8/25/2025

Faculty of science

postgraduate diploma in data science

Predicting Medical Costs Using Linear Regression

**Table of Contents**

1. Introduction……………………………………1
2. Method………..……………………………….1
   1. Data description……………………………1
   2. Tools……………………………………….1
   3. Procedural steps for analysis………………1
3. Results……………………………………..…..2
   1. Summary statistics………………………...2
   2. Correlation analysis……………………….2
   3. Univariate distribution analysis…………...2
   4. Bivariate interactions……………………...3
   5. Model Evaluation………………………….5
4. Discussion………………………………….….5
5. Conclusion…………………………………….5
6. **Introduction**

Linear regression is a statistical method for examining and simulating the relationship between a dependent variable and one or more independent variables (Montgomery, Peck and Vining, 2021). Linear regression provides an interpretable foundation for connecting expenses to demographic and behavioural characteristics. In order to forecast charges, this study uses linear regression with factors from the dataset, including age, sex, BMI, children, smoker, and region. I used a clear process that includes model specification, training, validation, critical evaluation, exploratory data analysis (EDA), and data description. To find determinants of medical charges, “Insurance dataset” is analysed.

**This report's primary goal is to**:

1. Conduct a descriptive analysis of the dataset.

2. Determine the connections between medical expenses and predictors.

3. Develop and assess a cost prediction model using linear regression.

1. **Method**
   1. **Data Description**

There are 1,338 observations in the insurance dataset, along with the following variables:

* Numeric: age (years), bmi (kg/m²), children (number of dependents), charges (USD)
* Categorical: smoker (yes/no), sex (female/male), region (northeast, northwest, southeast, southwest)
  1. **Tools**

The following Python libraries were used to perform the analysis: Scikit-learn for regression modelling and assessment; Matplotlib and Seaborn for data visualization; Pandas and NumPy for data processing.

