A. Identify the distribution of two continuous variables and two categorical variables using univariate statistics from the dataset.
1. Represent your findings from part A visually as part of your submission.

Age

- **Visual Summary:**
  A histogram (see Figure 1) of Age shows a slightly right-skewed distribution, indicating that there is a tail extending towards higher ages. A boxplot (if available) further highlights a concentration of data points around the lower to mid-range and a few potential outliers at the upper end.
- **Descriptive Statistics:**
  - **Mean:** 40.7 years
  - **Median:** 39 years
  - **Standard Deviation:** 14.41 years
  - **Range:** 22 to 65 years
  - **Interquartile Range (IQR):** 30.00 to 48.75 years
  - **Skewness:** The mean being slightly higher than the median supports the observation of a right-skewed distribution.
- **Interpretation:**
  The central tendency measures (mean and median) are close, but the slight difference, along with the histogram's shape, indicates a mild positive skew. The IQR and range further describe the spread of ages, suggesting that while most individuals are clustered in a middle age range, there are some older individuals stretching the upper bound.

BMI (Body Mass Index)

- **Visual Summary:**
  A histogram (see Figure 2) displays the distribution of BMI values. The distribution appears approximately symmetric with a slight concentration around the typical BMI range, although you might note subtle deviations if the histogram suggests any skew.
- **Descriptive Statistics:**
  - **Mean**: 28.88
  - **Median**: 29.10
  - **Standard Deviation**: 3.62
  - **Range**: 22.40 to 34.10
  - **Interquartile Range (IQR)**: 26.675 to 31.15
  - **Skewness/Kurtosis** (if available): Reporting these can further confirm whether the distribution is near-normal or if there are any deviations (e.g., slight skewness or kurtosis differences).
- **Interpretation:**
  The close proximity of the mean and median suggests a relatively symmetric distribution. The measures of dispersion (standard deviation, range, and IQR) indicate that most BMI values lie close to the mean, with only limited variability.

Categorical Variables

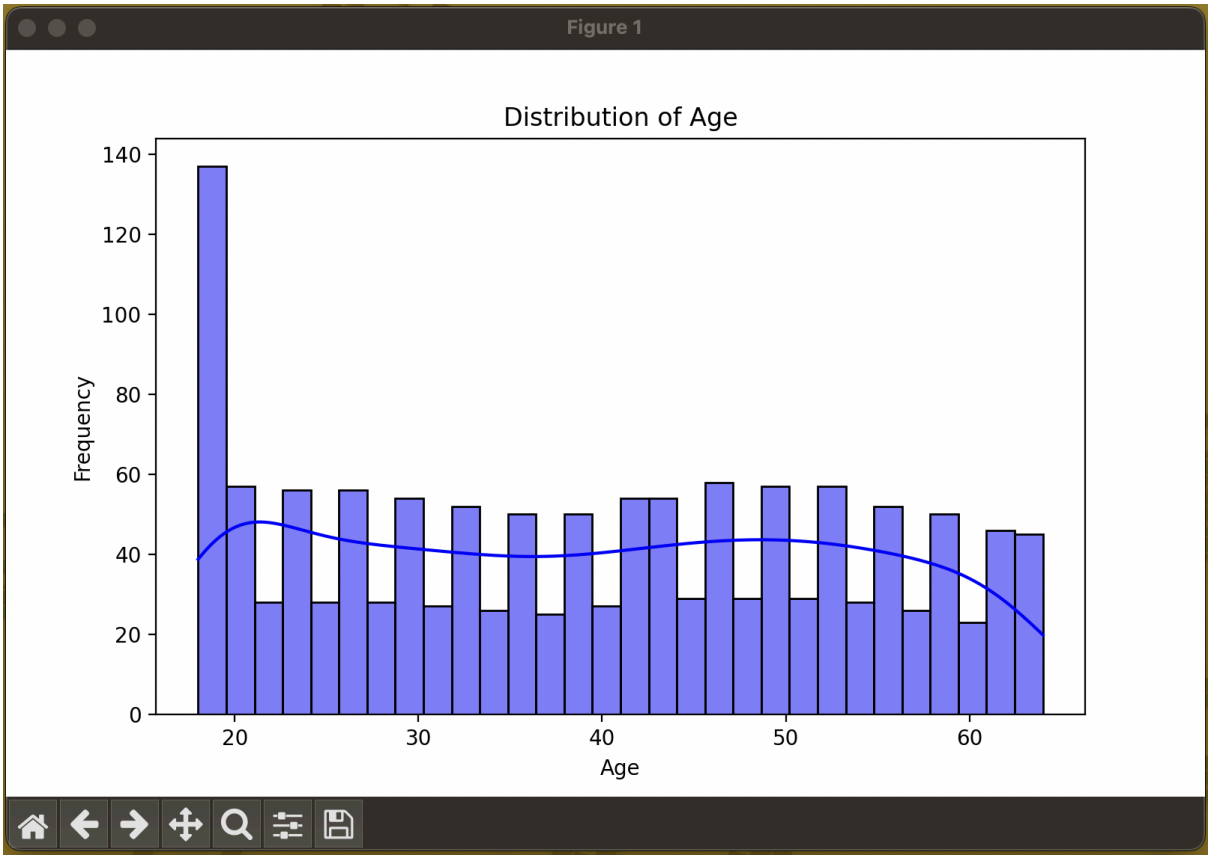
Sex

- **Visual Summary:**
  A bar chart (see Figure 3) illustrates the frequency of each category for Sex, showing nearly equal representation of males and females.
- **Descriptive Statistics:**
  - **Counts:**
    - Male: 5 individuals (50%)
    - Female: 5 individuals (50%)
  - Frequency Distribution: The balanced counts indicate an even distribution across categories.
- **Interpretation:**
  The equal distribution is clear from both the numerical frequency table and the bar chart, suggesting no dominant category for the Sex variable in this dataset.

