自动化数据分析报告

- Run ID: 20260109_171722_91ea88
- 生成时间: 2026-01-09T17:17:27
- 数据集: chicago_taxi_demo.csv
- 行/列: 200000 / 23
- 本次载入行数(用于分析): 200000

1. 数据概览

1.1 原始数据缺失率最高的列（Top 10）

- Dropoff Census Tract: 37.93%（dtype=float64，role=categorical）
- Pickup Census Tract: 37.49%（dtype=float64，role=categorical）
- Dropoff Community Area: 11.02%（dtype=float64，role=categorical）
- Dropoff Centroid Latitude: 10.66%（dtype=float64，role=numeric）
- Dropoff Centroid Longitude: 10.66%（dtype=float64，role=numeric）
- Dropoff Centroid Location: 10.66%（dtype=object，role=text）
- Pickup Community Area: 8.68%（dtype=float64，role=categorical）
- Pickup Centroid Latitude: 8.67%（dtype=float64，role=numeric）
- Pickup Centroid Longitude: 8.67%（dtype=float64，role=numeric）
- Pickup Centroid Location: 8.67%（dtype=object，role=text）

2. 清洗与特征工程日志

- 去除重复行: 0
- 删除列: 0

2.1 类型转换（共 2 项）

- Trip Start Timestamp -> datetime
- Trip End Timestamp -> datetime

2.2 缺失值填补（共 19 列）

列	方法	填充值
Taxi ID	missing_category	Missing
Trip Seconds	median	600.0
Trip Miles	median	1.2
Pickup Census Tract	missing_category	Missing
Dropoff Census Tract	missing_category	Missing

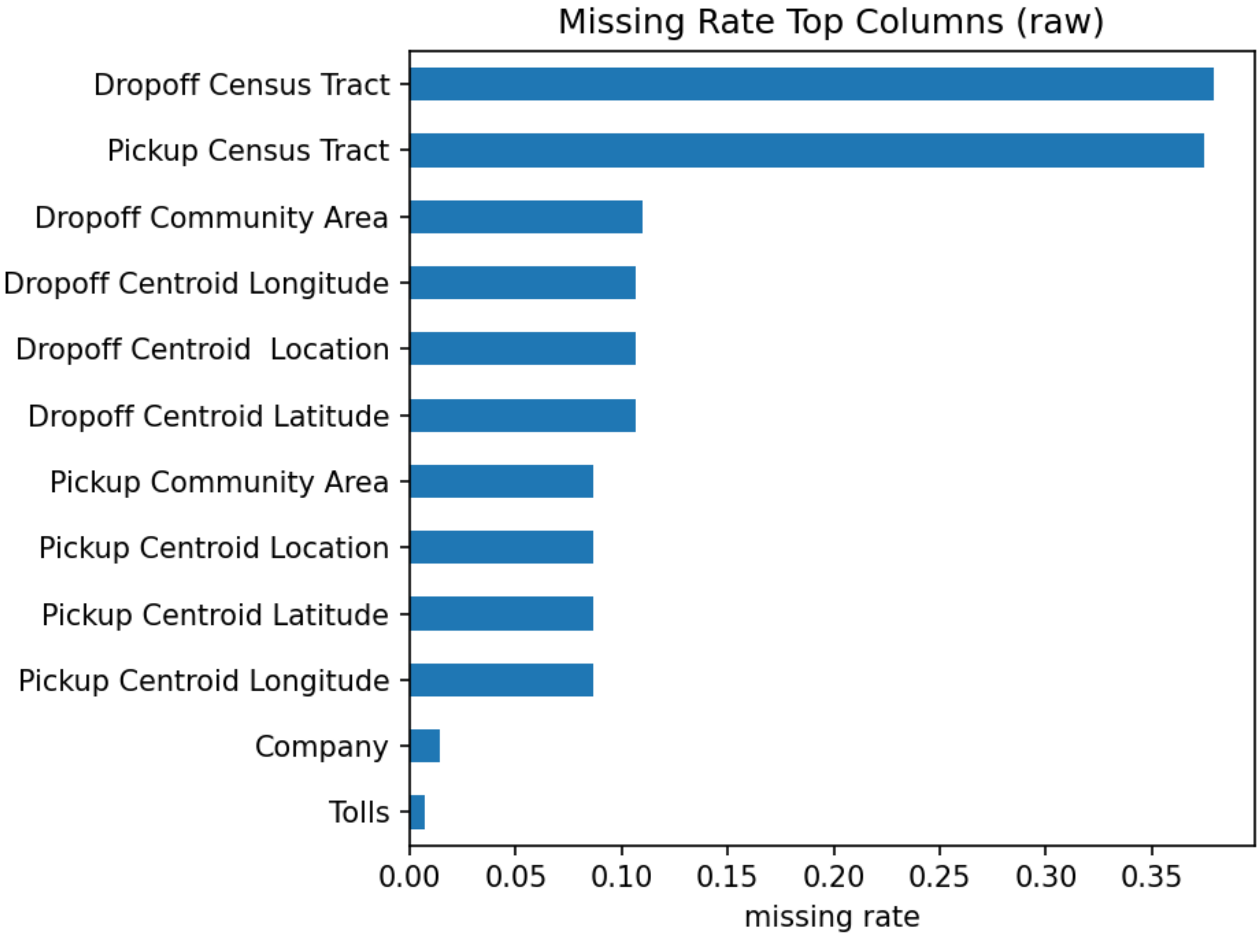
列	方法	填充值
Pickup Community Area	missing_category	Missing
Dropoff Community Area	missing_category	Missing
Fare	median	8.25
Tips	median	0.0
Tolls	median	0.0
Extras	median	0.0
Trip Total	median	10.0
> 仅展示前12列填补记录，完整记录见 analysis.json。		

