Data Exploration

Trayvonious Pendleton

WGU

Course Number: D599

STUDENT ID: 011205284

**Part 1: Univariate and Bivariate Statistical Analysis**

**A: Univariate Descriptive Statistics**

**Descriptive Statistics for Continuous Variables**

**Descriptive Statistics for Age:**

|  |  |
| --- | --- |
| Count | 1338.00 |
| Mean | **39.21** |
| STD | **14.05** |
| Min | **18.00** |
| 25% | **27.00** |
| 50% | **39.00** |
| 75% | **51.00** |
| Max | **64.00** |
| Skew | **0.06** |
| Mode | **18.00** |
| Name | **Age** |
| Dtype | **float 64** |

**Key Observations for Age:**

* The distribution is nearly symmetric
* The mode (18) indicates a concentration of younger individuals
* The 50% (39) suggests a balanced spread of ages
* One peak suggests an unimodal distribution
* The symmetric shape confirms a nearly normal distribution

**Descriptive Statistics for Charges:**

|  |  |
| --- | --- |
| Count | 1338.00 |
| Mean | 13270.42 |
| Std | 12110.01 |
| Min | 1121.87 |
| 25% | 4740.29 |
| 50% | 9382.03 |
| 75% | 16639.91 |
| Max | 63370.43 |
| Skew | 1.52 |
| Mode | 1639.56 |
| Name | charges |
| Dtype | float64 |

**Key Observation for Charges:**

* The charges distribution is unimodal
* Positively skewed, right skew
* Mode (1639.56) is above 50%, confirming the right skew

**Relative Frequency for Sex:**

|  |  |
| --- | --- |
| Male | 50.52% |
| Female | 49.48% |

**Key Observation for Sex:**

* The dataset contains almost equal proportions of males and females
* The mode for sex is Female
* The slight difference suggests no significant gender imbalance in the dataset.

**Relative Frequency for Smokers:**

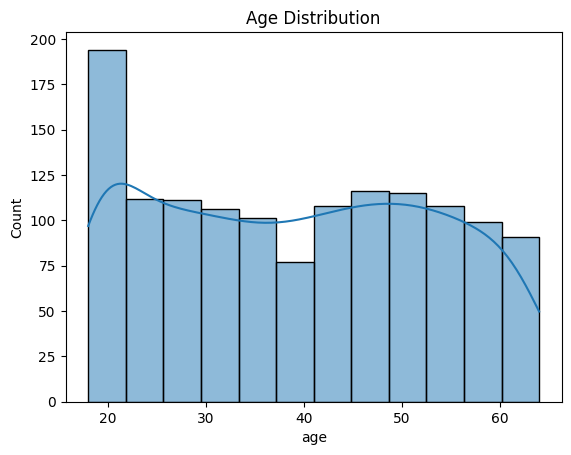
|  |  |
| --- | --- |
| No | 79.52% |
| Yes | 20.48% |

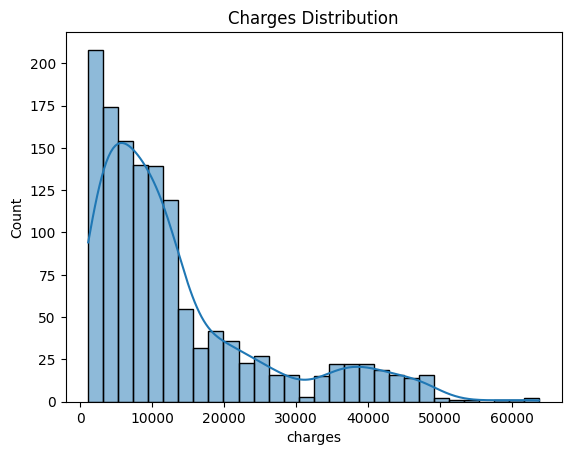
**Key Observation for smokers:**

* The mode for smokers is No
* Most individuals (79.52%) are non-smokers
* Only about 20% of individuals are smokers

