Decision Making

Section 1: Used Visualization Tools

We have selected Tableau Desktop and Tableau Prep Builder as the primary technology stack for this dashboard. Below is the justification for these choices and our reasoning for preferring them over alternative tools like Microsoft Excel or standard Python plotting libraries.

1. Tableau Prep Builder (Data Pre-processing)

We are using Tableau Prep Builder to handle the ETL (Extract, Transform, Load) process before the data ever reaches the dashboard. This includes cleaning null values, categorizing the 1 million BMI records, and generating the custom "Risk Score."

- Why we prefer it over Excel: Microsoft Excel has a hard row limit of roughly 1 million rows (1,048,576). Since our dataset hits this ceiling, opening and manipulating the file in Excel causes crashes and extreme lag. Tableau Prep handles millions of rows efficiently without freezing.
- Why we prefer it over Tableau Desktop (Direct Connection): While Tableau Desktop *can* create calculated fields, doing so for 1 million rows requires the processor to recalculate the math every time a user filters the dashboard. By using Prep to "pre-calculate" fields like BMI_Category and outputting a static .hyper file, we shift the processing load offline. This ensures the final dashboard is responsive and lag-free for the end-user.

2. Tableau Desktop (Visualization & Storytelling)

Tableau Desktop is our main tool for constructing the interactive dashboard, the "Story" narrative, and performing K-Means Clustering.

- Why we prefer it over Python (Matplotlib/Seaborn): While we used Python for our initial static data exploration, it lacks the user-friendly interactivity required by our stakeholders (Insurance Underwriters). Tableau allows us to build dynamic features—such as the "What-If" Parameters for smoking cessation analysis—without needing to write complex front-end code. This interactivity is crucial for the "Decision Making" phase of the project.
- Why we prefer it over Power BI: We prefer Tableau for this specific project due to its superior handling of complex, non-standard charts. Our project plan involves a Butterfly Chart (Diverging Bar) and Density Maps to visualize regional risk. Tableau's "drag-and-drop" marks card offers more flexibility to build these custom visuals compared to Power BI's more rigid template structure. Additionally, Tableau's ".hyper" extract engine provides faster performance for our large 1GB+ dataset compared to standard import modes.

Section 2: Explanation of Required Data Pre-processing

Our initial Python analysis of the insurance_dataset.csv revealed that the data is clean with zero missing values across all 1 million records. However, to enhance the "Risk Analysis" and "Population Health" user requirements, the following preprocessing steps (feature engineering) will be performed in Tableau Prep:

1. BMI Categorization: The raw bmi column (e.g., 24.5, 31.2) is too granular for high-level executive summaries. We will create a calculated field BMI_Category to group values into standard medical ranges: *Underweight (<18.5)*, *Normal (18.5-24.9)*, *Overweight (25-29.9)*, and *Obese (>30)*.
2. Age Binning: We will create Age_Group bins (e.g., *Young Adult 18-30*, *Middle Age 31-50*, *Senior 50+*) to allow for easier demographic segmentation.
3. Encoding Binary Variables: While Tableau handles strings well, we may create a binary flag for Is_Smoker (0/1) to facilitate calculating "Correlation with Charges" matrices.

