Acko Health Insurance: Data-Driven Insurance Premium Pricing Strategy

Introduction:

Acko is an online insurance company in India that offers a variety of insurance products, including car, bike, health, and travel insurance.

Acko, a digital insurance provider, is launching a new health insurance product and requires a data-driven approach to determine the optimal insurance premium pricing for different customer segments.

Objective:

- To determine an optimal premium pricing strategy that balances business profitability with customer affordability, leveraging available demographic, financial, and behavioral data.
- The primary objective is to develop a rule-based logical formula to calculate the final premium price for each customer based on their demographic, risk and affordability profile.

Business Impact:

Increased Profitability:Ensures premiums reflect customer risk levels, reducing losses from underpricing and maximizing revenue.

- Better Customer Acquisition & Retention: Affordable, customized premiums attract new customers and improve loyalty by offering fair pricing.
- Faster Market Launch:A simple rule-based model allows quick product rollout and easy updates for new customer segments.
- Enhanced Trust & Compliance:Transparent pricing builds customer trust and ensures alignment with regulatory standards.

Dataset Information:

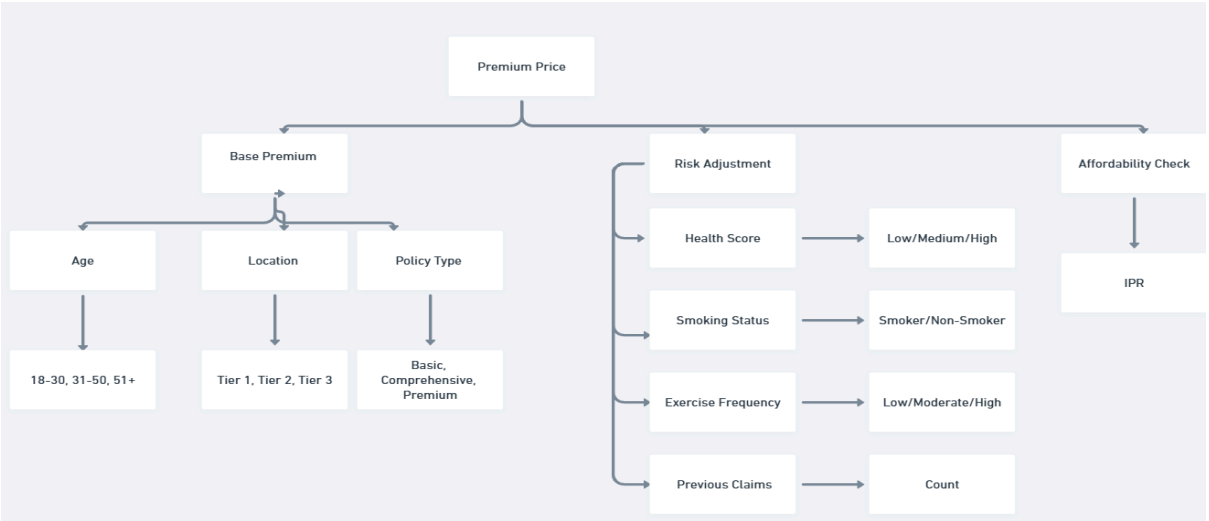
- Dataset: Acko Dataset
- Count of Rows : 1200000
- Count of Columns: 20

Explanation of Data Columns:

- **id (Customer ID)**: A unique identifier assigned to each customer in the dataset.
- **Age (Customer Age)**: The age of the customer in years at the time of policy purchase.
- **Gender (Customer Gender)**: The gender identity of the customer, which can be "Man" or "Woman."
- **Annual Income (Yearly Earnings in INR)**: The total income earned by the customer in a year, measured in Indian Rupees (INR).
- **Marital Status (Customer's Marital Condition)**: The marital status of the customer, such as "Spouse Present," "Not Married," or "Formerly Married."
- **Number of Dependents (People Financially Dependent on Customer)**: The number of dependents (such as children, parents, or others) that rely on the customer financially.
- **Education Level (Highest Education Attained)**: The highest level of education completed by the customer, such as "Undergraduate" or "Post Graduate."
- **Occupation (Customer's Job Type)**: The profession or employment category of the customer. In some cases, this data may be missing.
- **Health Score (Overall Health Indicator)**: A numerical score representing the customer's health condition based on lifestyle factors and medical history.Less score mean fit, more means unfit.
- **Location (Customer's Residence Tier)**: he classification of the customer's residence area into tiers such as Tier-1, Tier-2, or Tier-3 cities.
- **Policy Type (Type of Insurance Policy Chosen)**: The category of insurance policy purchased by the customer, such as "Basic," "Premium," or "Comprehensive."

