# **Insurance Claim Analysis: Effect of Lifestyle on Medical Costs**

1. Introduction

Health insurance pricing depends heavily on how well insurers understand risk. Different people pose different levels of risk depending on lifestyle and health choices. The aim of this project was to explore whether smoking and body mass index (BMI) have a measurable impact on insurance claim amounts, using a dataset of 1,338 individuals.

This analysis is meant to replicate the type of early-stage exploratory work actuaries or data analysts do before building more formal risk models. The main question driving the project was:  
**Do smokers incur higher claim costs, and how does BMI affect this relationship?**

1. Data Description

The dataset used is publicly available on Kaggle and contains the following information for each individual:

* **Age**
* **Sex**
* **BMI (Body Mass Index)**
* **Children (number of dependents)**
* **Smoker status (yes/no)**
* **Region**
* **Charges (medical claim amount in USD)**

For the purposes of this project, attention was mainly given to **BMI, smoker status, and charges**, since these are most directly linked to lifestyle and health-related risk.

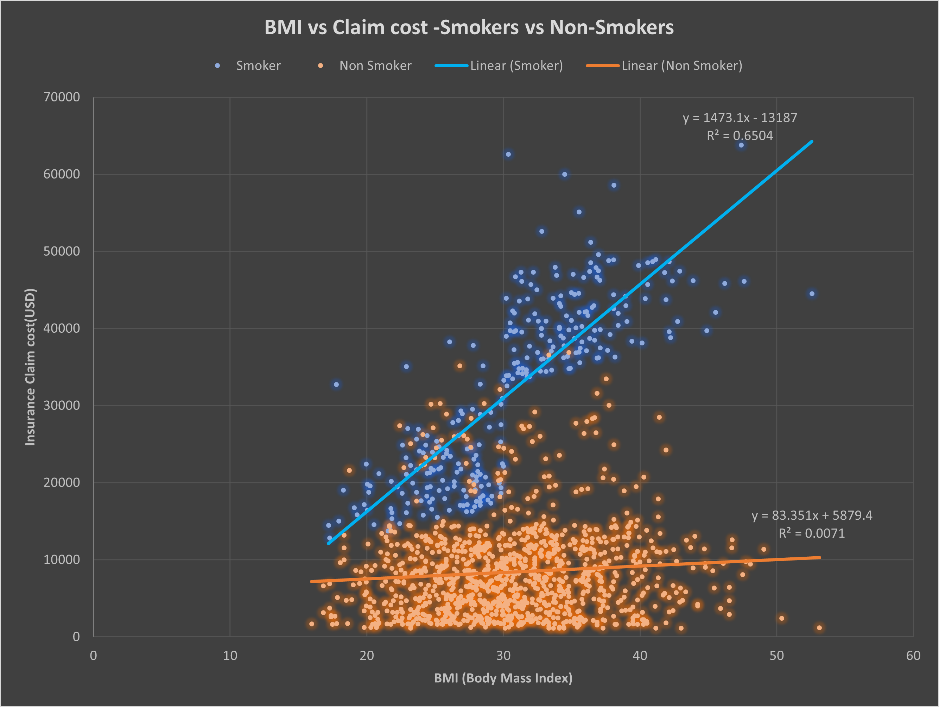
1. Methodology

The work was carried out in Excel using the following steps:

1. **Pivot Tables** were used to compare average claim costs between smokers and non-smokers, and later between different age groups.
2. **Scatter Plots** were created to visualize the relationship between BMI (x-axis) and charges (y-axis). Two series were plotted separately for smokers and non-smokers.
3. **Trendlines** were added to each series to compare the slopes, showing how costs rise with BMI for each group.
4. **R² values and equations** were taken from the trendlines to understand how strongly BMI explains claim costs.
5. **Average Costs** were calculated using Excel formulas to quantify the differences between smokers and non-smokers.
6. Findings
7. **Smokers pay substantially higher claims**
   * On average, smokers’ claims were roughly three times higher than those of non-smokers.
   * In the scatter plot, smokers clustered at the top of the cost distribution, often with claims above $30,000, while most non-smokers remained below $15,000.
   * As per Fig. 1 it was also calculated that average charges for smokers are 3 times more than those of non-smokers.

(Charges actually represent the **medical claim cost** (the amount billed/paid for healthcare services), **not the premium**)  
 (**Fig.1)**

|  |  |
| --- | --- |
| **Row Labels** | **Average of charges** |
| no | 8,434.268 |
| yes | 32,050.232 |
| **Grand Total** | **13270.42227** |

1. **BMI has a stronger effect on smokers**
   * The smoker trendline in Fig. 1 showed a slope of **about 1,200**, compared to only **250** for non-smokers. This means that for each 1-point increase in BMI, smokers’ costs rise nearly five times faster.
   * The R² value for smokers was around **0.40**, while for non-smokers it was closer to **0.10** as depicted in Fig. 2. In plain terms, this means BMI is a much stronger predictor of claim costs for smokers than it is for non-smokers. For non-smokers, other factors (such as age or family size) appear to matter more.  
     **(Figure2)**-   
     
2. **Two distinct cost clusters appear**
   * A **low to moderate cost cluster**, primarily non-smokers.
   * A **high-cost cluster**, dominated by smokers.

This separation shows that lifestyle choices are a critical driver of medical expenses, not just demographics like age or gender.

1. Conclusion

The project highlights how basic Excel tools can be used to uncover important business insights from raw data. The results clearly show that smoking has a major impact on claim costs, and that BMI further amplifies this effect among smokers.

The inclusion of slope and R² values also depicted Figure 2 from the trendlines adds statistical weight to these findings. A steeper slope for smokers confirms that their costs rise faster with BMI, while the higher R² suggests that BMI alone explains a meaningful part of their expenses. For non-smokers, the weaker R² indicates their claim amounts are influenced by a wider mix of variables.

For insurers, the implication is straightforward: lifestyle-related variables such as smoking status should be weighted heavily when determining premiums. For an aspiring actuary or analyst, this project demonstrates an ability to clean data, apply Excel-based analysis, and communicate results in both numerical and visual form.

1. Referencing
2. <https://www.kaggle.com/datasets/mirichoi0218/insurance?resource=download>