Analysis Project Documentation

1. Title

Insurance Policies Analysis Project

2. Abstract Summary

This project analyses an insurance policies dataset to uncover key factors influencing insurance claims and to identify distinct customer segments. The primary objectives are to understand how demographics, car features, and geographic location relate to claim frequency and amount, and to profile policyholders. The methodology includes data loading and cleaning, exploratory data analysis (EDA) using dplyr, and customer segmentation using k-means clustering. Key analyses focus on demographic claim patterns, the impact of driving children, regional claim trends, and customer segmentation based on age and income. The findings are visualized using ggplot2 to illustrate patterns, such as the relationship between driving children and claim frequency, and to visualize customer clusters. The conclusions from this analysis are intended to inform risk assessment, pricing strategies, and targeted marketing for the insurance provider.

---

3. Introduction

- **Background: The insurance industry relies heavily on data to assess risk, set premiums, and understand customer behaviour. By analysing historical policy and claim data, insurers can identify patterns that lead to more accurate pricing models and improved customer outreach. This project focuses on a dataset of insurance policies to extract such actionable insights.**

- **Problem Statement: The insurer needs to better understand the key drivers of insurance claims (both frequency and amount) among its policyholders. Furthermore, the company wishes to segment its customer base to identify different profiles, which can help in tailoring products and marketing efforts.**
- **Objectives:**
    - **To analyse the relationship between policyholder demographics (age, gender, marital status) and claim behaviour.**
    - **To determine the impact of car characteristics (make, model, use, year) on claims.**
    - **To segment customers based on a combination of demographic and economic factors (age, household income).**
    - **To quantify the effect of policyholders having children who drive on claim frequency and amount.**
    - **To identify and compare claim trends across different geographic coverage zones.**

---

4. Data Description

- **Data Sources: The data was sourced from an Excel file named Insurance Policies Project.xlsx, specifically from the sheet Insurance Policies Project.**
- **Data Collection: The method of data collection is not specified in the analysis script.**

- **Data Characteristics: The dataset contains information on insurance policies. After loading, rows with any missing values were omitted (na.omit()).**
    - **Categorical Variables (converted to factors): MaritalStatus, CarUse, Gender, Parent, Education, CarMake, CarModel, CarColor, CoverageZone.**
    - **Numerical Variables (used in analysis): ClaimFreq (claim frequency), ClaimAmount (claim amount), Age, HouseholdIncome, KidsDriving, CarYear.**
    - **The script runs str(Insurance_Policies) and summary(Insurance_Policies) to inspect the structure and basic statistics of the data.**

---

5. Methodology

- **Data Cleaning: The primary cleaning step involved removing all rows with missing values (na.omit()). Additionally, several character or numerical columns representing categorical data were converted to factors to ensure they are treated correctly in statistical models and plots.**
- **Exploratory Data Analysis (EDA): EDA was performed using the dplyr package. This involved grouping the data by various factors (e.g., demographics, car features, region) and then calculating summary statistics, specifically the mean claim frequency (AvgClaimFreq) and mean claim amount (AvgClaimAmount).**
- **Statistical Analysis: The primary statistical method used was k-means clustering (from the cluster library). This unsupervised**

learning technique was applied to the Age and HouseholdIncome variables to partition policyholders into distinct groups (specified as 3 clusters). The results are then prepared for visualization.

---

6. Analysis

The analysis is structured into several distinct parts, each addressing a specific question:

- **Analysis 1: Claim Analysis by Demographics**
    - **Question: How do claim frequency and amount vary by policyholder age, gender, and marital status?**
    - **Approach: The data was grouped by Age, Gender, and MaritalStatus. The mean ClaimFreq and ClaimAmount were calculated for each unique group.**
    - **Findings: The script generates a summary table (Claims_by_Demographics) containing these results, which can be viewed to identify demographic profiles with higher or lower claim rates and costs.**
- **Analysis 2: Claim Frequency by Car Features**
    - **Question: What is the relationship between a car's make, model, primary use, and year, and its claim metrics?**
    - **Approach: The data was grouped by CarMake, CarModel, CarUse, and CarYear. The mean ClaimFreq and ClaimAmount were calculated for each combination.**
    - **Findings: This analysis produces a summary table (claims_by_car_summary) to identify if specific types of**

vehicles or vehicle use are associated with more frequent or expensive claims.

