# Training Prep - Based on Data Engineering & Data Analysis

1. **Are there any insights or interesting findings in the data that would be important to share with your data scientist partner?**

* Age and Claims: The age group 20-29 has the highest total claim payout, which could be valuable for risk assessment and targeted marketing campaigns.
* State Variations: There is a variation in average age across different states (up to 26 years difference between highest and lowest). This could be relevant for setting premiums based on location.
* Car Age: Newer cars (2020-2024) have the highest representation in the data, potentially indicating a bias towards newer car owners or a lack of data for older vehicles.
* Safety Discounts: A small portion of cars have zero safety discounts, potentially indicating inherently safer vehicles or those with additional safety features.

I plotted the following histograms with the final data:

A graph of a distribution of vehicles

Description automatically generated

Vehicle Value: The histogram distribution for vehicle value is right-skewed, indicating that most customers have vehicles valued at $2,000 or less. This suggests that the majority of the customer base consists of lower-value vehicles, which may influence pricing strategies for insurance premiums.

A graph of a distribution of miles driven

Description automatically generated

Annual Miles Driven: Similarly, the histogram for annual miles driven is also right-skewed, with the majority of customers driving under 20,000 miles per year. This finding indicates that low-mileage drivers represent a significant portion of the dataset, potentially signifying lower risk profiles for insurance coverage.

A graph of a distribution of premium amount

Description automatically generated

6-Month Premium Amount: The histogram distribution for the 6-month premium amount shows a concentrated range, with most customers having premiums between $0 and $500. This wide concentration suggests that a significant number of customers fall within a similar pricing bracket, which may indicate a standard risk level among this group.

1. **What strategy did you use for dealing with the missing and duplicate data?**

To handle missing and duplicate data, I implemented the following strategies:

* Missing Data: After checking for missing values across the dataset, I found that there were no missing values in any of the columns. Therefore, no imputation or removal was necessary.
* Duplicate Data: There were no duplicate rows. However, I identified and removed duplicate entries based on key columns (CAR\_ID, CUST\_ID, HH\_ID) to ensure that analyses were not skewed by repeated records. This step helps maintain the integrity of the data and improves the accuracy of any model developed using this dataset.

This strategy ensures that the dataset is clean and ready for further analysis, reducing potential biases in subsequent insights.

1. **Thinking about this from an insurance standpoint, what additional features would you like to add to this data?**

From an insurance standpoint, several additional features could enhance the dataset:

* Driver History: Information regarding the driver's previous claims and driving violations could provide insights into risk assessment.
* Driver's License Information: Years of driving experience, license type (learner's permit, regular license, etc.) and any driving violations.
* Vehicle Safety Ratings: Including safety ratings for each vehicle model could help assess the likelihood of claims based on vehicle safety features.
* Claim History: Past claim details (date, type of claim, amount paid).
* Seasonal Usage: Understanding if the vehicle is used seasonally (e.g., for vacations) may affect the risk profile and premium pricing.
* Vehicle Usage: Daily/weekly mileage, primary purpose (commute, leisure, business), garaged overnight.
* Medical History: This would require careful consideration regarding privacy concerns, but certain medical conditions might be relevant for risk assessment (e.g., epilepsy, vision problems)

Adding these features would allow for a more robust risk assessment and could lead to more tailored insurance products.

1. **What features (if any) would you recommend removing from the final data set? Why?**

* In analyzing the dataset, I would recommend considering the removal of the following features:
* Phone Number: Unless anonymized, this raises privacy concerns and might not be directly relevant for most insurance models.
* ZIP Code: While location is important, ZIP code might be too granular. State could be a better option depending on model requirements and privacy considerations.