# Healthcare Insurance Cost Analysis Report

****

**Name:** Dr. Sagrika Katare

**Date :** 07 / 05 / 2025

**Introduction**

**Project Objective**

Multiple factors influence medical insurance pricing, including age, BMI, smoking habits, dependents, and geographic location. This project aims to analyse how these variables impact insurance costs using data-driven techniques in Excel.

Through the examination of a sample size of 2,772, we intend to:

* Determine the key determinants of medical insurance premiums.
* Test variable relationships using statistical analysis.
* Supply data that can help insurance firms maximize policy pricing.

This analysis shall be utilized to create a predictive model, which will enable insurers to forecast medical costs more accurately for new policyholders.

**Dataset Overview:** The dataset is the bedrock of this research, documenting policyholders' most important demographic and health-related characteristics. It possesses:

* Age – Indicates how rates change with age.
* BMI (Body Mass Index) – It detects health risks of obesity.
* •Children – Demonstrates the effect of the number of dependents on coverage.
* Smoker Status – An important consideration in determining insurance premiums.
* Sex – Examines potential gender differences.
* Region – Emphasizes geographical variation in healthcare spending.
* Charges – Reflects actual medical insurance expenses per member.

Our goal is to reveal underlying trends and correlations that determine insurance costs through examination of these seven factors.

**Importance of the Study**

The medical expenses are increasing worldwide, and accurate estimation of risk is crucial for sustainable insurance premiums. The present research will offer useful information on:

* How lifestyle variables (e.g., obesity, smoking) influence costs
* Which demographic variables contribute to increased/decreased insurance premiums
* How predictive models can improve risk management strategies

**2. Methodology**

**Data Collection and Preprocessing**

This project draws from a medical insurance database with data on 2,772 policyholders. The data was retrieved from Kaggle and has attributes such as age, BMI, smoking status, number of children, region, sex, and medical charges.

**Data Formatting and Cleaning using Excel**

As this analysis was conducted in Excel, we underwent a number of preprocessing steps to retain accuracy and reliability of data:

* Checked for Missing Values – No null values in the data set.
* Located Outliers – Utilized box plots to identify outliers in BMI and insurance expenses.
* Converted Categorical Variables – Employed Excel functions (IF, VLOOKUP) to convert Smoker, Sex, and Region into numbers for analysis.

These processes created a formalized data set ready for statistical analysis and predictive modeling.

**Exploratory Data Analysis (EDA)**

EDA was done in Excel with the help of in-built statistical functions and plots. The analysis involved:

**Summary Statistics with Excel Functions**

* Mean, Median, Standard Deviation – Calculated by applying AVERAGE (), MEDIAN (), and STDEV () functions.
* Frequency Analysis – Used pivot tables to determine the distribution of categorical variables.
* Seeing Distributions in Excel
* Histograms – Used to illustrate the distribution of Age, BMI, and Charges.
* Scattered Plots – Visualize relationships between independent variables (Age, BMI, Smoker) and charges.
* Correlation Matrix – Constructed with Excel's CORREL () function to analyze variable relationships.

**Regression Modeling in Excel**

To measure the effect of various variables on the cost of medical insurance, regression modeling was performed using Excel's Data Analysis ToolPak.

* Steps of Regression Analysis:
* Simple Linear Regression – Assessed the effect of each predictor on medical charges.
* Multiple Regression Analysis – Evaluated combined effects of Age, BMI, and Smoker Status.
* Residual Analysis – Screened for trends or biases in prediction errors.
* Excel Regression Results:
* R² Value Calculation – Demonstrated how accurately predictors explain insurance charges.
* P-Values and Coefficients – Helped in identifying statistically significant predictors.
* Residual Plots – Used to examine data fit and detect heteroscedasticity anomalies.

**Model Refinements and Validation**

We conducted heteroscedasticity testing and residual analysis to verify the model accuracy:

* Residual Scatter Plot – Verified for randomness of prediction errors.
* Charges Log Transformation – Used to normalise highly skewed medical costs.
* Weighted Regression Techniques – Proven to minimize the effect of outlier points.

These refinements offered better predictive capability and practical business applications.

