INSURANCE ANALYSIS

Introduction

The dataset used in this project represents health insurance policyholder information along with the corresponding insurance charges incurred by the company. Each record in the dataset corresponds to an individual insurance policy and captures a combination of demographic attributes, lifestyle indicators, health-related metrics, and financial outcomes. The primary objective of the dataset is to analyse the key factors that influence insurance claim amounts and to support data-driven decision-making in insurance underwriting and pricing.

The dataset includes variables such as age, gender, body mass index (BMI), smoking status, number of dependents, and geographic region, which are widely recognized risk determinants in the insurance domain. The target variable, insurance charges (in INR), reflects the total cost incurred by the insurer for each policyholder. A unique policy number is maintained for identification and record linkage purposes.

This dataset is well-suited for exploratory data analysis (EDA) to uncover patterns and relationships between risk factors and insurance costs, as well as for predictive modelling to estimate future insurance expenditures. The structured nature of the data enables both descriptive analytics and machine learning applications, making it an effective foundation for developing business insights, pricing strategies, and risk assessment frameworks in the health insurance sector.

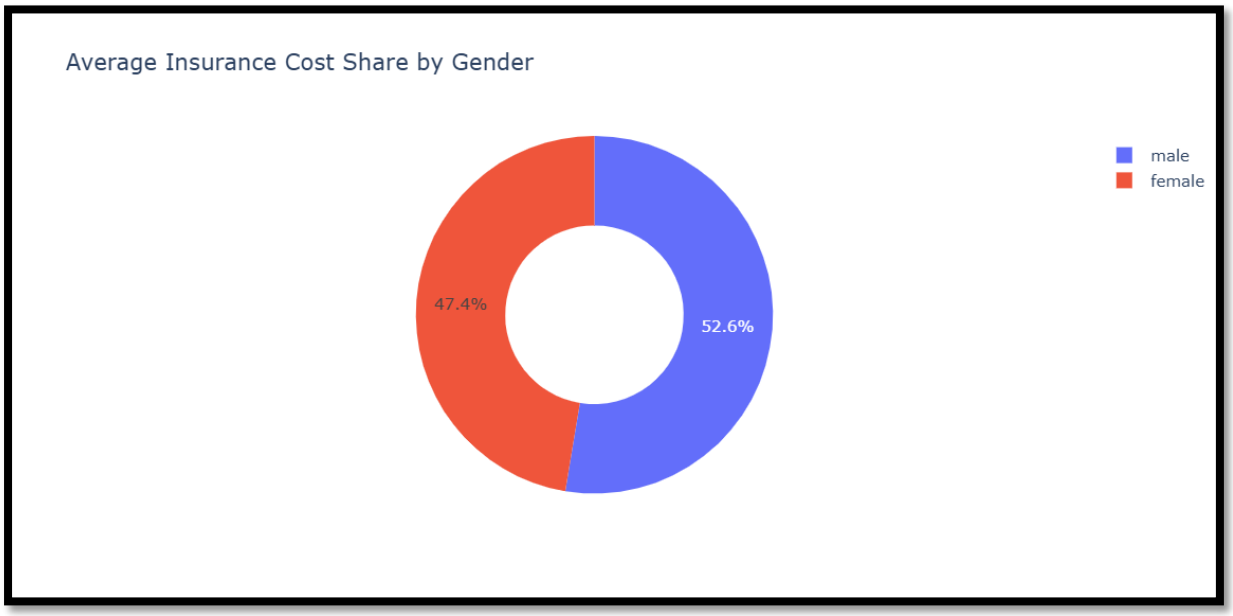
Objectives:

- Does the gender of a policyholder significantly influence the insurance cost incurred by the company?
- What is the average amount spent by the insurer per policy, and how consistent is this cost across customers?
- Is there a meaningful variation in insurance charges across different geographic regions that justifies region-specific policies?
- Does the number of dependents covered under a policy lead to higher insurance claims?
- Can BMI be used as a reliable indicator to estimate potential insurance risk and claim amounts?
- How strongly does smoking behaviour impact insurance costs compared to non-smokers?
- Does age act as a barrier or risk escalator in determining the insurance amount claimed?
- Can insurers design targeted discounts or premium adjustments based on health indicators such as BMI and smoking status?

