DATA PREPARATION AND EXPLORATORY ANALYSIS OF INSURANCE CUSTOMER DATA

PAMPATI SREYA
CHRIST UNIVERSITY

TABLE OF CONTENTS

INTRODUCTION

- The insurance industry is highly data-driven.
- Customer insights help retain clients and reduce churn.

Key focus areas:

- Customer demographics
- Policy behavior
- Claim patterns
- Satisfaction levels

Goal: Extract actionable insights from customer data.

Business impact:

- Improve satisfaction & loyalty
- Personalize services
- Design targeted campaigns
- Optimize pricing & product offerings

DATASET OVERVIEW

	Customer_ID	Age	Gender	Policy_Type	Premium	Claim_Count	Region	Date_Joined	Customer_Satisfaction
0	CUST07402	37	male	Auto	668.47	1.0	West	2019-08-17	1.0
1	CUST05835	46	Other	health	146.62	4.0	north	2017-04-26	6.0
2	CUST02123	21	M	travel	810.64	1.0	east	2018-01-04	4.0
3	CUST08789	30	m	Travel	675.57	2.0	East	2013-05-20	6.0
4	CUST00305	49	male	home	723.36	0.0	E	2016-11-24	4.0

- Synthetic dataset with 10,000 rows generated for insurance customer analysis
- 1% duplicated rows added to simulate real-world redundancy
- Inconsistent entries in Gender, Region, and Policy_Type
- Outliers present in Premium, Claim_Count, and Customer_Satisfaction
- Missing values in Gender, Claim_Count, and Customer_Satisfaction

DATASET OVERVIEW

Why Synthetic Data?

- Allows intentional inclusion of issues like missing values, outliers, and inconsistencies for testing data cleaning workflows.
- Enables working with large datasets (10,000+ rows) to simulate real-world scenarios.
- Helps design targeted features (e.g., tenure, satisfaction levels) for deeper segmentation analysis.

DATA QUALITY CHALLENGES

- Missing Values in key fields like Gender, Claim Count, and Customer Satisfaction.
- Duplicate Records (~1% duplicated rows).
- Inconsistent Categorical Formats – e.g., different casing or typos in Gender, Policy_Type, Region.
- Outliers in Premium, Claim Count, and Customer Satisfaction impacting averages and distributions.
- Date Issues – need to extract meaningful features like tenure, join year, etc.

Why This Matters:

- Inaccurate data can mislead the analysis, skewing results and insights.
- Proper formatting ensures reliable aggregations and accurate visualization.
- Clean data is essential for effective segmentation, modeling, and decision-making.
- Outliers and duplicates can inflate metrics and hide true trends.

METHODOLOGY

- Data Cleaning & Transformation
- Feature Engineering
- Demographic Insights
- Customer Segmentation Insights
- Regional Insights

DATA CLEANING & TRANSFORMATION

1. Removed Duplicates
    - Identified and dropped ~1% duplicate rows to ensure data integrity
2. Handled Missing Values
    - Gender:
        - Standardized inconsistent entries (e.g., "FEMALE", "f") into uniform categories.
        - Imputed missing values with "Other" to avoid skewing toward the most frequent category.
    - Claim Count:
        - Applied a K-Nearest Neighbors (KNN) inspired imputation
        - Strategy varied by Policy_Type:
            - Used nearest values by Premium and Age for Health, Auto, and Life
            - Used nearest values by Premium and Region for Home and Travel
    - Customer Satisfaction: Imputed missing values using the mode grouped by Region

DATA CLEANING & TRANSFORMATION

3. Resolving Inconsistencies

- Normalized categorical values across Gender, Region, and Policy_Type
- Corrected case mismatches and typos (e.g., "nortH" → "North")

4. Outlier Treatment

- Identified outliers in Premium, Claim Count, and Customer Satisfaction
- Chose to retain outliers in Premium and Claim Count as they reflect real-world behavior (e.g., high-risk policies or high-value customers)
- Clipped outliers in Customer Satisfaction by capping values >10 to 10 & <1 to 1

FEATURE ENGINEERING

Tenure-Based Features

- Extracted Tenure_Days from Date_Joined
- Created Tenure_Bracket with bins: <1 Year, 1–3 Years, 3–5 Years, >5 Years

