Problem Framing & Objective

Problem statement

Build a model to predict medical insurance charges and identify the key factors influencing increasing healthcare insurance costs.

Objectives

Enables accurate predictions of medical insurance costs to promote fair and transparent insurance charges, benefitting both insurers and customers, while providing insights on factors that contribute to high healthcare costs.

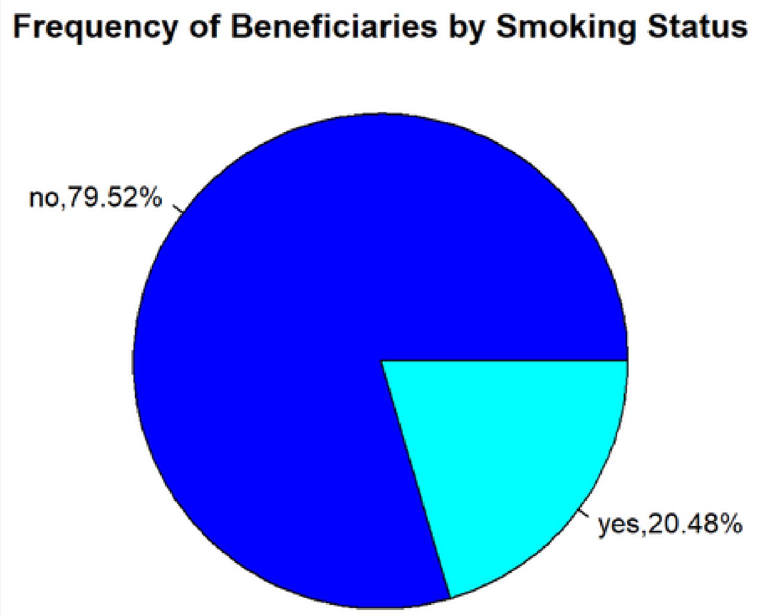
Variable Types

The 6 variables we are interested in exploring:

Categorical	Ratio
Sex	Age
Smoking status	Number of children
Residential region	BMI

