Analyzing Medical Insurance Cost Trends

Samuel Diaz, Osman Salehi, Bryan Sabangan, Jahaziel Sanchez

BA/MIS 649-01: Business Analytics

San Diego State University

October 28, 2024

Contributions

Samuel Diaz:Part 1, Part 2, Layout and Formatting

Osman Salehi: Part 3, Part 4

Bryan Sabangan: Part 1, Part 2, Part 3

Jahaziel Sanchez: Part 4

Table of Contents

Part I: Introduction

Background of Research Problem

Medical insurance charges are influenced by demographic, behavioral, and health-related factors. Understanding these relationships is essential for insurers to assess risk, set premiums, and design interventions, while policymakers can use this knowledge to allocate resources and improve health outcomes. This study examines these factors to predict medical costs and classify individuals into risk levels, enabling more effective cost management.

The dataset includes demographic variables (age, sex, region), lifestyle indicators (BMI, smoking status), and household information (number of children). These variables allow for an analysis of key drivers of medical costs. For regression, the focus is on identifying significant predictors of charges, such as age, BMI, smoking status, and regional differences. For classification, individuals are categorized into low, medium, or high risk levels based on these factors.

By exploring the relationships between medical costs and individual characteristics, this study offers actionable insights for insurers to refine premium structures and for policymakers to develop targeted health interventions. The findings aim to support better risk assessment, resource allocation, and cost management in the insurance and healthcare sectors.

Regression Problem

- **Research Question**: What factors most significantly influence the medical insurance charges for individuals?
- **Dependent Variable**: Charges (medical insurance charges).
- **Independent Variable(s):**
    - **age**: Could impact both charges and smoking habits.
    - **bmi**: A proxy for health and lifestyle, likely correlated with charges and smoking.
    - **children**: The number of dependents might influence charges.
    - **sex**: Gender differences may impact charges and health behavior.
    - **region**: Could reflect socioeconomic and cultural influences.
    - **smoker**: Smoking status
- **Relevance in Practice**:
    - **For Insurance Companies**: Predicting medical charges enables insurers to better assess risk and design premium structures.

Classification Problem

- **Research Question**: How can we predict an individual's medical cost risk level based on their demographic, lifestyle, and geographic factors?
- **Dependent Variable:**
    - **Risk Level:** a categorical variable with three classes derived from the **charges** column in the dataset.:
        - **Low Risk**: Charges < $5,000
        - **Medium Risk**: Charges between $5,000 and $15,000
        - **High Risk**: Charges > $15,000.
- **Independent Variables(s):**
    - **Demographics**: Age, Sex
    - **Lifestyle**: BMI (body mass index), Smoker status
    - **Household Information**: Number of children
    - **Geographic Information**: Region
- **Relevance in practice:**
    - **For Insurance Companies:** Understanding risk levels helps insurers set appropriate premiums, identify high-risk individuals, and design targeted intervention strategies to reduce costs.
    - **For Healthcare Policy Makers:** Classifying individuals by risk level allows for resource allocation, preventive healthcare programs, and personalized recommendations to improve health outcomes and reduce healthcare spending.

Hypothesis

- **Regression Hypothesis**: Individuals who smoke or have a higher BMI are likely to have higher medical charges due to the compounded health risks and associated medical costs.
- **Classification Hypothesis**: Demographic factors (age, gender, and region) and lifestyle factors (BMI and smoking status) significantly predict an individual’s medical cost risk level (low, medium, or high).

