Health Insurance Premium Charges - Factors and Their Influence

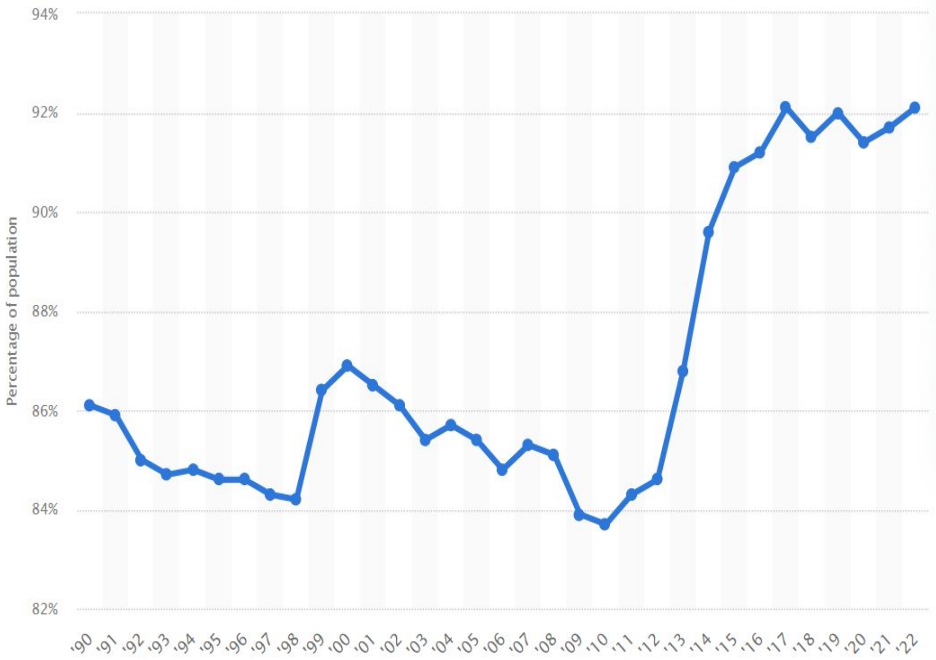
By Alex Ferrone and Akanksha Rai

Table of Contents

1. Introduction
2. Data Collection
3. Data Cleaning
4. Exploratory Data Analysis
5. Regression Models
6. Validation Testing