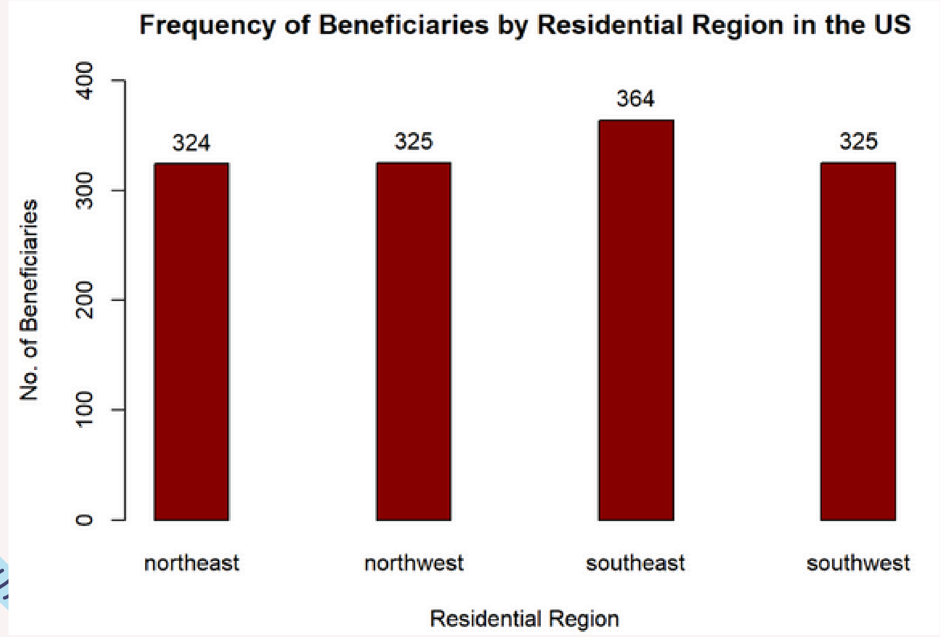
Data Overview

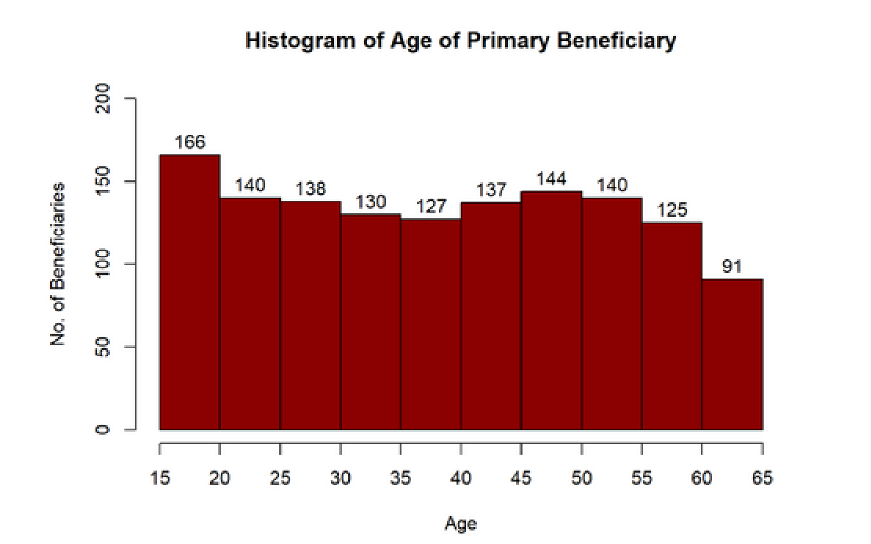
Number of observations: 1338



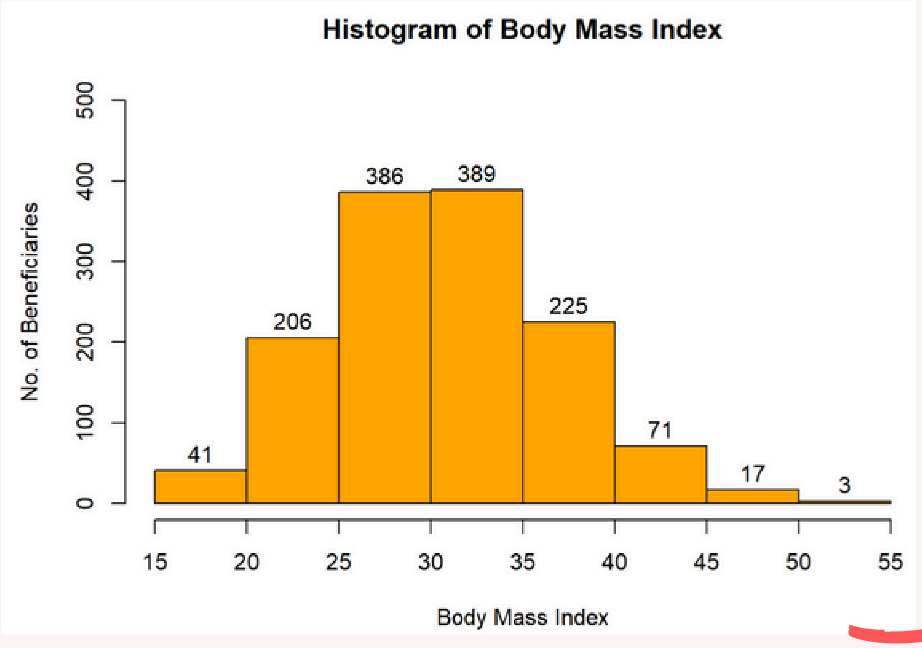
Smoking status
- Majority do not smoke

Region
- Frequency of beneficiaries across residential regions in the US is relatively balanced

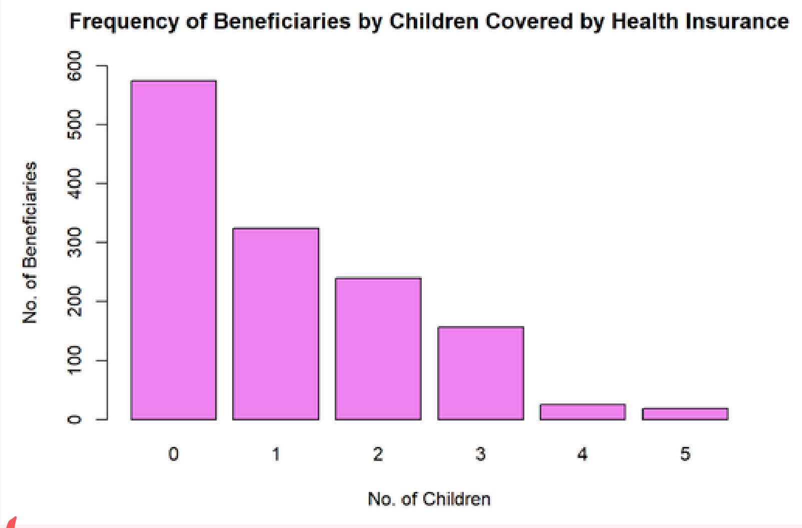




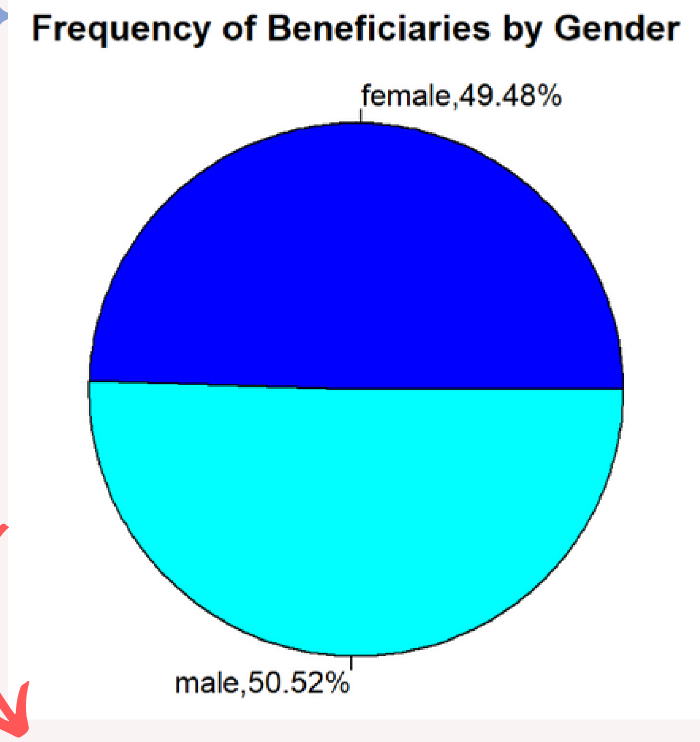
Age
- Relatively balanced between the ages of 20 and 60
- Number of beneficiaries aged between 60 and 65 is almost half of those who are aged between 15 and 20



BMI
- Majority of primary beneficiaries have a BMI of between 20 and 45.
- 3 extreme points between 50 and 55 → outliers