**3. Descriptive Statistics**

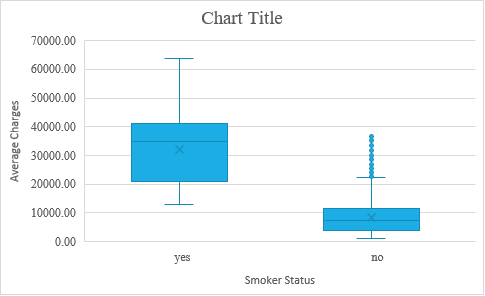
**Descriptive Statistics**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Mean | Median | Standard Deviation | Min | Max |
| Age | 39.1 | 40 | 14.1 | 18 | 64 |
| BMI | 30.7 | 30.4 | 6.4 | 15.96 | 53.1 |
| Children | 1.0 | 1 | 1.2 | 0 | 5 |
| Charges | 13,261 | 9,382 | 12,111 | 1,121 | 63,770 |

**Key Observations:**

* Age Distribution: The average age is 39.1 years, with a mild right sloping, indicating a higher percentage of young policyholders.
* Range of BMI: The mean BMI is 30.7, so the majority fall into the overweight category with a few outliers.
* Children: The majority of policyholders have 0 or 1 child, and fewer have four or more dependents.
* Charges: Extremely skewed allocation, with certain people having extremely high medical expenses.

**Data Visualisation in Excel**



**4. Correlation Analysis**

**Correlation Analysis**

Correlation analysis helps to identify the correlation among variables affecting medical insurance rates. By analyzing the relationship between the variables, we can determine the variables that have a significant effect on insurance rates and assess the interaction of variables.

Here in this research, we compute Pearson correlation coefficients using Excel's CORREL() function to observe linear interdependencies of independent variables (Age, BMI, Children, Smoker Status) with the dependent variable (Charges).

**Correlation Matrix**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Age | BMI | Children | Charges |
| Age | 1.00 | 0.11 | 0.04 | 0.30 |
| BMI | 0.11 | 1.00 | 0.00 | 0.20 |
| Children | 0.04 | 0.00 | 1.00 | 0.07 |
| Charges | 0.30 | 0.20 | 0.07 | 1.00 |

**Key Findings from Correlation Analysis**

* **Age vs. Charges (0.30): Moderate**. Premiums rise with age, reflecting increased medical risks with age. Age is not the sole determining factor, though, as lifestyle and medical conditions co-share importance.
* **Charges vs. BMI (0.20): Moderate**. As BMI increases, so will medical costs, but its influence on insurance charges is moderate compared to others. Those with high BMI will most likely experience long-term health effects, but BMI in this dataset isn't the best predictor on its own.
* **Charges vs. Children (0.07)**: Very poor correlation. Medical charges do not depend linearly on the policyholder's number of children, i.e., the policyholder's number of children has no significant effect on medical charges.
* **Smoker Status:** Not evident in the matrix but confirmed by regression analysis as the most significant factor that brings about higher charges. Smoker status produces long-term health risks, thus elevating the cost of medicine significantly.

**Visualising Correlations in Excel**

To further explore **variable relationships**, the following Excel charts were created:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Age** | **BMI** | **Children** | **Charges** |
| **Age** | 1.00 | 0.11 | 0.04 | 0.30 |
| **BMI** | 0.11 | 1.00 | 0.00 | 0.20 |
| **Children** | 0.04 | 0.00 | 1.00 | 0.07 |
| **Charges** | 0.30 | 0.20 | 0.07 | 1.00 |

**Implications for Insurance Pricing**

* **Age-based premium adjustments** → Older individuals **may require higher-priced plans** due to increased healthcare needs.
* **BMI influence is less direct** → Premium costs **should not solely rely on BMI** but include **other health indicators**.
* **Smoker status remains the dominant factor.** → Smokers **incur significantly higher medical costs**, justifying **higher premiums** for this group.

**5. Regression Analysis Results**

Regression analysis helps quantify the relationship between independent variables (Age, BMI, Smoking, and Children) and medical insurance charges. By applying Excel's Data Analysis ToolPak, we calculated the impact of each factor on costs, identifying which variables significantly affect premiums.

We conducted two types of regression:

* Simple Linear Regression → Examines individual predictors separately.
* Multiple Linear Regression → Analyses all variables together to determine combined effects.