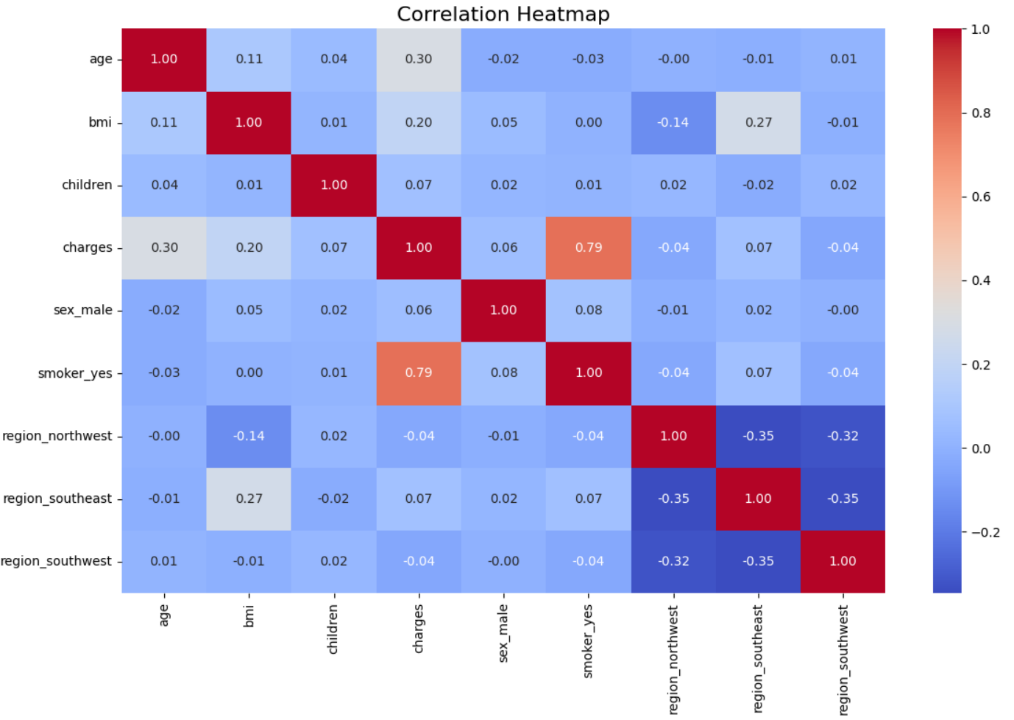
* 1. **Procedural steps for analysis**

1. Descriptive statistics were created in order to comprehend the distribution of variables.
2. Correlation analysis looked at how continuous variables are related to one another.
3. For visualizations- age, gender, charge, and smoking distributions were incorporated.
4. Medical expenses for male and female groups were compared using a boxplot.
5. To examine age versus charges, scatterplot with regression lines was employed.
6. Test (20%) and training (80%) sets of the dataset were separated.
7. The training data was fitted to a linear regression model.
8. Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R2 score were used to evaluate the model.
9. **Results**
   1. **Table 1. Summary statistics**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Age | BMI | Children | Charges (USD) |
| Count | 1338.0000 | 1338.0000 | 1338.0000 | 1338.0000 |
| Mean | 39.2070 | 30.6634 | 1.0949 | 13,270.4223 |
| Std | 14.0500 | 6.0982 | 1.2055 | 12,110.0112 |
| Min | 18.0000 | 15.96000 | 0.0000 | 1121.8739 |
| 25% | 27.0000 | 26.2963 | 0.0000 | 4740.2872 |
| 50% | 39.0000 | 30.4000 | 1.0000 | 9382.0330 |
| 75% | 51.0000 | 34.6938 | 2.0000 | 16,639.9125 |
| Max | 64.0000 | 53.1300 | 5.0000 | 63,770.4280 |

Interpretation: Both areas and sexes were evenly distributed in the dataset. The mean age was about 39 years old, and the mean BMI was 30.7 kg/m2, which is considered overweight. Charges were wildly disproportionate, with some people paying exorbitant fees.

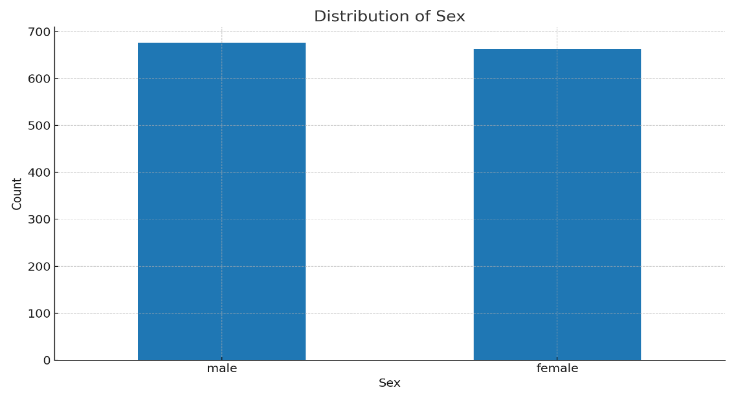
**3.2 figure 1. Correlation analysis**



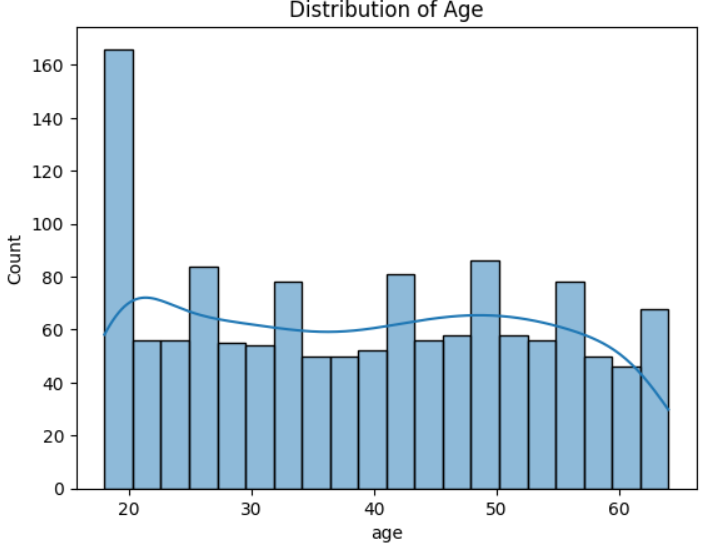
Interpretation: According to the correlation study, charges are positively connected with age, BMI, and smoking status, suggesting that medical expenses are often greater for smokers, older people, and those with higher BMIs. Smoking status showed the strongest correlation, indicating that it is a significant cost driver. On the other hand, characteristics like region showed smaller connections, suggesting that they had little effect on medical costs.