Age Grouping

Categorized customers into Age_Group: 18–30, 31–45, 46–60, 61–80

Premium Range Classification

Binned Premium values into:
Low, Medium, High, Very High based on quantiles

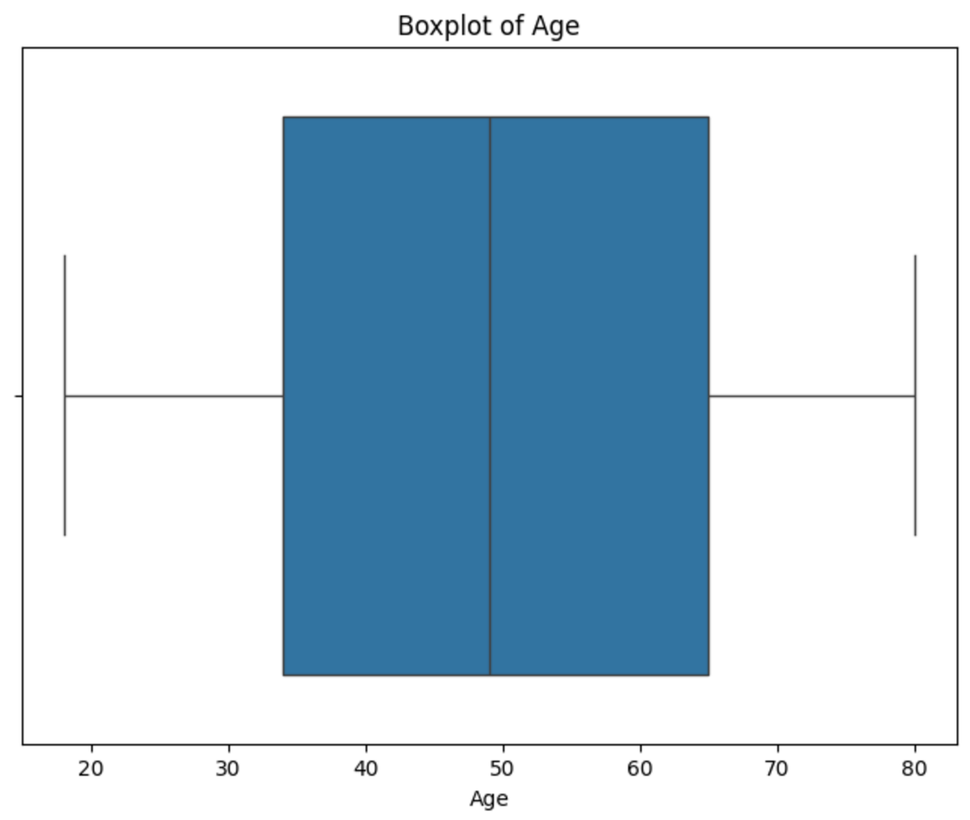
High Claim Flag

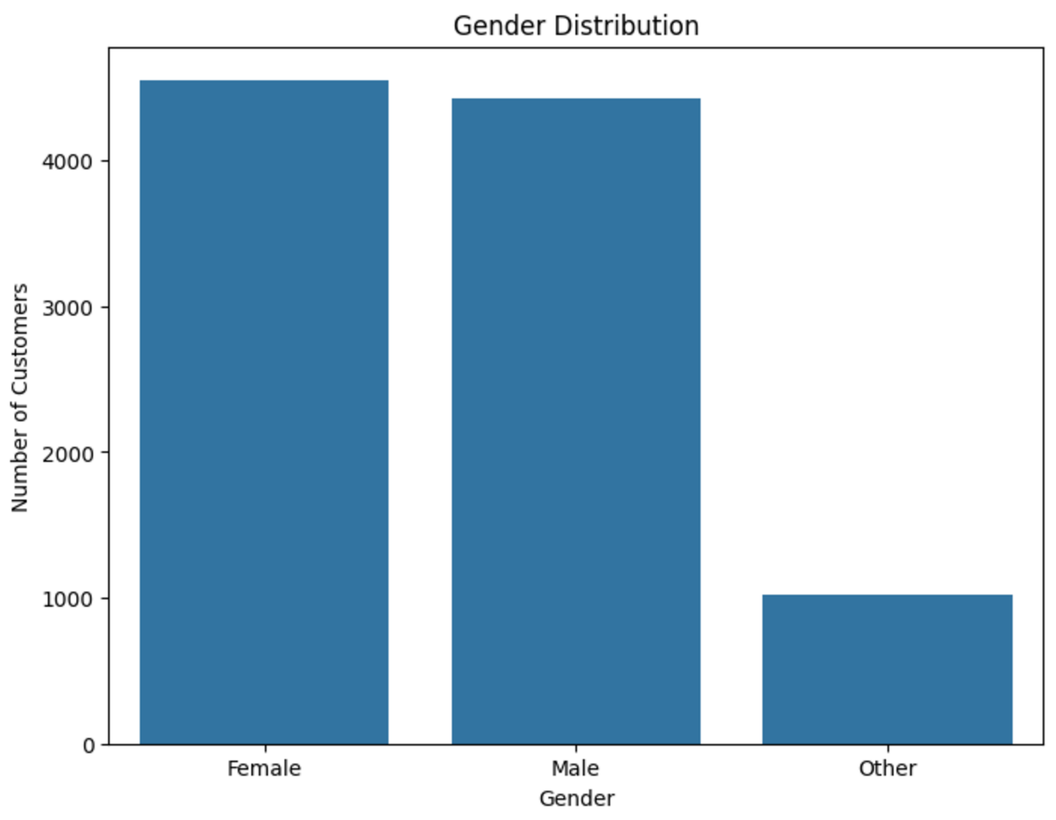
Identified customers with Claim_Count >= 3 and tenure in <1 Year or 1–3 Years

Satisfaction Level Encoding

- Binned Customer_Satisfaction into: Low (≤4), Medium (5–7), High (≥8)
- Retained the column name as Customer_Satisfaction for consistency