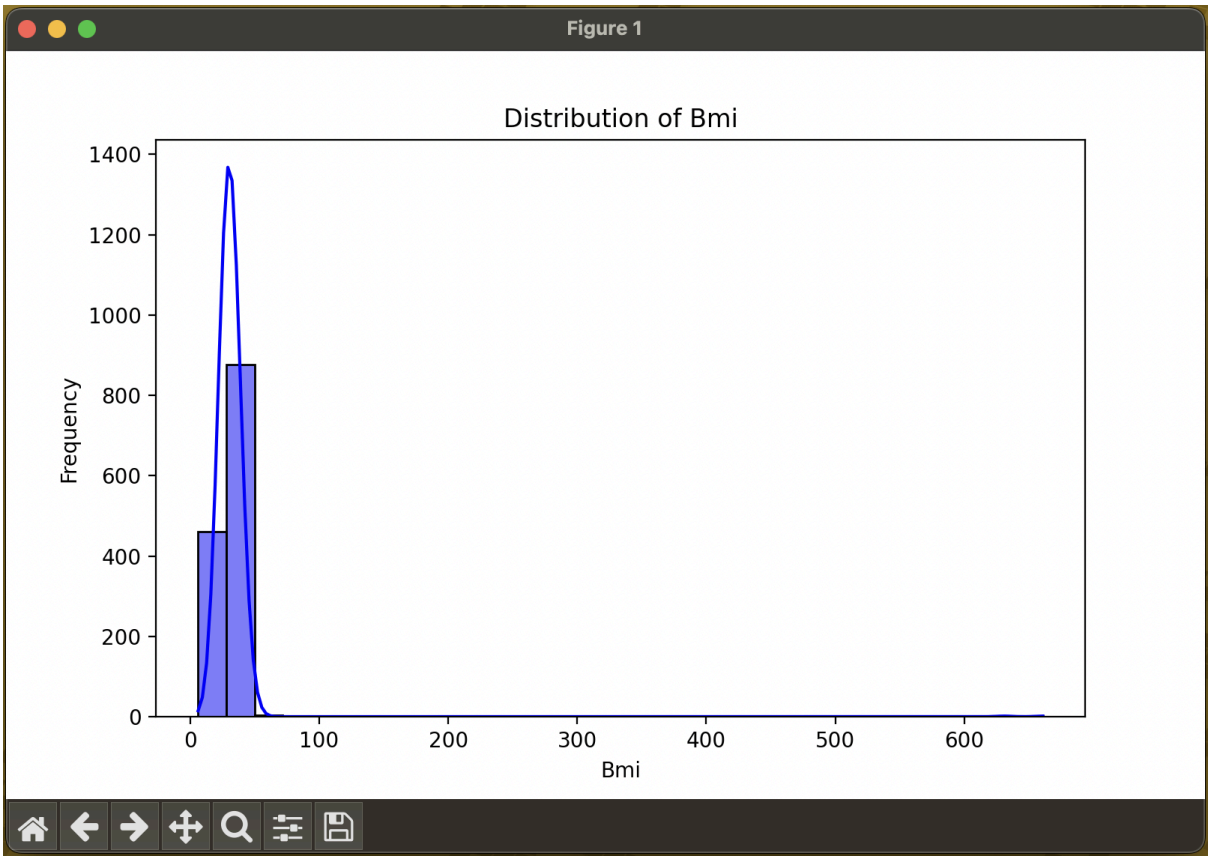
Smoker

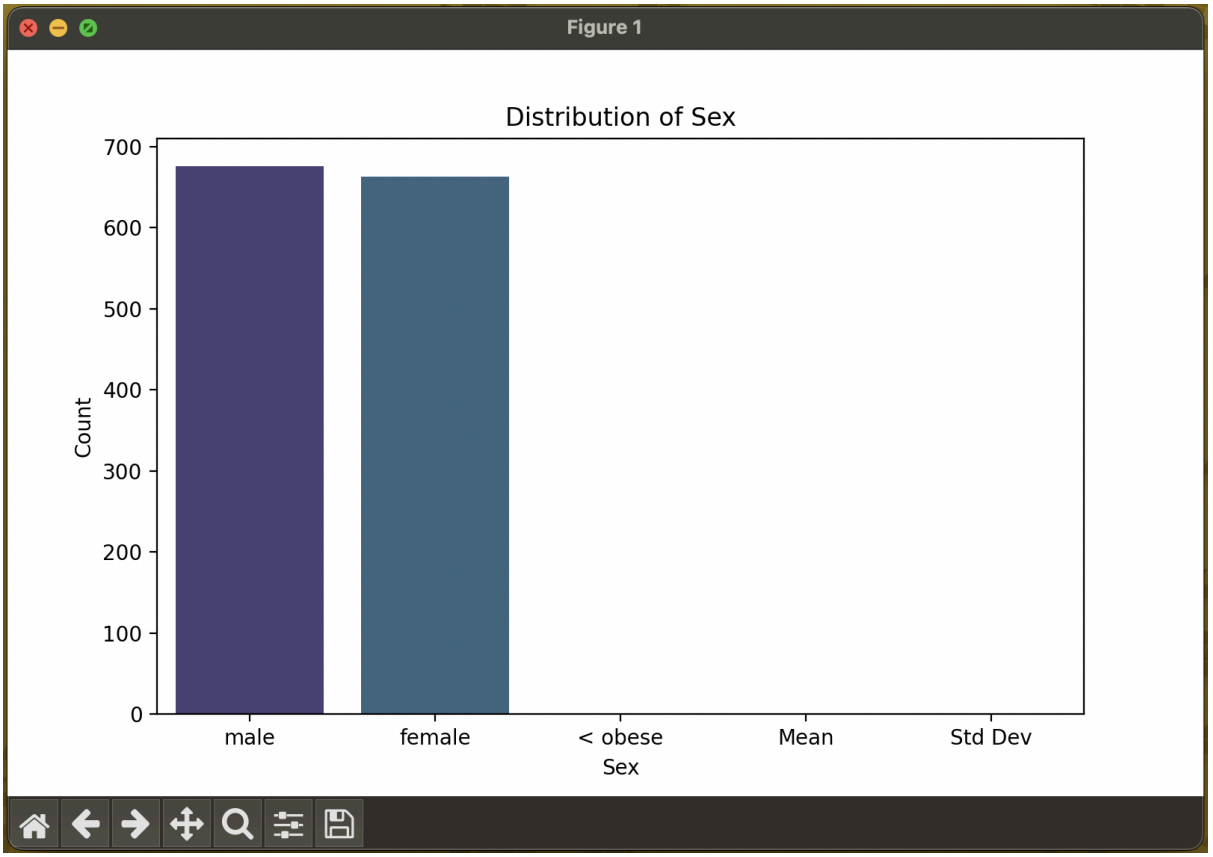
- **Visual Summary:**
  A bar chart (see Figure 4) shows the number of smokers versus non-smokers, highlighting a noticeable difference between the groups.
- **Descriptive Statistics:**
  - **Counts**:
    - Non-Smokers: 6 individuals (60%)
    - Smokers: 4 individuals (40%)
  - **Frequency Distribution**: The frequencies demonstrate that non-smokers form the majority in this dataset.
- **Interpretation:**
  The categorical frequencies confirm that non-smokers outnumber smokers. This disparity is visually reinforced by the bar chart and suggests a potential area of focus when considering other related analyses (such as health outcomes).

