US Insurance Charges

Hio Wa (Grace) Mak

10-14-2024

Data

- The dataset used in this project is from Kaggle
- Consists of 7 columns and 1337 observations, with no missing data.

Data Dictionary
Age: The insured person's age.
Sex: Gender (male or female) of the insured.
BMI (Body Mass Index): A measure of body fat based on height and weight.
Children: The number of dependents covered.
Smoker: Whether the insured is a smoker (yes or no).
Region: The geographic area of coverage.
Charges: The medical insurance costs incurred by the insured person.

Primary Objective

- Perform exploratory data analysis (EDA) to explore how demographic and health-related factors are associated with insurance charges

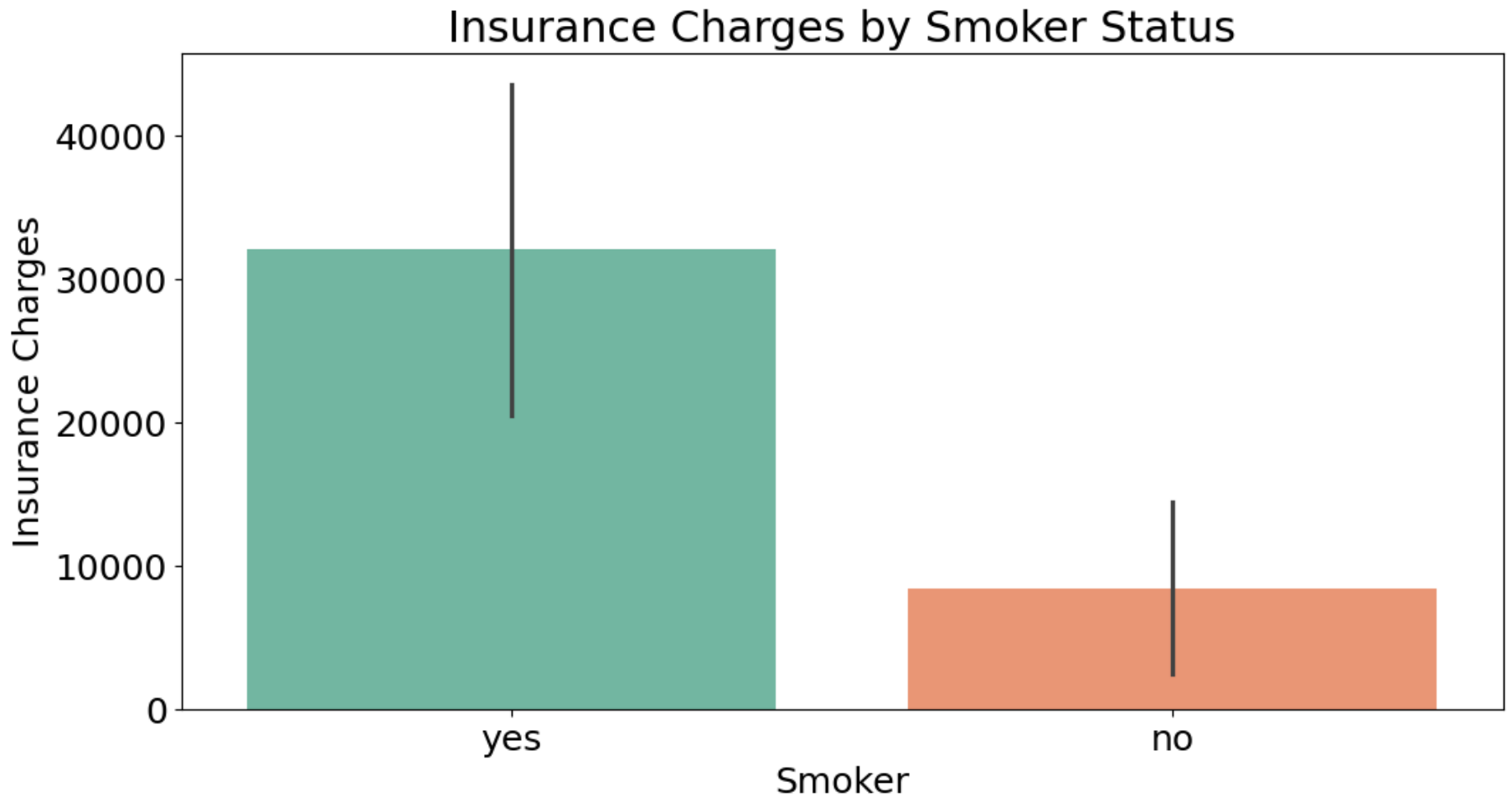
Main Questions

- What demographic and health factors are associated with higher insurance charges?
- Are insurance charges related to the number of risk factors a person has?