- **Previous Claims (Number of Past Insurance Claims)**: The total number of insurance claims the customer has made before purchasing the current policy.
- **Credit Score (Financial Responsibility Indicator)**: A numerical representation of the customer's creditworthiness, indicating their ability to manage finances and make timely payments.
- **Insurance Duration (Policy Tenure in Years):**: The number of years the customer has held the insurance policy.
- **Policy Start Date (Date When Policy Became Active)**: The date when the insurance policy was purchased and became active.
- **Customer Feedback (Customer Satisfaction Rating)**: The rating or feedback provided by the customer about their experience with the insurance policy.
- **Smoking Status (Whether the Customer Smokes)**: Indicates whether the customer is a smoker or not, as smoking impacts health risks and insurance premiums.
- **Exercise Frequency (How Often the Customer Exercises)**:The frequency of physical exercise performed by the customer, which can impact their health score.
- **Property Type (Type of Residence)**: The type of home the customer resides in, such as a detached home or an apartment.
- **Premium Amount (Final Insurance Premium in INR)**: The amount the customer pays for their health insurance policy, measured in Indian Rupees (INR)

KPI Tree:



Data Cleaning and Preparation:

1.Data Import: The CSV file containing Acko Health Insurance data was imported from the drive into Google Colab for further processing.
2.Finding null Values: The Dataset contains no null values in all columns. No modification needed here.

3.Checking Datatypes: Verifying the data types of each column is important to ensure they are correctly formatted.

4.Timestamp: The Policy Start Date column was in object datatype. We converted it to a consistent datetime format

5.Missing Data Handling : We filled the missing values by grouping data with relevant columns to capture patterns and applied the overall median/mode as a fallback where needed. This approach preserves data consistency and reflects real-world relationships. For age Filled missing values using the median within Occupation and Policy Type groups, For Credit Scoremedian within Occupation groups.
6.Handling Outliers: We use capping to handle outliers by limiting extreme values within a defined range. This prevents unusually high or low values from skewing the analysis and ensures the data remains realistic and consistent.For Annual Income, we applied 5th-95th percentile,For Premium Amount, we used 4th-96th percentile capping,For Previous Claims, we capped values at 6 to prevent excessive claims

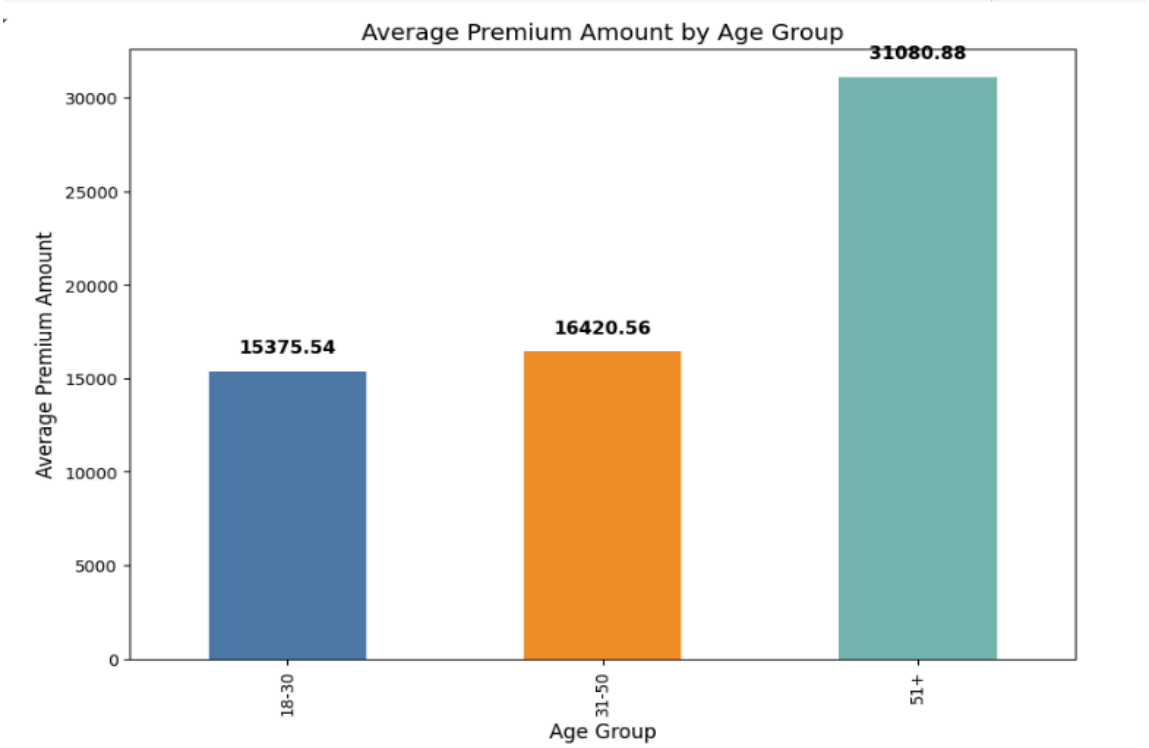
7.Creating new Columns: We created two new columns for easier analysis Age Group: Categorized ages into groups like 18-30, 31-50, 51+,Categorize Health Score: Classified health scores into categories like Low, Medium, and High