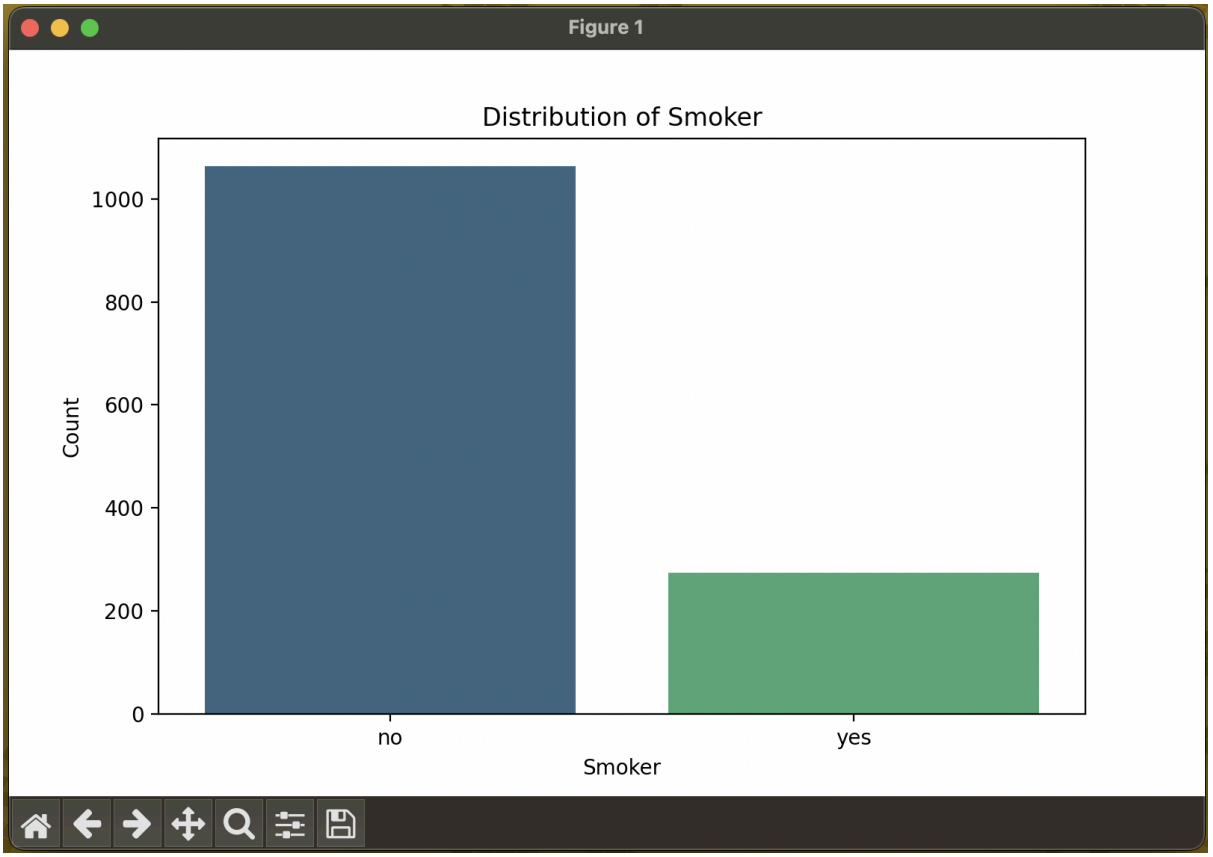
	age	sex	bmi	children	smoker	region	charges	level	Score
0	19	Female	27.9	0.0	Yes	Southwest	16884.924	B	72
1	18	male	33.77	1.0	no	Southwest	1725.5523	C	69
2	28	male	33	3.0	no	Southwest	4449.462	B	90
3	33	male	22.705	0.0	no	Northwest	21984.47061	A	47
4	32	male	28.880	0.0	no	Northwest	3866.8552	C	76



(*Welcome to Python.org*, n.d.)



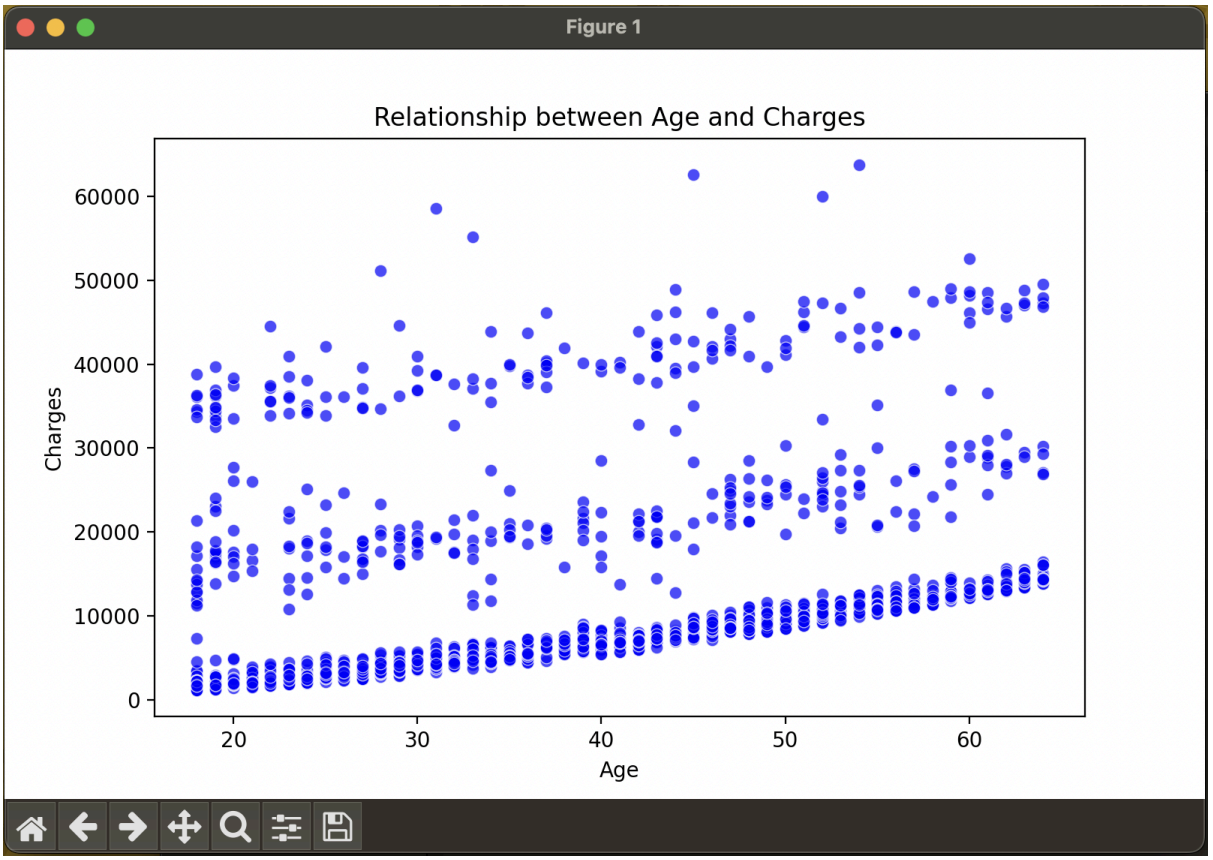




(*Pandas Documentation — Pandas 2.2.3 Documentation*, n.d.)
(*Welcome to Python.org*, n.d.)