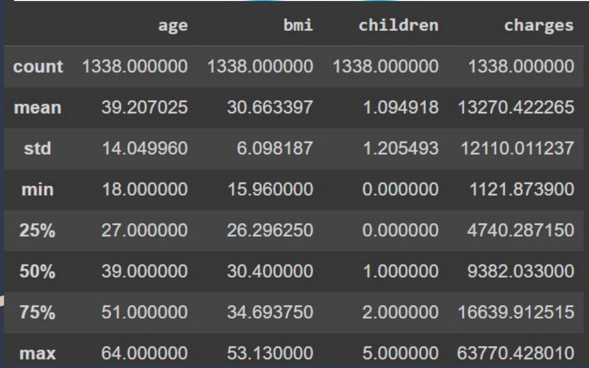
The Complexity of Risk Underwriting

% of people in the US with any type of Health Insurance from 1990 to 2022



Source: Statista

Data Collection



	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010

Age

Sex

BMI

Children

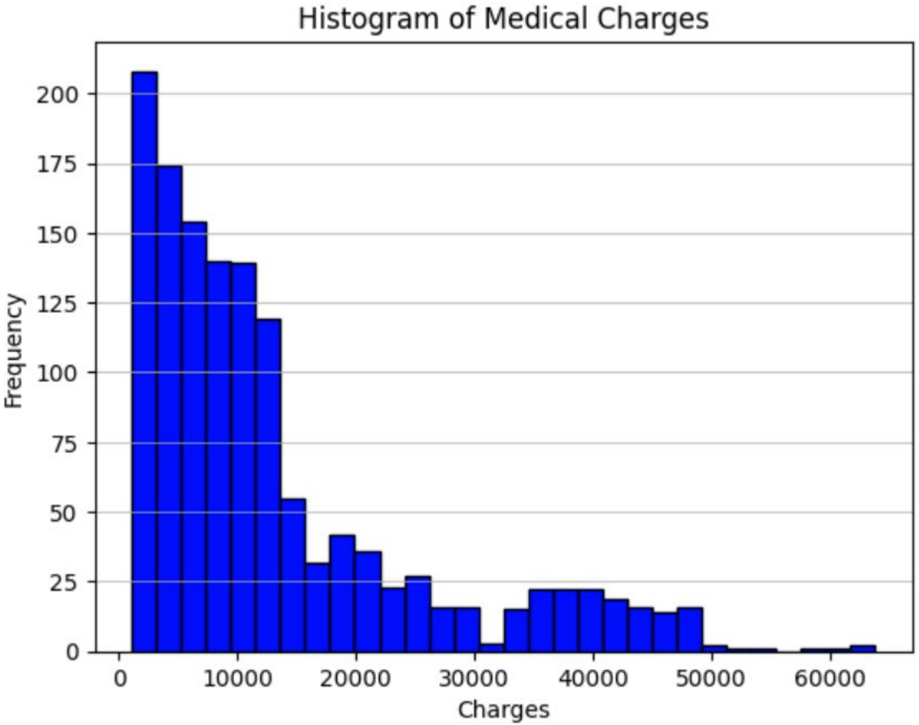
Smoker

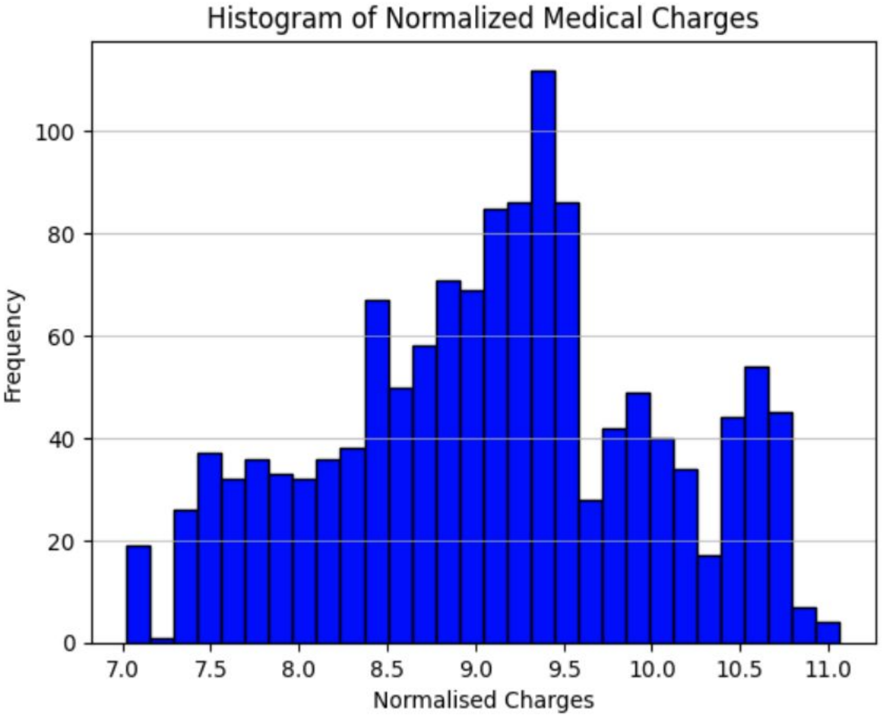
Region

Charges - dependent variable

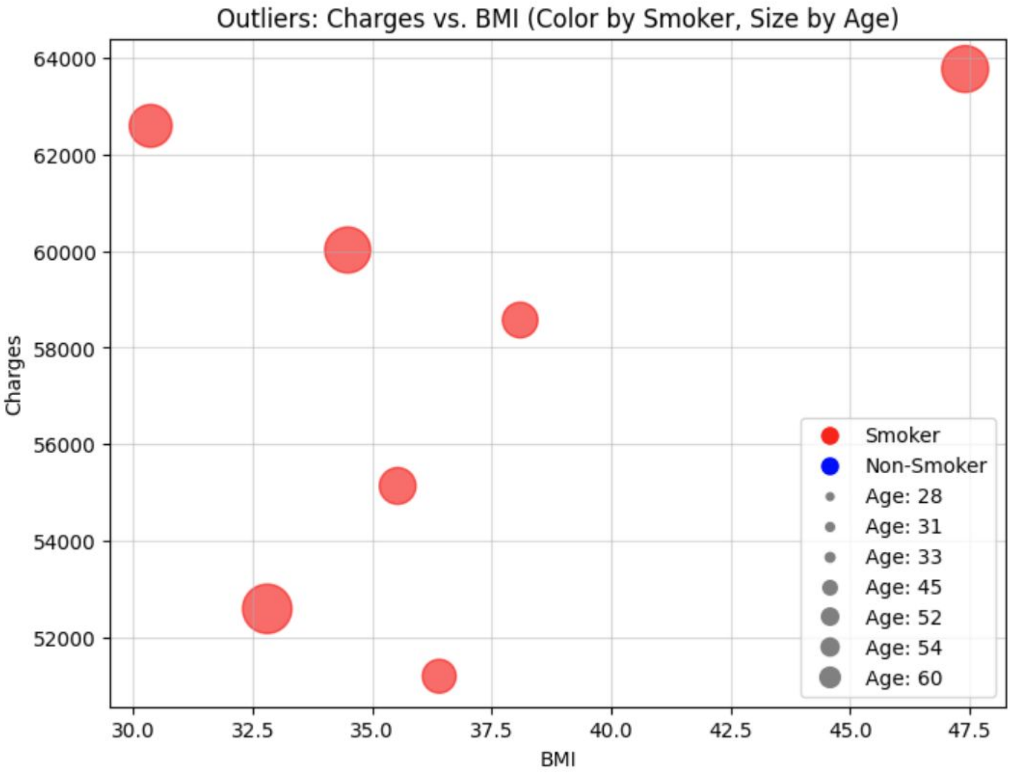
