INSURANCE CLAIM ANALYSIS

Presented by: Sarah Moin

PROJECT OVERVIEW

- Objective: Analyze insurance claims and predict claim severity.
- Importance: Improve risk assessment, reduce fraud, and optimize operations.
- Tools Used: Python, Pandas, Matplotlib, Scikit-learn.

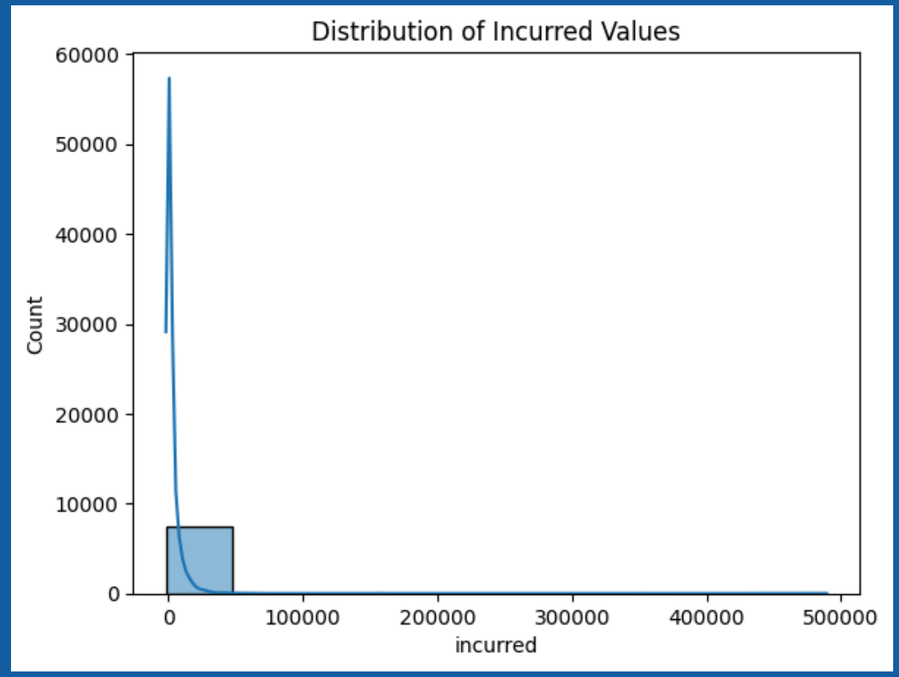
PROBLEM STATEMENT

- Challenges in understanding claim behaviour.
- Need for identifying predictive features.
- Ensuring data quality and readiness for weekly updates.

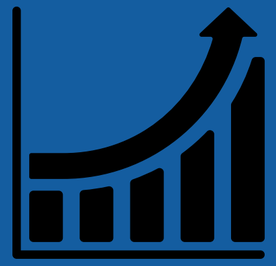
Data Exploration

Understanding the Dataset

- Dataset contains important columns: claim_count, incurred, vehicle_make, net_earned_premium, etc.
- Initial Observations:
    a. Missing values in Incurred.
    b. Outliers in Incurred (negative values).
- Visualization: Visualizations for some of the columns are shown in the next pages.