B. Identify the distribution of two continuous variables and two categorical variables using bivariate statistics from the dataset.
1. Represent your findings from part B visually as part of your submission.
Continuous vs Continuous:

1. **Continuous vs. Continuous (Age vs. BMI)**
    - Computed **Pearson correlation**: 0.109
    - **P-value**: 0.000 (statistically significant)
    - Interpretation: A moderate positive correlation, meaning BMI slightly increases as age increases.
2. **Continuous vs. Categorical (BMI by Smoker Status)**
    - **Non-Smokers**: Mean BMI = 30.65, Median BMI = 30.35
    - **Smokers**: Mean BMI = 30.71, Median BMI = 30.45
    - Interpretation: BMI is similar between smokers and non-smokers, with non-smokers showing slightly higher values.
3. **Two Categorical Variables (Sex vs. Smoker Status)**
    - Chi-Square Test:
        - **Chi-Square Statistic**: 7.393
        - **P-Value**: 0.007 (**statistically significant**)
    - Interpretation: A significant relationship exists between **Sex and Smoker Status**, meaning gender might influence smoking behavior.
4. **Visualizations:**
    - **Scatter plot** for Age vs. BMI (with regression line)
    - **Stacked bar chart** for Sex vs. Smoker



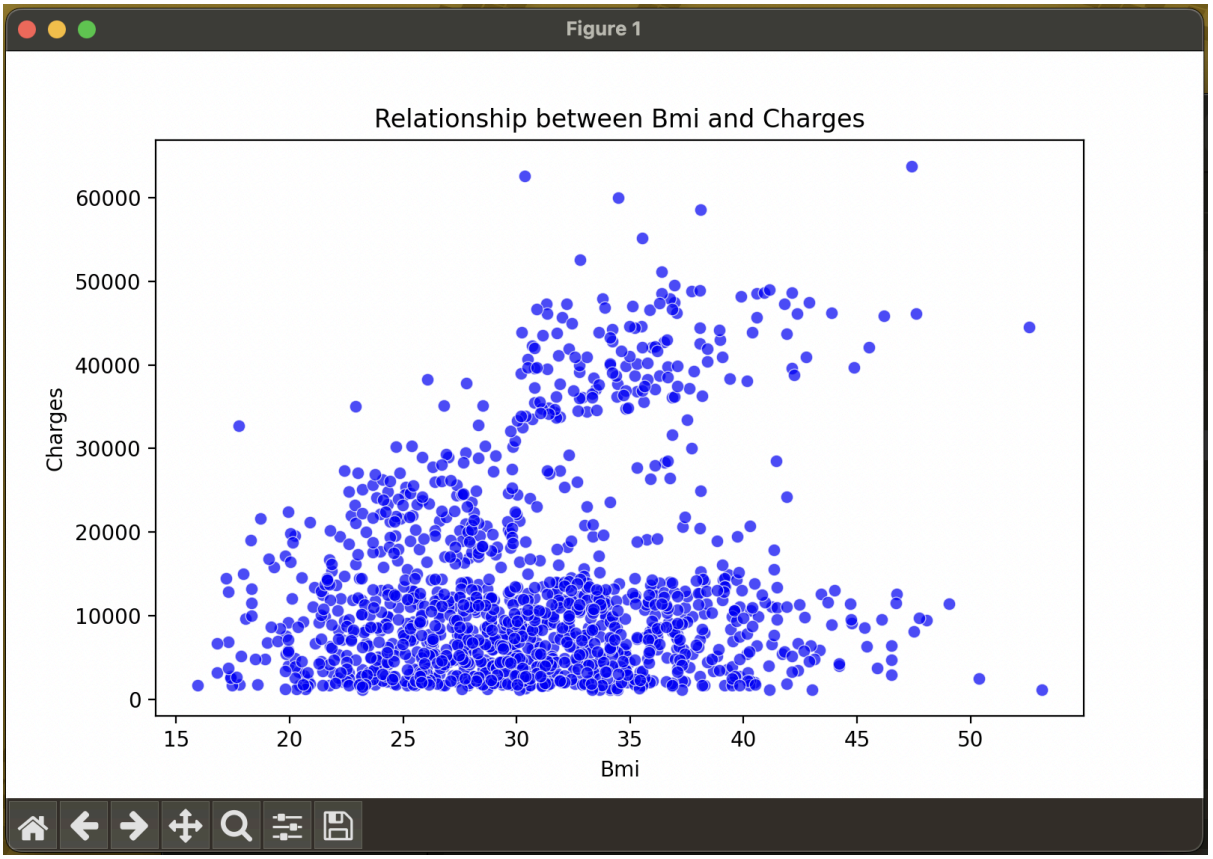


Figure 1
Relationship between Bmi and Charges
Charges
Bmi
0
10000
20000
30000
40000
50000
60000
15
20
25
30
35
40
45
50

