Exploratory Data Analysis: Health Insurance Premium Charges

Investigating Risk Underwriting Dynamics

Pruthviraj Santosh Sapate - SCFU123010

Dataset Explanation

The dataset comprises 1338 insured individuals and explores the dynamics of risk underwriting in health insurance. It includes key attributes influencing premiums:

- **age**: Age of the primary beneficiary (18–64 years, mean: 39.21, std: 14.05)
- **sex**: Gender (male/female)
- **bmi**: Body Mass Index (mean: 30.66, std: 6.10; range: 16–53.13)
- **children**: Number of children/dependents (0–5, mean: 1.09)
- **smoker**: Smoking status (yes/no)
- **region**: Geographic region (northeast, northwest, southeast, southwest)
- **expenses**: Individual medical costs billed by health insurance (mean: $13,270.42, std: $12,110.01; range: $1,121.87–$63,770.43)

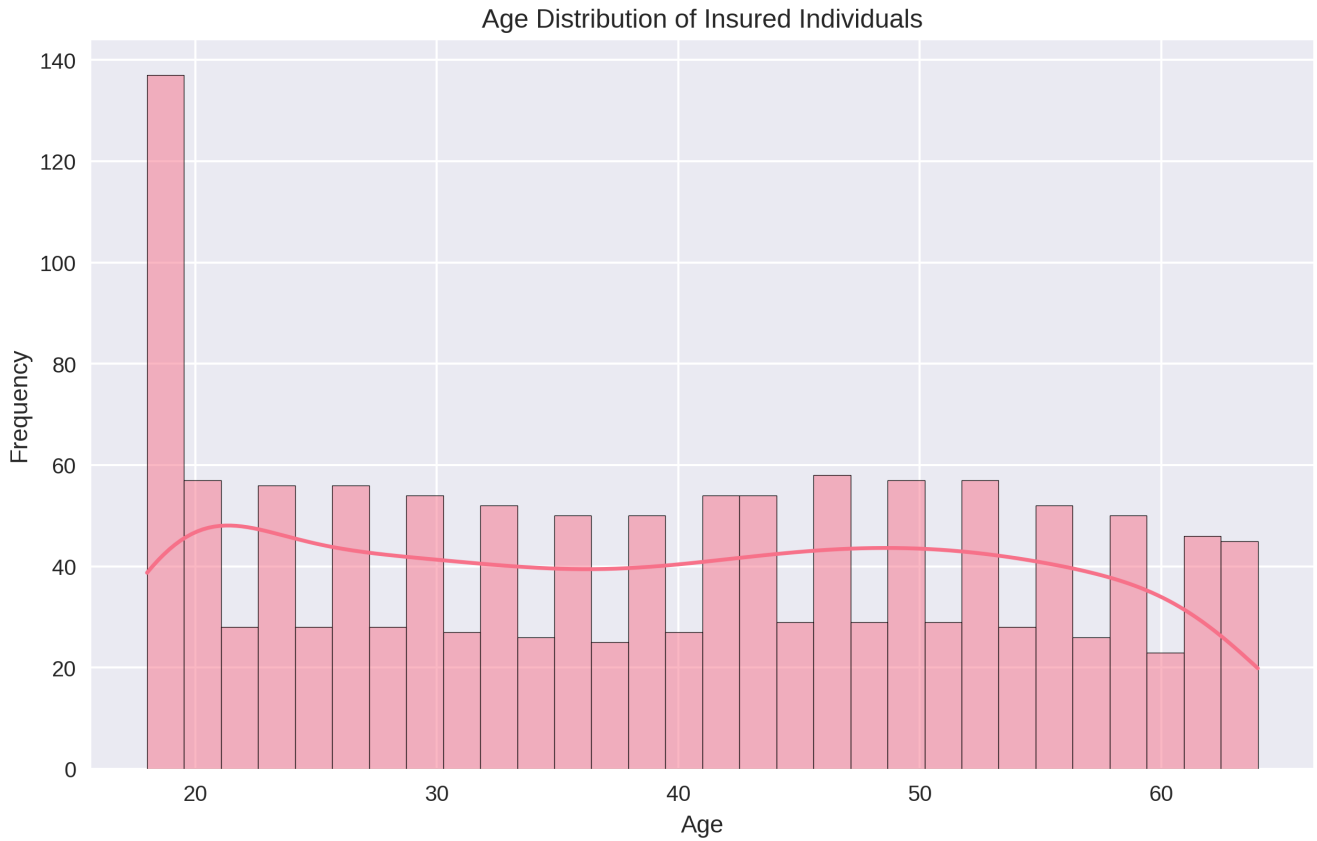
No missing values; mix of numerical and categorical features. Goal: Uncover relationships between attributes and expenses via visualizations.