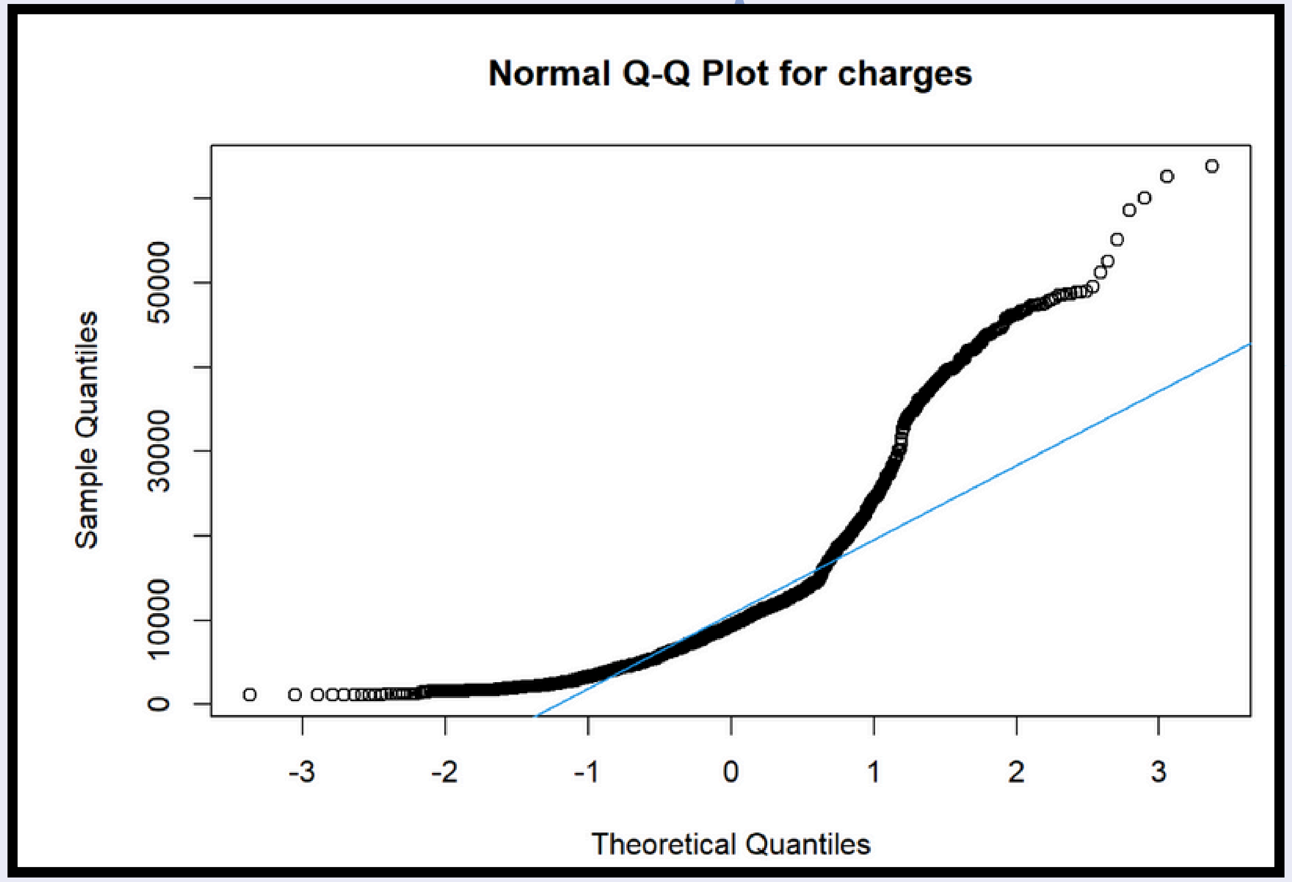
Children
- Number of primary beneficiaries is the highest among those with no children and lowest among those with 5 children.
- Majority have 0 to 3 children



Sex
- Percentage of female insurance beneficiaries is approximately equal to those who are male → sex has little impact on probability of buying insurance

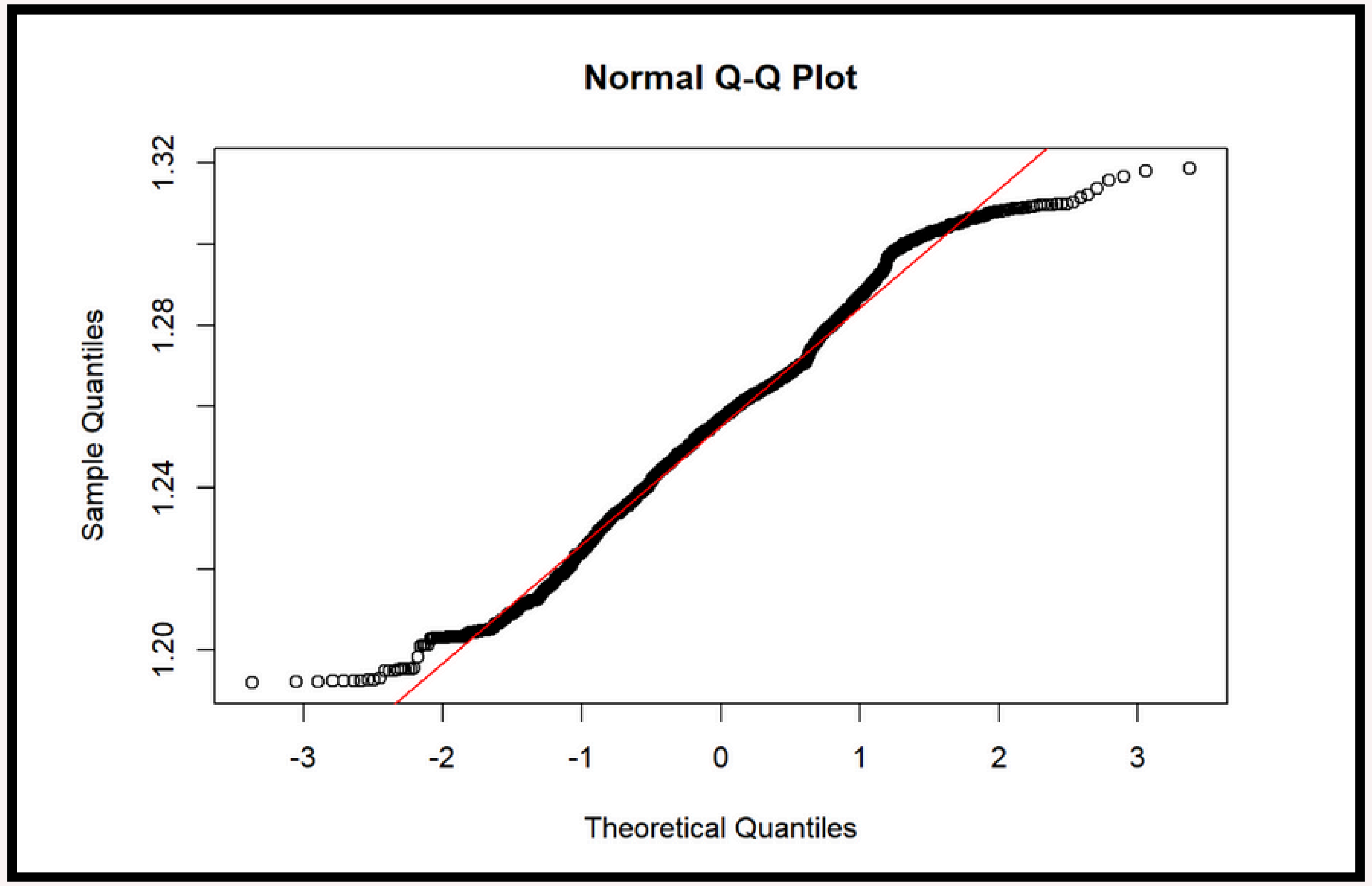
EDA - Response Variable

Response Variable:
Medical insurance charges



Charges show strong right skew.

Addressing skewness, we applied Tukey power transformation ($\mathbf{x}^{0.025}$, where x == df1$charges) to normalize distribution.

