US Healthcare Charges

Background

- Rising healthcare costs and insurance affordability are key concerns globally.
- Analyzing health insurance data helps identify cost drivers, risk factors, and policy gaps.
- The dataset provides insights into broader industry challenges and potential areas for policy improvement.

Hypothesis

- Higher age and BMI directly correlate with increased health risks (e.g., obesity, cardiovascular disease).

Project Objectives

- Provide data-driven insights into healthcare costs for insurers and -policymakers.
- Identify and analyze key risk factors using practice-set data.

Intended Impact

- Investigating data-driven evidence to validate or challenge assumptions.
- Ensuring fair and accurate insurance models that avoid oversimplified risk assessments.

- Policymakers: Track healthcare cost trends and implement timely interventions.
- Insurers: Use insights for budgeting and risk assessment.
- Both: Develop preventive measures to support healthier populations.

Data

- Practice set provided by Kaggle (https://www.kaggle.com/datasets/mirichoi0218/insurance/data)

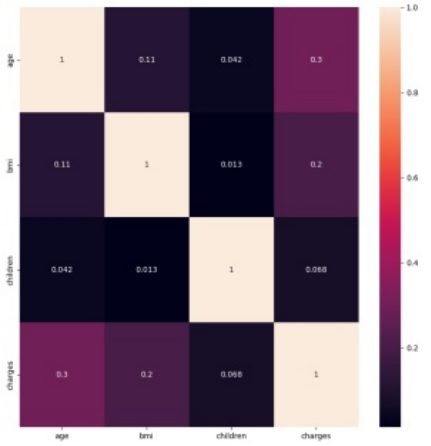
US Healthcare Charges

Health Charge Determinants

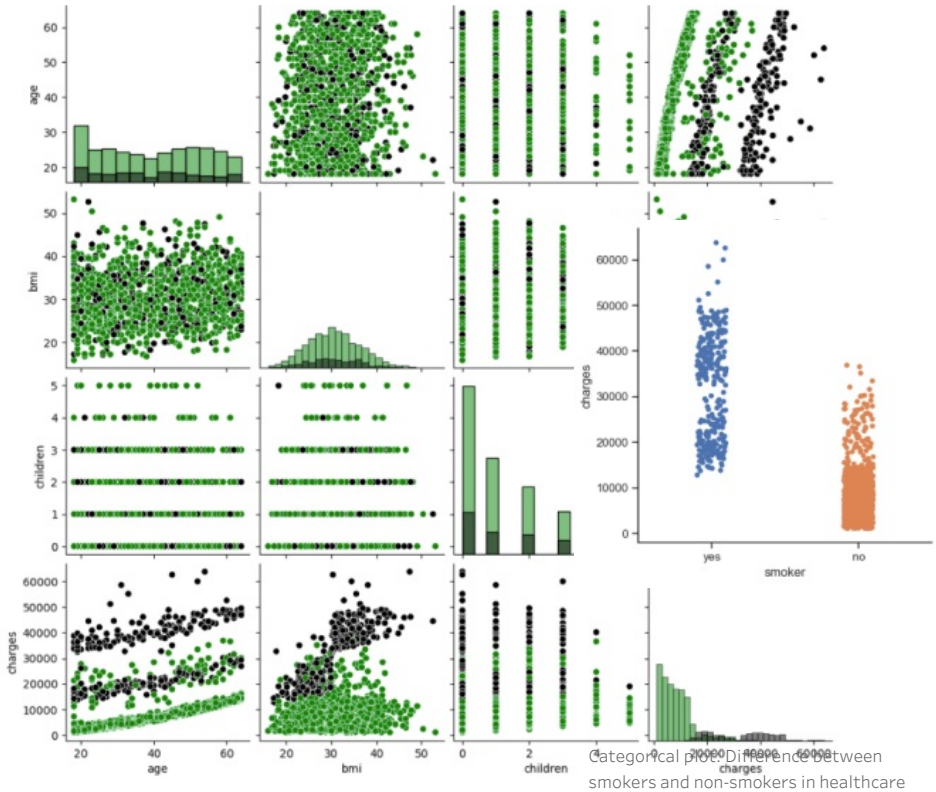
Exploratory Analysis: The data set provided information on categorical variables (sex, smoking). Smoking has shown to be an important confounding variable, gender did not show to be correlated with either numerical variables.

Numerical variables: charges, age, BMI, no. of children.

Overall variables show suprising weak linear correlation with charges. No correlation is stronge then 0,3 but relative to considered variables, the correlation is strongest for BMI, age with charges.



Correlation matrix with all numerical variables



Pair plot with all numerical variables

Categorical plot: Difference between smokers and non-smokers in healthcare charges.