Data Statistics



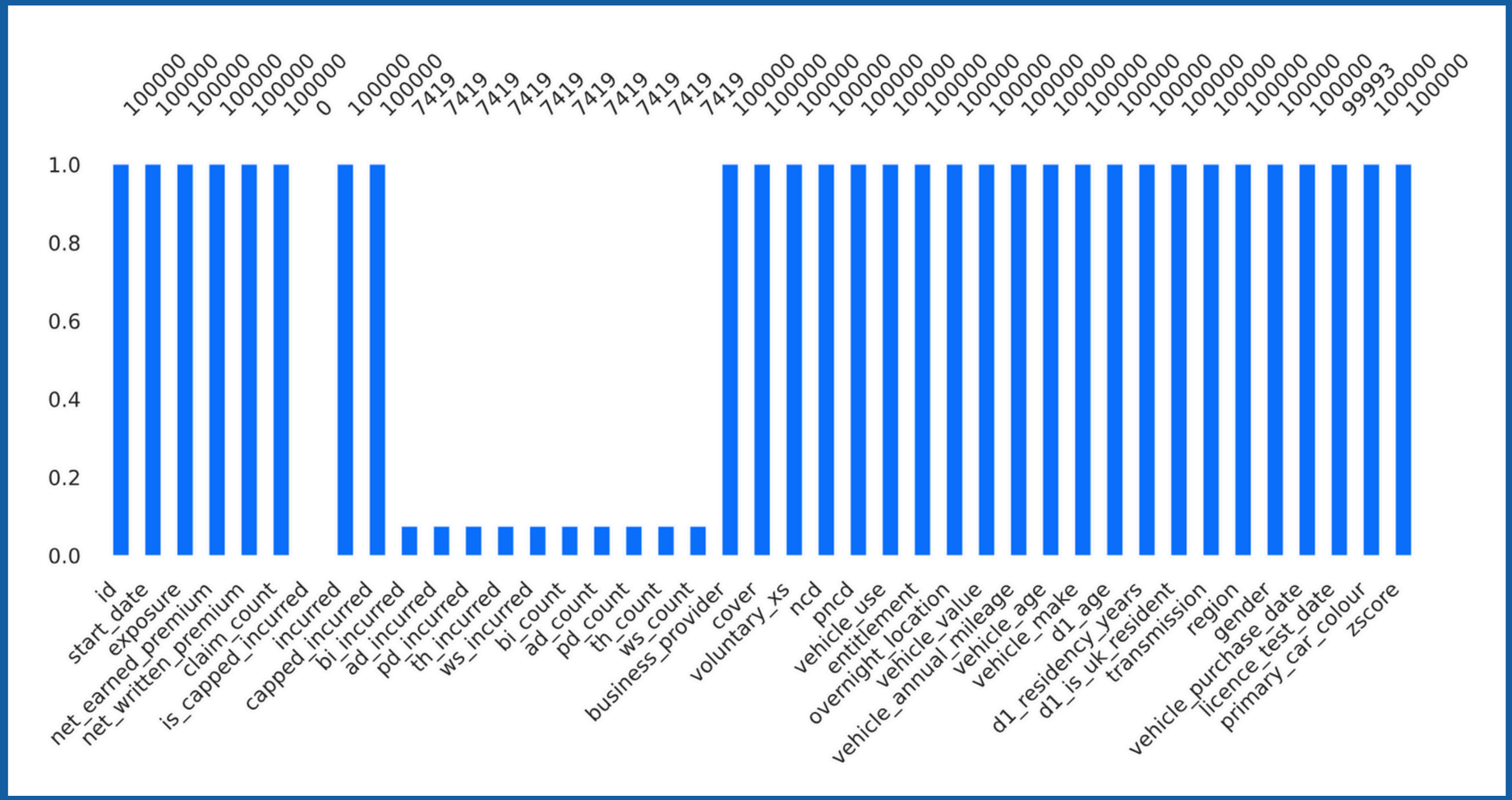
Dataset statistics

Number of variables	41
Number of observations	100000
Missing cells	1025817
Missing cells (%)	25.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	88.6 MiB
Average record size in memory	928.6 B

Variable types

Numeric	19
DateTime	3
Categorical	15
Unsupported	1
Text	2
Boolean	1

Missing Value Plot



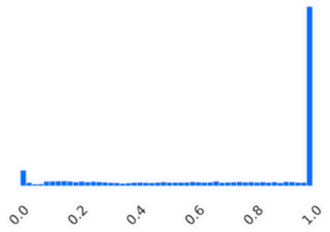
Exposure & Net premium earned

exposure

Real number (ℝ)

High correlation

Distinct	1046	**Minimum**	0.00273224
Distinct (%)	1.0%	**Maximum**	1
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	0.71959474	**Memory size**	781.4 KiB

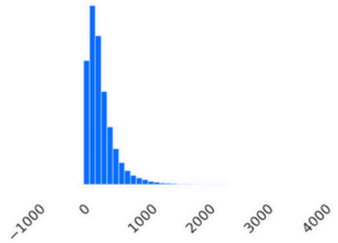


net_earned_premium

Real number (ℝ)

High correlation Zeros

Distinct	84182	**Minimum**	-1134.753
Distinct (%)	84.2%	**Maximum**	3912.24
Missing	0	**Zeros**	3579
Missing (%)	0.0%	**Zeros (%)**	3.6%
Infinite	0	**Negative**	39
Infinite (%)	0.0%	**Negative (%)**	< 0.1%
Mean	272.43219	**Memory size**	781.4 KiB



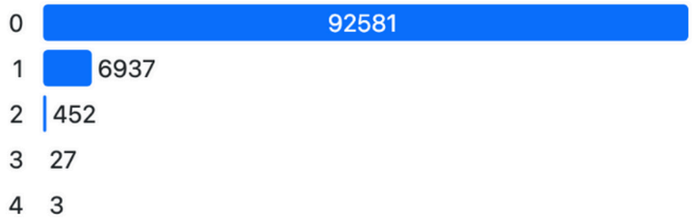
Claim count & Incurred

claim_count

Categorical

Imbalance

Distinct	5
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	5.5 MiB

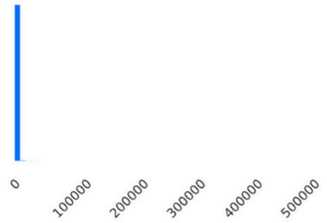


incurred

Real number (ℝ)

High correlation Skewed Zeros

Distinct	3121	**Minimum**	-1854
Distinct (%)	3.1%	**Maximum**	489297
Missing	0	**Zeros**	93077
Missing (%)	0.0%	**Zeros (%)**	93.1%
Infinite	0	**Negative**	1073
Infinite (%)	0.0%	**Negative (%)**	1.1%
Mean	207.49274	**Memory size**	781.4 KiB



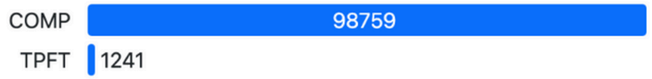
Cover & Vehicle use

cover

Categorical

Imbalance

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	5.8 MiB



vehicle_use

Categorical

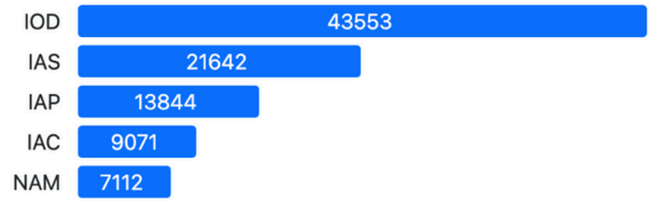
Distinct	3
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	5.5 MiB

