Exploratory Data Analysis & Initial Statistics Report Task - 1

Objective

The aim of Task 1 was to perform initial data cleaning, parsing, and exploratory analysis to understand the structure and relationships within the financial insurance dataset.

Key Activities

1. **Date Field Parsing and Handling Missing Entries**
   - Parsed and converted date fields such as `VehicleIntroDate`.
   - Handled missing or malformed entries using `errors='coerce'`.
2. **Cleaning Numeric Variables**
   - Cleaned key numeric variables, particularly `TotalPremium`, `TotalClaims`, and `SumInsured`, ensuring consistent data types.
3. **Descriptive Statistics and Distributional Analysis**
   - Conducted descriptive statistics and distributional analysis, identifying outliers and skewed distributions (especially for `TotalPremium`).
   - Created a histogram to visualize the distribution of `TotalPremium`, revealing a right-skewed pattern with a concentration of values below 2000.
4. **Correlation Analysis**
   - Built a correlation heatmap to explore relationships between:
     - `TotalPremium`
     - `TotalClaims`
     - `SumInsured`
     - `CalculatedPremiumPerTerm`

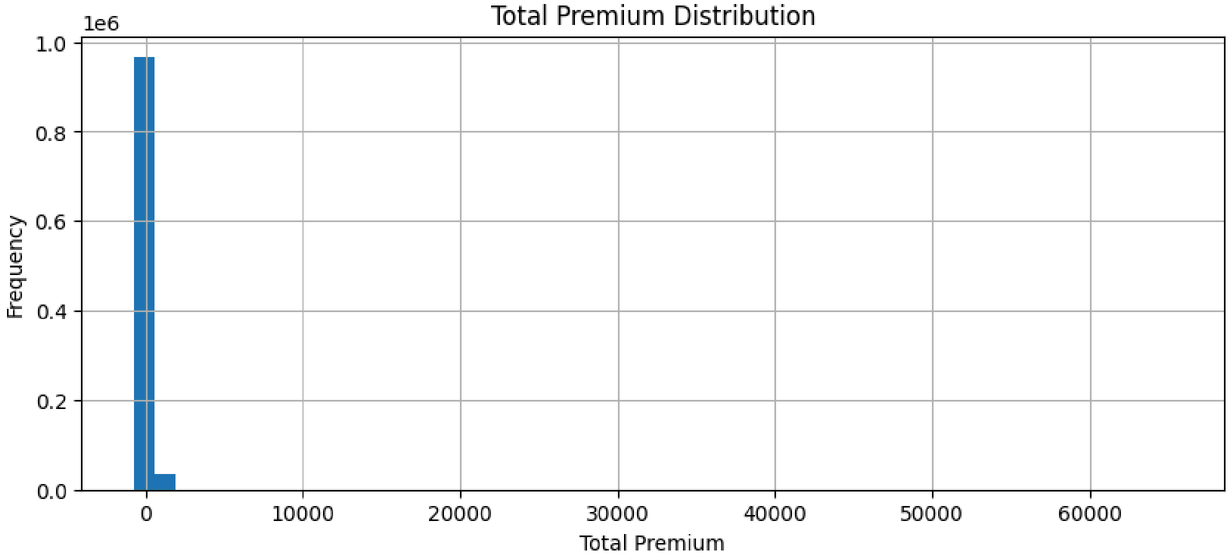
Key Findings

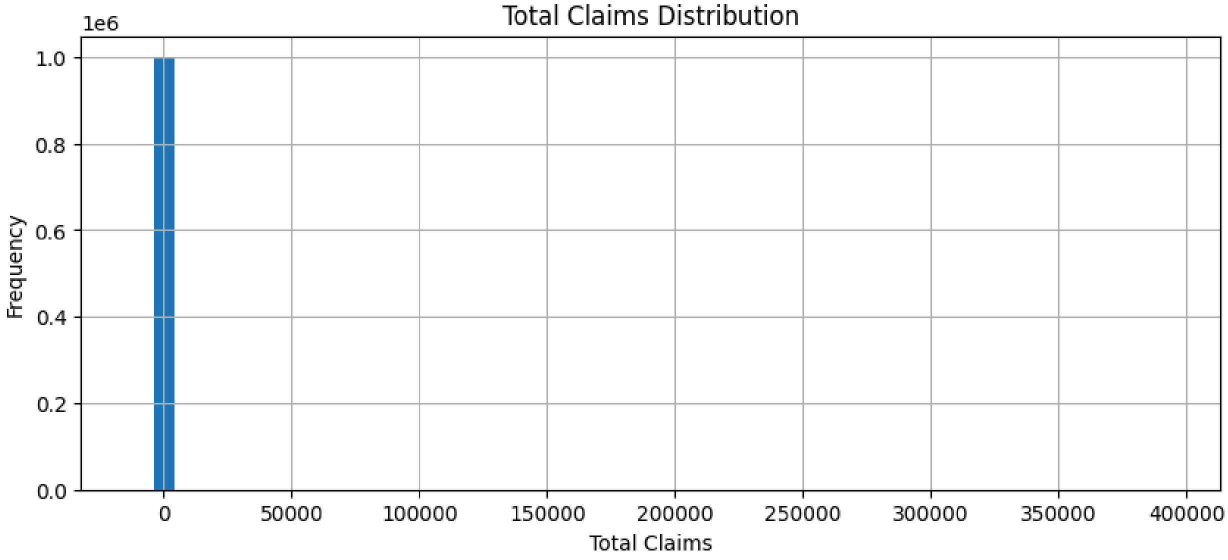
1. **Correlation Analysis**
   - Moderate positive correlation between `TotalPremium` and `CalculatedPremiumPerTerm` (r = 0.64).
   - Weak correlation between `TotalPremium` and `TotalClaims` (r = 0.12).