Part II: Data Preparation and EDA

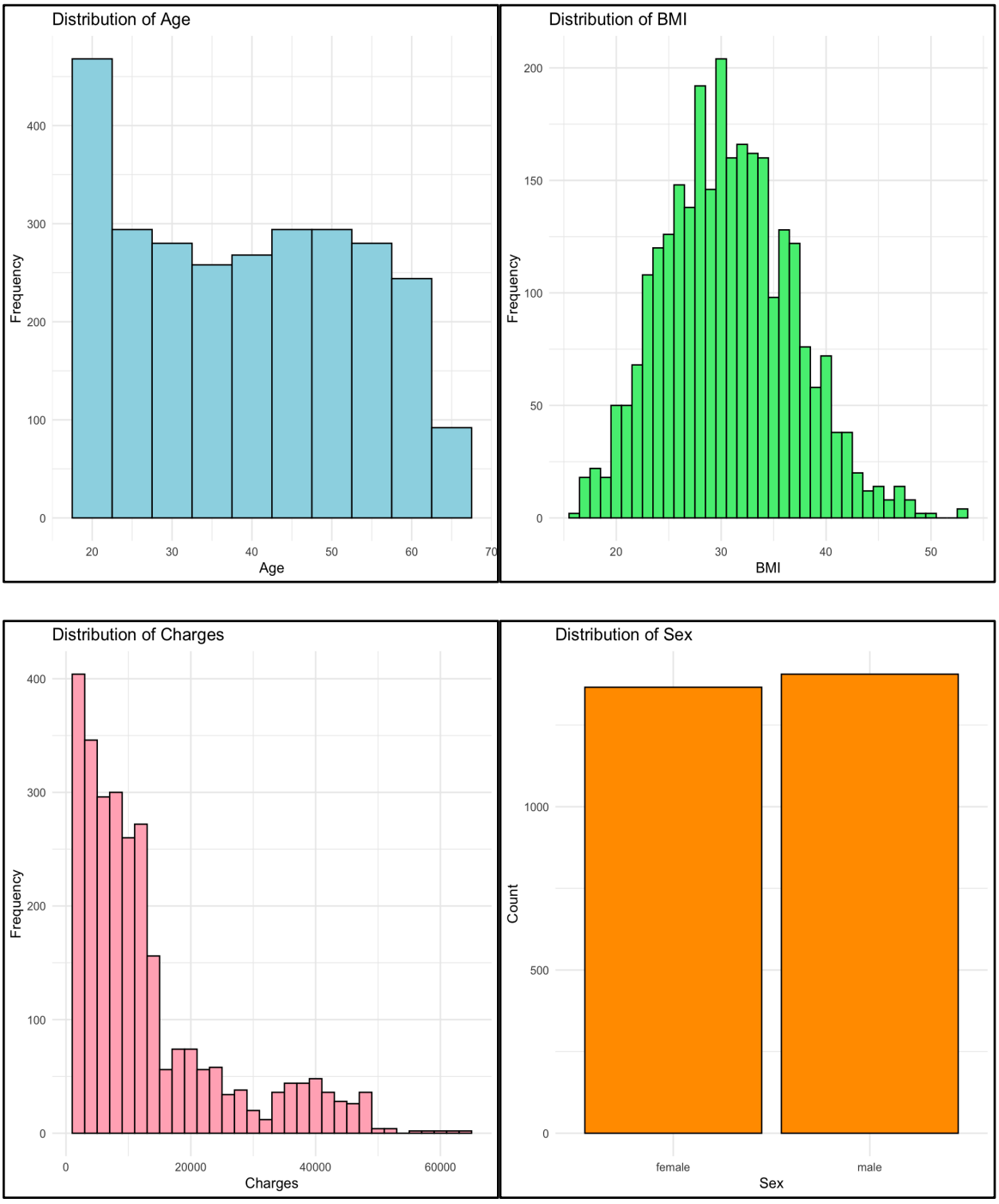
Data Summary

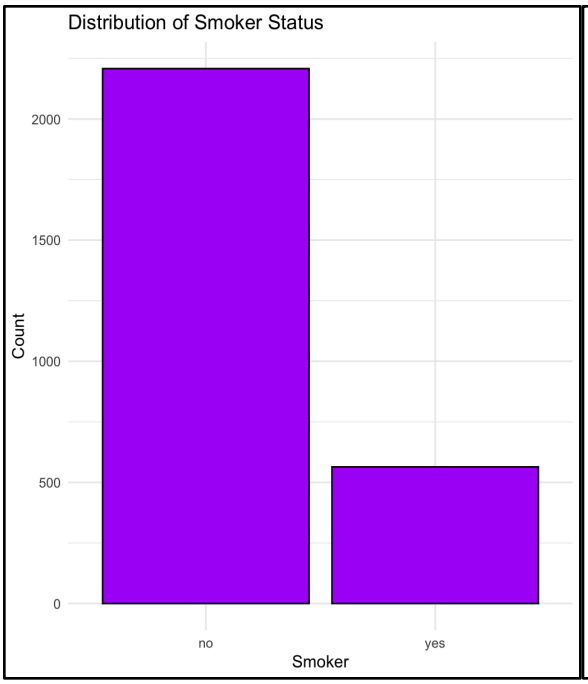
- **Source**: https://www.kaggle.com/datasets/rahulvyasm/medical-insurance-cost-prediction?resource=download
- **Observations**: There are 2,772 observations in the dataset. Each observation provides data for one unique individual, summarizing their demographic, health-related, and insurance cost details.
- **Variables**:

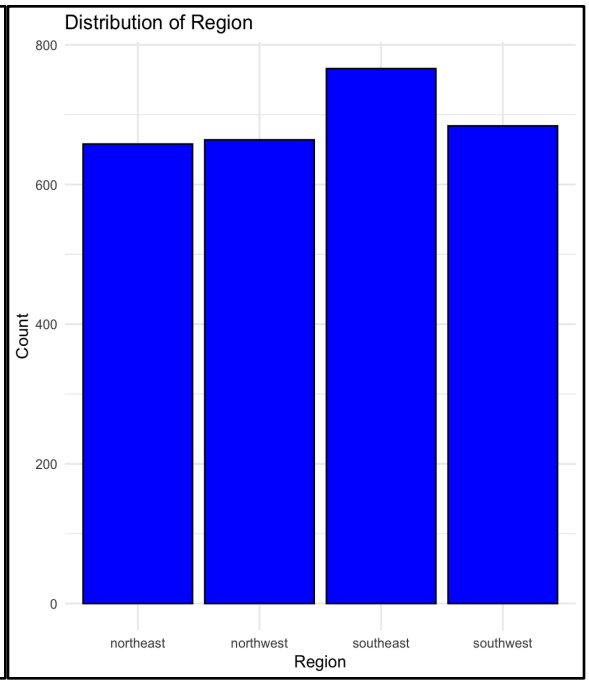
Variable	Description	Type
Age	The age of the individual in years.	Quantitative (Discrete)
Sex	The gender of the individual.	Qualitative (Nominal)
BMI	The Body Mass Index of the individual, a measure of body fat based on height and weight.	Quantitative (Continuous)
Children	The number of dependent children covered under the insurance.	Quantitative (Discrete)
Smoker	Whether the individual is a smoker (“yes” or “no”).	Qualitative (Nominal)
Region	The geographic region where the individual resides (e.g., “southwest,” “southeast,” etc.).	Qualitative (Nominal)
Charges	The medical insurance charges billed to the individual.	Quantitative (Continuous)

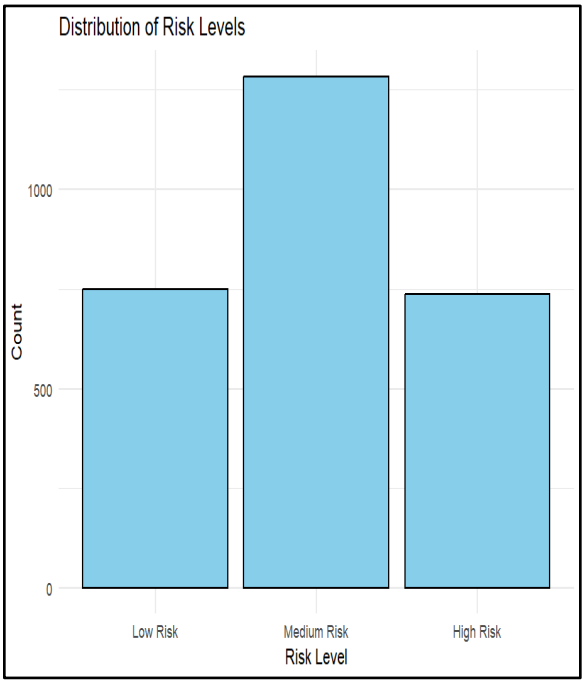
Univariate Analysis - Exploratory Data Analysis (EDA)

Outputs









Analysis for the Distribution of Data

- Numerical Variables
  - **Age**: Fairly uniform distribution, with no extreme outliers.
  - **BMI**: Slight right skew, with potential outliers on the higher end.
  - **Charges**: Strongly right-skewed, indicating a few individuals have significantly high medical costs.

- Categorical Variables:
    - **Sex**: Roughly balanced, with slightly more males.
    - **Smoker**: Majority of individuals are non-smokers.
    - **Region**: Southeast is the most frequent region.
    - **Risk Level:** Medium risk is the prominent category.

Missing Values

- None

Extreme Outliers

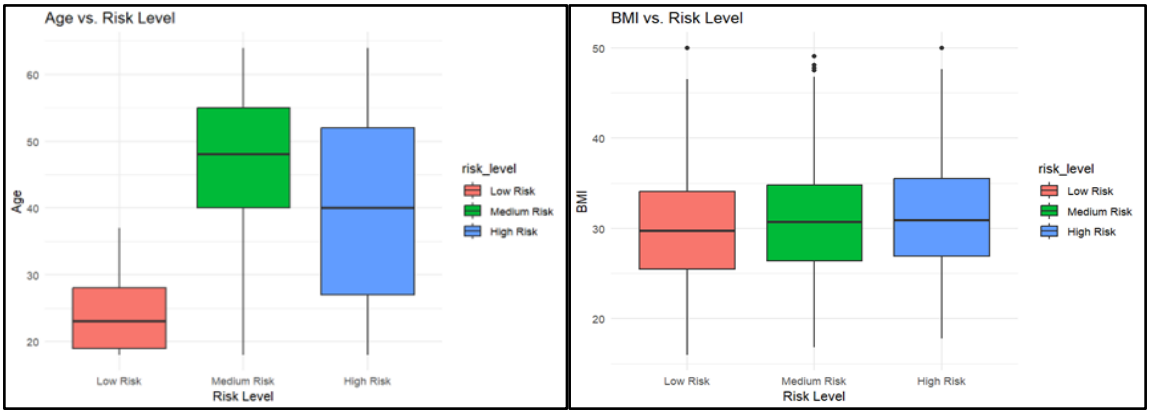
- **BMI**: Potential outliers above ~50.
- **Charges**: Significant outliers above 50,000.