Section 3: List of Final Sets of Questions

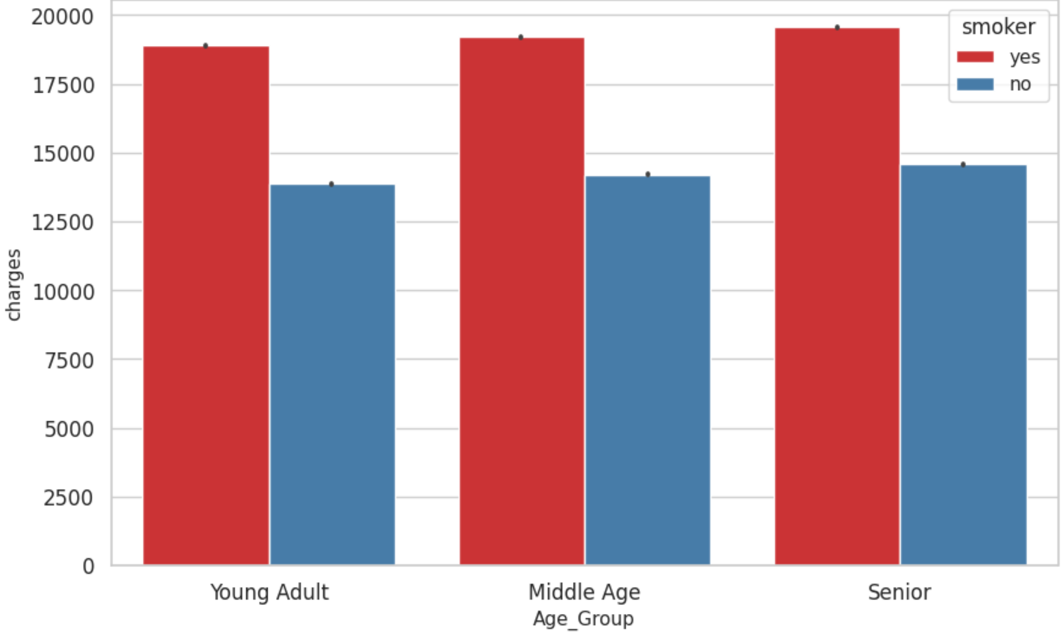
1. How does smoking status impact insurance charges across different age groups and what is the cost gap between smokers and non-smokers?
2. Is there a correlation between BMI categories and insurance costs?
3. How does exercise frequency moderate this relationship?
4. What are the top 5 risk factors contributing to high insurance costs?
5. How do average insurance charges vary across different regions and what factors drive regional disparities?
6. Which occupations have the highest average insurance costs?
7. What is the demographic profile of the highest risk group ?
8. How does the combination of medical history and family medical history influence insurance charges across coverage levels?
9. What is the distribution of medical conditions across the population and how do they vary by occupation and exercise frequency?
10. How do insurance charges progress with age, and what role do modifiable factors play in cost escalation?
11. Is there a significant difference in insurance charges between genders?

Section 4: Dashboard Plot Drafts

1. **How does smoking status impact insurance charges across different age groups?**

   Grouped Bar Chart

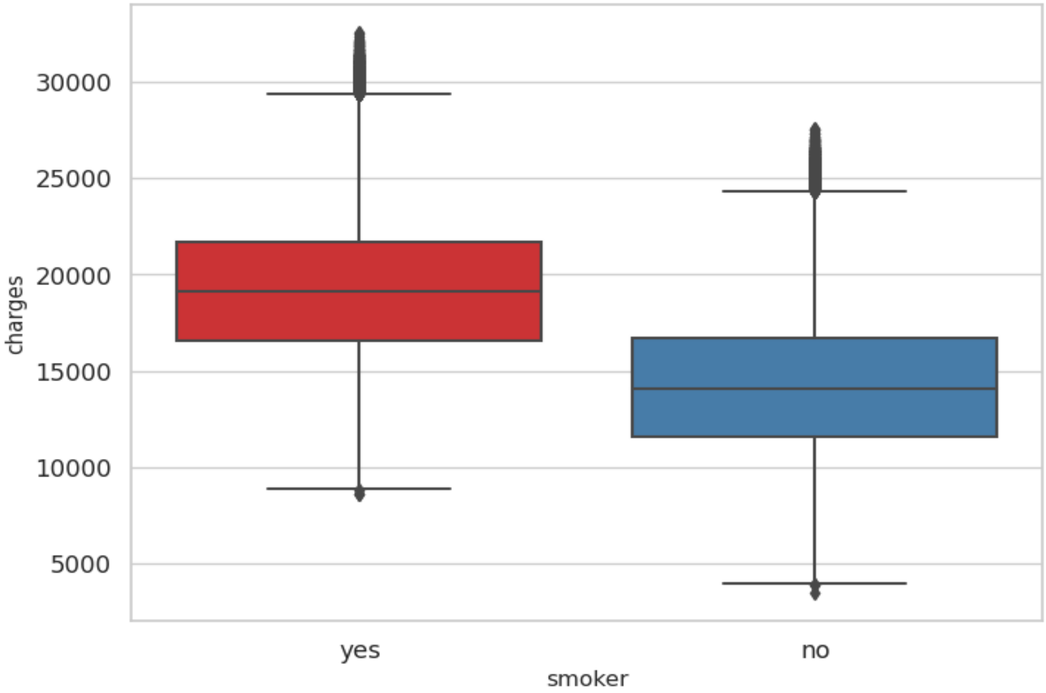
   - **Explanation:** The grouped bar chart segments the population into Young Adult, Middle Age, and Senior groups, displaying average insurance charges for Smokers and Non-Smokers side by side within each age category. It's easy to spot the price gap i.e smokers always pay more, no matter how old they are
   - **Pre- Attentive Attributes:**
     - Length (Charges)
     - Colour (Red for Smokers, Blue for Non-Smokers)
   - **Rough Sketch:**