2.3 新增特征（共 6 个）

- Trip Start Timestamp_year
- Trip Start Timestamp_month
- Trip Start Timestamp_dow
- Trip End Timestamp_year
- Trip End Timestamp_month
- Trip End Timestamp_dow

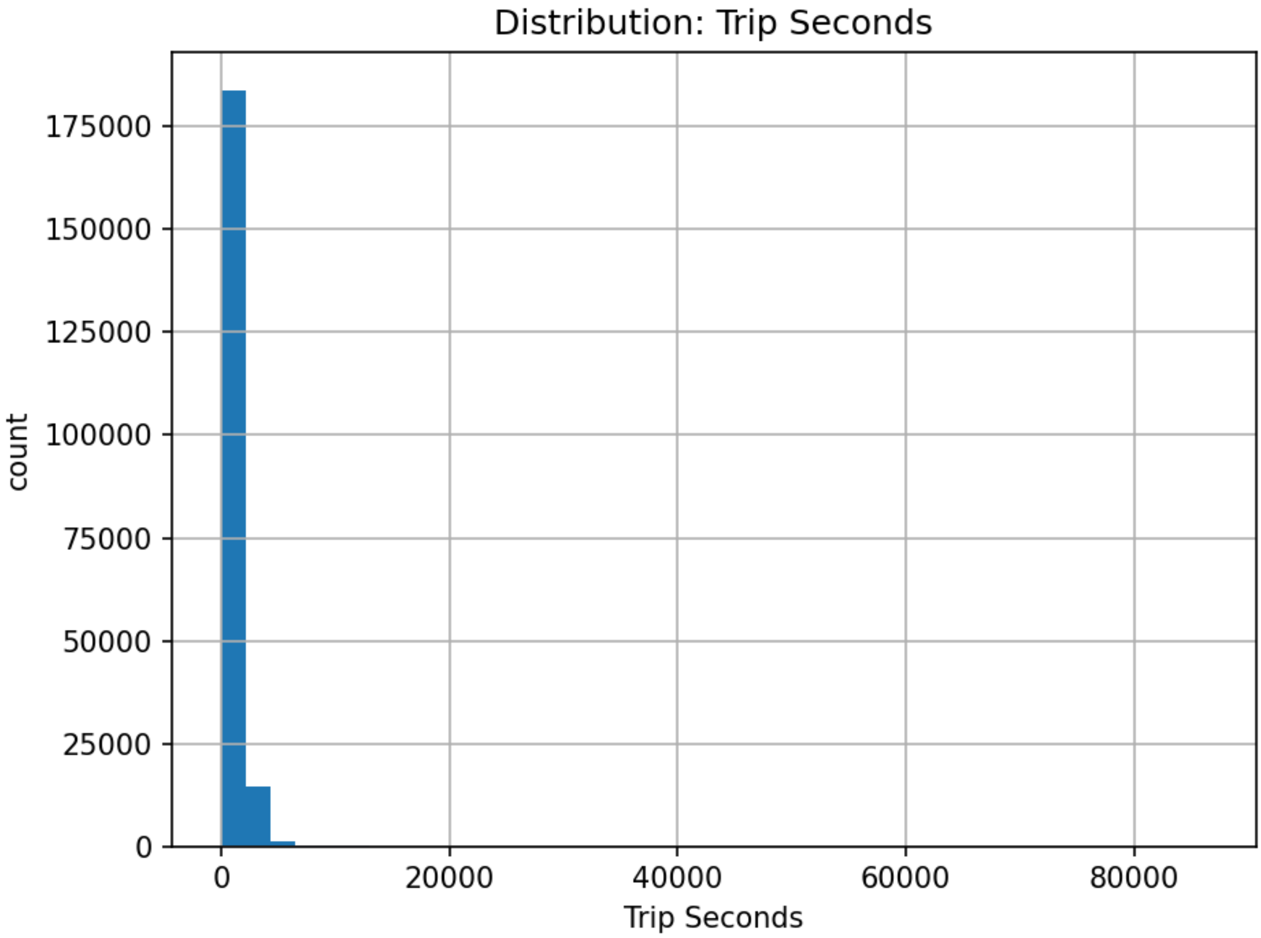
3. 图表

Figure 1：缺失率最高的列(TopK)



