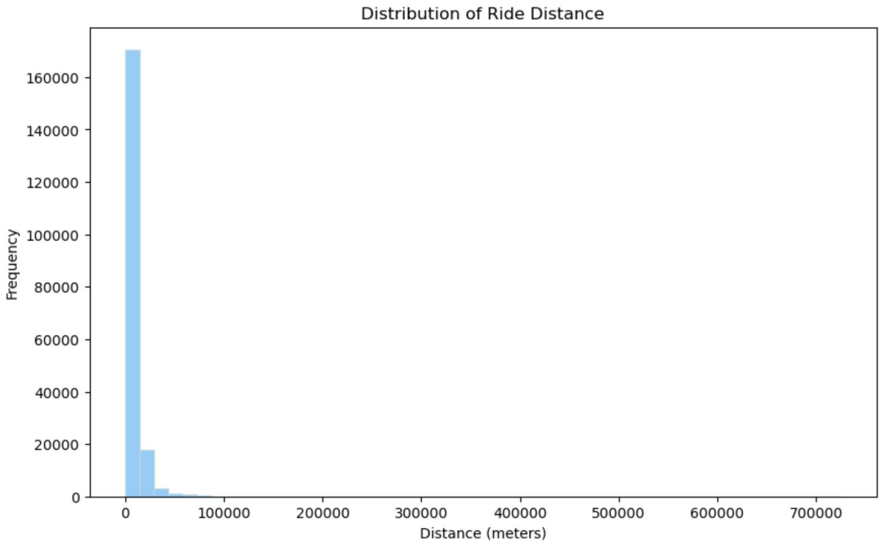
[Python Code Link]

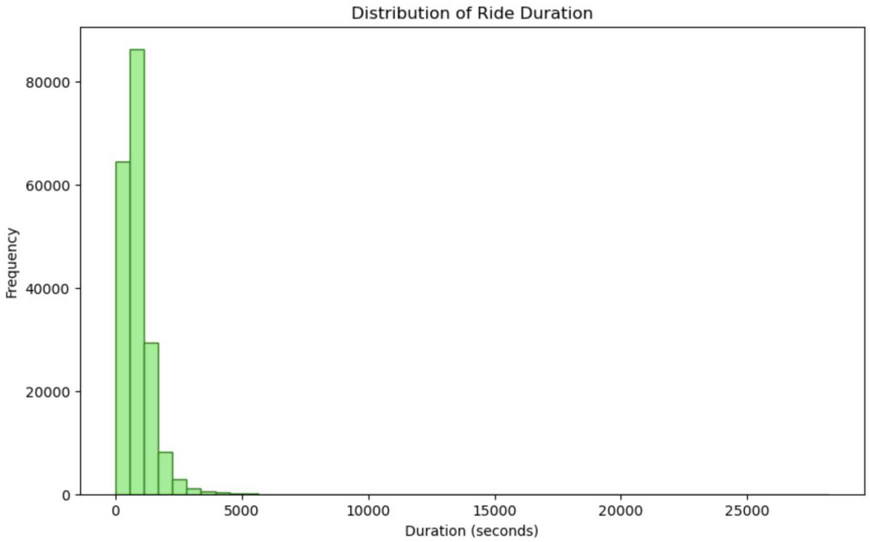
Thank You!

