Planning

Section 1: Dataset Description

Insurance Dataset for Predicting Health Insurance Premiums in the US

Dataset Description:

The dataset contains 1,000,000 entries with 12 attributes describing health insurance policyholders. It provides information such as age, gender, BMI, number of dependents, smoking status, geographic region, and associated medical expenses. This dataset is ideal for analyzing and visualizing relationships between demographic and lifestyle factors and healthcare costs.

Rows: 1,000,000

Columns: 12 (10 features + 1 target charges + 1 insurance_plan)

Column Analysis: Data Types & Domains:

Column Name	Data Type	Domain (Range or List of Values)
age	Ratio (Continuous)	18 - 65 years
gender	Categorical (Nominal)	male, female (Balanced ~50/50 split)
bmi	Ratio (Continuous)	18.00 - 50.00 (Uniformly distributed)
children	Ratio (Discrete)	0, 1, 2, 3, 4, 5
smoker	Categorical (Binary)	yes, no
region	Categorical (Nominal)	northeast, northwest, southeast, southwest

medical_history	Categorical (Nominal)	Diabetes, High blood pressure, Heart disease, None
family_medical_history	Categorical (Nominal)	Diabetes, High blood pressure, Heart disease, None
exercise_frequency	Categorical (Ordinal)	Never, Rarely, Occasionally, Frequently
occupation	Categorical (Nominal)	Blue collar, White collar, Student, Unemployed
insurance_plan	Categorical (Ordinal)	Basic, Premium, Standard
charges	Ratio (Continuous)	Target Variable (Currency amount, e.g., $2,000 - $60,000+)

Usage & Applications:

Data scientists use this specific dataset for:

1. Regression Modeling: Predicting the continuous variable charges based on risk factors (Smoking, BMI, Age).

2. Big Data Practice: Because it has 1 million rows, it is used to test algorithms that might be too slow on standard laptop CPUs, encouraging the use of Spark or GPU acceleration.

3. Risk Analysis: Analyzing how specific combinations (e.g., *Smoker* + *High BMI* + *Family History*) compound to drastically increase insurance costs.

Visualization Ideas:

1. Box Plot (Charges vs. Smoker Status): Compare the distribution of insurance premiums between smokers and non-smokers side-by-side. This clearly highlights the drastic difference in median costs and reveals extreme outliers, confirming smoking as a dominant risk factor.

2. Scatter Plot (BMI vs. Charges): Plot Body Mass Index against Charges to identify trends, particularly distinguishing how costs rise with obesity. Color-coding these points by "Smoker" status often reveals distinct clusters, showing how high BMI and smoking combine to exponentially increase premiums.

3. Correlation Heatmap: Generate a heatmap to display correlation coefficients between numerical variables like Age, BMI, Children, and Charges. This provides an instant visual summary of which features have the strongest linear relationships with the target variable, helping prioritize features for modeling.

Section 2: Prospective Dashboard Users

1. **Insurance Team:**

   - **Data-driven premium pricing:** Helps to refine premium models using key risk indicators (BMI, medical history, smoking status, etc.) to better reflect individual and group-level risk.

   - **Regional pricing insights:** Encourages to model more accurate region-specific pricing strategies by identifying geographic differences in healthcare costs and risk factors.

   - **Actuarial forecasting:** Studies historical charge data to predict future claims and determine appropriate reserve levels for financial stability.

   - **Risk segmentation & Coverage Plans:** Helps to analyze charge distribution across various plans to align pricing with risk patterns that could help to tailor specialized coverage plans or premium structures as per by high-risk groups.

   - **Strategic portfolio management:** Provides status of policyholder's profitability and helps to indicate risk exposure.

2. **Healthcare Providers or Hospitals:**

   - **Demand forecasting & capacity planning**: Helps to identify rising service demand across demographic groups to guide staffing, equipment purchases, and facility investments.

   - **Reimbursement negotiation:** Provides cost and demographic insights of the current medical scenarios to monitor current investments by patients in order to strengthen negotiations and cost-reflective reimbursement rates.

   - **Quality improvement**: Helps to detect cost outliers and variations in care, by prompting the refinement of clinical protocols to improve outcomes and reduce waste.

   - **Network and facility planning:** Helps to guide optimal placement of facilities and services based on regional health trends and local population needs.

3. **Government Health Agencies:**

   - **Early threat detection:** Identifies emerging health risks and shifts in disease burden for timely interventions or pandemics, outbreaks, or natural disasters in order to maintain healthcare system stability.

- **Targeted subsidies & investments:** To ensure financial support, clinic placement, staffing, and equipment investments are directed to areas of greatest need.

- **Cost-benefit & forecasting:** Provides data to plan sustainable programs by the impact of changing risk factors (e.g., smoking, obesity) for the future healthcare costs.

- **Socioeconomic integration:** Helps to link health outcomes and costs with social factors like income and occupation to guide broader policy decisions.

Section 3: List of User Requirements & Potential Questions.

User Requirements

The dashboard aims to serve multiple stakeholders insurance underwriters, healthcare providers, and government health agencies by providing actionable insights into the relationships between health, lifestyle, and cost factors. The following requirements outline what each group expects from the dashboard:

1. Insurance Underwriting Team

- **Risk Factor Evaluation:** Analyze how variables such as age, BMI, smoking habits, and medical history influence insurance premiums.
- **Premium Optimization:** Contribute to the development of data-driven pricing frameworks that reflect demographic and lifestyle differences.
- **High-Risk Profiling:** Identify customer segments with elevated financial risk to design customized coverage plans.
- **Regional Benchmarking:** Assess geographical disparities in medical costs to guide region-specific underwriting decisions
- **Visualize Forecasting:** Create interactive dashboards illustrating charge distributions, attribute correlations, and segment-based cost forecasts.