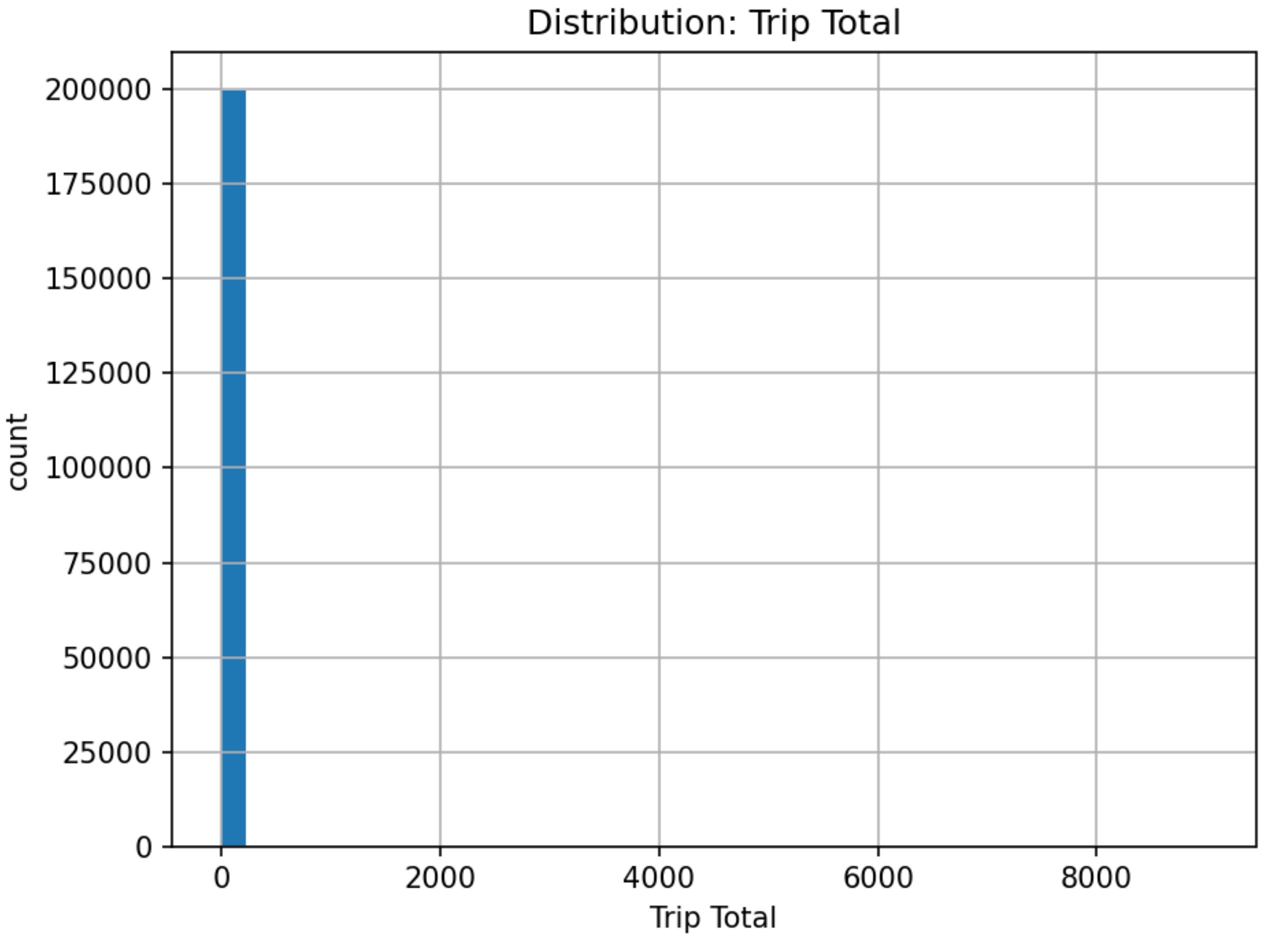
基于原始数据的缺失率统计。

Figure 2：数值列分布：Trip Seconds



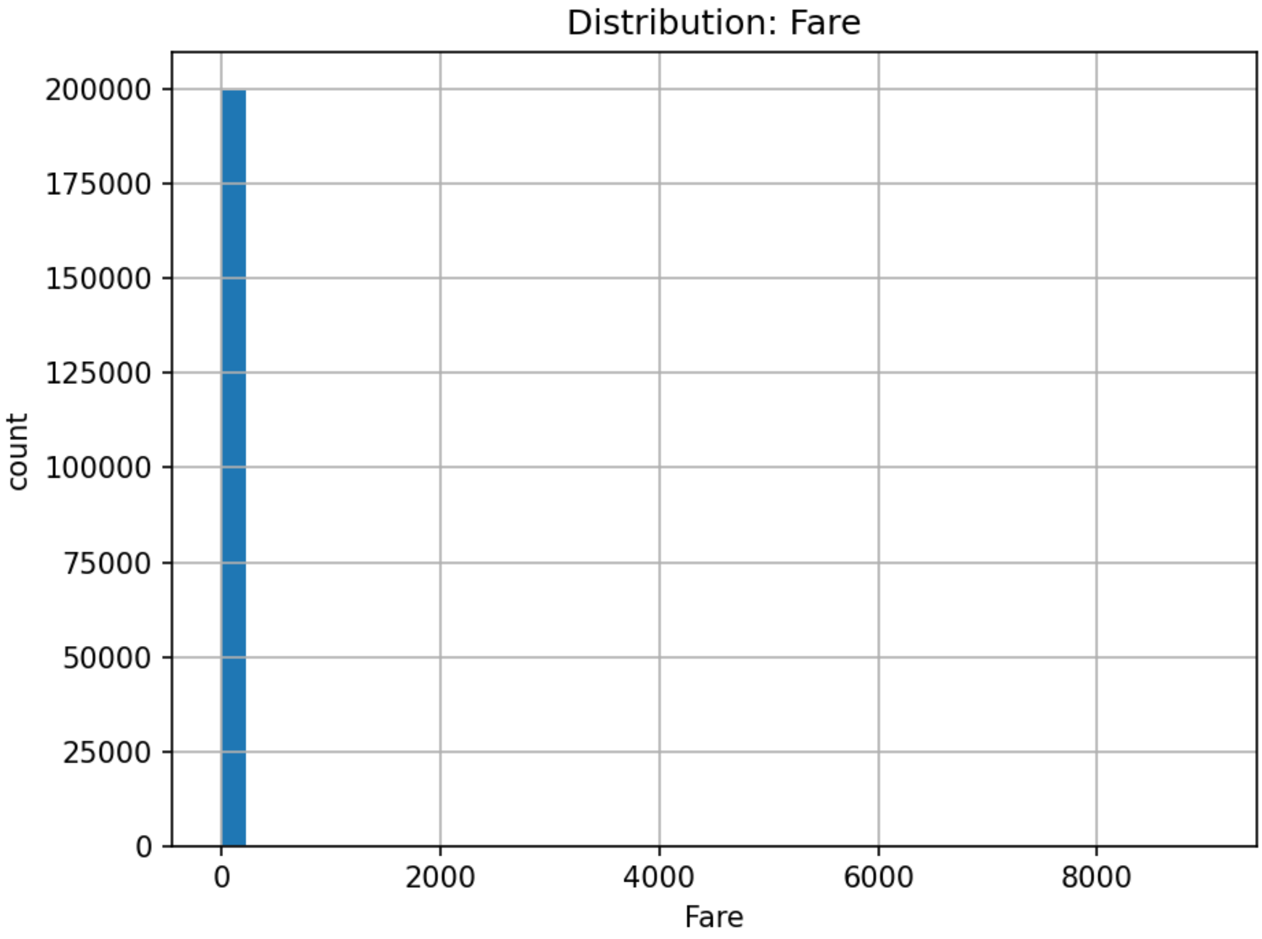