DEMOGRAPHIC INSIGHTS





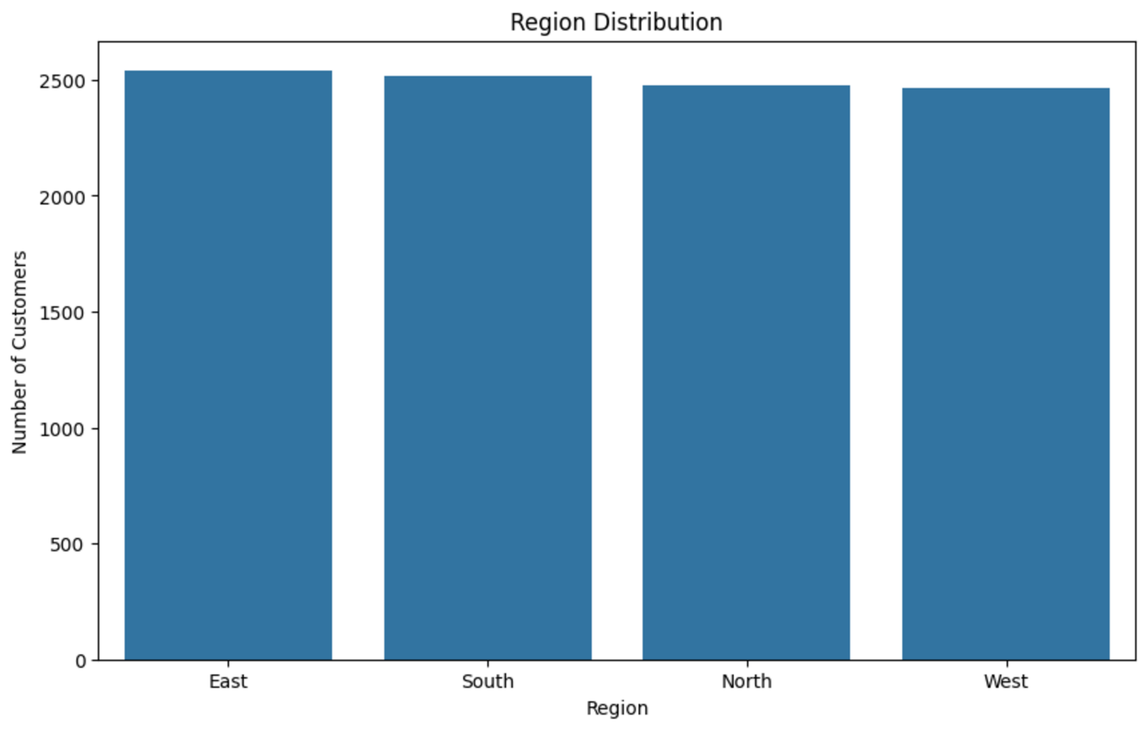
AGE DISTRIBUTION

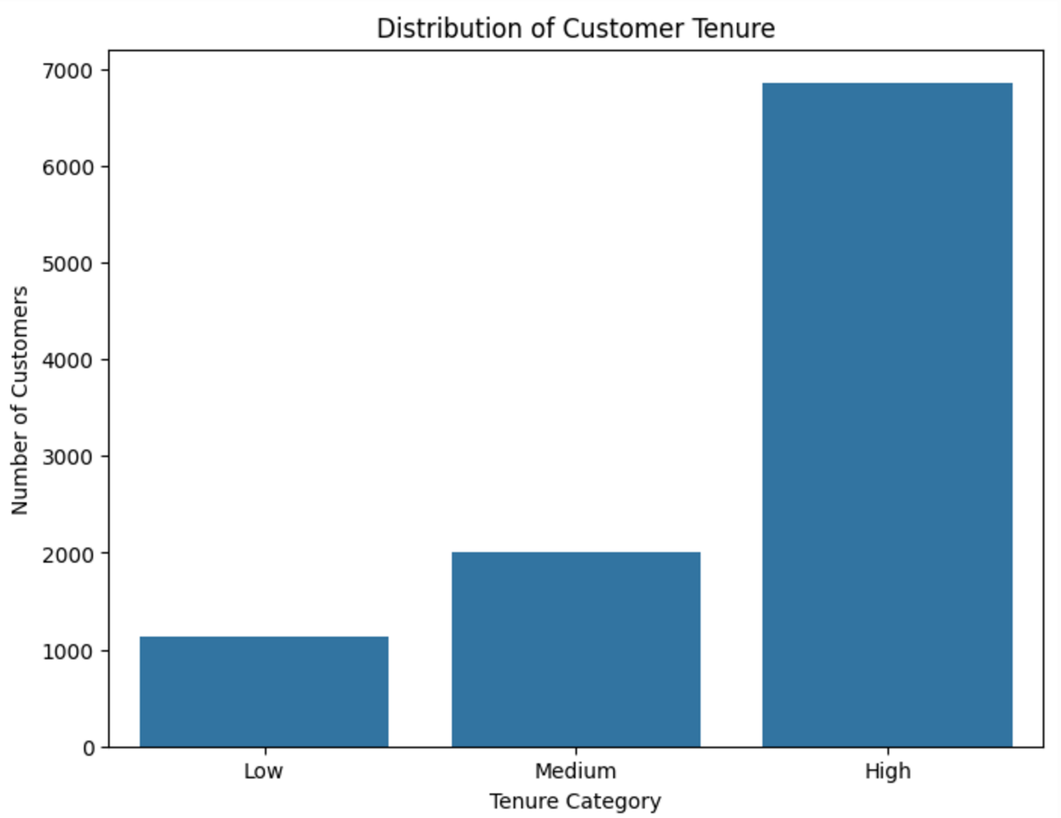
Majority customers fall in the 32–64 age range