2. **What is the cost gap between smokers and non-smokers?**

Box Plot

- **Explanation:** The box plot lays out insurance charges for smokers and non-smokers, showing things like medians, quartiles, and outliers for both groups. It's distinctly visible that smokers pay so much more than even their cheapest rates which makes it easier to study the population. It really shows how expensive smoking gets.
- **Pre- Attentive Attributes:**
  - 2D Position (Horizontal Marker lines like Median, Mean, Mode)

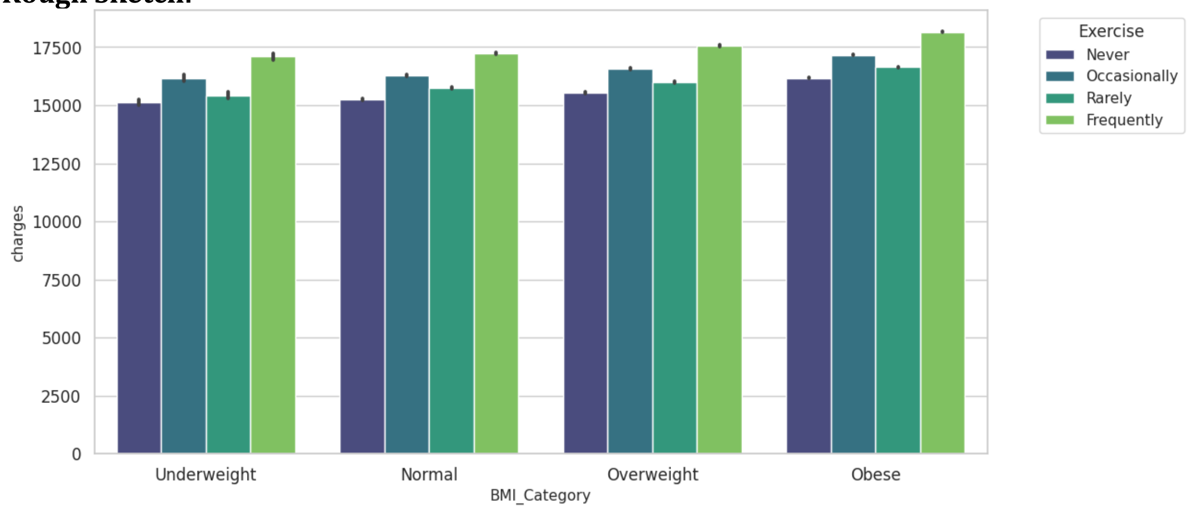


3. **Is there a correlation between BMI categories and insurance costs, and how does exercise frequency moderate this relationship?**

Clustered Bar Chart

- **Explanation:** The clustered bar chart displays BMI categories along the X-axis, segmented by exercise frequency i.e never, rarely, or frequently. So, we can quickly compare average insurance costs for each BMI group.
- **Pre- Attentive Attributes:**
  - Length (Charges)
  - Colour (Exercise Frequency Categories)