Child_bool*

Data Cleaning

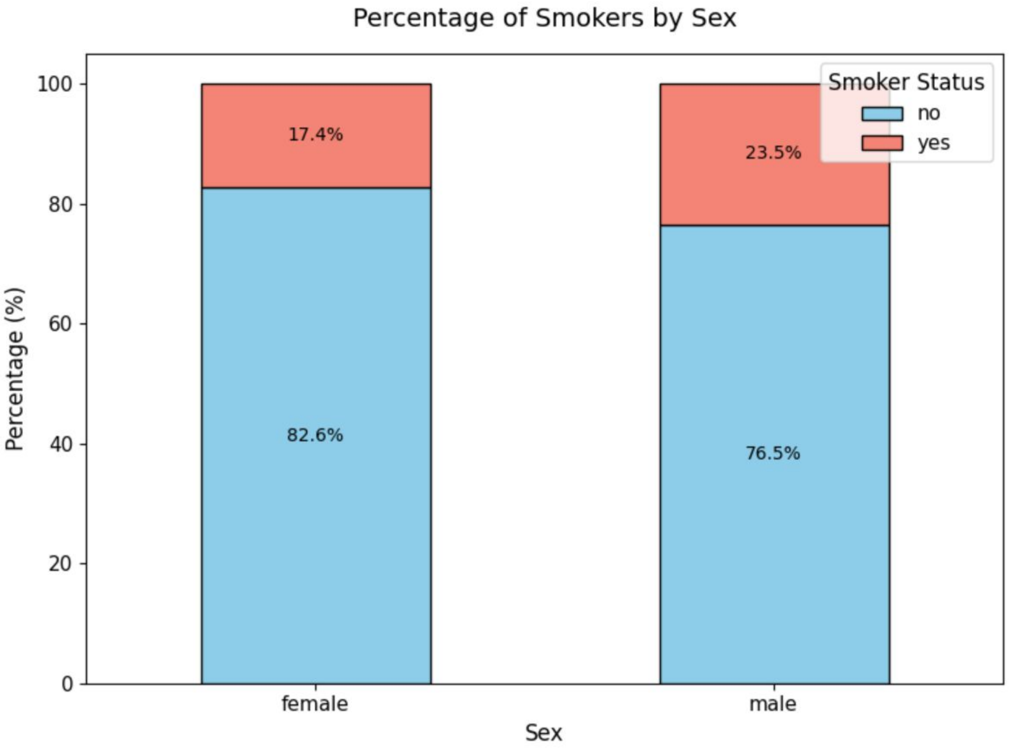




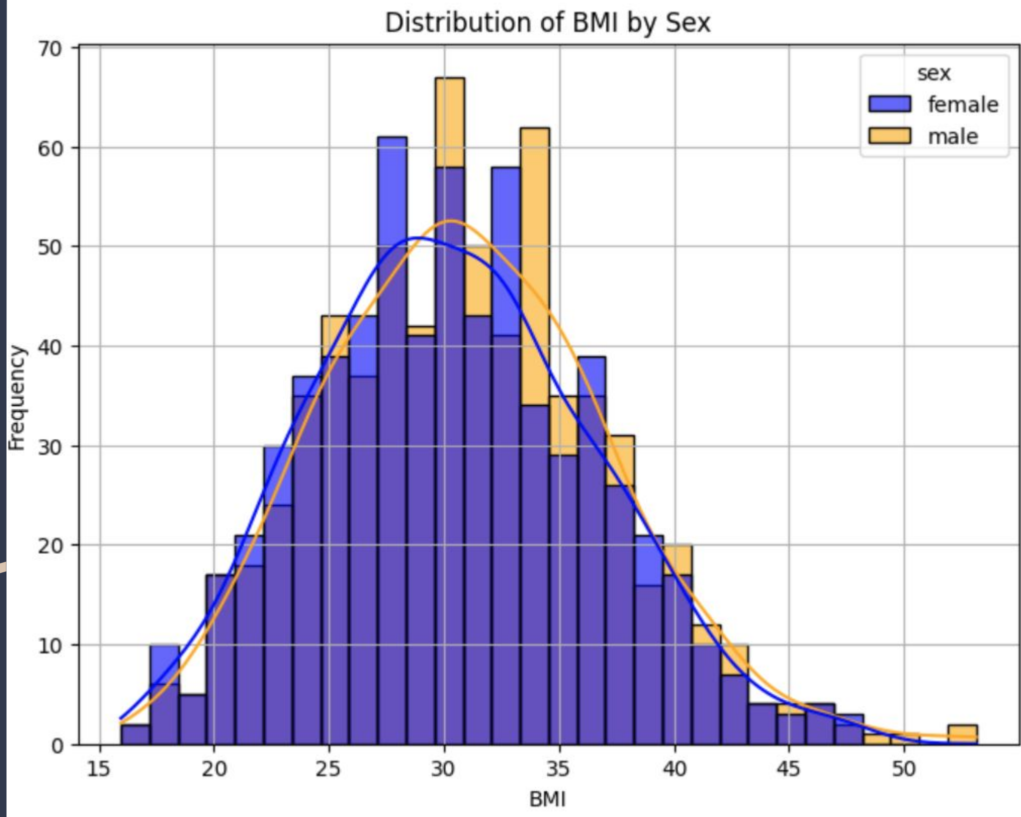
The outliers...



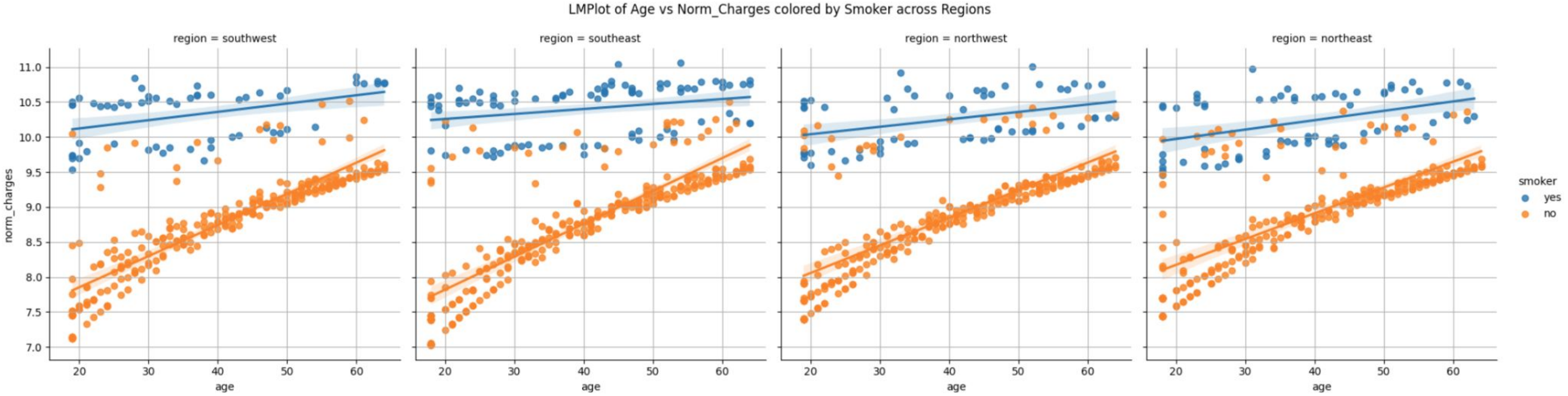
Who smokes more men or women?



