

自动化数据分析报告

- Run ID: 20251221_093948_d563d8
- 生成时间: 2025-12-21T09:39:55
- 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

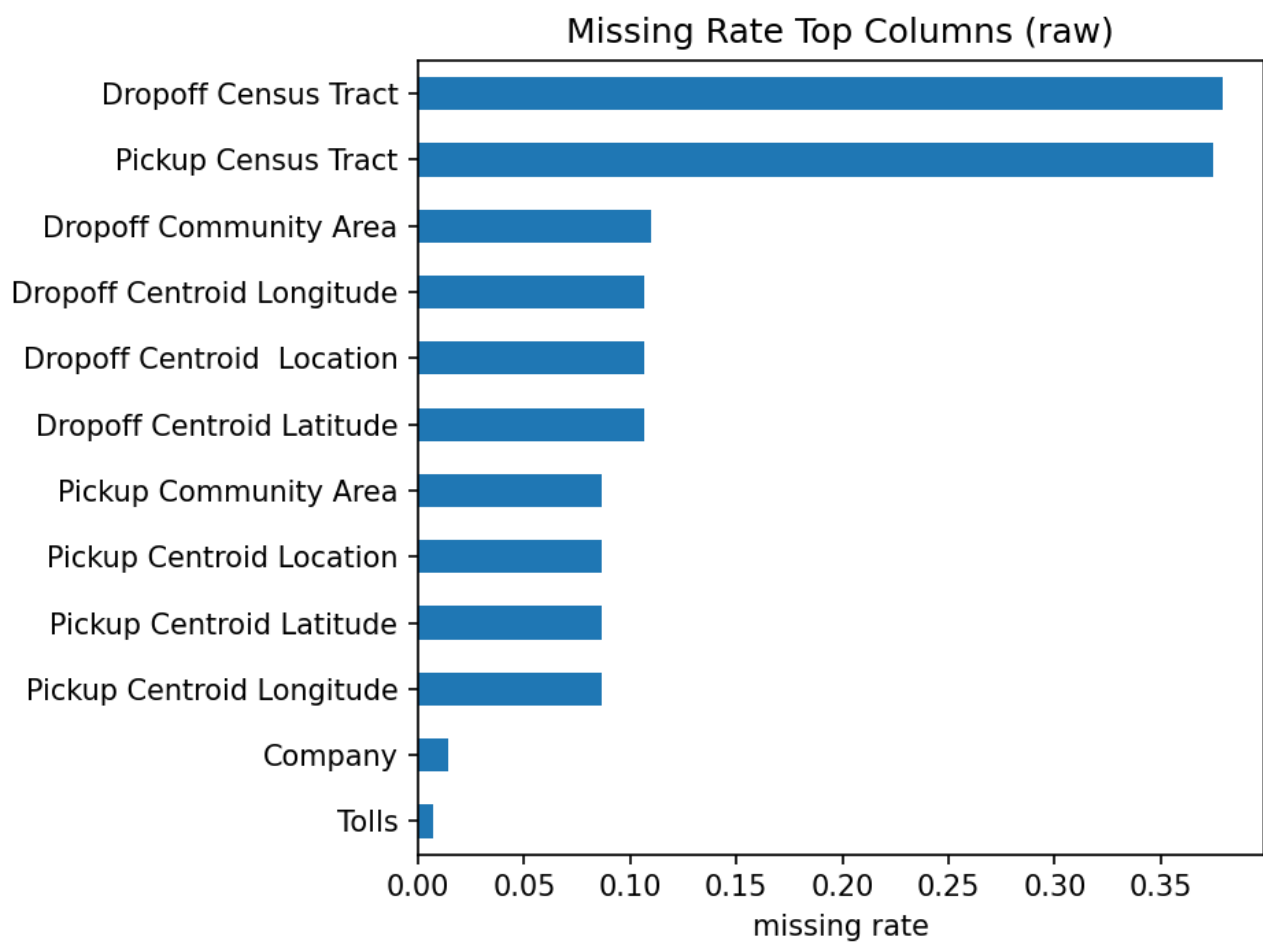
列	方法	填充值
Pickup Community Area	missing_category	Missing
Dropoff Community Area	missing_category	Missing
Fare	median	8.25
Tips	median	0.0
Tolls	median	0.0
Extras	median	0.0
Trip Total	median	10.0
> 仅展示前12列填补记录，完整记录见 analysis.json。		