* 1. **Univariate distribution analysis:**

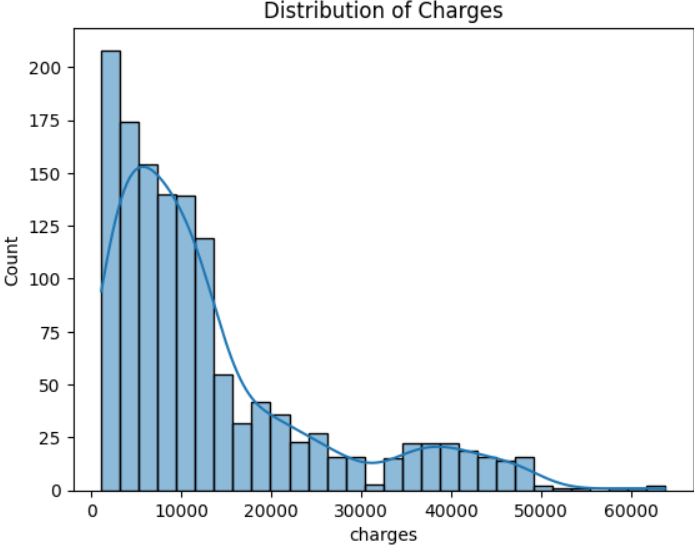
3.3.1 Figure 2. Sex: Almost balanced (female = 662; male = 676).



3.3.2 Figure 3. Age: Broadly uniform from 18 to 64; mild right‑tail at older ages.



3.3.3 Figure 4. Charges: Right‑skewed with a heavy tail and intense outliers



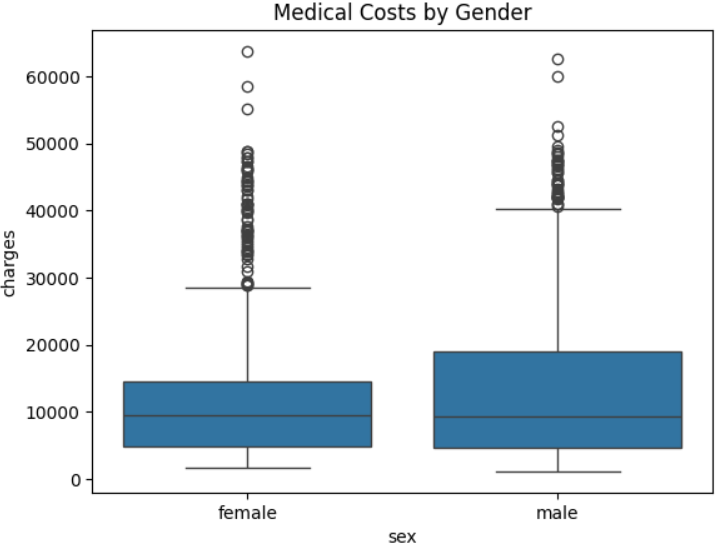
3.3.4 Figure 5. Smoking: majority are non-smokers, and smokers' charges were noticeably greater.

A graph with blue squares

AI-generated content may be incorrect.

