Capstone Project Report

MediBuddy is a digital healthcare platform offering inpatient hospitalisation, outpatient services, and corporate wellness benefits. It is an award-winning platform by Medi Assist that transforms the health insurance ecosystem by enabling seamless discovery, access, and monitoring of healthcare benefits. The company was founded in 2000 and is headquartered in Bangalore, Karnataka, India.

This project analyses two datasets provided by MediBuddy:

- Dataset 1: Personal information (policy number, children, smoker status, region)
- Dataset 2: Health and cost details (age, sex, BMI, charges)

Dataset Overview

After merging both datasets using **Policy no.**, the final dataset contained:

- Total records: **1338**
- Columns:
    - Policy no.
    - children
    - smoker
    - region
    - age
    - sex
    - bmi
    - charges

There were **no missing values**, and the dataset was clean and suitable for analysis.

Key summary statistics:

- Average age: **39 years**
- Average BMI: **30.66**
- Average charges: **Rs. 13,270**
- Maximum charges: **Rs. 63,770**
- Minimum charges: **Rs. 1,121**
- Female: Rs. 12,569
- Male: Rs. 13,956

The difference between male and female policyholders is relatively small (~Rs. 1,387).
Conclusion: Gender does not significantly impact insurance cost and should **not be used as a constraint** for extending policies.

Q1. Does the gender of the person matter for the company as a constraint for extending policies?

Average claim amount:

- Female: **Rs. 12,569**
- Male: **Rs. 13,956**

The difference is small **(~Rs. 1,387)**.
Conclusion: Gender does not significantly impact insurance cost and therefore should not be used as a constraint for extending policies.

Q2. What is the average amount of money the company spent on each policy cover?

The average amount spent per policyholder is: **Rs. 13,270**

Q3. Could you advise if the company needs to offer separate policies based on the geographic location of the person?

Average charges by region:

- Northeast: **Rs. 13,406**
- Northwest: **Rs. 12,417**
- Southeast: **Rs. 14,735**
- Southwest: **Rs. 12,346**

There are only minor differences across regions.
Conclusion: Geographic location does **not strongly affect** claim amount. Separate policies by region are **not required**.

Q4. Does the number of dependents make a difference in the amount claimed?

Average charges by number of children:

- 0 children → Rs. 12,365
- 1 child → Rs. 12,731
- 2 children → Rs. 15,073
- 3 children → Rs. 15,355
- 4 children → Rs. 13,850
- 5 children → Rs. 8,786

There is no consistent pattern.
Conclusion: The number of dependents has **only a weak impact** on claims and should not be a major pricing factor.

Q5. Does a study of a person's BMI give the company any idea of the insurance claim that it would extend?

Correlation between BMI and charges:
0.198 (positive relationship)

This means higher BMI tends to increase claim amount.
Conclusion: BMI provides **useful health risk insight** and can support underwriting decisions.

Q6. Is it needed for the company to understand whether the person covered is a smoker or a non-smoker?

Average charges:

- Non-smoker: Rs. 8,434
- Smoker: Rs. 32,050

Smokers claim almost **4 times more**.
Conclusion: Smoking status is the **most important factor** in predicting claims and must be considered for policy pricing.

Q7. Does age have any barrier to the insurance claim?

Correlation between age and charges:
0.299 (moderate positive relationship)

