Card Fraud Data: Fraud detection analysis based on merchant category, location, transaction time, and amount

Individual assignment

Kristina Kazlauskaitė

2025

Task overview

- **Dataset**: Card fraud data (1.2M+ transaction records)
- **Goal**: identify fraud characteristics using PySpark for big data processing. To examine fraud patterns across multiple features, including merchant categories, transaction amounts, geographic distribution, and time of day

Method and approach

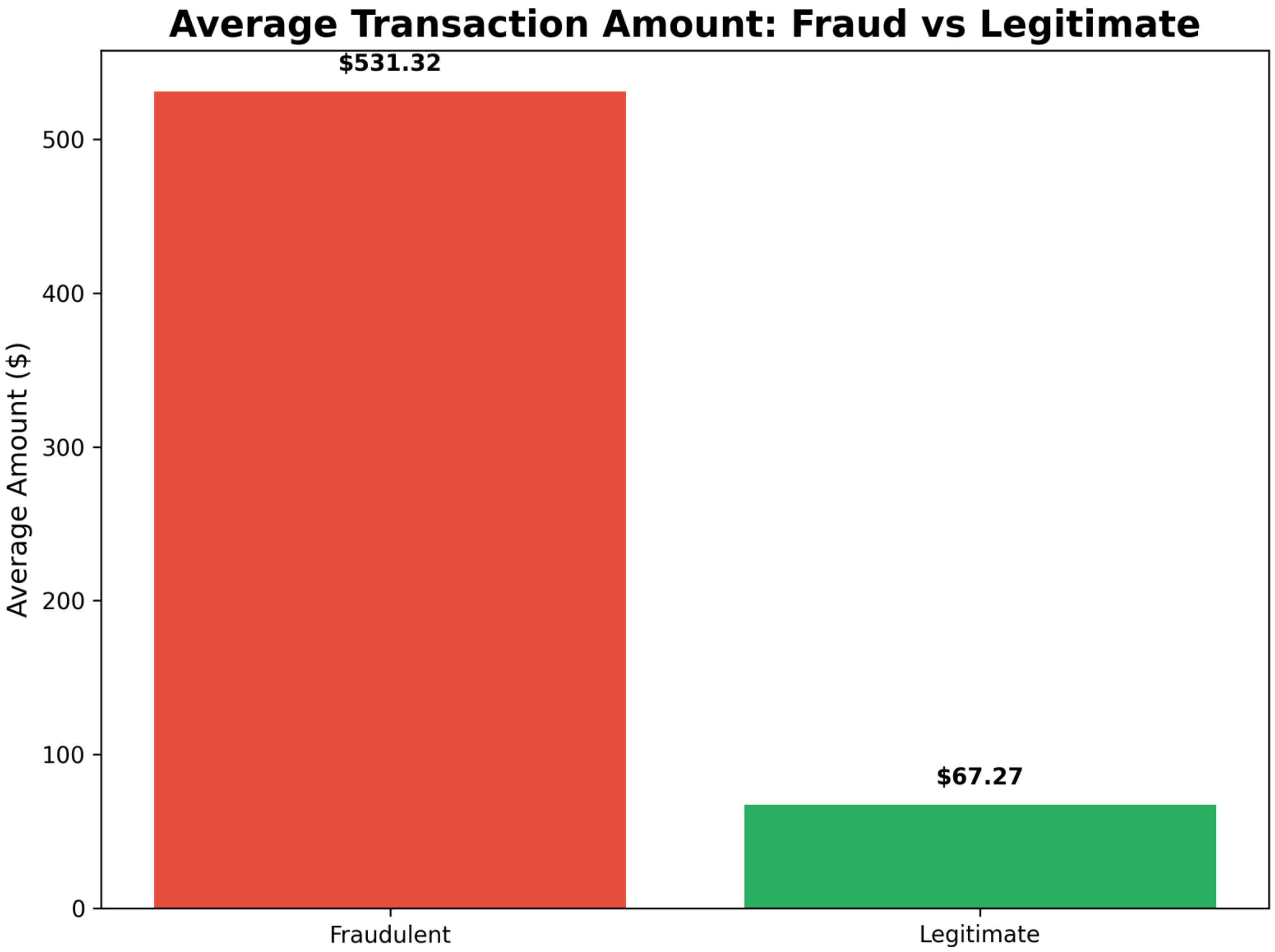
- **Parallel data processing** with PySpark DataFrames
- **Statistical aggregation** and grouping
- **Time-series analysis** for temporal patterns
- **Geographic analysis** for location-based insights
- **Data visualization** for result presentation

