Nicole Gallo
D599 – Data Preparation and Exploration
July 7, 2025
TCN1 Task 2: Data Exploration

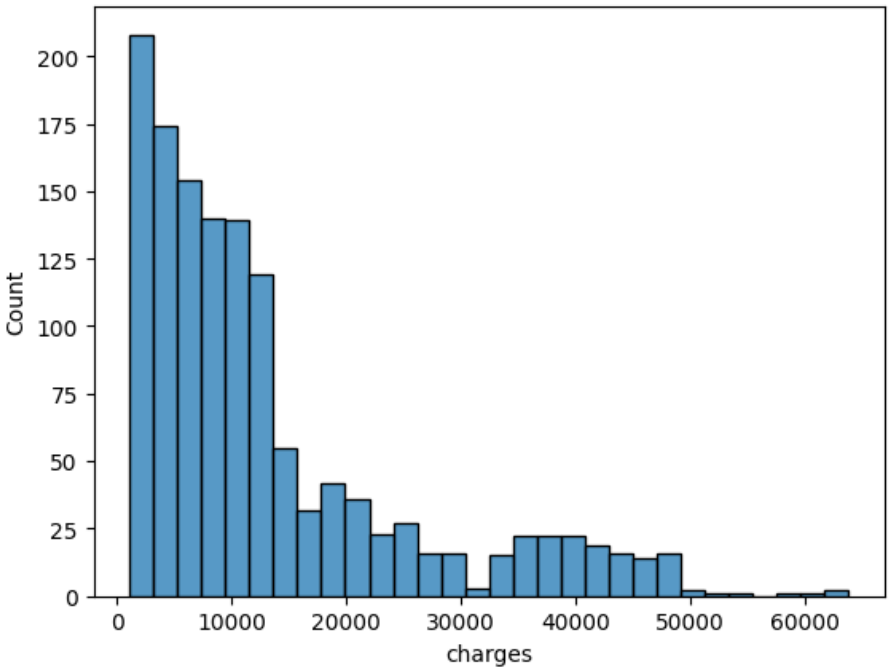
Part I: Univariate and Bivariate Statistical Analysis and Visualization

I chose two continuous variables (`Charges` and `BMI`) and two categorical variables (smoker and region). Using Python and Jupyter Lab environment, I performed the statistical calculations and generated the visualizations. I loaded the Excel file and did data cleaning to include the 1,338 valid rows in my analysis.

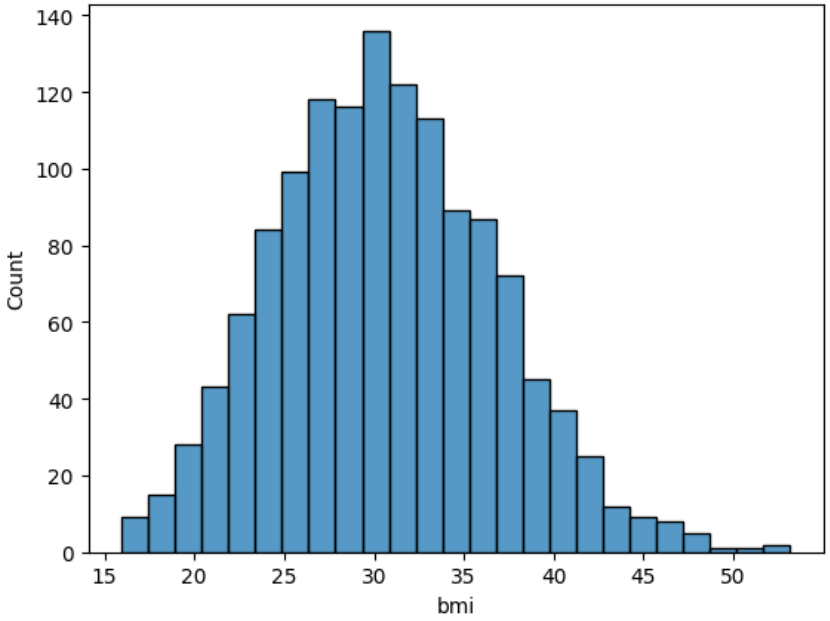
A. Identify the distribution of **two** continuous variables and **two** categorical variables using univariate statistics from the dataset.
   1. Represent your findings from part A visually as part of your submission.