Q&A

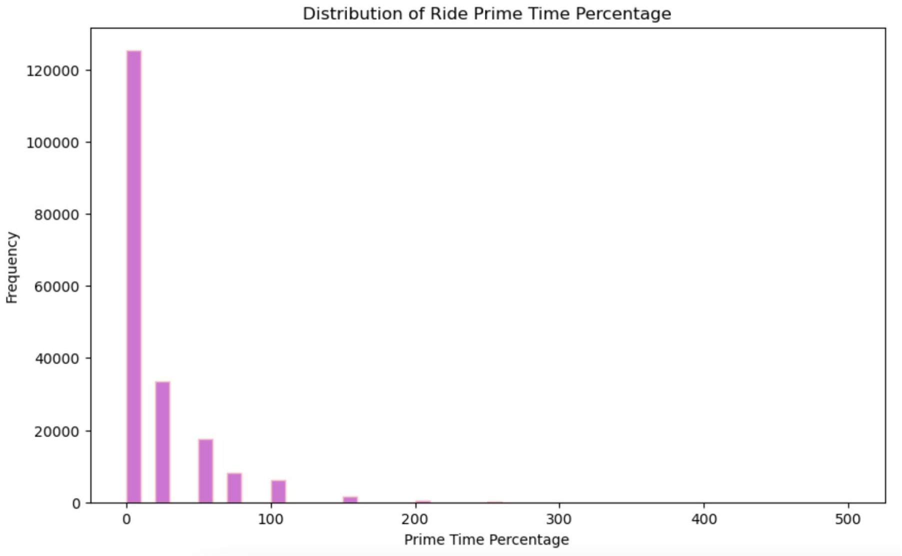
Backups

Backup #1.1: Analytics approach: Data Exploration

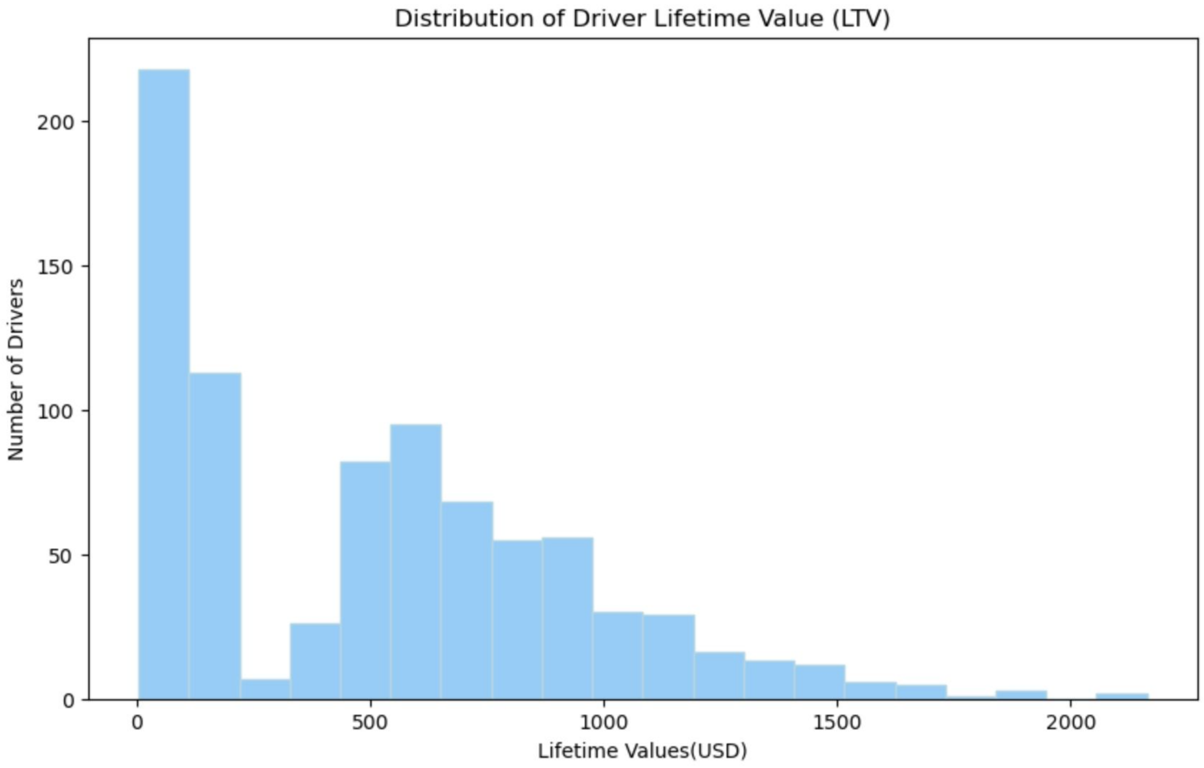




Backup #1.1: Analytics approach: Data Exploration



Backup #2.1: Analytics Insights: Average Lifetime Value of a Driver?



Backup #2.1: Main Factors Affect a Driver's LTV?

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    LTV   R-squared:                       1.000
Model:                            OLS   Adj. R-squared:                  1.000
Method:                 Least Squares   F-statistic:                 1.280e+32
Date:                Mon, 10 Jun 2024   Prob (F-statistic):               0.00
Time:                        11:24:36   Log-Likelihood:                 22339.
No. Observations:                 837   AIC:                        -4.467e+04
Df Residuals:                     833   BIC:                        -4.465e+04
Df Model:                           3
Covariance Type:            nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
const                    2.163e-14   8.84e-14      0.245      0.807   -1.52e-13    1.95e-13
ride_count                 -0.4375   8.61e-16  -5.08e+14      0.000      -0.437      -0.437
total_earnings              0.2500   7.59e-17   3.29e+15      0.000       0.250       0.250
avg_wait_before_arrived -8.891e-14   1.67e-14     -5.311      0.000   -1.22e-13   -5.61e-14
==============================================================================
Omnibus:                      337.132   Durbin-Watson:                   0.414
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1665.594
Skew:                           1.800   Prob(JB):                         0.00
Kurtosis:                       8.898   Cond. No.                     1.33e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.33e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
const                      2.162522e-14
ride_count                -4.375000e-01
total_earnings             2.500000e-01
avg_wait_before_arrived   -8.891485e-14
dtype: float64
```

Backup #2.3: Driver Churn Rate? Any predictive indicators for driver churn?