Sample Data (First 10 Rows)

age	sex	bmi	children	smoker	region	expenses
19	female	27.9	0	yes	southwest	16884.92
18	male	33.8	1	no	southeast	1725.55
28	male	33.0	3	no	southeast	4449.46
33	male	22.7	0	no	northwest	21984.47
32	male	28.9	0	no	northwest	3866.86
31	female	25.7	0	no	southeast	3756.62
46	female	33.4	1	no	southeast	8240.59
37	female	27.7	3	no	northwest	7281.51
37	male	29.8	2	no	northeast	6406.41
60	female	25.8	0	no	northwest	28923.14

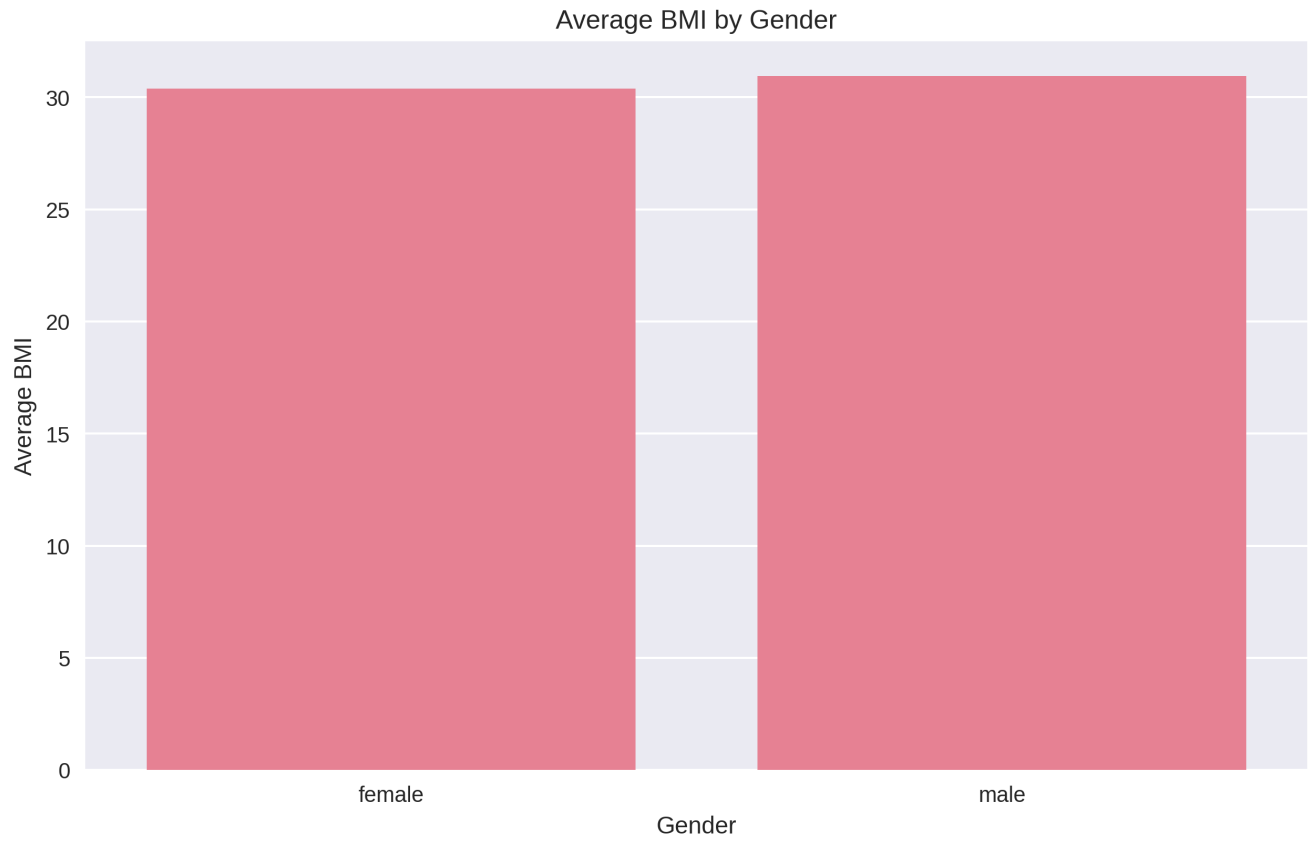
EDA Task 1: Age Distribution

This histogram visualizes the age range and spread of insured individuals, showing a roughly normal distribution centered around 39 years, helping detect demographic skews.