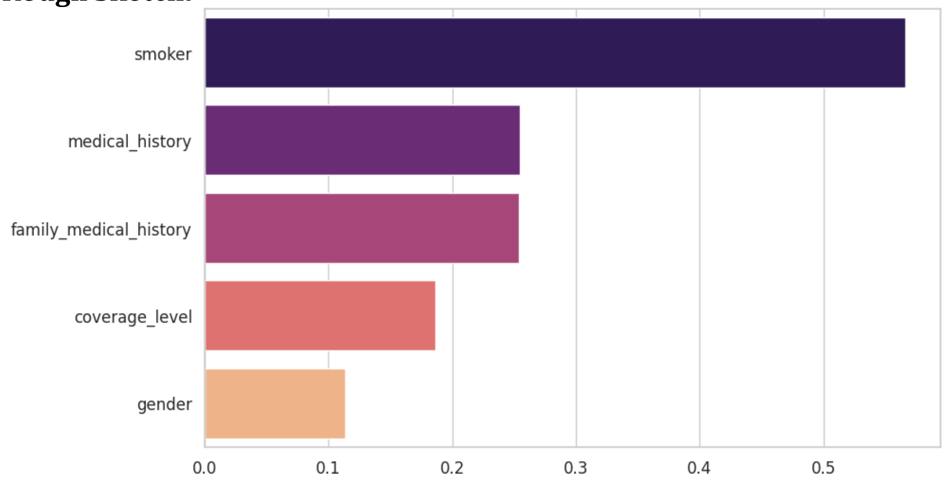
- **Rough Sketch:**



4. **What are the top 5 risk factors contributing to high insurance costs?**

   Horizontal Bar Chart

   - **Explanation:** This descending ordered horizontal bar chart lays out the main risk factors that strongly drive insurance costs, ranking them next to each other for a clear comparison. It combines things like smoking status, high BMI, and medical history into calculated fields. You can quickly spot the top five contributors and get a real sense of which health or lifestyle habits hit your wallet the hardest.
   - **Pre- Attentive Attributes:**
     - Length
   - **Rough Sketch:**

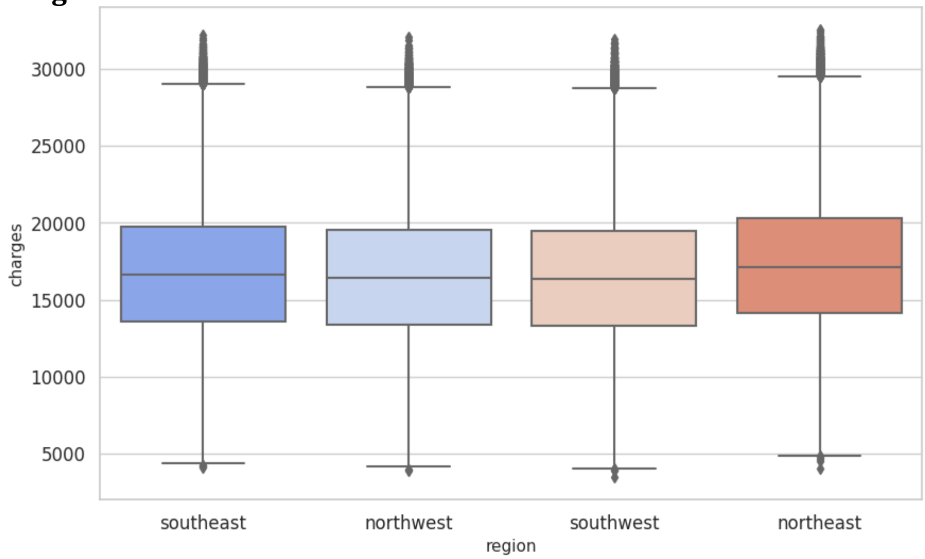


5. **How do average insurance charges vary across different regions and what factors drive regional disparities?**

   Box Plot

   - **Explanation:** This box plot shows how insurance charges vary in the four regions: NW, NE, SW, and SE. You can see the medians, quartiles, and outliers all at once. It's pretty easy to spot which regions have higher or more unpredictable costs. Furthermore, it gives you a starting point to dig deeper into what's causing these differences like whether it's lifestyle, access to healthcare, or something about the local population.
   - **Pre- Attentive Attributes:**

