INSURANCE PREMIUM ANALYSIS PROJECT

PREPARED BY:

ZAKARIA GATUNE

August 2025

ABSTRACT

This project presents a comprehensive analysis of customer data to uncover patterns in insurance premium payments, by examining variables such as age, gender, work status, marital status and customer tenure. The analysis employs descriptive statistics, groupwise comparison and visual summaries to uncover patterns and differences across customer segments. The findings aim to support data-driven decision making in pricing strategies, customer targeting and product design within the insurance industry.

Contents

Table of Figure

List of Tables

1.0TITLE:

REVENUE DRIVERS IN INSURANCE: ANALYZING FACTORS BEHIND HIGHER POLICY PREMIUMS

1.1 Objectives:

To identify the key customer characteristics that influence annual policy premium payments in an insurance dataset. The analysis helps the company understand its revenue drivers and segments high-value customers.

1.1 Data Source and Overview:

The dataset used in this project was obtained from Kaggle, a widely recognized platform for sharing datasets and data science competitions.

The datasets include 499 records with the following variables:

Table 2 Data Overview

Column (Variable)	Data Type	Description
Index	Integer	Unique row
Age	Integer	Age of the customer
Age-group	String	Grouped version of the age variable
Work_status	String	Employment status
Marital_status	String	Marital status of the customer (Eg, single or married)
Gender	String	Customer gender
Months_as_customer	Integer	Number of months the individual has been a customer
Policy_annual_premium	Float	Annual insurance premium paid by the customer
Incident date	Date	Date when a claim related incident occurred

Tools Used:

Excel – for data cleaning, grouping, pivot tables.

2.0 DATA CLEANING AND PREPARATION

Before analysis, the dataset underwent several cleaning and preprocessing steps to ensure accuracy, consistency and usability. This step focused on handling missing values, transforming data types and adding new features to enhance analytical insight.

I. Handling Missing Data

The work status column had 30 missing values (about 6% of the data set). Rather than dropping these records, which would lead to information loss, the missing values were replaced with the label “unknown” to preserve data integrity while retaining analytical flexibility.

II. Data Type Conversion

To ensure consistency and enable accurate analysis the following data transformations were performed:

Incident date – converted from string to proper date format to support time-based filter and potential trend analysis

Policy_Annual_Premium – converted to Kenyan Shillings (KES) for contextual relevance and easier interpretation of monetary values.

III. Creating New Variables

To enhance the dataset and enable better analysis new columns were added:

Age group – a new variable created by grouping customers into age brackets. This made it easier to analyze trends across different age segments.

Month as Customer range – a new variable created by grouping customers into levels, from level one to eight each having a range of 60 months (5 years).

3.0 EXPLORATORY DATA ANALYSIS (EDA)

This phased focused on understanding the structure patterns and relationships within the datasets.

3.1 Data Overview

The dataset contained 499 entries and 8 columns:

- Categorical variables: Gender, Marital_status, Work_status.
- Numerical Variables: Index, Age, Policy_Annual_Premium
- Date variable: Incident_date

3.2 Summary Statistics

Descriptive statistics were generated for key numerical features

- Age analysis

  The customers' age in the dataset ranged from a minimum of 17 years to a maximum of 90 years indicating a spectrum of insured individuals. I created a custom age grouping to categorize customers into a meaningful segment such as young adult, mid adults, seniors and other. Using Pivot Table in excel, I counted the number of customers in each group. This revealed that the young adult_25-34 years had the highest number of policy holders, while the Elderly_75-90 years had the fewest.
- Gender

  A pivot table was used to determine the number of customers by gender. The dataset includes both male and female. The majority of customers were males. Understanding gender distribution is important because it can influence targeted marketing strategies.
- Marital status

  The marital status field in the dataset includes categories: married, single and widowed. Using a pivotal table the distribution of marital status was analyzed. The majority of the customers were single, 52%, followed by single, 46%. The rest of the customers were widowed.
- Work Status

  The datasets include three work status categories: employed, unemployed and Unknown. A pivot table analysis showed that the majority of customer were

employed, 70% of the total customer. 24% were unemployed and the remaining percentage was unknown.

- Months as Customers

  This indicates how long each individual has been with the insurance company. In the dataset, customer tenure ranged from a minimum of 0 months to a maximum of 478 months. The average tenure was 204 months. This variation shows that the company serves both new and long-term clients.