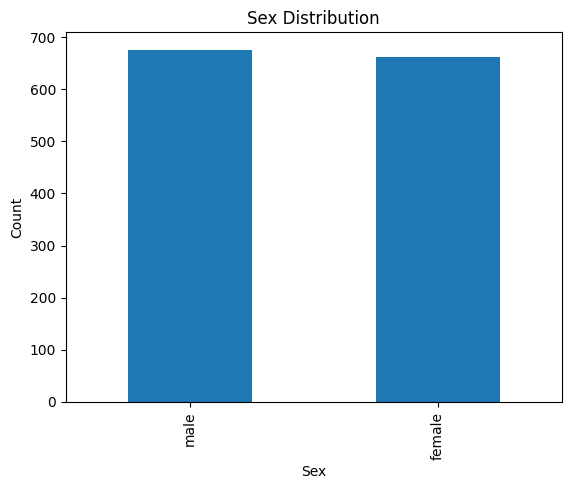
**A1: Univariate Statistical Analysis**

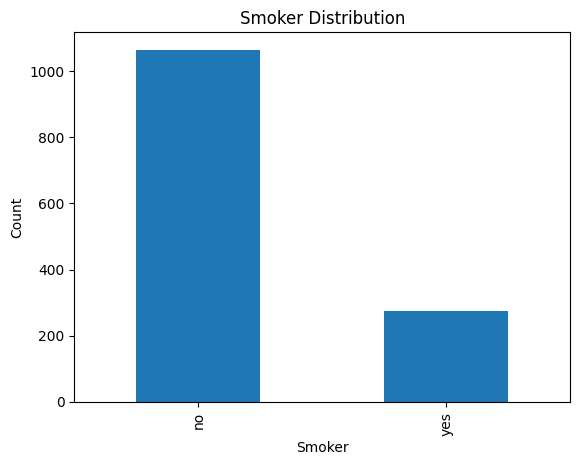
**Two Continuous Variables**





**Two Categorical Variables**

****



**B Bivariate Descriptive Analysis**

**Pearson correlation coefficient and p-value**

|  |  |
| --- | --- |
| Pearson Correlation | 0.198 |
| P-value | 0.000 |

**Key Observations:**

There is a weak but statistically significant positive relationship between BMI and insurance charges. This means that individuals with higher BMI tend to have higher insurance costs, but BMI alone is not a strong predictor of charges. The scatterplot shows much variation, meaning that other factors—such as smoking, age, or pre-existing conditions—might strongly influence medical costs.

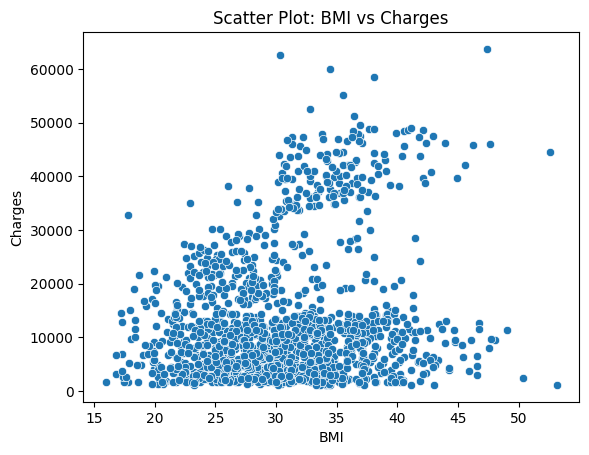
**Chi2-statistic, P-values, and Cramer’s V**

|  |  |
| --- | --- |
| Chi-Square Test | 7.343 |
| P-Value | 0.062 |
| Cramer’s V | 0.074 |

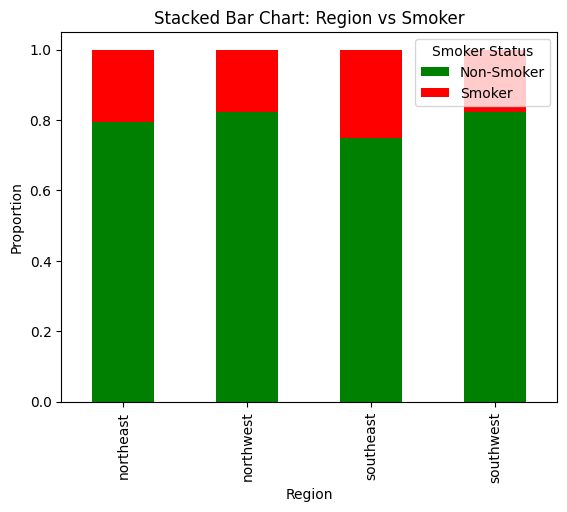
**Key Observations:**

There is no significant relationship between region and smoking status. The Chi-Square test resulted in a p-value of 0.062, indicating that any observed differences in smoking rates across regions are likely due to random chance. Cramér’s V is also only 0.074, confirming a very weak association. This means that an individual’s likelihood of smoking does not meaningfully depend on their geographic region. The stacked bar chart supports this conclusion, as the proportion of smokers and non-smokers appears similar across all regions.

