Predicting Outcomes with Data

Su Zheng

June 12, 2025

DATAANA 310 D

Project Overview

- **Dataset Chosen:** Medical Insurance Costs
- **Dataset Source:** Kaggle - Public Domain Datasets
- **Objective:** Build a regression model to predict medical insurance charges.
  - Target: charges *(continuous, numeric variable)*
  - Dependent variables: age, sex, BMI, children, smoker status, and region
- **Research Questions:**
  - Which demographic and lifestyle variables are most strongly associated with predicting insurance charges?
  - To what extent does BMI (body mass index) predict higher insurance costs?
  - Is there a significant difference in medical insurance charges between smokers and non-smokers?
- **Hypothesis:**
  - **H1:** Individuals who smoke will have significantly higher medical insurance charges than non-smokers.

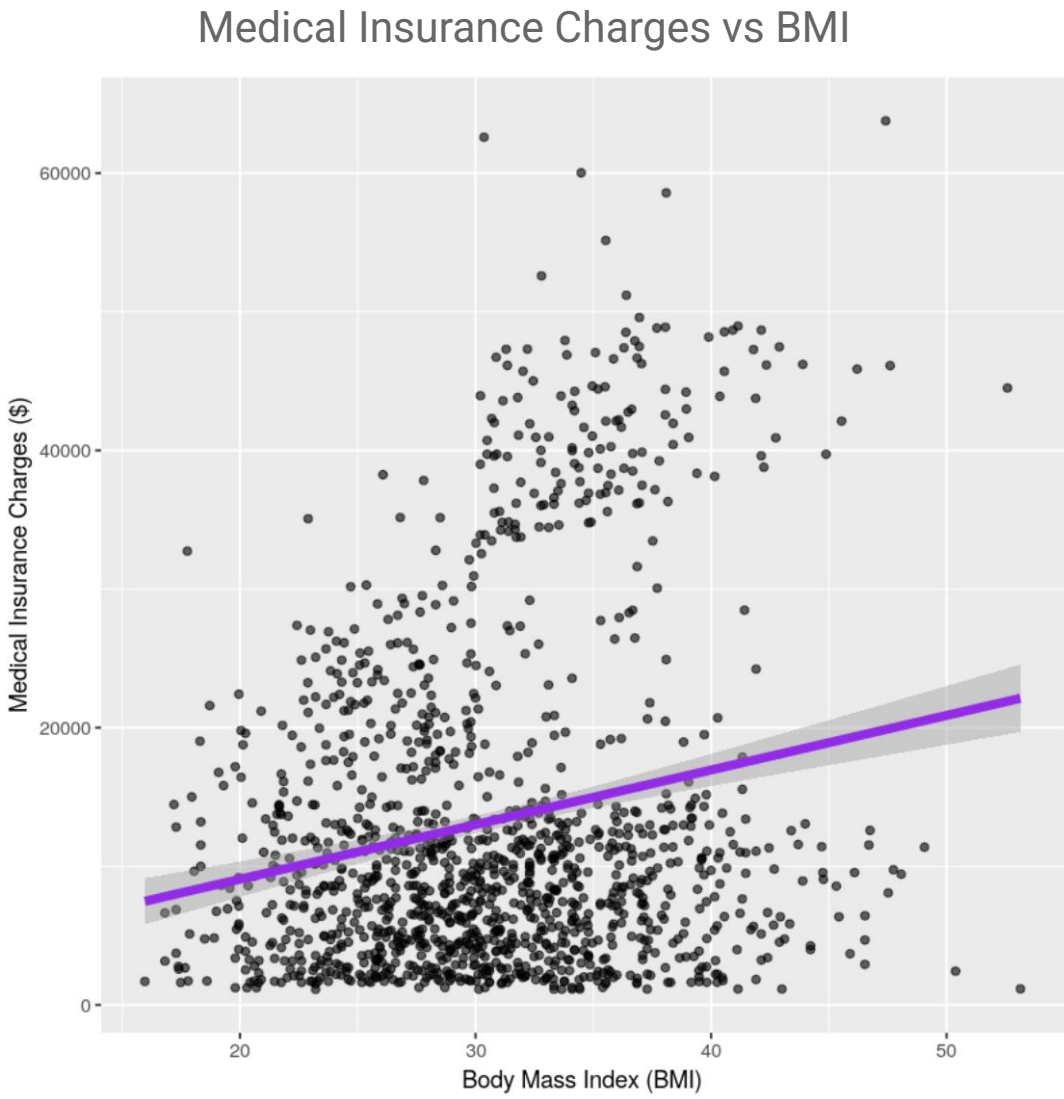
Data Exploration Highlights

- **Demographics Overview:**