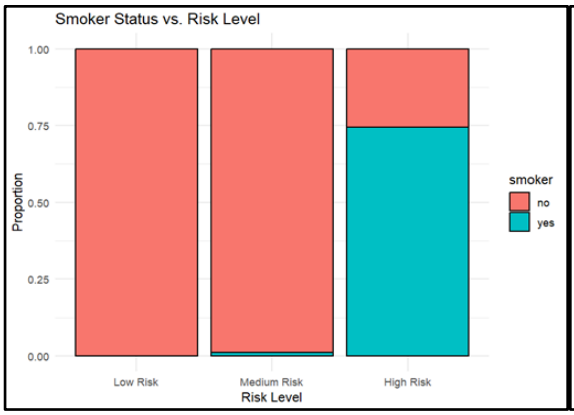
Bivariate Analysis - Exploratory Data Analysis (EDA)

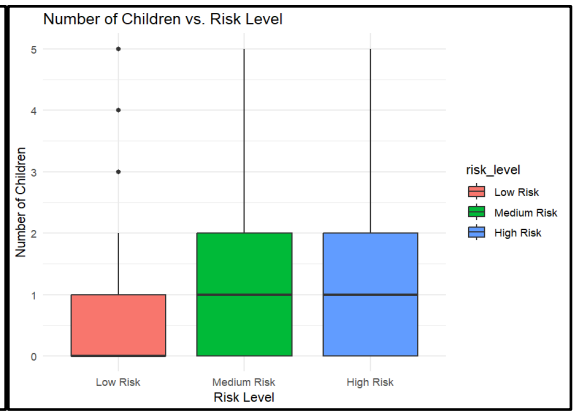
Outputs

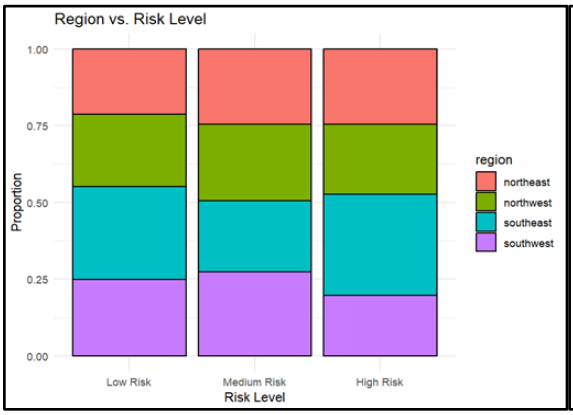
- **Correlation Matrix Formatted (rounded to two decimal places)**

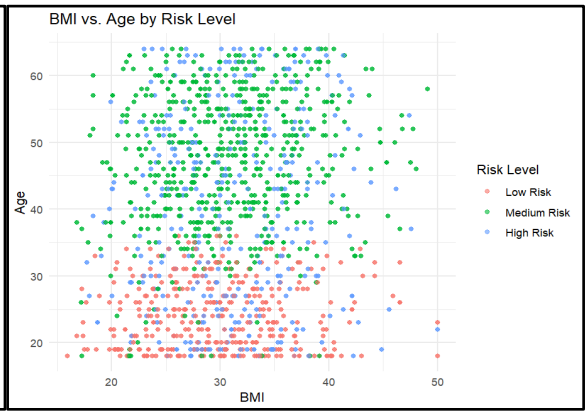
	Age	BMI	Children	Risk Level
Age	1.00	0.11	0.04	0.41
BMI	0.11	1.00	0.00	0.07
Children	0.04	0.00	1.00	0.07
Risk level	0.41	0.07	0.15	1.00

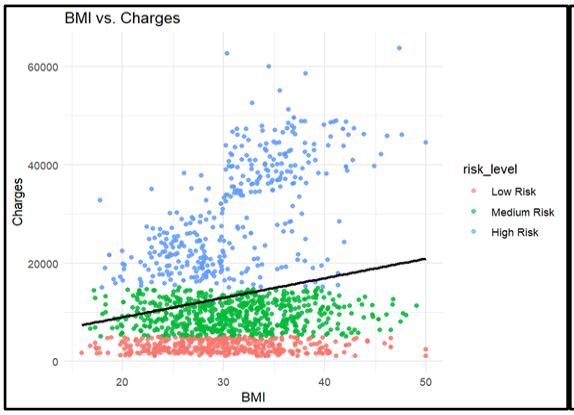


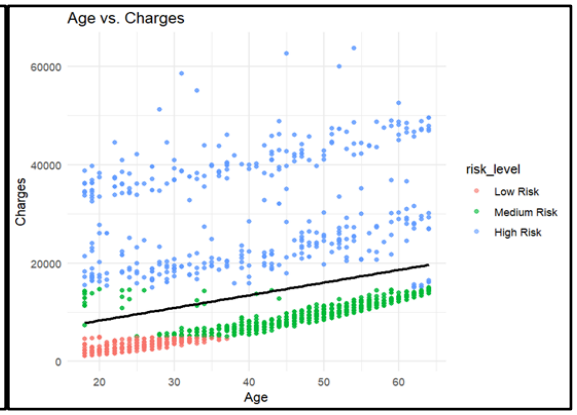

Smoker Status vs. Risk Level
Proportion
1.00
0.75
0.50
0.25
0.00
Low Risk
Medium Risk
High Risk
Risk Level
smoker
no
yes

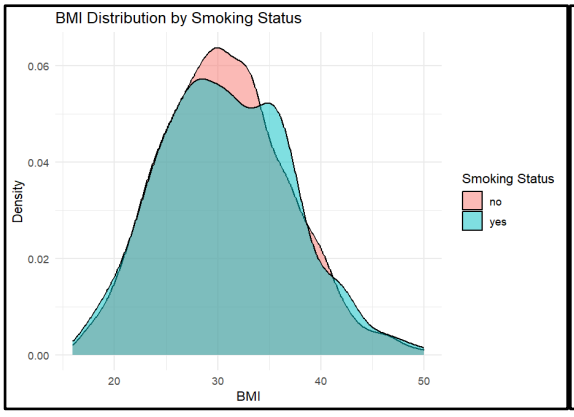

Number of Children vs. Risk Level
Number of Children
5
4
3
2
1
0
Low Risk
Medium Risk
High Risk
Risk Level
risk_level
Low Risk
Medium Risk
High Risk

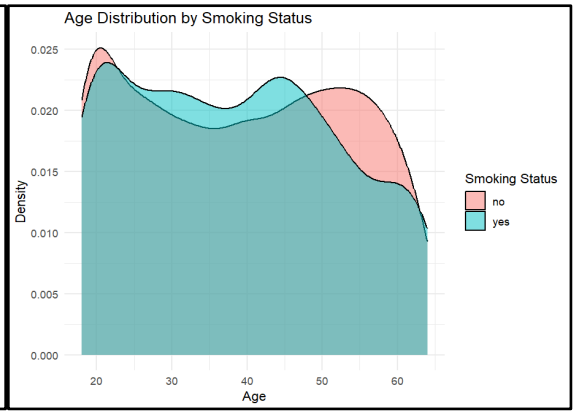

Region vs. Risk Level
Proportion
1.00
0.75
0.50
0.25
0.00
Low Risk
Medium Risk
High Risk
Risk Level
region
northeast
northwest
southeast
southwest