US Healthcare Charges

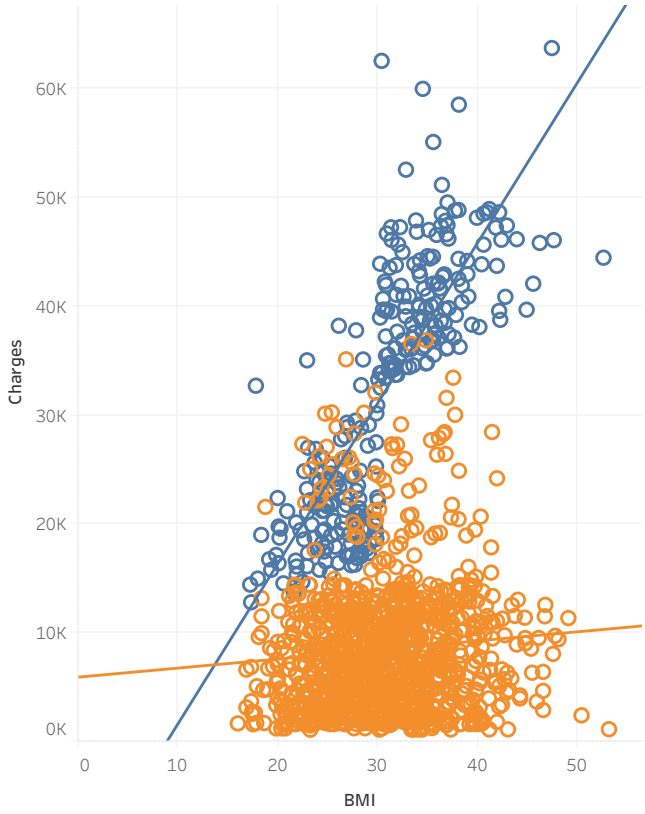
Age and BMI vs. Healthcare Charges

Based on previous exploratory analysis: BMI and charges are the strongest predictors of healthcare charges - allowing to focus on these two key factors.

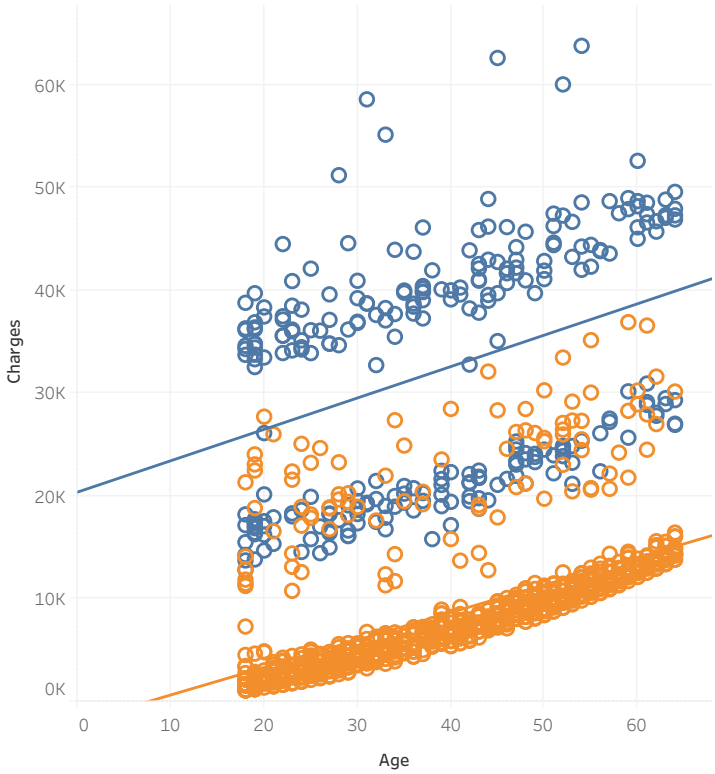
Smoking has strong influence on both factors, increasing health charges strongly.

Smoker no yes

BMI vs. Charges



Age vs. Charges



US Healthcare Charges

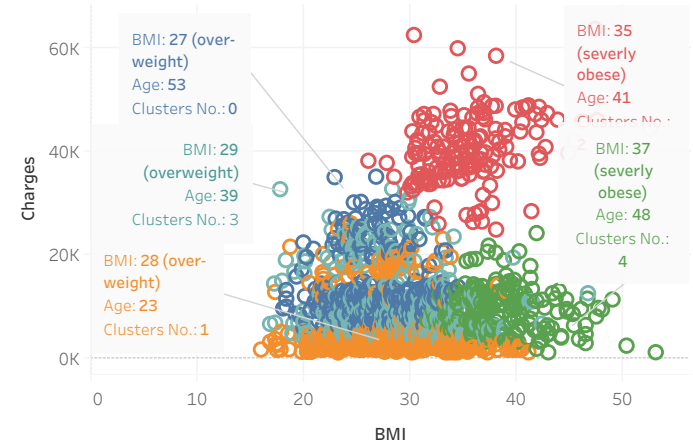
Risk Clusters based on BMI & Age

BMI and age form distinguishable clusters: the clusters with highest charges have in common that they are severly obese or overweight. If these factors are combined with smoking, charges increase significantly.