Exploratory Data Analysis(EDA):

Hypothesis 1:

Older Age Groups Pay Higher Health Insurance Premiums:

Age Group	count	mean	median
18-30	319083	15375.54	10718.00
31-50	505676	16420.56	11247.49
51+	373681	31080.88	23690.35



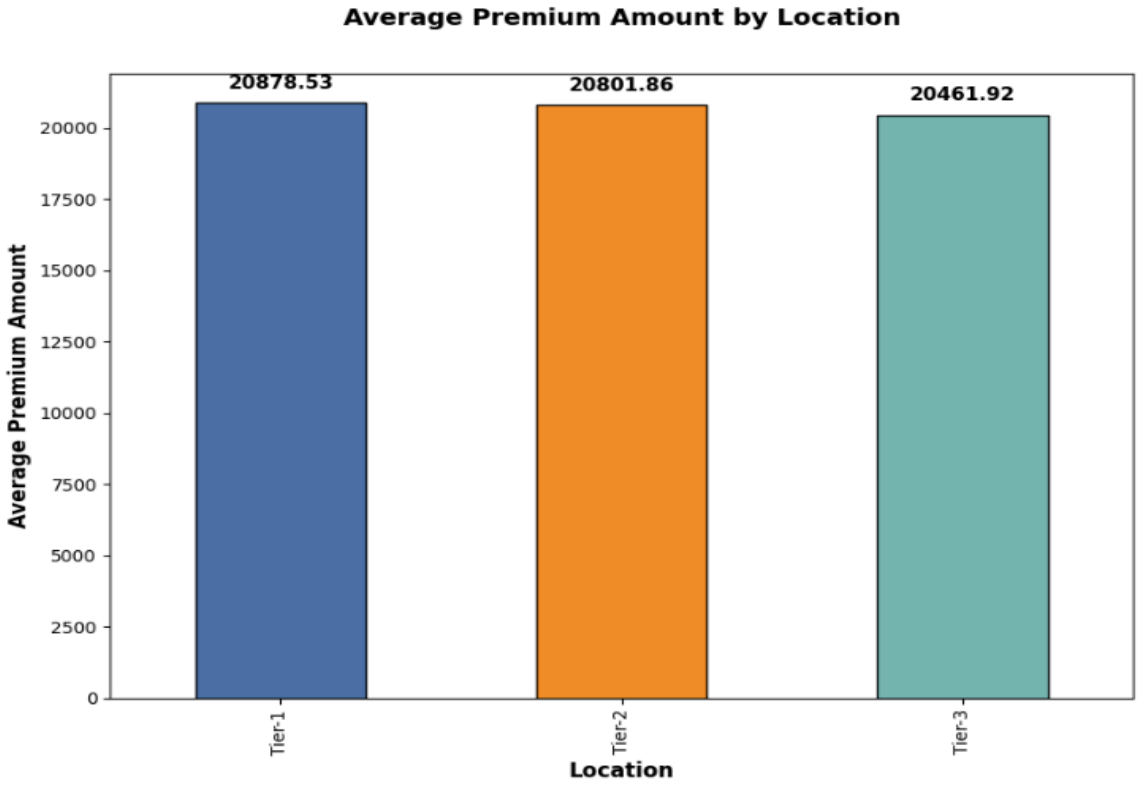
Findings: We analyzed the relationship between age groups and health insurance premiums by comparing the mean premiums across age categories. The findings confirm that older age groups pay significantly higher premiums, with the 51+ group paying nearly twice the median premium of the 18-30 group.

