Predicting Insurance Costs & Risk Factors Using Data Analytics

Context:
This capstone project explores how demographic and health factors impact medical insurance costs. By leveraging data analytics and machine learning, the project identifies key cost drivers, predicts high-risk individuals, and provides actionable insights for insurance companies, healthcare professionals, and individuals.

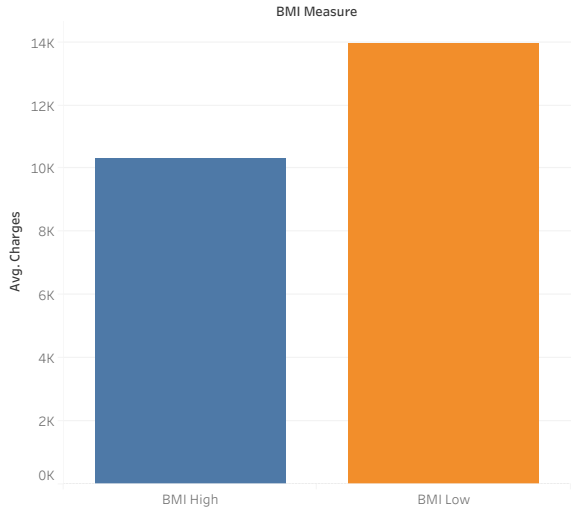
Objectives:
- Analyze key factors affecting insurance costs (Age, BMI, Smoking, Number of Children).
- Predict high-risk individuals using machine learning classification.
- Provide insights on cost variations by demographics and geographic regions.

Key Analytical Approaches:
- Descriptive Analysis: Understanding cost distribution and major influencing factors.
- Inferential Analysis: Identifying statistically significant relationships (e.g., smoker vs. non-smoker costs).
- Predictive Modeling: Using Random Forest Classifier to classify individuals into high risk vs. low risk.
- Tableau Data Visualization: Creating interactive dashboards for insights.

Data Source: https://www.kaggle.com/datasets/willianoliveiragibin/healthcare-insurance
By Ronak Singh

Correlation Analysis

Factors	age	bmi	sex	smoker	charges
age	1.000	0.109	-0.021	-0.025	0.299
bmi	0.109	1.000	0.046	0.004	0.198
sex	-0.021	0.046	1.000	0.076	0.057
smoker	-0.025	0.004	0.076	1.000	0.787
charges	0.299	0.198	0.057	0.787	1.000



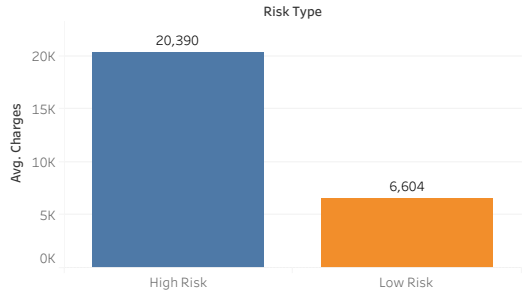
- **Smoking:** The correlation analysis shows that smoking and insurance costs have a direct relationship. 0.787 shows a robust positive correlation. Therefore, people who smoke tend to pay higher charges.

- **Age:** Moderate to strong positive relationship (0.299). Therefore, as age increases, so does insurance cost, but not as much as smoking.

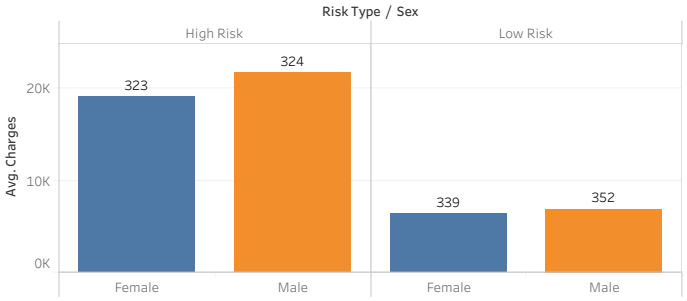
- BMI: Moderate positive relationship with insurance charges. However, this analysis will show that the population with lower BMI tends to have a higher average cost.

- **Sex and Region:** Tend to have a near zero to no correlation with charges.

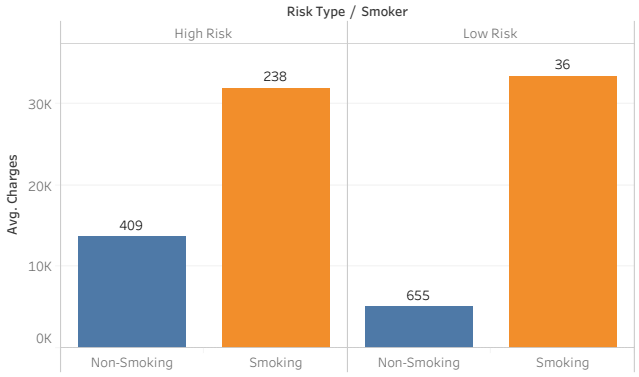
Risk Analysis off All Factors



- We see that High-risk populations, on average, pay 33% more than Low-Risk populations.