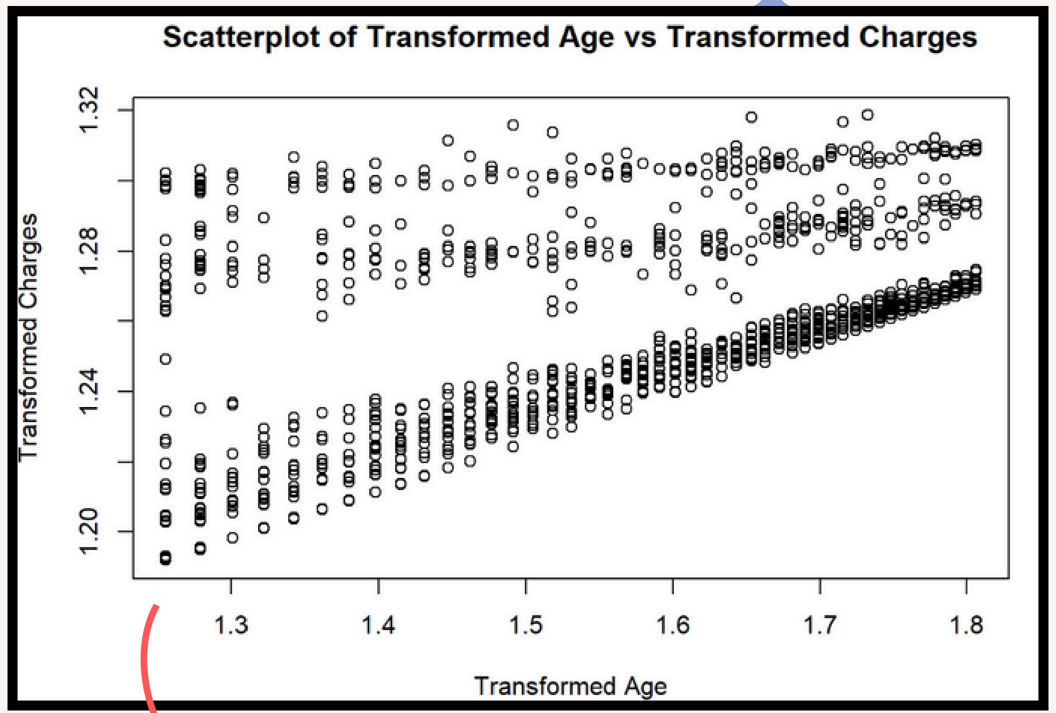


QQ-plot after transformation closely follows diagonal, indicating improved normality.

EDA - Key Predictors

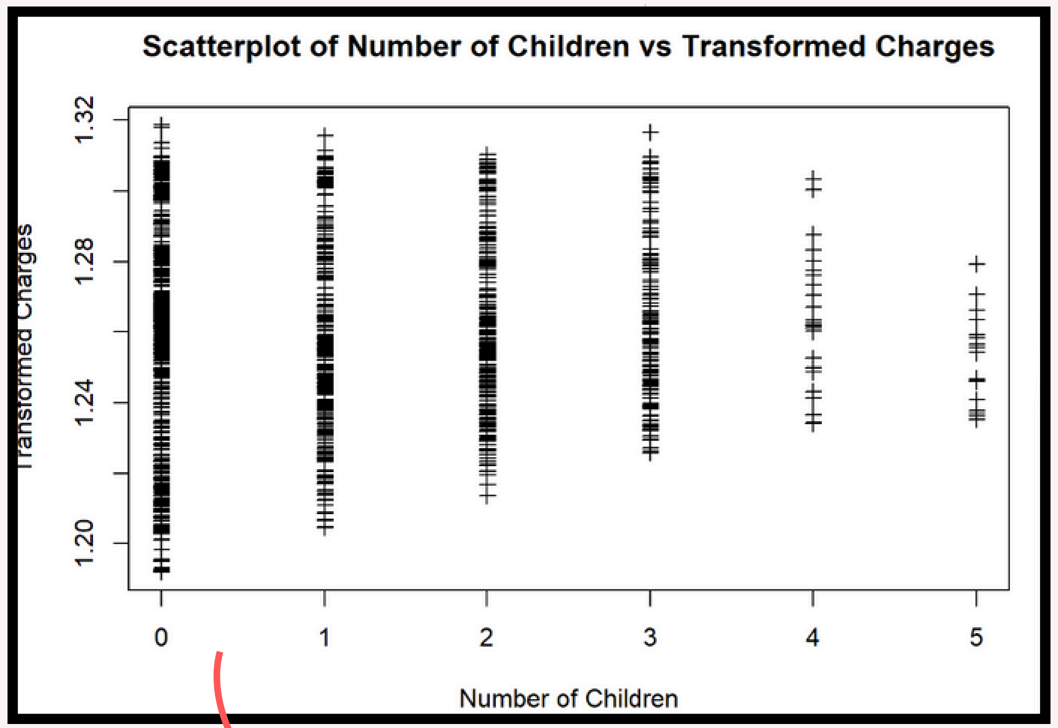
Key Predictors:

Insights



Age Transformation

- Applied transformation to ensure linearity with transformed charges.
- Scatterplot with fitted trend confirms an approximately linear relationship, validating use in linear regression.

