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
     1. **Scatter plot between charges and age**

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 |

Table 1 Summarized correlation tabel for linear regression between age and charge

To improve the best fit line accuracy, separting each segment into individual graph is essential to have a better focus.

1. **Low charge (0 – 17,000)**

Linear regression is not enough to fit the regerssion trend between age and low charge. Therefore, a polynomial regression should be applied.

Without making any procedures of cleaning data, the polynomial regression between age and charge is:

A graph with a red line

AI-generated content may be incorrect.

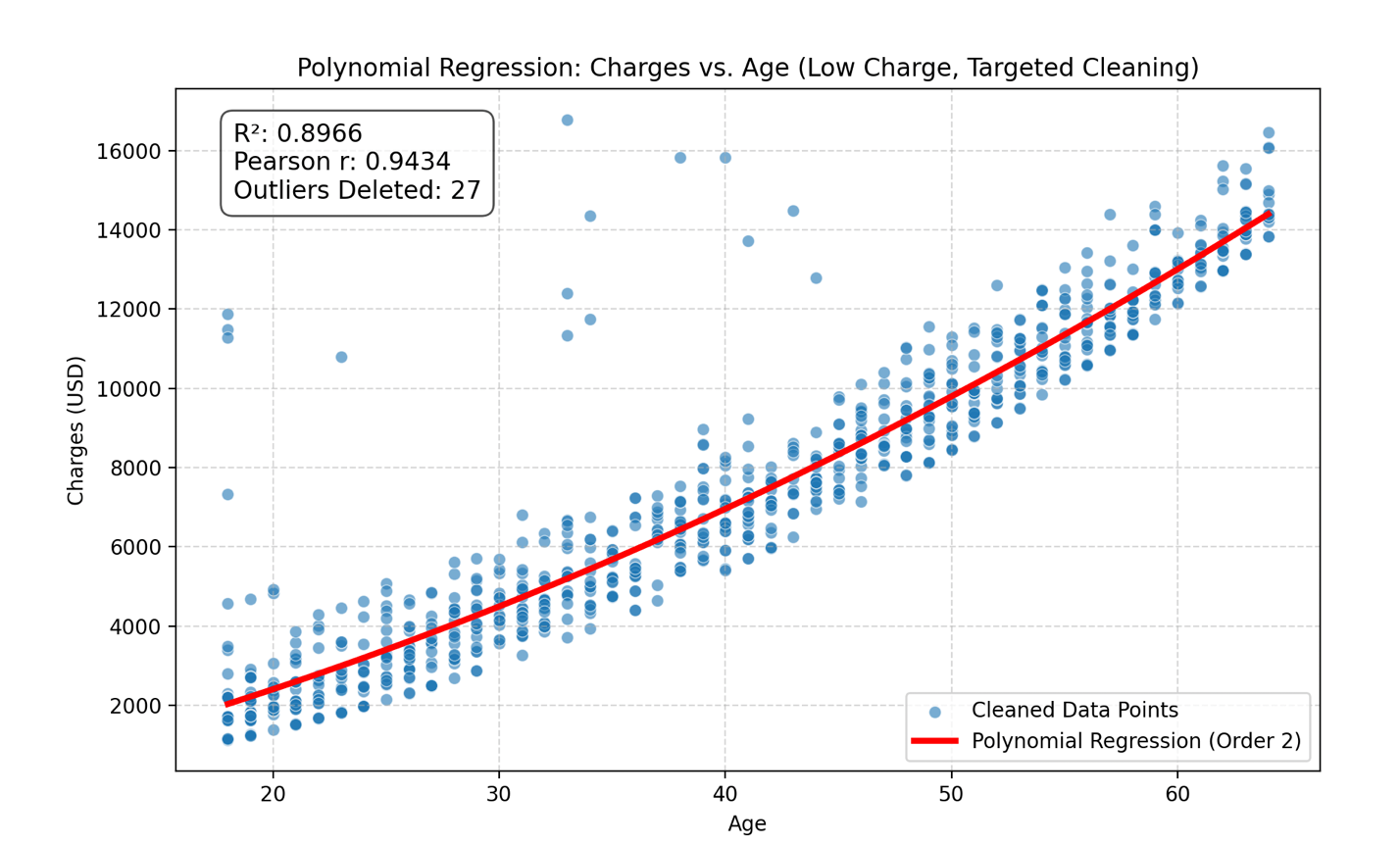
Figure 3 Polynomial regression between age and charge for low charge segment

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Segment | N | Polynomial regression Equation |  | Pearson r |
| Low charge  (0-17,000) | 1007 |  | 0.6817 | 0.8155 |

Table 2 Correlationa result for polynomial regression between age and low charge

Intuitively, the starting point of polynomial regression line should be located at the point around 2,100 to 2,600. The reason why current regression linear starts from 3,700 might be hugely influenced by noise data points on the left corner side. (age < 30 && charge >= 12,000). Therefore, a procedure of cleaning data is essential to be considered.

By using standard 1.5 \* IQR on both age and charges, no extreme outliers distorted the boundaries. Secondly, correspond to previous observation, adding a highly specific filter based on the visual scatter plot is essential: Age < 30 and Charges > 12,000.