BMI vs. Age by Risk Level
Age
60
50
40
30
20
20
30
40
50
BMI
Risk Level
Low Risk
Medium Risk
High Risk


BMI vs. Charges
Charges
60000
40000
20000
0
20
30
40
50
BMI
risk_level
Low Risk
Medium Risk
High Risk


Age vs. Charges
Charges
60000
40000
20000
0
20
30
40
50
60
Age
risk_level
Low Risk
Medium Risk
High Risk


BMI Distribution by Smoking Status
Density
0.06
0.04
0.02
0.00
20
30
40
50
BMI
Smoking Status
no
yes


Age Distribution by Smoking Status
Density
0.025
0.020
0.015
0.010
0.005
0.000
20
30
40
50
60
Age
Smoking Status
no
yes

Analysis for the Distribution of Data

- **BoxPlots**
    - **Age vs. Risk Level:** Older individuals are more likely to fall into the High Risk category. Younger individuals dominate the Low Risk group
    - **BMI vs Risk Level:** BMI has a weak relationship with risk level.
    - **Children vs Risk Level:** The number of children shows a weak association with risk level, as the presence of children does not strongly influence charges.
- **Stacked Bar Graphs**
    - **Smoker vs. Risk Level**: Smokers are heavily represented in High Risk, indicating smoking as a critical predictor.
    - **Region vs. Risk Level:** The Southeast region shows a higher proportion of individuals in the High Risk category. Regional disparities in risk levels may highlight socioeconomic or healthcare access differences.
- **Scatterplots**
    - **Age vs. Charges:** A slight upward trend in charges with increasing age. Older individuals might have higher medical costs, but the relationship is not very strong or linear.
    - **BMI vs. Charges:** Charges tend to increase as BMI rises, especially above 30 (indicative of obesity). Obesity could be a major factor in driving up medical costs.
- **Density Plots**
    - **BMI Distribution by Smoking Status:** Smokers tend to have slightly higher BMI values on average compared to non-smokers. BMI might help in classifying smoking status, though the distinction is not stark.
    - **Age Distribution by Smoking Status:** Both smokers and non-smokers have similar age distributions. Age may not strongly distinguish smokers from non-smokers.

Missing Values

- None

Extreme Outliers

- **BMI**: There are extreme BMI values above 50, which are unusually high. These might correspond to a small number of severely obese individuals and could significantly influence the model. They should be reviewed for data integrity or handled (e.g., capped or transformed) during modeling.
- **Charges**: Extreme medical charges above 50,000. These cases likely represent individuals with severe medical conditions or high-risk profiles. These high charges could heavily skew the regression model.

Data Cleaning and Preparation Based of EDA

1. **Missing Values**: No missing values were detected in the dataset, so no action is needed.
2. **Derived Variables**: Created Risk Level categorical variable based on charges.
    - Low Risk: Charges < $5,000
    - Medium Risk: Charges between $5,000 and $15,000
    - High Risk: Charges > $15,000
3. **Handling Outliers**: Extreme outliers mitigated as deemed necessary
    - **BMI**: Winsorized extreme BMI values above 50 to mitigate the influence of extreme values.
    - **Charges**: Applied a log transformation to mitigate the influence of extreme values (log_charge).
4. **Interaction Terms**: Added interaction terms to capture the joint effect of certain predictors (i.e., bmi_smoker and age_smoker).
5. **Data Transformation**: Normalized age (age_norm) and bmi (bmi_norm) to ensure all variables are on a comparable scale, which is particularly useful for KNN.

Part III: Analysis and Findings

Regression using Generalized Regression Models

Regression List of Predictor Terms

- **Original Predictors**: age, bmi, children, sex, region, smoker
- **Transformed Predictors:** log_charges (target), age_norm, bmi_norm.
- **Interaction Terms:** bmi_smoker (BMI and smoker status) age_smoker (Age and smoker status).

Regularization and Variable Selection

1. Subset Selection based on Adjusted R^2
    - **Selected Predictor Terms and Coefficient Estimate**s:

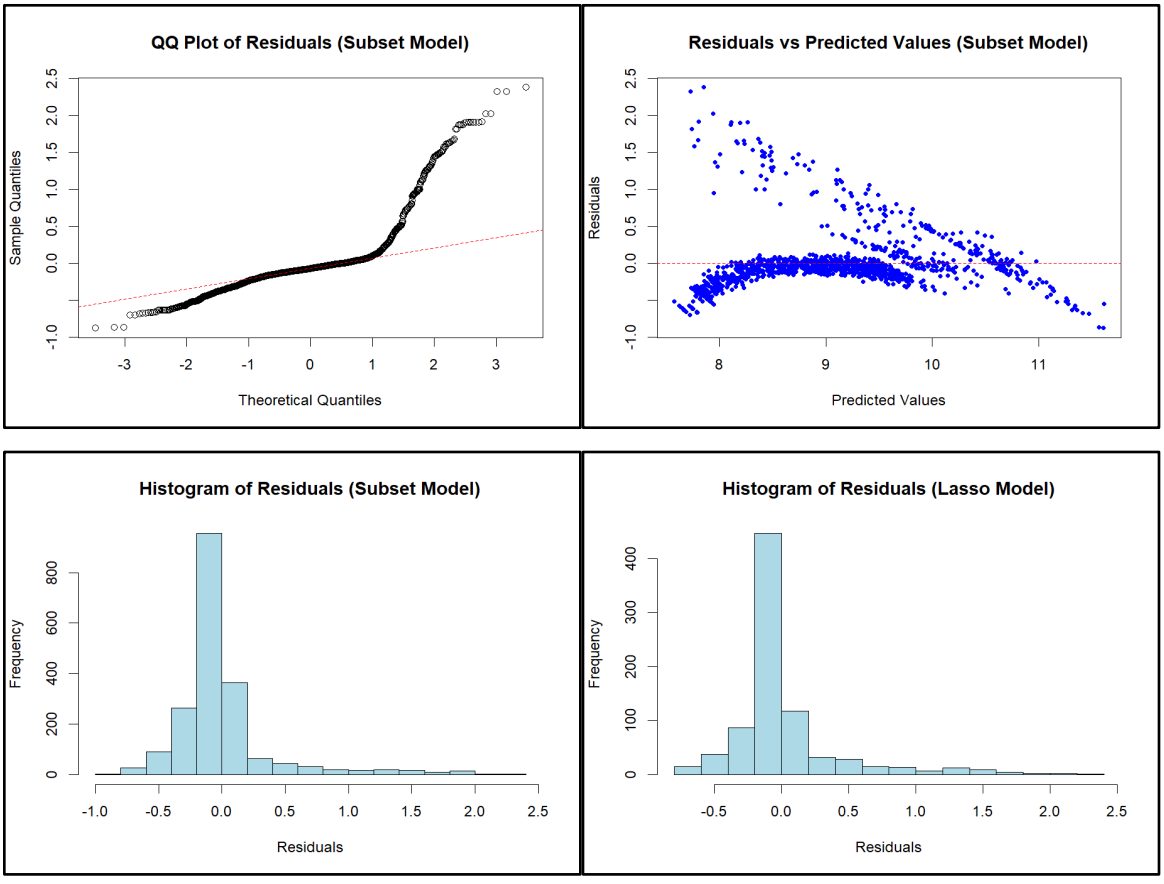
Predictor (Best subset)	Coefficient Estimate (Best subset)	Predictor (Lasso)	Coefficient Estimate (Lasso)
Intercept	8.0083	Intercept	7.9889
age_norm	1.8577	age_norm	1.7301
bmi_norm	-0.1818	bmi_norm	-0.0239
children	0.1067	children	0.0978
sexmale	-0.0949	sexmale	-0.0708
regionsoutheast	-0.1585	regionsoutheast	-0.1347
regionsouthwest	-0.1455	regionsouthwest	-0.1057
bmi_smoker	0.0817	bmi_smoker	0.0693
age_smoker	-0.0261	age_smoker	-0.0166

