Insurance Risk Claim Prediction Report

1. Executive Summary

This report explores a machine learning prediction analysis of medical insurance claims. The dataset includes individual-level information such as age, sex, BMI, number of children, smoking status, and residential region, alongside the target variable charges, which is the medical insurance cost based on these features. The analysis aims to build a predictive model that accurately estimates medical charges based on personal and demographic attributes.

This dataset is particularly suitable for linear regression due to a continuous numeric explanatory variable ‘Charges’, a set of statistically significant response variables and the dataset’s real-world relevance, making interpretability important in data analysis.

2. Planning of Analysis

The following steps were carried out to explore, preprocess, and model the data:

1. **Exploratory Data Analysis (EDA)**:
   This section aims to understand the data structure, as well as check for missing values, outliers, and correlations. In addition, visualizations of data, its distributions and relationships between features and the target variable (\`charges\`) will be explored. Potential outliers, and visualized relationships (e.g., correlation heatmap) will be created to further understand the information.
2. **Feature Engineering & Selection**:
   This section identified statistically significant features using p-values and other domain knowledge. This was completed using backward elimination to refine the model and drop insignificant variables. Categorical variables were converted to numeric using one-hot encoding and the use of dummy variables. Features were selected using backward elimination based on OLS regression p-values.
3. **Model Training**: The dataset was split into 70% training and 30% testing. An Ordinary Least Squares (OLS) regression model was trained on the selected features.
4. **Model Evaluation & Residual Analysis**: Model performance was assessed using R-squared and RMSE on the test set. Residuals were checked for normality, homoscedasticity, and independence. Multicollinearity was examined using Variance Inflation Factors (VIFs).

