Module 2 Final Project - Milestone 1: EDA

**Exploratory Data Analysis for Car Insurance Claim Prediction**

****

**Mohit Maithani**

Northeastern University: College of Professional Studies

ALY 6040: Data Mining

Professor Justin Grosz

Group Member: Ayush Anand and Mohit Maithani

March 4, 2025

**Abstract**

This report presents an exploratory data analysis (EDA) of the Car Insurance Claim Prediction dataset, aiming to **identify key risk factors that influence the likelihood of a policyholder filing an insurance claim within six months**. Through **data cleaning, statistical summaries, and data visualizations**, we analyze the impact of critical variables such as **age of car, policy tenure, and population density** on claims. These insights will help insurance companies refine their **risk assessment models, pricing strategies, and customer engagement plans**.

## **1. Introduction**

Insurance companies rely heavily on **predictive analytics** to assess risks and **optimize premium pricing**. This study explores a **real-world insurance dataset** containing **policyholder details, car attributes, and claim history**, mirroring the use of data-driven decision-making seen in industries like streaming services (e.g., Amazon Prime optimizing content recommendations).

The core objective is to **analyze trends in claim behavior** and provide **actionable insights for risk-based premium adjustments, fraud detection, and personalized insurance policies**.

## **2. Data Description and Preparation**

The dataset consists of **over 100,000 policyholder records**, including features like **policy tenure, car specifications, claim history, and population density**. To ensure **data integrity**, the following preprocessing steps were applied:

### **2.1 Identifying and Handling Missing Data**

* **'age\_of\_car'**: Missing values were imputed using the **median** (less sensitive to outliers).
* **'is\_claim'**: Missing values were **filtered out**, as they belong to test dataset entries that do not contain claim outcomes.

### **2.2 Data Cleaning & Transformation**

* **Non-Numeric Data**: Categorical variables such as **‘make’, ‘segment’, and ‘fuel\_type’** were encoded into numerical values for analysis.
* **Outliers Detection**: Used **boxplots and IQR (Interquartile Range)** to identify and cap extreme values.
* **Feature Engineering**: Created **new metrics** such as **risk scoring** based on car age and previous claims.

## **3. Metrics Identification**

We selected **three key metrics** that significantly influence insurance claim:

1. **Age of Car**: Older cars are expected to have a higher likelihood of claims due to wear and tear.
2. **Policy Tenure**: Longer policy durations might correlate with **loyalty** but also **increased risk exposure** over time.
3. **Population Density**: Higher-density areas may have **higher accident rates** due to increased traffic congestion.

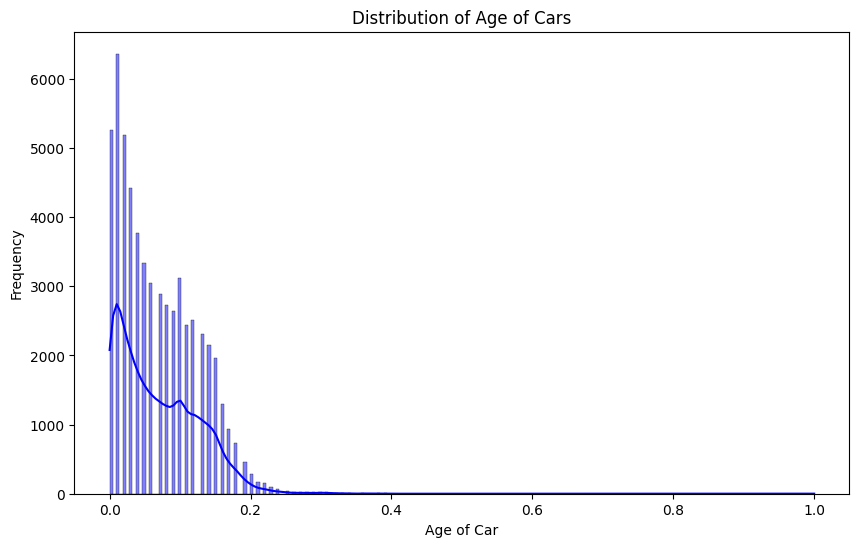
## **4. Data Analysis and Visualizations**

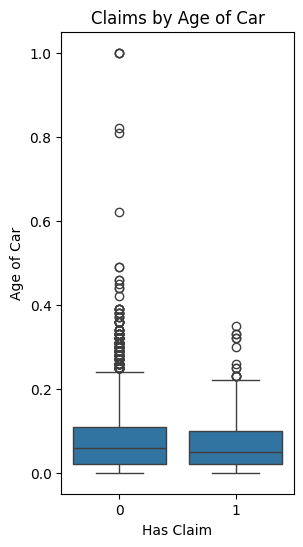
To extract meaningful insights, we employed various **data visualization techniques**:

### **4.1 Distribution of Car Ages**

*Figure 1: Histogram of Car Age Distribution*

🔹 **Key Insights**:

* The dataset contains mostly **newer cars** (right-skewed distribution).
* This suggests a **lower overall risk profile** but indicates that **older cars may require additional risk assessment**.****