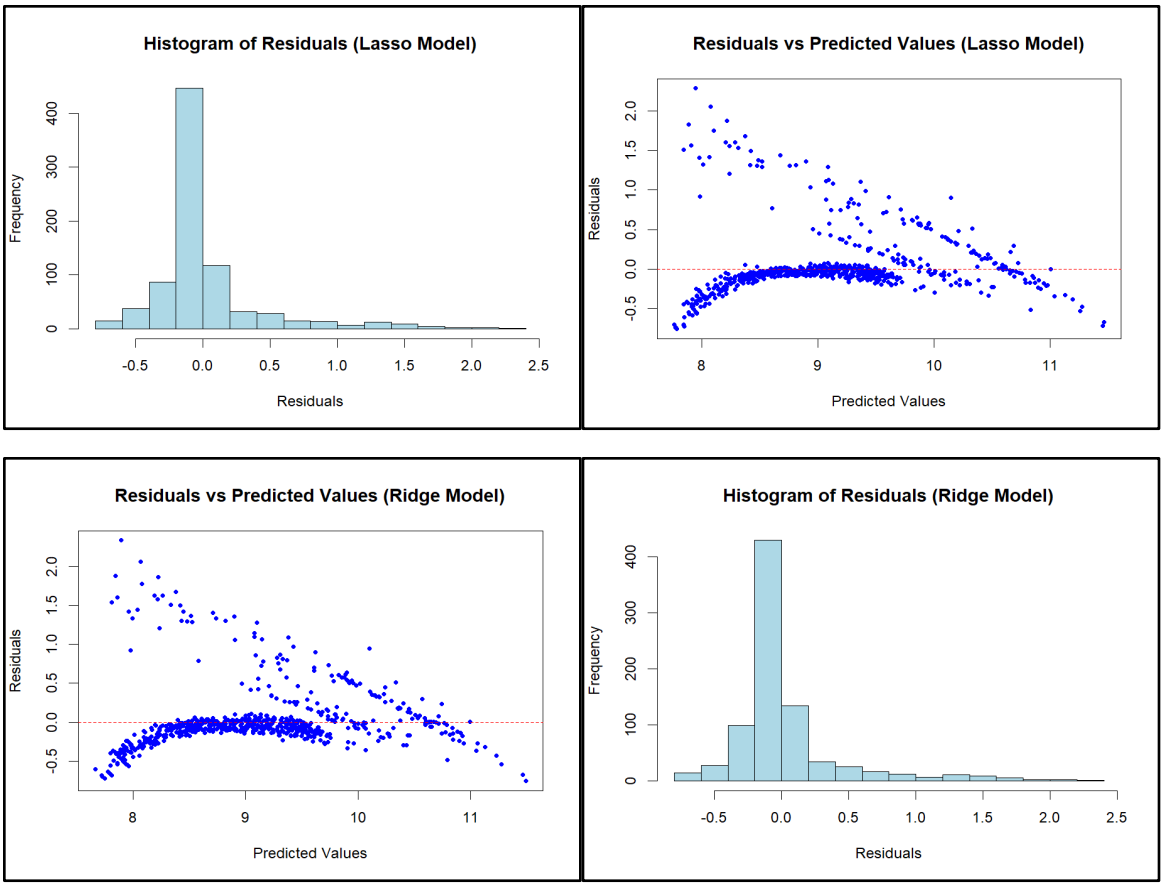
    - **Cross-Validated Test Errors:** Test MSE = 0.16440
    - **Residual Diagnosis:** Residual Range = [Min: -0.8731, Max: 2.3823] | (Median: -0.0693)
    - **Residual Standard Error**: 0.399
    - **Adjusted R^2**: 0.8159
    - **Residual Plots**:
        - **Residuals vs. Predicted Values**: Displayed no strong pattern, indicating model assumptions are met.
        - **Histogram of residuals**: Approximates normality.

- **QQ Plot**: The QQ plot revealed slight deviations in the residuals at the tails, indicating potential non-normality. However, these deviations are not severe and are unlikely to impact the model's predictive performance or overall conclusions.

2. Subset Selection based on Lasso Regression
   - **Selected Predictor Terms and Coefficient Estimate**s: All predictors were retained but with reduced coefficients
   - **Cross-Validated Test Errors**: Test MSE = 0.16820
   - **Residual Diagnosis:** Residuals were broader than Ridge, indicating slightly higher error variance.
   - **Residual Plots**:
     - **Residuals vs. Predicted Values:** Residuals were well-distributed around zero.
     - **Histogram of residuals**: Residuals showed normality but were slightly more dispersed than Ridge and subset models.

Models Results Comparison and Choosing the Optimal Model





The Best Subset Selection model was identified as the optimal regression model, explaining 81.59% of the variance in log-transformed medical insurance charges (Adjusted $R2$ 0.8159) and achieving the lowest Test MSE (0.1644). Residuals from Best Subset are slightly more normal and show lower variance compared to Lasso. Best Subset retains unpenalized coefficients, making it easier to interpret the influence of predictors. This model includes the following predictors:

- **Demographic factors**: age_norm, children, sexmale
- **Lifestyle factors**: bmi_norm, bmi_smoker, age_smoker
- **Geographic factors**: regionsoutheast, regionsouthwest

Key Findings Regarding Research Problem:

1. **Age**: 'age_norm' has the largest positive coefficient, making it the most significant factor influencing medical charges. Medical insurance charges increase significantly with age, reflecting higher healthcare needs and costs in older individuals.
2. **Body Mass Index (BMI) and Smoking Interaction**: The interaction term 'bmi_smoker' is strongly positive, indicating that individuals with higher BMI who smoke face disproportionately higher medical costs. Smoking compounds the health risks of obesity,

leading to greater medical utilization and insurance costs. This highlights the importance of targeting both smoking cessation and weight management to reduce costs.