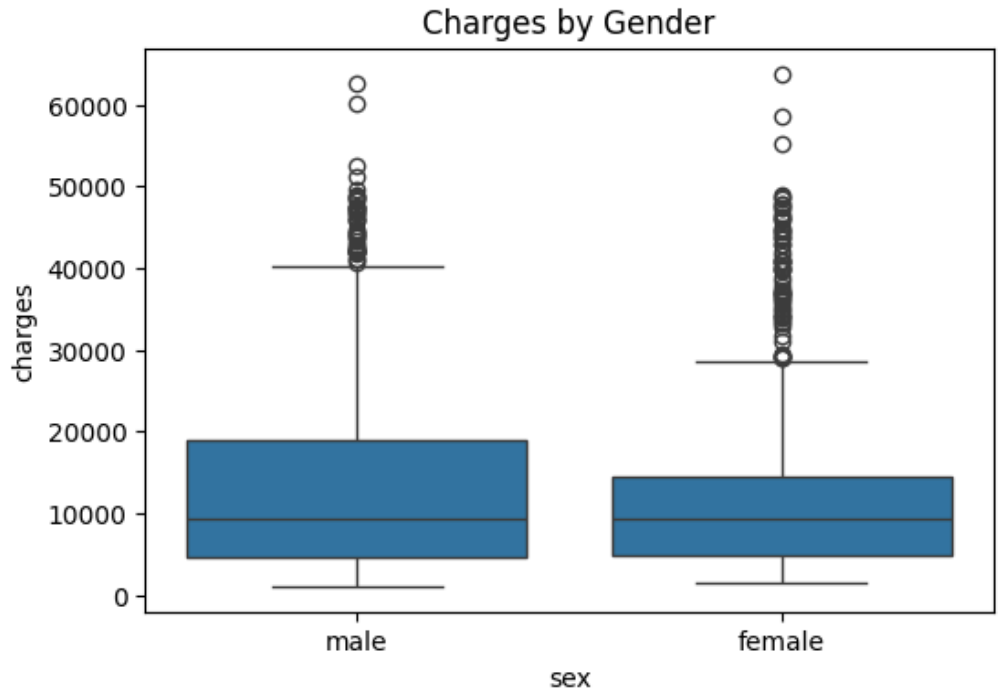
Older individuals tend to claim more.
Conclusion: Age has a **noticeable impact** on claims and should be taken into consideration when determining premiums.

Q8. Can the company extend certain discounts after checking the health status (BMI) In this case?

Since people with lower BMI generally claim less, rewarding healthy individuals is logical.
Conclusion: Yes, the company can introduce **wellness-based discounts** for customers with a healthy BMI to promote preventive healthcare.

Followings are a few Insights:

1. Charges vs Gender
Graph used: Boxplot of sex vs charges

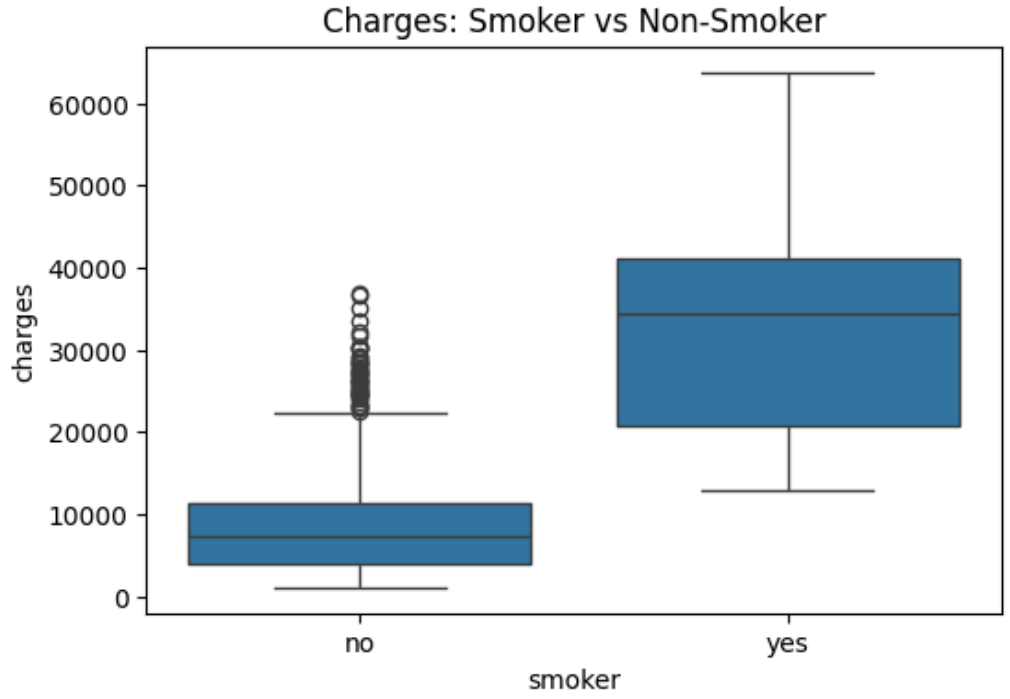


The boxplot shows that the distribution of insurance charges for males and females is quite similar, with only a small difference in average values. This indicates that gender does not strongly influence claim amounts. Therefore, the company should avoid using gender as a restriction for extending insurance policies.