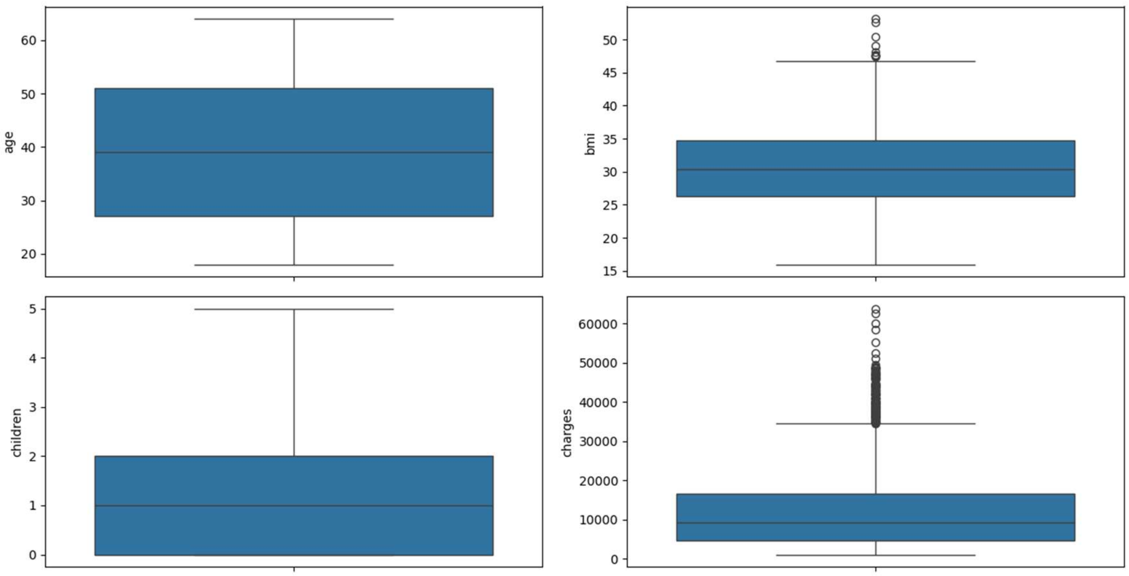
3. Analysis

3.1 Exploratory Data Analysis (EDA)

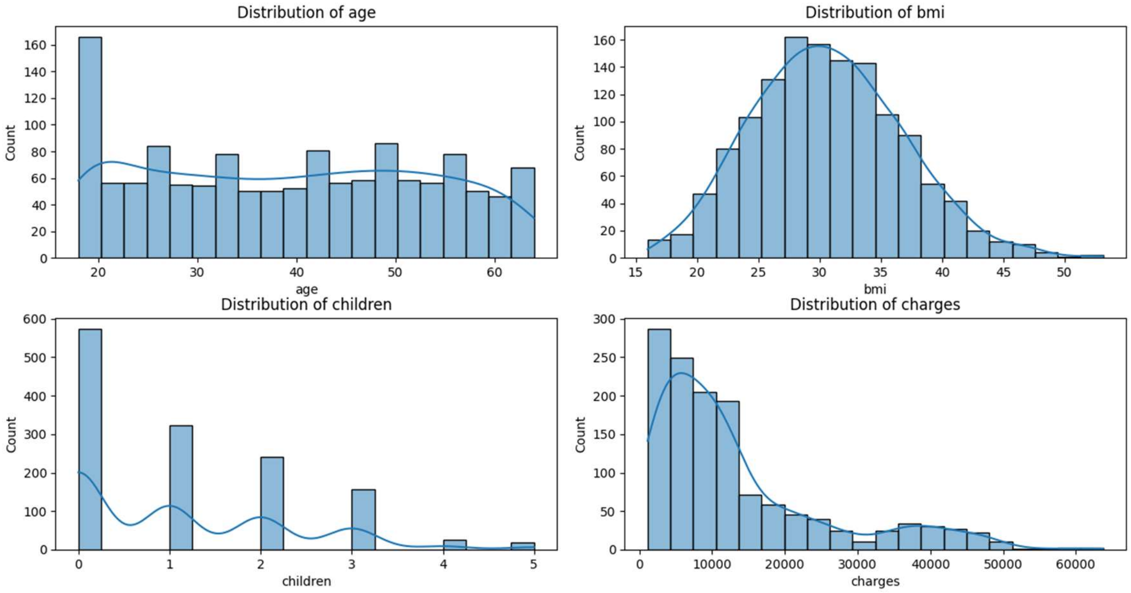
The dataset comprises 1,338 observations with no missing values across all variables. Summary statistics reveal that the average age is approximately 39 years, with a BMI mean of 30.66, suggesting a population leaning toward being overweight. Most individuals are non-smokers (1,064), and the majority reside in the Southeast region (364). The number of children ranges from 0 to 5, with charges spanning widely from around $1,122 to over $63,000. It has a mean of $13,270 and a high standard deviation, indicating substantial variability in insurance costs. This information about the data highlights key demographic and health-related patterns relevant to predicting charges.