Continuous Variables

1. `Charges`
   a. Python used: `df['charges'].describe()` and `sns.histplot(data=df, x='charges')`
   b. Univariate Statistics:
      i. Count: 1338.00
      ii. Mean: 13270.42
      iii. Median: 9382.03
      iv. Min / Max: 1121.87 to 63770.43
   c. Distribution: Right-skewed, meaning "long right tail", distribution slopes downward to the right, therefore, more people are being charged less in their insurance claims for medical expenses
   d. Visualization:

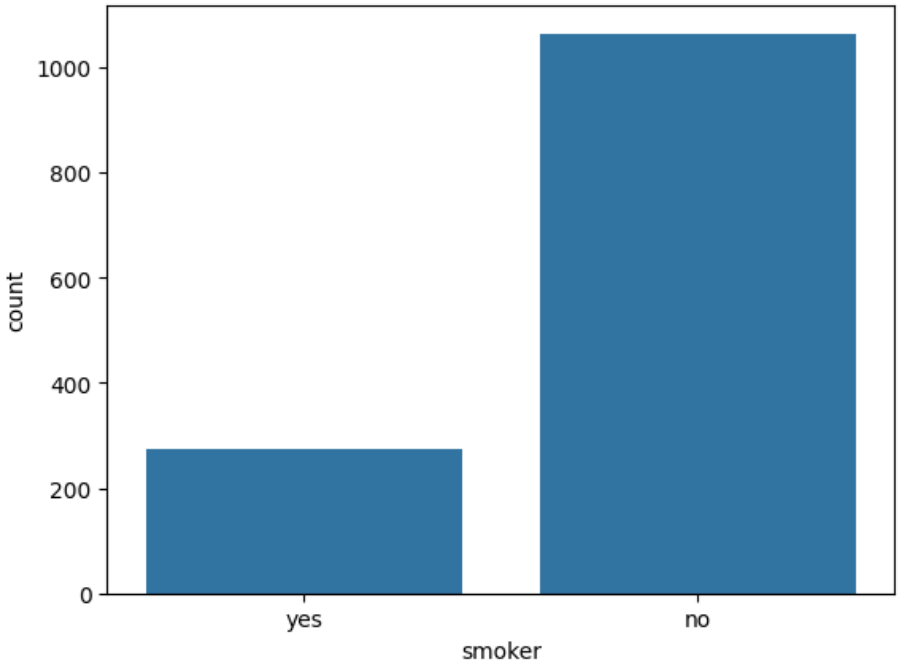


2. `BMI` (Body Mass Index)
    a. Python used: `df['bmi'].describe()` and `sns.histplot(data=df, x='bmi')`
    b. Univariate Statistics:
        i. Count: 1338
        ii. Mean: 30.66
        iii. Median: 30.40
        iv. Min / Max: 15.96 to 53.13
    c. Distribution: Roughly bell-shaped with a slight right skew; most BMI values falling between the 25-35 range
    d. Visualization:

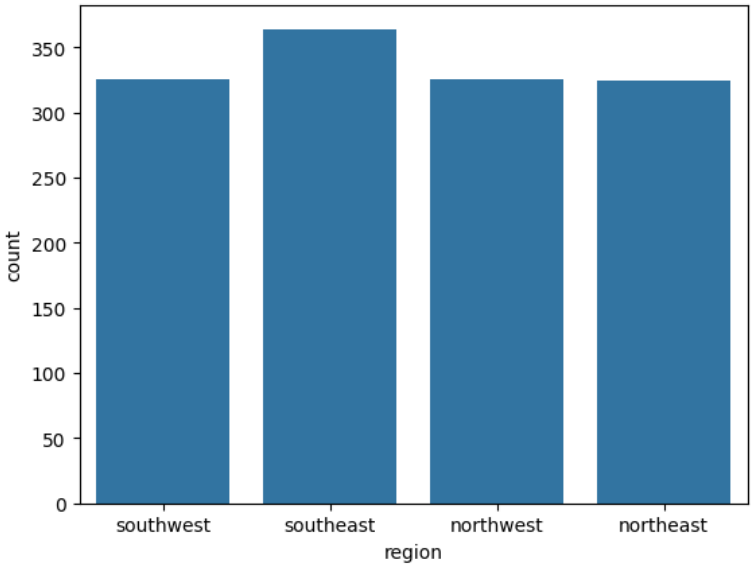


Categorical Variables

1. `Smoker`
    a. Python Used: `df.value_counts('smoker')` and `sns.countplot(data=df, x='smoker')`
    b. Counts:
        i. No: 1064
        ii. Yes: 274
    c. Distribution: Most people are non-smokers compared to those individuals who do smoke.
    d. Visualization:



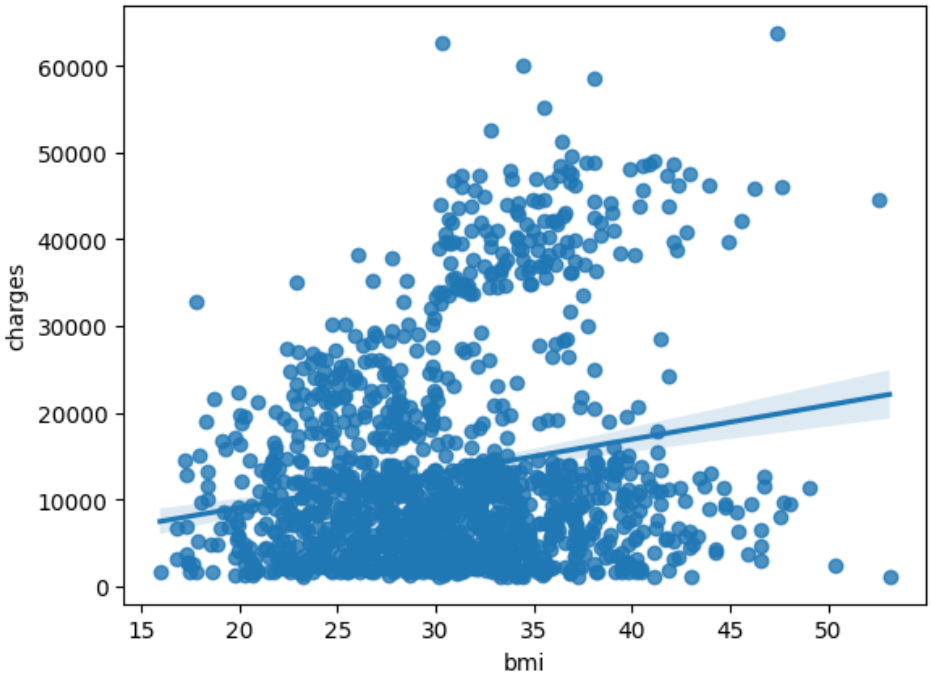
2. Region
    a. Python Used: `df.value_counts('region')` and `sns.countplot(data=df, x='region')`
    b. Counts:
        i. Southeast: 364
        ii. Northwest: 325
        iii. Southwest: 325
        iv. Northeast: 324
    c. Distribution: Fairly evenly distributed across all 4 regions, with slightly more of the beneficiaries' residences being located in the Southeast
    d. Visualization:



B. Identify the distribution of **two** continuous variables and **two** categorical variables using bivariate statistics from the dataset.