---

2. Smoker vs Non-Smoker

Graph used: Boxplot of smoker vs no-smoker

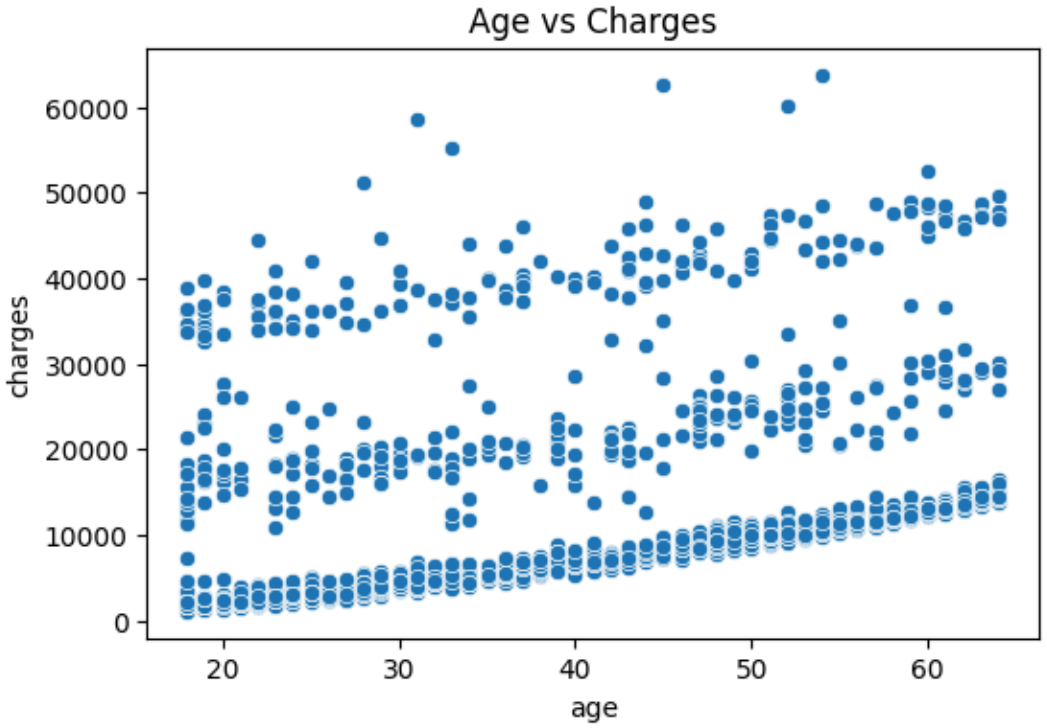


The visualisation clearly shows that smokers have significantly higher insurance charges compared to non-smokers. The median and spread of charges for smokers are much higher. This implies that smoking status is a major risk factor and should be strongly considered when pricing insurance premiums.

---

3. Age vs Charges

Graph used: Scatter plot of age vs charges

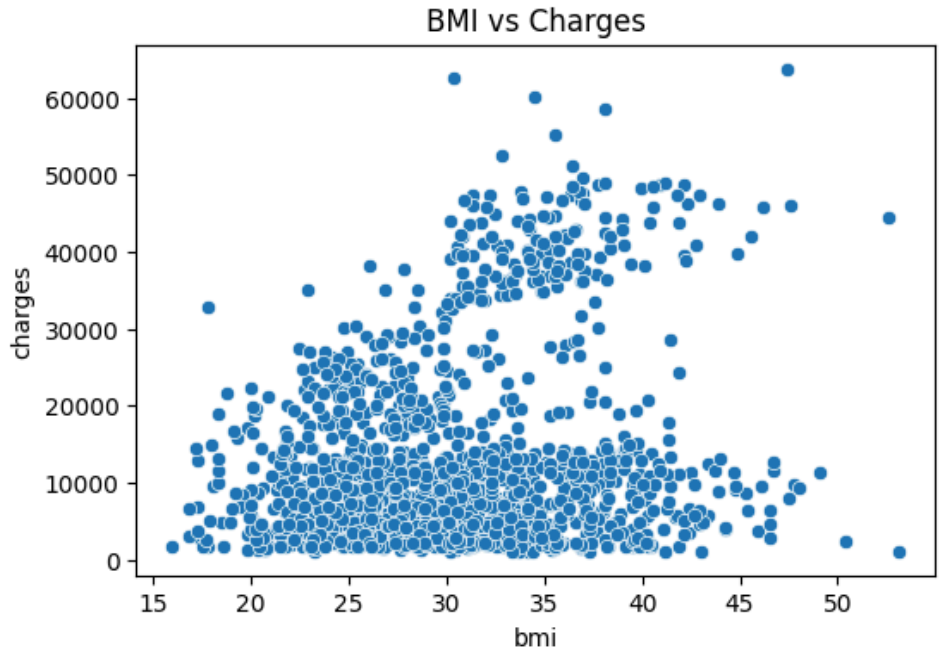


The scatter plot shows an upward trend, where insurance charges tend to increase as age increases. This suggests that older individuals generally incur higher healthcare costs. The company should consider age-based risk while designing premium slabs.