Recommendation: Our analysis indicates that the average premium price increases with age. Therefore, we have incorporated the age factor into our premium calculation formula."

Hypothesis 2:

Health insurance premiums vary across locations due to differences in healthcare access and costs:

Location	count	mean	median
Tier-1	397001	20878.53	12756.78
Tier-2	400416	20801.86	12952.51
Tier-3	401023	20461.92	12915.80



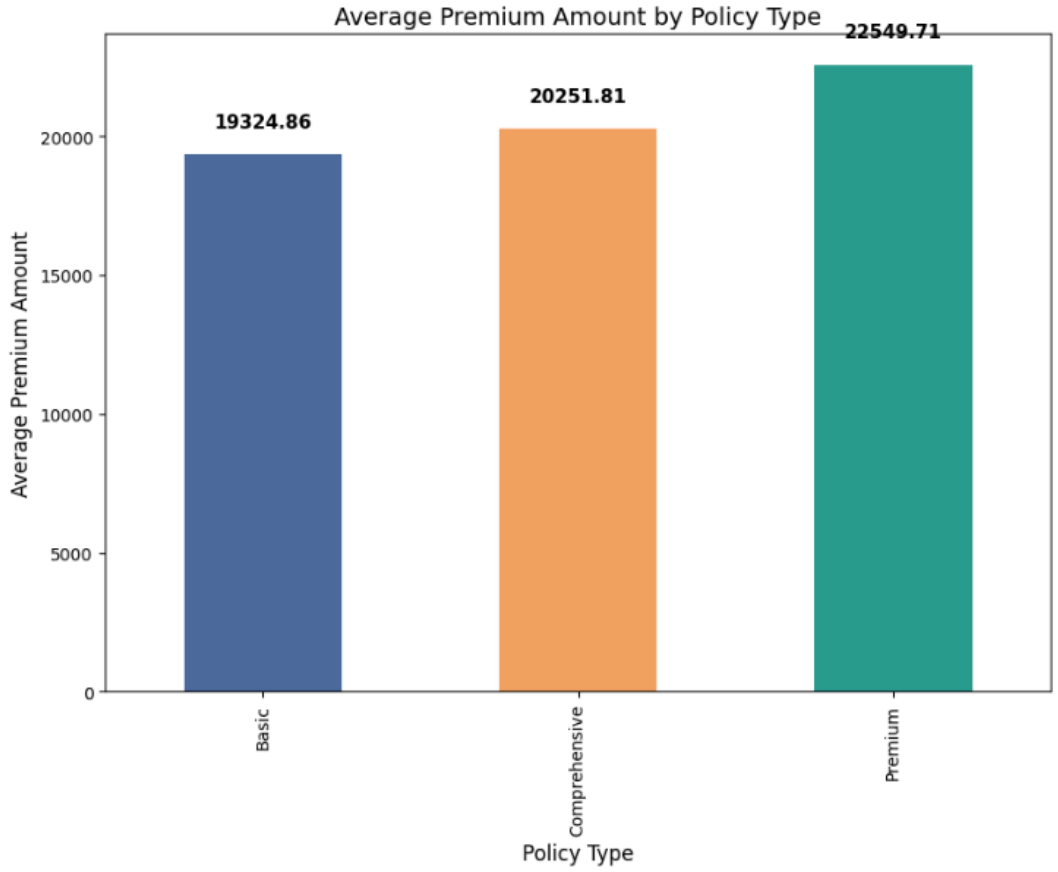
Findings: We analyzed the relationship between location tiers and health insurance premiums by comparing mean values across Tier-1, Tier-2, and Tier-3. The analysis shows that Tier-1 customers pay slightly higher premiums, while Tier-2 and Tier-3 premiums are nearly identical. This partially supports our hypothesis, as only Tier-1 shows a clear premium increase.

Recommendation: Our analysis shows that Tier-1 customers pay slightly higher premiums, while Tier-2 and Tier-3 premiums are similar, making our hypothesis partially true. Given this insight, we recommend incorporating the location factor into our premium calculation formula, with a location-based adjustment for Tier-1 customers.