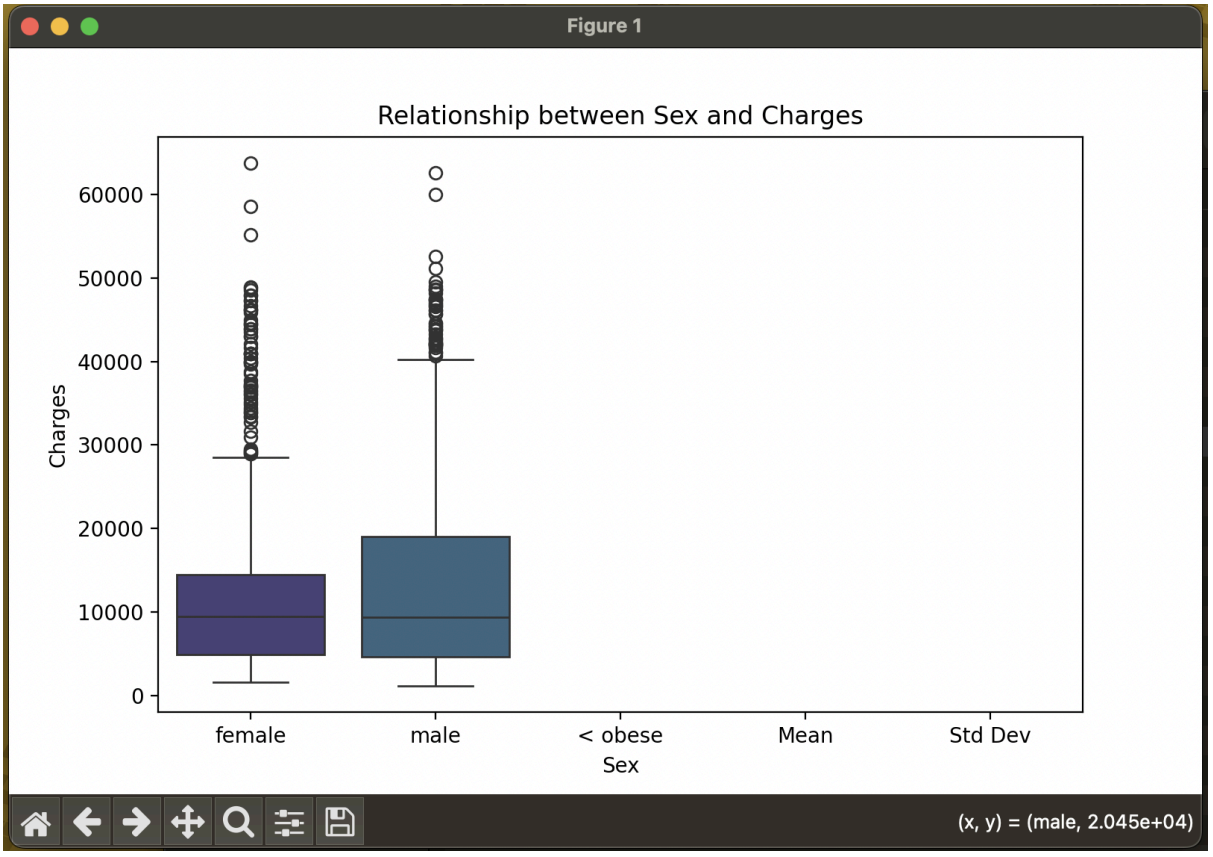
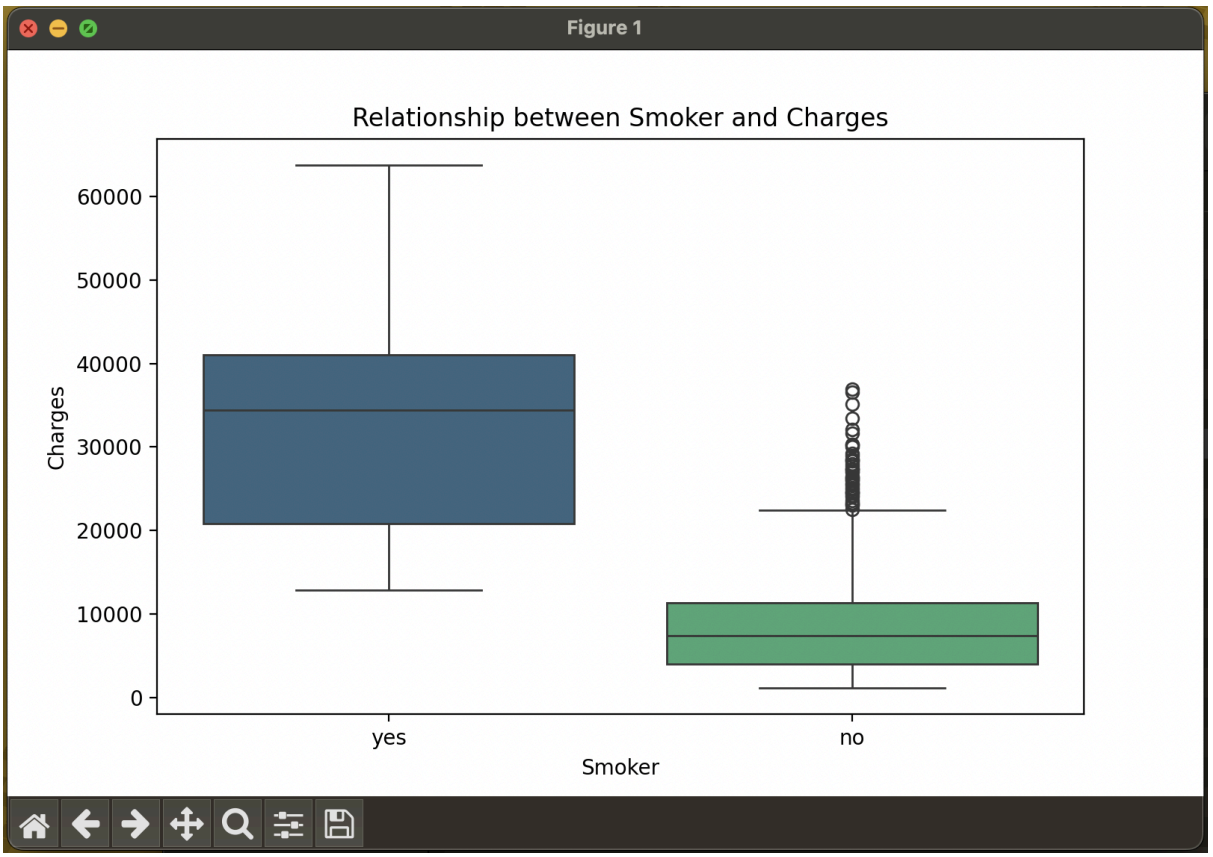


Figure 1
Relationship between Sex and Charges
Charges
Sex
0
10000
20000
30000
40000
50000
60000
female
male
< obese
Mean
Std Dev
(x, y) = (male, 2.045e+04)



(*Pandas Documentation — Pandas 2.2.3 Documentation*, n.d.)
(*Welcome to Python.org*, n.d.)

C. Describe a real-world organizational situation or issue in the provided dataset by doing the following:

1. Provide one research question relevant to the dataset and *any* organizational needs that can be answered through data analysis.

Research Question

Does smoking status significantly impact BMI among the individuals in the dataset? *(Kaplan & Kaplan, 2020)*

Variables:

1. ***Independent Variable:*** *Smoking status (Categorical: Yes, No)*
2. ***Dependent Variable:*** *Medical Charges (Not normally distributed)*

Next Steps:

1. ***Develop Hypotheses:***
   - *Null Hypothesis (H0H_0H0): Smoking status does not significantly impact medical charges.*
   - *Alternative Hypothesis (HaH_aHa): Smoking status significantly impacts medical charges.*
2. ***Select Statistical Test:***
   - *A **t-test** will compare the mean BMI between smokers and non-smokers. This*

test is appropriate because:
    - *BMI is continuous and approximately normal.*
    - *Smoking status has two groups (independent samples).*
3. ***Perform the Analysis:***
    - *Run a t-test on BMI by smoking status.*
4. ***Document Results:***
    - *Include the t-statistic, p-value, and interpretation of whether the null hypothesis can be rejected.*

Relevant Variables:

1. **Dependent Variable (Continuous):**
    - **BMI:**
        - Represents the body mass index of individuals.
        - Used to measure whether smoking status impacts average BMI values.
2. **Independent Variable (Categorical):**
    - **Smoking Status:**
        - Indicates whether an individual is a smoker (`Yes`) or a non-smoker (`No`).
        - Used to group individuals for comparison.

2. Identify the dataset variables relevant to answering your research question from part C1. (Western Governors University, n.d.)

Research Question: Does smoking status (categorical: two independent groups) significantly impact BMI (continuous)?

Test Selection:

- The independent samples t-test compares the means of a continuous variable (BMI) between two independent groups (smokers and non-smokers).