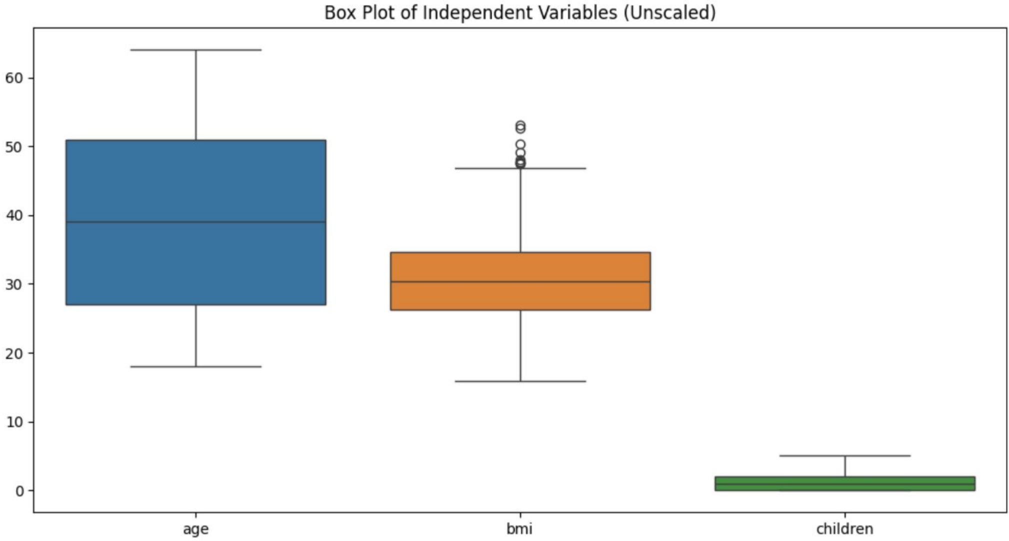
Who has a higher BMI men or women?



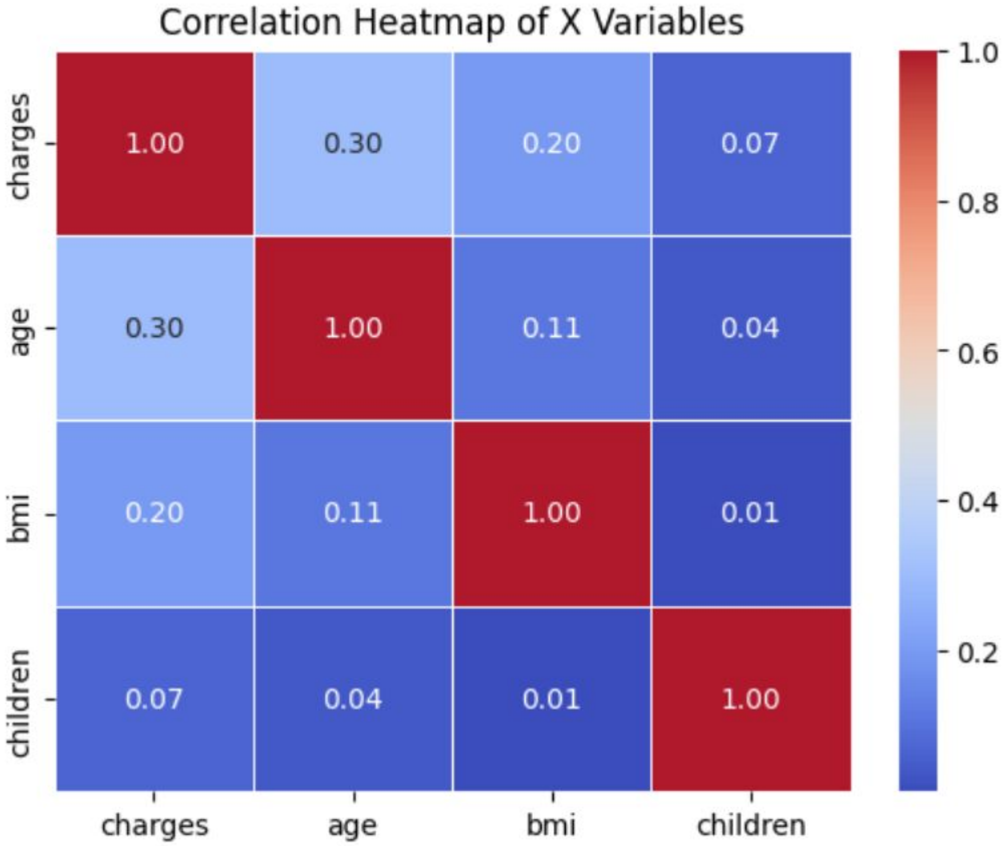
The largest contributor to price of premiums