Hypothesis 3:

Policy Type Influences Health Insurance Premium Amounts:

Policy Type	count	mean	median
Basic	398008	19324.86	10825.14
Comprehensive	399087	20251.81	11397.64
Premium	401345	22549.71	12952.51



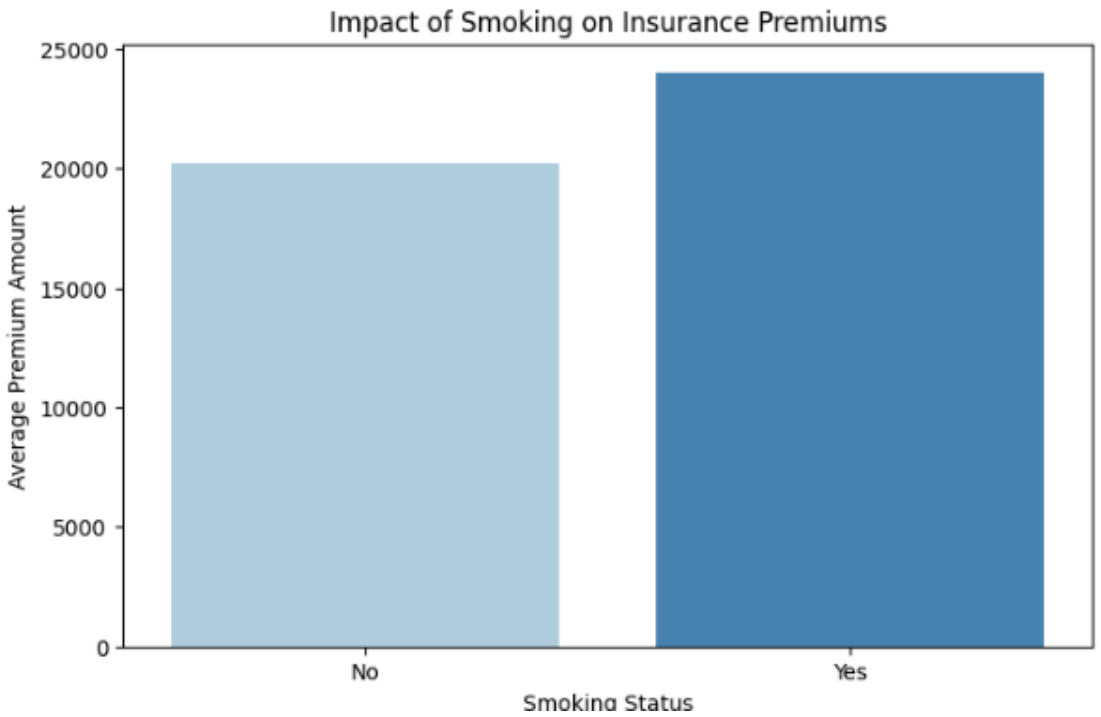
Findings: We analyzed the relationship between policy type and health insurance premiums by comparing the mean premiums across different policy categories. The findings confirm that certain policy types are associated with significantly higher premiums,

Recommendation: Incorporate policy type as a key factor in the premium calculation formula by assigning multipliers based on coverage levels, ensuring premiums reflect the cost differences across policy categories.adjustment for Tier-1 customers.

Hypothesis 4:

Smokers tend to have higher insurance premiums compared to non-smokers.

	Smoking Status	Premium Amount
0	No	19405.28
1	Yes	22013.74



Finding: Smokers pay higher health insurance premiums (₹22,011.66) compared to non-smokers (₹19,401.32).

Recommendation: True Smoking status increases the premium amount, we can use this factor in our formula

Hypothesis 5:

Individuals who exercise more frequently (Daily/Weekly) tend to have lower health insurance premiums compared to those who exercise less frequently (Monthly/Rarely).