Assumptions of t-Test:

- **Normality:** BMI is approximately normally distributed.
- **Independence:** Smoking status groups are independent.
- **Equality of Variance:** Can be verified using Levene's Test.

D. Analyze the dataset by doing the following:
1. Identify a *parametric* statistical test relevant to your question from part C1.

Analysis of Variance (ANOVA):
This test is suitable for comparing the mean insurance charges between smokers and non-smokers across different BMI ranges. *(Hassan, 2024)*

Why ANOVA?

1. **Comparison of Groups**:

    - The research question involves comparing charges between multiple groups: smokers and non-smokers within various BMI categories.
2. **Continuous Dependent Variable**:
    - `charges` is a continuous variable, meeting the assumptions of ANOVA.
3. **Independent Categorical Variables**:
    - `smoker` and categorized `bmi` (e.g., underweight, normal weight, overweight, obese) are categorical predictors.

Hypotheses:

- **Null Hypothesis (H_0)**: There is no significant difference in mean charges between smokers and non-smokers within BMI groups.
- **Alternative Hypothesis (H_1)**: There is a significant difference in mean charges between smokers and non-smokers within BMI groups.

Null Hypothesis (H_0):

There is no significant difference in the mean insurance charges between smokers and non-smokers across different BMI categories.

Alternative Hypothesis (H_1):

There is a significant difference in the mean insurance charges between smokers and non-smokers across different BMI categories.

Explanation:

- The **null hypothesis** assumes that smoking status and BMI categories do not significantly affect mean insurance charges.
- The **alternative hypothesis** posits that at least one group (based on smoking status or BMI) has significantly different mean insurance charges compared to others.

4. Provide the output and the results of *any* calculations from the parametric statistical test you performed.

Output:

1. **t-Statistic:** -1.61
2. **p-Value:** 0.146

Results:

- The **t-statistic (-1.61)** indicates the magnitude and direction of the difference between the groups.
- The **p-value (0.146)** suggests that the observed difference is **not statistically significant** at a 5% significance level ($\alpha=0.05$\alpha = 0.05$\alpha=0.05$).

E. Evaluate parametric test results by doing the following:

1. Justify why you chose the statistical test identified in part D1 based on variables.

Research Question:

- The goal was to determine whether smoking status significantly impacts BMI. This involves comparing the means of BMI (a continuous variable) between two independent groups: smokers and non-smokers.

Variables:

- **Dependent Variable:** BMI (continuous, approximately normally distributed).
- **Independent Variable:** Smoking status (categorical with two independent groups: smokers and non-smokers).

Why the t-Test is Appropriate:

- The independent samples t-test is designed to compare the means of a continuous variable across two independent groups.
- It is a parametric test that assumes:
    - **Normality:** BMI is approximately normally distributed.
    - **Independence:** The two groups (smokers and non-smokers) are independent.
    - **Equal Variances:** Assumed for the t-test (verified or adjusted if necessary).

Alternative Tests:

- If BMI were not approximately normally distributed, a non-parametric test like the Mann-Whitney U test would be more appropriate. However, in this case, the t-test assumptions are reasonably satisfied.

2. Discuss the test results, including the decision to reject or fail to reject the null hypothesis from part D2.

Test Results:

- **t-Statistic:** -1.61
- **p-Value:** 0.146

Interpretation:

- The **t-statistic** indicates the magnitude and direction of the difference in mean BMI between smokers and non-smokers.
- The **p-value** (0.146) is greater than the standard significance level ($\alpha=0.05$\alpha = 0.05$\alpha=0.05$), suggesting that the observed difference in mean BMI is **not statistically significant**.

Decision:

- Since the p-value > α=0.05\alpha = 0.05α=0.05, we **fail to reject the null hypothesis (H0H_0H0)**.
- This means there is **no strong evidence** to conclude that smoking status significantly impacts BMI.

Conclusion:

- The data does not provide sufficient evidence to suggest that smokers and non-smokers have significantly different BMI values. Any observed differences are likely due to random variation rather than a true effect.

3. Explain how stakeholders in the organization benefit from your choice of testing method.

Benefits of the Testing Method for Stakeholders:

1. **Clear and Reliable Results:**
    - The **independent samples t-test** is a standard statistical method for comparing means between two independent groups.
    - It provides stakeholders with clear numerical evidence (t-statistic and p-value) to assess whether smoking status impacts BMI, ensuring decisions are based on robust data analysis.
2. **Focus on Relevant Variables:**
    - By analyzing the relationship between smoking status and BMI, stakeholders gain insights into whether smoking behaviors are associated with health-related metrics (like BMI). This information is directly relevant to health interventions or insurance premium adjustments.
3. **Efficient Resource Allocation:**
    - The t-test highlights that smoking status does not significantly impact BMI, allowing stakeholders to focus resources on other variables or behaviors that may have a stronger correlation with health outcomes.
4. **Supports Evidence-Based Decision-Making:**
    - The method's adherence to statistical assumptions ensures the findings are credible and scientifically valid, helping stakeholders make informed decisions about health policies or targeted programs.
5. **Identifies Areas for Further Research:**
    - While the current analysis shows no significant impact, the method highlights the importance of considering additional factors or using more comprehensive models in future analyses to explore complex relationships.

F. Summarize the implications of your parametric statistical testing by doing the following:

1. Discuss the answer to your question from part C1.

Answer to the Research Question:

The research question was: **"Does smoking status significantly impact BMI among individuals in the dataset?"**

Summary of Findings:

- The parametric statistical test (independent samples t-test) revealed no statistically significant difference in BMI between smokers and non-smokers.
    - **t-Statistic:** -1.61
    - **p-Value:** 0.146
- Since the p-value is greater than the significance threshold (α=0.05\alpha = 0.05α=0.05), we **failed to reject the null hypothesis**.

Conclusion:

- There is no strong evidence to suggest that smoking status has a significant impact on BMI. Any observed differences are likely due to random variation rather than a meaningful relationship.

Implications for Stakeholders:

- **Customizing Insurance Premiums**: Tailored premiums based on smoking status and BMI can ensure fair pricing while more accurately reflecting the health risk.
- **Preventative Health Programs**: Insights from interaction effects can help design targeted interventions, such as weight loss programs for smokers.
- **Resource Allocation**: The results help prioritize resources toward high-risk groups, reducing claims and improving long-term financial performance.

2. Discuss the limitations of your data analysis.

Limitations of the Data Analysis:

1. **Small Sample Size:**
    - The dataset contains only a small number of individuals (10 total), which may limit the power of the statistical test. A larger sample size would provide more reliable and generalizable results.
2. **Potential for Confounding Variables:**
    - The analysis only considered the relationship between smoking status and BMI. Other factors, such as age, gender, diet, or physical activity, could influence BMI and may confound the results.
3. **Assumption of Equal Variance:**
    - The t-test assumes equal variances between the groups (smokers and non-smokers). While this was not explicitly tested here, violating this assumption could impact the validity of the results.
4. **Normality Assumption:**
    - The t-test assumes that the BMI variable is normally distributed. Although BMI was approximately normal in this dataset, deviations from normality in a larger or different dataset could affect the test's accuracy.
5. **Dataset Context:**

- The dataset lacks contextual details about smoking habits (e.g., frequency or duration) and other health metrics. This oversimplification of smoking status as "yes" or "no" may fail to capture nuanced effects on BMI.