清洗后数据的直方图分布。

Figure 3：数值列分布：Trip Total



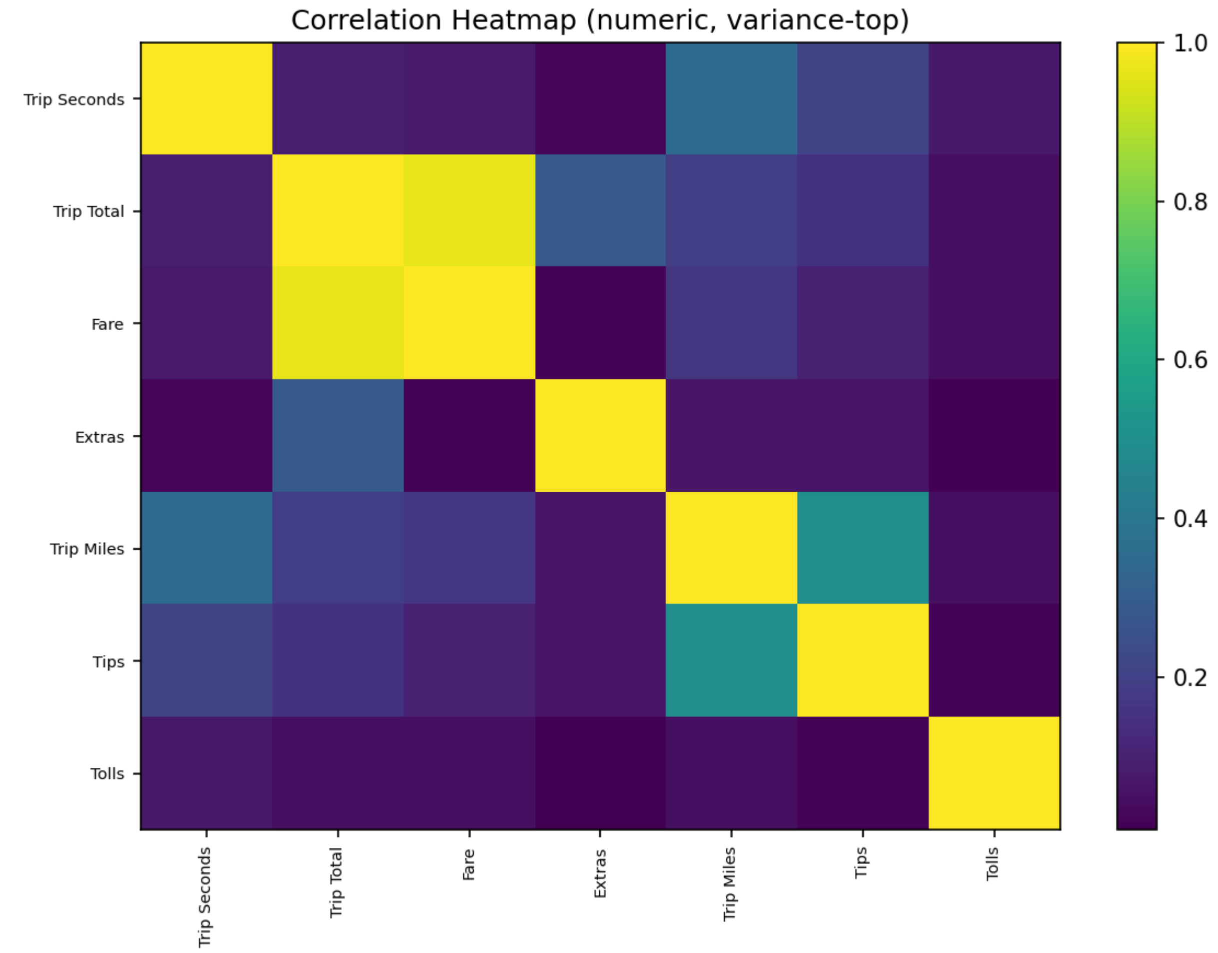
清洗后数据的直方图分布。