- Policy Annual Premium

  This represents the amount (in KES) paid yearly by customer for their insurance policy. The minimum premium was KSH 433.33, while the maximum premium was KSH 2,047.59 The average premium was KSH 1,251.58. The distribution indicates that most customers fall within a premium range.

The Min, Max and Average values were obtained using Excel's *MIN, MAX* and *AVERAGE* functions.

4.0 DATA ANALYSIS AND MODELLING

4.1 Group-Wise Premium Analysis

1. AGE GROUP

To explore how insurance premium payments vary across age brackets, I used an excel pivot table to compute the average and count of policy annual premium for each defined age group. This allowed me to identify which age groups have the highest number of customers and pay the highest average premiums. The chart below shows the count and average policy annual premium for all age group.

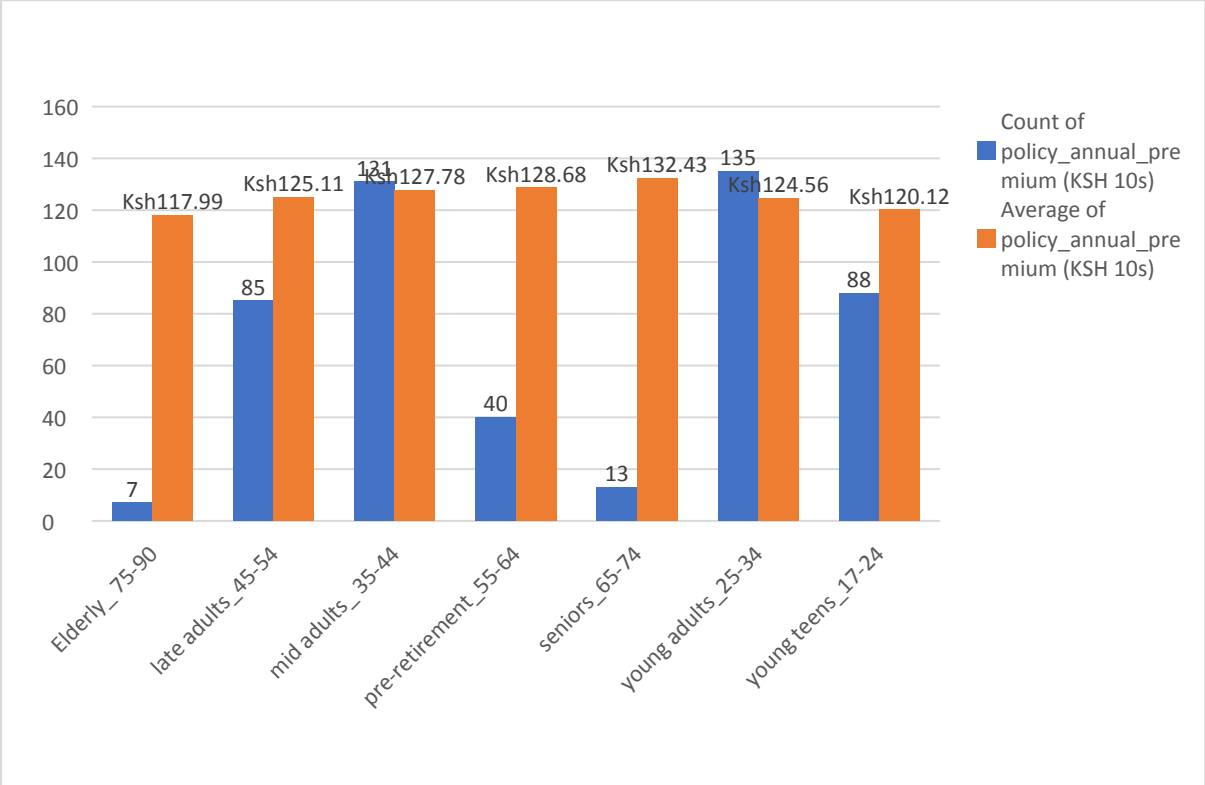


Figure 1 Count and Average of Policy Annual Premium by Age Group

From the chart it evident that young adults from the age of 25 to 34 years have the highest number of policy holders, indicating this group may be a key target market. Groups like elderly with 75 years and above and seniors with the age of 65 to 74 years have the lowest count, suggesting fewer people in these brackets are insured.

The highest average premiums are observed the seniors (65-74) and pre-retirement (55-64). This may reflect high income level, more valuable assets or broader coverage needs. The youngest and the oldest group tends to have lower average premiums, possibly due to lower income levels or limited insurance needs.