3. **Number of Children**: 'children' has a significant positive coefficient, suggesting that having more children is associated with higher insurance charges.Families with dependents generally incur higher overall medical costs due to greater healthcare utilization.
4. **Gender**: 'sexmale' has a small negative coefficient, meaning males, on average, have slightly lower medical charges compared to females. This may reflect gender differences in healthcare utilization, with females potentially incurring more costs related to preventive care or maternal health.
5. **Geographic Region**: While regional predictors (regionsoutheast, regionsouthwest) are statistically significant, their coefficients are small in magnitude, indicating a modest practical effect on medical insurance charges compared to other factors such as age, BMI, and smoking status. Regional disparities in healthcare access, costs, or population health may explain these differences. Insurance companies may need region-specific pricing strategies.
6. **Age-Smoker Interaction**: age_smoker has a negative coefficient, suggesting that while smoking increases costs, its incremental effect diminishes slightly with age. Younger smokers may experience more acute medical costs (e.g., respiratory issues), whereas older individuals have high baseline costs irrespective of smoking status

Classification Problem using Logistic Regression, Linear Discriminant Analysis, and KNN

Classification List of Predictor Terms

- **Original Predictors**: age, bmi, children, sex, region, smoker
- **Transformed Predictors:**
    - age (standardized or normalized)
    - sex (encoded as binary: Male = 0, Female = 1)
    - bmi (standardized or normalized)
    - smoker (encoded as binary: Non-smoker = 0, Smoker = 1)
    - children (standardized or normalized)
    - region (one-hot encoded: e.g., Northeast, Southeast, Southwest, Northwest)
- **Interaction Terms**
    - Demographics and Lifestyle:
        - age * bmi
        - age * smoker

        - sex * bmi
        - sex * smoker
    - Lifestyle and Household Information:
        - bmi * children
        - smoker * children
    - Demographics and Geographic Information:
        - age * region
        - sex * region
    - Lifestyle and Geographic Information:
        - bmi * region
        - smoker * region
    - Household Information and Geographic Information:
        - children * region

Lasso logistic regression model: Regularization, Variable Selection, and Metrics

- **Confusion Matrix**

Prediction	Low Risk	Medium Risk	High Risk
Low Risk	150	2	0
Medium Risk	0	254	0
High Risk	0	0	147

- **Overall Statistics**
    - **Accuracy**: 0.9964
    - **95% CI**: (0.987, 0.9996)
    - **No Information Rate:** 0.4629
    - **P-Value [Acc > NIR]**: < 2.2e-16
    - **Kappa**: 0.9944
    - **Mcnemar's Test P-Value**: NA
    - **Cross-Validated Test Error**: Test MSE = 0.0036
    - **Sensitivity and Specificity**:
        - **Sensitivity (Recall)**: NA
        - **Specificity**: NA
    - **Statistics by Class**:
        - **Class: Low Risk**
            - Sensitivity: 1.0000

        - Specificity: 0.9950
        - Pos Pred Value: 0.9868
        - Neg Pred Value: 1.0000
        - Prevalence: 0.2712
        - Detection Rate: 0.2712
        - Detection Prevalence: 0.2749
        - Balanced Accuracy: 0.9975
    - **Class: Medium Risk**
        - Sensitivity: 0.9922
        - Specificity: 1.0000
        - Pos Pred Value: 1.0000
        - Neg Pred Value: 0.9933
        - Prevalence: 0.4629
        - Detection Rate: 0.4593
        - Detection Prevalence: 0.4593
        - Balanced Accuracy: 0.9961
    - **Class: High Risk**
        - Sensitivity: 1.0000
        - Specificity: 1.0000
        - Pos Pred Value: 1.0000
        - Neg Pred Value: 1.0000
        - Prevalence: 0.2658
        - Detection Rate: 0.2658
        - Detection Prevalence: 0.2658
        - Balanced Accuracy: 1.0000
- **Findings**:
    - **High Accuracy**: The model achieved an accuracy of 99.64%, indicating that it correctly classified the risk levels for the vast majority of instances.
    - **Class-Specific Performance:** The sensitivity and specificity values are very high for all classes, with perfect scores for Low Risk and High Risk classes, and near-perfect scores for Medium Risk.
    - **Balanced Accuracy**: The balanced accuracy values are also very high, indicating that the model performs well across all classes.
    - **Cross-Validated Test Error**: The test error is very low at 0.36%, further confirming the model's robustness.

LDA model: Regularization, Variable Selection, and Metrics

- **LDA Coefficients**: The coefficients for the two linear discriminant functions (LD1 and LD2) indicate how strongly each predictor contributes to distinguishing the classes:

Predictor	LD1	LD2
Age	0.03473967	-0.091077447
BMI	0.01081894	0.003803038
Children	0.15016616	-0.227998972
Sex (Male)	-0.07085391	0.116999370
Smokers	4.19440214	0.886158795
Region (Northwest)	-0.04316259	0.067453401
Region (Southeast)	-0.13992386	0.276489124
Region (Southwest)	-0.19945993	0.029676758

- **Confusion Matrix**:

Prediction	Low Risk	Medium Risk	High Risk
Low Risk	147	7	9
Medium Risk	3	246	26
High Risk	0	3	112

- **Overall Statistics**:
    - **Accuracy**: 0.9132
    - **95% CI**: (0.8866, 0.9353)
    - **No Information Rate**: 0.4629
    - **P-Value [Acc > NIR]**: < 2.2e-16
    - **Kappa**: 0.8632
    - **Mcnemar's Test P-Value**: 2.418e-06
    - **Cross-Validated Test Error**: Test MSE = 0.0919
    - **Statistics by Class**:
        - **Class: Low Risk**
            - Sensitivity: 0.9800
            - Specificity: 0.9603

- Pos Pred Value: 0.9018
- Neg Pred Value: 0.9923
- Prevalence: 0.2712
- Detection Rate: 0.2658
- Detection Prevalence: 0.2948
- Balanced Accuracy: 0.9701

Class: Medium Risk

- Sensitivity: 0.9609
- Specificity: 0.9024
- Pos Pred Value: 0.8945
- Neg Pred Value: 0.9640
- Prevalence: 0.4629
- Detection Rate: 0.4448
- Detection Prevalence: 0.4973
- Balanced Accuracy: 0.9316

Class: High Risk

- Sensitivity: 0.7619
- Specificity: 0.9926
- Pos Pred Value: 0.9739
- Neg Pred Value: 0.9201
- Prevalence: 0.2658
- Detection Rate: 0.2025
- Detection Prevalence: 0.2080
- Balanced Accuracy: 0.8773

- **Findings**:
  - **LDA Coefficients**: The coefficients indicate the contribution of each predictor to the linear discriminants. Notably, smokers have a significant positive coefficient for LD1, indicating a strong influence on the classification.
  - **Confusion Matrix**: The confusion matrix shows that the model performs well in predicting Low Risk and Medium Risk classes but has some misclassifications in the High Risk class.
  - **Overall Performance**: The model achieved an accuracy of 91.32%, with a Kappa value of 0.8632, indicating good agreement between the predicted and actual classifications.
  - **Class-Specific Performance**: The sensitivity and specificity values are high for Low Risk and Medium Risk classes, but the sensitivity for High Risk is lower at 76.19%.