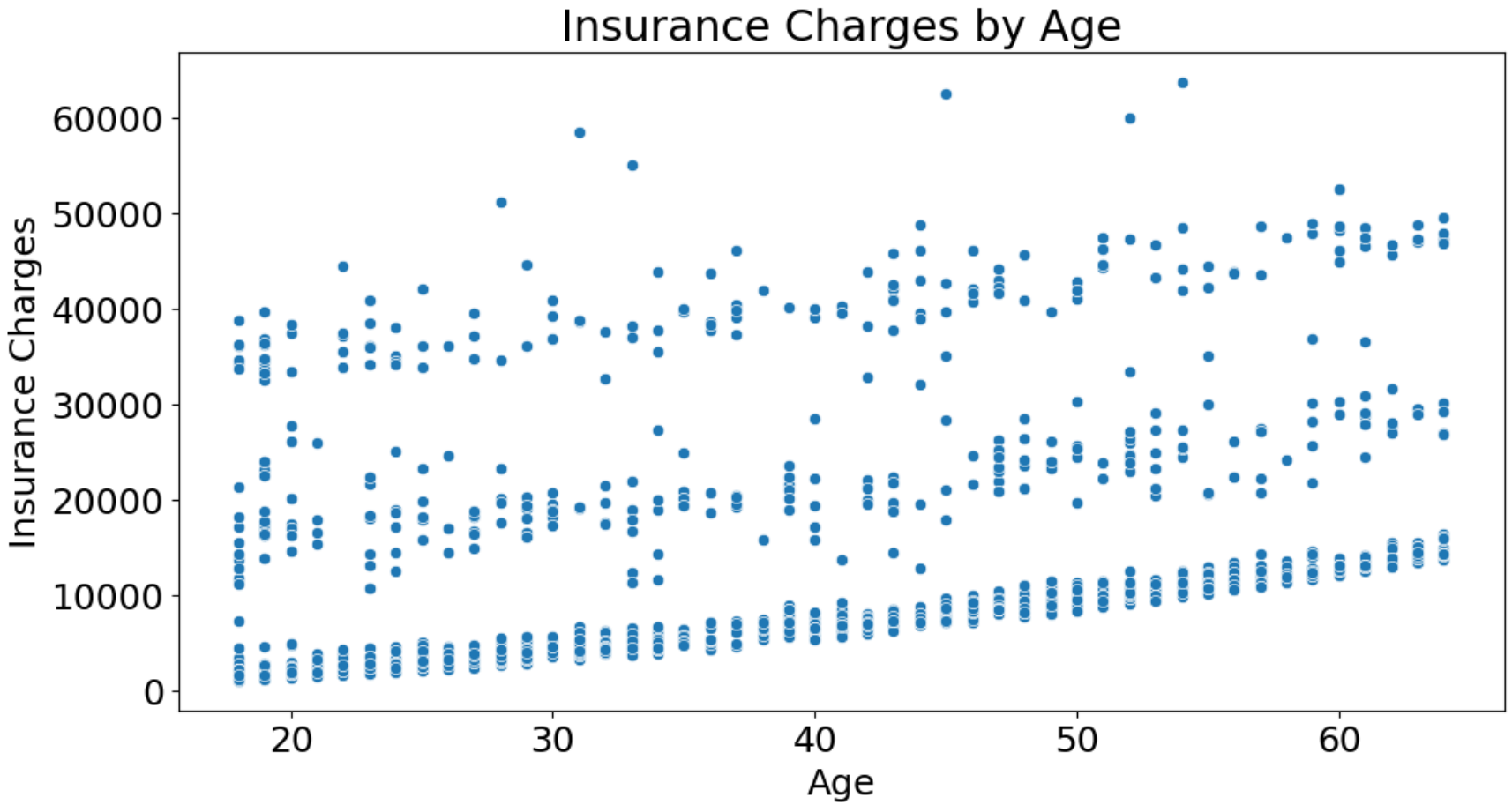
Findings

- Minor differences in insurance charges by sex, region, and number of children
- Major differences in insurance charges by smoker status, age, and BMI