**B1: Bivariate Statistical Analysis**

**Continuous vs Continuous: Correlation Between Charges and BMI**

**Categorical vs Categorical: Smoker Status by Region**



**Part II: Parametric Statistical Testing Research Question:**

**C1: Research Question**

A key research question relevant to the dataset and organizational needs is:  
**"Does BMI significantly differ between male and female policyholders?"**

This question is important for health insurance companies and policy analysts as BMI is a widely used indicator of health risks. Understanding whether BMI varies by gender can help insurers assess potential risk factors associated with different demographics and adjust premium structures accordingly. Additionally, this analysis can inform public health strategies by identifying whether gender-based health interventions may be necessary. By analyzing BMI distribution across genders, insurers can gain data-driven insights into how biological or lifestyle factors may contribute to health-related costs. This research question will be answered through statistical comparison methods, ensuring an evidence-based approach to insurance pricing and risk assessment.

### C2: Relevant Variables

To answer this research question, we will analyze two key variables:

* **Independent Variable (bmi)** – A continuous numeric variable representing the body mass index of policyholders.
* **Dependent Variable (sex)** – A categorical variable indicating the gender of the policyholder (male or female).

Since this analysis involves comparing BMI across two categorical groups, we will determine whether BMI distribution differs significantly between males and females. If the data meets the assumptions of parametric testing, an Independent T-test will be conducted. Otherwise, a Mann-Whitney U test will be used to evaluate whether there is a statistically significant difference in BMI based on gender

D1: Identify a Parametric Statistical Test

To determine whether BMI significantly differs between male and female policyholders, a parametric statistical test is required—provided that the assumption of normality is met. The appropriate parametric test for this analysis is the Independent Samples T-Test

D2: **Develop Null and Alternative Hypotheses**

**Null Hypothesis:** There is no significant difference in BMI between male and female policyholders.

**Alternative Hypothesis:** There is a significant difference in BMI between male and female policyholders**.**

**D3: Conduct the T-Test**

import scipy.stats as stats

# Ensure BMI column is numeric

df['bmi'] = pd.to\_numeric(df['bmi'], errors='coerce')

# Separate BMI data by gender

male\_bmi = df[df['sex'] == 'male']['bmi']

female\_bmi = df[df['sex'] == 'female']['bmi']

# Perform Shapiro-Wilk test for normality

shapiro\_male = stats.shapiro(male\_bmi)

shapiro\_female = stats.shapiro(female\_bmi)

print(f"Shapiro-Wilk Test (Male BMI): p-value = {shapiro\_male.pvalue:.3f}")

print(f"Shapiro-Wilk Test (Female BMI): p-value = {shapiro\_female.pvalue:.3f}")

# Perform Levene’s Test for equal variance

levene\_test = stats.levene(male\_bmi, female\_bmi)

print(f"Levene’s Test p-value: {levene\_test.pvalue:.3f}")

# Perform T-Test (Welch’s T-Test if variances are unequal)

t\_stat, p\_value = stats.ttest\_ind(male\_bmi, female\_bmi, equal\_var=(levene\_test.pvalue > 0.05))

print(f"T-Test Results: T-Statistic = {t\_stat:.3f}, P-Value = {p\_value:.3f}")

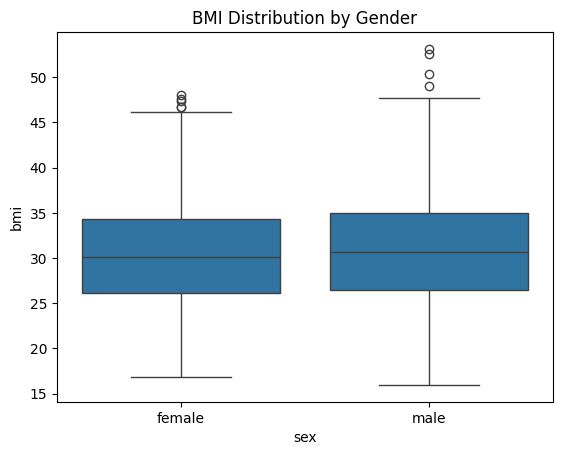
# Boxplot for BMI by gender

sns.boxplot(x=df['sex'], y=df['bmi'])

plt.title('BMI Distribution by Gender')

plt.show()