GENDER DISTRIBUTION

Slight female dominance; inconsistent entries cleaned and imputed

DEMOGRAPHIC INSIGHTS





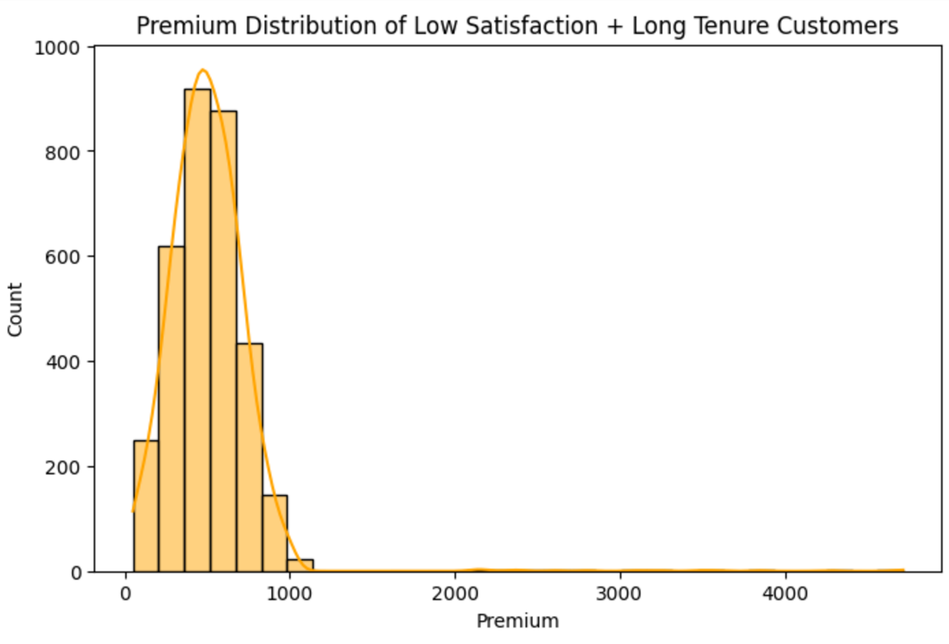
REGION SPREAD

Fairly balanced across East, West, North, and South

TENURE

Many customers with high (>5 yrs) tenure

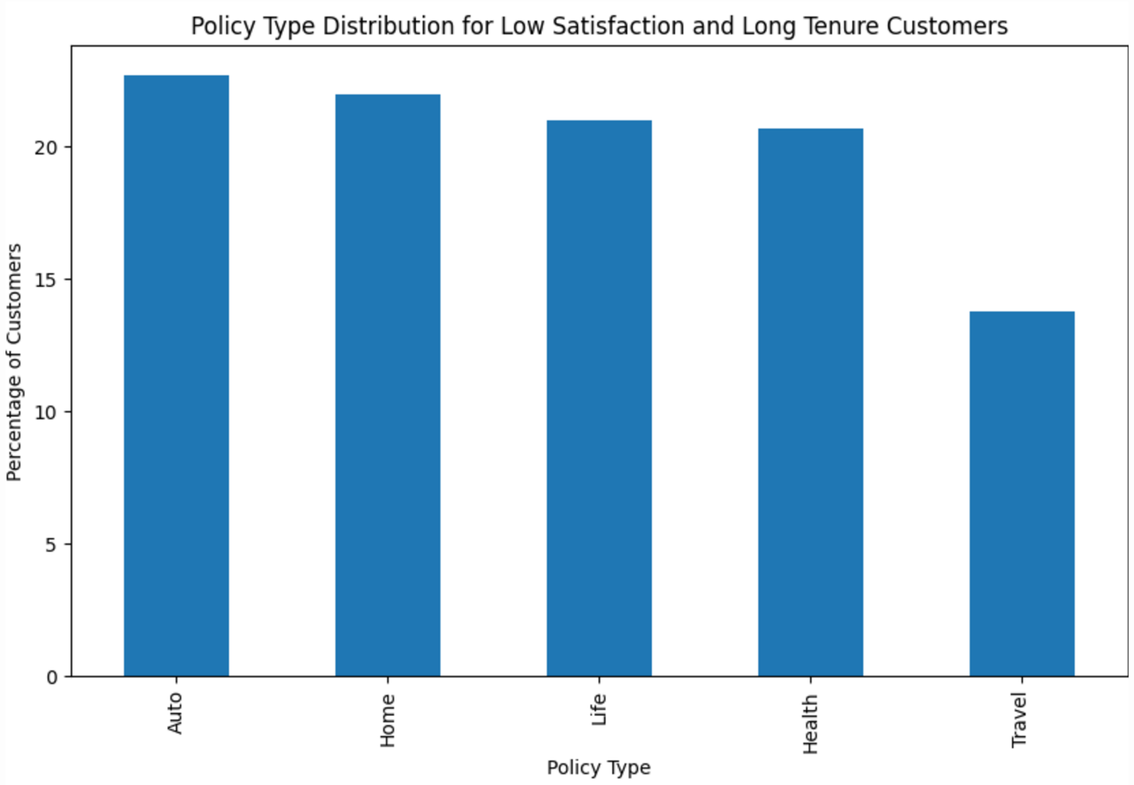
CUSTOMER SEGMENTATION BASED ON SATISFACTION & TENURE



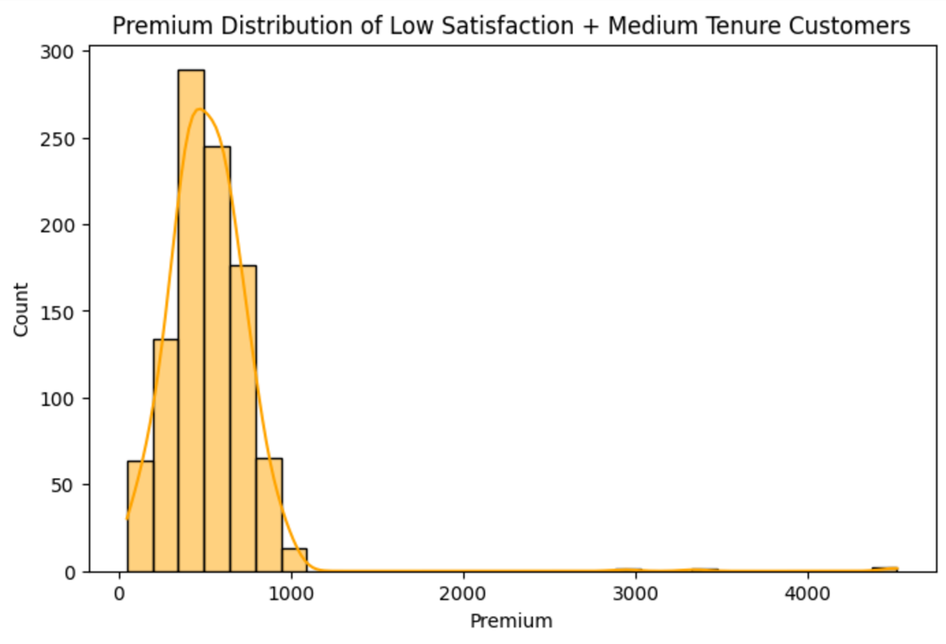
- Customers with tenure > 5 years show persistent low satisfaction levels.
- Medium premium range customers are almost evenly distributed, with Auto Insurance slightly dominant.
- Their loyalty contrasts with their satisfaction — suggesting gaps in engagement, policy relevance, or service experience.

Recommendations:

- Run targeted feedback campaigns
- Offer loyalty perks (discounts, benefits)
- Improve long-tenure customer support
- Review and align policy offerings