Entitlement & Vehicle value

entitlement

Categorical

Distinct	6
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	5.7 MiB

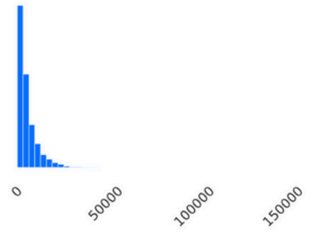


vehicle_value

Real number (ℝ)

High correlation

Distinct	5702	**Minimum**	1
Distinct (%)	5.7%	**Maximum**	163995
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	5525.1672	**Memory size**	781.4 KiB



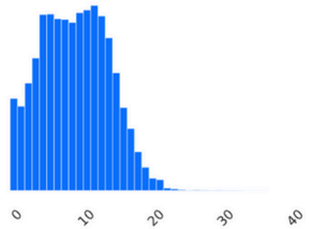
Vehicle age & Vehicle make

vehicle_age

Real number (ℝ)

High correlation | Zeros

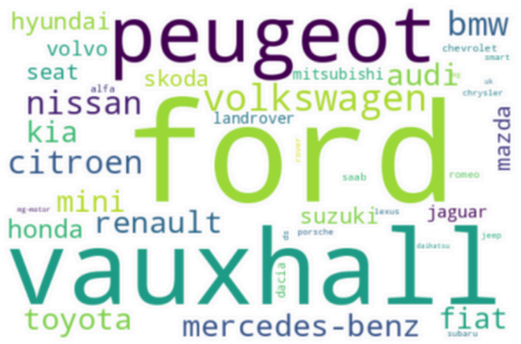
Distinct	40	**Minimum**	-1
Distinct (%)	< 0.1%	**Maximum**	41
Missing	0	**Zeros**	3339
Missing (%)	0.0%	**Zeros (%)**	3.3%
Infinite	0	**Negative**	336
Infinite (%)	0.0%	**Negative (%)**	0.3%
Mean	8.37847	**Memory size**	781.4 KiB