2. WORK STATUS

To access how employment status correlates with premium payments, I analyzed the count and average policy annual premium for each status category using pivot table and chart in excel as shown below.

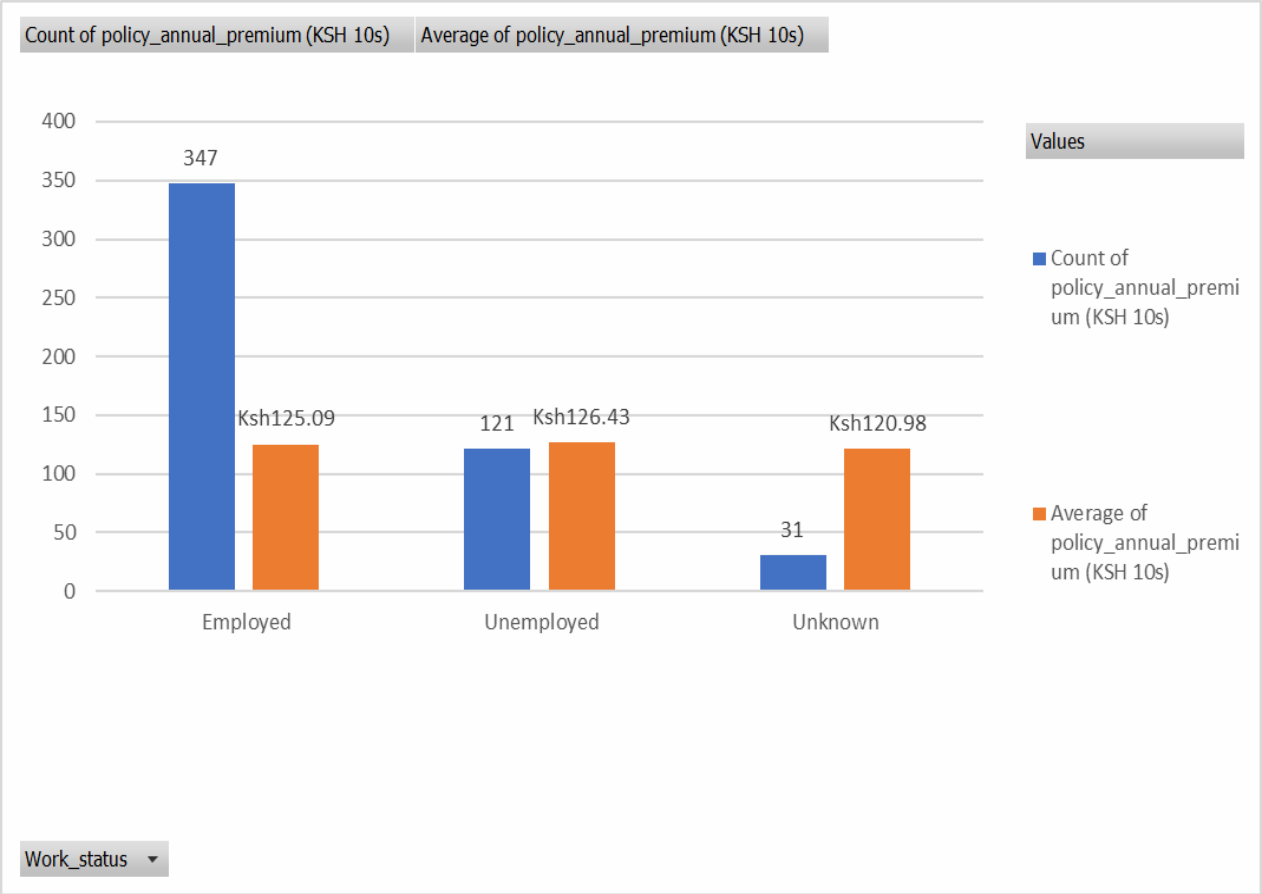


Figure 2 Count and Average of Policy Annual Premium by Work Status

The majority of the customers were employed accounting for 347 individuals but had an average premium of KSH 1,250.9, which is lower than that of unemployed customers who had the highest average premium of KSH 1,264.3. A total of 121 individuals were unemployed. The unknown group had the lowest premium of KSH 1,209.8 and smallest customer base of 31 individuals.

This suggest that employment status does not strongly influence premium size in this dataset, although further analysis could explore why unemployed individuals might have higher premiums.