Plot & Graph Analysis

1. Numerical Variables

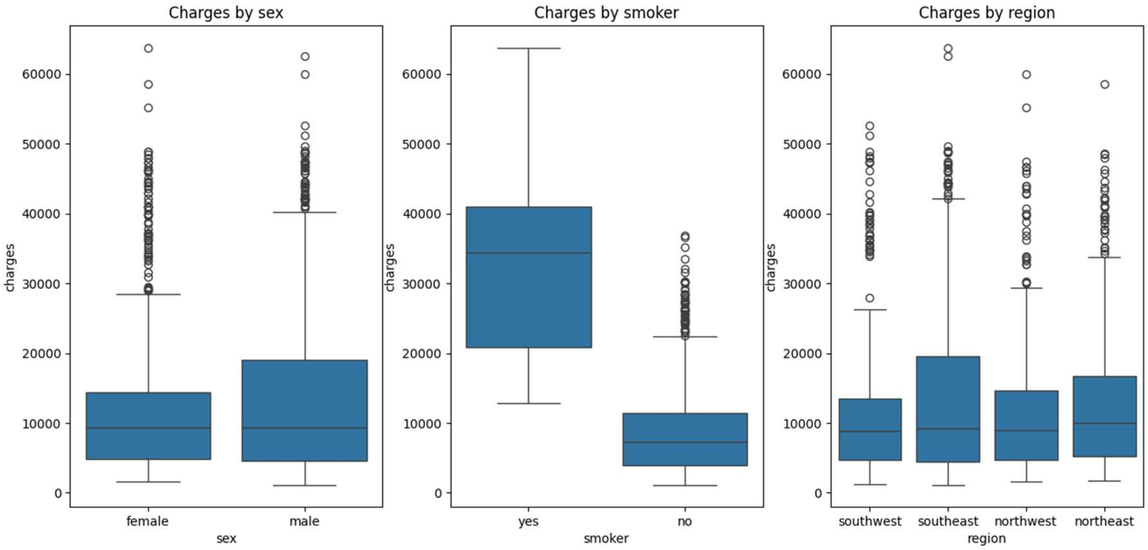


The **boxplots** help to identify the central tendency, variability, and presence of outliers. For example, charges display many high-value outliers, indicating that a small portion of individuals get significantly higher medical expenses. The BMI variable shows a few mild outliers on the higher end, suggesting a limited number of individuals with very high body mass index values. Meanwhile, age and children appear to be more symmetrically distributed, with minimal extreme values.



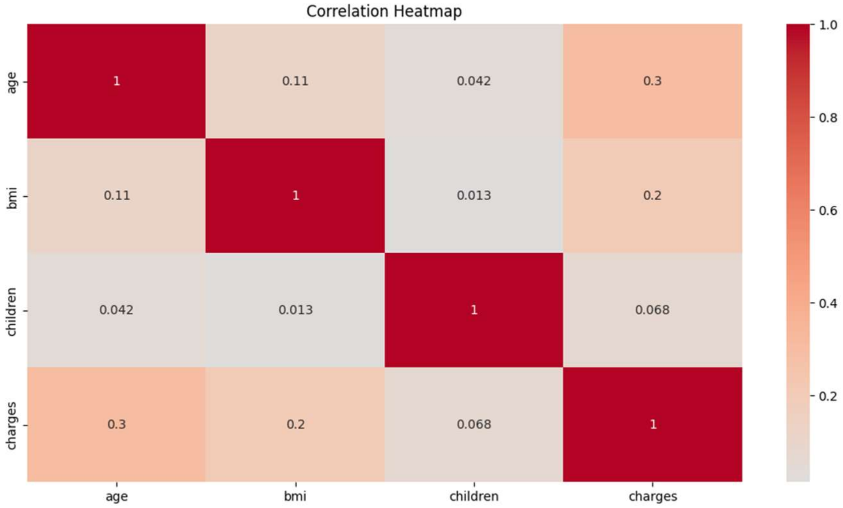
The **distributions** further support this by showing the frequency distribution of each variable. Age is roughly uniformly distributed between 20 and 60, while BMI has a bell-shaped distribution, peaking around 30. The ‘children’ variable is skewed towards zero. The charges histogram is right-skewed, showing that although most people incur lower charges, a few individuals face very high costs—likely influenced by factors such as smoking or chronic conditions.

2. Categorical Variables



The counts reveal that the dataset is balanced by sex, but heavily skewed by smoking category. The remaining people are non-smokers. The region variable is well-distributed, with the highest number of individuals residing in the Southeast and the lowest in the Northwest.