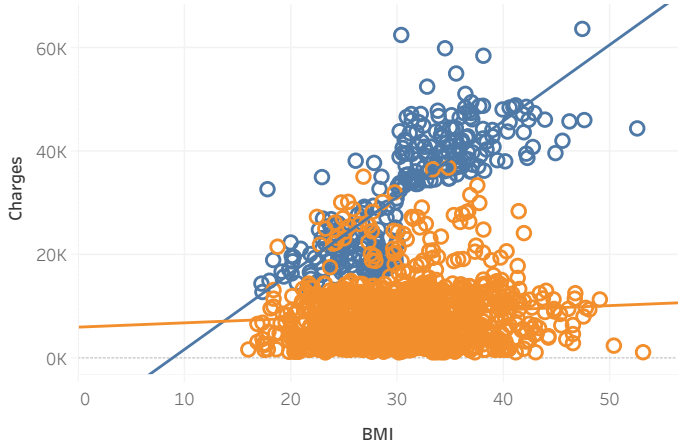
Group 4 vs Group 2 are both severly obese, yet Group 2 has higher charges, very likely due to smoking. The smoking effect even seems to offset the higher age of Group 4.

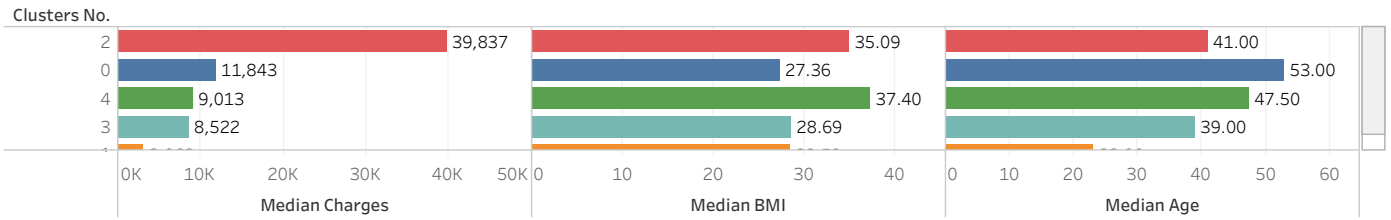
These findings highlight clear risk groups for insurance companies, with smokers consistently having higher costs regardless of other factors.

Five Clusters (with median values)