3. GENDER

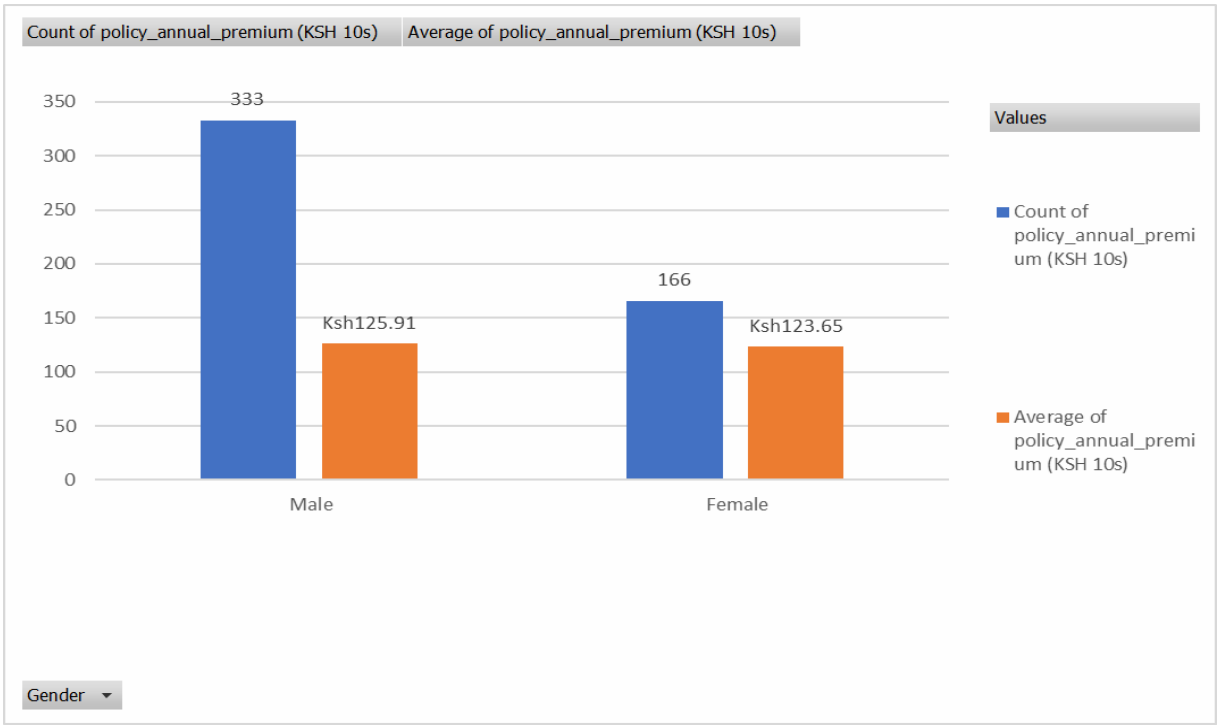


Figure 3 Count and Average of Policy Annual Premium by Gender

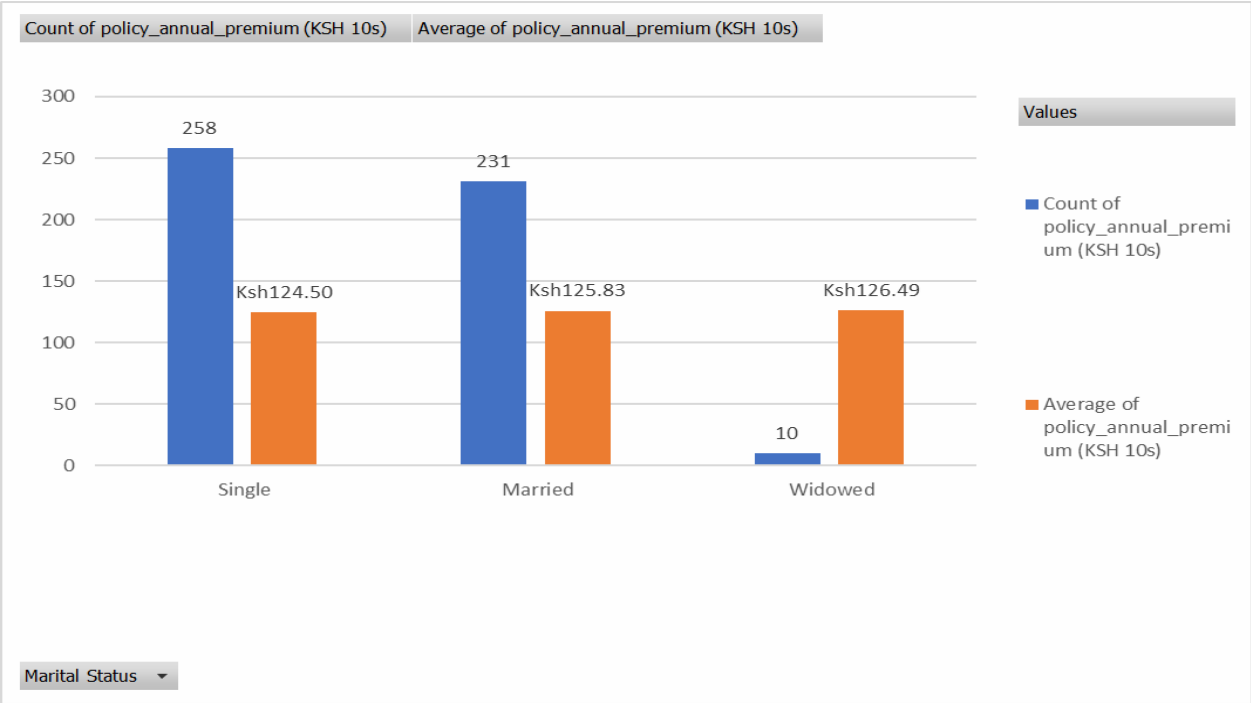
To assess if gender influences premium payments, I analyzed the count and average policy annual premium across gender categories using pivot table and charts in excel as shown above.

Males had the highest number of individuals (333), while female were 166 individuals. Although males were majority in the dataset, the difference in the average premium is minimal, indicating that gender may not significantly influence premium pricing. However, the higher number of males policy holders could have a moderate effect on the overall revenue contribution.

4. MARITAL STATUS

Figure 4 Count and Average of Policy Annual Premium by Marital Status

A pivot chart show that the majority of the policy holders are single, followed by married and small number of widowed as shown in the figure below.