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

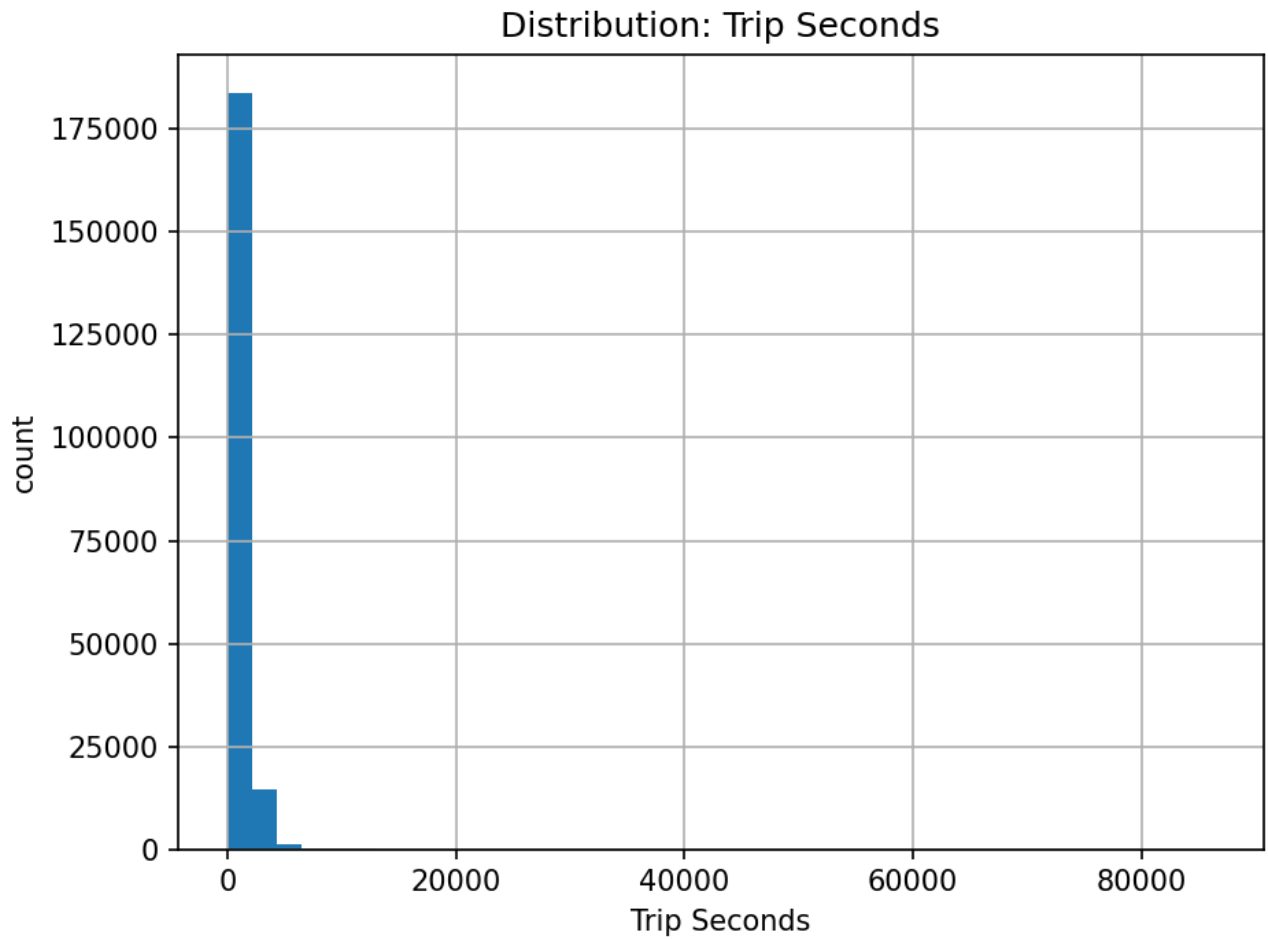
3. 图表

Figure 1： 缺失率最高的列(TopK)



