Summary:

1. Splitting the data with only “High” and “Very High”
2. Visualizing to check which Style has the highest number claims (4,3,7,1,5,2,9,6,8,12,11,16,10,14)
3. Visualizing the Attributes as per the highest option codes present in them

(Attribute wise)

1. The original data sets after merging has the following structure:

**Low (99,428)** and **Very Low (15,527)** dominate the dataset.

**Medium (6,485)** has a moderate presence.

**High (1,510)** and **Very High (411)** are significantly underrepresented

Which sum up to less than 2% of the entire data

1. Using synthetic data to fill the dataset will result in 5lac approximate columns - not feasible as it requires high computational power.
2. Since computational time and memory are concerns, using **Class Weights(more priority on )** along with SMOTE will provide proportional sample subsets of other classes to match high and very high as well as generate the synthetic data required
3. Here is the conducted procedure:
4. Load and merge the data sets
5. Dropping "Scale Labor Cost" as it is not contributing to claims
6. Filtering the required classes and taking sample proportional subsets from Low, Very Low, and Medium.
7. Combining High/Very High claims with the sampled subset
8. Encoding categorical and target variables
9. Converting to boolean columns to integers and applying SMOTE to balance dataset

Stepwise:  
### \*\*Step-by-Step Procedure with Descriptions\*\*

1. \*\*Load and Merge Data\*\*:

- Import the claims and options datasets.

- Merge them using the `Truck Number` column to create a combined dataset for analysis.

2. \*\*Drop Irrelevant Columns\*\*:

- Remove columns like `Scale Labor Cost` that do not contribute meaningfully to the analysis.

3. \*\*Encode Target Variable\*\*:

- Convert the `Scale Claim Cost` values (e.g., `Very Low`, `Low`, `Medium`, etc.) into numerical categories for modeling.

4. \*\*Visualize Claim Distribution\*\*:

- Plot the distribution of claim severity to understand the imbalance in the dataset.

5. \*\*Filter and Sample Data\*\*:

- Retain all rows with `High` and `Very High` claims.

- For other claim levels (`Very Low`, `Low`, and `Medium`), sample a proportional subset to reduce computational load while preserving class diversity.

6. \*\*One-Hot Encode Categorical Variables\*\*:

- Convert categorical columns into numerical columns using one-hot encoding for model compatibility.

7. \*\*Define Features (X) and Target (y)\*\*:

- Separate independent variables (X) and the target variable (y).

- Drop irrelevant columns like truck identifiers (`Truck Number`, `Claim Number`).

8. \*\*Handle Class Imbalance with SMOTE\*\*:

- Apply Synthetic Minority Oversampling Technique (SMOTE) to create synthetic samples for underrepresented classes, ensuring a balanced dataset.

9. \*\*Split Data into Training, Validation, and Test Sets\*\*:

- Divide the dataset into training, validation, and test subsets while maintaining class proportions (`stratify=y`).

10. \*\*Train Logistic Regression Model\*\*:

- Train a simple and interpretable logistic regression model.

- Evaluate its performance on the validation set using metrics like precision, recall, F1-score, and a confusion matrix.

11. \*\*Train Decision Tree Model\*\*:

- Use a decision tree for more nuanced insights into the relationships between attributes.

- Evaluate its performance on the validation set.

12. \*\*Train Random Forest Model\*\*:

- Train a Random Forest classifier for robust and accurate predictions.

- Use class weights to handle any remaining class imbalance.

13. \*\*Evaluate Best Model on the Test Set\*\*:

- Evaluate the model with the best validation performance on the test set to measure its real-world predictive capability.

14. \*\*Feature Importance Analysis\*\*:

- Extract and visualize the most important features contributing to predictions (e.g., using Random Forest's `feature\_importances\_`).

15. \*\*Attribute Pair Analysis\*\*:

- Identify significant attribute pairings using pairwise combinations.

- Analyze the frequency of attribute pair occurrences and their contribution to warranty claims.

- Visualize the top attribute pairs contributing to high warranty claims.

16. \*\*Save Results\*\*:

- Save cleaned datasets and analytical outputs (e.g., significant attribute pairs, feature importance) to Excel for further reference and reporting.

17. \*\*Visualization and Insights\*\*:

- Use bar plots and heatmaps to present the results of feature importance and confusion matrices.

- Highlight actionable insights and recommendations based on the analysis.

18. \*\*Generate Recommendations\*\*:

- Summarize findings in terms of attributes or attribute pairs associated with higher claims.

- Suggest potential strategies for optimizing future vehicle configurations based on the analysis.

This structured approach ensures clarity in workflow, effective modeling, and actionable insights.