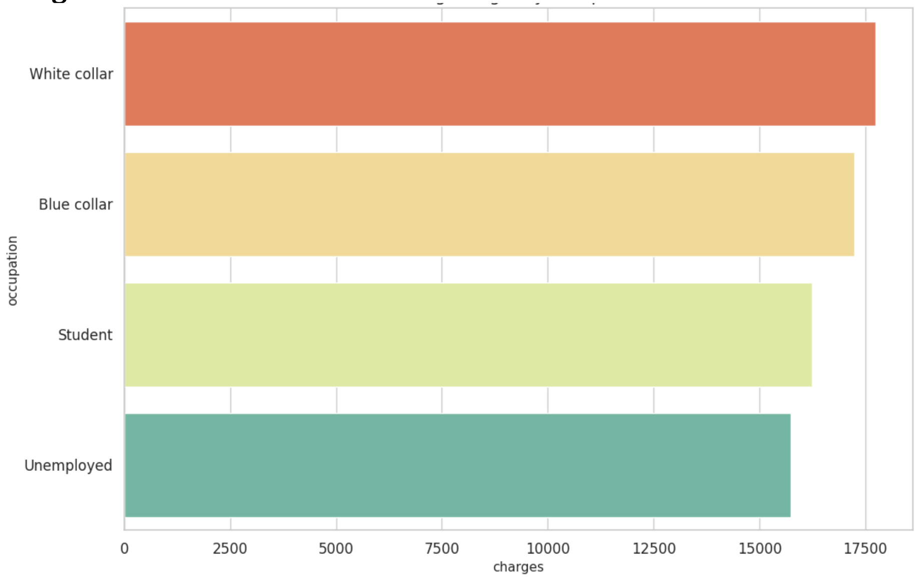
    - 2D Position
- **Rough Sketch:**



6. **Which occupations have the highest average insurance costs ?**

    Bar Chart

    - **Explanation:** The descending sorted bar chart lines up jobs like Blue Collar,White Collar, Student, and Unemployed by their average insurance costs. It makes it easy to identify which groups end up paying the most. Just scanning the bars we can know where the biggest differences are.
    - **Pre- Attentive Attributes:**
        - Length
    - **Rough Sketch:**

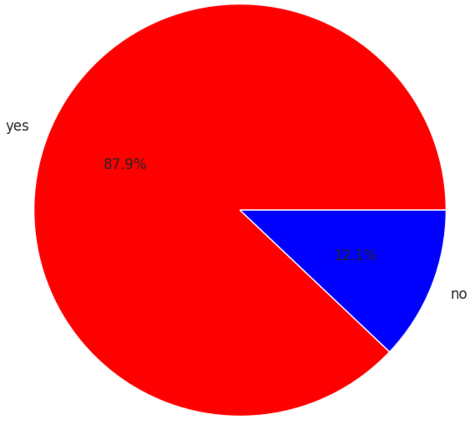


7. **What is the demographic profile of the highest risk group?**

    Pie Chart

    - **Explanation:** This pie chart zooms in on the top 20% of insurance claims, putting the biggest spenders front and center. It splits these high-cost policyholders into Smokers and Non-Smokers, so we can instantly spot which group racks up more expensive claims. The size of each slice tells us how much each subgroup contributes.

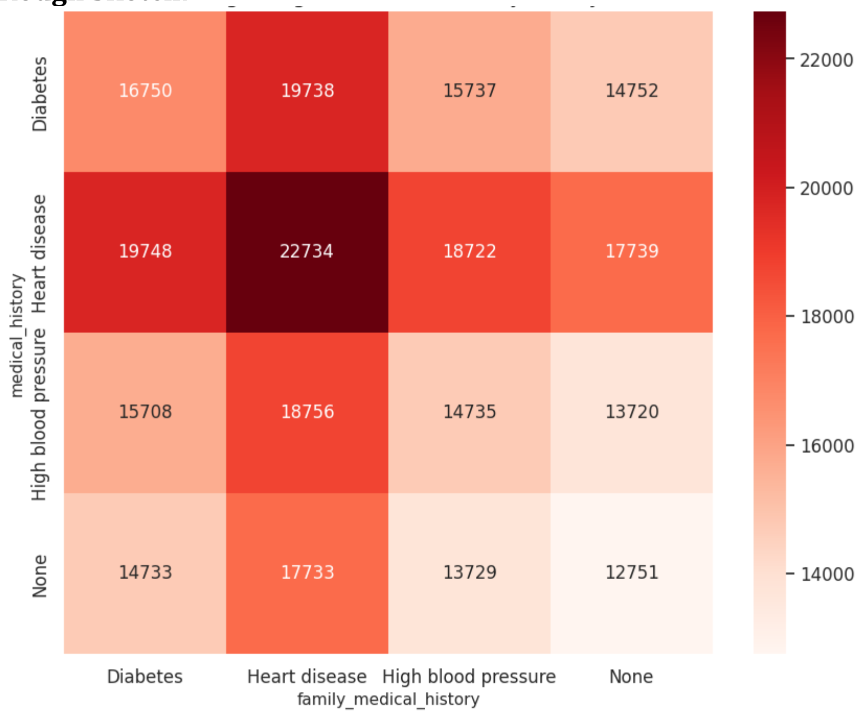
- **Pre- Attentive Attributes:**
  - Area
  - Colour
- **Rough Sketch:**