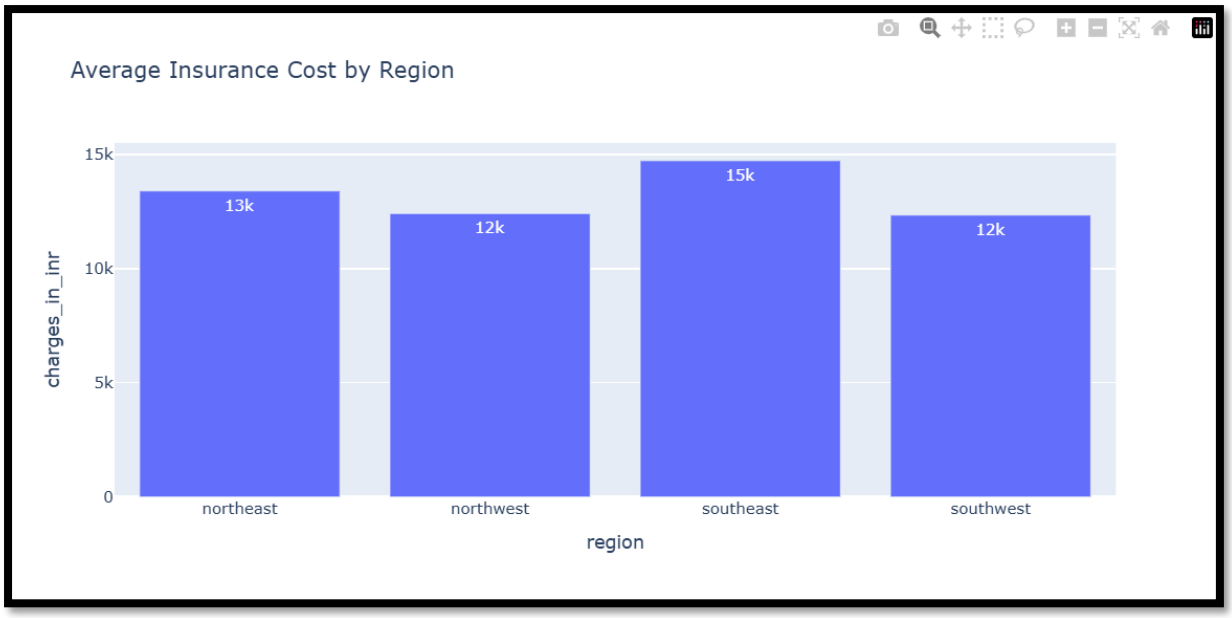
Analyses:

Methodologies:

- SQL has been used for joining the two tables
- Python has been used for EDA
- Machine Learning has been used for per year pricing predictions



The above pie chart shows that male policyholders account for a slightly higher share of average insurance costs (about 53%) compared to female policyholders (about 47%). Gender has a minor impact on insurance costs. While males contribute marginally more to overall charges, the difference is not significant enough to be a primary pricing driver on its own.



The Southeast region has the highest average insurance cost (~₹15k).

Reason: This may indicate higher health risk exposure, higher claim frequency, or higher medical costs in that region.

The Northeast region follows with an average cost of ~₹13k.

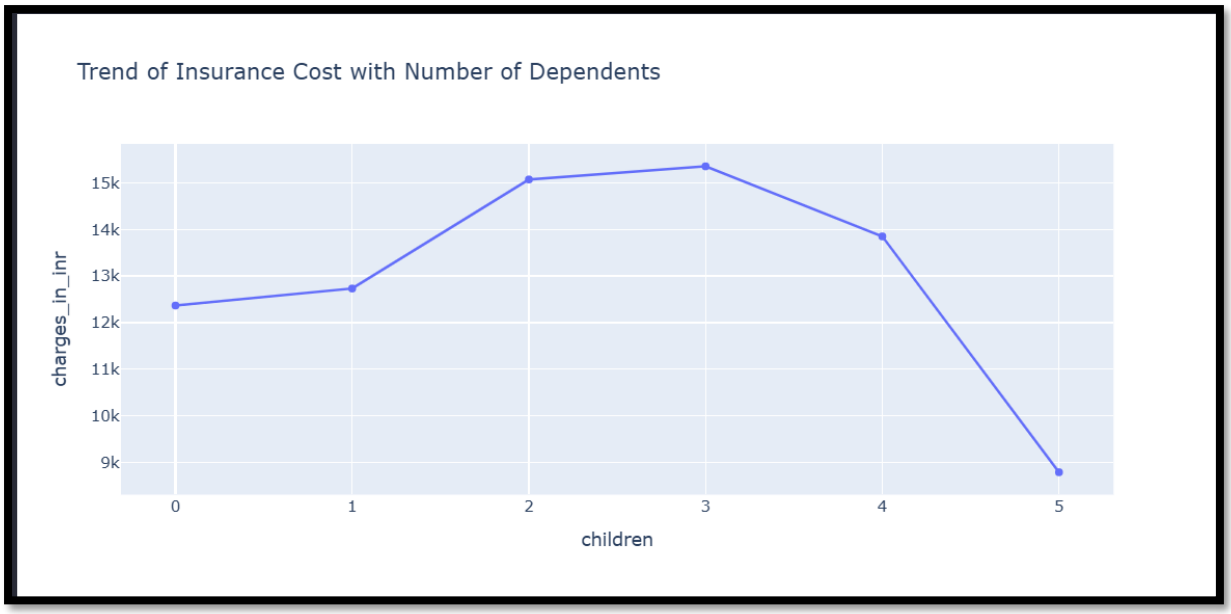
Reason: Moderate claim behaviour and healthcare cost levels.