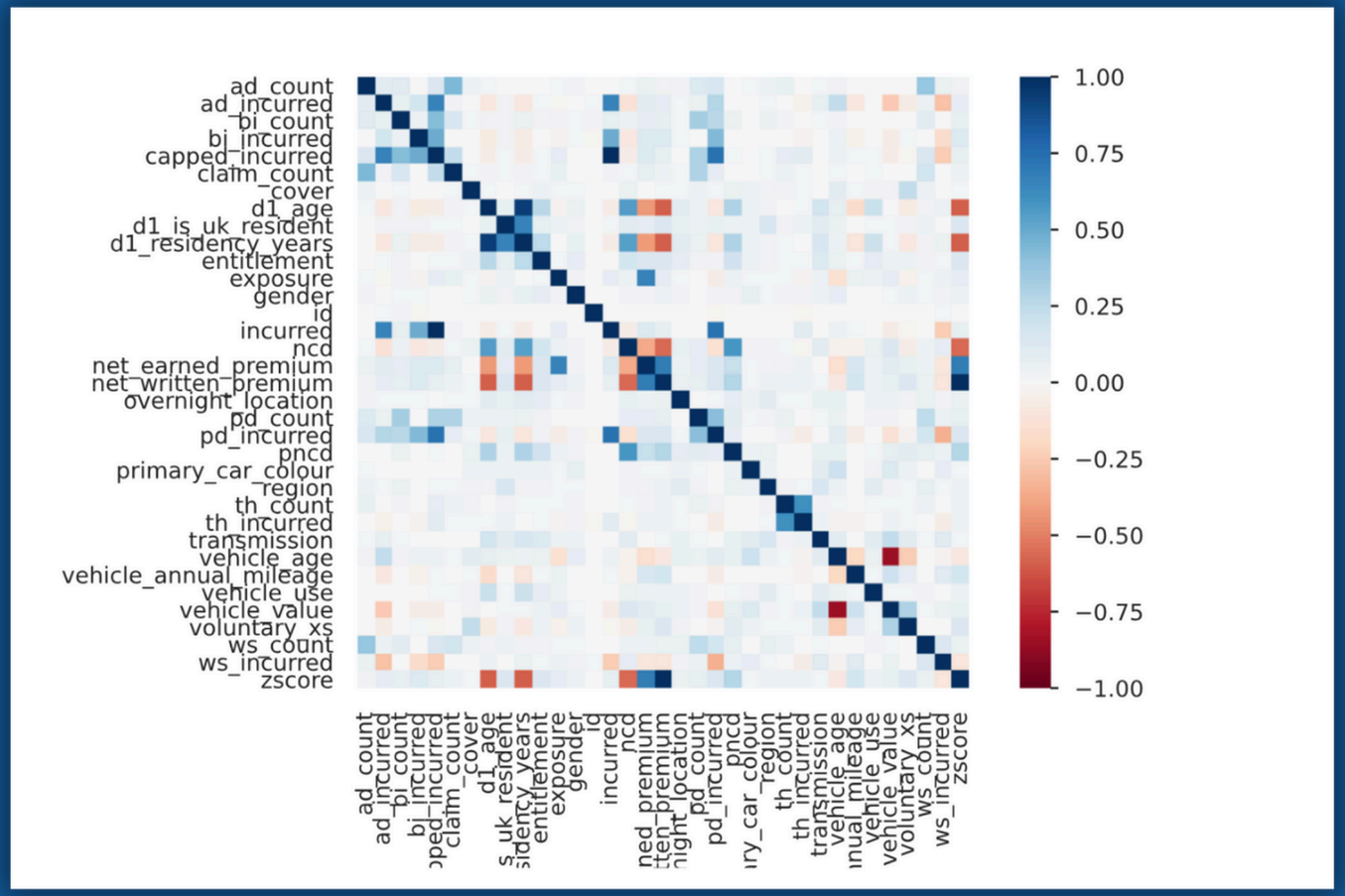
vehicle_make

Text

Distinct	67
Distinct (%)	0.1%
Missing	0
Missing (%)	0.0%
Memory size	6.0 MiB



Correlation Matrix



Data Preprocessing

Cleaning

1. Addressed missing values:
   - Categorical: Replaced with mode.
   - Numerical: Imputed with median.

2. Handled outliers (e.g., negative incurred values set to 0).

3. Transformed data for consistency:
   - Binned vehicle_age.
   - Binned vehicle value.

Feature Engineering

4. Feature Engineering: Created few extra columns based on understanding of the dataset.
   - policy duration
   - Claim severity
   - Driver experience

Feature Engineering

New Features Added:

- Claim Severity = incurred / claim_count
- Premium per Exposure = net_earned_premium / exposure
- Vehicle make categories transformed using one-hot encoding.
- Date features (e.g., policy_duration extracted from start_date).

Top Predictive Features

Key Insights from Analysis:

Top Features:

1. Vehicle Value
2. Vehicle Age Group
3. Vehicle Make

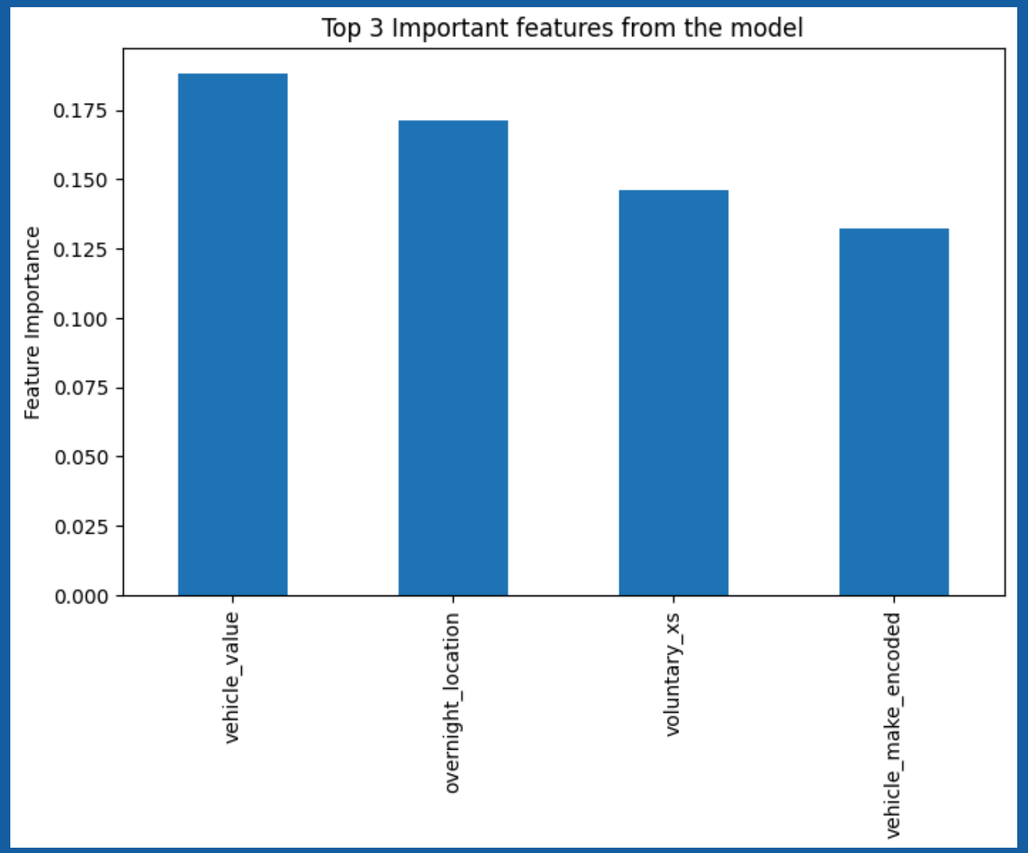
Data Quality Checks

Implemented tests for:

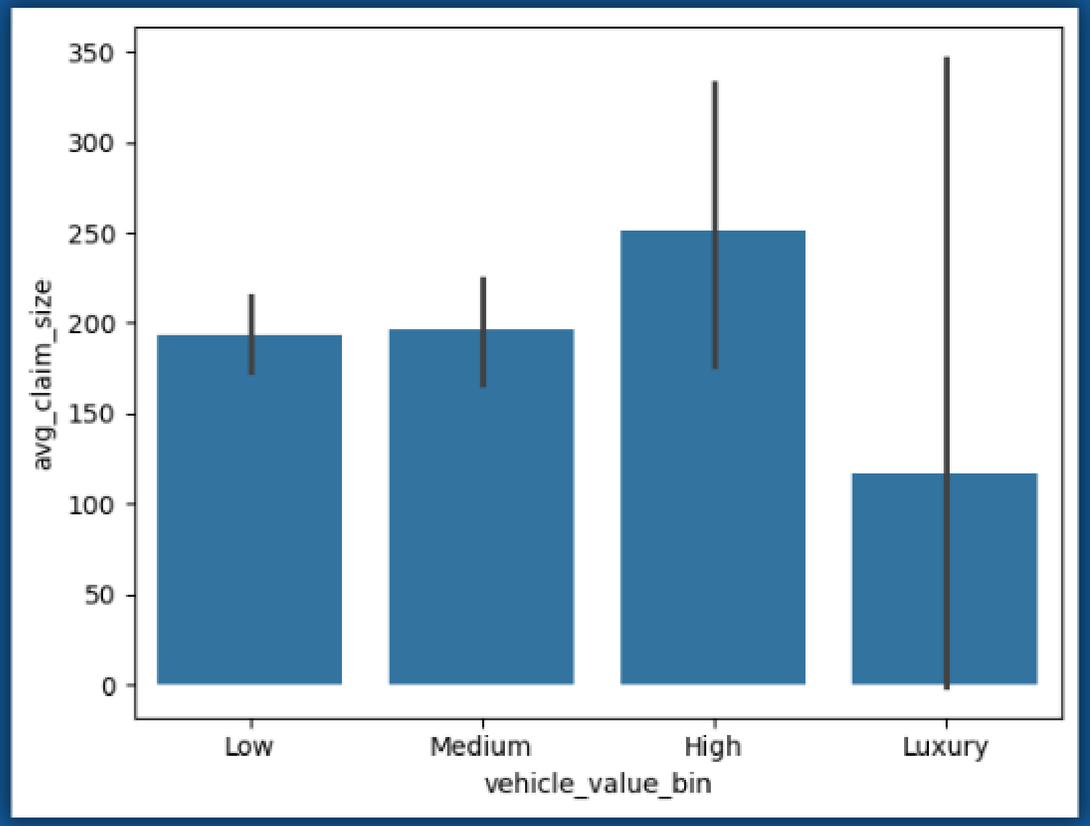
- Missing values.
- Outliers.
- Duplicates.
- Automated flags for invalid data (e.g., negative incurred values).

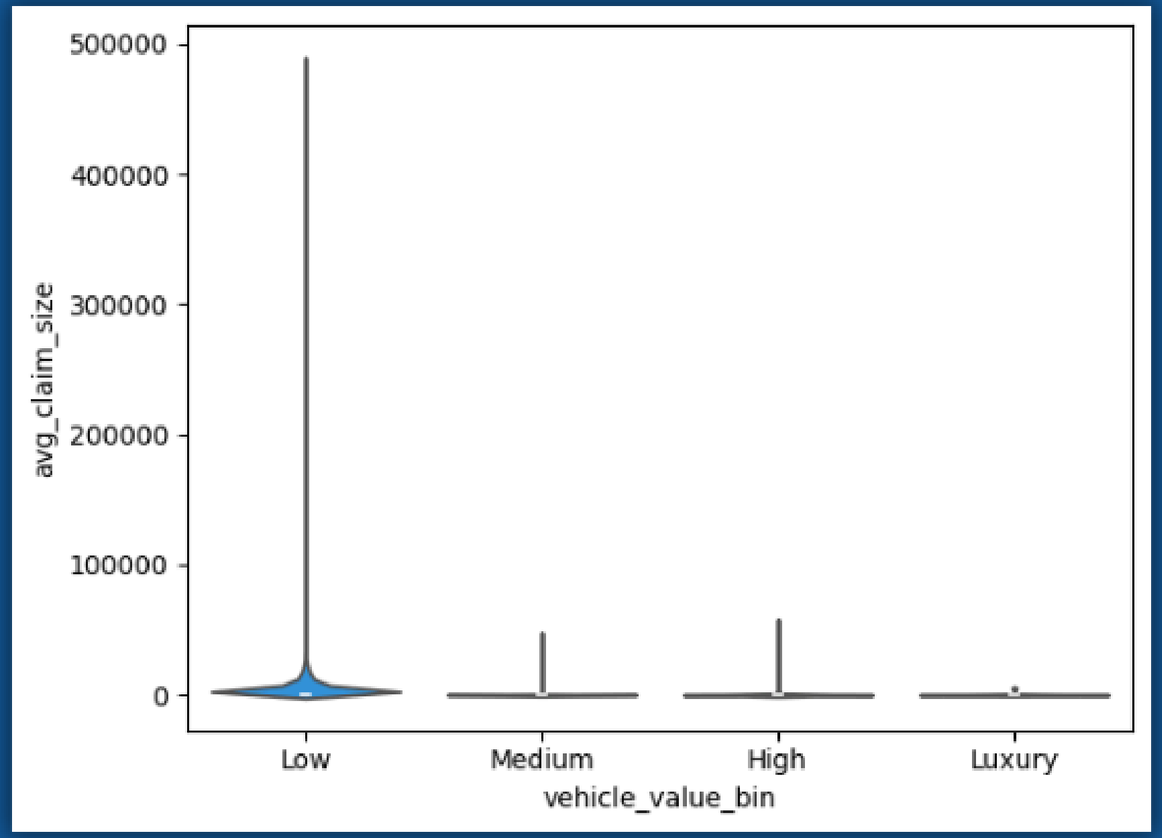
Model Building

- Random Forest used for feature importance.
- Identified top features.