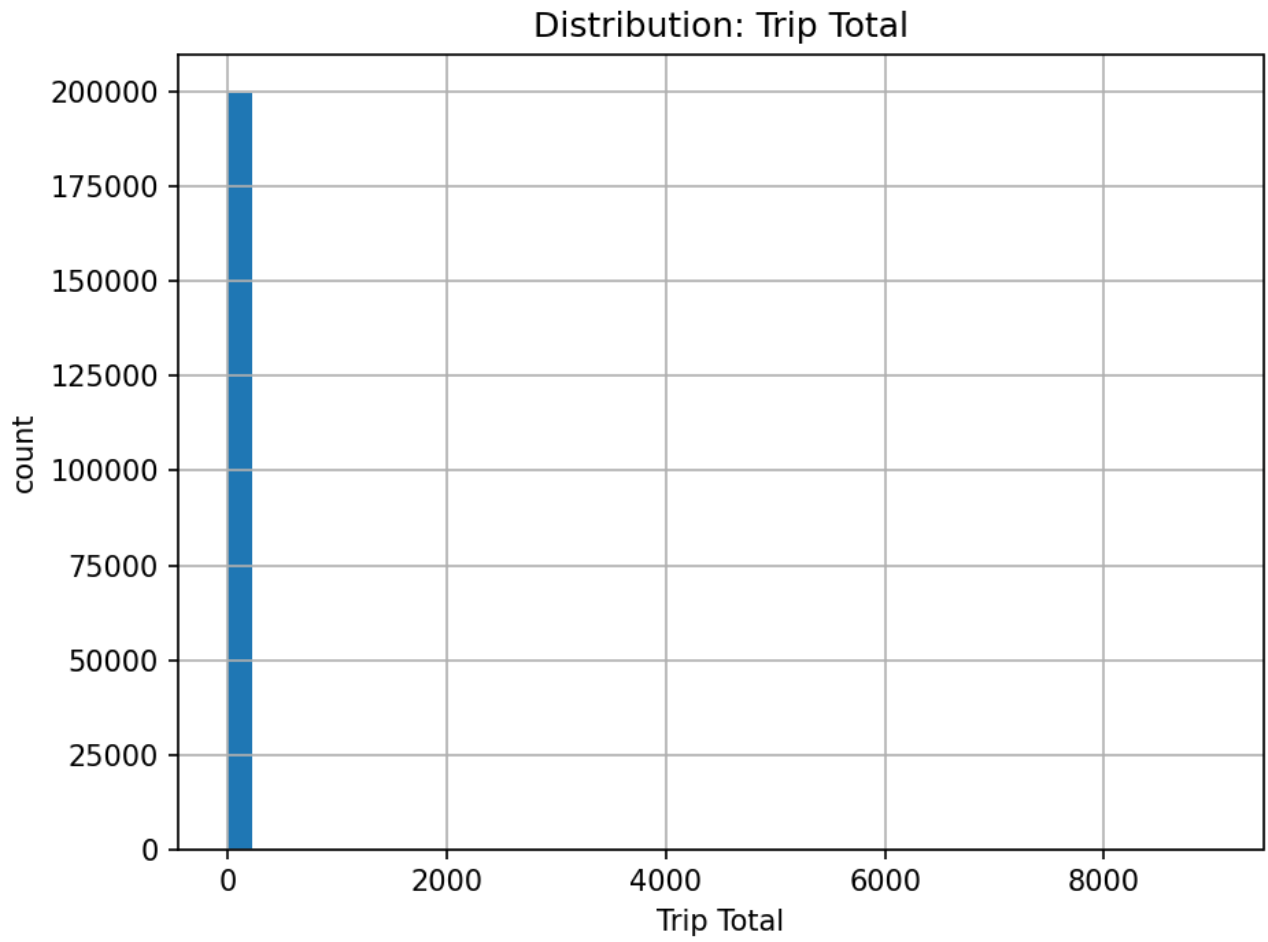
基于原始数据的缺失率统计。

Figure 2: 数值列分布: Trip Seconds



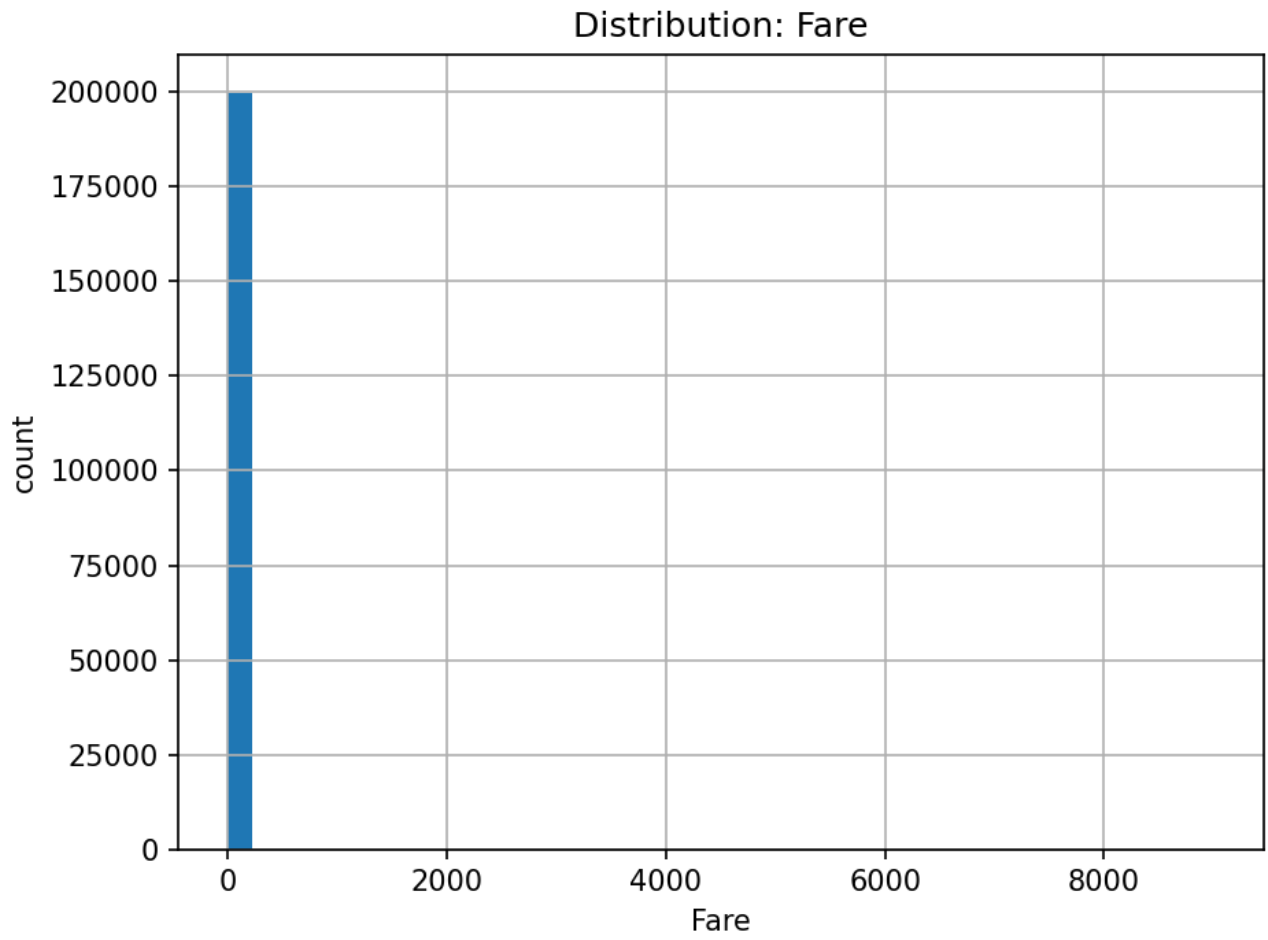
清洗后数据的直方图分布。

Figure 3: 数值列分布: Trip Total



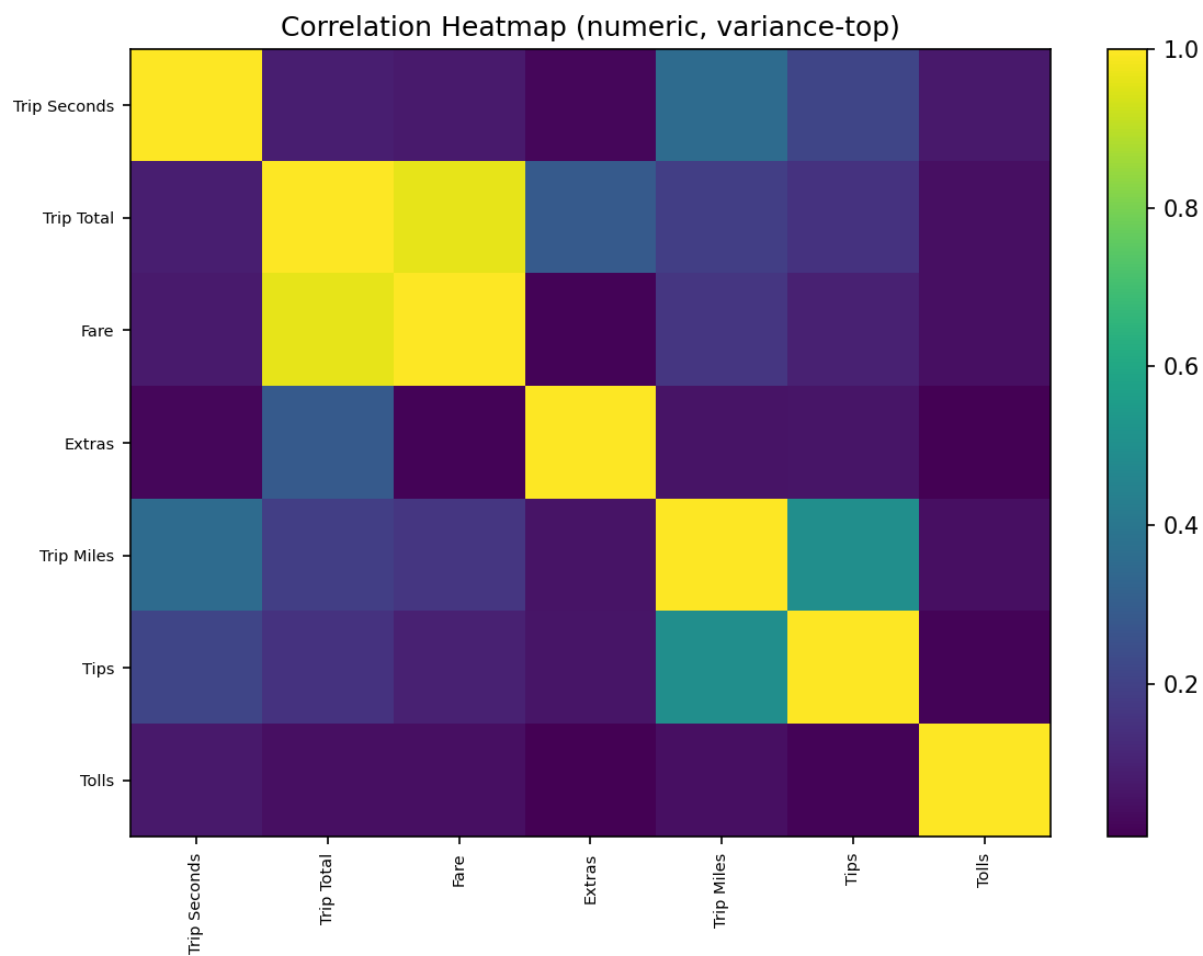
清洗后数据的直方图分布。

Figure 4: 数值列分布: Fare