The Northwest and Southwest regions have the lowest average costs (~₹12k each).

Reason: These regions may have healthier policyholders, fewer claims, or better cost control.

Eastern regions show higher insurance costs than western regions, the company can take the following actions to control and reduce costs:

- **Review pricing and premiums** in eastern regions to better reflect higher risk levels.
- **Promote preventive health programs** (regular check-ups, wellness plans) to reduce claim frequency.
- **Introduce stricter underwriting norms** for high-risk profiles in those regions.
- **Encourage network hospitals and negotiated rates** to control treatment costs.



1. Rising costs up to 3 dependents

2. More dependents generally mean higher healthcare utilization (more coverage needs, higher probability of claims).

3. Families with 2–3 dependents often fall into prime earning-age groups, enabling them to afford broader or higher-value insurance plans, which raises average charges.

4. Decline after 3 dependents

5. Households with 4–5 dependents are fewer in the dataset, making averages more sensitive to variation.

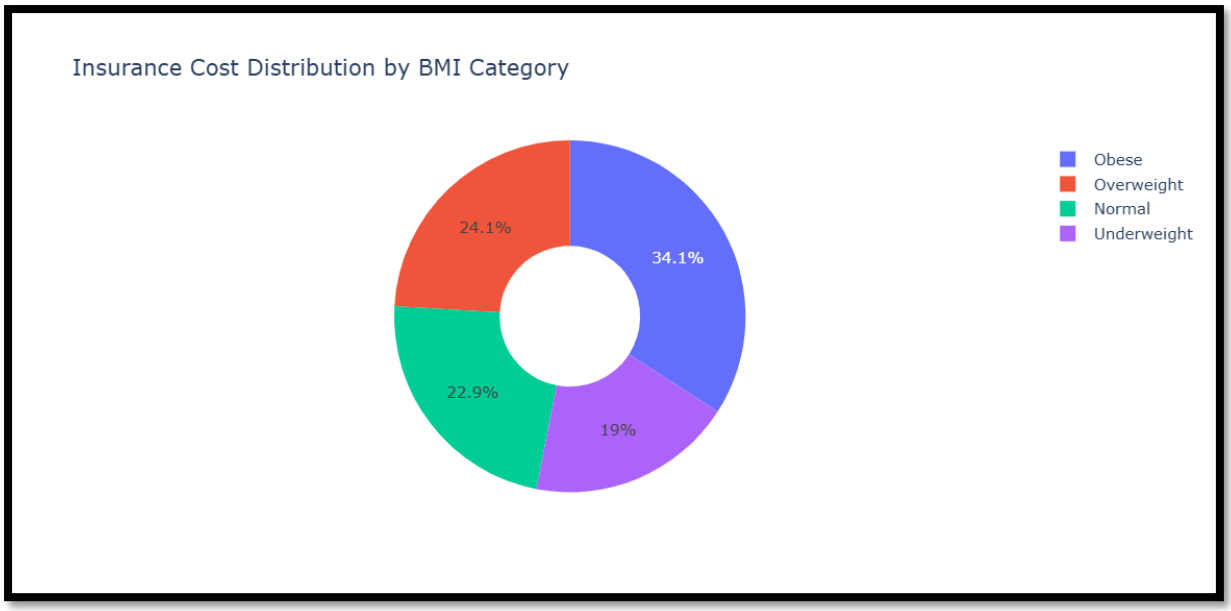
6. Such families may opt for cost-controlled or basic plans due to budget constraints, reducing average charges.

7.There may be a demographic overlap with younger parents or non-smokers, which lowers risk-based premiums.