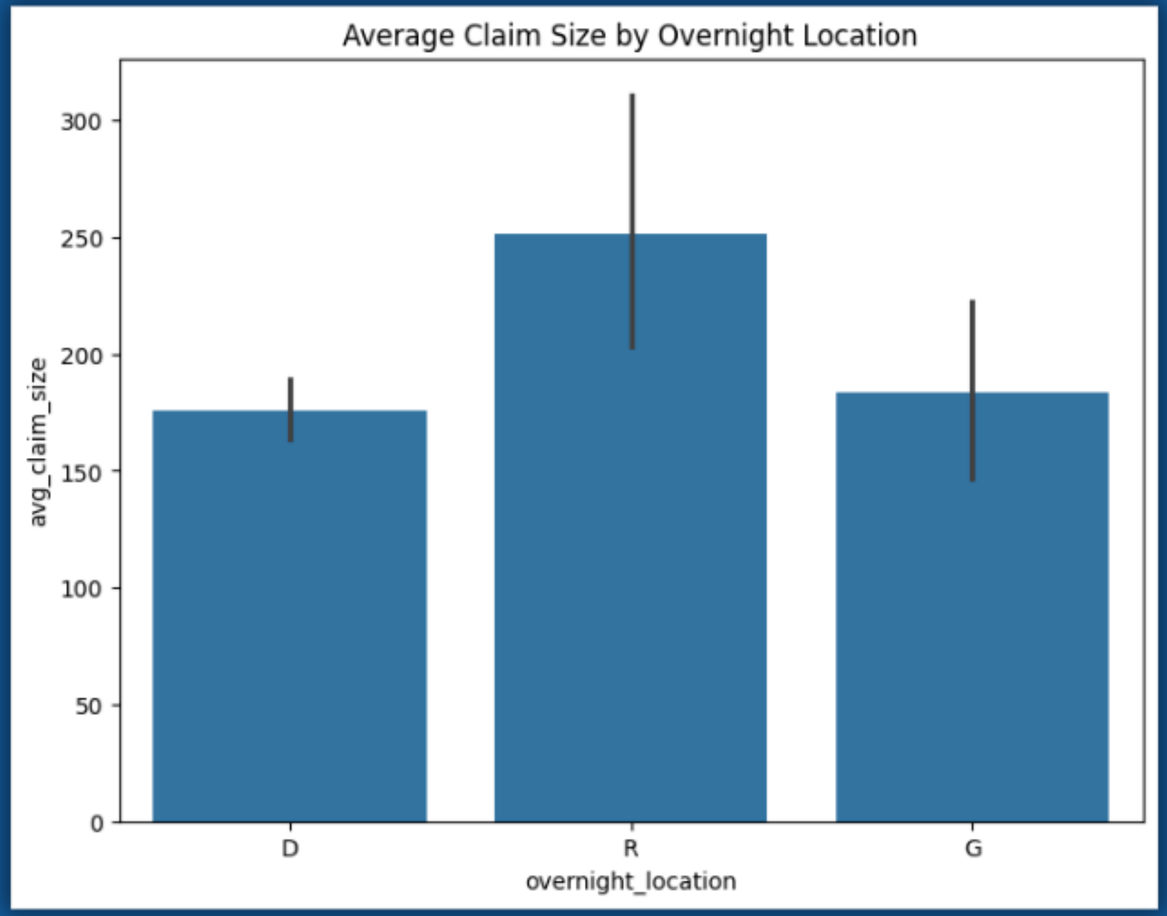


Vehicle Value

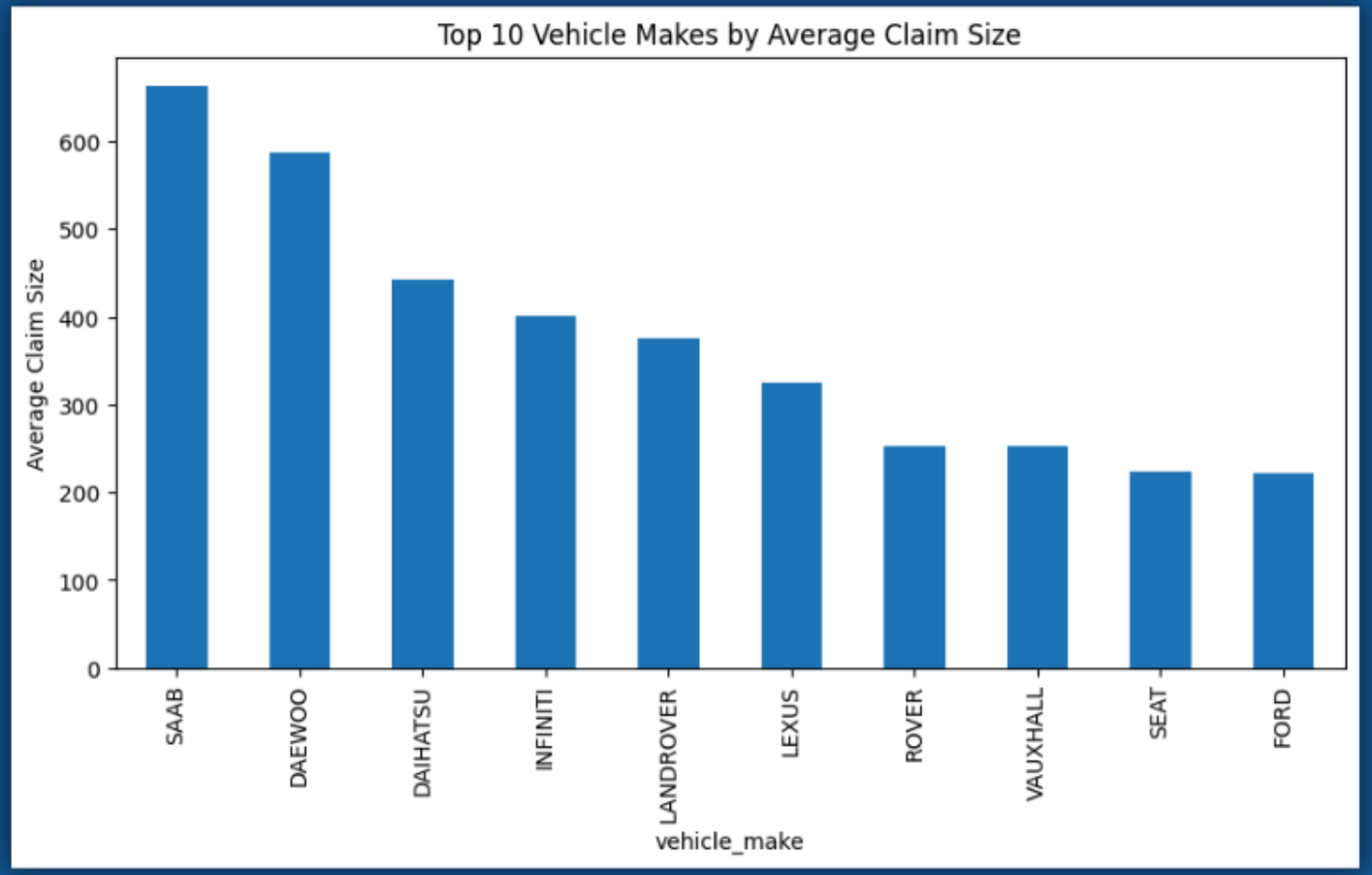


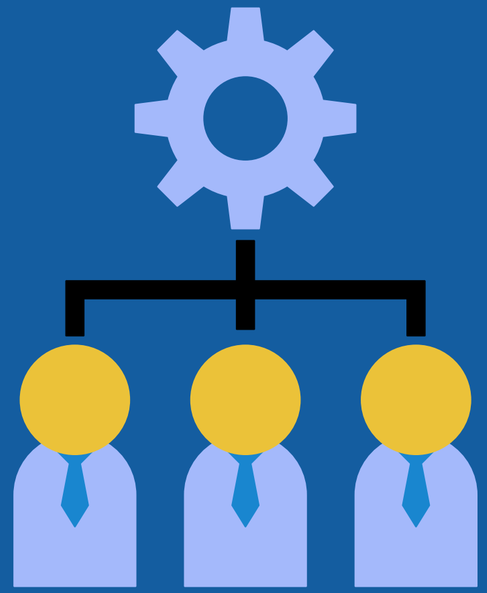


Overnight Location

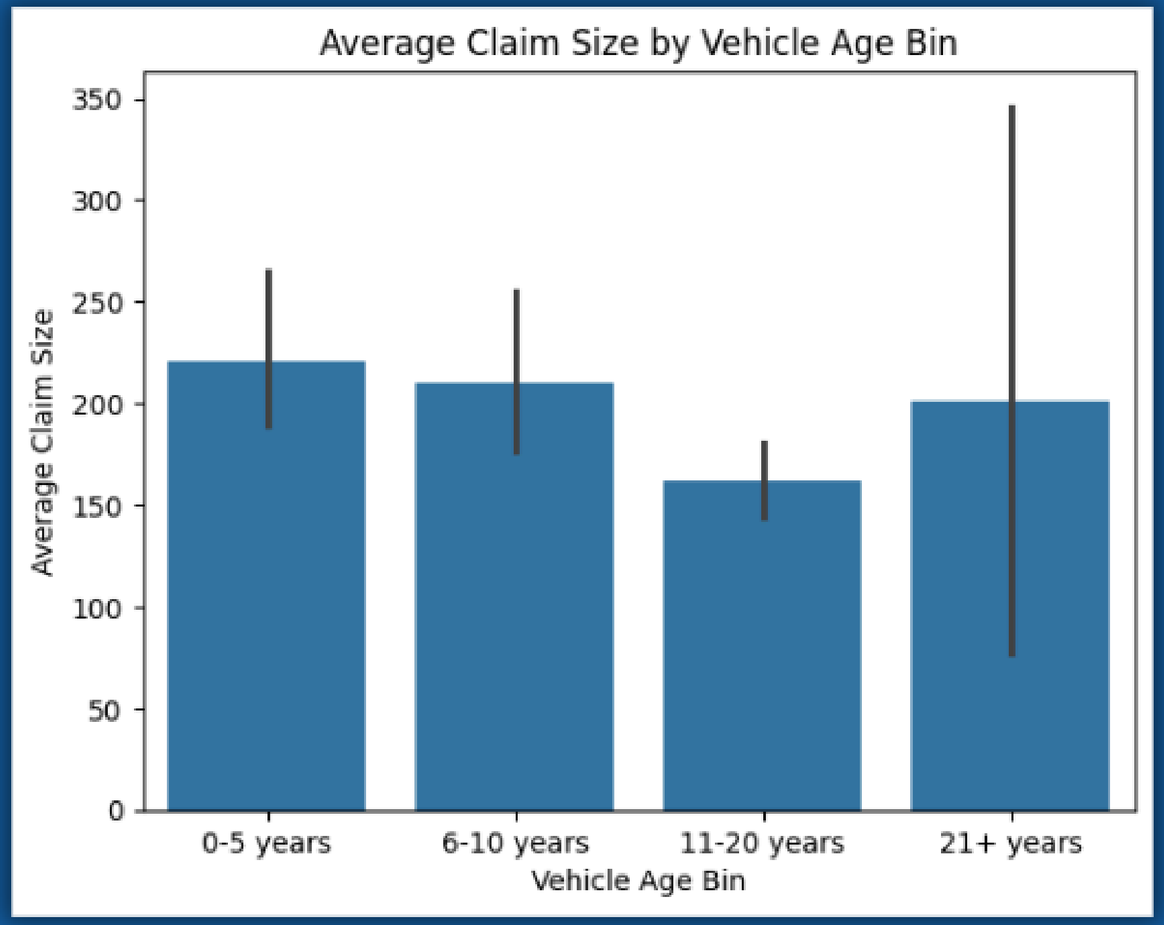


Vehicle Make



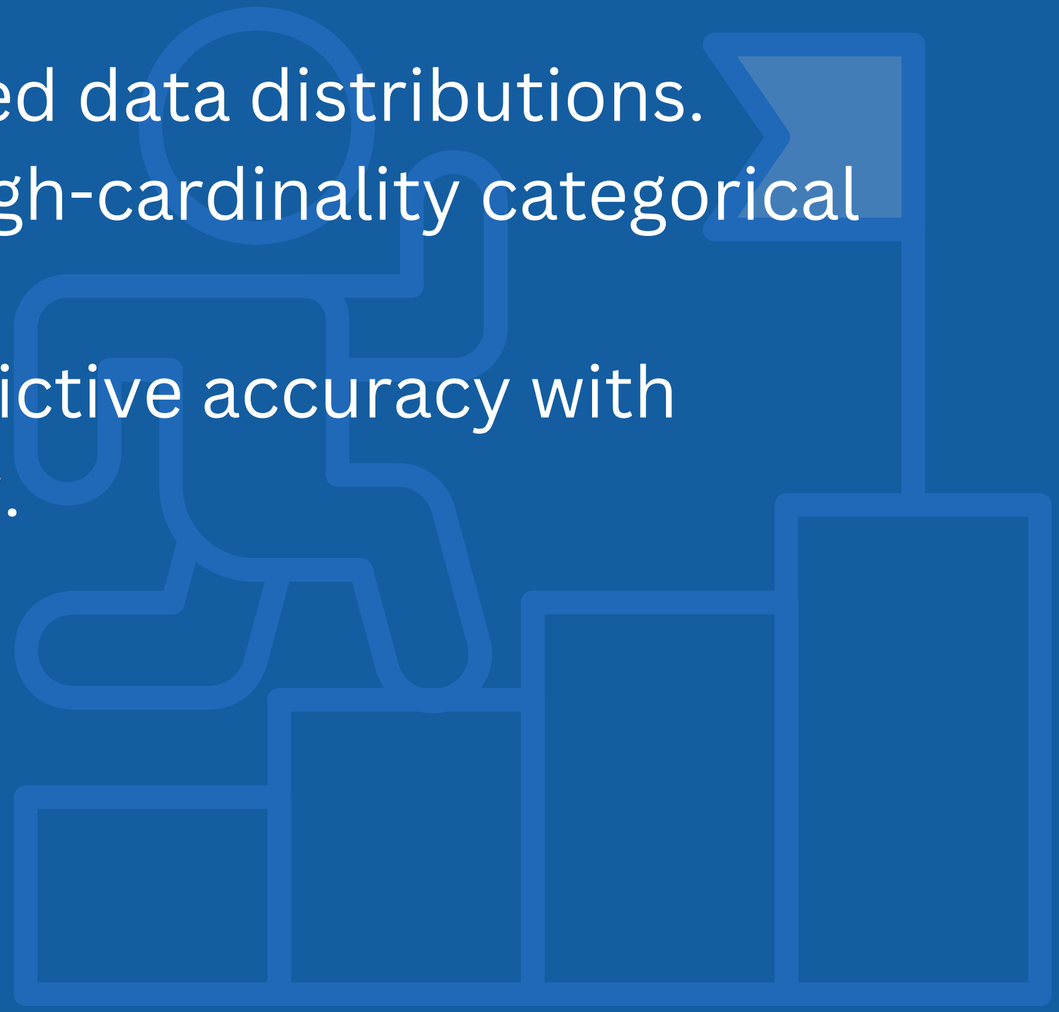


Vehicle Age


Average Claim Size by Vehicle Age Bin
Average Claim Size
350
300
250
200
150
100
50
0
0-5 years
6-10 years
11-20 years
21+ years
Vehicle Age Bin

Challenges

- Handling skewed data distributions.
- Dealing with high-cardinality categorical features.
- Balancing predictive accuracy with interpretability.



Future Work

- Implement real-time data pipelines.
- Use advanced models (e.g., Gradient Boosting).
- Explore external data sources for enrichment.

Conclusion

- Data transformation and feature engineering improved the data quality and hence in future will be used for model building purpose also.
- Top 3 feature were explored in details.