Figure 4：数值列分布：Fare



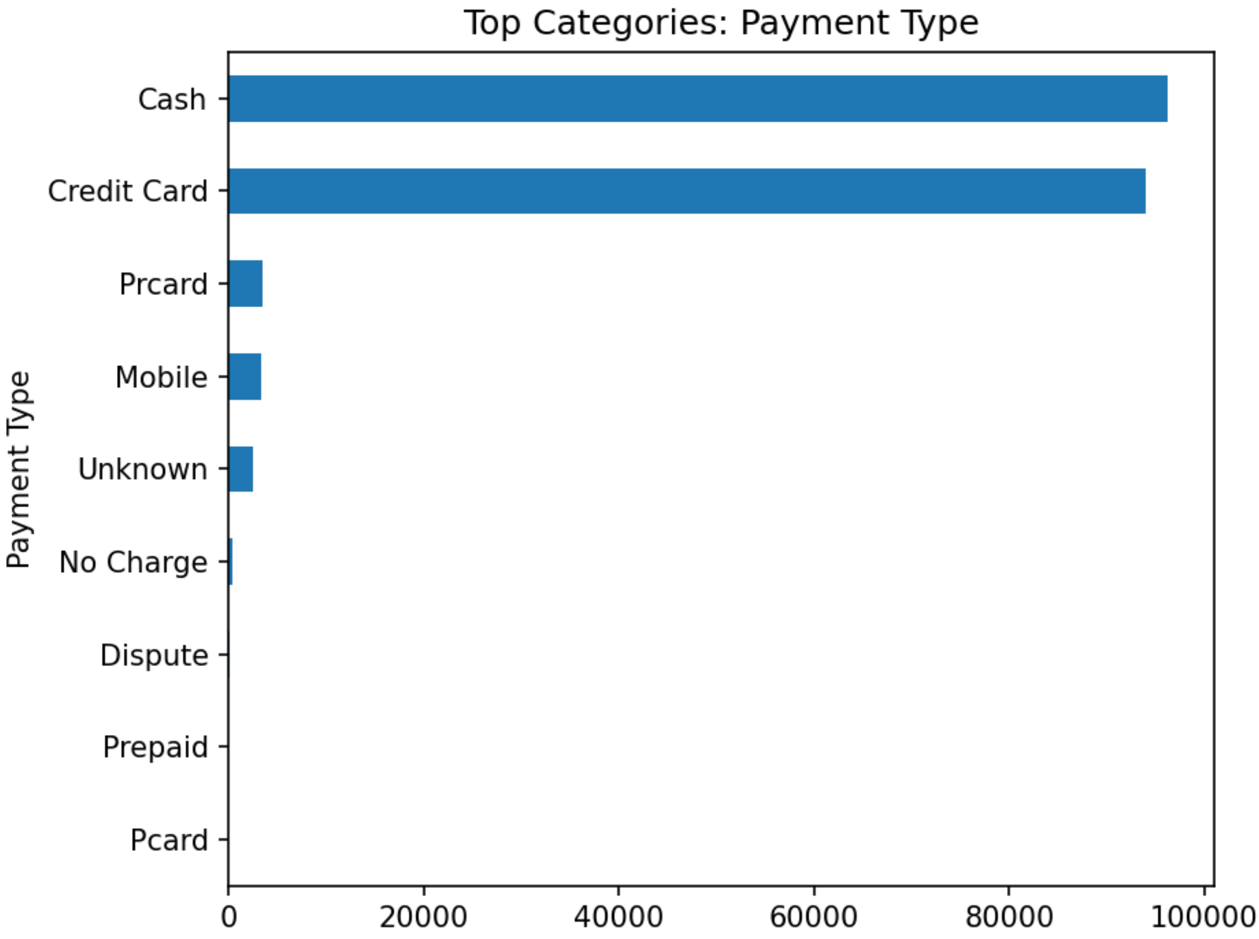
清洗后数据的直方图分布。

Figure 5：数值特征相关性热力图(方差TopN,已过滤经纬度/编码列)



