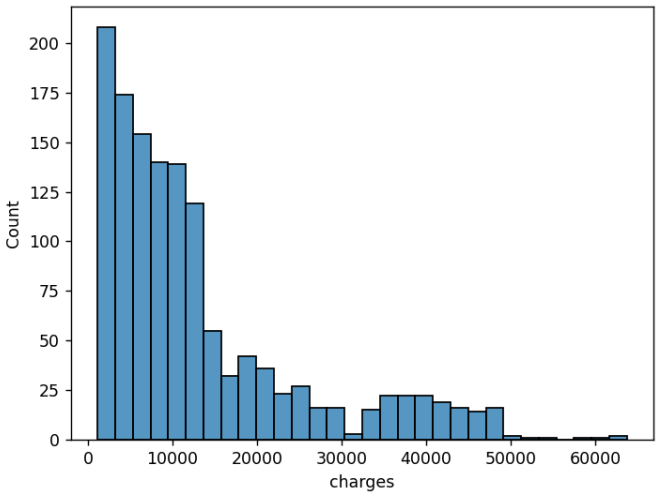
D599 Task 2

Part 1:

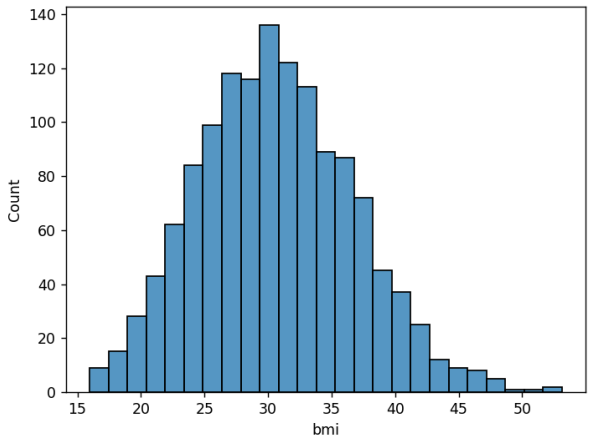
Part A:

1. Select four variables (e.g., two quantitative/numeric variables and two qualitative/categorical variables) and provide univariate visualizations for each variable selected.

Quantitative 1: Charges
Distribution: Right-skewed, due to the long right tail, telling us that most policyholders have fewer than 10000 in charges.

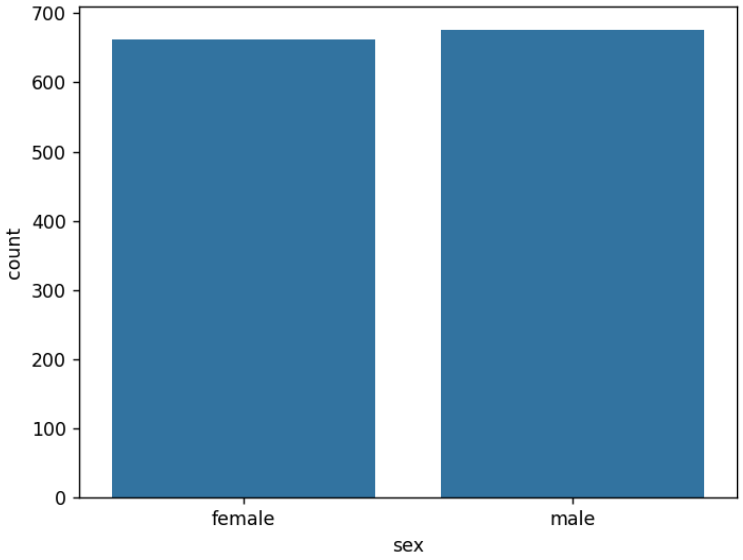


Quantitative 2: BMI
Distribution: Roughly bell-shaped with a slight right-skew. Most BMIs are between 25-35.