3. Correlation of Response Variables to Feature Variables



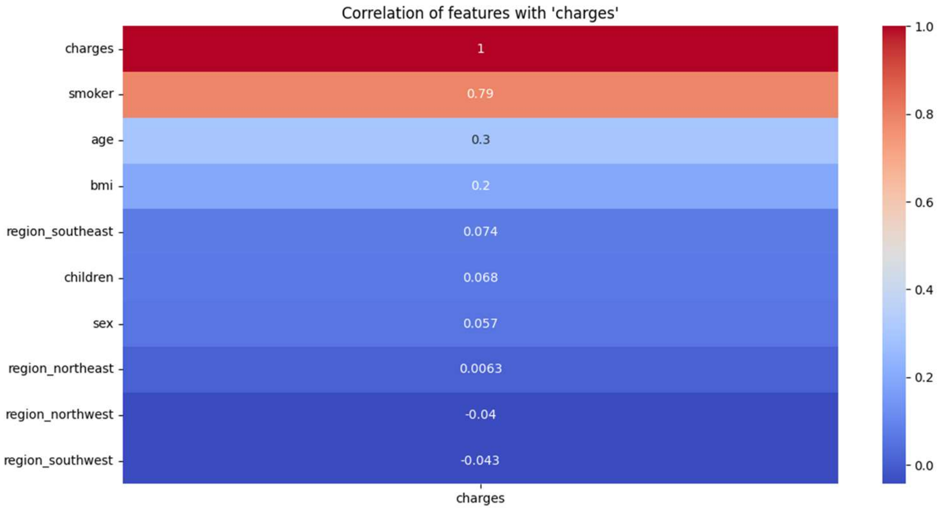
The figure shows a **correlation heatmap** for the numerical variables. It visually represents the strength and direction of linear relationships between them. The most notable correlation is between charges and smokers, which is strong and positive. This indicates that being a smoker is closely associated with higher medical costs. Age and BMI also show moderate positive correlations with charges, suggesting that older individuals and those with higher BMI tend to have higher charges.

3.2 Feature Engineering & Selection

This section involved preparing the region column for use in the linear regression model. It was converted into dummy variables using one-hot encoding. Since regression models require numerical inputs, categorical variables like region (which contained four distinct areas) must be transformed. To avoid multicollinearity in the model, one of these dummy columns was dropped, so as to avoid the dummy variable trap.

In addition, the dataset was split into training and test sets. The data was divided using a 70-30 split: 70% for training, and 30% for testing. An OLS model was fitted using the selected predictor variables from the training set.

The plot below visualizes the correlation of each feature with the target variable charges, using a bar plot. It reveals that a smoker has the highest positive correlation with charges, suggesting that being a smoker greatly increases insurance costs. Age and BMI show moderate but positive correlations. On the other hand, variables like sex and children have very weak correlations, implying a minimal direct effect on the cost.



Backward Elimination using OLS

To select features, backward elimination using p-values from an Ordinary Least Squares (OLS) regression was used to refine the model by removing statistically insignificant predictors. It was an iterative process that removes predictors with p-values greater than 0.05, as they do not contribute to predicting charges. In this case, the sex variable was eliminated due to its high p-value (~0.69), indicating that gender does not significantly affect medical charges in this dataset.

Final model predictors: Age, BMI, Children, Smoker, Region Dummy Variables.

3.3 Model Training

The final OLS regression model after backward elimination was used with the final predictors. Notably, the variables age, BMI, children, smoker have very low p-values, confirming their statistical significance in predicting medical charges. The model's **R-squared value** is relatively high, indicating that the predictors explain 74.2% of the variance in charges. The coefficients give insight into the direction and magnitude of influence for each variable. For instance, being a smoker is associated with an increase in predicted charges, while region-related dummy variables capture modest regional effects.