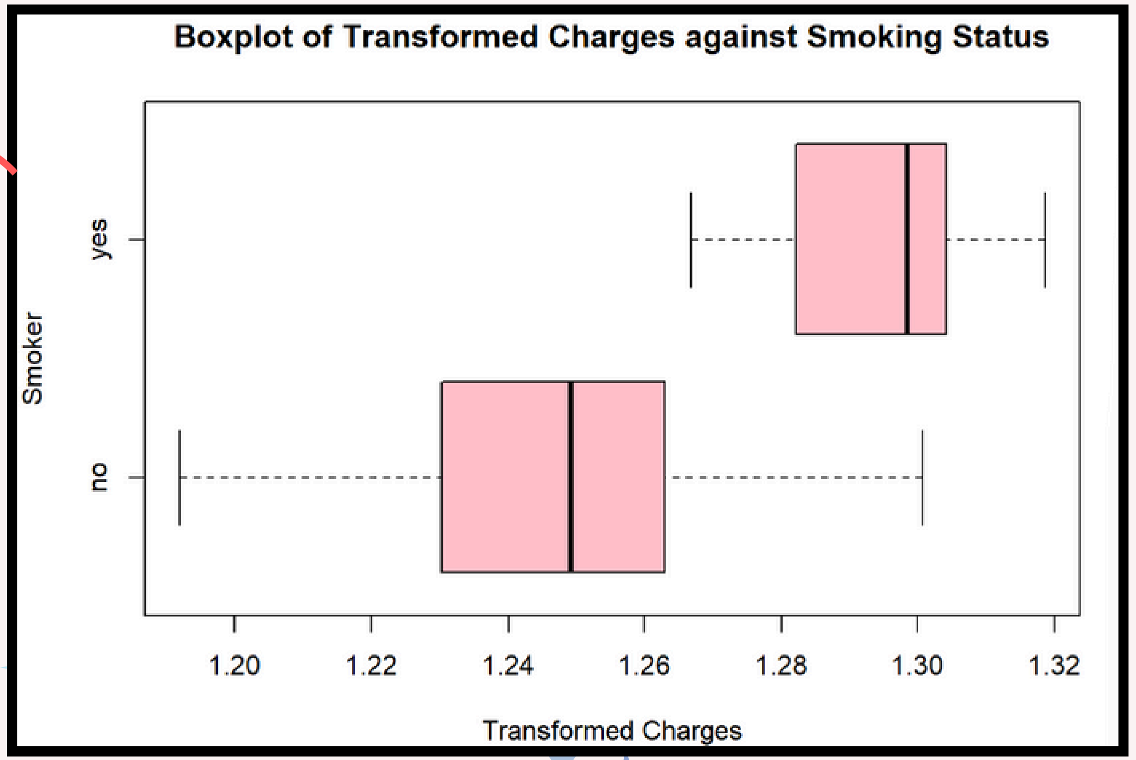


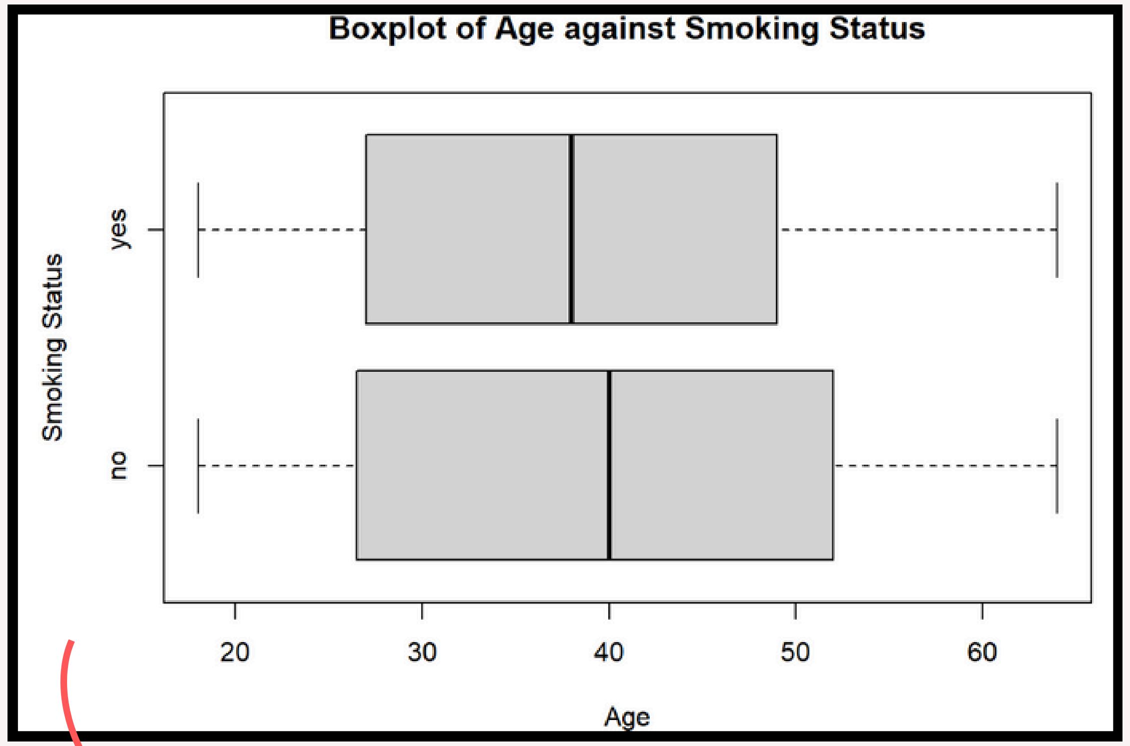
Children

Charges remain stable across numbers of children, with a slight upward trend.

Categorical Variables

- Sex & Region: Similar medians and spreads indicates no association with transformed charges. (Their boxplots were omitted in this slide)
- Smoker: Distinct medians and spreads indicates strong association with transformed charges.





Interaction Term

The median and spread are different in each group. Therefore, age is associated with smoking status.

Baseline Model

Obtained using the response variable, transformed charges, against most strongly correlated variables, smoker status and transformed age

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.155283   0.004283  269.77   <2e-16 ***
smokerno    -0.048976   0.001090  -44.94   <2e-16 ***
age.t        0.089331   0.002681   33.31   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01462 on 1067 degrees of freedom
Multiple R-squared:  0.7406,    Adjusted R-squared:  0.7401
F-statistic:  1523 on 2 and 1067 DF,  p-value: < 2.2e-16
```

p-value < 0.05

- Both are statistically significant predictors at 0.05 level of significance

Adjusted R-squared of 0.740

- 74.0% of variance can be explained by the baseline model

Advanced Model

```
Step:  AIC=-9317.63
charges.t ~ smoker + age.t + children + bmi + region + sex +
    smoker:age.t

              Df Sum of Sq      RSS     AIC
<none>                      0.17354 -9317.6
+ bmi:region   3 0.0007127 0.17283 -9316.0
```

How?

Forward stepwise selection applied on baseline model

- AIC penalises overfitting
- Model with lowest value of AIC (-9317.63) selected

Results

Regression Equation: charges.t = 1.25 – 0.179*smokerno + 0.0204*age.t + 0.00257*children + 0.000431*bmi – 0.00214*regionnorthwest – 0.00385*regionsoutheast – 0.00441*regionsouthwest – 0.00213*sexmale + 0.0834* smokerno*age.t

For an additional unit increase in predictor, there is an β associated increase. on average, holding all other factors constant, where β == regression coefficient estimate for predictor

For categorical variables (smoker, region, sex), this increase occurs in comparison to the values of charges.t at the corresponding reference levels