- **Analysis 3: Customer Segmentation**
    - **Question: Can distinct customer segments be identified based on age and household income?**
    - **Approach: K-means clustering was performed on the Age and HouseholdIncome variables to create 3 clusters. The resulting cluster assignment was added back to the original dataset.**
    - **Findings: The analysis successfully segments customers into three groups. These groups are then visualized to show the separation.**
- **Analysis 4: Policyholders with Children Driving**
    - **Question: Do policyholders with children of driving age (or who are driving) have different claim patterns?**
    - **Approach: A new binary variable, HasChildrenDriving, was created ("Yes" if KidsDriving > 0, "No" otherwise). The data was then grouped by this new variable, and the average claim frequency and amount were calculated for each group.**
    - **Findings: A summary table (children_driving) is generated, which is then used to create a bar chart comparing the average claim frequency for those with and without children driving.**
- **Analysis 5: Regional Claim Trends**
    - **Question: How do claim frequency and amount vary by the policyholder's CoverageZone?**

- **Approach: The data was grouped by the CoverageZone factor. The mean ClaimFreq and ClaimAmount were calculated for each zone.**
- **Findings: The script generates a summary table (regional_summary) and prepares a ggplot visualization (likely a bar chart) to compare these metrics across different regions.**

---

7. Results

- **Summary of Findings: The analysis produces several key data tables and visualizations designed to uncover insights. It identifies customer segments, quantifies the difference in claim frequency for policyholders with driving children, and breaks down claim metrics by demographics, car features, and region. The specific results (e.g., "Cluster 1 has the highest income") are contained within the tables and plots generated by the script.**
- **Visualizations: The R script generates the following visualizations using ggplot2:**
    1. **Customer Segmentation Plot: A scatter plot of Household Income vs. Age, with points colored by their assigned cluster. This helps visualize the distinct profiles of the 3 segments.**
    2. **Claim Frequency by Children Driving: A bar chart showing the average claim frequency for policyholders *with* children driving versus those *without*.**

3. **Regional Claim Trends Plot: A plot (type unspecified in the snippet, likely a bar chart) to visualize the average claim frequency and/or amount by CoverageZone.**

---

8. Discussion

- **Interpretation of Results: The results from this analysis would allow an insurer to interpret *which* factors are most strongly correlated with claims. For example, the "Children Driving" analysis directly tests the hypothesis that this group is higher risk. The segmentation plot helps to create personas (e.g., "Young, Low-Income," "Older, High-Income") that the business can understand and target. The regional analysis can pinpoint zones with unexpectedly high claim rates, meriting further investigation.**
- **Implications:**
    - **Risk & Pricing: Findings can be used to refine underwriting rules and pricing models. For instance, if certain demographics or car models are consistently linked to higher claim amounts, premiums can be adjusted accordingly.**
    - **Marketing: The customer segments can be used for targeted marketing. Low-risk segments (e.g., low claim frequency) could be offered discounts, while high-risk segments could be targeted with safety information or different policy structures.**
    - **Operations: Identifying high-claim regions might suggest a need for more claims adjusters in that area or a review of local-level fraud or risk factors.**

---

9. Conclusion

- **Summary: This project successfully analyzes an insurance dataset to profile customers and understand the drivers of claim frequency and amount. By examining demographics, car features, driving children, and geographic zones, the analysis provides a multi-faceted view of the policyholder base. The k-means clustering provides a clear segmentation of customers based on age and income.**
- **Recommendations: Based on the *types* of analyses performed, the following actions would be recommended:**
    1. **Review Findings: Closely examine the output tables and plots to identify the specific high-risk and low-risk groups.**
    2. **Refine Pricing Models: Use the statistically significant findings (e.g., impact of driving children, high-risk car models) as inputs for reviewing and adjusting premium calculations.**
    3. **Develop Targeted Strategies: Create distinct marketing and retention strategies for the identified customer segments (e.g., "Segment 1," "Segment 2," "Segment 3").**
    4. **Investigate Regional Anomalies: Conduct a deeper dive into any CoverageZone that shows significantly higher claim rates than the average.**