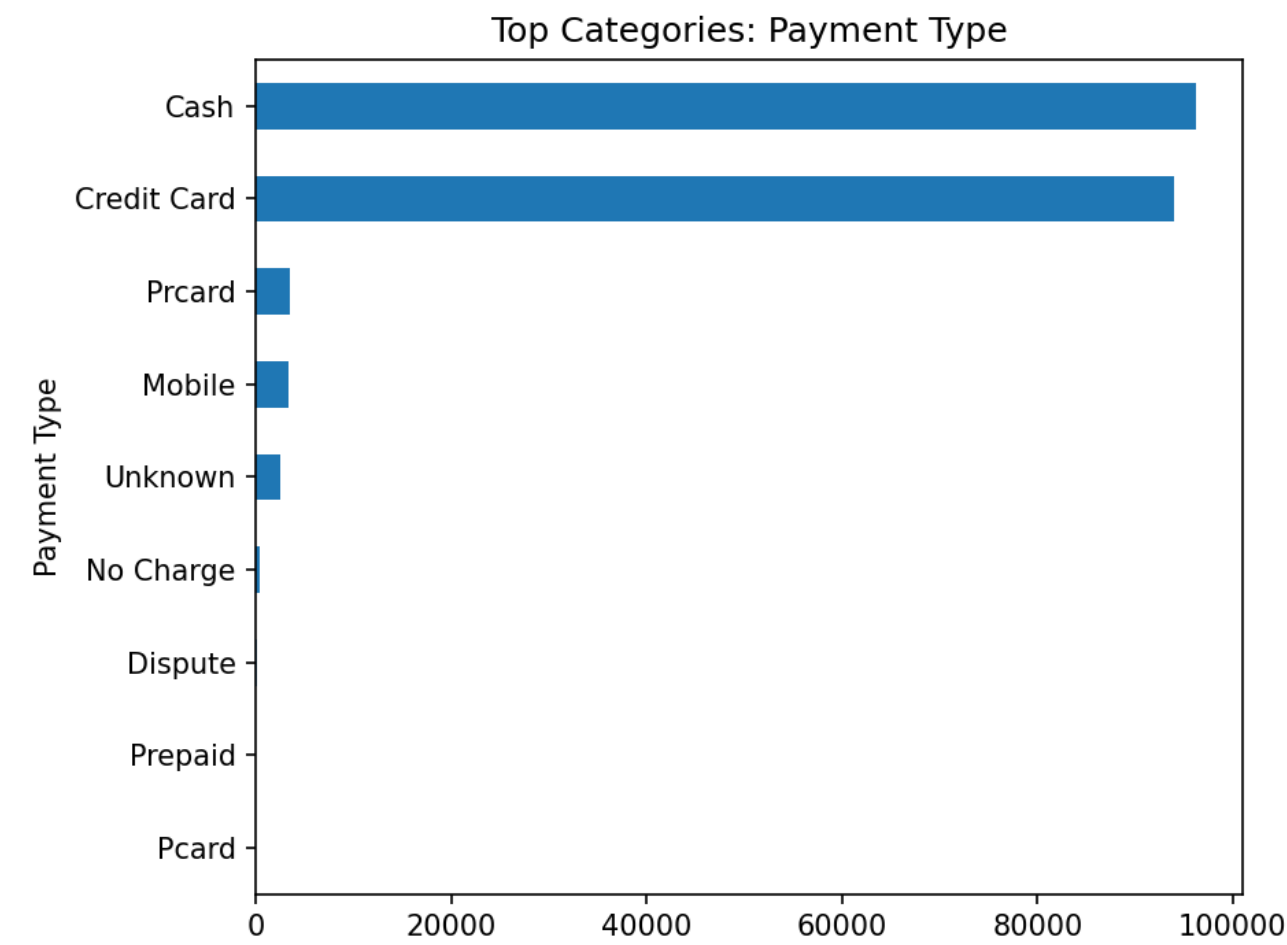
清洗后数据的直方图分布。

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



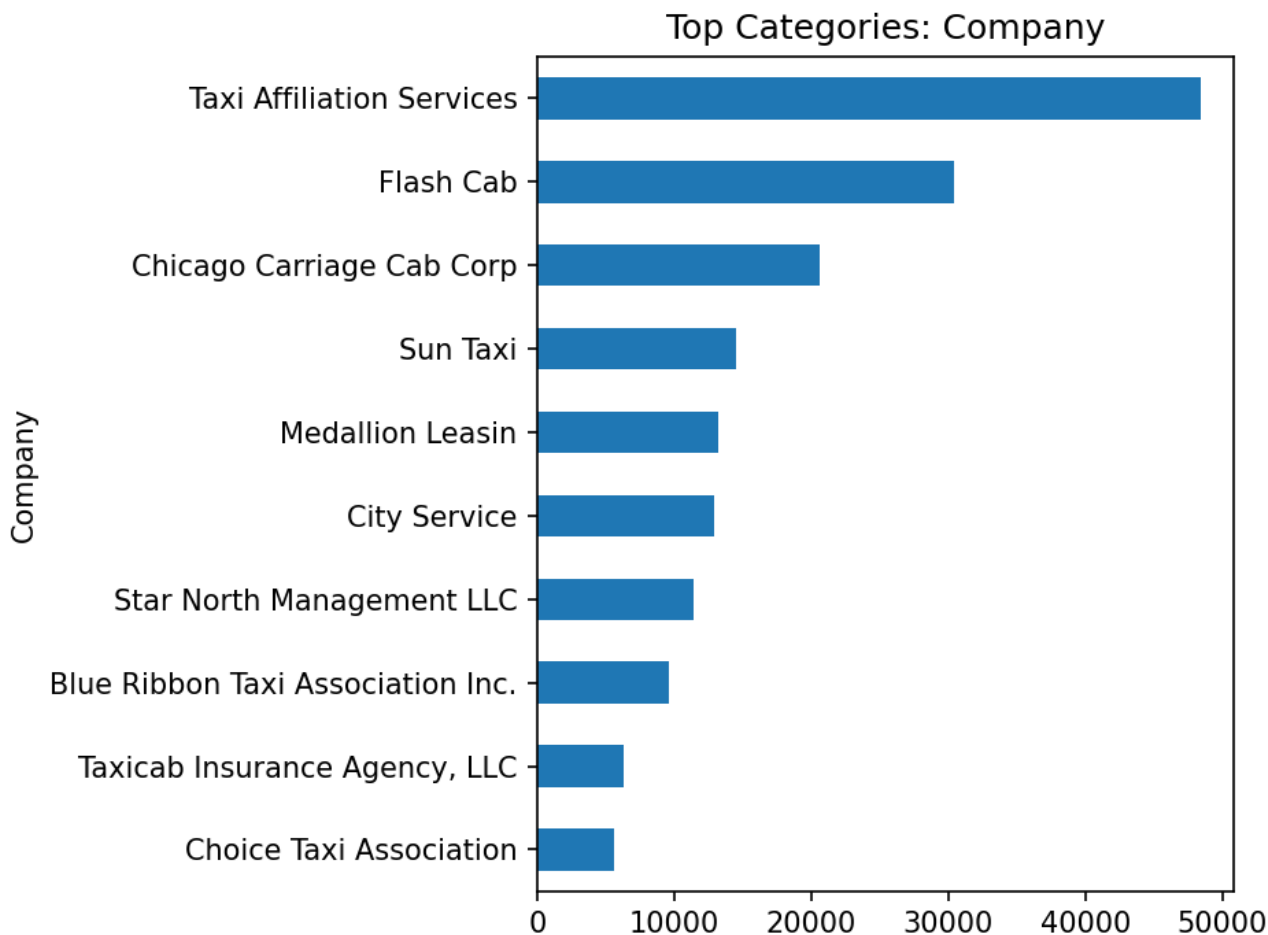
清洗后数值列相关性。列集合按方差TopN选择, 并排除经纬度/编码列。

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, $\text{stat}=0.4989$, $p=<1e-300$