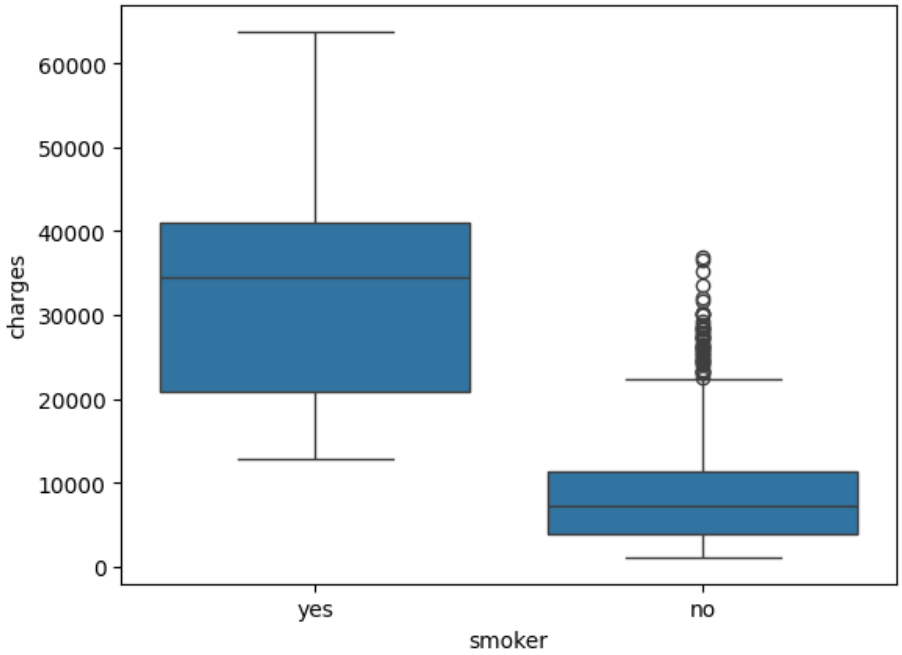
1. Represent your findings from part B visually as part of your submission.

    1. `Charges vs BMI`
        a. Type: Continuous vs. Continuous
        b. Python: `sns.regplot(data=df, x='bmi', y='charges')`
        c. Visualization:



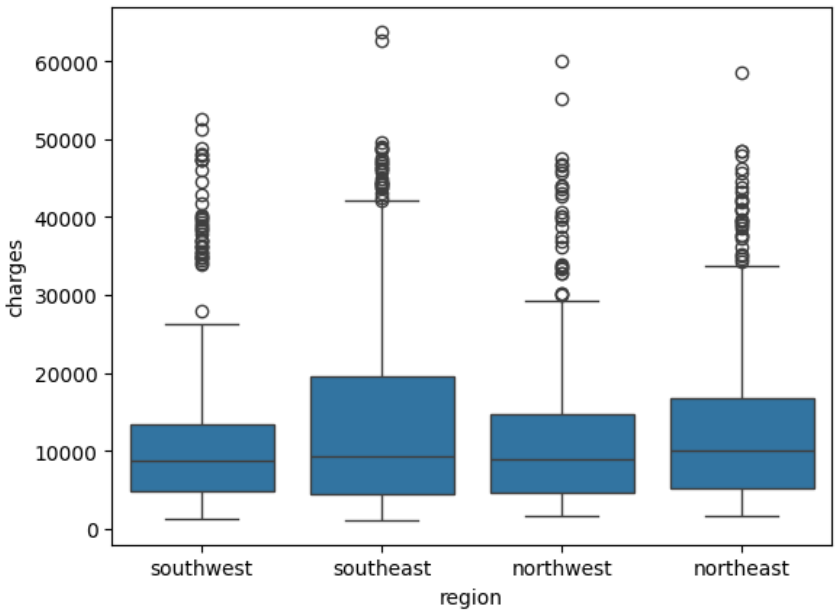
        d. Interpretation: Does BMI lead to higher charges? This scatterplot shows a slight positive trend, where people with higher BMIs tend to have higher insurance charges or claims.
        r=0.198 meaning r value is considered weak. The correlation isn't strong, but the regression line does show a general upward trend.

    2. `Smoker vs Charges`
        a. Type: Categorical vs. Continuous
        b. Python: sns.boxplot(data=df, x='smoker', y='charges')
        c. Visualization:



