**STEP 3:**

1. **A few insights and findings in the data that should be shared with the data scientist partner:**
2. **Lien/Lease vs. Antique Vehicles:**
   * There were instances where lien/lease = 1 and antique = 1.
   * Antique vehicles (over 25 years old) should typically not have liens or leases due to their ownership status and usage patterns. Would you agree with this reasoning?
3. **Customer Cancellation Records:**
   * Identified 31,708 records with status = 'Customer Cancellation' still holding discounts enabled.
   * **Action Taken:** Discounts have been converted to 0 to prevent impact on aggregations. Do you want to revert this change?
4. **Employment Data Issues:**
   * Found 5,464 rows in the CUSTOMERS data with employment\_type and income marked as “**#REF!**”.
   * **Action Taken:** Converted these to NULL for analysis. Should we drop these records, set a default value, or impute income based on age?
5. **Data Partitioning:**
   * Partitioned data by state due to even distribution and unique car\_ids per state.
   * **Question:** What columns do you frequently query on? Should we consider partitioning by a different column to enhance efficiency?
6. **Handling Duplicates and Null values:**
7. **Null Values in CUSTOMERS.csv:**
   * In the CUSTOMERS.csv file, 15 columns from \_c5 to \_c19 had all NULL values, so I deleted them safely.
   * I checked for null values and planned to drop them after joining the datasets to avoid prematurely dropping records and losing valuable information. Luckily, there were no more null values at all.
8. **Duplicate Records Check:**
   * I then checked for duplicates to handle them before merging datasets and causing unintended joins. The results are:
   * The ***CARS*** DataFrame has 95,051 duplicate records based on car\_id out of 499,999 total records.
   * The ***CUSTOMERS*** DataFrame has 149 duplicate records based on cust\_id out of 500,000 total records.
   * The ***HOUSEHOLDS*** DataFrame has 131,352 duplicate records based on hh\_id out of 499,999 total records.
9. **Strategy for Handling Duplicates**

* I identified that car\_id was unique within each state instead of globally. Therefore, I used a combination of car\_id + state to join instead of dropping duplicates in ***CARS***.
* ***CUSTOMERS*** duplicates could be justified, as one customer could have added their family member to their insurance policy. Before joining, I created a new DataFrame retaining only the oldest member’s record.
* ***HOUSEHOLDS*** duplicates were also justified, as one household may have more than one customer and car. A combination of hh\_id + cust\_id was used as the primary key.

1. **Data Quality Checks**
   * Upon performing data quality checks, the ***HOUSEHOLDS*** DataFrame had one **cust\_id** that was not present in the ***CUSTOMERS*** DataFrame.
   * As a result, that record was excluded, as we didn’t have any information about that cust\_id.
2. **Thinking about this from an insurance standpoint, some additional features I would like to add to the data are:**
3. **Driving Behavior Data**
   * + **Telematics Score**: A score based on driving habits (speeding, hard braking, etc.), providing a more nuanced understanding of a driver’s behavior.
     + **Average Speed**: The average speed driven over the past year, which can help assess risk levels more accurately.
4. **Vehicle Safety Features**
   * + **Safety Equipment**: Detailed information on safety features such as airbags and ABS, which can influence both risk and premiums.
     + **Anti-Theft Devices**: A record of the presence of anti-theft devices in the vehicle, contributing to risk assessments and potential discounts.
5. **Geographic Risk Factors**
   * + **Crime Rate**: The crime rate in the area where the vehicle is registered, as higher crime rates can increase the likelihood of theft or vandalism.
     + **Natural Disaster Risk**: Assessment of the risk of natural disasters (flood, fire, etc.) in the region, which can impact coverage needs and pricing.

Currently, we have **vehicle\_safety\_discount** and **driver\_safety\_discount**, which are binary (**0 or 1**) values. More nuanced data like the above would allow for better risk assessment and more tailored insurance offerings.

1. **Thoughts on features to remove:**

**Data is the new oil**, and the more we collect, the greater the potential benefits for the organization in various ways. However, without specific use cases outlined, it becomes challenging and suboptimal to justify the removal of certain features.

That said, I recommend considering the removal of the following features due to their limited relevance for assessing risk in an insurance context:

1. **Phone Number**: Useful for contact purposes, but it does not contribute to risk assessment or underwriting decisions.

2. **Referral Source**: Primarily relevant for marketing and customer acquisition strategies rather than risk evaluation.

3. **Antique Vehicle**: It may lack the nuance needed to impact risk assessments, especially since vehicle age is already captured in model\_year.

4. **Business Use**: It may not provide enough granularity to effectively assess risk. Without details about the nature of business use (e.g., type of business, frequency of use for business purposes), it may not significantly influence risk evaluations.

5. **Status**: If it merely indicates whether a policy is active or inactive without further context, its contribution to risk assessment may be minimal.