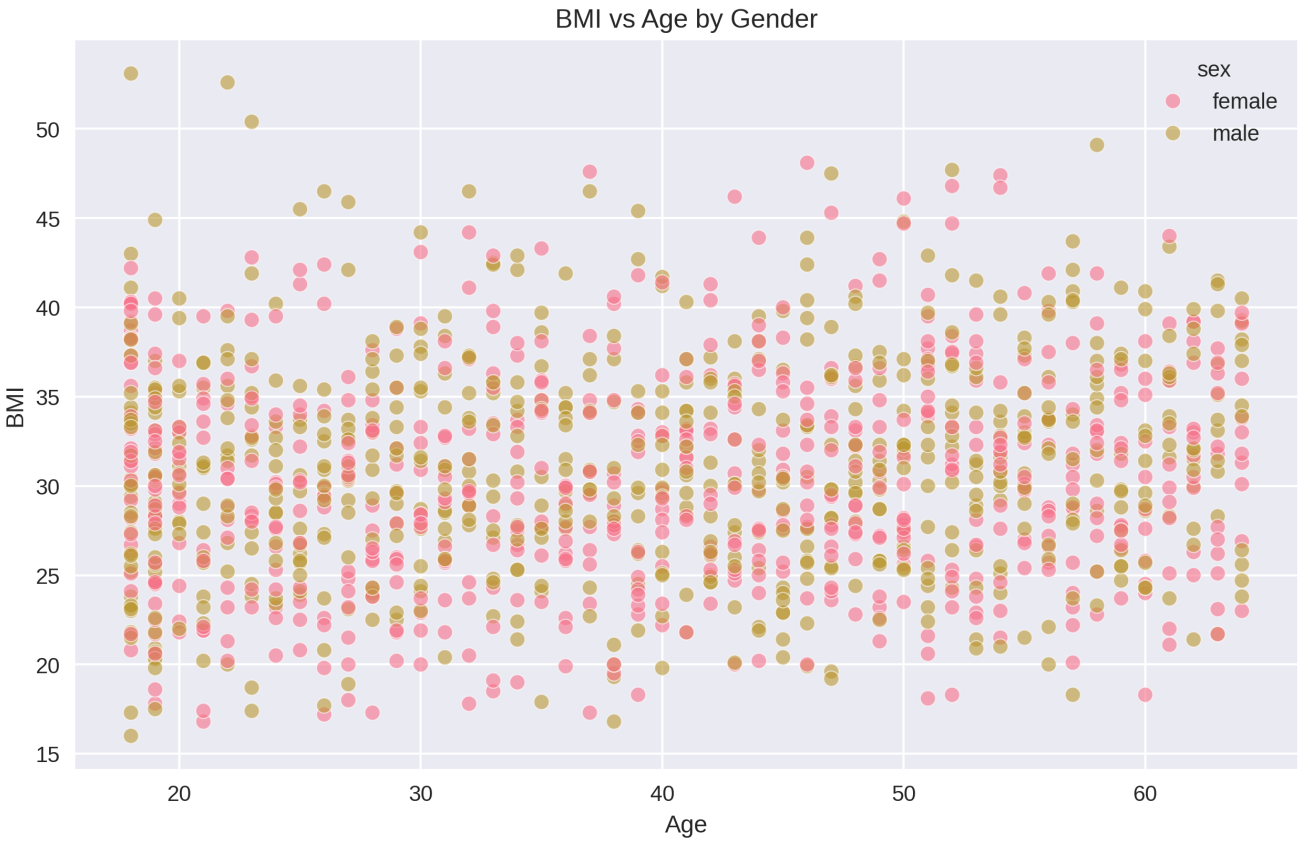
EDA Task 2: Average BMI by Gender

This bar plot compares average BMI across genders, showing slight differences (e.g., males slightly higher). It highlights gender-based health trends influencing risk assessment.