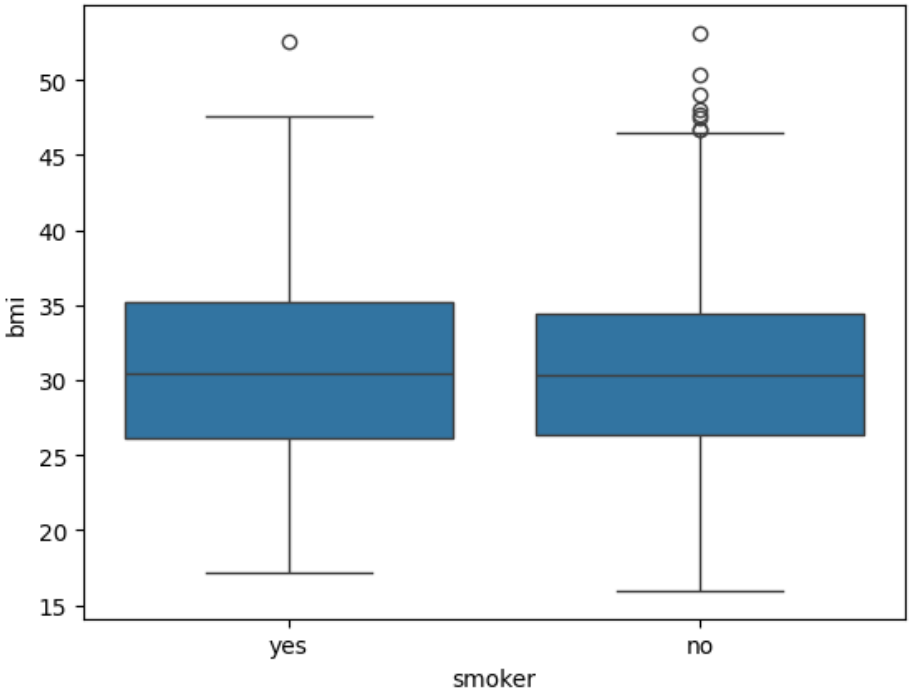
d. Interpretation: Are smokers charged more compared to non-smokers? Yes, smokers have significantly higher charges compared to non-smokers. The median (middle line in each boxplot) charges for smokers are much higher, and the range of values is broader due to outliers (dots above the box) and wider amounts of charges. This indicates that smoking is strongly associated with increased healthcare costs.

3. `Region vs Charges`
   a. Type: Categorical vs. Continuous
   b. Python: `sns.boxplot(data=df, x='region', y='charges')`
   c. Visualization:



d. Interpretation: Does region influence insurance claims? No. Average charges vary slightly across each region. The Southeast region has a slightly higher distribution of charges, but overall, charges appear relatively consistent among regions. In addition, the regions are similar in median value as well as the number of outliers within their data.

4. `Smoker vs BMI`
    a. Type: Categorical vs. Continuous
    b. Python: `sns.boxplot(data=df, x='smoker', y='bmi')`
    c. Visualization:



    d. Interpretation: Does being a smoker have an impact on the individuals' BMI? No, not a massive correlation between smoking and BMI. Smokers do appear to have a slightly higher median BMI than non-smokers, but not enough of a difference for it to have a meaningful impact. There are outliers in both smokers and non-smokers, and the BMI distribution for smokers is slightly wider in comparison.

Part II: Parametric Statistical Testing

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.

    Do smokers have significantly higher medical charges than non-smokers?

This question is highly relevant to health insurance organizations because smoking has been known to contribute to a variety of chronic illnesses and health complications. This can lead to increased medical costs for individuals.
Understanding whether there is a statistically significant difference in charges between smokers and non-smokers can help the health insurance organization:

- Better assess risk when pricing insurance policies
- Develop targeted wellness or how-to-stop-smoking programs
- Inform premium strategies and preventative outreach efforts

2. Identify the variables in the dataset that are relevant to answering your research question from part C1.

   The relevant variables we will be using to answer our research question are as follows:
   - `smoker` (categorical) – indicates whether the individual is a smoker (`yes` or `no`)
   - `charges` (continuous) – total medical charges claimed by the individual

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

   The parametric statistical test that we will be using is the **independent samples t-test**, which will compare the means (averages) of the two groups:
   - Group 1: smokers
   - Group 2: non-smokers

   This test answers:
   Are the average charges significantly different between smokers and non-smokers?