---

4. BMI vs Charges

Graph used: Scatter plot of BMI vs charges

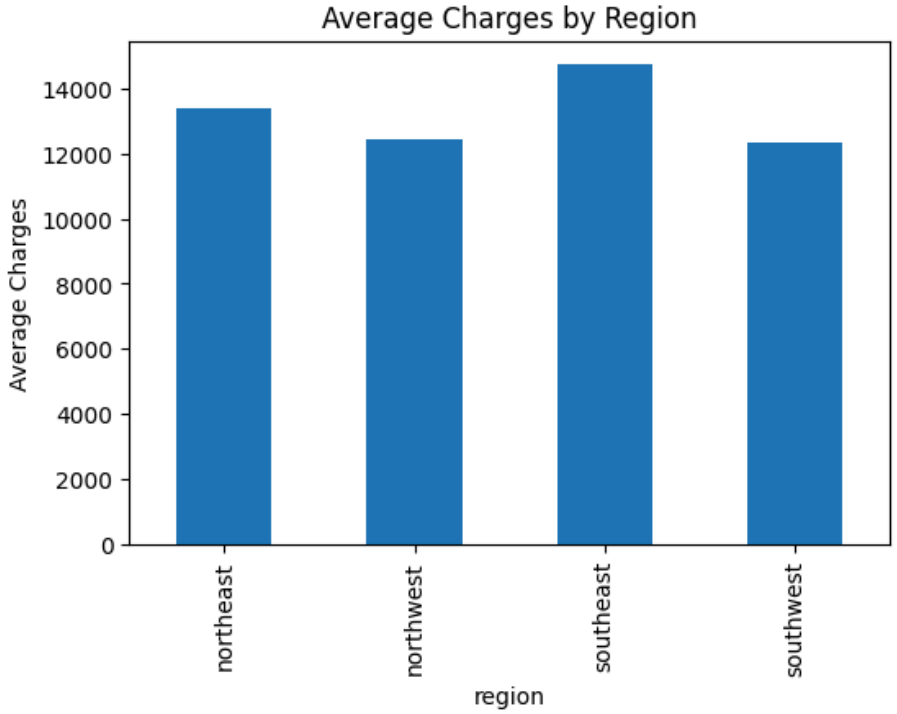


The chart shows that individuals with higher BMI values tend to have higher insurance charges. Although the relationship is moderate, it still indicates that health condition plays a role in claim amounts. This suggests BMI can be used as a supporting health indicator during policy assessment.