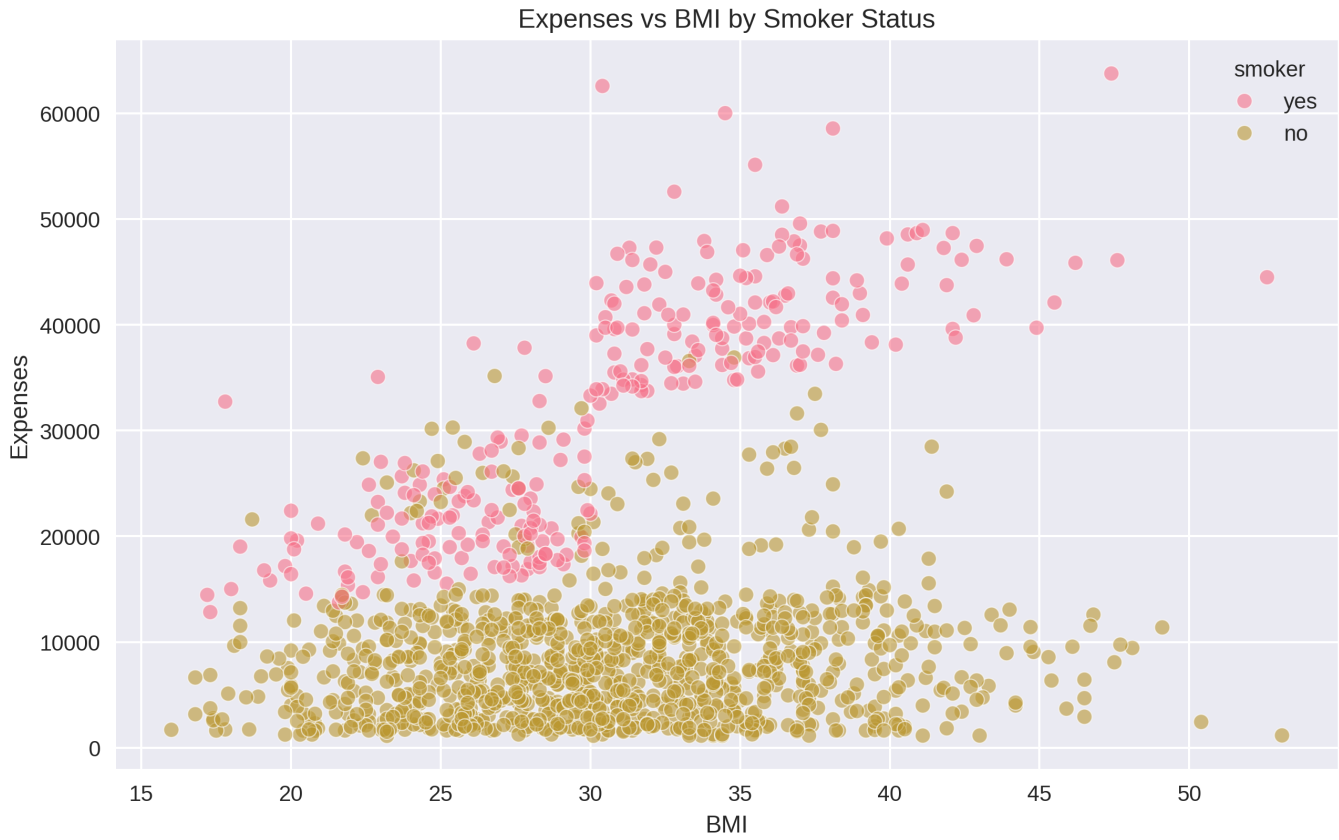
EDA Task 3: BMI vs Age by Gender

This scatter plot examines BMI variation with age for each gender, revealing a gradual increase with age and minimal gender divergence, aiding in age-related risk modeling.



EDA Task 4: Expenses vs BMI by Smoker

This scatter plot shows how smoking status impacts premium costs relative to BMI. Smokers exhibit a steeper upward trend, indicating compounded risk from obesity and smoking.