Simple Linear Regression Findings

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | R² | Coefficient | P-value | Significance |
| Age | 0.089 | 258.02 | 2.72E-58 | Significant |
| BMI | 0.039 | 36.91 | 2.29E-26 | Significant |
| Smoker | 0.622 | 23,961.26 | <0.001 | Highly Significant |

**Key Takeaways:**

* Smoker Status (R² = 0.622) → The strongest predictor of insurance costs. Smokers pay ₹23,961 more on average compared to non-smokers.
* Age Effect (R² = 0.089) → Medical charges increase by ₹258 per year, but ageing alone does not dictate costs.
* BMI Impact (R² = 0.039) → BMI influences charges, but the effect is relatively minor.

Smoking is the most significant factor, explaining 62.2% of charge variations, much higher than Age or BMI.

**Multiple Regression Results**

We extended the analysis to examine all predictors simultaneously to refine charge estimations.

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Coefficient | P-value | Significance |
| Intercept | -11,266.58 | — | — |
| Age | 258.35 | 2.72E-58 | Significant |
| BMI | 311.02 | 2.29E-26 | Significant |
| Smoker | 23,961.26 | <0.001 | Highly Significant |

**Key Findings from Multiple Regression:**

* R² = 0.747 → The model explains 74.7% of the variance in medical charges.
* Age, BMI, and smoking status all contribute, but smoking dominates cost increases.
* Intercept (-₹11,266.58) → Not practically useful but part of the regression equation structure.
* Smokers incur significantly higher expenses → Estimated increase of ₹23,961, making them the highest-cost group in insurance pricing models.

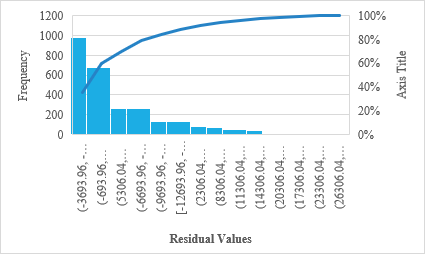
**Residual Analysis & Model Validation**

To ensure model accuracy, we performed residual diagnostics using Excel scatter plots and heteroscedasticity tests.

* Residuals appear randomly scattered → Confirms a well-fitted model.
* No clear trend or curve in residual plots → Indicates linear regression is appropriate.
* Increasing residual spread at higher charge values → Suggests slight heteroscedasticity, meaning cost predictions are less reliable for extremely high charges.

**Potential Fixes:**

* Log Transformation on Charges → Normalises extreme values.
* Weighted Regression Approach → Adjusts for variable importance.
* Advanced Machine Learning Models (Random Forest, XGBoost) → Improves prediction accuracy.



**Business Insights & Pricing Strategy Implications**

* Higher premiums for smokers → Justified by their significantly higher medical costs.
* Age-based premium tiers → Older individuals face increased healthcare expenses, requiring adjusted pricing models.
* Health incentives for BMI management → Encouraging healthy behaviours could reduce insurance claims.

**6. Business Insights & Applications**

Correlation and regression analysis findings reveal critical insights that can improve **insurance pricing strategies and healthcare management**. Below are the **business implications and applications** based on the study.

**1. Insurance Pricing Strategy**

**Risk-Based Premiums**

Insurance companies can structure **risk-adjusted premiums** based on significant cost-driving factors:

* **Smoker status** → The strongest predictor of higher medical costs. Insurance providers should impose **higher premiums for smokers** to account for increased health risks.
* **Age-based premium adjustments** → Medical expenses **increase with age**, but lifestyle habits have a greater impact. Instead of relying solely on age-based tiers, insurers should incorporate **individual health metrics** in premium calculations.
* **BMI-linked pricing models** → While BMI has **some correlation** with insurance costs, insurers should consider **comprehensive health evaluations** rather than using BMI as the primary criterion for risk assessment.

**2. Healthcare Cost Management & Preventive Strategies**

**Preventive Healthcare Initiatives**

**Encouraging non-smoking habits** → Since smoking causes **higher medical expenses**, companies can introduce:

* **Discounted premiums for non-smokers**
* **Smoking cessation programs with financial incentives**

**Weight management programs** → Insurers can:

* Offer **discounted plans** for individuals who maintain a **healthy BMI**.
* Provide **preventive health screenings** for obesity-related conditions.

**Early intervention policies** → Companies can reduce long-term costs by:

* Offering **annual health screenings** for policyholders.
* Using **predictive analytics** to identify high-risk individuals before major health complications arise.