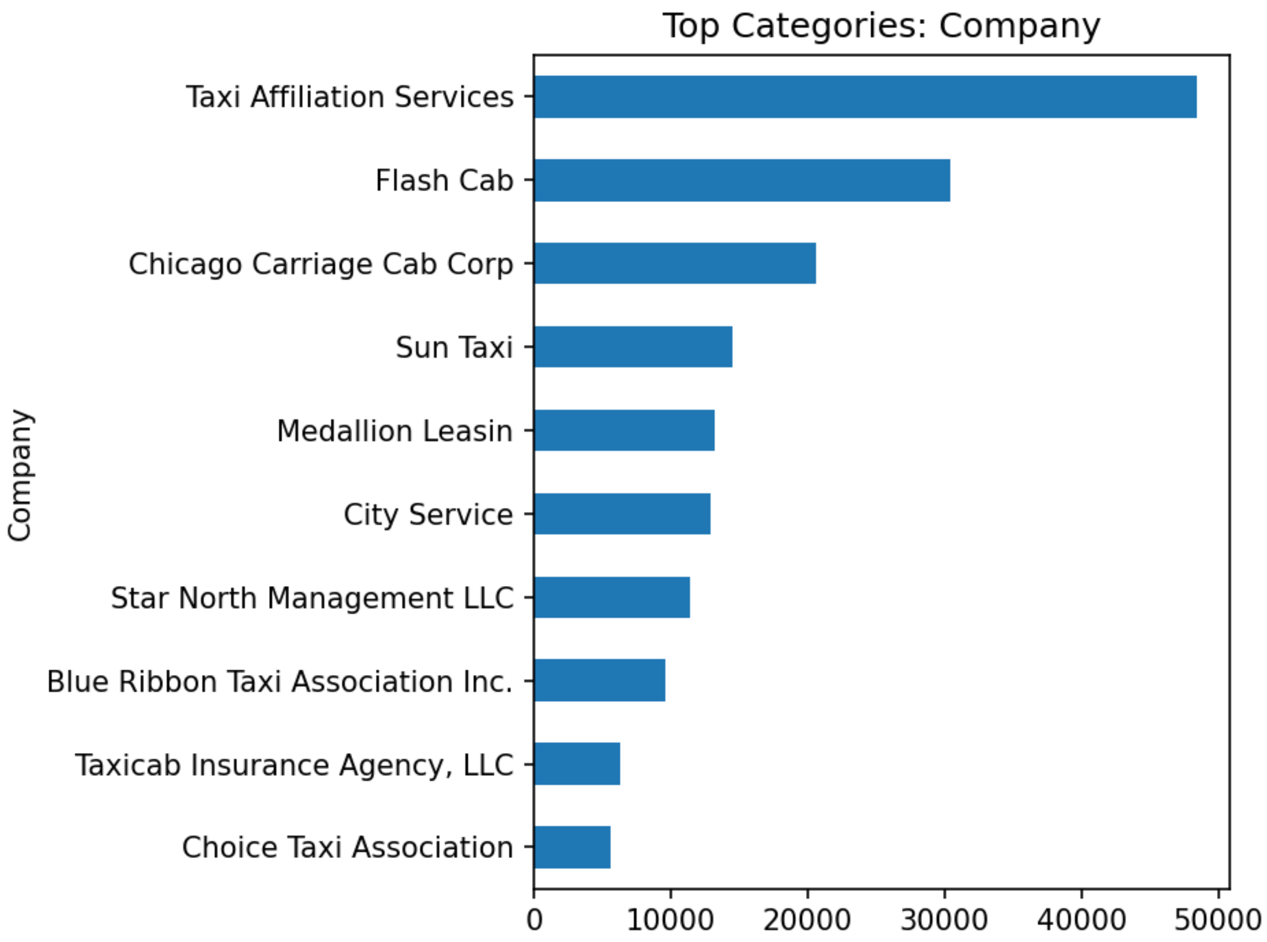
清洗后数值列相关性。列集合按方差TopN选择, 并排除经纬度/编码列。

Figure 6：类别TopN：Payment Type



清洗后类别频次TopN。

Figure 7：类别TopN：Company



清洗后类别频次TopN。

4. 统计检验

- pearsonr：Trip Miles vs Tips，stat=0.4989，p=<1e-300
- 数值-数值相关显著性检验(Pearson,已过滤 trivial total/component 列对)。
- anova：Payment Type vs Trip Seconds，stat=121.8682，p=4.65e-104
- 多组均值差异检验(ANOVA,组数做了上限)。
- chi2：Payment Type vs Company，stat=29887.0889，p=<1e-300
- 类别-类别独立性检验（卡方）。

5. 建模结果

- 目标列：Trip Seconds
- 任务类型：regression
- 模型：LinearRegression
- Train/Test：160000 / 40000
- 指标：R^2=0.1556, MAE=392.5056