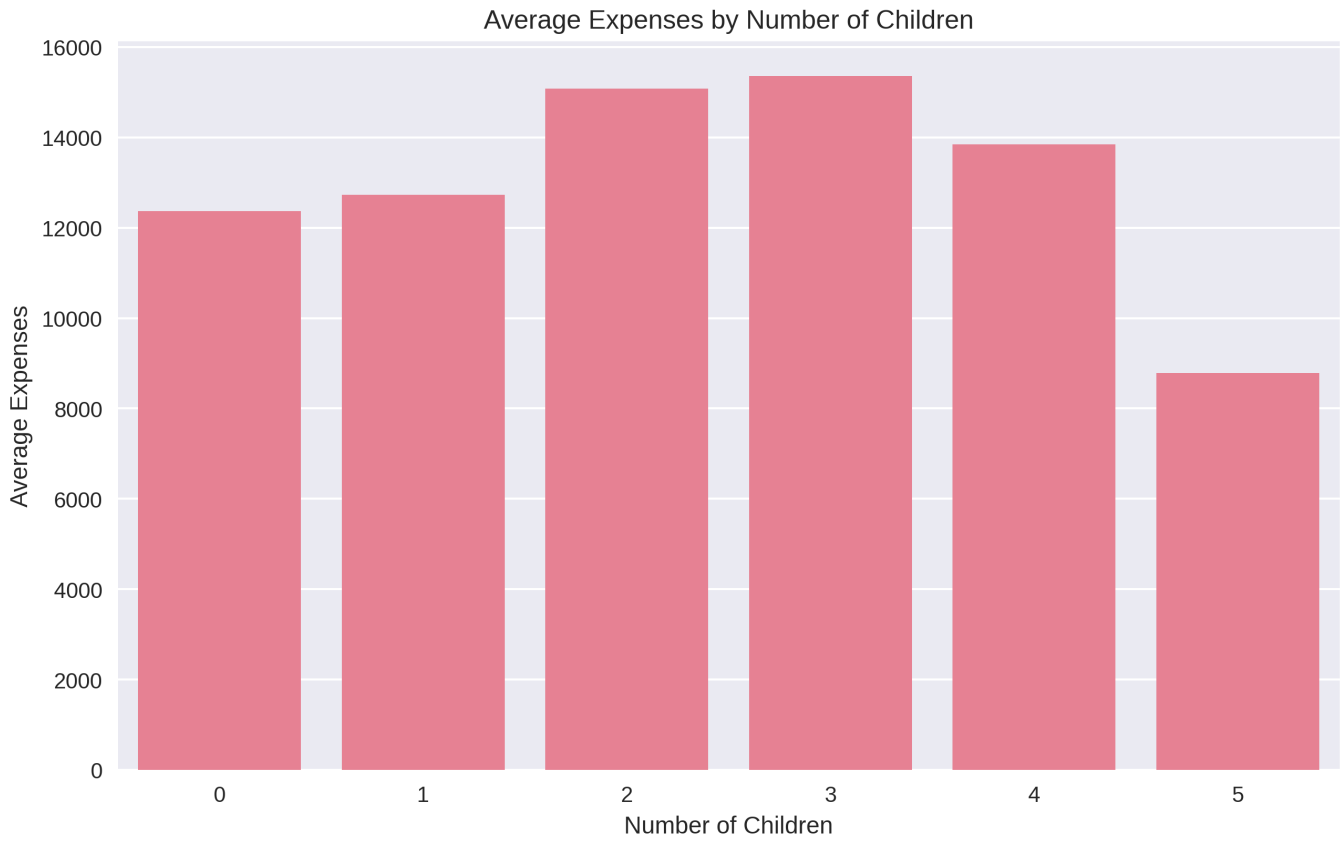


EDA Task 5: Expenses vs Age by Smoker

This scatter plot highlights the combined effect of age and smoking on expenses. Non-smokers show linear growth, while smokers show exponential increases with age.

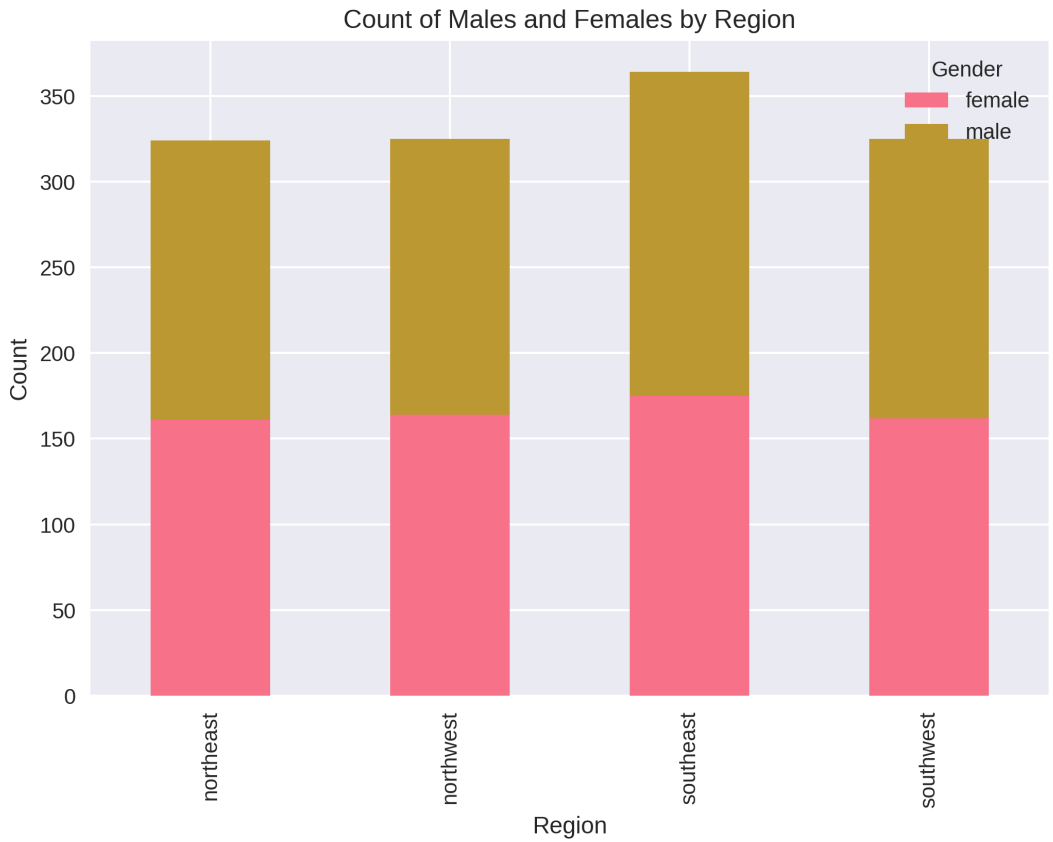
EDA Task 6: Average Expenses by Children