Implementation

```
fraud_analysis.py
├── Java setup configuration
├── Spark session initialization
├── Data loading and preprocessing
├── Fraud analysis functions:
│   ├── analyze_fraud_by_category()
│   ├── analyze_fraud_amounts()
│   ├── analyze_geographic_patterns()
│   └── analyze_time_patterns()
├── Visualization generation
└── Summary table generation
```

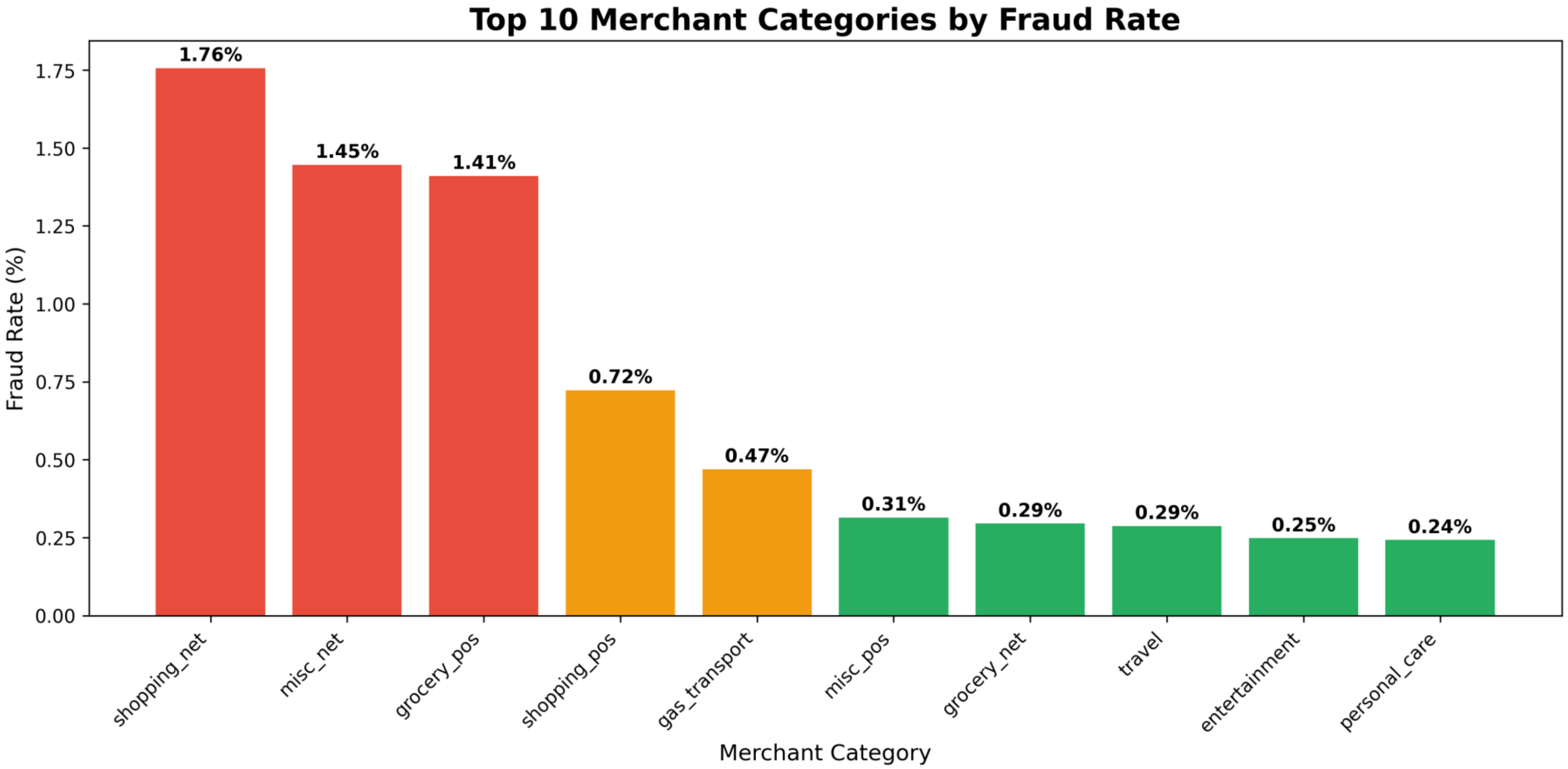
Results I

- Fraud rate: 0.58%
- Fraudulent transactions are 8 times larger than legitimate ones.