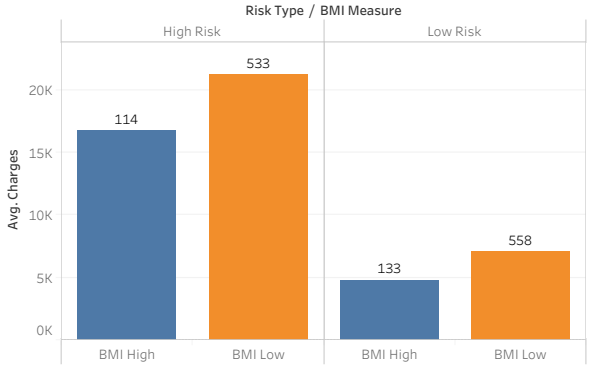


- For High-Risk populations, there are 324 Males and 323 Females, with Males paying an average charge of $21,693$, while Females pay $19,083$. For Low-Risk populations, there are 352 men and 339 women. Males pay $6,836 and Females pay $6,394



- We see that High-Risk populations have 238 smokers whilst having 409 nonsmokers, with smokers paying 43% more ($31,863) on average.

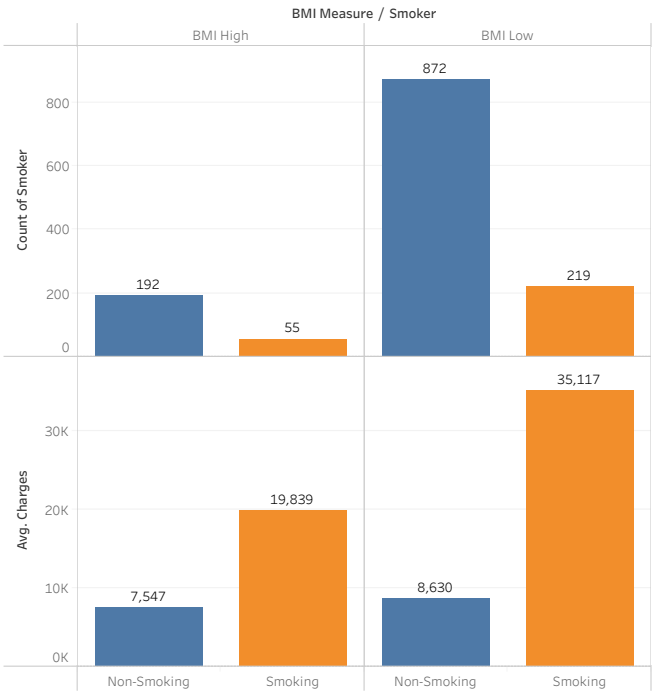
- For Low-Risk populations, there are 655 non-smokers and 36 smokers. Furthermore, smokers pay $33,289, whilst non-smokers pay $5,938.

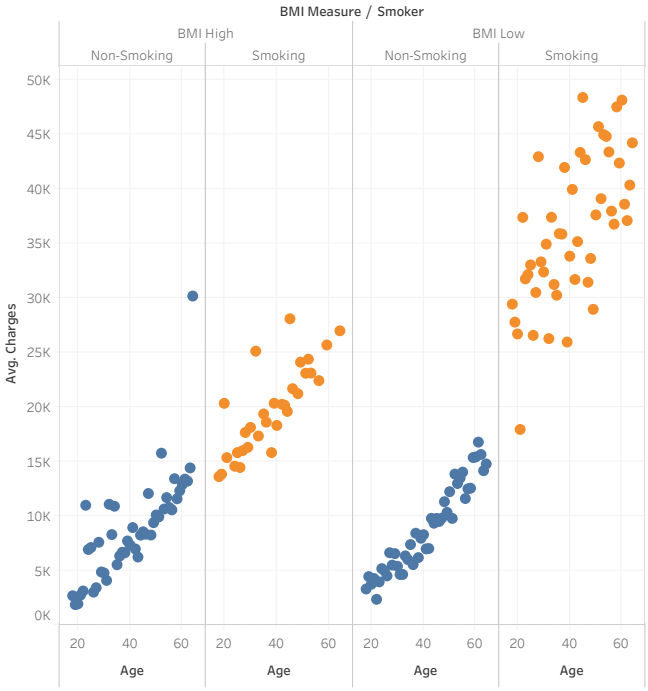


- Low BMI populations are charged more in both High-Risk and Low-Risk populations. There are 533 low-BMI people ($21,172$) and 114 high-BMI people ($16,733 average charge) who are High-Risk.