This bar plot examines how dependents affect premiums. A modest increase with more children suggests family size as a mild risk factor.



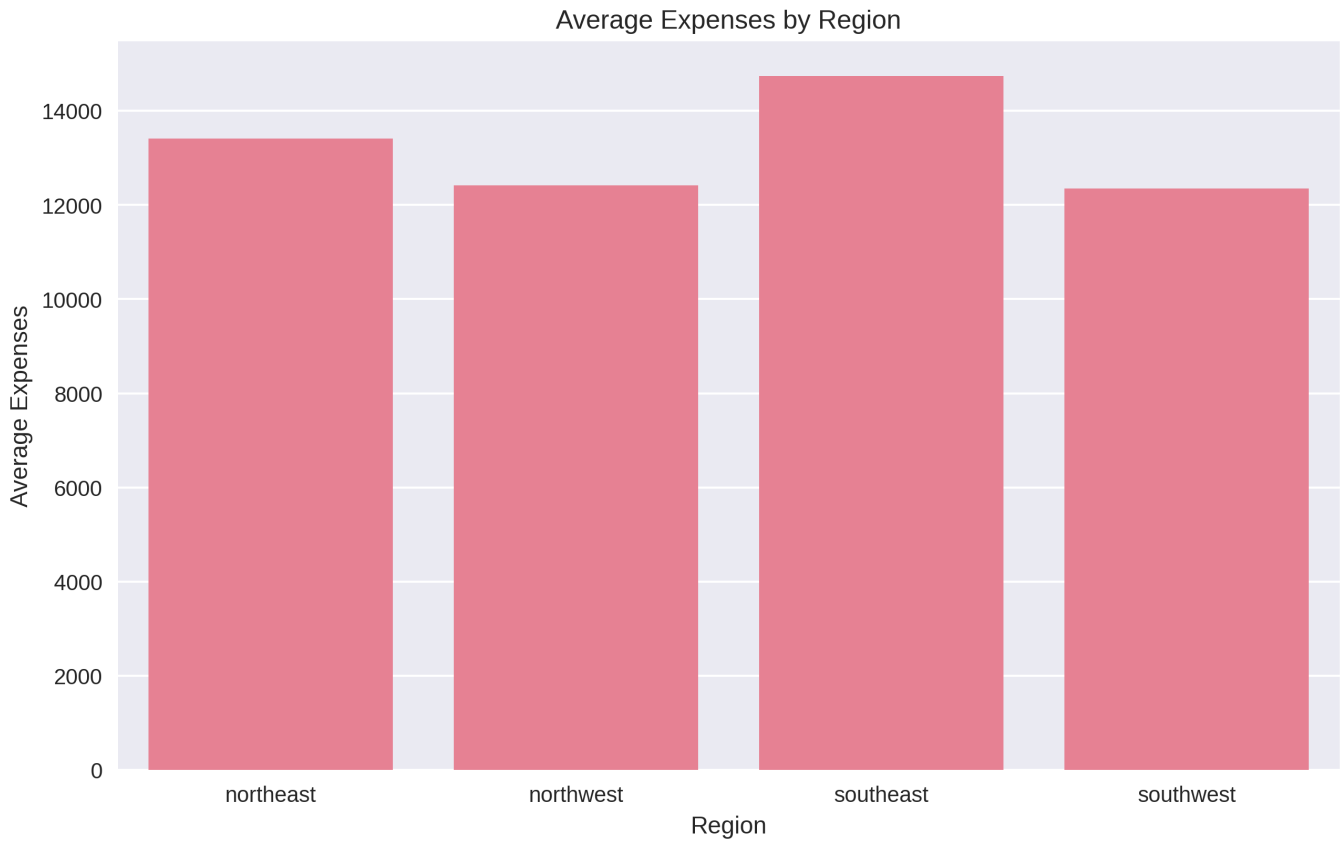
EDA Task 7: Gender Count by Region

This stacked bar plot shows gender distribution by region. Balanced ratios