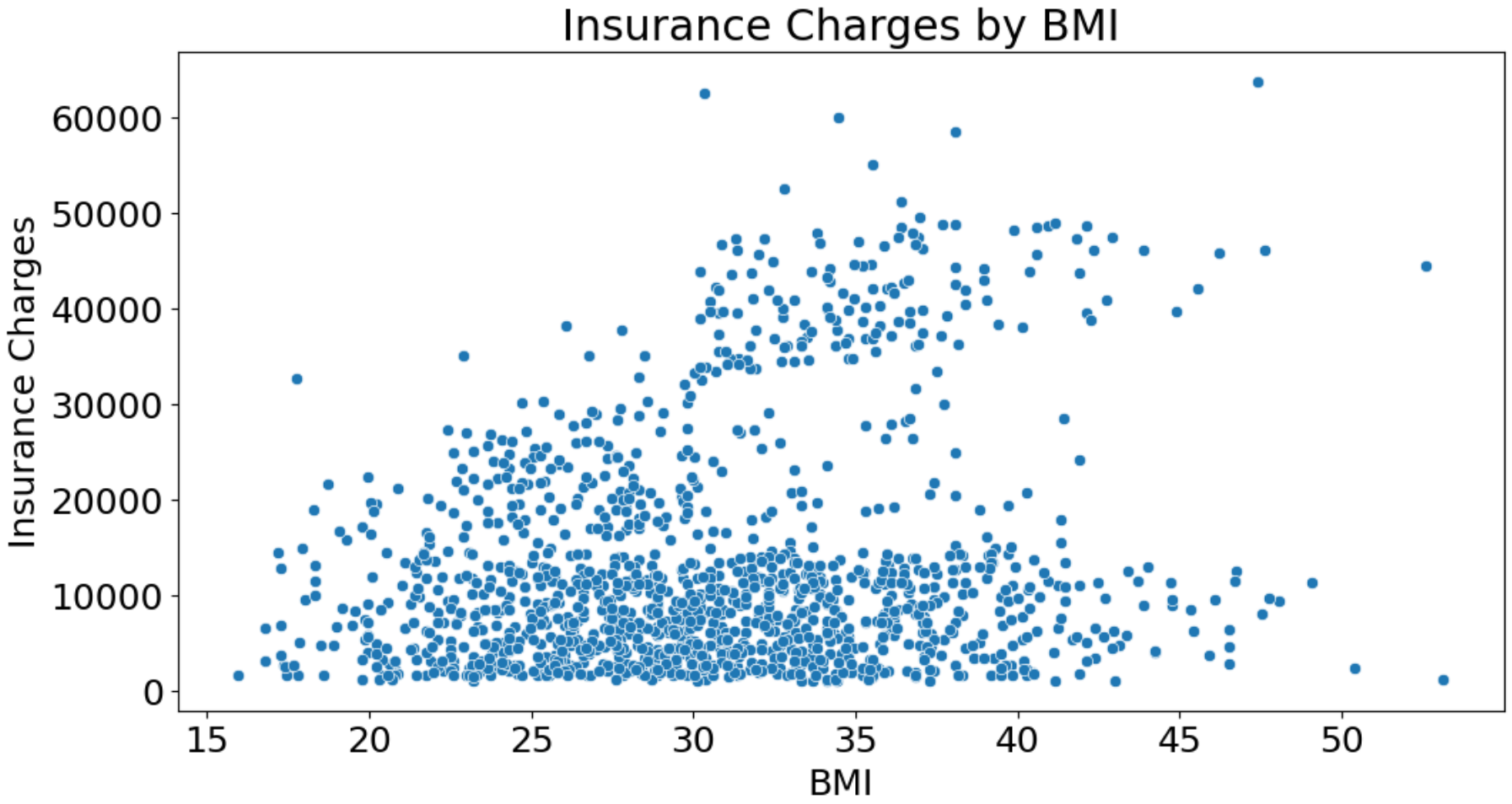
Insurance Charges by Smoker Status



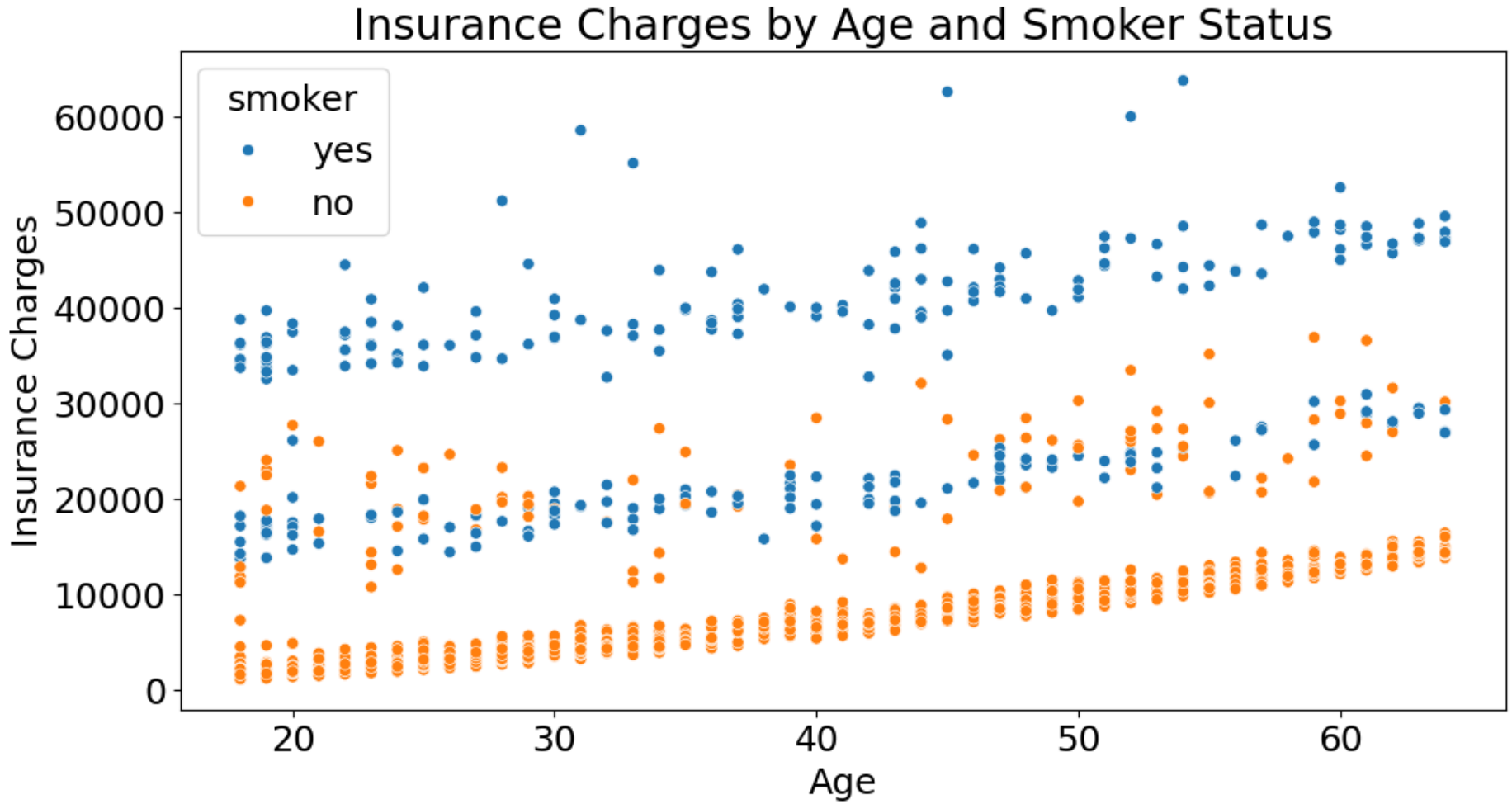
Insurance Charges by Age



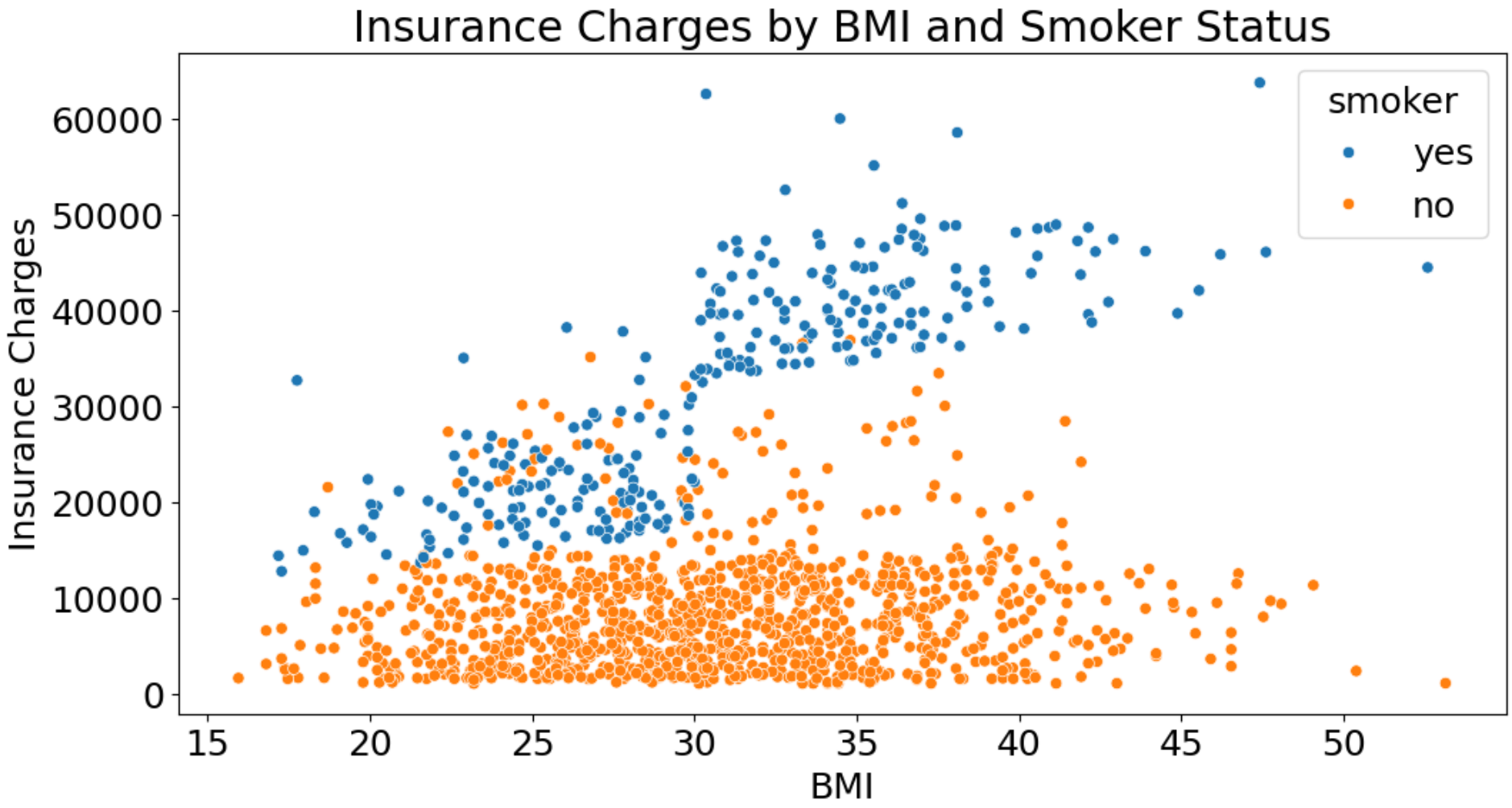
Insurance Charges by BMI



Insurance Charges by Age and Smoker Status



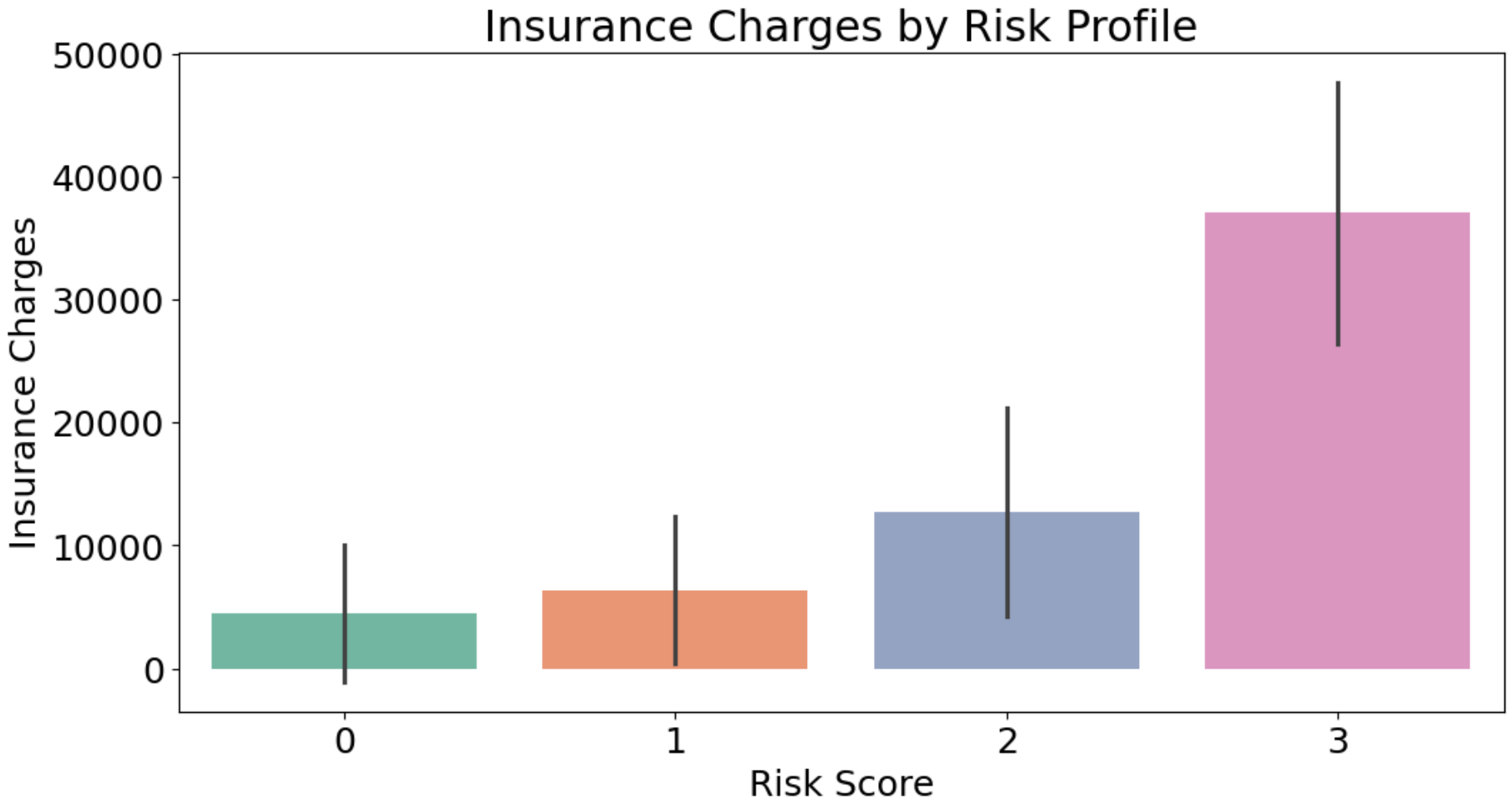
Insurance Charges by BMI and Smoker Status



Constructing a Risk Profile Index

- Individuals classified as underweight, overweight, or obese (i.e., BMI < 18.5 or BMI > 25) received a risk score of 1
- Individuals aged 30 or older received a risk score of 1
- Smokers received a risk score of 1
- **Total risk score** for each person = summing the risk scores
  - Ranges from 0 to 3

Insurance Charges by Risk Profile



Conclusion

- Smokers, older individuals, and those with higher BMI levels tend to incur greater insurance charges
- Insurance charges increased with risk scores, with individuals having the highest risk scores incurring the highest charges
- Insurance companies can adjust plan premiums based on individuals' varying levels of risk.
- **Next steps:** use multiple regression and other machine learning (ML) techniques to predict insurance charges for individuals based on their specific risk profiles (allowing different weights for different risk factors).