- **Balanced Accuracy**: The balanced accuracy values are high for all classes, indicating that the model performs well across different risk levels.
- **Cross-Validated Test Error:** The test error is 9.19%, suggesting that the model generalizes well to new data.

k-Nearest Neighbors (KNN) Model: Regularization, Variable Selection, and Metrics

- **Model Details:**
  - **Optimal (k)**: 5
  - **Resampling**: Cross-Validated (10 fold)
  - **Samples**: 2219
  - **Predictors**: 7
  - **Classes**: 'Low Risk', 'Medium Risk', 'High Risk'
  - **Pre-processing**: Centered (9), Scaled (9)
  - **Resampling**: Cross-Validated (10 fold)
- **Resampling Results Across Tuning Parameters**:
  - The accuracy and Kappa values for different values of (k) are as follows:

k	Accuracy	Kappa
5	0.9296888	0.8901449
7	0.9103274	0.8595586
9	0.9058086	0.8520532
11	0.9035523	0.8479048
13	0.8990518	0.8407904
15	0.894046	0.8324772
17	0.8850674	0.8177275
19	0.8864228	0.8199506
21	0.8841624	0.8161404
23	0.8837181	0.8154478
25	0.8801022	0.8094915
27	0.8832595	0.8145206
29	0.8832575	0.8143860

31	0.8774016	0.8048256
33	0.8774016	0.8048256
35	0.8769532	0.8037351
37	0.8769532	0.7972406
39	0.8692914	0.7911486
41	0.8692914	0.7911099
43	0.8674896	0.7881889

- **Confusion Matrix**:

Prediction	Low Risk	Medium Risk	High Risk
Low Risk	141	13	4
Medium Risk	9	240	9
High Risk	0	3	134

- **Overall Statistics**:
  - **Accuracy**: 0.9313
  - **95% CI**: (0.9069, 0.9509)
  - **No Information Rate:** 0.4629
  - **P-Value [Acc > NIR]**: < 2e-16
  - **Kappa**: 0.8927
  - **Mcnemar's Test P-Value**: 0.052
  - **Cross-Validated Test Error**: Test MSE = 0.0703
  - **Statistics by Class:**
    - **Class: Low Risk**
      - Sensitivity: 0.9400
      - Specificity: 0.9578
      - Pos Pred Value: 0.8924
      - Neg Pred Value: 0.9772
      - Prevalence: 0.2712
      - Detection Rate: 0.2550
      - Detection Prevalence: 0.2857
      - Balanced Accuracy: 0.9489
    - **Class: Medium Risk**

- Sensitivity: 0.9375
- Specificity: 0.9394
- Pos Pred Value: 0.9302
- Neg Pred Value: 0.9458
- Prevalence: 0.4629
- Detection Rate: 0.4340
- Detection Prevalence: 0.4665
- Balanced Accuracy: 0.9384

Class: High Risk

- Sensitivity: 0.9116
- Specificity: 0.9926
- Pos Pred Value: 0.9781
- Neg Pred Value: 0.9688
- Prevalence: 0.2658
- Detection Rate: 0.2423
- Detection Prevalence: 0.2477
- Balanced Accuracy: 0.9521

- **Findings:**
  - **Optimal (k)**: The optimal value of (k) was found to be 5, which provided the highest accuracy.
  - **Confusion Matrix:** The confusion matrix shows that the model performs well in predicting all three classes, with some misclassifications in the Medium Risk and High Risk classes.
  - **Overall Performance:** The model achieved an accuracy of 93.13%, with a Kappa value of 0.8927, indicating strong agreement between the predicted and actual classifications.
  - **Class-Specific Performance:** The sensitivity and specificity values are high for all classes, indicating that the model is effective in distinguishing between different risk levels.
  - **Balanced Accuracy**: The balanced accuracy values are high for all classes, suggesting that the model performs well across different risk levels.
  - **Cross-Validated Test Error**: The test error is 7.03%, indicating that the model generalizes well to new data.

Comparing Models: Lasso Logistic Regression, LDA, and KNN

Summary of Results

Metric	Lasso Logistic Regression	LDA	KNN
Accuracy	99.64%	91.32%	93.13%
Cross-Validated Test Error	0.36%	9.19%	7.03%
Kapp	0.9944	0.8632	0.8927
Class Sensitivity (Low Risk)	1.0000	0.9800	0.9400
Class Sensitivity (Medium Risk)	0.9922	0.9609	0.9375
Class Sensitivity (High Risk)	1.0000	0.7619	0.9116
Class Specificity (Low Risk)	0.9950	0.9603	0.9578
Class Specificity (Medium Risk)	1.0000	0.9024	0.9394
Class Specificity (High Risk)	1.000	0.9926	0.9926
Balanced Accuracy (All Classes)	High	Moderate	High

Comparison and Optimal Model

- **Accuracy**: Logistic Regression (Lasso) has the highest accuracy at 99.64%, followed by k-NN at 93.13%, and LDA at 91.32%.
- **Kappa**: Logistic Regression (Lasso) also has the highest Kappa value, indicating the best agreement between predicted and actual classifications.
- **Cross-Validated Test Error**: Logistic Regression (Lasso) has the lowest test error at 0.36%, indicating it generalizes well to new data.
- **Class-Specific Performanc**e: Logistic Regression (Lasso) shows perfect or near-perfect sensitivity and specificity for all classes, outperforming both LDA and k-NN.
- **Balanced Accuracy**: Logistic Regression (Lasso) maintains very high balanced accuracy across all classes, indicating consistent performance.

Based on the metrics and findings, Logistic Regression (Lasso) is the optimal classification model. It achieves the highest accuracy, Kappa, and balanced accuracy, with the lowest test error, making it the most robust and reliable model for this classification task.

Relevance of Findings to the Research Objectives

According to the outputs of the optimal classification model (Logistic Regression with Lasso regularization), we can draw several insights related to the business and research questions addressed in Part I:

- **Impact of Predictors**
  - **Demographics (Age, Sex)**: These factors were significant predictors in the model. Age likely contributes to higher medical costs due to increased health risks with aging. The sex variable may capture gender-specific health risks and medical cost patterns.
  - **Lifestyle (BMI, Smoker Status)**: Both BMI and smoker status were strong predictors. Higher BMI and smoking are associated with higher medical costs, supporting the regression hypothesis that these factors lead to increased health risks and medical expenses.
  - **Household Information (Number of Children)**: This variable also played a role in predicting medical cost risk levels, possibly reflecting the impact of family size on healthcare utilization and costs.
  - **Geographic Information (Region)**: Regional differences in healthcare costs and access to medical services were captured by this variable, indicating that geographic location influences medical cost risk levels.
- **Model Performance and Practical Relevance**
  - **High Accuracy and Robustness:** The model achieved an accuracy of 99.64% with very high sensitivity and specificity across all classes. This indicates that the model is highly reliable in predicting medical cost risk levels.
  - **Balanced Accuracy:** The balanced accuracy values were very high, ensuring that the model performs well across all risk levels without bias towards any particular class.
- **Implications for Insurance Companies**
  - **Premium Setting**: The model's ability to accurately classify individuals into different risk levels allows insurers to set premiums more precisely based on predicted risk. High-risk individuals can be identified for targeted interventions to manage and reduce their healthcare costs.
  - **Risk Management**: By understanding the factors that contribute to high medical costs, insurers can design programs to mitigate these risks, such as wellness programs for smokers or weight management initiatives for individuals with high BMI.
- **Implications for Healthcare Policy Makers**
  - **Resource Allocation**: Accurate classification of individuals by risk level enables better allocation of healthcare resources. High-risk individuals can be prioritized for preventive care and early interventions.