indicate no strong geographic bias in enrollment.



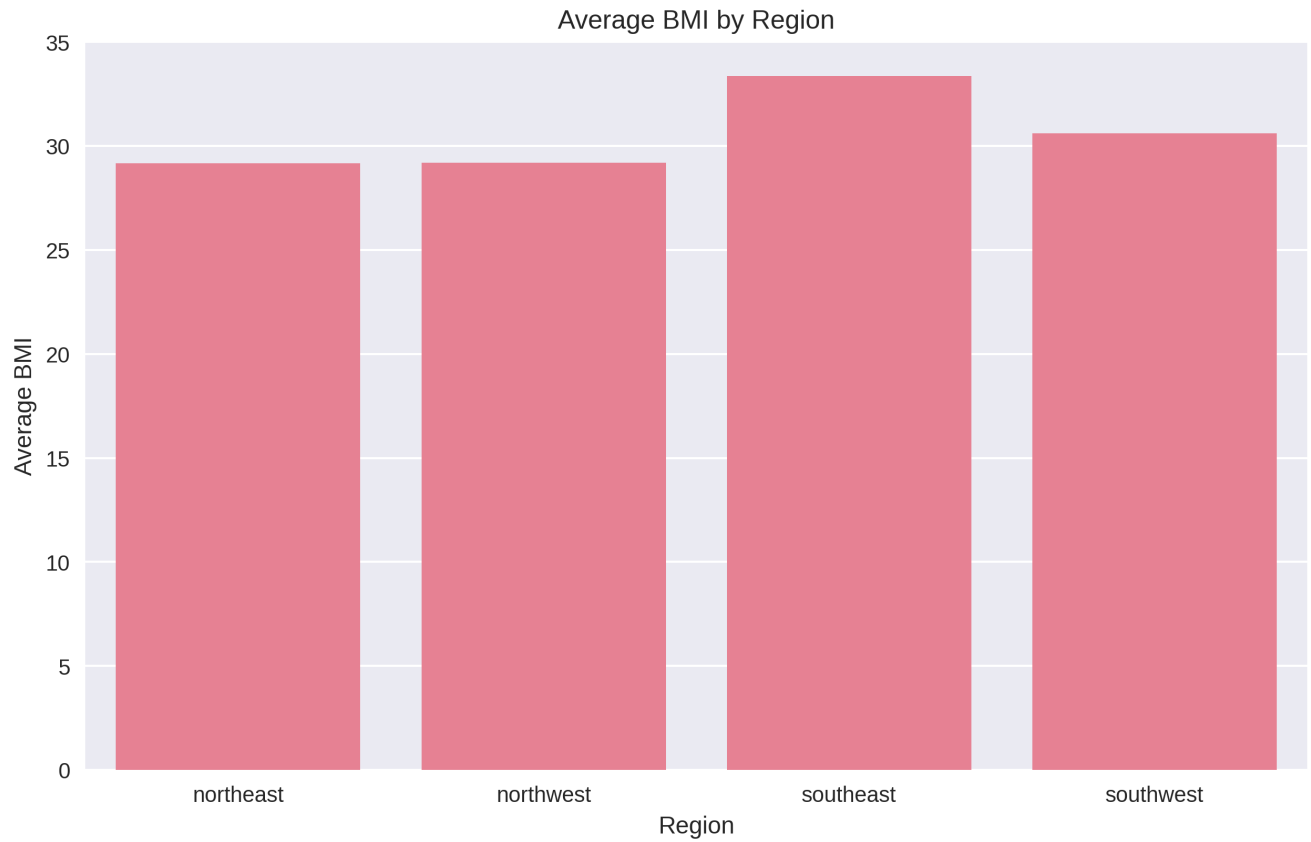
EDA Task 8: Average Expenses by Region

This bar plot reveals regional variations in insurance costs, with the southeast showing higher averages, suggesting regional health cost differences.