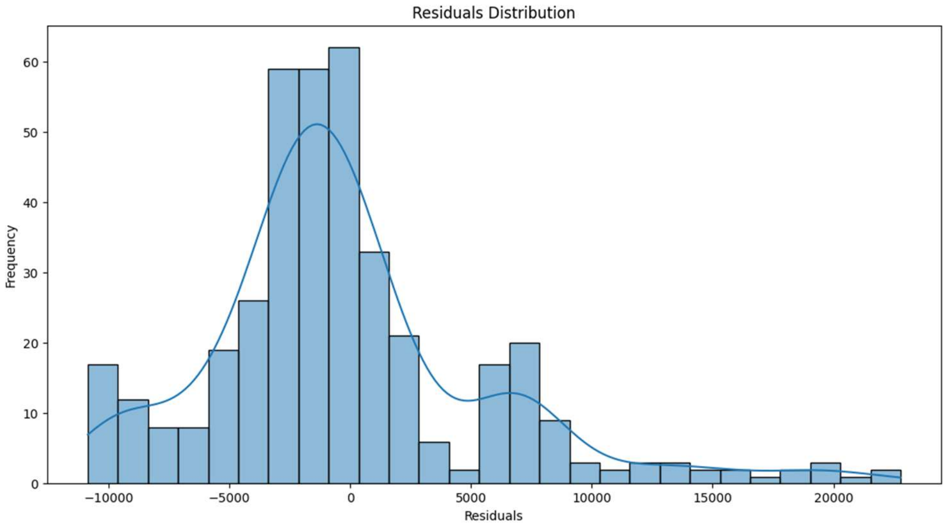
This table provides detailed statistical metrics for each feature retained in the model:

Dep. Variable:	charges	R-squared:	0.742
Model:	OLS	Adj. R-squared:	0.740
Method:	Least Squares	F-statistic:	381.8
Date:	Mon, 07 Apr 2025	Prob (F-statistic):	3.94e-268
Time:	21:45:43	Log-Likelihood:	-9493.1
No. Observations:	936	AIC:	1.900e+04
Df Residuals:	928	BIC:	1.904e+04
Df Model:	7		

3.4 Model Evaluation & Residual Analysis

Variance Inflation Factors (VIFs) were used for each feature in the model to assess multicollinearity. Multicollinearity occurs when predictors are highly correlated with each other, which can distort the interpretation of coefficients. Fortunately, all VIF values are below 5, indicating that multicollinearity is not a problem in the final model. This confirms that the independent variables do not overlap significantly in the information they contribute, and the model coefficients are stable and reliable.

The Breusch-Pagan test was performed, to check for heteroskedasticity. It is when the variance of residuals is not constant across all levels of the predicted values. The p-value for the test is less than 0.05, thus it suggests that there is no significant heteroskedasticity. This means the model satisfies one of the key assumptions of linear regression: constant variance of error terms. variance is reasonably satisfied.



4. Evaluation of Model & Results

The final model's Root Mean Squared Error (RMSE) is 5807.15, indicating that, on average, the model's predictions differ from the actual insurance charges by around $5807. This is a relatively reasonable error considering the wide range of charges in the dataset (from $1,122 to over $63,000). The R-squared value of 0.77 suggests that **77%** of the variability in medical insurance charges is explained by the model's features. This represents a strong case, making the model a good predictor for insurance charges.

5. Conclusion

The linear regression model developed in this analysis demonstrates strong predictive performance and interpretability. Key predictors such as smoking status, age, and BMI were statistically significant, with smoking having the most substantial impact on charges. Diagnostic tests and plots confirmed that model assumptions such as linearity, homoscedasticity, and lack of multicollinearity were satisfied. With an R^2 of 0.77 and a modest RMSE, the model offers a robust and transparent approach to estimating medical insurance costs.

References

Kennedy, W. B., 2025. *An Overview of Feature Selection.* [Online]
Available at: https://towardsdatascience.com/an-overview-of-feature-selection-1c50965551dd/
[Accessed 6 April 2025].

MOUFAD, B., 2022. *Towards Data Science.* [Online]
Available at: https://towardsdatascience.com/beyond-linear-regression-467a7fc3bafb/
[Accessed 8 April 2025].

Prabhu, T. N., 2019. *Explanatory Data Analysis In Python.* [Online]
Available at: https://towardsdatascience.com/exploratory-data-analysis-in-python-c9a77dfa39ce/
[Accessed 7 April 2025].

StatsModels, 2024. *statsmodels.regression.linear_model.OLS.* [Online]
Available at:
https://www.statsmodels.org/stable/generated/statsmodels.regression.linear_model.OLS.html
[Accessed 8 April 2025].