```
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.2503161  0.0082755 151.087  < 2e-16 ***
## smokerno        -0.1790058  0.0090398 -19.802  < 2e-16 ***
## age.t            0.0203757  0.0051167   3.982 7.29e-05 ***
## children         0.0025735  0.0003265   7.882 7.99e-15 ***
## bmi              0.0004307  0.0000680   6.335 3.51e-10 ***
## regionnorthwest -0.0021368  0.0011236  -1.902 0.057475 .
## regionsoutheast -0.0038532  0.0011296  -3.411 0.000671 ***
## regionsouthwest -0.0044114  0.0011286  -3.909 9.87e-05 ***
## sexmale         -0.0021305  0.0007884  -2.702 0.006993 **
## smokerno:age.t   0.0833601  0.0057706  14.446  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0128 on 1060 degrees of freedom
## Multiple R-squared:  0.8025, Adjusted R-squared:  0.8009
## F-statistic: 478.7 on 9 and 1060 DF,  p-value: < 2.2e-16
```

p-value < 0.05

- Predictors (excluding regionnorthwest) are all statistically significant at 0.05 level of significance
- Region is still a statistically significant predictor and hence is left in the model

```
##              GVIF Df GVIF^(1/(2*Df)) Interacts With
## smoker   1.036683  3        1.006022          age.t
## age.t    1.036683  3        1.006022         smoker
## children 1.009832  1        1.004904             --
## bmi      1.115715  1        1.056274             --
## region   1.106000  3        1.016933             --
## sex      1.014108  1        1.007029             --
```

Adjusted R-squared of 0.801

- 80.1% of variance can be explained by the advanced model

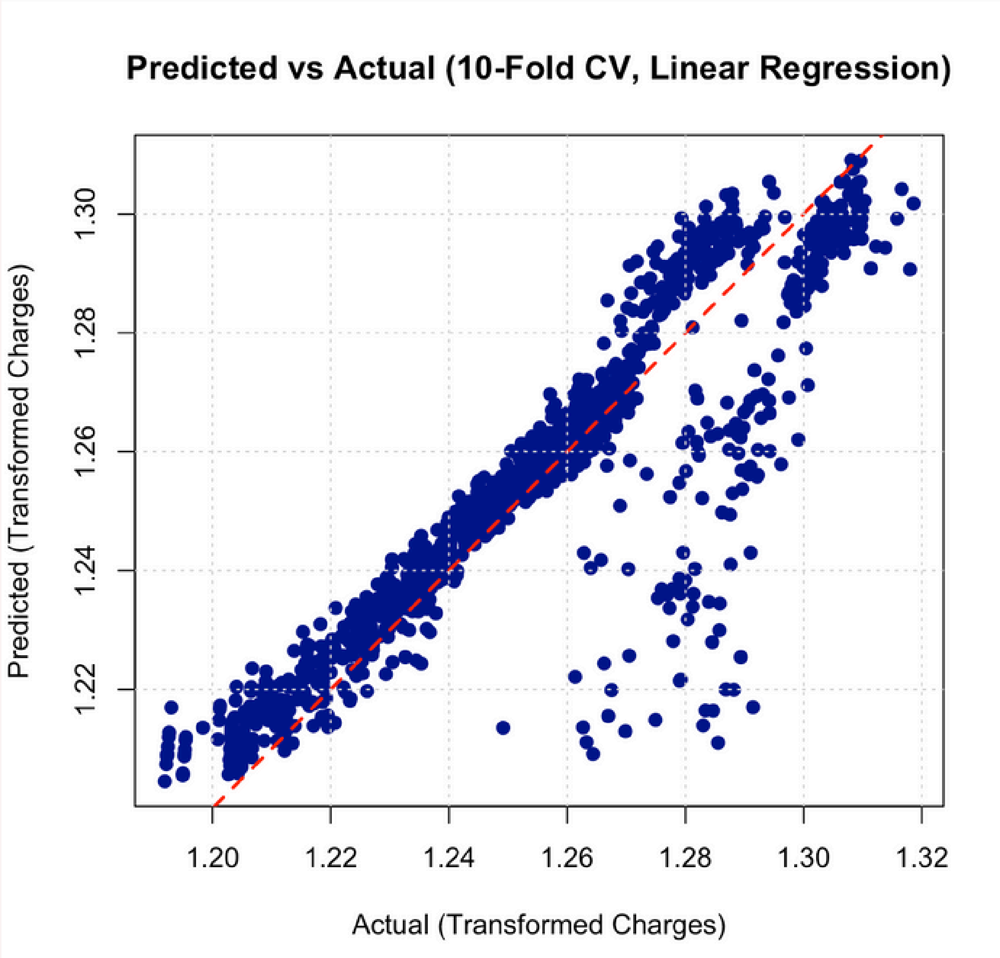
GVIF < 5

- A general threshold of GVIF < 5 is used
- Since GVIF < 5 for all variables, multicollinearity is not an issue here

Linear regression model: Feature Importance and fairness analysis

Feature importance:

- **High R^2 (≈0.82)** and **low RMSE** indicate that the model **explains most of the variability** in transformed medical charges.
- **Smoking** status remains the **single strongest determinant** of insurance charges. Smokers, on average, have substantially higher costs due to the increased risk of chronic diseases.
- **Age** shows a **positive**, **approximately logarithmic relationship** with **charges**. Older individuals tend to have higher predicted costs, consistent with health risk progression.
- The **interaction term (smoker × age)** is **significant**: **smoking amplifies** the **effect of age on charges**, meaning older smokers face disproportionately higher costs.
- **BMI** contributes **moderately.** Higher BMI values are associated with increased medical expenses, reflecting obesity-related risks.
- **Children**, **sex**, and **region** show **relatively minor effects**.



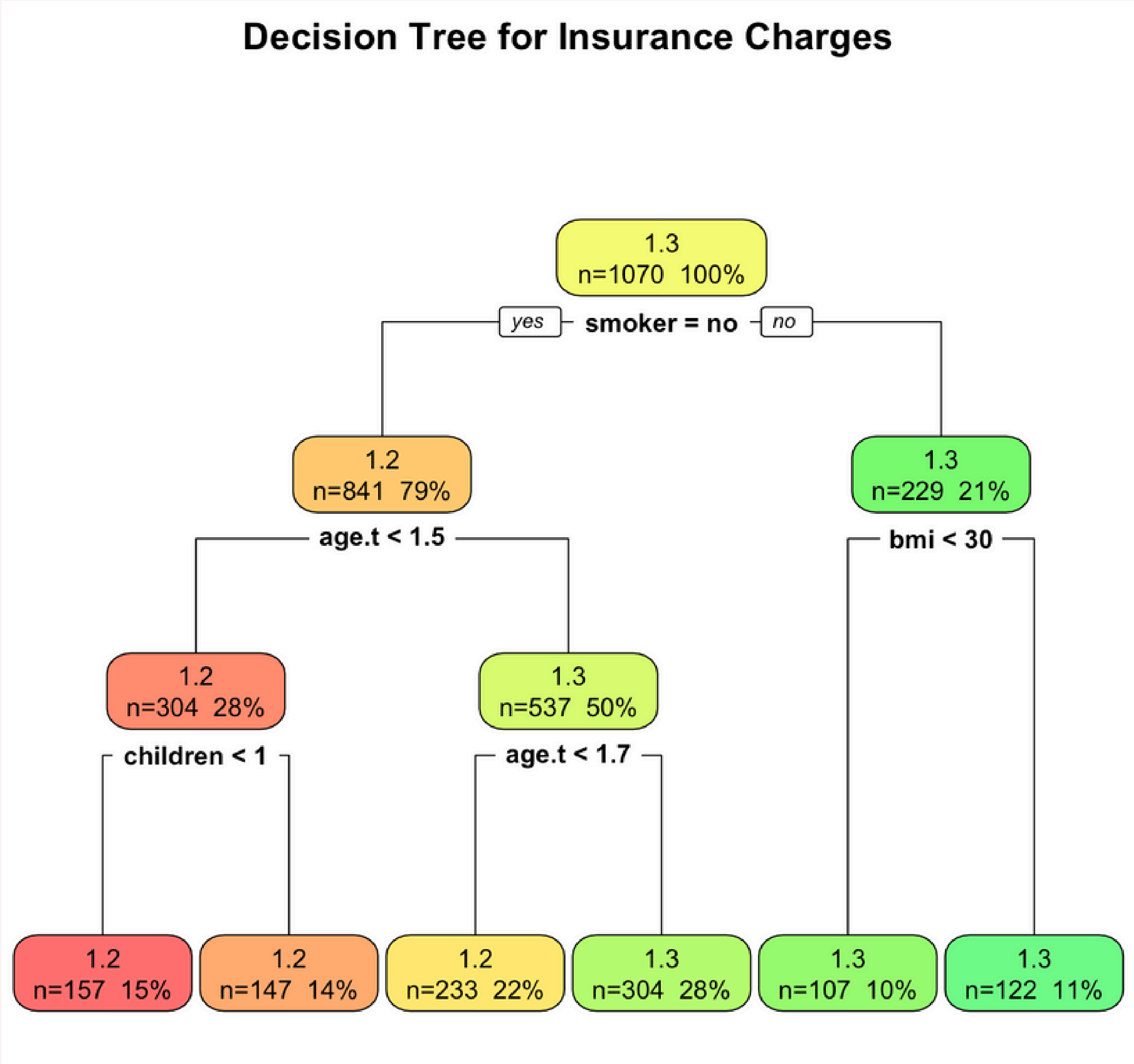