2. Healthcare Providers or Hospitals

- **Population Health Analysis:** Monitor relationships between lifestyle factors such as smoking, physical activity, and occupation and medical conditions.
- **Preventive Care Strategy:** Helps to identify lifestyle patterns associated with chronic diseases to design effective wellness and prevention programs.
- **Resource Planning:** Leverage demographic and cost data to optimize staffing, facility management, and allocation of medical resources.
- **Cost and Performance Benchmarking:** Compare treatment outcomes and claim patterns against regional or national standards to assess efficiency.

3. Government Health Agencies

- **Public Health Trend Analysis:** Track key indicators such as obesity rates, smoking prevalence, and their economic impact on the healthcare system.
- **Policy Effectiveness Evaluation:** Analyze data to measure the outcomes of existing health policies and subsidy initiatives.

- **Regional Inequality Assessment:** Detects regions or populations with elevated medical expenditures to inform equitable resource distribution.
- **Public Awareness Insights:** Examine how lifestyle behaviors influence healthcare costs to shape targeted awareness and education campaigns.
- **Strategic Planning Support:** Provide analytical foundations for long-term decisions related to healthcare funding, insurance frameworks, and national policy design.

Potential Questions

Q1. Is there a correlation between BMI and insurance costs?

Why: To explore whether body mass index has a measurable influence on healthcare spending and insurance charges. Understanding this relationship can highlight the financial burden associated with obesity-related health issues.

Usefulness: Provides valuable insights for underwriters to visualize how BMI impacts overall risk, enabling more equitable premium adjustments and promoting health-conscious policy incentives.

Q2. How do average charges vary across different regions?

Why: Regional variations in medical costs may stem from differences in healthcare access, living standards, or lifestyle habits. Examining these differences helps reveal broader patterns in healthcare affordability.

Usefulness: Equips insurers, policymakers, and healthcare planners with data to identify high-cost regions and design region-specific pricing, resource distribution, or wellness initiatives.

Q3. How does age influence the progression of insurance charges?

Why: Medical expenses tend to rise with age due to an increased likelihood of chronic conditions and healthcare needs. Investigating this trend provides a clearer view of age-related cost progression.

Usefulness: Supports visualizations that demonstrate how healthcare costs evolve across age groups, informing premium structuring and future resource allocation strategies.

Q4. How does smoking status impact insurance charges?

Why: Smoking is a major health risk factor that directly contributes to higher treatment and insurance costs. Analyzing this variable helps quantify the economic impact of tobacco use.

Usefulness: Enables insurers to visualize and compare cost patterns between smokers and non-smokers, improving policy differentiation and health risk assessments.

Q5. Is there any significant difference between insurance charges for males and females?

Why: Gender can influence medical spending due to biological, behavioral, and lifestyle factors. Identifying these differences helps ensure fairness in insurance evaluations.

Usefulness: Assists in assessing gender-based pricing equity and detecting potential biases in premium calculations or predictive models.

Q6. What does the demographic of the highest risk group look like?

Why: Understanding the profile of high-risk individuals such as their age, BMI, and habits can reveal the main contributors to elevated healthcare costs.

Usefulness: Helps insurers and public health organizations design targeted interventions, customized coverage options, and preventive wellness programs.

Q7. What is the percentage breakdown of individuals with a specific medical condition?

Why: Knowing how common certain diseases are across the population helps in prioritizing public health and insurance initiatives.

Usefulness: Provides actionable insights for healthcare providers and government agencies to develop prevention campaigns, awareness programs, and better treatment resource planning.

Q8. How does the frequency of exercise and the type of occupation influence the presence of medical conditions?

Why: Lifestyle and professional environment play a crucial role in determining health outcomes and risk exposure. Analyzing these relationships uncovers key behavioral patterns.

Usefulness: Allows stakeholders to visualize how activity levels and job types affect long-term health, encouraging workplace wellness policies and active lifestyle initiatives.

Q9. Which occupations have the highest average insurance costs?

Why: Certain professions may involve physical strain, stress, or exposure to hazards that increase medical expenses. Recognizing these trends provides deeper insights into occupational health risks.

Usefulness: Enables insurers to design fair occupation-based premium structures and helps employers implement targeted wellness and safety programs.

Q10. How does exercise frequency relate to BMI and health risk levels?

Why: Regular physical activity has a proven correlation with healthier BMI levels and reduced medical costs. Exploring this relationship highlights the value of active living.

Usefulness: Supports the creation of visual dashboards that show how exercise habits influence weight management and overall health, guiding wellness-driven policy decisions.

Q11. What are the top 5 factors contributing to high insurance costs?

Why: Identifying the main cost drivers such as age, BMI, or medical history helps in understanding the financial landscape of health insurance.

Usefulness: Provides a foundation for predictive modeling and cost-control strategies, enabling insurers to enhance pricing efficiency and customer segmentation.

Q12. Are there any seasonal or regional trends in insurance cost distribution?

Why: Medical expenses can fluctuate across seasons or regions due to environmental, behavioral, or policy-related factors. Detecting these trends helps forecast demand patterns.

Usefulness: Aids insurers and public agencies in anticipating healthcare surges, planning preventive measures, and refining policy or resource allocation strategies accordingly.

Section 4: References

Link to the Dataset

Link to the Mural