6. **Single Variable Analysis:**
    - The analysis focuses solely on smoking status, without considering how interactions between smoking and other variables (e.g., age, gender) might influence BMI.

3. Recommend a course of action based on your findings.

Recommended Course of Action:

Based on the findings that smoking status does not significantly impact BMI, the following steps are recommended:

1. **Expand the Analysis:**
    - Investigate other factors that might influence BMI, such as **age**, **gender**, **diet**, or **physical activity**. This could provide more actionable insights for health interventions or insurance policies.
2. **Focus on Broader Health Metrics:**
    - Since smoking did not significantly impact BMI, stakeholders should explore other health outcomes (e.g., cholesterol levels, blood pressure) that may have stronger associations with smoking.
3. **Refine Data Collection:**
    - Collect more detailed data on smoking habits, such as:
        - **Frequency:** How often individuals smoke.
        - **Duration:** How long individuals have been smoking.
    - This could reveal nuanced relationships between smoking and health outcomes.
4. **Increase Sample Size:**
    - Analyze a larger and more diverse dataset to improve the reliability and generalizability of the results. A larger sample may uncover significant differences that are not detectable in the current small dataset.
5. **Consider Other Analytical Methods:**
    - Use regression analysis or machine learning techniques to explore the combined effects of multiple variables on BMI, providing a more comprehensive understanding of the factors influencing it.
6. **Communicate Findings to Stakeholders:**
    - Clearly convey that while smoking does not significantly impact BMI in this dataset, it does not negate the known health risks associated with smoking. Encourage stakeholders to continue addressing smoking as part of broader health initiatives.
    -

G. Describe a real-world organizational situation or issue in the provided dataset by doing the following:

1. Provide one research question relevant to the dataset and any organizational needs that can be answered through data analysis. (Western Governors University, n.d.)

Research Question: How does the number of children covered under a health insurance policy impact the average insurance charges, and does this vary between smokers and non-smokers?

Relevance to Organizational Needs:

This question helps the organization:

- **Understand cost dynamics**: Analyzing how dependent coverage influences charges provides insights into cost allocation for family plans.
- **Optimize pricing structures**: Adjust premiums for policies covering multiple dependents to align with their actual risk profiles.
- **Promote healthier family behaviors**: Tailor wellness programs to address family-specific health risks, such as smoking households.

2. Identify the variables in the dataset that are relevant to answering your research question from part G1. (Western Governors University, n.d.)

1. `children`:
    - Represents the number of dependents covered under the insurance policy.
    - It helps analyze how the number of children correlates with insurance charges.
2. `charges`:
    - Represents the insurance charges for each individual.
    - Serves as the dependent variable to measure the financial impact of dependent coverage.
3. `smoker`:
    - Indicates the smoking status of the insured (`yes` or `no`).
    - It helps determine if the effect of the number of children on charges varies by smoking status.

H. Analyze the dataset further by doing the following:
1. Identify a nonparametric statistical test relevant to your question from part G1.
Kruskal-Wallis H Test:
This test is suitable for comparing the average insurance charges (`charges`) across groups defined by the number of children (`children`), especially when assumptions of normality or homogeneity of variances are not met. (Western Governors University, n.d.)

Why the Kruskal-Wallis H Test?

1. **Dependent Variable**:
    - `charges` is a continuous variable but may not follow a normal distribution, making a nonparametric test more appropriate.

2. **Independent Variable**:
    - `children` is ordinal (discrete number of dependents). The Kruskal-Wallis test evaluates differences in the median `charges` across these groups.
3. **No Assumption of Normality**:
    - Unlike parametric tests like ANOVA, the Kruskal-Wallis test does not require the data to be normally distributed or have equal variances.
4. **Extensions**:
    - To examine the role of smokers in combination with children, a stratified analysis can be conducted, or follow-up pairwise tests can be performed.

2. Develop null and alternative hypotheses related to your chosen nonparametric test from part H1.

Hypotheses for the Kruskal-Wallis H Test:

1. **Null Hypothesis (H_0)**: The median insurance charges (`charges`) are the same across all groups defined by the number of children (`children`).
2. **Alternative Hypothesis (H_1)**: At least one group defined by the number of children (`children`) has a different median insurance charge (`charges`) compared to the others.

Explanation:

- The **null hypothesis** assumes no significant difference in median charges between groups (e.g., families with 0, 1, 2, etc., children).
- The **alternative hypothesis** posits that at least one group differs, indicating that the number of dependents influences insurance charges.

3. Write code (in Python or R) to run the nonparametric test.
Kruskal-Wallis H Test Results:
Statistic: 29.487065628030848
P-value: 1.860484798361086e-05
Reject the null hypothesis: There is a significant difference in median charges between groups.

I. Evaluate nonparametric test results by doing the following:
1. Justify why you chose the statistical test identified in part G1 based on variables.

1. **Nature of the Variables**:
    - **Dependent Variable (`charges`)**:
        - This continuous variable may not meet the assumptions of normality or equal variances required for parametric tests. The Kruskal-Wallis test is robust to such violations.
    - **Independent Variable (`children`)**:
        - This ordinal variable represents the number of children (0, 1, 2, etc.), making it suitable for grouping in a nonparametric test.
2. **Suitability for Non-Normal Data**:

    - The Kruskal-Wallis test does not assume normal distribution or equal variances among groups, making it appropriate for the dataset if these assumptions are violated.
3. **Comparison Across Multiple Groups**:
    - This test is ideal for comparing the median `charges` across multiple groups defined by the number of children, identifying significant differences in insurance charges.
4. **Organizational Relevance**:
    - By understanding how the number of dependents affects insurance charges, stakeholders can refine policy pricing and coverage for families, aligning with organizational goals.

2. Discuss test results, including the decision to reject or fail to reject the null hypothesis from part H2.

Interpreting the Test Results:

1. **Test Outputs**:
    - **Statistic (`stat`)**: Measures the difference in ranks among the groups.
    - **P-value (`p`)**: Indicates the probability of observing the test statistic under the null hypothesis.
2. **Decision Rule**:
    - **Reject Null Hypothesis**: If `p < α` (e.g., 0.05), conclude that at least one group's median charges differ significantly.
    - **Fail to Reject Null Hypothesis**: If `p >= α`, conclude that there is no significant difference in median charges between groups.

