Interim Report: B5W3 Insurance Risk Analytics

Submitted for the Interim Deadline: June 13, 2025

Repository: https://github.com/moheranus/B5W3-Insurance-Risk-Analytics

Branch: task-1

Submitted on: June 12, 2025

By: Daniel Shobe

1 Overview

This interim report summarizes Task 1 (Git Setup & Exploratory Data Analysis) for the B5W3: End-to-End Insurance Risk Analytics & Predictive Modeling project. Task 1 established a Git repository, performed EDA on the insurance dataset, and generated visualizations to uncover risk and profitability patterns, addressing key performance indicators (KPIs) and guiding questions.

2 Task 1: Git and GitHub Setup

2.1 Achievements

- Repository: https://github.com/moheranus/B5W3-Insurance-Risk-Analytics
- Branch: task-1
- Structure:
  – data/: Hosts MachineLearningRating_v3.txt (1M rows, 503.89 MB).
  – src/: Contains eda_insurance_analysis.py.
  – visualizations/: Stores 10 PNG visualizations, archived as visualizations.zip
  – reports/: Includes this report.
- README: Project overview, setup, and usage.
- Commits: Multiple on June 12, 2025 (e.g., visualizations, .gitignore).
- CI/CD: Planned using GitHub Actions.

2.2 KPIs

- Dev Environment: Configured Python with pandas, seaborn, matplotlib; Git initialized.
- Skills: Demonstrated Git and data analysis proficiency

3 Task 1: Exploratory Data Analysis & Statistics

3.1 Dataset

- File: MachineLearningRating_v3.txt (1,000,098 rows, 52 columns)
- Period: February 2014–August 2015.
- Features: TotalPremium, TotalClaims, Province, VehicleType, Gender, make, TransactionMonth.

3.2 EDA Execution

EDA was conducted using src/eda_insurance_analysis.py

3.2.1 Data Summarization

- Descriptive Statistics:
  – TotalPremium: Mean 61.91, Std 230.28, Range [-782.58, 65,282.60].
  – TotalClaims: Mean 64.86, Std 2,384.08, Range [-12,002.41, 393,092.10].
  – CustomValueEstimate: Mean 225,531, Std 564,516, 78% missing.
- Data Structure: 52 columns (int64, float64, object, bool); TransactionMonth as datetime.

3.2.2 Data Quality Assessment

- Missing Values:
  – NumberOfVehiclesInFleet: 100%.
  – CrossBorder: 99.93%.
  – CustomValueEstimate: 77.96%.
  – WrittenOff, Rebuilt, Converted: 64.18%.

3.2.3 Univariate Analysis

- Numerical: Histograms (numerical_distributions.png) show skewed distributions.
- Categorical: Bar charts (categorical_distributions.png) for Province, VehicleType, Gender.

3.2.4 Bivariate/Multivariate Analysis

- Loss Ratio: Visualized by Province, VehicleType, Gender (loss_ratio_*.png).
- Correlations: Heatmap (correlation_matrix.png).
- ZipCode: Scatter plot (premium_vs_claims_postalcode.png).

3.2.5 Temporal Trends

- Monthly claims and premiums (monthly_trends.png) show seasonal patterns

3.2.6 Outlier Detection

- Box plots (box_plots.png) identify outliers in TotalClaims.

3.2.7 Vehicle Make Analysis

- Top 10 makes by claims (top_makes_claims.png).

3.3 Creative Visualizations

- loss_ratio_province.png: Regional risk differences.
- monthly_trends.png: Seasonal patterns.
- top_makes_claims.png: High-risk makes.

3.4 Statistical Thinking

- Distributions: Histograms with KDE.
- Correlations: Pearson matrix.
- References: “Python for Data Analysis” (Wes McKinney), seaborn docs

3.5 KPIs

- Proactivity: Resolved large file issues, optimized EDA.
- EDA Techniques: Comprehensive visualizations.
- Stats: Actionable insights from loss ratios and trends.

4 Challenges and Solutions

- Large File: Removed using git filter-repo.
- Performance: Sampled 10,000 rows
- Data Quality: Negative values, missing data noted.

5 Task 2: DVC Setup

• Status: Planned.

• Plan: Initialize DVC, track dataset.

6 Next Steps

• Task 5: DVC setup.

• Task 3: A/B testing.

• Task 4: Modeling.

7 Limitations

• Negative values and missing data.

• Sampled visualizations.

• Mixed types in CapitalOutstanding, CrossBorder.

8 Visualizations

In visualizations.zip on task-1.

9 Conclusion

Task 1 delivered a robust EDA, uncovering risk insights. The repository is ready for Task 2.