**D4: Results of the T-Test**



**Shapiro-Wilk Test (Male BMI): p-value = 0.003**

**Shapiro-Wilk Test (Female BMI): p-value = 0.004**

**Levene’s Test p-value: 0.950**

**T-Test Results: T-Statistic = 1.697, P-Value = 0.090**

**E1: Justification for Choosing the Statistical Test**

The Independent Samples T-Test was selected to analyze whether BMI significantly differs between male and female policyholders. This test is appropriate because it compares the means of a continuous variable (bmi) across two independent groups (sex: Male and Female). Since we are determining whether gender impacts BMI levels, the T-test allows us to assess if the average BMI of males is significantly different from that of females. For the test to be valid, the assumptions of normality and homogeneity of variances must be met. The Shapiro-Wilk test is used to check if BMI is normally distributed for each gender, while Levene’s test assesses whether the variance in BMI is equal between the two groups. If both assumptions hold, we use the standard Independent T-Test. However, if BMI is not normally distributed, a Mann-Whitney U test, a nonparametric alternative, will be used instead. This statistical approach ensures that the comparison is robust and appropriate for evaluating whether gender differences in BMI exist. Understanding this relationship can provide insights for health insurance pricing strategies, risk assessments, and public health initiatives.

**E2: Discussion of Test Results**

The **Shapiro-Wilk test results** indicate that BMI is **not normally distributed** for both males (**p = 0.003**) and females (**p = 0.004**), since the p-values are **less than 0.05**. This confirms a **violation of the normality assumption**, which is a key requirement for conducting a parametric **Independent T-Test**. Despite this, the **Levene’s test p-value (0.950)** suggests that the **variances of BMI between males and females are equal**, meaning the assumption of homogeneity of variances is met.

The Independent T-Test results show a T-statistic of 1.697 and a p-value of 0.090. Since the p-value is greater than 0.05, we fail to reject the null hypothesis. This means that there is no statistically significant difference in BMI between male and female policyholders at the 5% significance level.

Looking at the boxplot of BMI distribution by gender, the median BMI values for males and females appear to be very close, further supporting the test results that BMI does not significantly differ between genders. The spread of BMI values is also similar, with a few outliers present at the higher BMI range for both genders.

Although the T-test fails to show a significant difference, the violation of normality suggests that a nonparametric test (Mann-Whitney U test) would be more appropriate to validate this conclusion. However, based on these results, gender does not appear to be a major determinant of BMI among policyholders, and BMI-based risk assessments in insurance policies may not need to vary significantly based on gender alone.

**E3: Implications for Stakeholders**

The Independent Samples T-Test, despite its limitations due to the violation of normality, provides a statistically rigorous method for comparing BMI between male and female policyholders. The results from this test directly benefit various stakeholders within the organization, including insurance analysts, policy underwriters, risk assessment teams, and healthcare strategists.

For insurance analysts and underwriters, the findings help in evaluating whether gender should be considered as a factor in BMI-related insurance pricing. Since the test results indicate no significant difference in BMI between males and females, insurers may decide to avoid gender-based premium adjustments solely based on BMI, ensuring that policies remain fair and data-driven.

For risk assessment teams, the analysis confirms that BMI distributions are similar across genders, allowing them to focus on more impactful risk factors, such as smoking status, pre-existing conditions, and age, rather than gender-based BMI variations. This enhances the accuracy of risk modeling and pricing strategies.

For healthcare strategists and wellness program designers, the insights from the test suggest that gender-specific weight management programs may not be necessary, and broader health initiatives can be gender-neutral while focusing on other determinants of BMI fluctuations.

Additionally, by selecting an appropriate statistical test, the organization ensures credibility in data-driven decision-making, aligning with regulatory requirements and ethical fairness in policy pricing. The test results ultimately support evidence-based underwriting decisions, improve customer trust, and optimize company-wide resource allocation in risk assessment strategies.

**F1: Answer to the Research Question**

"Does BMI significantly differ between male and female policyholders?"

After conducting the Independent Samples T-Test, the results indicate that there is no statistically significant difference in BMI between males and females. The T-test p-value (0.090) is greater than 0.05, meaning we fail to reject the null hypothesis, which states that BMI does not significantly differ by gender.