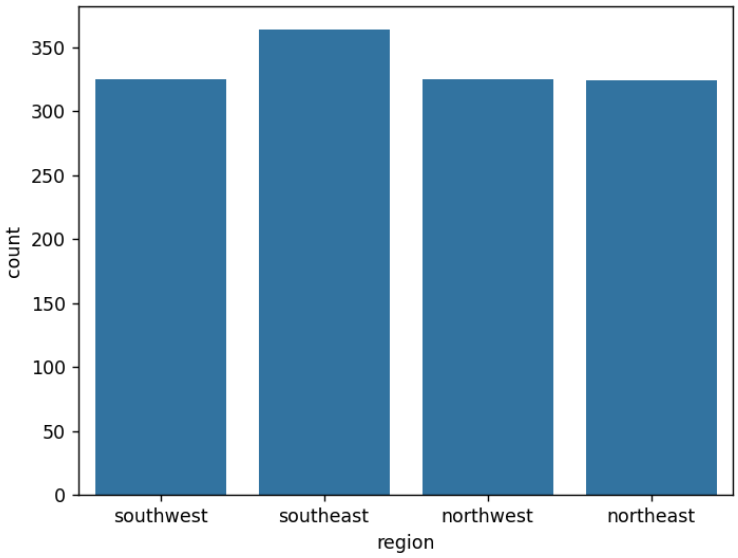
Qualitative 1: Sex

Distribution: Roughly the same amount of male and female policyholders with slightly more males.



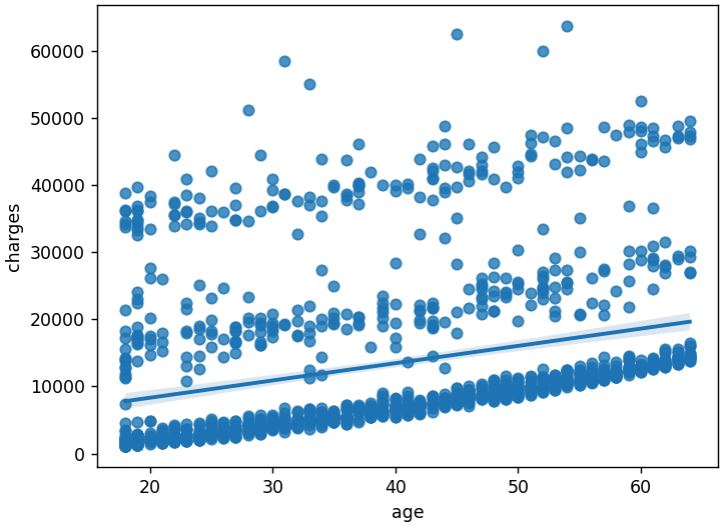
Qualitative 2: Region

Distribution: Roughly the same number of policyholders across all 4 regions with slightly more residing in the southeast region.