Checking for normalization of independent variables



Final preparation for linear regression



Multi-Linear Regression

Adj R-Squared: 0.778

F-Statistic: 468.2

Jarque-Bera (JB): 1225.146

Mean Absolute Error: 0.2999

Coefficients: All p_values < 0.05

Smoker_yes: 1.5348

Age: 0.4948

BMI: 0.0790

Children: 0.1267

Sex_male: -0.0715

Region_northwest: -0.0951

Region_southeast: -0.1465

Region_southwest: -0.1420

Polynomial Regression

Degrees = 2

Adj R-Squared: 0.851

F-Statistic: 164.5

Jarque-Bera (JB): 8340.519

Mean Absolute Error: 0.2149

Smallest Eigenvalue: 1.09e-29

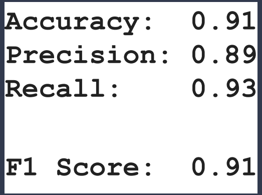
- Might indicate strong multicollinearity
- Or design matrix is singular

VIF and Durbin watson statistics to assess the model

feature	VIF
const	5.4290
age	1.0168
bmi	1.1066
children	1.004
sex_male	1.0089
smoker_yes	1.01207
region_northwest	1.5188
region_southeast	1.6522
region_southwest	1.5294

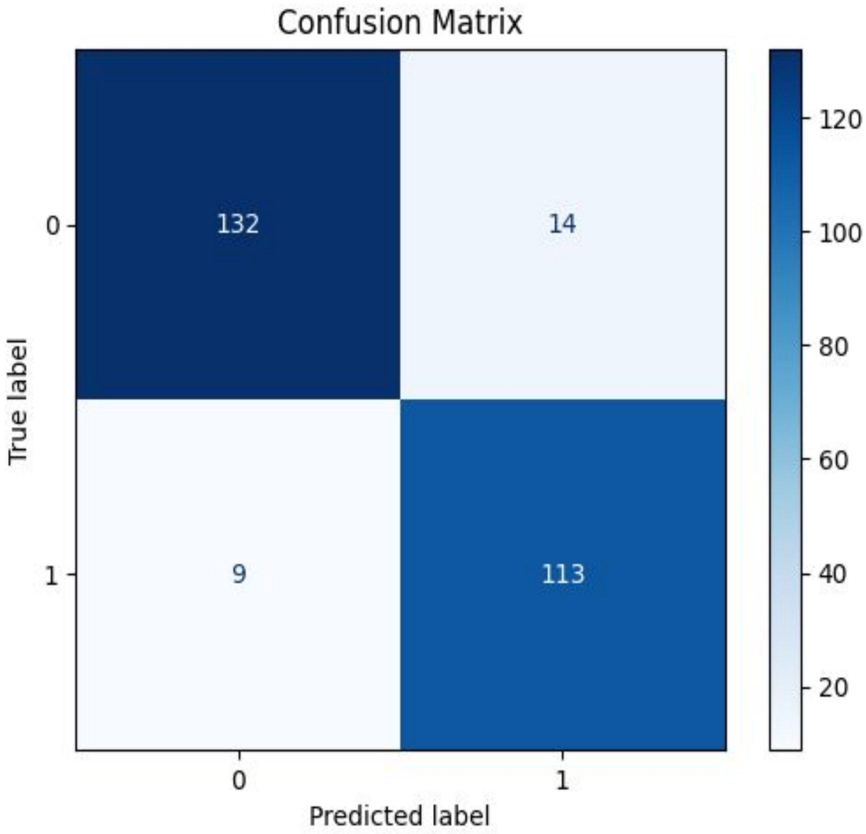
Durbin-Watson Statistic: 2.0464

Confusion Matrix

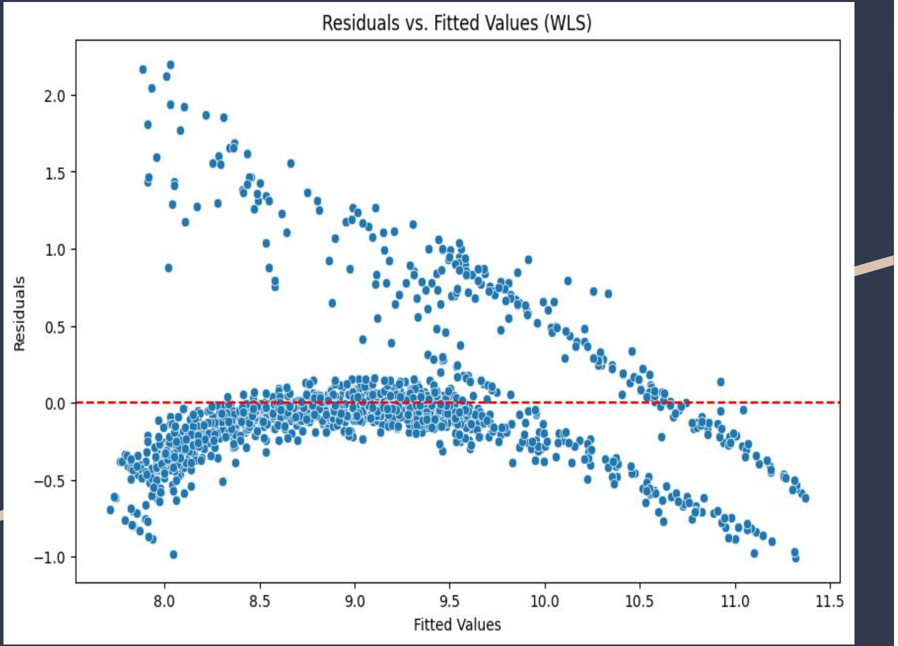


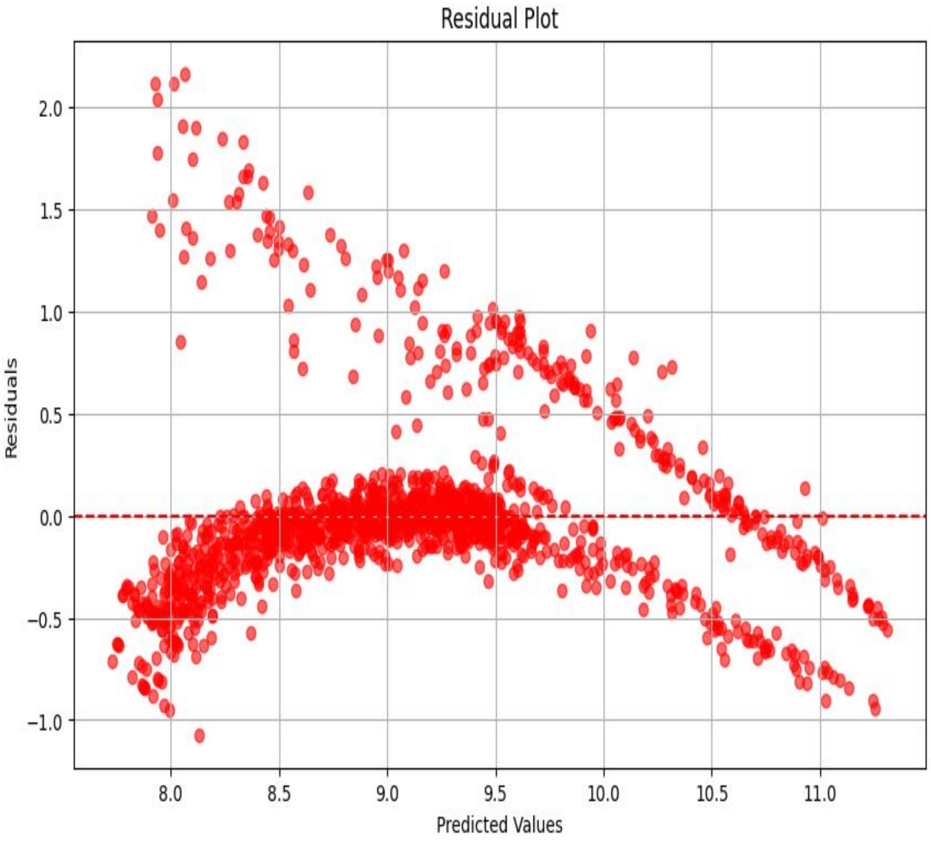
```
Accuracy:  0.91
Precision: 0.89
Recall:    0.93


F1 Score:  0.91
```