8. **How does the combination of medical history and family medical history influence insurance charges across coverage levels?**

Heatmap

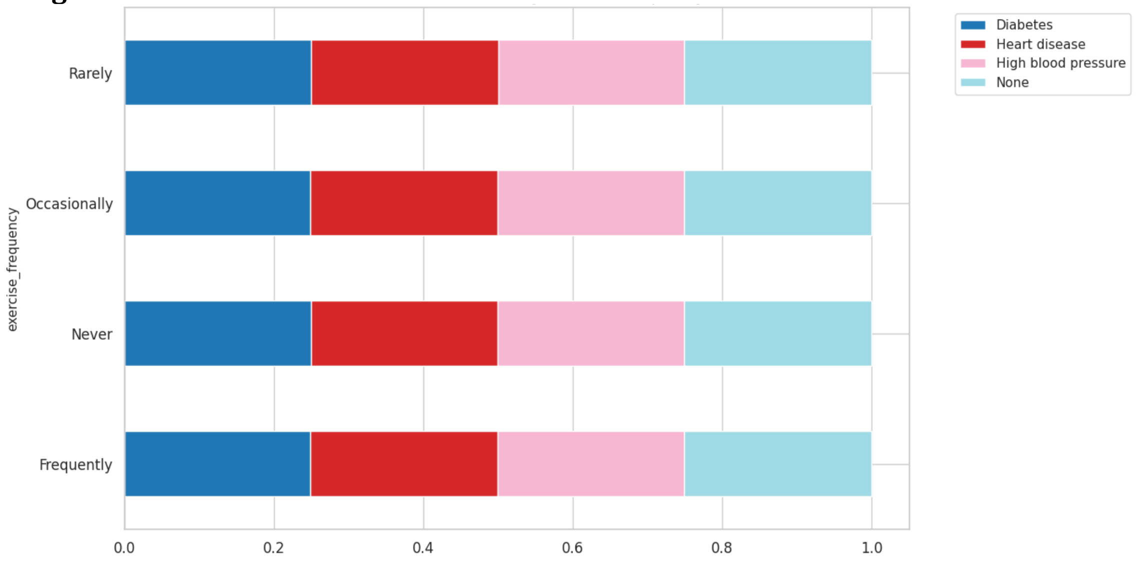
- **Explanation:** The heatmap lines up personal medical history on the Y-axis and family medical history on the X-axis. The color gets darker as insurance charges go up, so those high-cost “toxic combinations” such as certain personal conditions coupled with related family health risks. If you filter by coverage level, you can see how these risky pairings affect charges in Basic, Standard, or Premium plans.
- **Pre- Attentive Attributes:**
  - Colour (A Sequential colour palette for cost)
- **Rough Sketch:**



9. **What is the distribution of medical conditions across the population and how do they vary by occupation and exercise frequency?**

A 100% stacked bar chart

- **Explanation:** This 100% stacked bar chart sorts people by how often they exercise like never or frequently and shows what percentage in each group has different medical conditions. We can notice how heart disease or obesity show up more in some groups than others. It's a simple way to spot the connection between exercise habits and health, and you can quickly tell if people who rarely move have more health problems.
- **Pre- Attentive Attributes:**
  - Length (Segment size)
  - Color Hue (Condition)
- **Rough Sketch:**