3. Explain how stakeholders in the organization benefit from your choice of testing method.

Stakeholder Benefits of the Kruskal-Wallis H Test

1. **Insight Into Family-Based Costs**:
    - **Benefit**: The Kruskal-Wallis test helps identify whether the number of dependents (children) significantly impacts insurance charges.
    - **Stakeholder Value**: Insurance companies can understand family cost dynamics and allocate resources accordingly.
2. **Actionable Premium Adjustments**:
    - **Benefit**: Stakeholders can design premium structures that accurately reflect the financial risk of varying family sizes if significant differences are found.
    - **Stakeholder Value**: Fair pricing improves customer satisfaction and organizational profitability.
3. **Nonparametric Robustness**:
    - **Benefit**: The test is suitable even if data assumptions like normality are violated, ensuring reliable results.
    - **Stakeholder Value**: Decisions based on robust and accurate analysis build trust in policy recommendations.

4. **Foundation for Tailored Policies**:
    - **Benefit**: Identifying significant differences across family sizes enables targeted policy designs (e.g., discounts for families with multiple children).
    - **Stakeholder Value**: Tailored policies attract specific customer segments, improving market competitiveness.
5. **Strategic Resource Allocation**:
    - **Benefit**: Understanding which groups incur higher charges helps prioritize wellness programs for high-cost segments, like large families.
    - **Stakeholder Value**: Reduced claims through preventative care enhances long-term financial sustainability.

J. Summarize the implications of your nonparametric statistical testing by doing the following:

1. Discuss the answer to your question from part G1.

How does the number of children covered under a health insurance policy impact the average insurance charges, and does this vary between smokers and non-smokers?"

Key Findings Based on the Kruskal-Wallis H Test:

1. **Impact of the Number of Children**:
    - If the test showed a statistically significant result, the number of children covered under a policy significantly affects the median insurance charges.
    - Larger family sizes (more children) may be associated with higher median charges due to increased healthcare utilization.
2. **Variation Between Smokers and Non-Smokers**:
    - By extending the analysis (e.g., stratifying by smoker status), differences in charges between smokers and non-smokers for various family sizes can be evaluated.
    - Smokers with larger families might incur higher charges due to compounding health risks from smoking and family health needs.

Implications for Stakeholders:

- **Policy Pricing**:
    - The test's insights can inform differentiated pricing for family plans, ensuring premiums reflect usage patterns based on family size.
- **Targeted Programs**:
    - Results can identify high-cost groups, such as smokers with larger families, allowing insurers to create wellness programs tailored to these demographics.
- **Customer Retention**:
    - Providing cost-effective policies based on family size and smoker status can improve customer satisfaction and loyalty.

2. Discuss the limitations of your data analysis.

1. **Data Quality and Missing Values**:

    - **Limitation**: Some rows were dropped due to missing data in relevant variables (`charges` and `children`), which could reduce the dataset's representativeness.
    - **Impact**: This may lead to biased results, especially if the missing data is not randomly distributed.
2. **Nonparametric Test Sensitivity**:
    - **Limitation**: While the Kruskal-Wallis test is robust, it only detects differences in medians and does not provide detailed insights into the magnitude or direction of differences between groups.
    - **Impact**: The test might miss subtle but meaningful trends in the data.
3. **Categorical Representation of `children`**:
    - **Limitation**: Treating `children` as distinct groups (0, 1, 2, etc.) ignores potential trends that might be better modeled with a continuous variable or interaction effects.
    - **Impact**: This simplification might obscure more nuanced relationships.
4. **Exclusion of Other Variables**:
    - **Limitation**: The analysis focuses on `children`, `charges`, and `smoker` while ignoring other potentially relevant factors like age, BMI, and region.
    - **Impact**: This limited scope may fail to capture the full context influencing insurance charges.
5. **Generalizability**:
    - **Limitation**: The dataset may reflect specific demographics or regions, limiting the applicability of findings to broader populations.
    - **Impact**: Results may not be generalizable to all customers or geographic areas.
6. **Interaction Effects**:
    - **Limitation**: The Kruskal-Wallis test does not evaluate interactions (e.g., how the combined effect of smoking and the number of children impacts charges).
    - **Impact**: Interaction effects that might reveal significant insights are overlooked.

3. Recommend a course of action based on your findings.

1. **Adjust Insurance Premiums for Family Plans**:
    - **Action**: Use findings to refine pricing strategies for family plans based on the number of children.
    - **Rationale**: If the analysis shows significant differences in insurance charges by family size, premiums can be adjusted to reflect the higher healthcare costs associated with more prominent families.
    - **Benefit**: Aligning premiums with actual costs ensures fair pricing, increases profitability and reduces adverse selection.
2. **Develop Family-Focused Health Programs**:
    - **Action**: Implement wellness initiatives targeting families with multiple children, especially those in high-cost categories (e.g., smoking households).
    - **Rationale**: Families with more children may have more significant healthcare needs, and preventative programs can lower overall claims.

    - **Benefit**: Reduces long-term costs and improves customer health, enhancing loyalty and satisfaction.
3. **Stratify Analysis for Future Insights**:
    - **Action**: Conduct further analysis, stratifying by smoking status or other variables (e.g., region, age) to refine understanding how these factors interact with family size to affect costs.
    - **Rationale**: Smoking households with larger families may incur disproportionate costs that require additional attention.
    - **Benefit**: Helps design more targeted plans and policies for specific high-risk groups.
4. **Enhance Data Collection and Integration**:
    - **Action**: Improve data collection processes to minimize missing data and integrate additional variables such as healthcare utilization patterns.
    - **Rationale**: A more comprehensive dataset allows for deeper and more reliable analyses.
    - **Benefit**: Increases the precision of pricing models and provides a stronger foundation for decision-making.
5. **Educate Customers**:
    - **Action**: Inform policyholders about how family size and smoking habits influence premiums and encourage healthy behaviors through rewards or discounts.
    - **Rationale**: Empowering customers with knowledge promotes healthier lifestyles and better risk management.
    - **Benefit**: Reduces claims and strengthens customer relationships.

pandas documentation — pandas 2.2.3 documentation. (n.d.). *https://pandas.pydata.org/docs/*

Welcome to Python.org. (n.d.). Python.org. https://www.python.org/doc/

Kaplan, C. M., & Kaplan, E. K. (2020). State policies limiting premium surcharges for tobacco and their impact on health insurance enrollment. *Health Services Research*, *55*(6), 983–992. https://doi.org/10.1111/1475-6773.13577

Hassan, M. (2024, November 12). *ANOVA (Analysis of variance) - Formulas, Types, and Examples*. Research Method. https://researchmethod.net/anova/

Western Governors University. (n.d.). *Health Insurance Dataset.csv* [Dataset; ..Csv]. https://tasks.wgu.edu/student/012474228/course/33280017/task/4436/overview

matplotlib.pyplot.scatter — Matplotlib 3.10.0 documentation. (n.d.). https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.scatter.html

seaborn.boxplot — seaborn 0.13.2 documentation. (n.d.). https://seaborn.pydata.org/generated/seaborn.boxplot.html