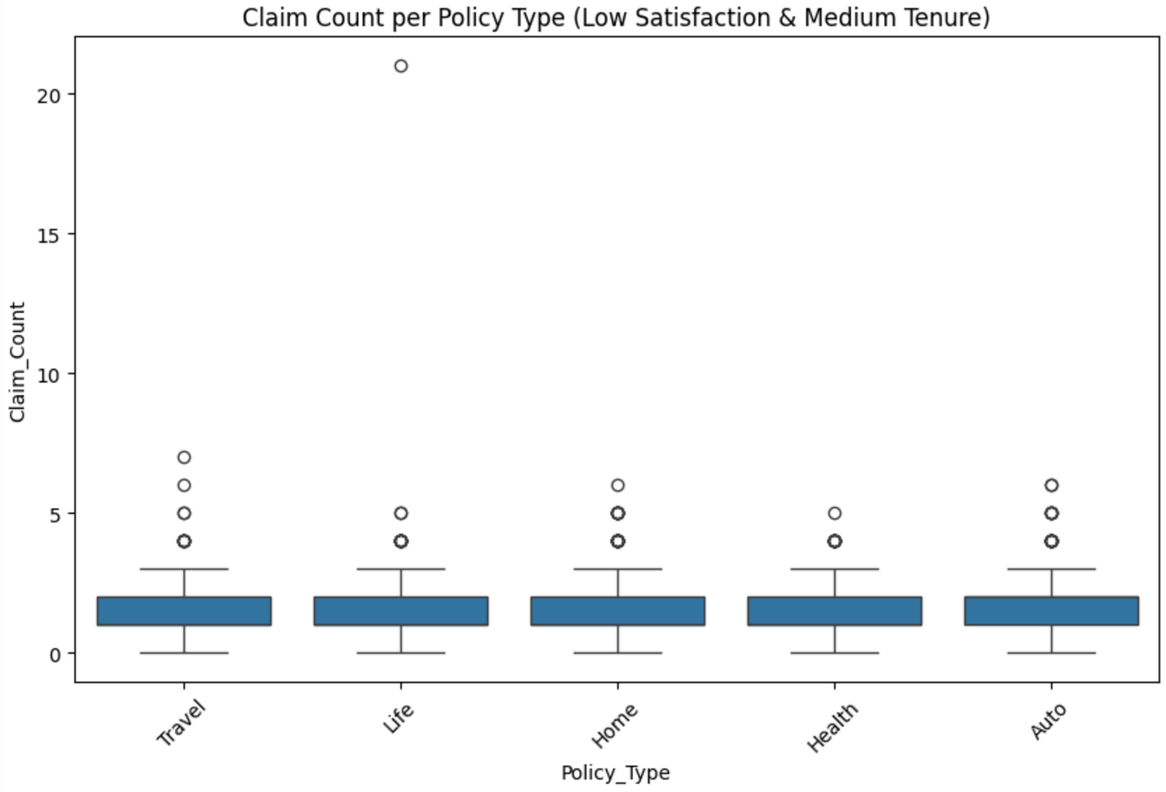
CUSTOMER SEGMENTATION BASED ON SATISFACTION & TENURE



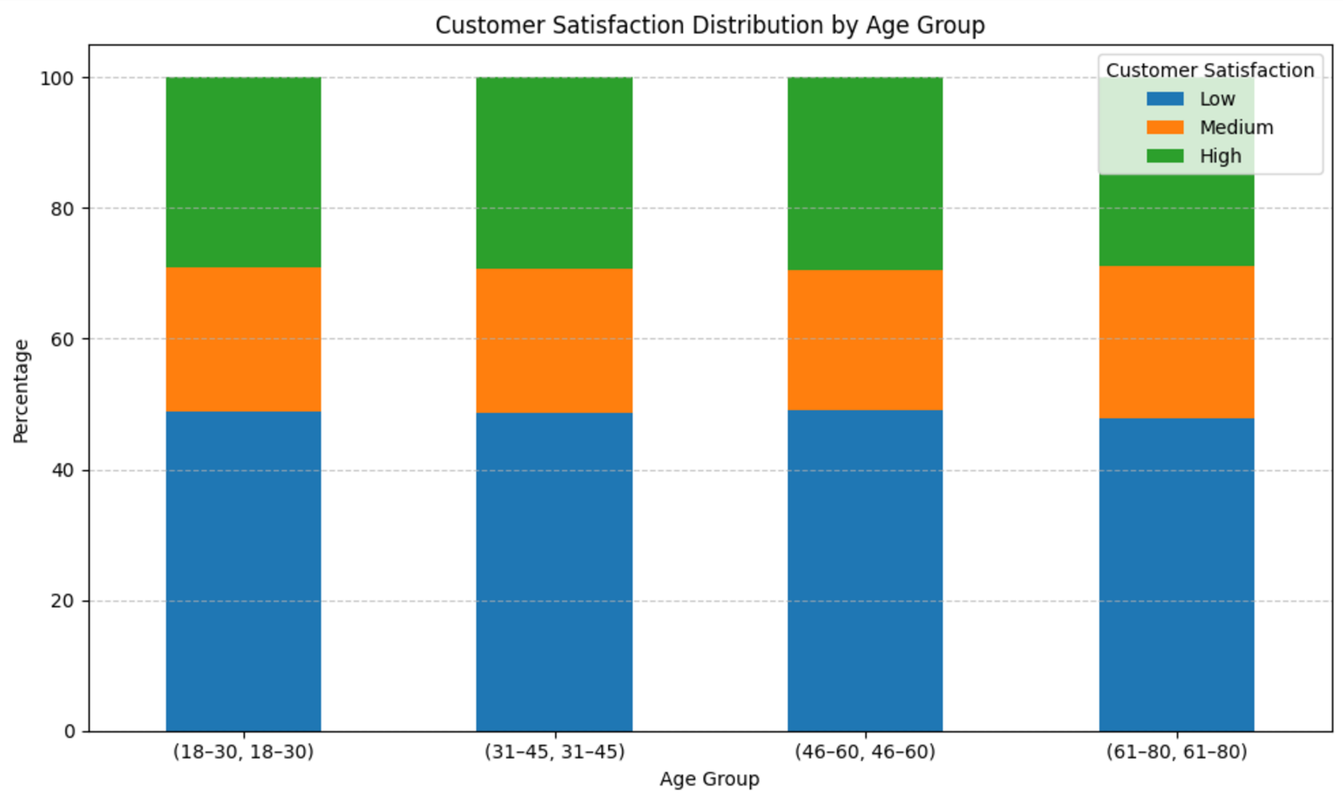
- These customers pay slightly above-average premiums, indicating long-term investment.
- Claim counts are not abnormally high, suggesting issues may lie in service experience or perceived value.

Recommendations:

- Run targeted feedback campaigns to identify service gaps.
- Offer discounts or gift cards to improve customer goodwill.
- For low claim users, review claim support and turnaround experience.