  - Fair distribution across **age**, **sex**, and **region**
  - Female vs Male proportions are balanced
  - Median age: 39 years
- **Region:** Each region includes approximately 324 to 364 individuals
  - Indicating even geographic distribution
- **Body Mass Index (BMI):** 16 - 33, notable portion exceeding obesity threshold of 30
- **Smoking Status:** 1,064 non-smokers and 274 smokers
- **Children in Household:** Mean number of children: 1.1
- **Insurance Charges**
  - Range: $1,122 to $63,770
  - Mean exceeds median, indicating a right-skewed distribution
- **Variable Types**
  - Categorical: sex, smoker, region
  - Numerical: age, bmi, children, charges
- **Visualizations Used:**
  - Scatterplot: Charges vs. BMI
  - Boxplot: Charges by Smoker Status

Scatterplot:

Medical Insurance Charges vs BMI

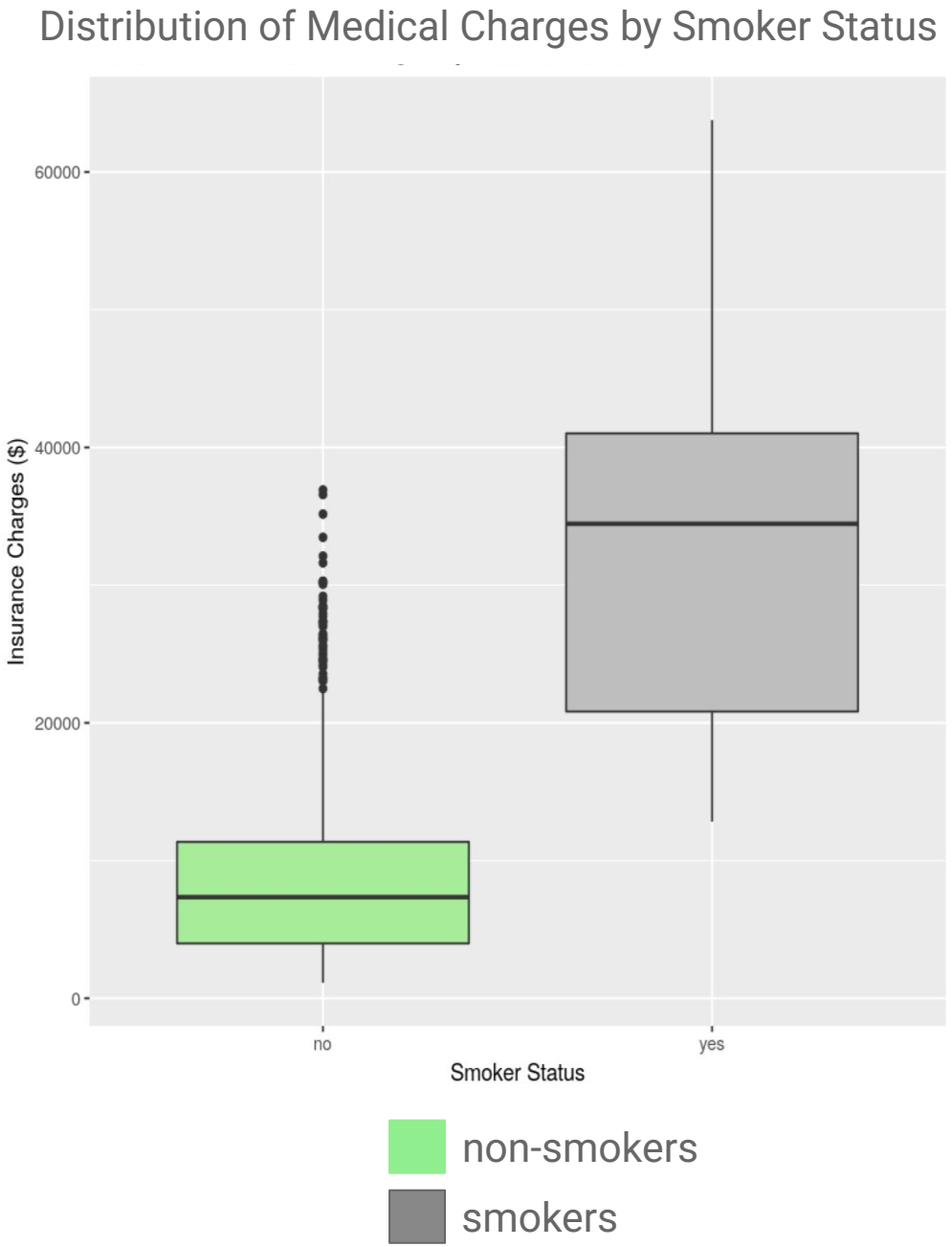
- Weak slope
- Positive, upward trend
- BMI increases → charges increases
- Several upper outliers