    - No significant correlation between `SumInsured` and other variables.
2. **Distribution Analysis**
    - Strong data imbalance in `TotalPremium` distribution — this may require log transformation or outlier treatment in future modeling steps.

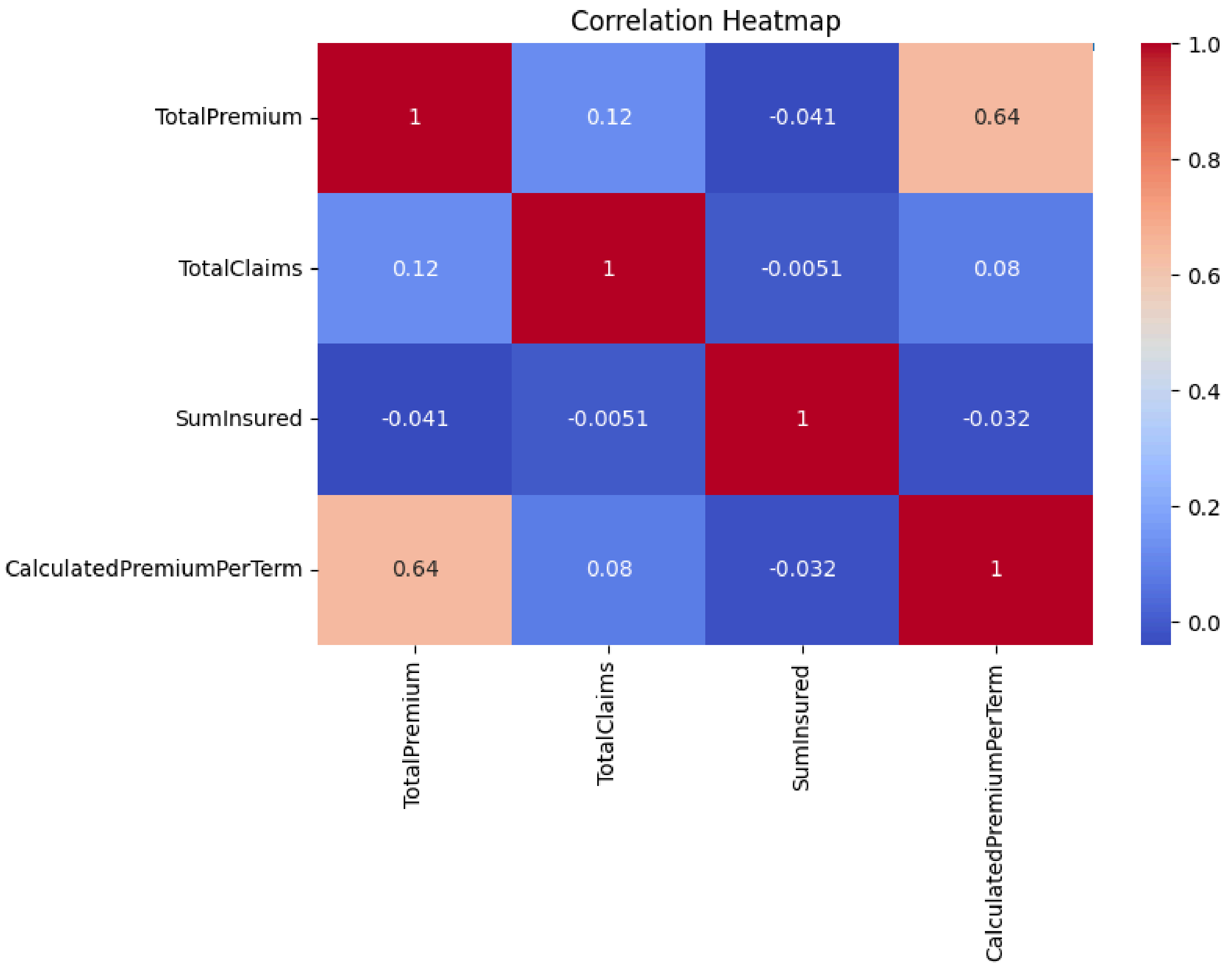
Visualizations

- **Histogram of TotalPremium**
    - Reveals a right-skewed pattern with a concentration of values below 2000.





- **Correlation Heatmap**
    - Visualizes relationships between key variables.



Recommendations

- Consider log transformation or outlier treatment for `TotalPremium` to address data imbalance.
- Further investigate the weak correlation between `TotalPremium` and `TotalClaims` to understand underlying factors.