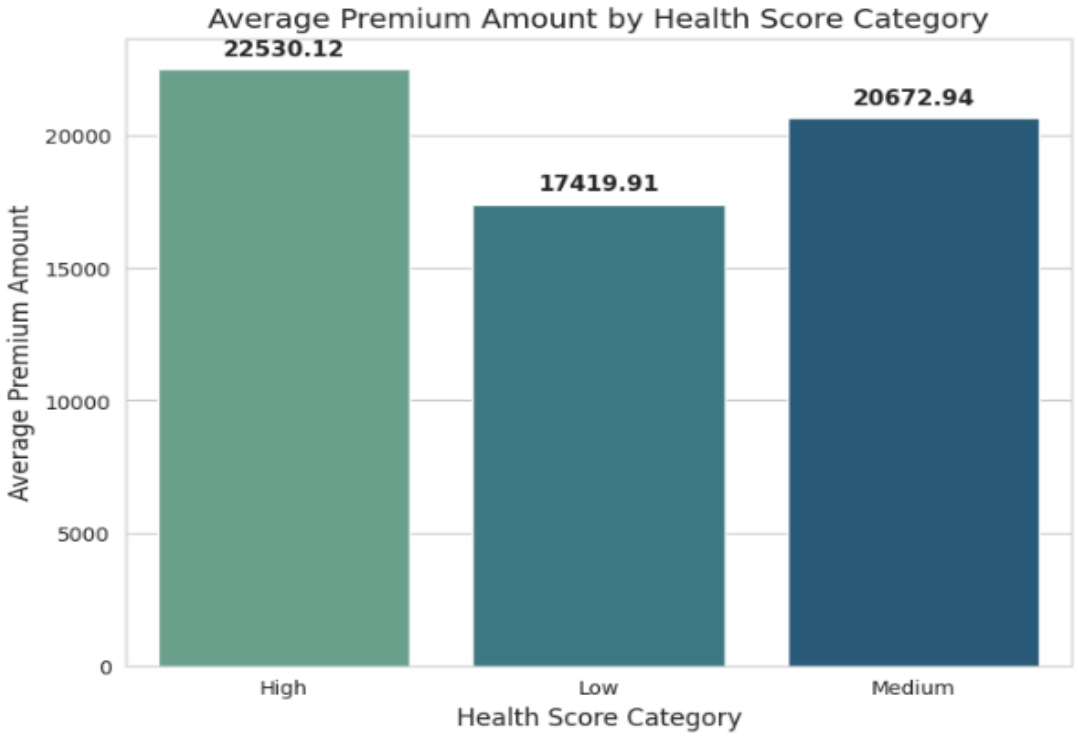
	Exercise Frequency	Premium Amount
0	Daily	18960.46
3	Weekly	18968.28
1	Monthly	19879.45
2	Rarely	25057.54

Findings: Individuals who exercise less frequently (Monthly or Rarely) have higher average health insurance premiums, with the "Rarely" group paying the most.

Recommendation: True – More frequent exercise (Daily/Weekly) is associated with lower health insurance premiums,we can use this factor in our formula

Hypothesis 6:

Higher the Health Score, higher the Premium Amount.



Findings: As mentioned in data lower the health score means more fit, we find that lower the health score also have lower average premium price and higher the health score higher the average premium price.

Recommendation: we can use this factor in formula as the by category average premium prices differ.

Hypothesis 7:

Higher Previous Claims Lead to Increased Premium Prices.

	Previous Claims	Premium Amount
0	0.0	23097.32
1	1.0	18835.50
2	2.0	22888.42
3	3.0	23065.29
4	4.0	23487.99
5	5.0	24176.75
6	6.0	24572.97

Finding:Premium amounts generally increase as the number of previous claims rises, with a noticeable upward trend from 1 to 6 claims.

Recommendation:Partially True – While premium amounts increase with higher claims, the premium for 1 claim is lower than for 0 claims, indicating some variation in the trend, we can use this factor in our formula.

Formula Making:

Based on our hypothesis testing, we found that age group, location, policy type, smoking status, health score, exercise frequency, and previous claims all significantly impact premium pricing. Therefore, we have incorporated these factors into our premium calculation formula to ensure more accurate and fair pricing.

Here are the steps of the formula:

1.Base Premium Calculation(B):

We calculate the median premium for each Policy Type (Basic, Comprehensive, Premium).

This Base Premium acts as a reference point to measure premium variations across Age Groups and Location.

2.Age Factor Calculation:

We group the data by Age Group (18-30, 31-50, 51+). For each group, we calculate the mean premium.

Age Factor is computed using the formula:

$$\text{Age Factor} = \frac{\text{Mean Premium for Age Group} - \text{Base Premium}}{\text{Base Premium}}$$

This measures how much **Age** influences the premium relative to the **Base Premium**.

3.Location Factor Calculation:

We group the data by Location (Tier-1, Tier-2, Tier-3).For each location, we calculate the mean premium.