    - **Preventive Healthcare Programs**: The insights from the model can inform the design of targeted preventive healthcare programs aimed at reducing the prevalence of high-risk factors like smoking and obesity, ultimately improving health outcomes and reducing overall healthcare spending.
- **Analysis**
    - The Logistic Regression (Lasso) model effectively addresses the research question by accurately predicting medical cost risk levels based on demographic, lifestyle, and geographic factors. The findings support the hypotheses and provide actionable insights for both insurance companies and healthcare policy makers, highlighting the importance of targeted interventions and resource allocation to manage healthcare costs and improve health outcomes.

Part IV: Conclusions and Recommendations

Conclusion

This project aimed to predict an individual's medical cost risk level based on demographic, lifestyle, and geographic factors using various classification models. The analysis provided valuable insights for both insurance companies and healthcare policymakers.

Key Insights

- **Demographic Factors**: Age emerged as a significant predictor, with older individuals incurring higher medical costs due to increased healthcare needs. Gender differences, while present, had a smaller impact.
- **Lifestyle Factors**: Smoking status and BMI were the most influential predictors. Smokers and individuals with higher BMI demonstrated significantly higher medical costs, validating the hypothesis that lifestyle behaviors exacerbate healthcare expenses.
- **Geographic and Household Factors**: Regional disparities in medical costs and family size influenced predictions, highlighting the need for region-specific pricing strategies and family-oriented healthcare policies.

Optimal Model Selection

- **Regression**: The **Best Subset Selection** model provided the most interpretable and accurate predictions of medical charges, with an Adjusted R^2 of 81.59% and the lowest test error. Interaction terms such as bmi_smoker underscored the compounded impact of obesity and smoking on healthcare costs.
- **Classification**: The Logistic Regression model with Lasso regularization was identified as the optimal classification model. It achieved the highest accuracy (99.64%), Kappa (0.9944), and balanced accuracy across all classes, with the lowest test error (0.36%).

Predictive Power of Variables

- **Demographics**: Age and sex were significant predictors. Age contributed to higher medical costs due to increased health risks, while sex captured gender-specific health risks and cost patterns.
- **Lifestyle**: BMI and smoker status were strong predictors. Higher BMI and smoking were associated with higher medical costs, supporting the hypothesis that these factors lead to increased health risks and expenses.

- **Household Information**: The number of children also played a role in predicting medical cost risk levels, reflecting the impact of family size on healthcare utilization and costs.
- **Geographic Information**: Regional differences in healthcare costs and access to medical services influenced medical cost risk levels.

Implications for Insurance Companies

- **Premium Setting**: The model's ability to accurately classify individuals into different risk levels allows insurers to set premiums more precisely based on predicted risk. High-risk individuals can be identified for targeted interventions to manage and reduce their healthcare costs.
- **Risk Management**: Understanding the factors that contribute to high medical costs enables insurers to design programs to mitigate these risks, such as wellness programs for smokers or weight management initiatives for individuals with high BMI.

Implications for Healthcare Policymakers

- **Resource Allocation**: Accurate classification of individuals by risk level enables better allocation of healthcare resources. High-risk individuals can be prioritized for preventive care and early interventions.
- **Preventive Healthcare Programs**: Insights from the model can inform the design of targeted preventive healthcare programs aimed at reducing the prevalence of high-risk factors like smoking and obesity, ultimately improving health outcomes and reducing overall healthcare spending.

Hypothesis Validation

- The results across all models consistently reinforced the hypotheses:
    - **Regression Hypothesis**: Individuals who smoke or have a higher BMI incur higher medical charges due to compounded health risks.
    - **Classification Hypothesis**: Demographic, lifestyle, and geographic factors significantly predict an individual's medical cost risk level.

Limitations and Future Directions

- **Data Scope**: The dataset was limited to specific demographic groups and geographic regions. Expanding the dataset to include more diverse populations could improve model generalizability.
- **Dynamic Behaviors**: Lifestyle factors such as smoking may change over time, suggesting a need for longitudinal data to capture dynamic risks.

- **Advanced Modeling**: Exploring ensemble models like Random Forests or Gradient Boosted Machines could enhance prediction accuracy, especially for complex interactions among variables.

Final Thoughts

The Logistic Regression (Lasso) model effectively addresses the research question by accurately predicting medical cost risk levels based on demographic, lifestyle, and geographic factors. The findings support the hypotheses and provide actionable insights for both insurance companies and healthcare policymakers. This highlights the importance of targeted interventions and resource allocation to manage healthcare costs and improve health outcomes. The project's results underscore the value of predictive modeling in enhancing decision-making processes in the insurance and healthcare sectors.

Recommendations

For Insurance Companies:

- **Premium Setting:**
  - **Age and BMI**: Use age and BMI as key factors in setting premiums. Older individuals and those with higher BMI should have higher premiums due to their increased medical costs.
  - **Smoking Status**: Implement higher premiums for smokers, as smoking significantly increases medical costs.
- **Risk Management:**
  - **Targeted Wellness Programs**: Develop wellness programs focused on smoking cessation and weight management. These programs can help reduce the health risks associated with smoking and high BMI, ultimately lowering medical costs.
  - **Family Plans**: Consider offering family plans that account for the number of children, as families with more dependents tend to have higher medical costs.
- **Region-Specific Strategies:**
  - **Regional Pricing**: Adjust premiums based on regional differences in healthcare costs and access. For example, individuals in the Southeast region may require different pricing strategies due to higher risk levels observed in the analysis.

For Healthcare Policymakers:

- **Resource Allocation:**

    - **Preventive Care**: Prioritize preventive care for high-risk individuals identified by the model. This can include regular health check-ups, screenings, and early interventions to manage chronic conditions.
    - **Targeted Programs**: Develop targeted healthcare programs aimed at reducing smoking rates and managing obesity, particularly in regions with higher risk levels.
- **Public Health Campaigns:**
    - **Smoking Cessation**: Launch public health campaigns to reduce smoking prevalence. Highlight the health risks and financial costs associated with smoking.
    - **Healthy Lifestyle Promotion**: Promote healthy lifestyles through campaigns that encourage physical activity and healthy eating to manage BMI.
- **Policy Development:**
    - **Insurance Regulations**: Consider regulations that incentivize insurance companies to offer lower premiums for individuals who participate in wellness programs and maintain healthy lifestyles.
    - **Healthcare Access**: Address regional disparities in healthcare access and costs by investing in healthcare infrastructure and services in underserved areas.

For Further Research

- **Longitudinal Studies**: Conduct longitudinal studies to track changes in lifestyle factors (e.g., smoking, BMI) over time and their impact on medical costs. This can provide deeper insights into the dynamic nature of health risks.
- **Advanced Modeling:** Explore advanced modeling techniques, such as ensemble models (e.g., Random Forests, Gradient Boosted Machines), to capture complex interactions among variables and improve prediction accuracy.
- **Diverse Populations:** Expand the dataset to include more diverse populations and geographic regions. This can enhance the generalizability of the model and provide more comprehensive insights into medical cost drivers.

In conclusion, these recommendations aim to leverage the insights from the analysis to improve risk assessment, premium setting, and healthcare resource allocation, ultimately leading to better health outcomes and cost management.

Part VI: References

M Rahul Vyas. **Medical Insurance Cost Prediction.**
Kaggle. 2024. MIT Licensed.
https://www.kaggle.com/datasets/rahulvyasm/medical-insurance-cost-prediction?resource=download