**URS project:**

1. **Kaggle Medical insurance dataset**

Source: <https://www.kaggle.com/datasets/mosapabdelghany/medical-insurance-cost-dataset>

Variables:

* age: Age of primary beneficiary (int)
* sex: Gender of beneficiary (male, female)
* bmi: Body Mass Index, a measure of body fat based on height and weight (float)
* children: Number of children covered by health insurance (int)
* smoker: Smoking status of the beneficiary (yes, no)
* region: Residential region in the US (northeast, northwest, southeast, southwest)
* charges: Medical insurance cost billed to the beneficiary (float)
  1. **Generate scatter plot between charges and age/sex/bmi/children/smoker/region**

Since the variables sex, smoker, and region are categorical rather than numerical, their data distributions require different visualization techniques. To explore these distributions, I initially used a swarm plot, which spreads individual data points horizontally to avoid overlap. This type of plot is particularly effective for displaying the distribution of numerical values across categorical groups. Interestingly, the overall shape of the swarm plot resembles that of a violin plot, which led me to further investigate the dataset using violin plots. Violin plots not only show the distribution density but also provide a more compact and informative summary of the data across categories.

A graph of a number of blue dots

AI-generated content may be incorrect.

Three distinct linear patterns between age and three charges range.

Figure 1 Scatter Plot between age and charges

*A graph of a graph of a number of points

AI-generated content may be incorrect.*

Figure 2 Segmented linear regression line between age and chages

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Segment | N | Linear regression Equation |  | Pearson r |
| Low charge  (0-17,000) | 1007 | Charges = -1803.15 + 235.71 \* Age | 0.665 | 0.816 |
| Medium charge  (15,000-32,000) | 203 | Charges = 15894.96 + 147.19 \* Age | 0.283 | 0.532 |
| High Charge (31,000-60,000) | 153 | Charges = 31783.55 + 227.14 \* Age | 0.394 | 0.628 |

A diagram of blue dots

AI-generated content may be incorrect.

Below BMI 30: increasing variance

Below BMI 30: clustered

BMI Index

**Overweight**

**Normal weight**

**Extremely obese**

**Underweight**

**Obese**

Figure 3 Scatter plot between bmi and charges

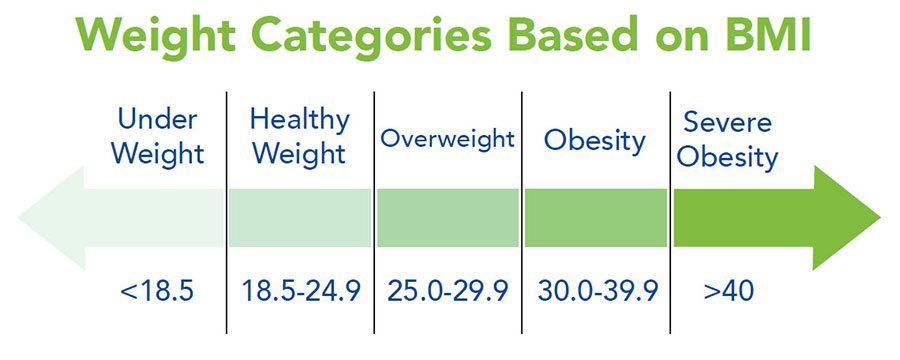


Figure 4 Weight categories based on BMI

1. **The BMI Threshold at 30**

The most striking feature is the dramatic change in the data's pattern around (the clinical definition of **obesity**).

* **Below :** Charges are relatively low, tightly clustered, and primarily follow a single, gentle upward trend.
* **Above (Obesity):** The costs suddenly **bifurcate** (split). A large segment of the population remains on the lower charge trend, but a new, distinct population appears, showing a steep increase in maximum and average charges.

1. **Increased Variance**

Beyond , the **spread or vertical distance** between the dots drastically increases.

* This means that for someone with a high BMI, we are much **less certain** of their exact charge. They could be anywhere from the low-cost baseline to the highest-cost bracket.

1. **Further investigation**

* Since BMI is a comprehensive indicator of body weight that also reflects factors such as age, sex, and other physiological characteristics, its relationship with medical charges is complex. In this project, I focus on the variables sex, age, and smoking status to better understand this relationship. The association between BMI and charges cannot be adequately captured by a simple linear or polynomial regression model. Instead, a more nuanced approach is required, potentially involving multiple regression, piecewise regression, or other advanced statistical techniques, to account for the interplay of multiple influencing factors.

A graph of blue dots

AI-generated content may be incorrect.

Figure 5 Scatter plot between children and charges

* 1. A counter-intuitive observation: Increasing number of children covered by medical health insurance doesn’t correlate with higher charges. Plus, there is a huge drop from 0-3 children to 4-5 children. Thus, only focusing on two variables at a time is not enough. By combining factors of age, smoking status, and age together, we could conclude this table:

|  |  |  |  |
| --- | --- | --- | --- |
| Children Group | Typical Age Profile | Highest Risk Individuals | Maximum Charge |
| 0 Children | Bimodal: Includes very young adults (low risk) AND very old adults (highest risk). | Oldest smokers (50-64 years). | Very High |
| 1-3 Children | Primarily middle-aged adults (30s-50s) who are still accumulating medical issues. | Middle-aged smokers (30s-50s). | Very High |
| 4-5 Children | Tends to be dominated by younger to middle-aged adults (20s-40s) due to the physiological constraints of having many children. | Fewer/No older adults (50s-64s). | Lower (Capped) |

Table 1 Multiple factors are incorporated into the analysis of children and charges

* 1. Further investigation: regression modeling for children vs. children
     + Use One-Hot Encoding (or dummy variables). You would create a dummy column for each category, using as the baseline.
     + use standard Multiple Linear Regression (MLR), assuming a linear relationship.A graph of a person and person

       AI-generated content may be incorrect.A graph of a diagram

       AI-generated content may be incorrect.A graph of different colored lines

       AI-generated content may be incorrect.