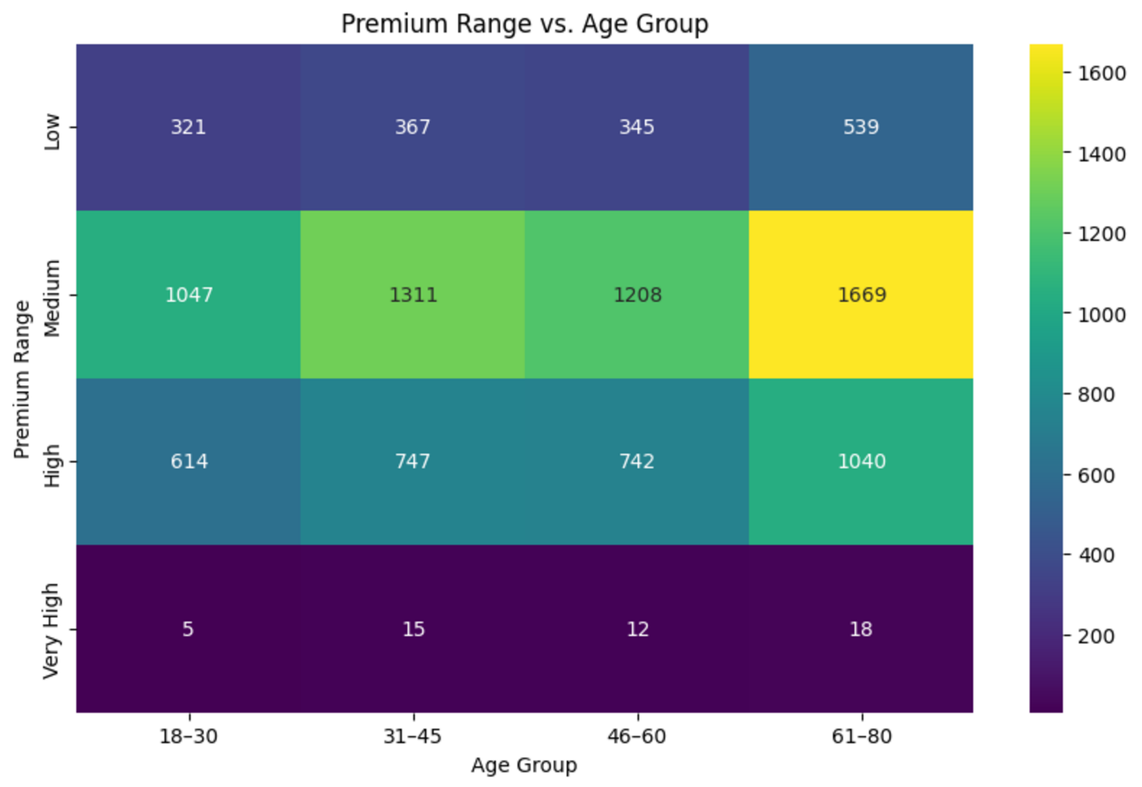


CUSTOMER SEGMENTATION BASED ON SATISFACTION & AGE GROUP



- Older age groups (61–80) tend to report slightly better satisfaction compared to the younger ones.
- Suggests generational differences in expectations or policy relevance.
- Recommendation: Tailored communication and flexible options may help improve satisfaction among younger customers.

CUSTOMER SEGMENTATION BASED ON AGE GROUP & PREMIUM RANGE

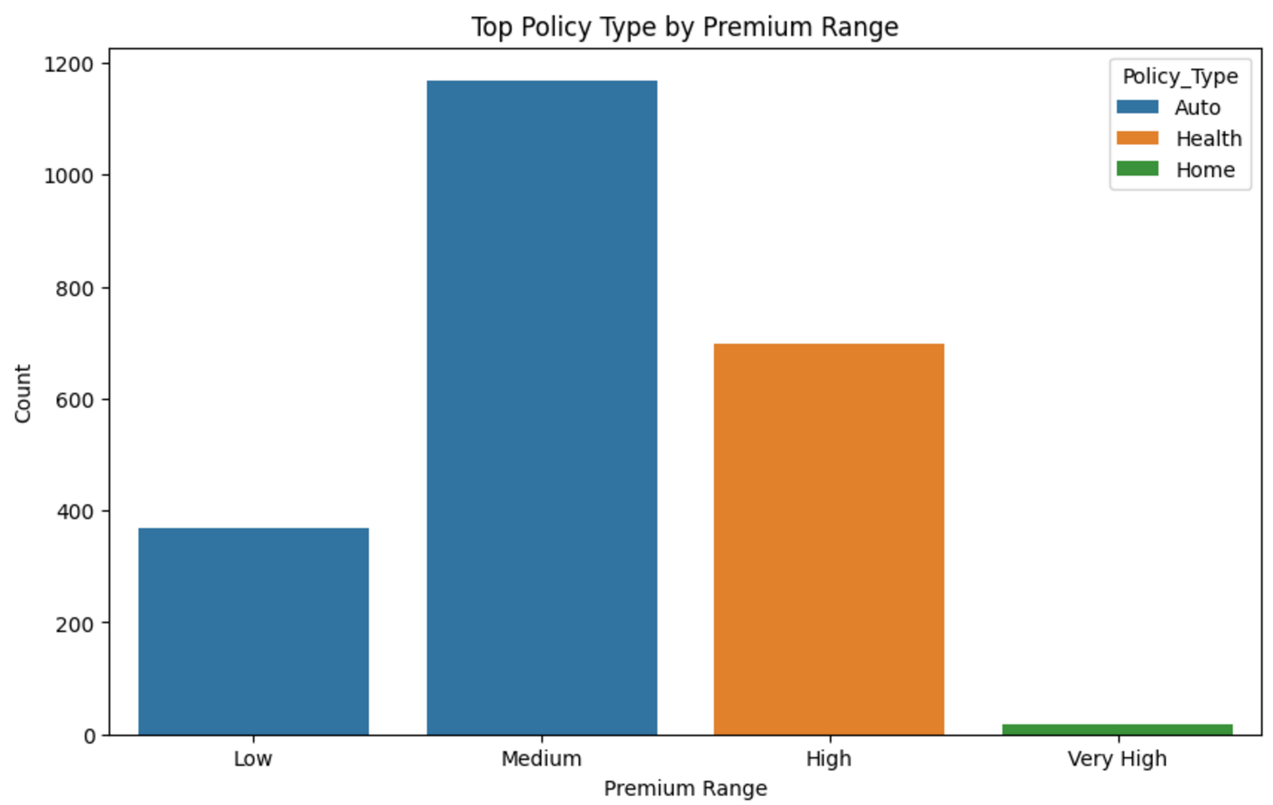


- Majority of customers across all age groups fall into the Medium Premium range.
- Premium increases with age, peaking in the 61–80 group with the highest counts in both High and Very High brackets.
- Younger customers (18–30) mostly hold Low or Medium Premium policies, suggesting cost sensitivity.