Additionally, the Shapiro-Wilk test revealed that BMI is not normally distributed for either gender, suggesting that a nonparametric test (Mann-Whitney U test) may be more appropriate to confirm this conclusion. However, the Levene’s test for equal variances (p = 0.950) confirmed that BMI variance is similar across genders, meaning the spread of BMI values is consistent between males and females.

Looking at the boxplot visualization, both genders have similar median BMI values and distributions, with some outliers at the higher end. This further supports the finding that gender does not play a major role in determining BMI among policyholders.

The results suggest that gender does not significantly influence BMI among insurance policyholders. This insight is valuable for insurance companies because it implies that BMI-based risk assessments and premium pricing strategies should not differ based on gender alone. Instead, other factors such as age, lifestyle, and smoking status may have a greater impact on BMI and overall health risk.

**F2: Limitations of the Data Analysis**

While the analysis provided valuable insights into the relationship between BMI and gender, several limitations must be considered when interpreting the results. One key limitation is the violation of the normality assumption, as indicated by the Shapiro-Wilk test results for both males and females. Since the Independent Samples T-Test assumes normality, the reliability of its results is questionable, and a nonparametric alternative (Mann-Whitney U test) should be conducted to confirm the findings. Additionally, the analysis only focused on BMI and gender, without considering other important factors such as age, dietary habits, physical activity, socioeconomic status, or pre-existing health conditions, all of which could influence BMI variations and insurance risk assessments.

Another limitation is potential sampling bias, as the dataset may not be fully representative of the broader policyholder population. If the data is skewed toward certain age groups, regions, or health conditions, the results may not generalize well. Furthermore, the dataset categorizes gender as binary (Male/Female), which does not account for BMI variations among non-binary or transgender individuals, limiting the inclusivity of the study. The presence of high-BMI outliers, as observed in the boxplot visualization, could also inflate variance estimates and influence statistical test results, even though Levene’s test confirmed equal variances. Lastly, if BMI values were missing and imputed, the method of imputation (e.g., using the mean or median) may introduce bias into the results.

Given these limitations, further research should incorporate a nonparametric test to confirm results, consider additional health-related factors, examine outlier effects, and ensure a more diverse dataset to improve the robustness and generalizability of the findings.

**F3: Recommended Course of Action**

Based on the findings of this analysis, it is recommended that insurance companies do not use gender as a determining factor for BMI-based risk assessments or premium pricing. The Independent Samples T-Test results indicated that there is no statistically significant difference in BMI between male and female policyholders (p-value = 0.090), meaning that gender alone does not appear to influence BMI. Since the Shapiro-Wilk test showed that BMI is not normally distributed, it is advisable to conduct a nonparametric Mann-Whitney U test to confirm this conclusion before making final policy decisions.

Additionally, instead of focusing on gender, insurance providers should consider other more influential factors such as smoking status, age, pre-existing health conditions, and lifestyle behaviors when assessing risk and determining insurance premiums. Future analyses should incorporate these variables to develop a more comprehensive risk model that accurately predicts healthcare costs. Moreover, given the presence of BMI outliers, further investigation into whether extreme BMI values impact medical costs could help refine risk assessment strategies.

From a public health perspective, gender-neutral health and wellness programs should be prioritized over gender-based interventions, as the findings suggest that BMI variations are not significantly different between males and females. Insurers could instead focus on preventive care initiatives that target high-risk BMI groups regardless of gender, promoting healthier lifestyles and potentially reducing long-term healthcare expenses. Lastly, insurance companies should ensure their datasets are diverse and representative to maintain fairness and accuracy in policy development.

**Part III: Nonparametric Statistical Testing**

**G1: Research Question**

A key organizational question that can be addressed using nonparametric statistical analysis is:

"Is there a significant association between smoking status and geographic region?"

This question is important for insurance companies and healthcare providers as it helps determine if smoking habits vary by region. If certain regions have higher concentrations of smokers, insurance providers may need to adjust regional pricing models or develop targeted health initiatives to reduce smoking rates in those areas. Additionally, understanding regional differences in smoking behavior can help public health agencies design more effective anti-smoking programs based on local trends.

**G2 : Relvant Variables**

To answer this research question, we need to analyze the relationship between two categorical variables:

1. Independent Variable: region – This represents the geographic region where the policyholder resides (Northeast, Southeast, Southwest, or Northwest).
2. Dependent Variable: smoker – This is a binary categorical variable (yes or no), indicating whether the individual is a smoker.