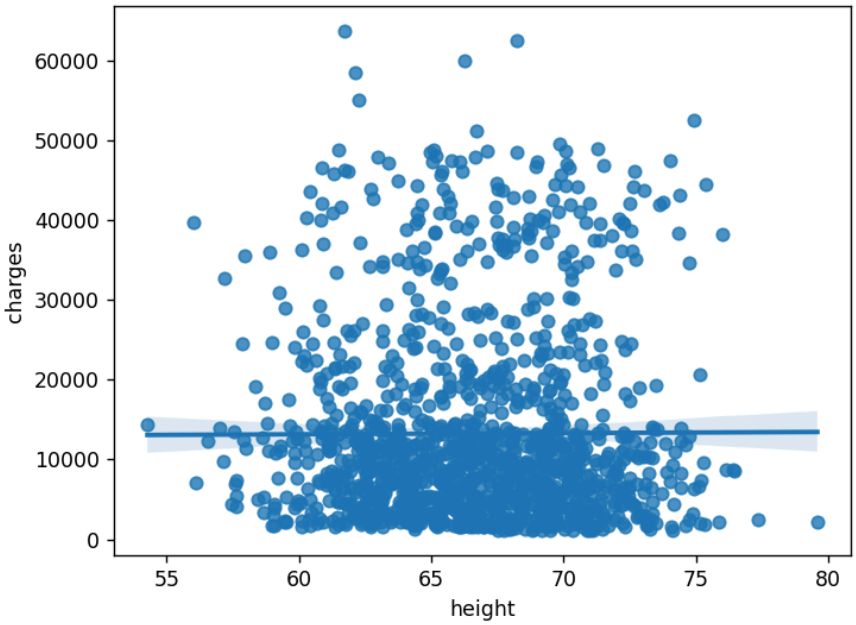


2. Provide two bivariate visualizations for each variable selected from part A1.

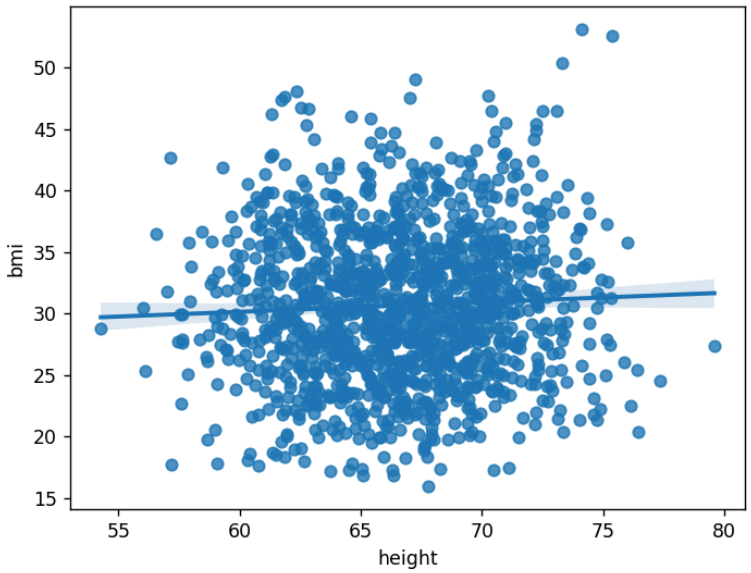
Charges v. Age



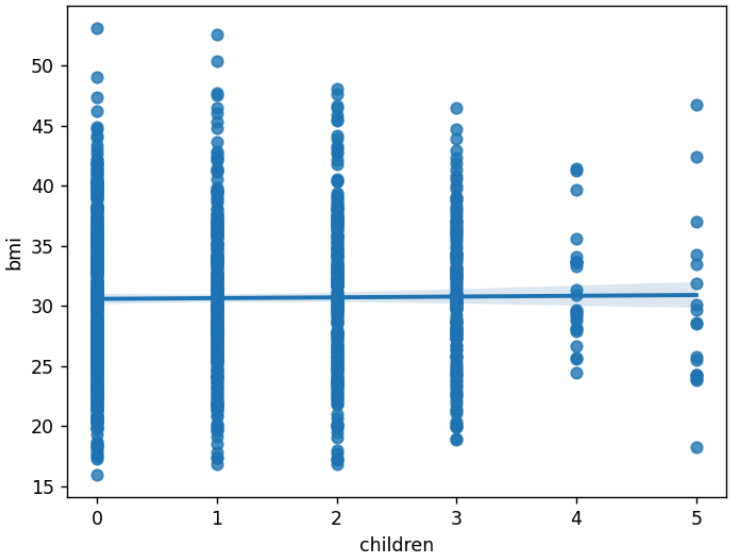
Charges v. Height



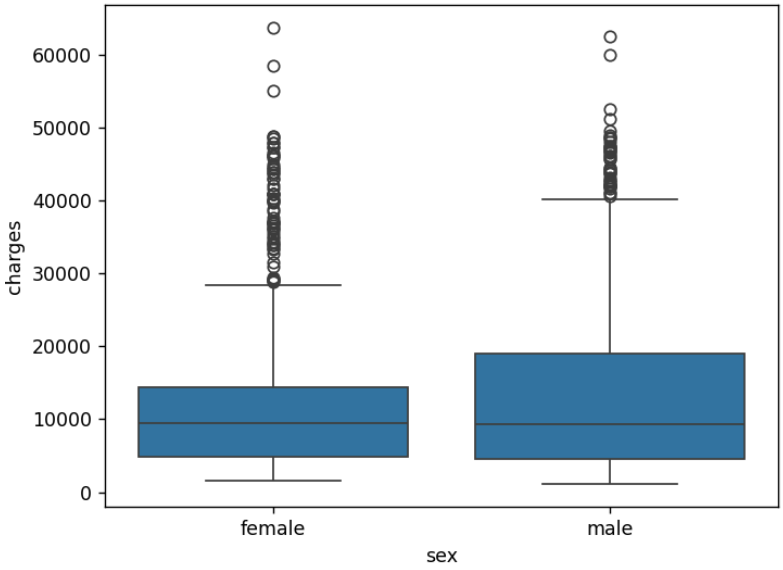
BMI v. Height



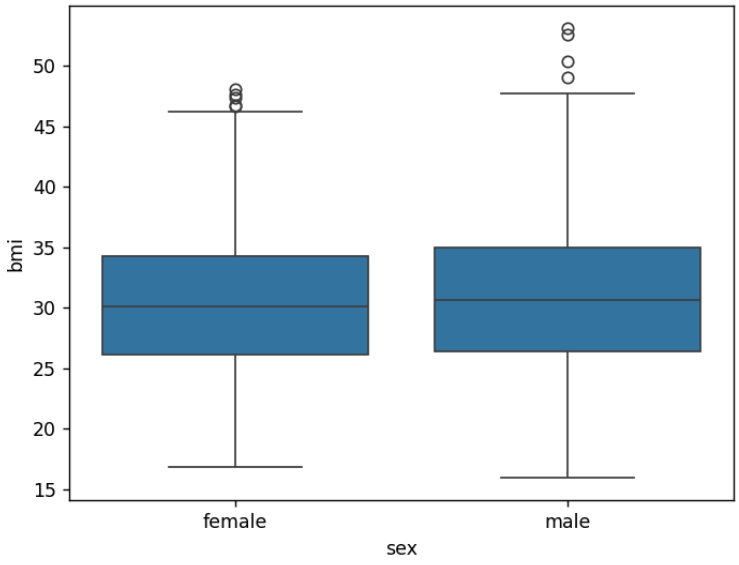
BMI v. Number of Children



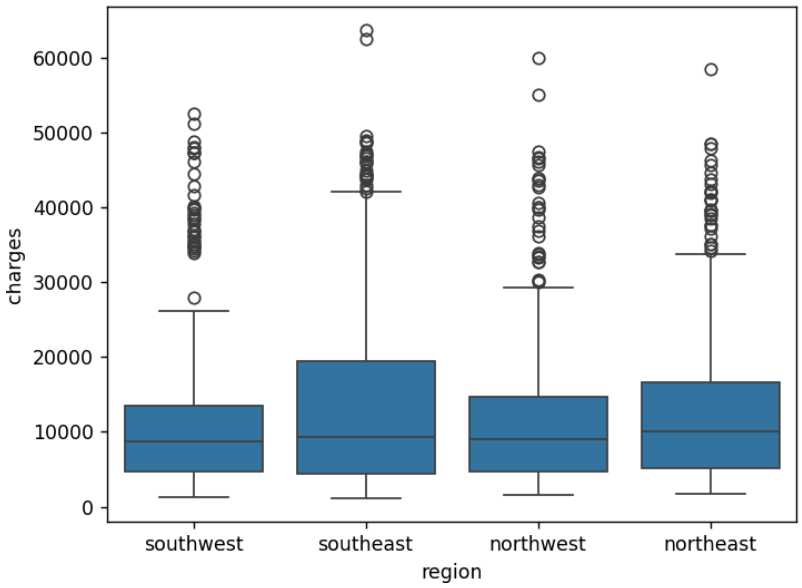
Charges v. Sex



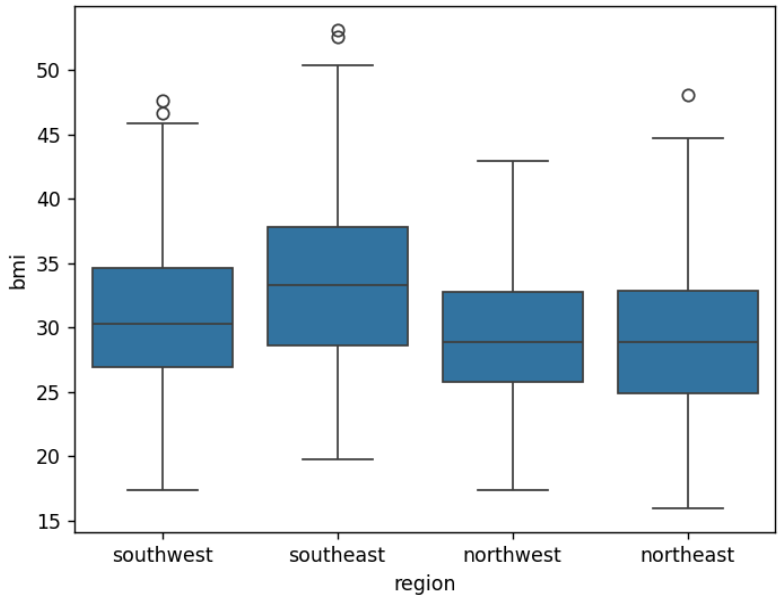
BMI v. Sex



Charges v. Region



BMI v. Region



B. Complete the following using the attached “Health Insurance Dataset” and R or Python:

1. Provide the descriptive statistics (e.g., mean, median, range, standard deviation, variance, percentiles, quartiles) for all quantitative (i.e., numeric) variables selected in the dataset.

Age:

count	1338.000000
mean	39.207025
median	39.000000
std	14.049960
var	197.25385
min	18.000000
25%(Q1)	27.000000
50%(Q2)	39.000000
75%(Q3)	51.000000
max(Q4)	64.000000

Charges:

count	1338.000000
mean	13270.422292
median	9382.033000
std	12110.011242
var	146542766.63
min	1121.873900
25%	4740.288500
50%	9382.033000
75%	16639.914427
max	63770.430000

BMI:

count	1338.000000
mean	30.663987
median	30.400000
std	6.098063
var	37.158582
min	15.960000
25%	26.296250
50%	30.400000
75%	34.693750
max	53.130000

Height:

count	1338.000000
mean	66.607682
median	66.601959
std	3.867210
var	14.944134
min	54.280000
25%	63.816668
50%	66.601959
75%	69.465980
max	79.578403

Children:

count	1338.000000
mean	1.094918
median	1.000000
std	1.205493
var	1.452127
min	0.000000
25%	0.000000
50%	1.000000
75%	2.000000
max	5.000000

Score:

count	1338.000000
mean	65.967115
median	66.000000
std	15.235684
var	231.95258
min	0.000000
25%	56.250000
50%	66.000000
75%	77.000000
max	100.000000

2. Provide the descriptive statistics (e.g., frequency counts and percentages) for all qualitative (i.e., categorical) variables in the dataset.

Sex:

male 676
female 662

male 50.523169%
female 49.476831%

Smoker:

no 1064
yes 274

no 79.521674%
yes 20.478326%

Region:

southeast 364
southwest 325
northwest 325
northeast 324

southeast 27.204783%
southwest 24.289985%
northwest 24.289985%
northeast 24.215247%

Level:

A 112
B 264
C 427
D 348
E 187

A 8.370703%
B 19.730942%
C 31.913303%
D 26.008969%

E 13.976084%

Part II: Parametric Statistical Testing

C. Describe a real-world organizational situation or issue in the attached "Health Insurance Dataset" by doing the following:

1. Create one research question that is relevant to the dataset and any organizational needs that can be answered through data analysis and is appropriate for parametric testing.

Research Question: Do smokers show significantly higher medical charges than those who don't smoke?

This question could prove very relevant to a health insurance company. It has been largely known that smoking can lead to a variety of health problems. This leads to increased medical expenses for the individual which is then passed along to the insurance company. Having clear evidence of the relationship between smoking and medical costs can allow the company to more accurately account for risk when pricing insurance policies or even create resources to help quit smoking which would reduce the company's costs.

D. Analyze the dataset by doing the following:

1. Identify a parametric statistical test that is relevant to your research question from part C1.

A useful parametric statistical test for working with this research question is an independent samples t-test.