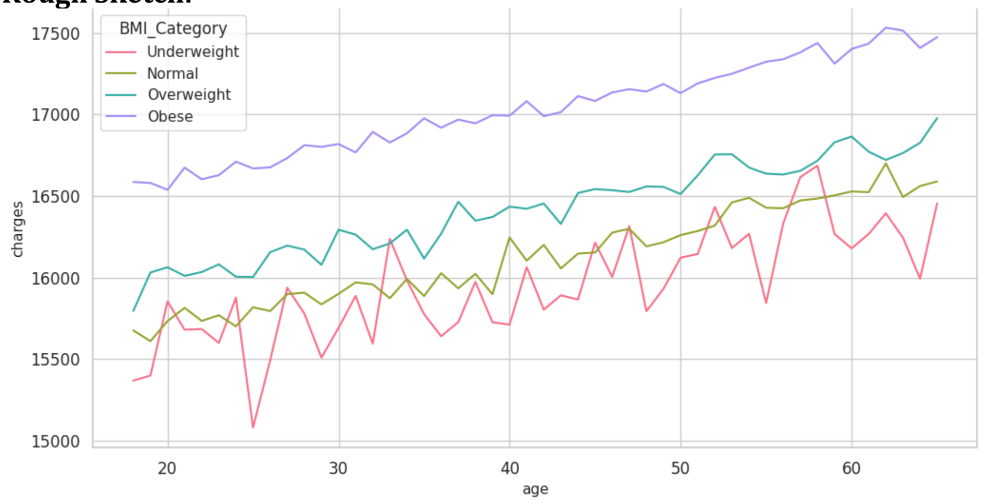


10. **How do insurance charges progress with age, and what role do modifiable factors play in cost escalation?**

Multiple Line Chart

- **Explanation:** This line chart plots age along the bottom and insurance charges up the vertical side.The lines are categorised by BMI categories. When we look at all the lines together, we can really spot how much extra weight ramps up long-term charges compared to others.
- **Pre- Attentive Attributes:**
  - Colour (BMI Category)
  - 2D Position

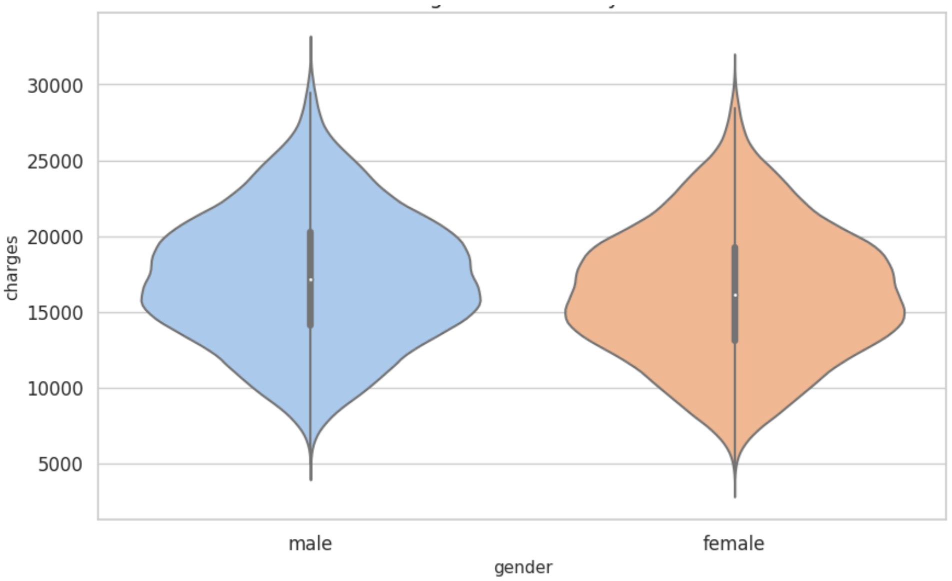
- **Rough Sketch:**



11. Is there a significant difference in insurance charges between genders?

Violin plot

- **Explanation:** A violin plot lays out how insurance charges distribution changes for men and women. The width of the plot at each cost level highlights where claims are concentrated and whether one gender has a larger share of high-cost cases, indicated by a “fatter tail."It's a quick way to see not just averages, but also who’s got more outliers and bigger swings in their costs.
- **Pre- Attentive Attributes:**
  - Shape (Distribution Width)
  - Position (Median Marker Line)
- **Rough Sketch:**



Section 5: Dashboard Interactivity

1. Region Filter (Dropdown)

Used For:
Filtering the dashboard based on the user's selected geographic region. This helps regional insurance teams evaluate local patterns in risk, medical conditions, and spending.

Connected Plots:
All 11 plots update when the region filter is applied.

Value Range:

- northeast
- northwest
- southeast
- Southwest

2. Smoker Scenario Toggle (Parameter)

Used For:
Performing a "What-If" analysis:

- "Current State" = actual smoker distribution
- "Simulated Non-Smoker" = everyone treated as a non-smoker

This helps model cost reduction if smoking decreases.