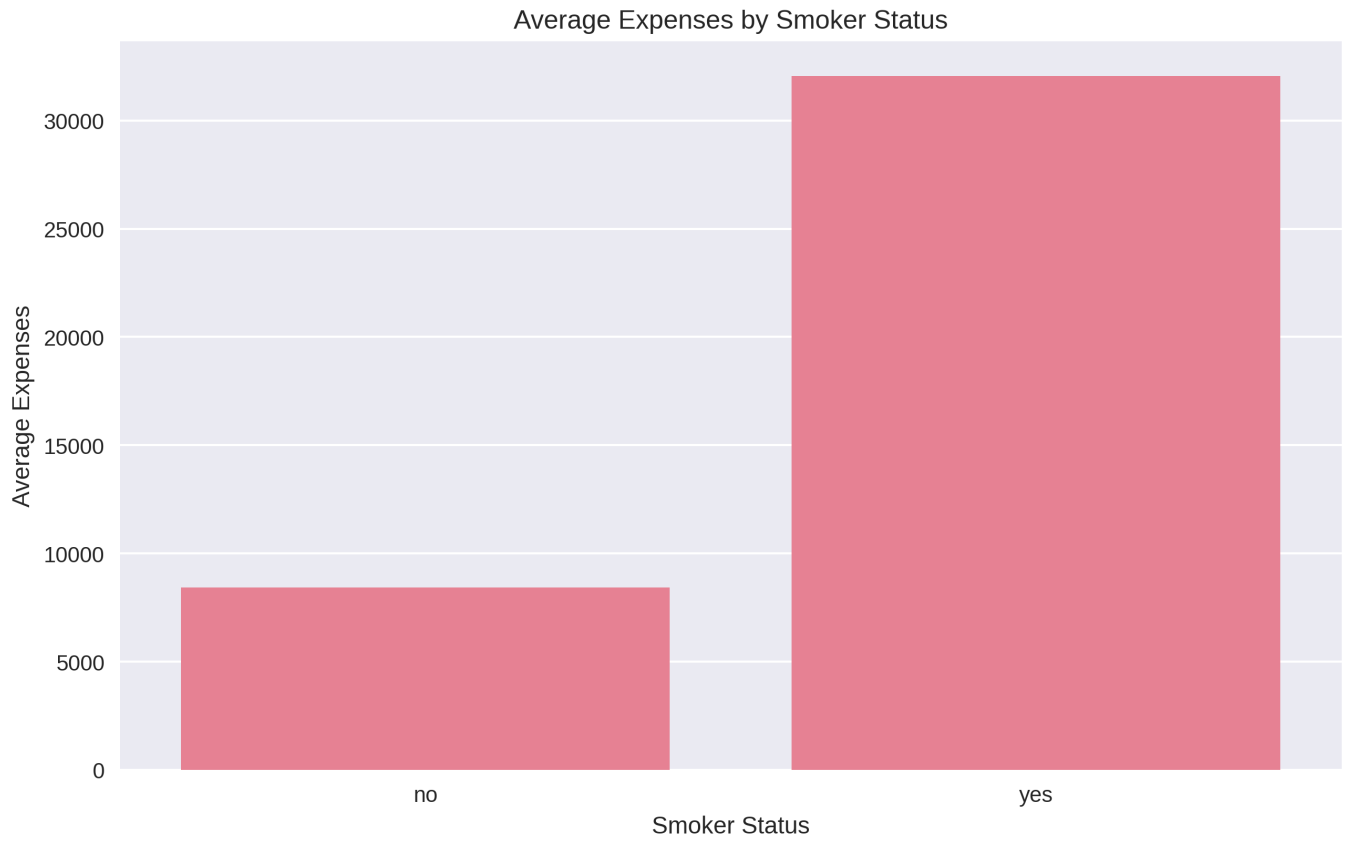
EDA Task 9: Average BMI by Region

This bar plot compares average BMI across regions, showing slightly higher BMI in the southeast, connecting regional lifestyle to premiums.



EDA Task 10: Average Expenses by Smoker

This bar plot emphasizes smoking's strong impact: smokers have 3–4x higher average expenses ($32k vs. $8.4k). Lifestyle is the primary cost driver.



Summary Insights

Main Findings

- Smoking dominates premium variance (150–200% increase).
- Age and BMI show strong positive correlation with expenses.
- Gender and region have minor influence; children mildly affect premiums.

These findings form a foundation for regression modeling and risk prediction.