2. Develop null and alternative hypotheses related to your chosen parametric test from part D1.

   - Null Hypothesis:
     There is **no significant difference** in average charges between smokers and non-smokers

   - Alternative Hypothesis:
     There **is a significant difference** in average charges between smokers and non-smokers

3. Write code (in either Python or R) to run the parametric test.

```
[88]: # Split into smoker and non-smoker groups
      smokers = df[df['smoker'] == 'yes']['charges']
      non_smokers = df[df['smoker'] == 'no']['charges']

[89]: # Run independent samples t-test (Welch's t-test for unequal variances)
      t_stat, p_value = ttest_ind(smokers, non_smokers, equal_var=False)

[90]: # Output results
      print("T-statistic:", t_stat)
      print("P-value:", p_value)

      T-statistic: 32.751887766341824
      P-value: 5.88946444671698e-103
```

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

   Parametric Statistical Test Results
   - T-statistic: 32.75
   - P-value: 5.89 x 10^{-103}

   The t-statistic being 32.75 means the difference in average charges between smokers and non-smokers is 32.75 standard errors away from zero. It gives us very strong evidence against the null hypothesis, meaning **it is extremely unlikely that smokers and non-smokers have the same average medical charges.**

   Due to the p-value being less than 0.05, we reject the null hypothesis. There is strong statistical evidence that **smokers have significantly different (higher) average charges than non-smokers** based on the parametric statistical test results.

E. Evaluate parametric test results by doing the following:

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

   I chose the independent samples t-test because the goal was to compare the mean medical charges between two independent groups—smokers and non-smokers. The `charges` variable is continuous, and the `smoker` variable is categorical with two groups (`yes` and `no`). The sample size is large, and Welch's t-test (with `equal_var=False`) was used to account for unequal variance between the groups, making this an appropriate parametric test.

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

   The t-test returned a **t-statistic of 32.75** and **a p-value of 5.89 x 10^{-103}**. Since the p-value is significantly less than the common significance threshold of 0.05, **the null hypothesis is rejected**. This means that there is **a statistically significant difference in average medical charges** between smokers and non-smokers, with smokers incurring higher costs on average.

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

   The stakeholders in the organization benefit from this analysis because it provides clear, data-driven evidence that smoking is strongly associated with increased medical costs.

   They can use this information to do the following:
   - Refine premium pricing models
   - Identify high-risk customer segments
   - Develop targeted how-to-stop-smoking programs to reduce long-term costs

F. Summarize the implications of your parametric statistical testing by doing the following:
1. Discuss the answer to your question from part C1.

   We asked the question: "**Do smokers have significantly higher medical charges than non-smokers?**" Based on our analysis, we provided strong statistical evidence that **smokers have significantly higher medical charges** than non-smokers. The p-value from the independent samples t-test was below the 0.05 threshold, allowing us to confidently reject the null hypothesis. This suggests that smoking is a key factor contributing to increased healthcare costs.

2. Discuss the limitations of your data analysis.

   The limitations of our analysis are as follows:
   - The dataset only includes a limited set of health and lifestyle variables
   - The data is cross-sectional, meaning we observe a snapshot in time rather than changes over time
   - Charges can be influenced by multiple factors like age, BMI, or region, which were not controlled for in this analysis
   - The smoking variable is binary (yes or no) and doesn't reflect frequency or duration, which may affect severity and cost.

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

   Based on our analysis and our findings, the organization should consider these recommendations for a course of action:

   - Incorporating smoking status as a risk factor in premium pricing models
   - Offering targeted how-to-stop-smoking programs or incentives to reduce long-term costs
   - Conducting a multivariate analysis that includes other variables, like age, BMI, region, to refine the model further.

4. Submit your code for each test.
   See attached file: d599_task2.py

Part III: Nonparametric Statistical Testing

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.

   Does the distribution of insurance charges differ significantly across the four US regions?

   This question is great for a nonparametric test because it compares two independent groups, the `charges` variable is not normally distributed, and the business relevancy of this question will help determine if location influences cost.