Side-By-Side Boxplots:

Insurance Charges ($) of Smokers vs Non-smokers

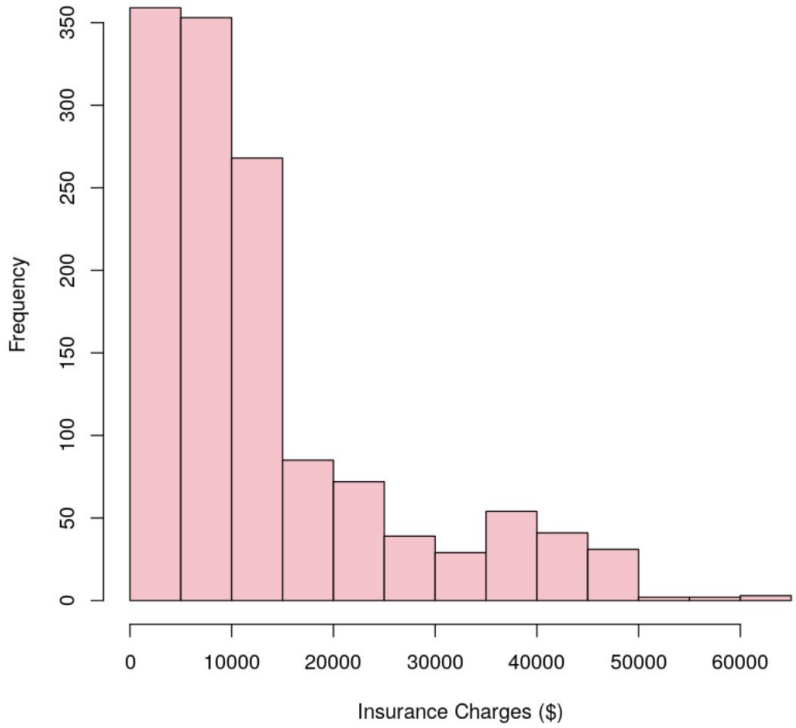
- Smokers show much higher median, IQR, and overall charge