BMI vs. Charges





US Healthcare Charges

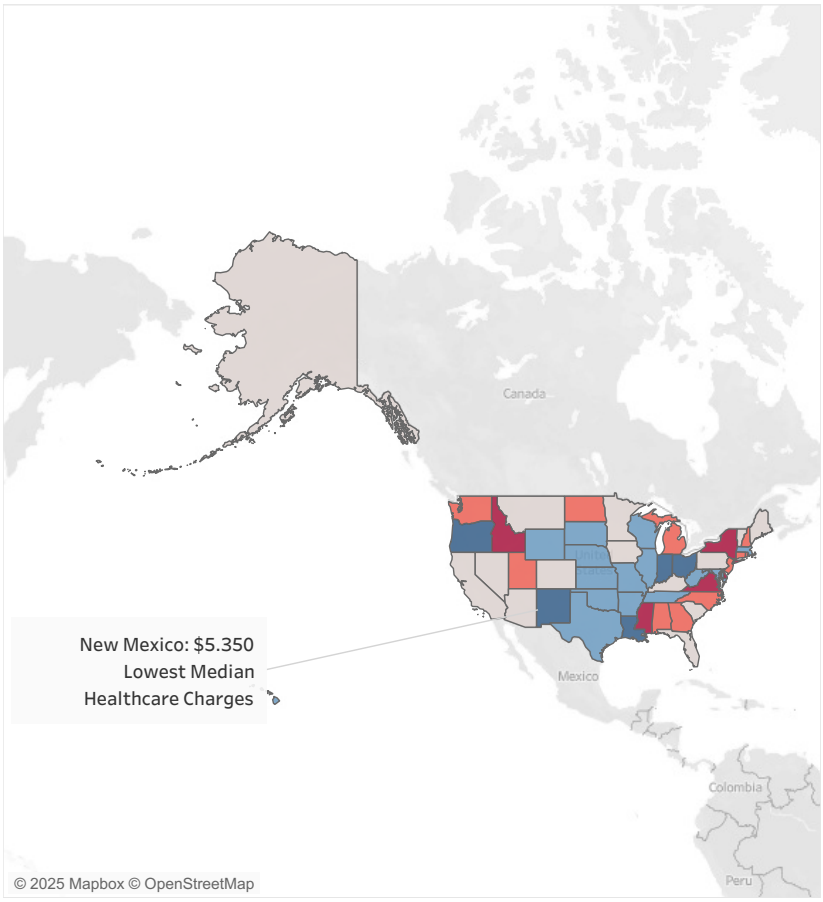
Healthcare Charges by State

To account for differences in population and extreme values of individuals, *median* health care charges are mapped.

Top 5 states highest healthcare charges
1. Virginia 13920
2. Mississippi 13458
3. New York 12731
4. Delaware 12558
5. Idaho 12302

Top 5 states with lowest healthcare charges
1. New Mexico 5350
2. Indiana 6078
3. Ohio 6238
4. Louisiana 6594
5. Oregon 6933

Key Results
- **Healthcare charges vary significantly by state**, with Virginia having median costs more than **2.6x** higher than New Mexico.
- **Geographic and policy factors** likely play a role in cost differences. States with higher costs may have fewer providers, higher hospital charges, or more chronic disease cases.
- Using the **median** instead of the **mean** provides a fairer compariso..



Median Charges
5,350 13,920

Select to Filter Map by State

State	
Virginia	18,
Mississippi	17,
Alabama	17,
Delaware	17,
Arizona	16,
Kentucky	16,
Texas	15,
North Carol..	15,
New York	15,
Rhode Island	15,
South Carol..	15,
New Hamps..	14,
Michigan	14,
Alaska	14,
Idaho	14,
Maryland	14,
Maine	14,
Georgia	13,
Nevada	13,
Pennsylvan..	13,
West Virgin..	13,
Utah	13,
Washington	13,
Tennessee	13,
Florida	13,
Oregon	13,
New Jersey	13,
Arkansas	12,
Wisconsin	12,
North Dako..	12,
Ohio	12,
Connecticut	12,

US Healthcare Charges

Key Results

The analysis supports the hypothesis that **higher age and BMI correlate with increased health risks** and costs. However, **smoking amplifies these effects**, significantly increasing healthcare charges regardless of other factors.

Risk Factors & Clusters

- **BMI and age** influence costs, but the highest charges occur in **severely obese or overweight** individuals—especially smokers.
- **Smoking is the strongest cost driver**, sometimes outweighing the effects of age.
- Distinct risk groups emerge, with **smokers consistently incurring higher costs**.

Regional Cost Differences

- **Healthcare charges vary significantly by state**, with Virginia's median costs **2.6x higher** than New Mexico's.
- Geographic and policy factors likely contribute to these differences.

Limitations and Ethical Considerations

- The dataset is **synthetic** and lacks a timeframe, limiting real-world applicability.
- **BMI** and **smoking** are simplified metrics that may not fully reflect health risks (e.g. neglect of body composition measures, amount of smoking was not captured).
- **State-level data was approximated**, making comparisons less precise.

Ethical concerns include **potential bias in insurance models**, requiring fairness and transparency.

Broader Implications

Findings highlight the need for **more nuanced insurance models** that account for the **combined impact of smoking, BMI, and age**. Insurers and policymakers should:..

US Healthcare Charges

Recommendations for Insurers & Policymakers

- **Refine Risk Models** – Incorporate smoking as a key cost driver alongside BMI and age to improve pricing accuracy.

- **Fair & Transparent Pricing** – Ensure models do not disproportionately impact vulnerable groups by addressing potential biases.

- **Targeted Prevention Strategies** – Develop wellness incentives, smoking cessation programs, and obesity interventions to lower long-term costs e.g. offer free fitness and health programs, subsidies for fitness and health-related expenses (gym, fitness-trackers).

- **State-Level Policy Adjustments** – Investigate regional disparities to improve healthcare access and affordability in high-cost states.

Next Steps for Further Analysis

- **Assess Real-World Data** – Validate findings using actual claims data with a defined timeframe and additional health factors.

- **Explore Non-Linear Relationships** – Investigate potential non-linear effects of BMI, age, and smoking on healthcare costs.

- **Deep Dive into Regional Differences** – Analyze policy, provider availability, and chronic disease prevalence by state.

- **Expand Predictors** – Include additional health indicators (e.g., chronic conditions, lifestyle factors) for a more comprehensive risk model.