- For the low-risk population, there are 558 people with Low-BMI ($7,758 average charge) and 133 with High-BMI ($4,757 average charge).

Smoking and BMI





- Here, we see the relationship between smoking and BMI. We can observe that even though the count of High and Low BMI populations can be disproportional in the count of the population, the charge, on average, for low-BMI individuals tends to be higher.

- This seems counterintuitive to our understanding of high-BMI individuals. However, when understanding the robust direct correlation to smoking on charges as well as the fact that smoking can be a factor in weight loss, the figures above have more accuracy.

-We also see this in the exponential price increase in Low-BMI smokers vs Non-Smokers. The average charges for Non-smoking populations tend to remain the same no matter the BMI. This follows from our moderate correlation seen between BMI and Charges in the correlation analysis.

Conclusion

The analysis highlights several key factors influencing insurance charges, with smoking, BMI, and risk classification playing significant roles in cost variation.

- **Smoking:** There is a strong positive correlation (0.787) between smoking and insurance charges. Smokers consistently pay significantly higher insurance costs than non-smokers, with high-risk smokers paying 43% more than their non-smoking counterparts.
- **Age**: A moderate to strong correlation (0.299) shows that as age increases, so do insurance charges, though not as drastically as smoking.
- **BMI:** While there is a moderate positive relationship between BMI and charges, the analysis reveals that low-BMI individuals tend to pay more on average across both high-risk and low-risk populations. This trend can be partially explained by the fact that smoking, which is linked to weight loss, also has a significant impact on increasing insurance costs.
-Sex & Region: These variables have little to no impact on insurance charges.

Risk Classification:
- High-risk populations pay 33% more on average than low-risk populations.

- Men consistently pay more than women across both risk groups, with high-risk males paying an average of $21,693, compared to $19,083 for females.

- Smokers pay substantially higher premiums, with low-risk smokers paying over 5 times more ($33,289) than non-smokers ($5,938).

- Smokers in the high-risk category pay the highest premiums, with an average charge of $31,863, while non-smokers in the same category pay significantly less ($22,285).

- Low-BMI individuals face higher average charges than high-BMI individuals across both risk groups, with high-risk low-BMI individuals paying $21,172, compared to $16,733 for high-risk high-BMI individuals. Similarly, in the low-risk category, low-BMI individuals pay $7,758, while high-BMI individuals pay $4,757.

- This pattern appears counterintuitive, as high BMI is typically linked to greater health risks. However, given the strong correlation between smoking and increased insurance costs, and the fact that smoking can contribute to lower BMI, these findings align with the overall trend of smoking-related price increases.

Policy Recommendations for Risk-Based Pricing & Management

- **Stronger Risk-Based Premium Adjustments**
  - Enhance risk classification models by emphasizing smoking status as a key pricing factor.
  - Introduce a graduated pricing system where smokers who quit over time see progressive premium reductions to incentivize behavioral change.
  - Adjust BMI-based risk models to avoid overcharging low-BMI individuals unless additional health risk factors justify higher costs.

- **Personalized Health & Wellness Programs for High-Risk Populations**
  - Given the strong correlation between smoking and higher insurance costs, insurance companies should expand smoking cessation programs with financial incentives.
  - Provide targeted interventions for low-BMI, high-risk individuals to help mitigate unexpected cost increases.

- **Gender-Based Risk Assessment Adjustments**
  - Since men pay consistently higher premiums, insurers should assess whether these cost differences reflect actual risk disparities or outdated pricing models.
  - Consider introducing unisex pricing tiers based on specific risk factors rather than blanket gender-based cost differences.

- **Fraud Prevention & Data Transparency**
  - Given the counterintuitive pricing trends for low-BMI individuals, insurers should audit their risk classification models to ensure that pricing is aligned with actual health risks rather than indirect correlations.
  - Improve data transparency by providing more apparent breakdowns of risk classification decisions, ensuring consumers understand the factors impacting their insurance costs.

- **More Granular Risk Segmentation to Improve Pricing Fairness**
  - Rather than broad risk categories, insurers should implement multi-tiered risk segmentation that incorporates Lifestyle factors (e.g., physical activity, diet, alcohol consumption), medical history, and preventive care participation

- **Longitudinal health improvements over time**
  - This approach would ensure that high-risk individuals who actively improve their health are rewarded with lower premiums rather than locked into high-cost plans indefinitely.