Since both variables are categorical, we cannot use traditional parametric tests like the T-test. Instead, we will use a Chi-Square Test for Independence, which is the most appropriate statistical test for determining whether an association exists between two categorical variables.

**H1: Identify a Nonparametric Statistical Test**

Since region and smoker are categorical variables, we cannot use parametric tests like the T-test. Instead, we use the Chi-Square Test for Independence, which determines whether a significant association exists between smoking status (smoker) and geographic region (region).

Why Chi-Square Test?

* It is designed for categorical data.
* It tests if two categorical variables are independent or associated.
* It does not assume a normal distribution.

If the test result is significant (p-value < 0.05), smoking behavior varies significantly by region, helping insurance companies adjust pricing or target health initiatives based on geographic trends.

**H2: Develop Null and Alternative Hypothesis**

**Null Hypothesis (H₀):** Smoking status is independent of region (i.e., there is no significant relationship between region and smoker).

**Alternative Hypothesis (H₁):** Smoking status is not independent of region (i.e., there is a significant relationship between region and smoker).

If we reject the null hypothesis, it suggests that smoking habits differ across regions, which could have important implications for insurance policy pricing, public health programs, and smoking cessation efforts.

**H3: Conduct the Chi-Square Test**

# Create a cross-tabulation table

cross\_tab = pd.crosstab(df['region'], df['smoker'])

# Perform Chi-Square Test

chi2\_stat, p\_value, dof, expected = chi2\_contingency(cross\_tab)

# Print results

print(f"Chi-Square Statistic: {chi2\_stat:.3f}")

print(f"P-Value: {p\_value:.3f}")

print(f"Degrees of Freedom: {dof}")

# Visualize cross-tab with a heatmap

plt.figure(figsize=(8, 6))

sns.heatmap(cross\_tab, annot=True, cmap="Blues", fmt="d")

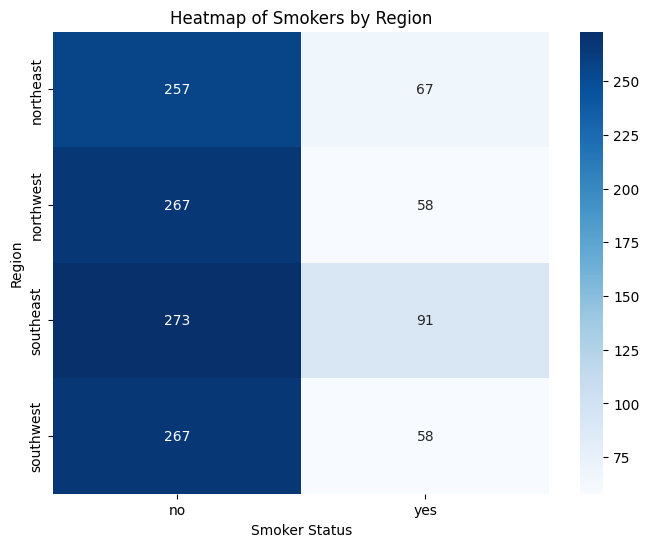
plt.title("Heatmap of Smokers by Region")

plt.xlabel("Smoker Status")

plt.ylabel("Region")

plt.show()

**H4: Results of the Chi-Square Test**



**Chi-Square Statistic:** 7.343

**P-Value:** 0.062

**Degrees of Freedom:** 3

**I1: Justification for Choosing the Statistical Test**

The Chi-Square Test for Independence was chosen because both smoking status (smoker) and geographic region (region) are categorical variables. The test is designed to determine whether there is a statistically significant association between two categorical variables. Unlike parametric tests that assume normality, the Chi-Square test does not require any specific distribution, making it ideal for analyzing categorical data.

Since we aimed to determine whether smoking prevalence varies significantly across different regions, the Chi-Square test was the most appropriate method for this analysis. The test allows insurance companies and public health policymakers to understand if certain regions have higher concentrations of smokers, which could impact insurance pricing models, regional healthcare policies, and smoking cessation initiatives.

**I2: Discussion of Test Results**

The Chi-Square test produced a Chi-Square statistic of 7.343, a p-value of 0.062, and 3 degrees of freedom. The p-value is greater than 0.05, which means we fail to reject the null hypothesis (H₀). This indicates no statistically significant association between smoking status and geographic region. In other words, the data does not provide strong enough evidence to suggest that smoking habits vary significantly across different areas.