- 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, $\text{stat}=121.8682$, $p=4.65e-104$
- 多组均值差异检验(ANOVA,组数做了上限)。
- chi2: Payment Type vs Company, $\text{stat}=29887.0889$, $p=<1e-300$
- 类别-类别独立性检验 (卡方) 。

5. 规则洞见 (无LLM)

- 缺失率最高的列是 Dropoff Census Tract ($\text{missing_rate}=37.93\%$) , 建议确认采集流程或考虑剔除/重采样。
- 统计检验 pearsonr 显示 Trip Miles 与 Tips 存在显著关系 ($p=<1e-300$) 。
- 统计检验 anova 显示 Payment Type 与 Trip Seconds 存在显著关系 ($p=4.65e-104$) 。
- 统计检验 chi2 显示 Payment Type 与 Company 存在显著关系 ($p=<1e-300$) 。
- Trip Miles 与 Tips 存在显著正相关, 可优化计费或服务策略。

- 不同 Payment Type 的 Trip Seconds 差异显著，需关注支付方式对行程时长的影响。
- Payment Type 与 Company 显著相关，可能涉及业务合作或运营模式差异。
- 建议进一步分析 Payment Type 对服务效率的影响，优化资源配置。
- 数据中缺失率较高的 Dropoff Census Tract 可能影响区域分析，需谨慎处理。
- 时间特征已提取，可用于分析时段与业务表现的关系。