5.1 Statsmodels 摘要（节选）

```
                          OLS Regression Results
```

```
==============================================================================
Dep. Variable: Trip Seconds R-squared: 0.191 Model: OLS Adj. R-squared: 0.188 Method: Least
Squares F-statistic: 69.04 Date: Fri, 09 Jan 2026 Prob (F-statistic): 1.23e-213 Time: 17:17:26 Log-
Likelihood: -43524. No. Observations: 5000 AIC: 8.708e+04 Df Residuals: 4982 BIC: 8.720e+04 Df
Model: 17
Covariance Type: nonrobust
==============================================================================
coef std err t P>|t| [0.025 0.975]
```

```
const -3.027e+04 4.63e+04 -0.654 0.513 -1.21e+05 6.05e+04 Trip Miles 104.2053 4.596 22.674 0.000
95.196 113.215 Fare 208.2090 110.029 1.892 0.059 -7.497 423.915 Tips 255.8074 109.936 2.327 0.020
40.284 471.331 Tolls -109.0237 612.046 -0.178 0.859 -1308.904 1090.857 Extras 199.3129 110.094
1.810 0.070 -16.519 415.145 Trip Total -207.8255 110.033 -1.889 0.059 -423.538 7.887 Trip Start
Timestamp_year 15.2133 22.940 0.663 0.507 -29.760 60.187 Trip Start Timestamp_month 10.8234
12.273 0.882 0.378 -13.236 34.883 Trip Start Timestamp_dow 2167.7464 152.637 14.202 0.000
1868.510 2466.983 Trip End Timestamp_dow -2178.7037 152.575 -14.280 0.000 -2477.817 -1879.590
Payment Type_Credit Card -132.9848 62.553 -2.126 0.034 -255.616 -10.354 Payment Type_Dispute
162.7164 845.452 0.192 0.847 -1494.743 1820.175 Payment Type_Mobile -101.5844 171.804 -0.591
0.554 -438.395 235.226 Payment Type_No Charge -132.9125 517.978 -0.257 0.797 -1148.377 882.552
Payment Type_Prcard 96.7451 154.111 0.628 0.530 -205.380 398.870 Payment Type_Prepaid -28.4190
1462.431 -0.019 0.984 -2895.428 2838.590 Payment Type_Unknown 274.9322 184.882 1.487 0.137
-87.518 637.383
==============================================================================
Omnibus: 12820.920 Durbin-Watson: 2.016 Prob(Omnibus): 0.000 Jarque-Bera (JB): 240176441.031
Skew: 28.761 Prob(JB): 0.00 Kurtosis: 1075.165 Cond. No. 4.52e+06
==============================================================================
```

Notes: [1] Standard Errors assume that the covariance matrix of the errors is correctly specified. [2] The condition number is large, 4.52e+06. This might indicate that there are strong multicollinearity or other numerical problems.

5.2 共线性诊断（VIF Top 10）

Trip Total: VIF=156612.39

Fare: VIF=150864.65

Trip End Timestamp_dow: VIF=604.5

Trip Start Timestamp_dow: VIF=604.36

Extras: VIF=390.45

Tips: VIF=333.95

Trip Start Timestamp_year: VIF=43.07

Trip Start Timestamp_month: VIF=39.36

Payment Type_Credit Card: VIF=4.18

Trip Miles: VIF=2.13

6. LLM 洞见

关键洞见

1. `Trip Miles` 与 `Tips` 存在显著正相关（Pearson r=0.4989），提示里程越长，小费越高。
2. 不同 `Payment Type` 的 `Trip Seconds` 均值存在显著差异（ANOVA p=4.65e-104），说明支付方式可能影响行程时长。
3. `Payment Type` 与 `Company` 之间存在强关联（卡方检验 p<1e-300），表明不同公司可能偏好特定支付方式。

数据质量风险

1. `Dropoff Census Tract` 和 `Pickup Census Tract` 缺失率较高（37.93% 和 37.49%），可能影响地理分析的准确性。
2. `Dropoff Centroid Latitude`、`Dropoff Centroid Longitude` 和 `Dropoff Centroid Location` 缺失率均超过10%，需确认数据完整性。

可行动建议

1. 针对高缺失率字段（如 `Dropoff Census Tract`），考虑补充数据或剔除相关分析以避免偏差。
2. 深入分析 `Payment Type` 与 `Company` 的关系，探索是否可优化支付流程或提升用户体验。

7. 额外分析表（Agent 自动补充）

分组汇总：Payment Type（金额相关 Top 12）

Payment Type	count	Fare_mean	Fare_median	Tips_mean	Tips_median	Trip Total_mean	Trip Total_median	tip_rate_mean
Cash	96177	13.7013	7.5000	0.0024	0.0000	14.6149	7.7500	0.0002
Credit Card	93946	16.5929	9.0000	3.7509	2.2000	22.2124	12.0000	0.1689
Prcard	3561	18.8587	18.0000	0.1583	0.0000	19.1647	18.2500	0.0083
Mobile	3394	14.9127	8.7500	2.9098	1.7500	19.2037	11.1000	0.1515
Unknown	2444	18.6564	18.0000	0.0367	0.0000	18.8940	18.2500	0.0019
No Charge	339	13.7513	7.5000	0.3054	0.0000	15.1547	8.2500	0.0202