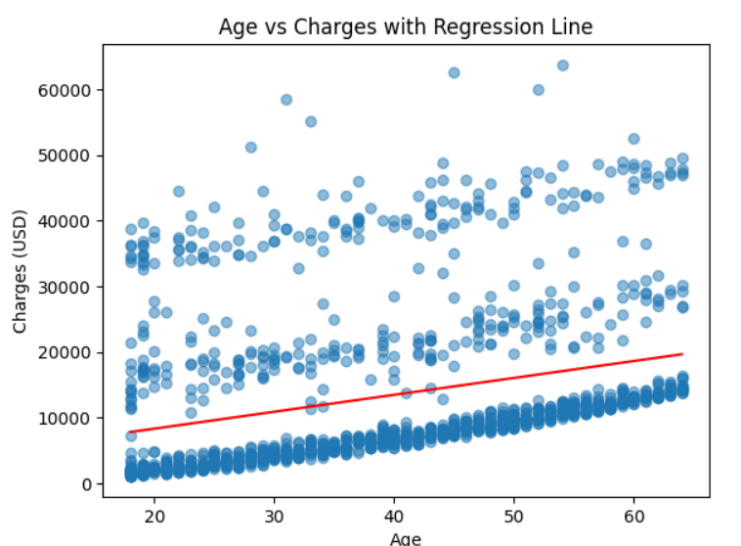
* 1. **Bivariate interactions**

3.4.1 Figure 6. Distribution of Medical Costs (Males vs Females)



Interpretation: Mean charges female ≈ $12,570, male ≈ $13,957; variability is high, and sex difference is modest. The median charges for males and females are comparable. For both categories, the interquartile range (spread) of charges is broad, indicating a significant variance in the expenses of individual medical care. There are numerous high outliers in both groups, particularly in the upper range of charges, indicating that a tiny percentage of people have extremely high medical expenses. The central tendency does not significantly differ between males and females, indicating that sex alone may not be the most reliable indicator of medical charges.

3.4.2 Figure 7. Age vs charges (simple regression line)



Interpretation: Age and medical charges were positively correlated, according to scatterplots and regression lines, although the variance rose with age. Additionally, the boxplot showed that smokers were consistently charged more across all age categories. Because smoking status and BMI were found to significantly correlate with charges in previous research, these factors have a major impact on the prevalence of large clusters of high charges among older adults (see Figure 1).

This trend shows that while age alone accounts for a portion of the variation in medical expenses, the relationship is significantly influenced by other lifestyle characteristics. Age is therefore not a reliable indicator of healthcare spending on its own, even though it has a positive linear relationship with costs.

* 1. **Table 2. Model evaluation:**

|  |  |
| --- | --- |
| Mean Square Error (MSE) | 33596915.85 |
| Root Mean Square Error (RMSE) | 5796.28 |
| R2 score | 0.7836 |

Interpretation: With an R2 score of 0.78, the linear regression model can account for 78% of the variation in medical bills. Given the skewed distribution of charges, the RMSE of 5796.28 indicates a moderate average forecast error.

1. **Discussion**

According to the correlation analysis, charges are strongly positively correlated with age, smoking, and BMI, which is consistent with the regression results. Expanding upon this, the subsequent conversation assesses the suitability and constraints of the fitted model.

In overall, a significant amount of the variance is explained by linear regression, with R2 = 0.78 indicating that the model accounts for about 78% of the variance in medical charges. However residual patterns indicate possible non-linearities. A log-link on charges may also help to increase residual normalcy and stabilize variation (James et al., 2021). From ethical point of view, including attributes such as smoking status carries risk of price discrimination and reputational harm. As articulated by Harrell (2015), regulatory and compliance audits must quantify and justify fairness metrics prior to production, paralleling the systematic assessment.

With regard to generalizability and bias, the cross-sectional structure of the dataset and its absence of temporal validation may restrict its applicability to populations in the future. Additionally, the stable distributions across groups assumed by encoded categorical variables could not hold true in practice (Wooldridge, 2016). In model risk concerns, supreme cases are hard to forecast due to heavy-tailed distributions in medical expenses. Since linear regression tends to underestimate such cases, alternative methods like quantile regression or generalized linear models (GLMs) might be more suited for insurance pricing (Kuhn and Johnson, 2013).

1. **Conclusion**

It can be concluded that the insurance dataset was well fitted by the linear regression model, which accounted for 78% of the variation in medical bills. Nonetheless, the heavy-tailed distribution of charges limited the model’s ability to accurately capture extreme costs, however, it did a respectable job of reflecting overall cost trends, especially for average expenses. Although the findings show that linear regression is a good baseline model, its drawbacks show that more advanced strategies, including ensemble methods or generalized linear models, are required to improve prediction accuracy for expensive outliers. Furthermore, when incorporating sensitive characteristics like smoking and BMI into healthcare pricing models, fairness and ethical issues need to be taken into account.

**References**

Harrell, F.E. (2015) Regression Modeling Strategies. 2nd edn. Cham: Springer.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2021) An Introduction to Statistical Learning with Applications in R. 2nd edn. New York: Springer.

Kuhn, M., and Johnson, K. (2013) Applied predictive modeling. New York: Springer.

Montgomery, D.C., Peck, E.A. and Vining, G.G. (2021) Introduction to Linear Regression Analysis. 6th edn. Hoboken, NJ: Wiley.

Wooldridge, J.M. (2016) Introductory econometrics: A modern approach. 6th edn. Boston: Cengage Learning.