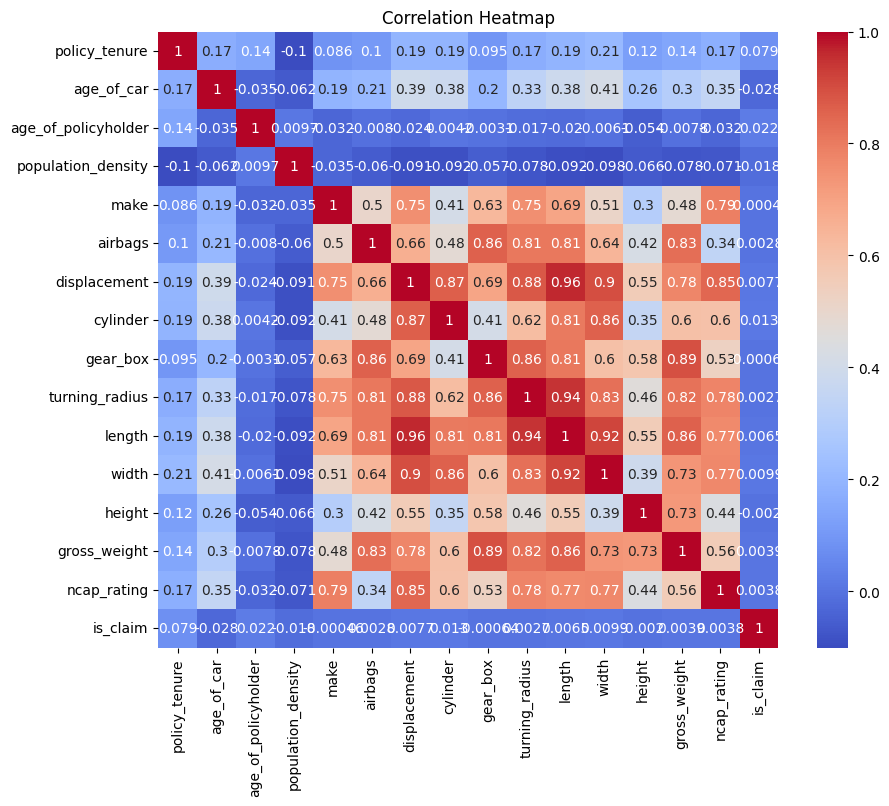
### **4.2 Claims by Age of Car**

*Figure 2: Boxplot of Claims by Age of Car*

🔹 **Key Insights**:

* Older cars **tend to have more claims**, validating the hypothesis that **vehicle wear and tear impacts claim probability**.
* The presence of **outliers** suggests a subset of vehicles with **exceptionally high claim frequencies**, warranting further investigation.

### **4.3 Correlation Between Variables**

*Figure 3: Correlation Heatmap*****

🔹 **Key Insights**:

* **Low correlation** between individual variables and claims suggests **multiple factors contribute to claim likelihood**.
* Strong correlations exist between **car specifications (e.g., weight, engine displacement, and turning radius)**, indicating **these factors should be considered together in risk assessment models**.

## **5. Key Findings and Business Recommendations**

Based on the EDA, we propose the following **data-driven strategies**:

### **5.1 Premium Adjustments Based on Risk**

* Introduce **tiered pricing models** that **increase premiums for older vehicles** due to their **higher claim probability**.
* Offer **discounted premiums for policyholders in lower-risk demographics** (e.g., newer cars, safe driving records).

### **5.2 Customer Retention & Maintenance Incentives**

* Provide **preventative maintenance discounts** for older vehicles to **reduce the likelihood of claims**.
* Implement **loyalty incentives** for long-term policyholders to **retain profitable customers**.

### **5.3 High-Risk Area Strategies**

* Increase **awareness campaigns** for policyholders in high-density areas to promote **safer driving**.
* Offer **customized coverage plans** based on the **historical claim frequency of a specific area**.

## **6. Conclusion and Reflection**

The findings from this study align with the fundamental principles of **predictive analytics** used in various industries, including **insurance and streaming platforms**.

* Similar to how **Netflix leverages user behavior** to enhance recommendations, **insurance companies must leverage claim patterns** to optimize **risk assessment, pricing strategies, and customer engagement**.

This analysis provided a **data-driven approach** to understanding car insurance claims, demonstrating that **older vehicles and high-density areas are key risk factors**. Future work could explore **machine learning models** to further refine claim prediction accuracy.

## **References**

* Python Software Foundation. (2020). *Python Language Reference, version 3.8*. Retrieved from<https://www.python.org>
* McKinney, W. (2012). *Pandas: A Foundational Python Library for Data Analysis and Statistics*. O'Reilly Media.
* Seaborn Developers. (2023). *Seaborn: Statistical Data Visualization*. Retrieved from https://seaborn.pydata.org/
* Matplotlib Developers. (2023). *Matplotlib: Visualization with Python*. Retrieved from<https://matplotlib.org/>
* Kaggle. (2023). *Car Insurance Claim Prediction Dataset*. Retrieved from [www.kaggle.com/datasets/ifteshanajnin/carinsuranceclaimprediction-classification](https://www.kaggle.com/datasets/ifteshanajnin/carinsuranceclaimprediction-classification)