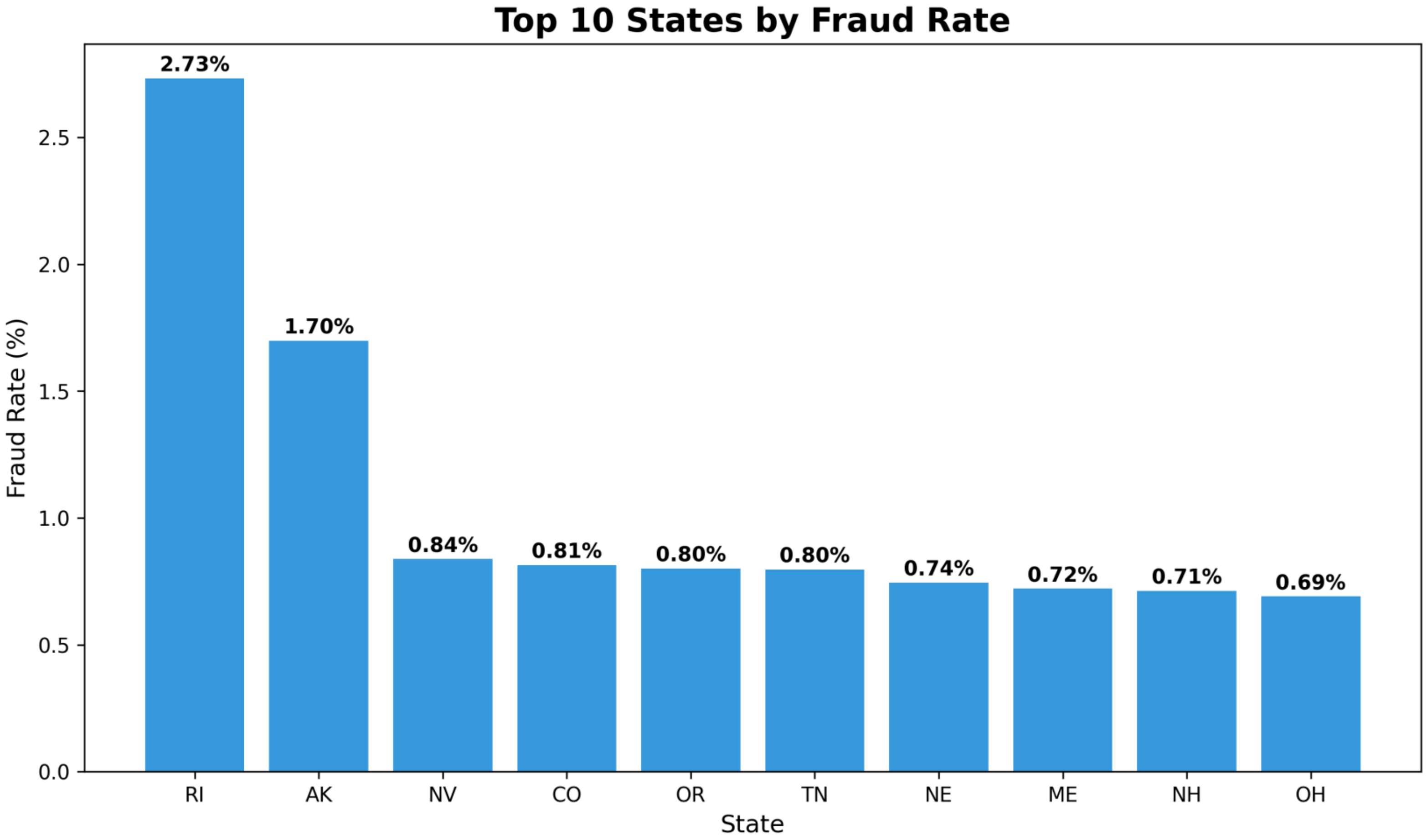
Results II

- Highest risk categories: Shopping NET, Misc NET, Grocery POS



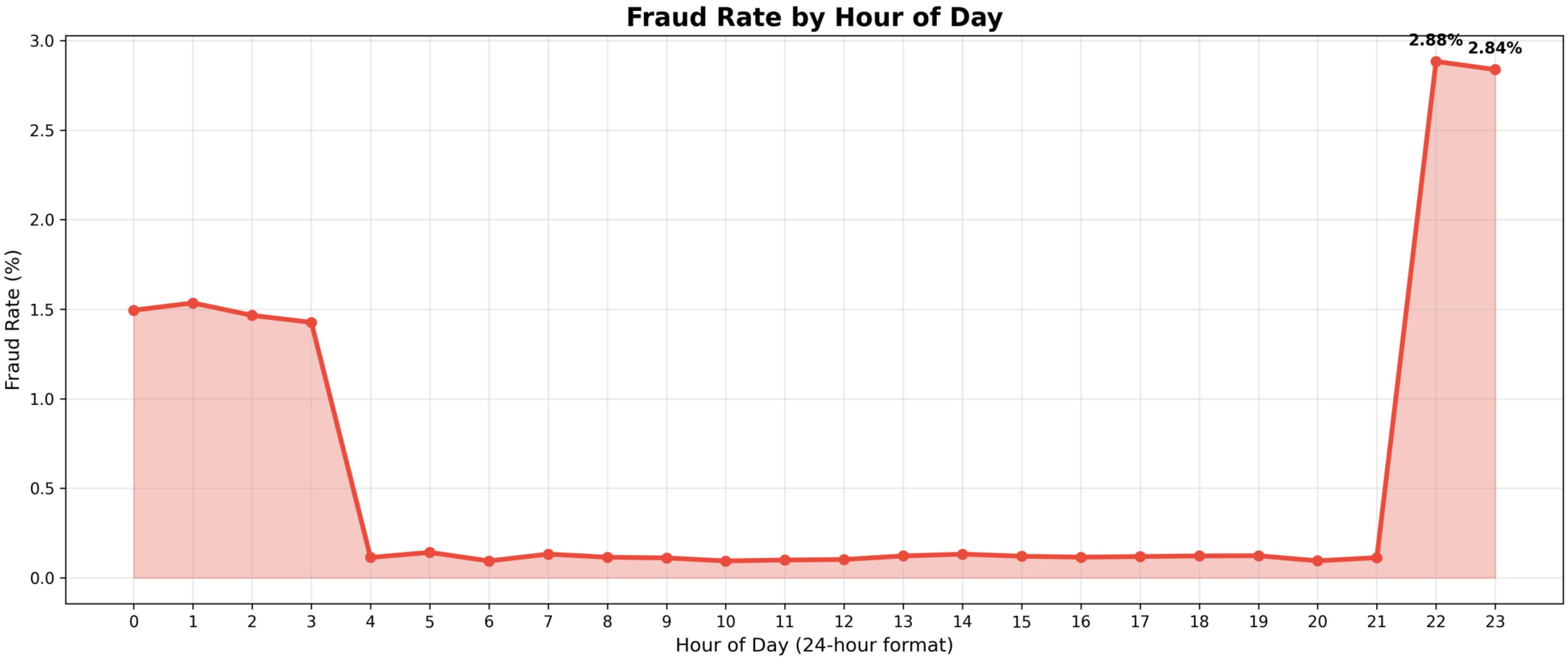
Results III

- Top states by fraud rate: Rhode Island, Alaska, Nevada



Results IV

- Peak fraud hours: 22:00-23:00


Fraud Rate by Hour of Day
Fraud Rate (%)
Hour of Day (24-hour format)