While there may be some slight differences in smoking prevalence across regions, they are not large enough to be considered statistically significant at the 5% level (p < 0.05). This suggests that smoking behavior is relatively uniform across different regions, and regional pricing adjustments for insurance based on smoking prevalence may not be necessary.

However, the p-value is relatively close to 0.05 (marginal significance), meaning that a larger dataset or a more detailed regional breakdown (e.g., by state instead of region) could reveal a significant relationship. Future research could consider adding more granular location data or incorporating other factors, such as urban vs. rural differences, to understand regional smoking trends better.

**I3: Implications for Stakeholders**

These results provide valuable insights for insurance companies, public health agencies, and policymakers. Since the analysis found no strong regional differences in smoking behavior, insurance companies may not need to adjust premiums based on geography alone. Instead, individual-level risk factors such as age, BMI, and pre-existing conditions might be more important determinants for premium adjustments.

The lack of regional variation for public health officials suggests that national or statewide anti-smoking campaigns may be equally effective across all regions rather than requiring region-specific strategies. However, given the marginal p-value (0.062), further studies with more location-specific data (such as state-level or urban vs. rural comparisons) could help identify whether smaller-scale trends exist.

Employers offering regional health insurance plans can also use this data to determine if regional smoking cessation programs are needed or if a one-size-fits-all approach to employee health benefits is more appropriate.

**J1: Answer to the Research Question**

The research question was: "Is there a significant association between smoking status and geographic region?" Based on the Chi-Square Test for Independence results, we found no statistically significant relationship between smoking status and geographic region. The test produced a Chi-Square statistic of 7.343, a p-value of 0.062, and 3 degrees of freedom. Since the p-value is greater than 0.05, we fail to reject the null hypothesis (H₀), meaning that smoking prevalence does not significantly vary across different regions in this dataset.

This result indicates that, within this insurance dataset, smoking habits appear to be relatively evenly distributed across geographic regions. No strong statistical evidence suggests that certain areas have higher or lower smoking rates. Consequently, regional pricing models based on smoking prevalence may not be necessary for insurance companies. However, the p-value is close to the 0.05 threshold, suggesting that a larger sample size or more granular regional data (e.g., state-level or urban vs. rural comparisons) could provide further insights**.**

**J2: Limitations of the Data Analysis**

Despite the insights gained from this analysis, several limitations exist. First, the dataset only includes four broad regions (Northeast, Southeast, Southwest, and Northwest), which may not accurately capture localized smoking trends within individual states or cities. Smoking behavior can vary significantly between urban and rural regions, which is not accounted for in this analysis. A more granular geographic breakdown might yield different results.

Second, the dataset does not control for other demographic or socioeconomic factors influencing smoking behavior and geographic distribution. Variables such as income level, education, occupation, and access to healthcare could play a role in regional smoking differences but were not included in this analysis. Additionally, self-reported smoking status could introduce reporting bias, as some individuals may not accurately disclose their smoking habits.

Lastly, while the Chi-Square Test is appropriate for categorical data, it only tells us whether an association exists—it does not measure the strength or direction of the relationship. If a weak association exists but does not reach statistical significance, the test may not detect it, especially if the sample size is not large enough.

**J3: Recommended Course of Action**

Given that the results suggest no significant relationship between region and smoking status, insurance companies may not need to adjust premium rates or risk assessments based on geographic location alone. Instead, insurers should focus on individual risk factors such as age, BMI, and pre-existing conditions, which are likely stronger predictors of healthcare costs than geographic regions. However, given the p-value of 0.062, which is close to the 0.05 significance level, further research with larger sample size or state-level data is recommended to confirm whether smaller-scale regional trends exist.

This analysis suggests that broad, nationwide smoking prevention campaigns are as effective for public health officials as region-specific programs. However, it may still be beneficial to conduct state-level or local studies to identify specific areas where smoking prevalence may be higher. Further analysis incorporating socioeconomic and behavioral factors could provide a more detailed understanding of the drivers behind smoking rates in different regions.

Employers and insurance providers should also consider expanding wellness programs to encourage smoking cessation rather than relying on regional data for pricing or policy decisions. Since smoking behavior does not significantly differ by region in this dataset, individual lifestyle factors and health habits should be prioritized when designing insurance pricing models and employee health benefits.