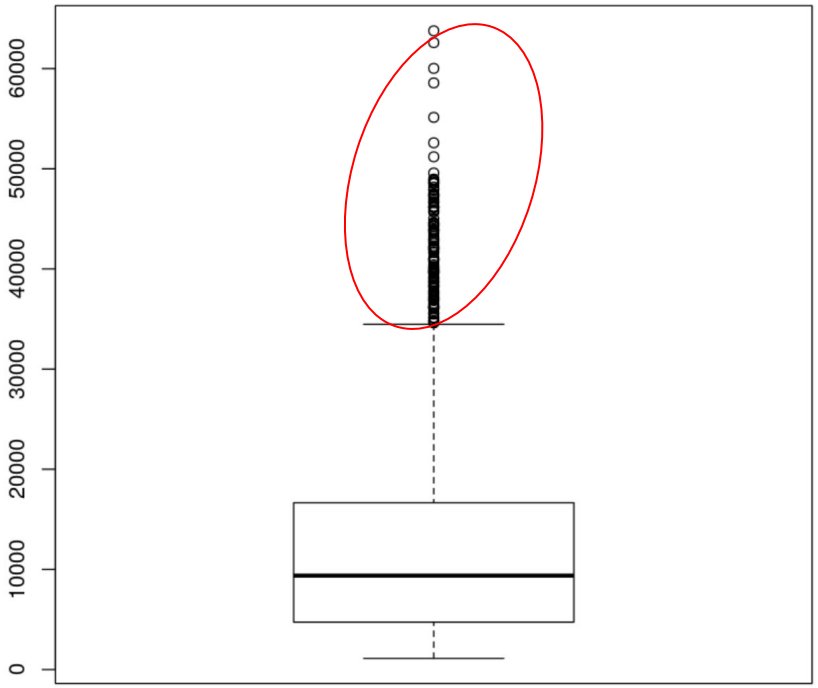


Data Wrangling

- Checked for missing values
- Viewed outliers
  - Used quantile() and boxplots to view distributions
  - BMI - realistic extremes
  - Charges - large variances
- Converted categorical variables to factors
  - Sex, smoker, region

Distribution of Medical Insurance Charges





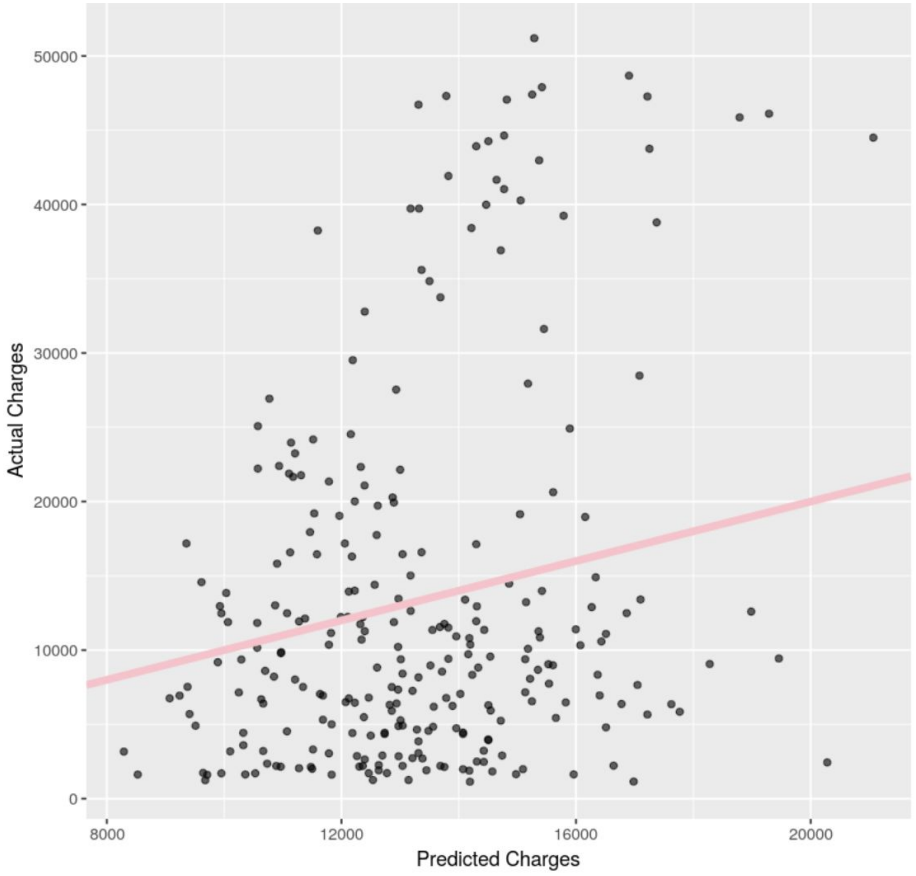
- Both histogram and boxplot illustrate right-skewness
- Several outliers beyond the upper-bound (charges > $34,490)