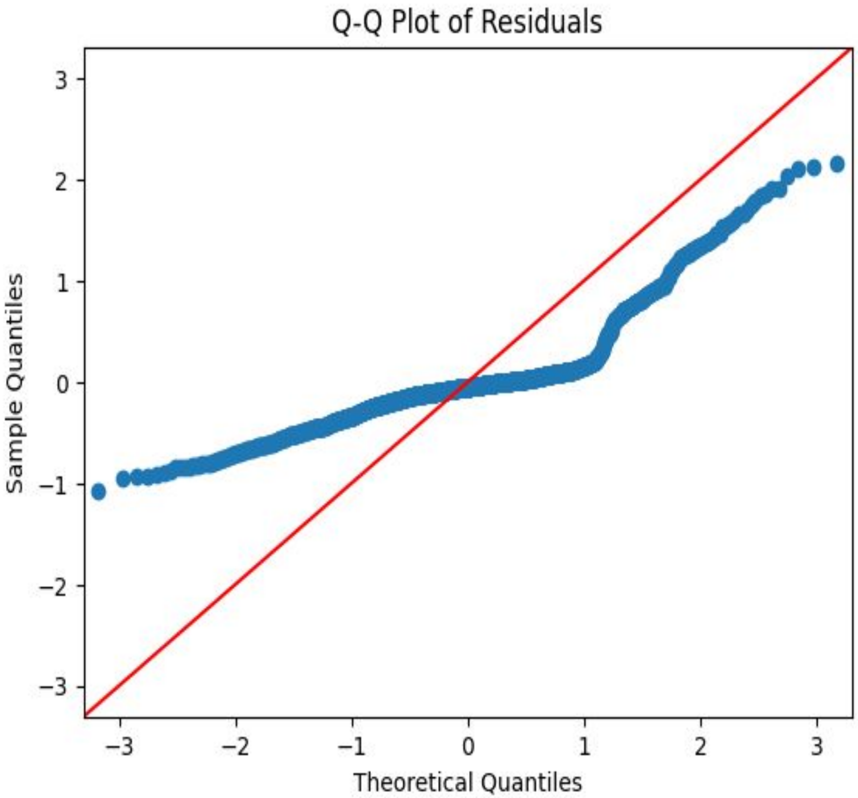


Residual plot to assess the performance of the model

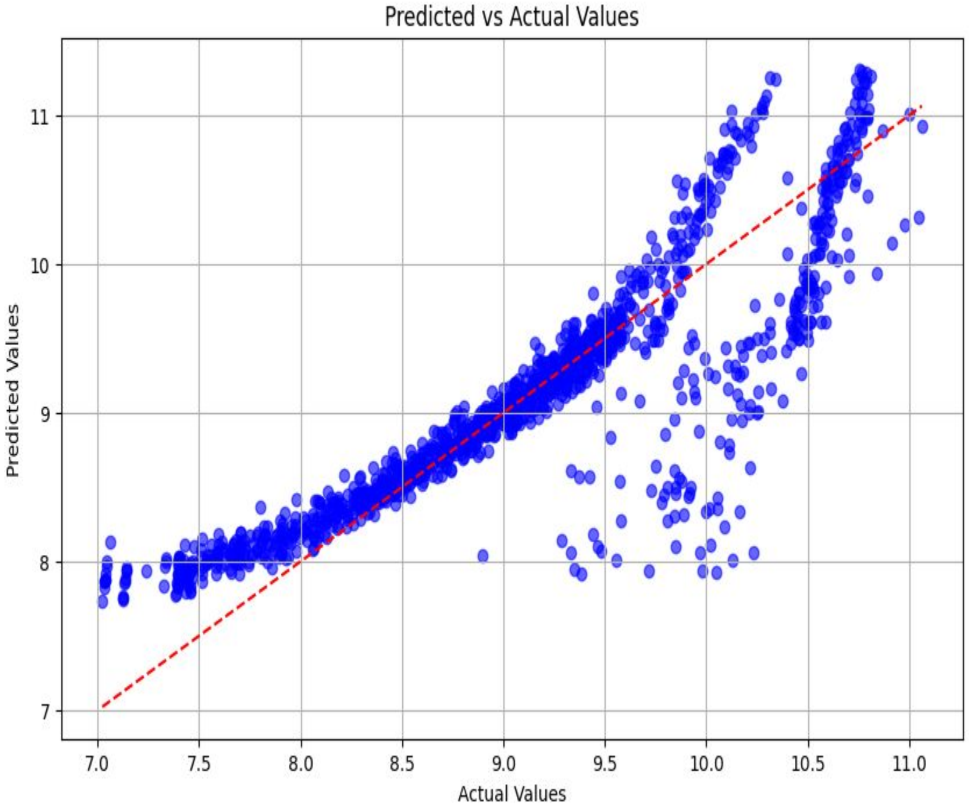




Q-Q Plot


Q-Q Plot of Residuals
Sample Quantiles
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3
Theoretical Quantiles

Predicted vs Actual values



Conclusion

Simple is sometimes better: Linear Regression

Additional Independent Variables

Demographic data

Medical History data

Policyholder Behavior data

Additional models to consider:

K-means clustering

Sentiment Analysis

Thank You !

Q&A

Any questions?

Sources

Statista

Kaggle

GeeksforGeeks

W3Schools

Collab.Research.Google

ChatGPT

MSBA 502 Slide Decks