---

5. Region vs Charges

Graph used: Bar chart of average charges by region

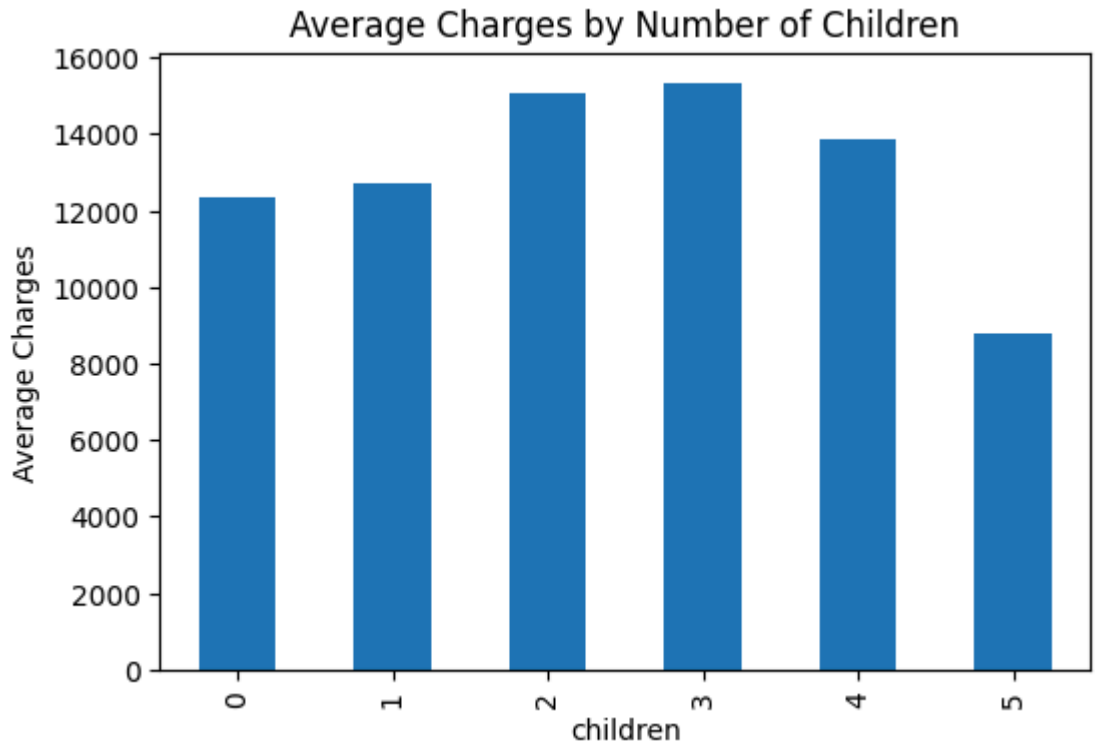


The bar chart shows that the average insurance charges across regions are relatively close, with only slight variation. This indicates that geographic location does not have a strong influence on claims. Therefore, offering separate policies based on region is not necessary.

---

6. Children vs Charges

Graph used: Bar chart of children vs average charges



The chart shows that while claims slightly increase for customers with 2–3 children, the pattern is inconsistent overall. This suggests that the number of dependents does not strongly impact insurance cost and should not be a major factor in pricing decisions.

Q9. Which factor has the strongest impact on insurance charges?

Among all variables, **smoking status** has the strongest impact on insurance charges. Smokers have an average claim of **Rs. 32,050**, while non-smokers have only **Rs. 8,434**.

This makes smoking the most influential risk factor in the dataset.

Business Insight: The company should prioritise lifestyle-based risk assessment (especially smoking) while designing premiums.

Q10. Among age, BMI, and children, which numerical factor impacts claims the most?

Correlation values:

- Age vs Charges → **0.299**
- BMI vs Charges → **0.198**
- Children vs Charges → Very weak trend

This shows that **age has a stronger influence than BMI**, while the number of children has minimal impact.

Business Insight: Age-based pricing slabs are more meaningful than dependent-based pricing.

Q11. Do younger people always have low insurance claims?

Not always. While younger individuals generally have lower claims, some young individuals (especially smokers or those with high BMI) also show high charges.

Business Insight: Risk profiling should be multi-dimensional. Age alone should not be used to judge a customer's risk.

Q12. Can two people of the same age have very different insurance charges?

Yes. The visualisations show that people of the same age group can have very different charges depending on smoking habits, BMI, and lifestyle.

Business Insight: Personalised pricing based on multiple health indicators is more effective than flat age-based pricing.

Q13. Is it possible to identify low-risk customers from the dataset?

Yes. Customers who are:

- Non-smokers
- Have a BMI below ~25–27
- Are below middle age
  tend to have significantly lower claim amounts.

Business Insight: The company can identify low-risk users and offer them loyalty benefits or discounted premiums.

Q14. What type of customers are the highest risk?

The highest risk customers are typically:

- Smokers
- Older age (50+)
- High BMI (30+)

These customers frequently appear in the high-claim region of the dataset.

Business Insight: These customers should be carefully priced to avoid financial losses.

Q15. Can this data help MediBuddy in a preventive healthcare strategy?

Yes. The analysis shows that unhealthy lifestyle indicators (smoking, high BMI) lead to higher claims.

Business Insight: MediBuddy can invest in preventive programs like fitness rewards, smoking cessation programs, and wellness tracking to reduce long-term claim burden.

Q16. Is this dataset suitable for real-world deployment?

The dataset is structured and realistic but limited in features (e.g., no medical history, income, or hospitalisation history).

Business Insight: With additional real-world features, this model can be further improved and used in real production systems.

Q17. What improvements can be made to the model in future?

Future improvements could include:

- Adding medical history features
- Including exercise/activity data
- Adding past claim history
- Using advanced models like XGBoost
- Collecting longitudinal (time-series) data

Business Insight: More data = better predictions = smarter pricing decisions.