- The plot above shows predicted values **closely aligned along the 45° line**, indicating **strong model calibration** and **consistent** generalisation across folds.
- Residual dispersion is minimal, suggesting **low bias** and **variance**.

Fairness analysis:

- **Gender fairness**
  - From the boxplot of transformed charges by sex, the **median** and **spread** were **nearly identical** across male and female beneficiaries.
  - **No significant difference** in **predicted** or **actual costs by gender**, indicating the model **does not exhibit gender bias** once other variables (e.g. smoking, age, BMI) are accounted for.
- **Regional fairness**
  - Boxplots and bar charts across four regions (northeast, northwest, southeast, southwest) show **relatively similar distributions of charges**.
  - **Slight variations** likely reflect **genuine regional price differences** rather than model bias.
- **Age fairness**
  - Age correlates with medical cost as expected.
  - The **log transformation of age** ensures **proportional**, **not exponential**, increases in predicted charges, reflecting data-driven, clinically valid effects **rather** than **unfair penalisation** of older individuals.

Conclusion: The linear model provides an interpretable, stable framework. Each **feature's contribution** is **monotonic** and **consistent with medical Intuition**: older, smoking, and higher-BMI individuals incur higher expected charges.

Decision tree model: Feature Importance and fairness analysis



Interpretation:

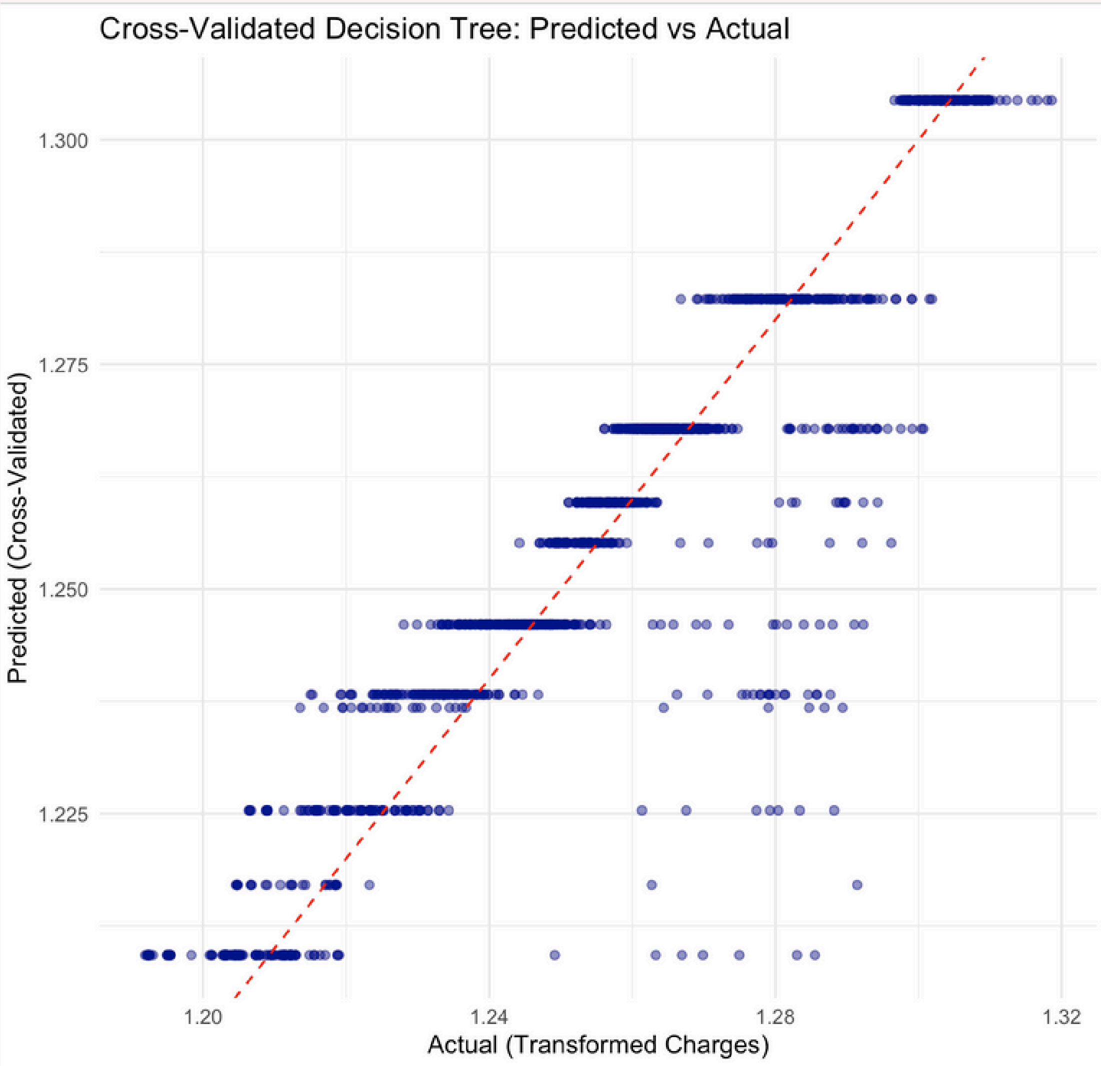
- The tree visually **reinforces the linear model's conclusions** while **exposing non-linear interactions** and thresholds.
- **Smoking status** drives the **largest partition** in data (≈80% of explained variance).
- **Age** and **BMI refine** within-group predictions.
- **Sex** and **region** are **absent** from top splits, implying **minimal predictive influence.**

Feature importance:

- The decision tree **divides** beneficiaries primarily by **smoking status**, followed by **age** and **BMI**, confirming these as the **dominant predictors**.
- Key splits observed:
  - **Root split (Smoker):** The first split separates smokers and non-smokers, underscoring the **large cost differential**.
  - Age threshold (≈1.5 on log-transformed scale): Among smokers, **age further stratifies risk**: older smokers have much higher predicted charges.
  - BMI and children: For non-smokers, **BMI < 30** identifies a group with **lower costs**; for smokers, **having more children** slightly **increases** predicted charges, possibly due to **shared insurance** or **lifestyle effects**.

Fairness analysis:

- **Gender Fairness**