8.Sharp drop at 5 dependents

9. Likely driven by small sample size and outlier effects.

Indicates that number of dependents alone is not a strong independent driver of insurance cost.



Obese (34.1%) → Highest insurance cost

Reason: Higher health risks lead to more medical expenses.

Overweight (24.1%) → Second highest cost

Reason: Increased chances of health issues compared to normal BMI.

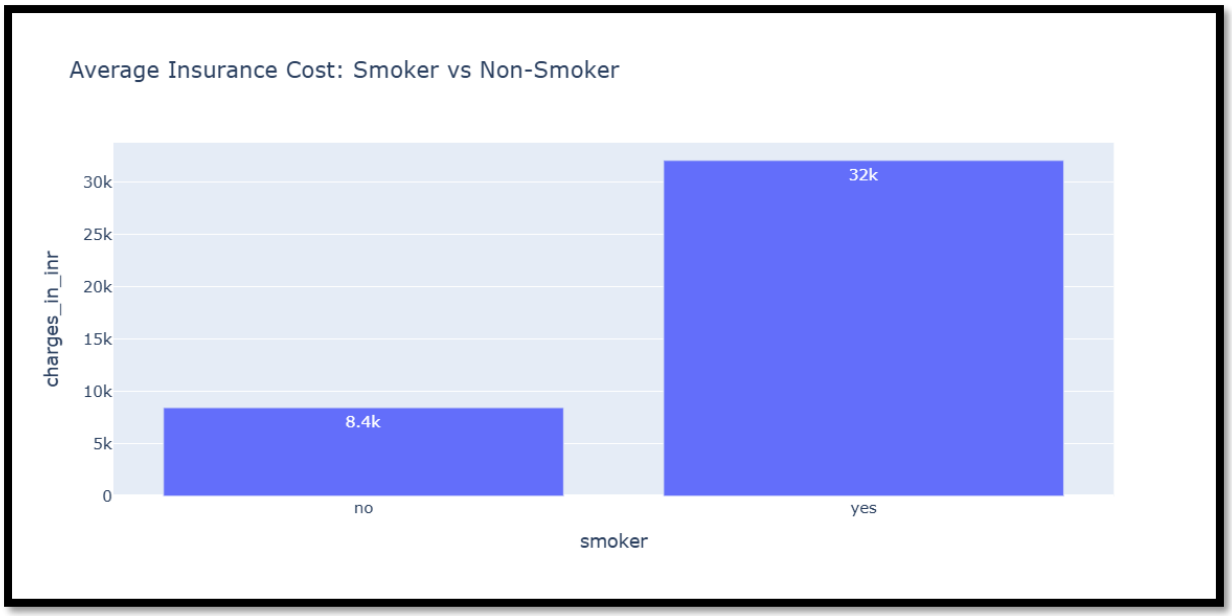
Normal (22.9%) → Moderate cost

Reason: Generally healthier, fewer medical claims.

Underweight (19%) → Lowest cost

Reason: Fewer costly health conditions in this data.

Overall Insight: Insurance costs increase as BMI increases, mainly due to higher health risks.