差异”。

Payment Type	count	Fare_mean	Fare_median	Tips_mean	Tips_median	Trip Total_mean	Trip Total_median	tip_rate_mean
Dispute	122	11.4795	7.7500	0.0000	0.0000	13.7242	8.0000	0.0000
Prepaid	13	18.7885	12.7500	0.0000	0.0000	19.3269	13.0000	0.0000
Pcard	4	8.2500	5.8500	0.0000	0.0000	8.6250	6.3500	0.0000
> 用于补充“不同支付方式在费用/小费/总额上的差异”。								

分组汇总：Payment Type（行程相关 Top 12）

Payment Type	count	Trip Miles_mean	Trip Miles_median	Trip Seconds_mean	Trip Seconds_median
Cash	96177	2.7421	1.0300	810.6204	533.0000
Credit Card	93946	4.3956	1.4000	986.6365	660.0000
Prcard	3561	5.9339	5.3400	1255.19	1094.00
Mobile	3394	4.5440	1.6700	923.8506	636.0000
Unknown	2444	3.7078	1.5000	1195.70	1080.00
No Charge	339	2.5838	0.8000	630.5422	420.0000
Dispute	122	2.6148	1.3000	689.5082	540.0000
Prepaid	13	6.4285	3.4200	924.5385	898.0000
Pcard	4	2.4500	1.2500	450.0000	210.0000
> 用于补充“不同支付方式在里程/时长上的差异”。					

公司汇总：Company（Top 10 by count）

Company	count	Fare_mean	Tips_mean	Trip Total_mean	Trip Miles_mean	Trip Seconds_mean
Taxi Affiliation Services	48380	13.9204	1.7598	17.1212	2.3305	808.0564
Flash Cab	30388	15.9762	1.3653	18.4488	5.0272	1059.14
Chicago Carriage Cab Corp	20640	14.2732	1.9020	17.6785	4.2105	940.3694
Sun Taxi	14529	15.2472	2.2925	19.1888	4.6633	1069.96
Medallion Leasin	13187	15.6578	1.8431	18.9671	3.8838	944.8587
City Service	12865	13.7844	2.0821	17.2961	4.0612	864.5765
Star North Management LLC	11436	13.5160	1.8829	16.6755	3.5356	814.9838
Blue Ribbon Taxi Association Inc.	9632	12.0867	1.5672	14.5915	0.1704	828.8116
Taxicab Insurance Agency, LLC	6295	13.7453	2.0284	17.0257	3.7169	822.8408
Choice Taxi Association	5619	14.6250	2.2601	18.2140	4.0174	877.8284
> 用于补充“头部公司在费用/里程/小费上的差异”。						

时间模式：按小时（Trip Start Timestamp）

_hour	count	Fare_mean	Tips_mean	Trip Total_mean	Trip Miles_mean	Trip Seconds_mean
0.0000	4454.00	14.3209	1.8792	18.0063	3.5642	743.9914
1.0000	3708.00	12.5640	1.5791	15.7698	2.8891	765.4683
2.0000	2754.00	11.8739	1.4059	14.7511	2.7348	662.0194
3.0000	2155.00	12.6139	1.2919	15.1541	2.9056	701.9041
4.0000	1772.00	15.0080	1.5466	17.5685	4.1335	775.0159
5.0000	1654.00	26.8350	2.0833	30.0193	6.1146	931.9570
6.0000	2736.00	18.2864	1.8806	21.0456	5.4869	951.5452
7.0000	5694.00	13.9389	1.5632	16.2640	3.7338	867.2727
8.0000	9055.00	12.6609	1.5053	14.9690	3.0590	856.9595
9.0000	10049.00	14.9897	1.6637	17.9613	3.2274	848.3350
10.0000	9846.00	14.6338	1.6464	17.3201	3.4904	811.3204
11.0000	10639.00	16.3358	1.6959	19.1447	3.6777	790.3296

> 仅展示前 12 行，完整可在代码里导出或提高展示上限。 用于补充“高峰时段/时段费用差异”。

时间模式：按星期（0=周一…6=周日）（Trip Start Timestamp）

_dow	count	Fare_mean	Tips_mean	Trip Total_mean	Trip Miles_mean	Trip Seconds_mean
0.0000	28153.00	15.9309	1.8879	19.3479	3.8167	880.7260
1.0000	32865.00	14.6086	1.8789	17.6968	3.4566	935.0830
2.0000	31312.00	14.7910	1.8671	17.7979	3.5205	882.1672
3.0000	30738.00	15.3340	1.9357	18.4563	3.6673	964.9879
4.0000	37721.00	15.0512	1.6888	18.1012	3.4363	918.9575
5.0000	21129.00	14.2401	1.4296	17.1080	3.2151	825.0614
6.0000	18082.00	17.3982	2.0189	21.2343	4.5356	915.2682

_dow	count	Fare_mean	Tips_mean	Trip Total_mean	Trip Miles_mean	Trip Seconds_mean
> 用于补充“工作日 vs 周末差异”。						