Recommendations:

- 18–30: Promote affordable plans with flexible coverage and digital support to encourage policy upgrades.
- 31–45 & 46–60: Offer bundled policies (e.g., home + life) and loyalty benefits to retain medium-high premium customers.
- 61–80: Provide priority service, health-focused add-ons, premium care perks, and timely renewal reminders.
- Focus on upselling Medium premium holders by communicating long-term value and benefits.

CUSTOMER SEGMENTATION BASED ON POLICY TYPE & PREMIUM RANGE

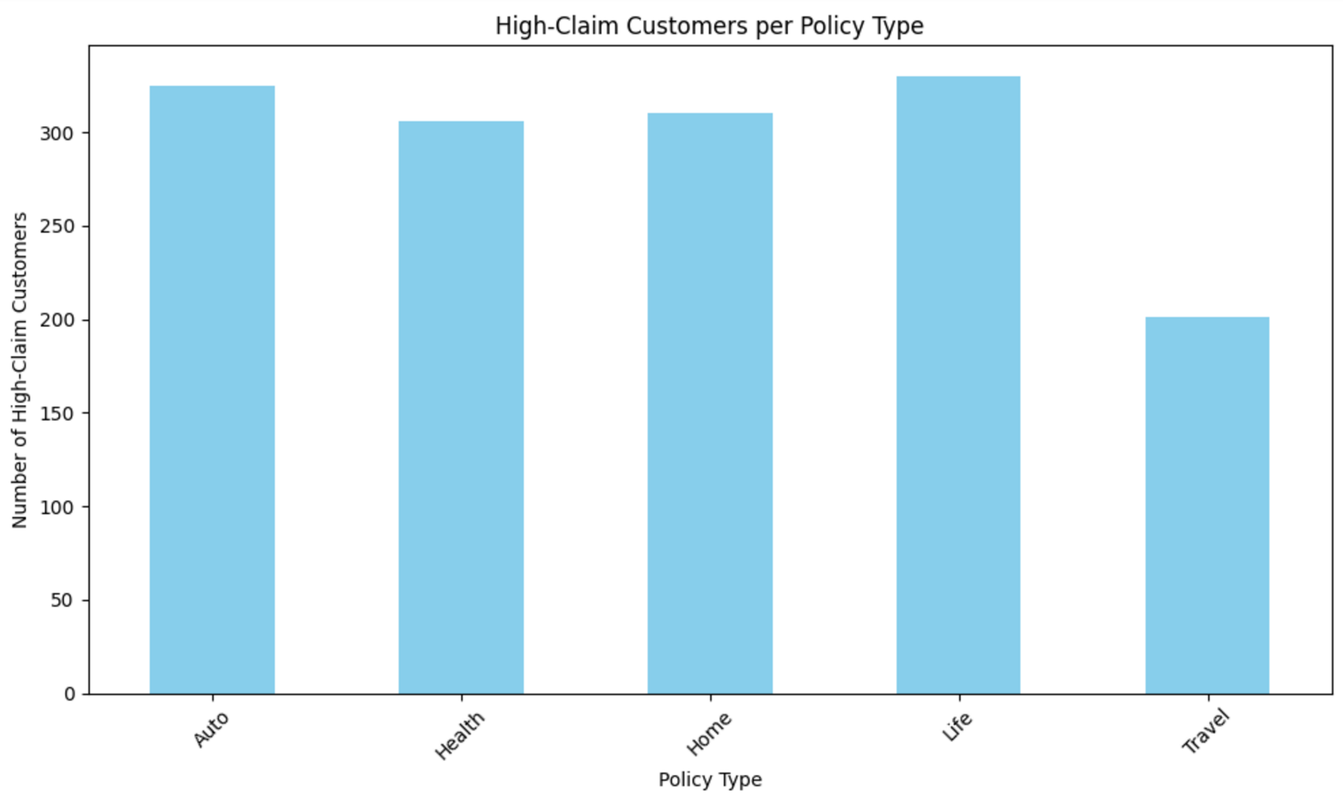


- Medium premium segment is the most populated, with Auto insurance dominating.
- High premium customers lean toward Health and Home policies.

Recommendations:

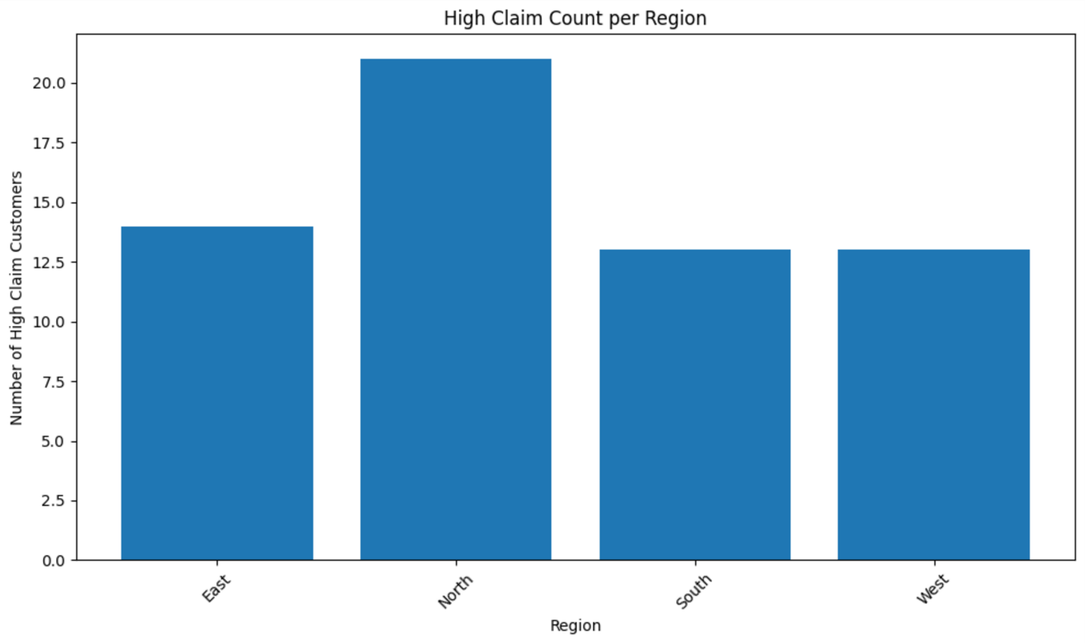
- Bundle offers or personalized upgrades for medium-tier Auto customers.
- For high-premium segments, emphasize value-added services and wellness perks.