Model Summary

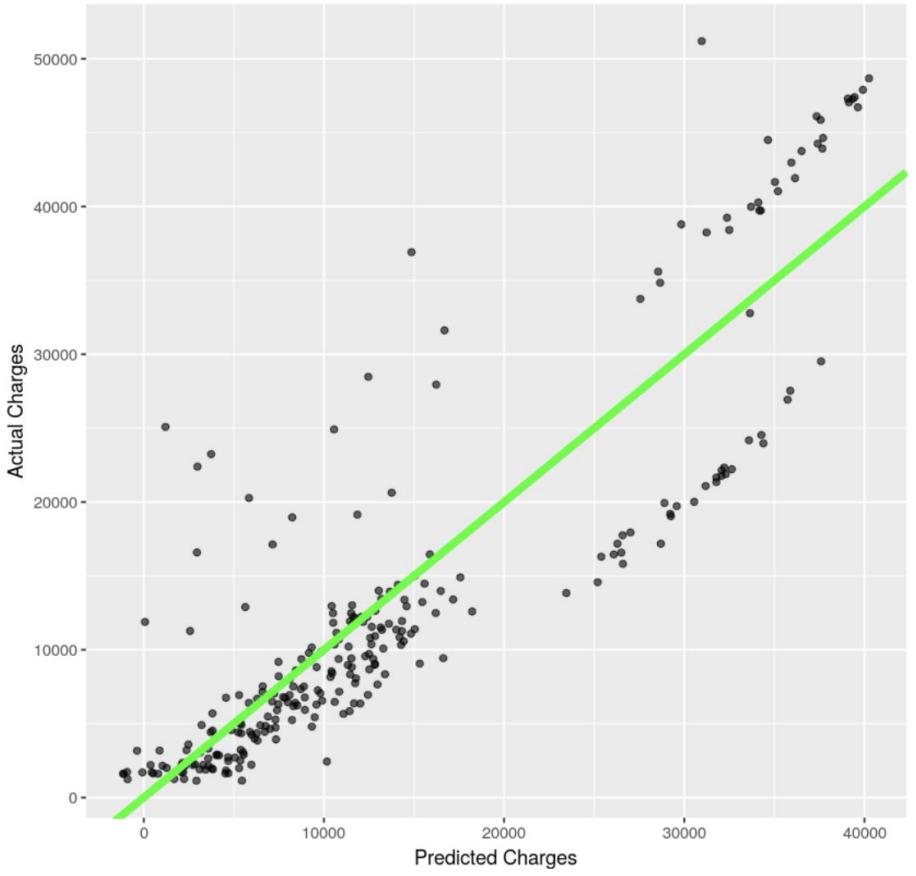
- **Model Chosen:** Linear Regression
  - Model #1 - simple linear regression (only BMI)
  - Model #2 - multiple linear regression (all variables, except Sex)
- **Key Predictors**
  - Strongest predictor = Smoker Status
  - Age, BMI, and number of children also significant variables
  - Sex removed from the model due to its low statistical significance.
- **Model Performance and Key Metrics**
  - **Model 1 - Simple**
    - BMI is a weak predictor alone (low $R^2 \approx 0.07$).
    - RMSE ≈ $12,092
  - **Model 2- Multivariate**
    - Strong relationships between several predictors and charges($R^2 \approx 0.77$).
    - RMSE ≈ $5,972

Visualizing Model Performance: Predicted vs Actual

Model #1 - Simple



Model #2 - Multiple



Experimentation & Analysis

- Removed the region variable in an alternative model to test its impact on prediction accuracy
  - RSME increased only slightly (5,972 to 5,988)
  - Adjusted R-squared decreased by just 0.07%
- Some regions were not statistically significant, so excluding region variable can simplify the model
  - Northwest p-value > 0.05

Key Findings & Insights

- **Statistical Analysis:**
  - Two-sample t-test to compare smokers vs non-smokers
  - **H0:** There is no significant difference in average insurance charges between smokers and non-smokers.
  - **H1:** Smokers have significantly higher average insurance charges than non-smokers.
  - The result is statistically significant. **Reject the null hypothesis.**
- BMI is positively correlated with charges but is a weak predictor alone
- Multivariate model explains 77% of charge variance
- Smoking is the most impactful predictor
  - Increases costs by ~$23,000 on average
- Age, number of children, and BMI also significantly impact charges

Conclusions

Key Takeaways:

- Strong difference in charges between smokers and non-smokers was greater than expected, highlighting the significant cost impact of smoking.
- While BMI alone wasn't a strong predictor, its influence is more meaningful when combined with other demographic and lifestyle features.

- **Challenges/Limitations:**
  - Large range, right-skewed distribution, and high RMSE for charges reduced model accuracy
- **Next Steps:**
  - Include additional lifestyle or demographic features (exercise, diet, occupation, race, etc) for more accurate predictions