  - **Predicted charges** for **male** and **female** beneficiaries show **almost identical distributions** and **error rates** (MAE and RMSE differ by <2%, $p > 0.05$).
  - Since sex does not appear as a split in the tree, predictions are effectively **gender-neutral**. The model treats both genders **equitably**.
- **Regional Fairness**
  - Across the four U.S. regions, **mean predicted charges vary minimally** (<0.02 on the transformed scale).
  - **Residuals do not differ significantly by region** ($p > 0.05$), and region appears **low** in the tree hierarchy. The model shows **regional parity** as no region is systematically over- or under-predicted.



Cross-validated decision tree performance:

- From the graph above: Predictions **cluster around** the **diagonal**, though **more discretised compared** to the **linear model**. This is a known characteristic of regression trees, which predict by leaf averages.
- The slightly **wider vertical spread** indicates **higher variance** and **marginally lower R^2** compared to the linear model, reflecting that trees sacrifice some accuracy for interpretability and flexibility.

Practical Recommendations (linear regression vs decision tree models)

Target Smoking Behaviour

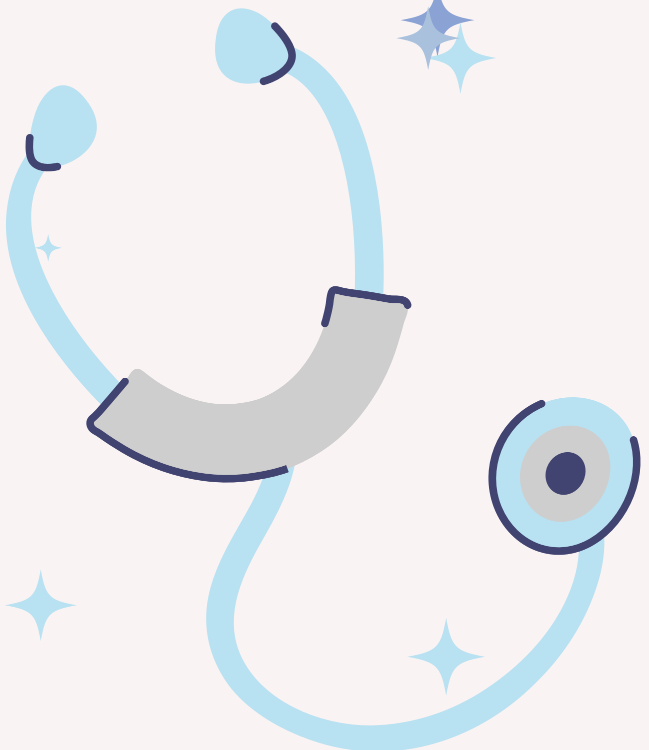
- Both models agree on smoking as the strongest cost driver.
- Implement targeted smoking cessation incentives or premium adjustments to manage high-risk groups effectively.

Age and BMI Risk Adjustment

- Linear regression model highlights steady increases in charges with age and BMI, while the decision tree captures sharp cost jumps at specific thresholds.
- Use age and BMI brackets for fairer pricing tiers.

Model Integration for Policy Design

- The linear regression model offers higher precision and interpretability, while the decision tree captures non-linear risk escalation more naturally but with slightly lower precision.
- Use linear regression for pricing accuracy, and decision trees for policy segmentation and communication.



Difficulties Faced and Methods Used to Overcome

Model Fitting

The dependent variable, charges, was highly right-skewed, violating normality assumptions for linear regression modelling.

We applied a Tukey transformation ($x^{0.025}$) to stabilise variance and achieve approximate normality, improving model fit.

Data Transformations

Several variables, namely BMI and children, showed non-linear relationships with charges.t even after multiple transformations.

They were retained in their original forms for interpretability and to preserve meaningful variable scales after confirming that no simple transformation improved linearity.

Appendix

Code File:

https://github.com/fucheng346/scitsitats

scitsitats (title card)

Medical Insurance Costs