Location Factor is computed using the same formula as the **Age Factor**:

$$\text{Location Factor} = \frac{\text{Mean Premium for Location} - \text{Base Premium}}{\text{Base Premium}}$$

4.Risk Adjustment Calculation:

We apply Risk Adjustment Factors to customize premium prices based on individual risk characteristics. These factors account for variations in health behavior and past claims.

1. **Smoking Status:**
    - We calculate the mean premium for Smokers and Non-Smokers.
    - Smokers are charged a higher premium based on the percentage difference from Non-Smokers.
2. **Health Score:**
    - Premiums are adjusted based on Low, Medium, or High health scores.
    - Low scores increase premiums, while High scores lower them, with Medium as the baseline.
3. **Exercise Frequency:**
    - Customers with Low exercise frequency face higher premiums, while High frequency reduces them.
    - Adjustments are based on the difference from the Moderate exercise group.
4. **Previous Claims:**
    - We compare the mean premium by the number of prior claims against no claims.
    - More claims lead to a higher premium due to increased risk.

These factors are integrated into the premium calculation formula, ensuring fair and risk-based pricing.

Final Formula - Final Premium = **Base Premium** × (1 + **Total Factor**)
Base Premium(B) - Median of particular policy type.
TotalFactor - age_factor + location_factor + smoking_factor + exercise_factor + health_score_factor + previous_claims_factor

Affordability Of Price

```
df['IPR'] = (df['Final Premium'] / df['Annual Income']) * 100
```

The Income-to-Premium Ratio (IPR) is a financial metric that measures how much of a person's Annual Income is spent on their Health Insurance Premium. It is expressed as a percentage and helps assess the affordability of the insurance policy.

The IPR (Insurance Premium to Income Ratio) helps measure how affordable a health insurance premium is relative to annual income. The thresholds (2, 4, 12) are based on financial guidelines:

≤ 2%: Higher Affordability – Minimal financial burden.

2% to 4%: High Affordability – Reasonable premium cost.

4% to 12%: Moderate Affordability – Manageable but higher expense.

12%: Low Affordability – Significant financial strain.

These values are chosen to reflect practical affordability levels and industry benchmarks.

```
Affordability_Category
Moderate Affordability   33.09
Higher Affordability     25.35
High Affordability       21.42
Low Affordability        20.14
Name: proportion, dtype: float64
```

The Income-to-Premium Ratio (IPR) analysis reveals how affordable health

insurance premiums are relative to annual income. Most customers fall within the 2% to 12% range, indicating that premiums are generally manageable. However, customers with IPR > 12% face significant financial strain, suggesting a need for targeted premium adjustments or customized policy options to improve affordability.

By incorporating IPR into our pricing model, we ensure balanced premium calculations that align with customer affordability and industry benchmarks, enhancing both customer retention and financial sustainability.

Overall Analysis:

- Age directly influences premium amounts—older individuals pay higher premiums—making age a true factor for premium calculation adjustments.
- Tier-1 cities have slightly higher premiums compared to Tier-2 **and* *Tier-3, but the variation is small. This indicates location has a minor effect on premium pricing.
- Premium policyholders pay the highest premiums, followed by Comprehensive, and Basic policies have the lowest premiums. Policy type directly affects premium calculations.
- Smokers consistently pay higher premiums than non-smokers, confirming a direct impact.
- Higher health scores result in higher premiums, while lower health scores correspond to lower premiums,
- Zero claims are associated with higher premiums, while individuals with 1 to 6 prior claims have increasing premium amounts.
- Regular exercise reduces health risks, leading to lower premiums, confirming that exercise frequency directly affects premium costs

Recommendations:

- Target Low & Moderate Affordability Groups (53.23% Total):Offer customized discounts or flexible payment plans to make premiums more manageable.
- Affordable Policy Variants: Design basic coverage plans tailored for Low and Moderate Affordability groups to capture a broader market.

- Income-Based Premium Plans:Implement tiered pricing tied to income brackets to improve affordability and attract new customers.
- Segment-Specific Marketing:Focus marketing efforts on the Low Affordability segment by highlighting cost-saving benefits and value-added services.
- Focus Offers on the Low Affordability Segment: Discounts & Payment Flexibility: Provide limited-time discounts or EMI options to ease the financial burden.
- Incentives for High Affordability Segment:Premium Upgrades: Offer value-added services (e.g., dental, vision, or wellness programs) to encourage policy upgrades.