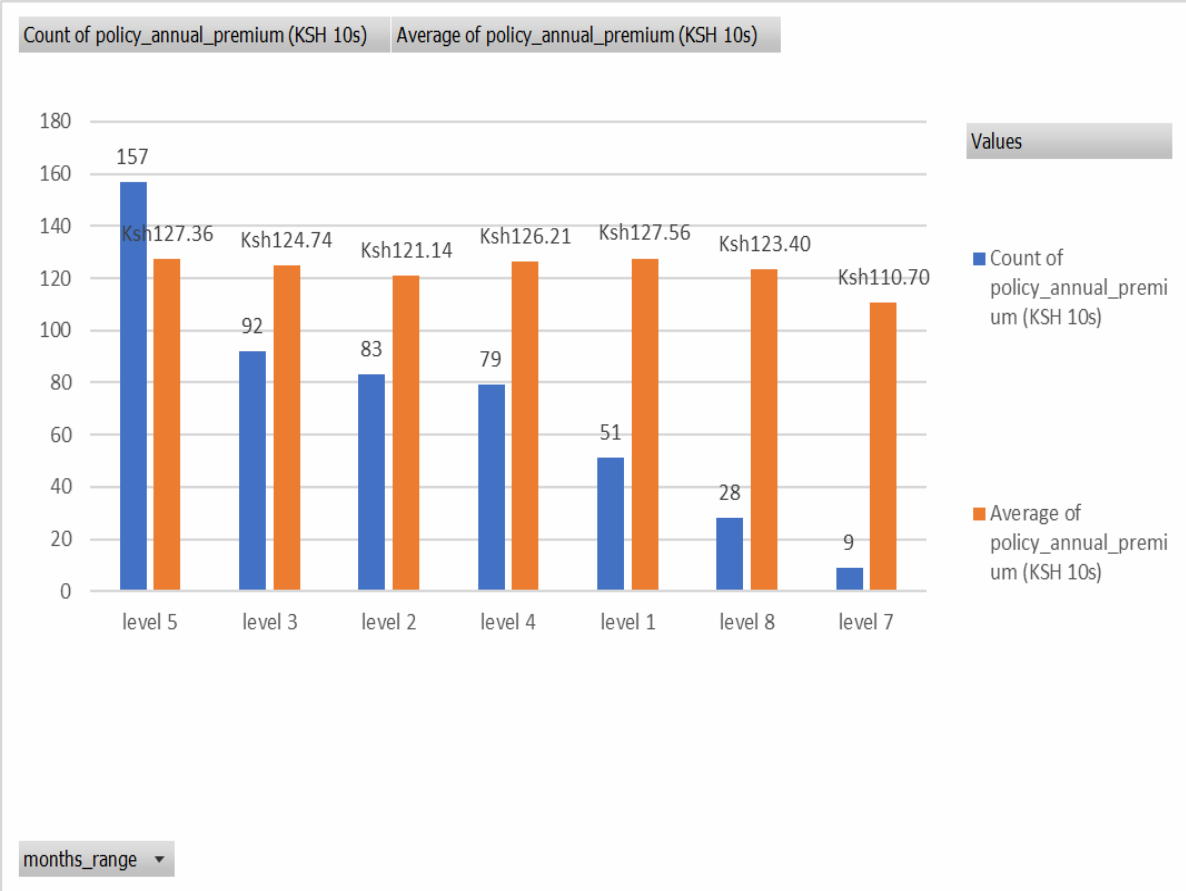


On average, widowed customers pay the highest annual premium at KSH 1,264.9, followed by married at KSH 1,258.3 and single at KSH 1,245.00. while the differences in premium payments are relatively small, this may suggest that married and widowed individuals tend to invest slightly more in their policies. However, since single people make up the largest portion of the policy holders, they remain a key demographic for targeting insurance products.

5. MONTH AS CUSTOMERS

The dataset was segmented into eight tenure levels each representing a five-year range of customers duration. Excel pivot chart was used to show the count and average of each level as shown below.

Figure 5 Count and Average of Policy Annual Premium by Customer Tenure



Level 5 (20-25 years) recorded the highest number of policyholders, with an average annual premium of KSH 1,273.60. Other significant groups include level 3 (10-15 years) with 92 customers and an average premium of KSH 1,247.4 and level 2 (5-10 years) with 83 customers and an average of KSH 1,211.40.

Notably, level 1(0-5 years) showed the highest average premium despite having a lower customer count. In contrast, level 7 (30-35 years) recorded the lowest average premium at KSH 1,107.00. While fluctuation in premiums exist across the tenure levels, no consistent trend was observed, suggesting that factors other than customer duration may influence premium pricing.

4.2 Conclusion

The analysis revealed that customers aged between 55 and 74 consistently paid higher insurance premiums, highlighting them as a key demographic insurance product. Although males formed the majority of the customers base, the difference in average premium between gender was minimal, suggesting equitable pricing across gender. Unemployed individuals recorded slightly higher average premiums than employed one, possibly reflecting specific risk profiles or product choices, marital status showed little variation in premium levels, with