2. Identify the variables in the dataset that are relevant to answering your research question from part G1.

   The relevant variables in the dataset that we will use to answer the research question are as follows:

   1. `region` (categorical with 4 levels) – Southeast, Northwest, Southwest, Northeast
   2. `charges` (continuous, skewed) – the total medical charges submitted by each person

H. Analyze the dataset further by doing the following:

1. Identify a **nonparametric** statistical test that is relevant to your question from part G1.

   Since we're comparing `charges` across 4 independent groups (`region`) and `charges` is a dependent, continuous, and not normally distributed variable, the **Kruskal-Wallis test** is what we will be using for this analysis.

   It will help us answer: **Are the median charges significantly different between regions?**

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

   Null Hypothesis:
   The distributions of charges are the **same across all regions**.

   Alternative Hypothesis:
   At least one region has a distribution of charges that is **significantly different**.

3. Write code (in either Python or R) to run the nonparametric test.

```
[91]: # Ensure 'charges' is numeric and drop rows with missing values
      df['charges'] = pd.to_numeric(df['charges'])
      df = df.dropna(subset=['charges', 'region'])

[92]: # Group charges by region
      regions = df['region'].unique()
      grouped_charges = [df[df['region'] == region]['charges'] for region in regions]

[93]: # Run the Kruskal-Wallis H-test
      h_stat, p_value = kruskal(*grouped_charges)

[94]: # Print results
      print("Kruskal-Wallis H-statistic:", h_stat)
      print("P-value:", p_value)

      Kruskal-Wallis H-statistic: 4.734181215658743
      P-value: 0.19232908072121044
```

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

   Nonparametric Statistical Test Results
   - H-statistic: 4.73
   - P-value: 0.192

   The results show an H-statistic of 4.73 and a p-value of 0.192. Since the p-value is greater than 0.05, we fail to reject the null hypothesis, meaning there is no statistically significant difference in medical charges among the 4 regions.

I. Evaluate nonparametric test results by doing the following:

1. Justify why you chose the statistical test identified in part G1 based on variables.

   I chose to use the Kruskal-Wallis test because the research question was focused on comparing the distributions of insurance charges across 4 independent groups. The charges variable is continuous but not normally distributed, making a nonparametric method more appropriate to use. This test does not assume equal variances or normality, making it good to use for real-world insurance data.

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

   The test results include a **H-statistic of 4.73** and a **p-value of 0.192**. Since the p-value is greater than the standard significance level of 0.05, we **fail to reject the null hypothesis**. This means there is **no statistically significant difference** in medical charges across the 4 US regions.

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

   Stakeholders in the organization will benefit from the Kruskal-Wallis test and its results because we can now conclude that geographic region does not appear to be a strong factor in determining medical charges.
   This information can also help with the following insights:

- Help simplify pricing models by not over-weighting geography
- Allow the organization to focus resources on more predictive factors like smoking, age, or BMI
- Reduce bias or assumptions about regional healthcare costs

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

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

   The research question is, "**Does the distribution of insurance charges differ significantly across the four US regions?"**. Based on the Kruskal-Wallis test, there is **no significant difference** in medical charges across the 4 geographic regions. This means that **region is not a key factor** in influencing the cost of medical care in this dataset.

2. Discuss the limitations of your data analysis.

   The limitations of our data analysis are as follows:
   - The test detects overall differences but doesn't identify which regions differ
   - The variable region may mask important intra-regional differences (e.g., urban vs rural)
   - Charges could be influenced by factors not included, such as healthcare provider networks, policy types, health history, etc.

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

   Since the regional differences are not statistically significant, our recommended course of action for the organization to explore is as follows:
   - Avoid over-weighting geography in cost models
   - Focus on more predictive variables such as smoking status, age, or BMI
   - Consider combining region with other features like lifestyle or accessibility to healthcare in multivariate modeling for further and deeper insights.

4. Submit your code for *each* test.
   See attached file: d599_task2.py

Part IV: Panopto Video Submission

K. https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=bae84fcd-955d-49f3-bf8b-b31400e3070f

Sources

L. Beyond WGU resources, I've used the following websites to confirm Python code structure for the visualizations in my analysis:
- https://www.datacamp.com/tutorial/how-to-make-a-seaborn-histogram
- https://seaborn.pydata.org/generated/seaborn.barplot.html
- https://seaborn.pydata.org/generated/seaborn.regplot.html