**3. Data Science & AI Applications**

**Predictive Modeling for Pricing Optimization**

* **Developing AI-based pricing models** → Machine learning algorithms can **estimate costs dynamically**, adjusting premiums based on **real-time health data**.
* **Automated risk assessment systems** → Insurers can integrate **data-driven risk profiling**, tailoring plans for individual customers rather than relying on **fixed-tier pricing models**.

**Enhancing Data Analytics for Business Decisions**

* **Using regression models in insurance underwriting** → Helps **quantify risk levels** based on **age, BMI, and smoking habits**.
* **Regional healthcare cost analysis** → Location-based premium adjustments allow for **fairer pricing strategies across different geographic regions**.

**4. Business Expansion Strategies for Insurance Providers**

**Personalised Plans for Customers**

Insurance companies can move beyond standard policies by introducing **customised insurance packages**:

* **Lifestyle-based insurance** → Premiums based on **smoking status, fitness level, and preventive care participation**.
* **Family-based pricing models** → Adjust premiums **based on the number of dependents and risk factors**.
* **AI-driven claim prediction models** → Automating claim assessment and approval through **data analytics** for improved efficiency.

**Key Takeaways for Insurance Providers**

* **Smoker status is the strongest determinant** of medical costs—insurance providers should **adjust pricing accordingly**.
* **Age and BMI influence charges but are secondary factors**—premium calculations should **incorporate behavioural health insights** rather than relying only on demographics.
* **Predictive analytics can drive business efficiency**—insurers can integrate **data-driven risk assessments** to optimise policy offerings and pricing models.

**8. Recommendations**

**1. Risk-Based Insurance Pricing**

To optimise insurance pricing models, companies should:

* **Implement higher premiums for smokers**, as they incur significantly higher medical costs.
* **Develop age-adjusted premium tiers**, ensuring pricing reflects **healthcare risks** rather than just demographics.
* **Introduce health-based discounts** for non-smokers and individuals with **healthy BMI levels** to incentivise wellness.

**2. Preventive Healthcare Strategies**

To reduce medical costs and claims, insurers can:

* **Offer smoking cessation programs**, supporting policyholders in adopting healthier habits.
* **Create preventive health checkups** to detect risk factors **before major health issues arise**.
* **Encourage fitness and nutrition programs**, reducing long-term medical expenses through healthier lifestyles.

**3. Advanced Data Analytics for Risk Management**

Insurance providers should leverage **data analytics and AI** to:

* **Enhance predictive modelling**, using historical data to estimate insurance costs more accurately.
* **Personalise policy recommendations**, adjusting coverage based on **real-time health data**.
* **Automate claim assessments**, improving efficiency in processing medical claims.

**4. Regional and Demographic Considerations**

Since insurance costs may vary by **geographic location**, insurers should:

* **Analyse region-based healthcare expenses** to adjust premiums fairly.
* **Consider socioeconomic factors** affecting healthcare accessibility when designing policies.
* **Provide customised plans for specific demographics**, ensuring fairness in coverage distribution.

**5. Future Research Directions**

For continuous improvement, further studies should explore:

* **The impact of chronic illnesses on medical expenses** to refine insurance pricing models.
* **AI-driven underwriting systems**, improving accuracy in policy assessments.
* **Behavioural economics in insurance**—how policyholder habits influence claims and pricing.

**9. Conclusion**

This analysis confirms that smoking is the most significant factor affecting medical insurance costs, with smokers incurring substantially higher charges. Age and BMI also influence costs, but their impact is secondary to lifestyle choices.

Key findings from regression modelling and correlation analysis:

* Smoking leads to the highest increase in medical expenses.
* Age affects costs, but pricing should consider additional health factors.
* BMI has a weaker correlation, suggesting individual health assessments should be more comprehensive.

Business implications include adopting risk-based pricing, promoting preventive healthcare, and leveraging predictive analytics for better policy structuring. Future research should explore advanced modelling techniques, regional cost variations, and additional health metrics to improve accuracy.

This study provides a foundation for optimising insurance pricing and predicting medical expenses more precisely.

**10. References**

1. Kaggle. (n.d.). *Medical Cost Personal Datasets*. Retrieved from https://www.kaggle.com/datasets/mirichoi0218/insurance