widowed clients slightly leading. Additionally, customers in the mid-tenure category contributed the highest counts and premium averages, indicating a strong segment for retention strategies. These insights can guide tailored marketing, pricing adjustment and customer relationship strategies to enhance business outcomes.

Customer Profile with the Highest Premium Potential

- Age Group: Pre-Retirement
- Work Status: Unemployed
- Gender: Male
- Marital Status: Widowed
- Month as Customer: Level 5 (20-25 years)

This composite profile helps the business identify high-value customer segments that may require tailored services, retention strategies or targeted marketing. Understanding what combination of demographic is associated with higher spending can guide product design and customer relationship strategies.

Customer Profile with the Lowest Premium:

- Age Group: Elderly 75-90 years
- Work Status: unknown
- Gender: Female
- Marital Status: single
- Month as Customer: Level 7 (30-35 years)

5.0 DOCUMENTATION

This project was executed in Microsoft Excel, leveraging its powerful data manipulation tools such as pivot tables, sorting and filtering and custom grouping features. The dataset underwent initial cleaning including removal of duplicates, handling of missing values and formatting numerical variables for consistency. Key functions and formulas like Averageif, Countif, If, Max, Min and Average statements were utilized during the analysis. Data visualization was performed using excel charts to highlight patterns in customer premiums across demographic segments. Pivot tables were essential for computing group wise statistics such as count, sum and average based on variables.

The project structure followed a logical analytical flow from problem definition through explanatory analysis to interpretation and business insight generation. All findings were documented clearly with visualization embedded and explained to support business making. The entire workflow and logic were maintained in a single workbook, and file naming was standardized for easy navigation. Overall, the documentation provides a transparent view of the methods used assumptions made and insights derived from the data.