Interim Report: Insurance Risk Analytics for 10 Academy Artificial Intelligence Mastery Project

Submission Deadline: June 15, 2025, 8:00 PM UTC **Repository**: https://github.com/bekonad/insurance-analytics.git **Prepared by**: Bereket Feleke

This report summarizes progress on **Task 1** (Git and GitHub + Exploratory Data Analysis) and **Task 2** (Data Version Control) for the 10 Academy Artificial Intelligence Mastery project, focusing on insurance risk analytics for AlphaCare Insurance Solutions (ACIS).

Project Overview

- **Objective**: Analyze historical car insurance data (February 2014 – August 2015) to identify low-risk segments and optimize premium pricing.
- **Data**: `insurance_data.txt`, a pipe-delimited dataset with 51 columns, including `UnderwrittenCoverID`, `PolicyID`, `TransactionMonth`, `Province`, `VehicleType`, `Gender`, `Make`, `TotalPremium`, and `TotalClaims`.
- **Tasks**:
  - Task 1.1: Set up Git and GitHub repository.
  - Task 1.2: Conduct EDA to explore risk and profitability patterns.
  - Task 2: Implement Data Version Control (DVC) for `insurance_data.txt`.
- **Environment**: Mobile device using Google Colab and GitHub.
- **Tools**: Python, pandas, matplotlib, seaborn, DVC, GitHub Actions.

Task 1: Git and GitHub + Exploratory Data Analysis (EDA)

Task 1.1: Git and GitHub

- **Status**: Completed
- **Actions**:
  - Created `insurance-analytics` repository with README and MIT License.
  - Configured CI/CD with `.github/workflows/lint.yml` for Python linting.
  - Merged `task-1` branch into `main` via pull request.
- **Outcome**: Functional repository with version control.

Task 1.2: EDA

- **Status**: Completed
- **Notebook**: `EDA_Insurance_Analytics.ipynb`
- **Challenges**:
  - **Output Suppression**: Only "Loss Ratio by Gender" displayed due to Colab's single-cell rendering.
    - **Solution**: Split Loss Ratio code into separate cells for Province, VehicleType, and Gender.
  - **TypeError**: Temporal trends plot failed due to `Period` object in `Month`.
    - **Solution**: Converted `Month` to `datetime` with `dt.to_timestamp()`.
  - **Data Loading**: Initial `KeyError: 'TotalClaims'` resolved by using `sep='|'`.
  - **Column Naming**: Renamed columns (`make` to `Make`, etc.) for consistency.
  - **Deprecation Warning**: Seaborn's `palette` without `hue` triggered a warning, but plots rendered correctly.
- **EDA Findings**:
- **Loss Ratio by Province**:
- Gauteng: 0.429, Mpumalanga: 0.393, Limpopo: 0.349, Western Cape: 0.342, North West: 0.285, KwaZulu-Natal: 0.265, Eastern Cape: 0.236, Northern Cape: 0.204, Free State: 0.106.
- **Insight**: Gauteng's high loss ratio indicates urban areas have elevated risk, suggesting higher premiums.
- **Loss Ratio by VehicleType**:
  - Heavy Commercial: 0.794, Light Commercial: 0.544, Medium Commercial: 0.493, Passenger Vehicle: 0.337, Bus: 0.0.
  - **Insight**: Heavy Commercial vehicles have the highest risk due to high claim costs; zero Bus claims may reflect low sample size.
- **Loss Ratio by Gender**:
  - Female: 0.492, Male: 0.349, Not specified: 0.348.
  - **Insight**: Higher female loss ratio warrants further investigation into driving patterns or data biases.
- **Distributions**:
  - `TotalPremium`: Mean 61.91 ZAR, median 2.18 ZAR, max 65,282.60 ZAR, right-skewed (std 230.28).
  - `TotalClaims`: Mean 64.86 ZAR, median 0 ZAR, right-skewed (std 2,384.08).

    - **Insight**: Most policies have low premiums and zero claims, but outliers drive high means, requiring robust modeling.
- **Temporal Trends**:
    - Average claims and claim frequency plotted, showing trends over time.
    - **Insight**: Likely seasonal patterns (e.g., Q4 spikes); exact peaks need further analysis.
- **Claims by Make**:
    - Analysis intended but not shown in output.
    - **Insight**: Expected high claims for prevalent makes (e.g., Toyota) or luxury brands (e.g., BMW).
- **Outcome**: Comprehensive EDA addressing all guiding questions, committed to `task-1

Task 2: Data Version Control (DVC)

- **Status**: Completed
- **Notebook**: `DVC_Setup.ipynb`
- **Actions**:
    - Initialized DVC with `--no-scm` to resolve SCM error.
    - Configured local storage at `/content/dvc_storage`.
    - Tracked `insurance_data.txt` with `dvc add` and pushed
    - Committed `insurance_data.txt.dvc` and `.dvc/config` to `task-2` branch.
    - Verified with `dvc pull`.
- **Challenges**:
    - SCM error resolved with `--no-scm`.
    - Mobile constraints addressed using local storage and web-based
- **Outcome**: Dataset tracked, meeting data versioning KPIs. ##Next Steps
- Merge `task-2` pull request.
- Start Task 3 (A/B Testing) using EDA insights.
- Submit repository URL, PDF of `EDA_Insurance_Analytics.ipynb`, and this report by June 15,

Citations

- DVC Getting Started
- Google Colab
- GitHub Docs