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Segment | N | Polynomial regression Equation |  | Pearson r |
| Low charge  (0-17,000) | 980 |  | 0.8966 | 0.9434 |

Figure 4 Correlationa result for polynomial regression between age and low charge after stricter data cleaning

1. **Meidum charge (15,000 – 32,000)**

According to the result in table 1, both the coefficient of determination and Pearson r values are low, which indicate low correlation between two variables. Therefore, to further improve the accuracy and relationship between these two variables, polynomial regression analysis is necessary to be included. Since, the data scattering pattern is not as clear as low charge one, incorporating various regression methods are necessary. In the following discussion, quadratic, cubic, and power regression.

A graph with colored lines and numbers

AI-generated content may be incorrect.

Figure 5 Non-Liner regression for medium charge segmented data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | N  (after cleaning) | Equation |  | Pearson r |
| Cubic polynomial | 203 |  | 0.3368 | 0.5322 |
| Quadratic polynomial | 203 |  | 0.2832 |
| Power law | 203 |  | 0.2542 |

Table 3 Correlationa result for polynomial regression between age and medium charge

The Pearson value of 0.5322 indicates a moderate positive linear relationship between Age and Charges in the raw data. Older people in this segment tend to have slightly higher costs.

However, the low value of (even for the best non-linear model, Cubic Polynomial) means that our chosen predictor, Age, accounts for only about 33.7% of the total variance in charges. The remaining of the cost variation must be explained by factors not included in the model, which will be further discussed.

1. High charge ( 31,000 – 60,000)

A graph with a line and a line

AI-generated content may be incorrect.

Figure 6 Non-Liner regression for high charge segmented data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | N  (after cleaning) | Equation |  | Pearson r |
| Cubic polynomial | 203 |  | 0.4502 | 0.6687 |
| Quadratic polynomial | 203 |  | 0.4474 |
| Power law | 203 |  | 0.4376 |

Table 4 Correlationa result for polynomial regression between age and high charge

The Pearson r value of 0.6687 indicates a moderately strong, positive linear correlation between Age and Charges in the raw data. However, the best-fitting still shows that more than half (55%) of the cost variation in this high-charge group is due to factors other than age.

* + 1. **Scatter plot between charges and bmi**

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 7 Scatter plot between bmi and charges

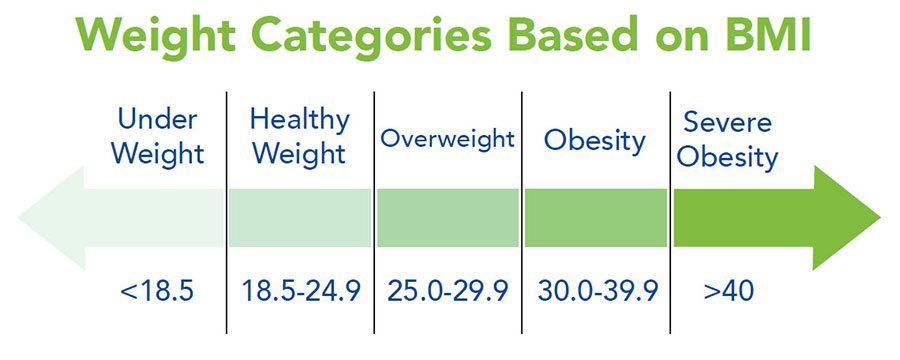


Figure 8 Weight categories based on BMI

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.
  + 1. **Scatter plot between charges and children**

A graph of blue dots

AI-generated content may be incorrect.

Figure 9 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 5 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). I would create a dummy column for each category, using as the baseline.
     + use standard Multiple Linear Regression (MLR), assuming a linear relationship.
     1. **Scatter plot between charges and sex**

The distribution for both male and female charges is heavily skewed, meaning most people have low charges, but there’s a long tail of very high charges. The density of points is highest in the to range for both sexes. The bulk of the points for male and female charges overlap almost entirely. It’s hard to visually identify a clear separation or group where one sex dominates the other.

A graph of a person and person

AI-generated content may be incorrect.

Figure 10 Swarm plot of sex and charges

The "violins" for both male and female charges are extremely wide at the bottom (low charges) and narrow quickly toward the top (high charges). This is definitive evidence of the high degree of positive skewness in insurance charges: most people incur low costs.

The most critical finding is that the shape and spread of the male and female violins are virtually identical. This confirms what was statistically observed: while the mean charges are statistically different by , the underlying risk distribution is not substantially different based on sex alone.

A diagram of two people

AI-generated content may be incorrect.

Figure Violin plot of the distribution between sex and charges

The visual evidence from the violin plot strongly supports the idea that sex is a weak predictor of charges.

A graph of a diagram

AI-generated content may be incorrect.

Figure 12 Swarm plot of smoker and charges

A diagram of a distribution of charge

AI-generated content may be incorrect.

Figure Violin plot of the distribution between smoking condition and charge

A graph of different colored lines

AI-generated content may be incorrect.

Figure 14 Swarm plots of region and charges

A diagram of different colored shapes

AI-generated content may be incorrect.

Figure Violin plot of the distribution between region and charges