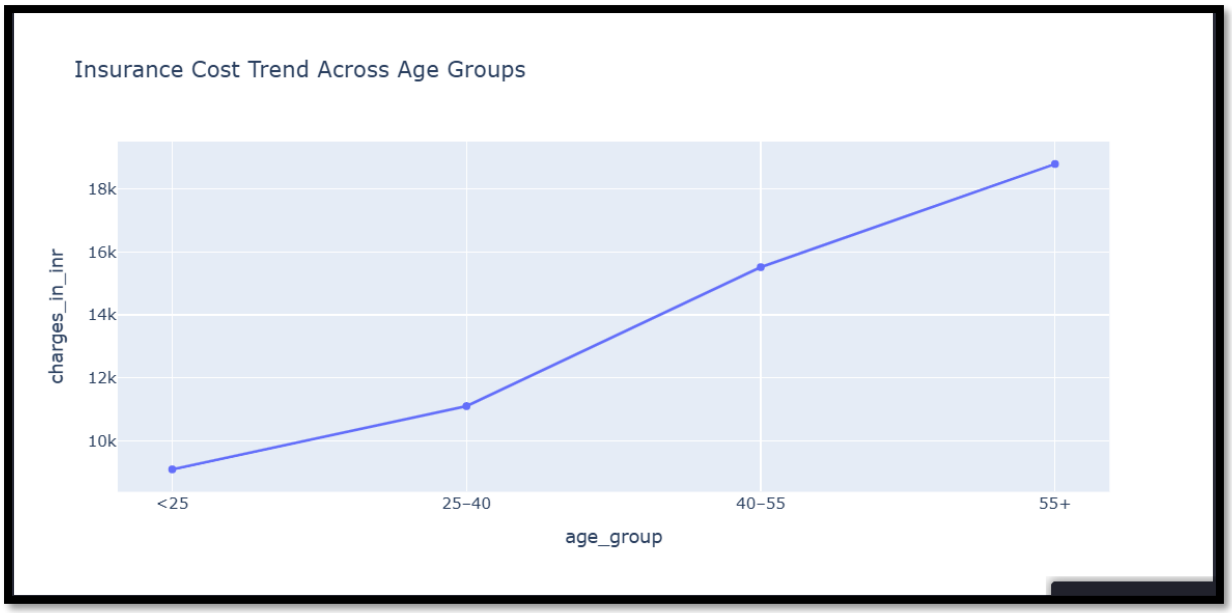


Smokers (~₹32k) pay much higher insurance costs than non-smokers.

Non-smokers (~₹8.4k) have significantly lower costs.

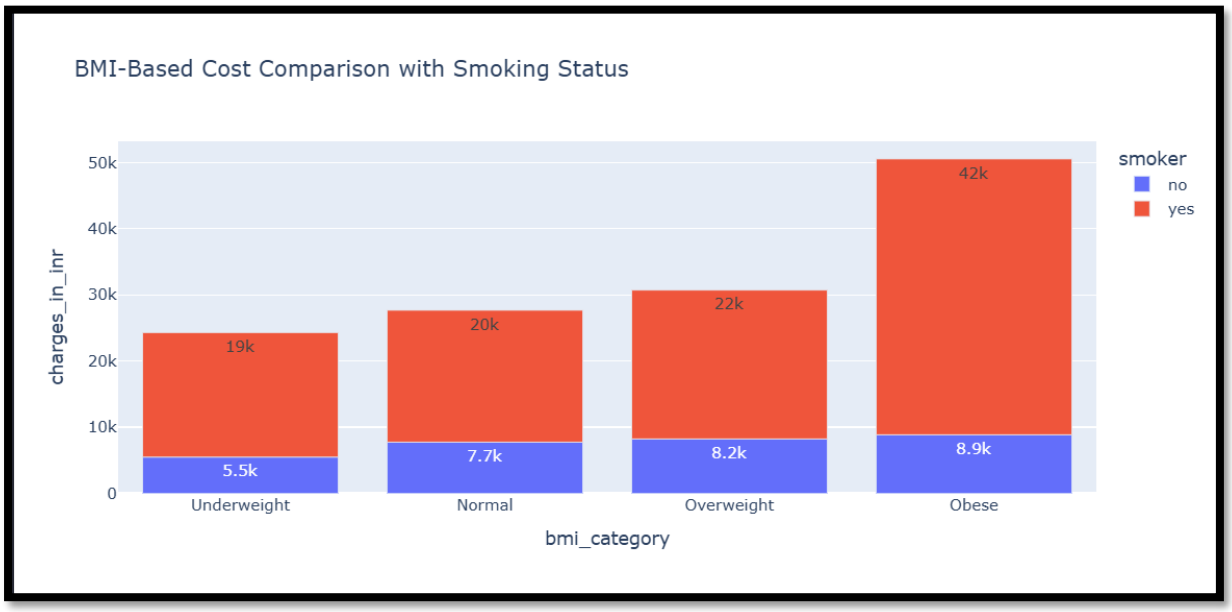
Reason: Smoking increases health risks and medical claims, so insurers charge higher premiums.

Key Insight: Smokers cost nearly 4 times more than non-smokers.



Older people have higher health risks, more medical needs, and frequent treatments, so insurance charges rise with age.

Key Insight: Age is a strong factor—insurance becomes more expensive as people get older.



Smokers have much higher insurance costs than non-smokers in every BMI group.

Insurance cost increases as BMI increases, especially for smokers.

By BMI category:

Underweight: Low cost for non-smokers, much higher for smokers

Normal & Overweight: Costs rise steadily, smokers always pay more

Obese: Highest cost, especially for smokers

Reason:

Smoking and high BMI both increase health risks.

When smoking + obesity combine, medical expenses rise sharply.

Key Insight: Smoking multiplies the cost impact of high BMI on insurance charges.

Streamlit dashboard - https://app-kdvw9gqijfnmp5d5zjdvsp.streamlit.app/