CUSTOMER SEGMENTATION BASED ON POLICY TYPE & CLAIM COUNT



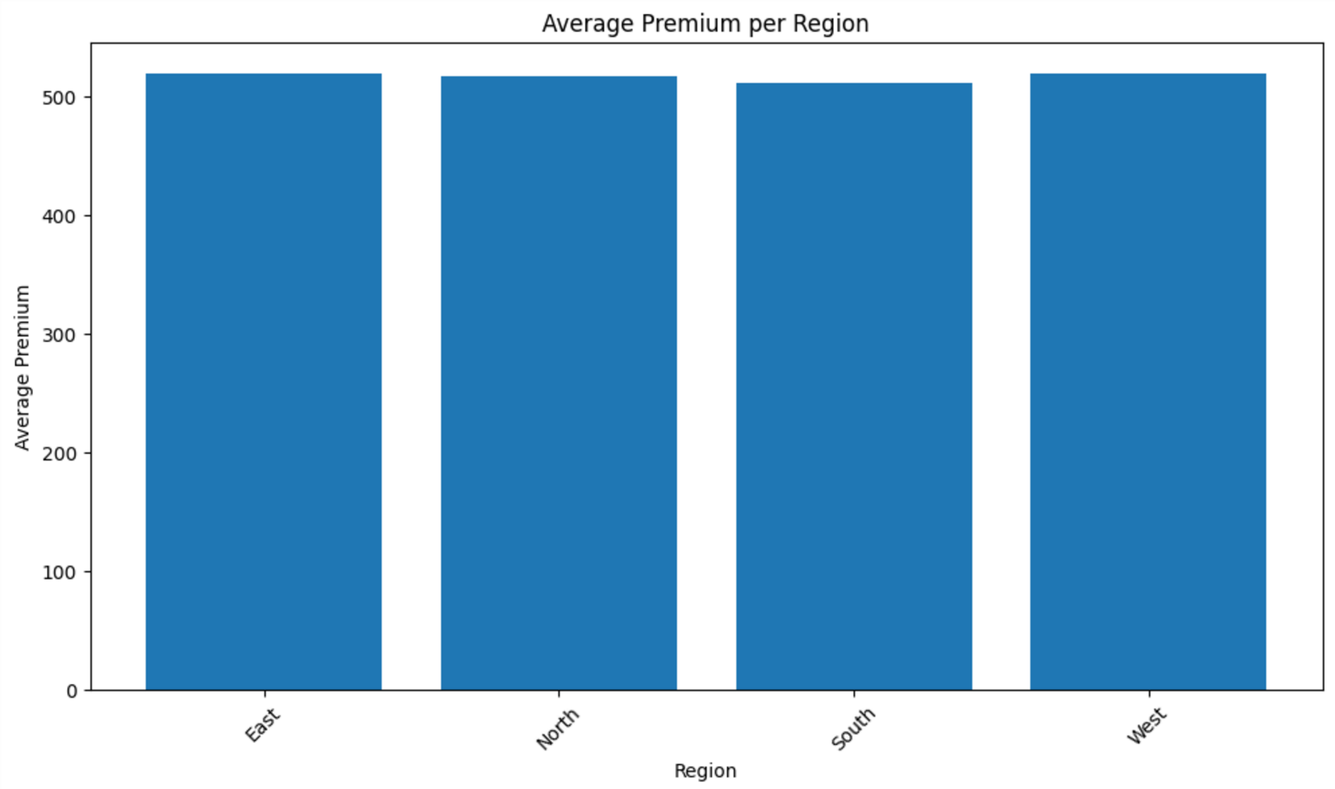
- Life and Auto policies exhibit the highest average claim counts.
- Travel policies show significantly lower claim activity.
- Reflects varying service demands and risk profiles across policy types.
- Recommendation: Prioritize efficient claim processing for Life and Auto segments, and evaluate Travel offerings to ensure competitiveness and customer engagement.

REGIONAL INSIGHTS: CLAIMS & PREMIUMS



- North and South regions have the highest average claim counts.
- Suggests higher risk exposure or more active policy usage.
- Recommendation: Monitor for potential fraud risks and ensure cost control.

- East and West have higher average premiums.
- May reflect higher-value policies or regional pricing strategies.



KEY FINDINGS

- Majority of customers are aged 31–64, with medium premiums being the most common.
- Auto and Life insurance show higher claim frequencies, especially in younger to middle-aged groups.
- Low satisfaction despite long tenure indicates issues with claim handling or service quality.
- Travel insurance sees low claim activity, possibly due to underutilization or policy gaps.
- Regions with high average premiums and high claim counts may require fraud risk monitoring.

CONCLUSION

- Cleaned and transformed a 10K-row synthetic insurance dataset.
- Extracted insights from demographics, premiums, claims, and satisfaction.
- Noted high claims in Auto/Life, low claims in Travel, and regional differences.
- Identified low satisfaction in long-tenure customers.
- Recommended feedback campaigns, perks, renewal reminders, and fraud monitoring.
- Enables data-driven strategies to improve customer retention and performance.

THANK YOU