Connected Plots:

- **Plot 1** – Avg Charges by Age Group & Smoker Status
- **Plot 2** – Smoker vs Non-Smoker Box Plot
- **Plot 8** – Smoker Composition in Highest Risk Group

Value Range:

- Current State
- Projected Non-Smoking

3. BMI Range Slider

Used For:
Filtering users based on BMI range (e.g., show only obese users BMI > 30).
Useful for obesity risk and cost impact studies.

Connected Plots:

- Plot 3 – BMI Charges Moderated by Exercise
- Plot 9 – Medical vs Family History Heatmap
- Plot 11 – Cost Progression with Age by BMI

Value Range:
BMI: 15.0 – 55.0

4. Exercise Frequency Filter (Dropdown)

Used For:
Segmenting users by exercise habit to identify lifestyle-related cost patterns.

Connected Plots:

- **Plot 3** – BMI × Exercise Frequency
- **Plot 10** – Medical Conditions by Exercise
- **Plot 11** – Cost Progression with Age (exercise impacts BMI trends)

Value Range:

- Never
- Rarely
- Occasionally
- Frequently

5. Age Group Selector (Button or Dropdown)

Used For:
Filtering visualizations based on age segments created in Tableau Prep.
Helps compare younger vs older population risk.

Connected Plots:

- **Plot 1** – Smoking vs Age Group
- **Plot 8** – High-Risk Group Composition
- **Plot 11** – Age Progression Line Chart

Value Range:

- 18–30 (Young Adult)
- 31–50 (Middle Age)
- 51–65 (Senior)

6. Occupation Filter (Dropdown)

Used For:
Allow users to analyze profession-based risk differences.

Connected Plots:

- **Plot 6** – Avg Charges by Occupation
- **Plot 7** – Medical Condition Prevalence by Occupation
- **Plot 10** – Medical Conditions by Exercise (optional cross-filter)

Value Range:

- Blue collar
- White collar
- Student
- Unemployed

7. Medical Condition Filter (Multi-Select)

Used For:
Filtering by specific health issues to evaluate risk.

Connected Plots:

- **Plot 7** – Condition Prevalence
- **Plot 9** – Medical × Family History Heatmap
- **Plot 10** – Condition by Exercise Frequency

Value Range:

- Diabetes
- High blood pressure
- Heart disease
- None

8. Family History Filter (Checkbox / Multi-select)

Used For:
Understanding genetic risk impact.

Connected Plots:

- **Plot 9** – Medical History vs Family History Heatmap

Value Range:

- Diabetes
- High blood pressure
- Heart disease
- None

9. Coverage Level Filter (Dropdown)

Used For:
Evaluating cost differences in:

- Basic
- Standard
- Premium

Connected Plots:

- **Plot 9** – Medical × Family History

Value Range:

- Basic
- Standard
- Premium

10. Metric Selector (Radio Button)

Used For:
Letting users switch the metric for consistency analysis:

- Average Charges
- Median Charges

This prevents outliers from skewing results.

Connected Plots:

- **Plot 1** – Smoker × Age
- **Plot 3** – BMI × Exercise
- **Plot 5** – Regional Cost Variations
- **Plot 6** – Occupation Costs

Value Range:

- Average
- Median

11. Tooltip Hover Interactivity (Automatic)

Used For:
Showing additional hidden details such as:

- Exact charges
- BMI values
- Region
- Medical condition
- Exercise level
- Count of records

Connected Plots:
All 11 plots.

12. Legend Click-to-Highlight (Built-in Tableau Feature)

Used For:
Allows users to click on a legend item (e.g., "Smoker") and dynamically highlight or mute it in the chart.

Connected Plots:

- **Plot 1** (Smoker legend)
- **Plot 3** (Exercise legend)
- **Plot 7** (Condition legend)
- **Plot 10** (Exercise × Condition)
- **Plot 11** (BMI Category legend)

Value Range:
Loaded automatically from marks card (smoker, BMI category, etc.)

Section 6: References

https://app.mural.co/t/group15dv8892/m/group15dv8892/1763500714125/4dc1aaa82f36983bc3e737de2ae9015bdf3eb7a0?sender=ueae1aa4802689266eefd8638

https://www.kaggle.com/datasets/sridharstreaks/insurance-data-for-machine-learning