This test will compare the means of the charges for smokers and the charges for non-smokers to determine if the average charges of smokers are significantly higher than the average charges of non-smokers.

2. List the dataset variables relevant to answering your research question from part C1.

The variables relevant to this statistical test are smoker(qualitative) and charges(quantitative).

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

An independent samples t-test will be perfect for answering our research question because we are comparing exactly two distinct groups: smokers and non-smokers. Our independent variable (smoker/non-smoker) is categorical and our dependent variable

(charges) is continuous. It is a large sample size. All of these point towards an independent samples t-test.

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

Null Hypothesis: There is no significant difference between the average charges of a smoker and a non-smoker.

Alternative hypothesis: There is a significant difference between the average charges of a smoker and a non-smoker.

5. Write error-free code in either Python or R to run the parametric test and provide the output and the results of all calculations from the parametric statistical test you perform.

```python
#import csv into Python
df = pd.read_csv(r"C:\Users\imret\Downloads\Health Insurance Dataset.csv")

#Split the data into smokers and non-smokers
smokers = df[df['smoker'] == 'yes']['charges']
non_smokers = df[df['smoker'] == 'no']['charges']

#Run independent samples t-test
t_stat, p_value = stats.ttest_ind(smokers, non_smokers, equal_var=False)

#Output
print("T-stat:",t_stat)
print("P-Value:",p_value)
```

```
PS C:\D599> & C:\Users\imret\AppData\Local\Programs\Python\Python313\python.exe "c:/D599/D599 Task 2.
py"
T-stat: 32.75188719942078
P-Value: 5.889493414643693e-103
```

E. Evaluate parametric test results by doing the following:

1. Discuss the test results, including the decision to reject or fail to reject the null hypothesis from part D4.

The two key results from the t-test are the T-stat and the P-value. The T-stat being so large (~32.75) tells us that the mean of the charges of smokers is significantly higher than the mean of the charges of non-smokers.

The P-value being so small (5x10^-103) gives us great confidence that the difference is real. This confidence allows us to reject the null hypothesis. There is strong statistical evidence that the charges of smokers are significantly different from the charges of non-smokers.

2. Create an answer to your research question from part C1 based on the decision to reject or fail to reject the null hypothesis.

The answer to the research question: Yes, smokers do show significantly higher medical charges than non-smokers.

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

Having statistical evidence that smokers generate higher medical charges than non-smokers can provide justification to stakeholders to raise prices for the insurance of smokers in order to offset increased charges. It can also justify implementing strategies to get policyholders to stop smoking, in order to save the company from having to pay increased charges.

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

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

Based on my findings, which state that smokers experience much higher medical charges than non-smokers, I would recommend that the company create a pricing structure that requires higher premiums for smokers, in order to offset the higher costs. Alongside this, I would recommend we launch a campaign to create resources to assist policyholders who would like to quit smoking. This could reduce long-term medical charges for the policyholder over time.

2. Discuss the limitations of your data analysis.

The biggest limitations of my data analysis include:

1. The analysis fails to take into account a large number of other variables related to smoking, such as, how often they smoke, how long they have been a smoker, were any of our non-smokers previously smokers who quit, etc.
2. The analysis only takes into account the smoking factor, not other factors that contribute to health, such as BMI or region, which were not controlled for in this analysis.

Part III: Nonparametric Statistical Testing

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

1. Create one research question that is relevant to the dataset and any organizational needs that can be answered through data analysis and is appropriate for nonparametric testing.

Research Question: Do charges significantly differ across the 4 different US regions?

H. Analyze the dataset further by doing the following:

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

The non parametric statistical test I will use to answer my question is a Kruskal-Wallis Test.

2. List the dataset variables relevant to answering your research question from part G1.

Relevant variables:
Charges (Quantitative)
Regions (Qualitative)

3. Justify why you chose the statistical test identified in part H1 based on variables.

A Kruskal-Wallis test works for this situation because we are dealing with multiple groups with the 4 different regions and the charges are not normally distributed, so we shouldn't use ANOVA.

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

Null Hypothesis: The distribution of charges across regions have no significant difference

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

5. Write error-free code in either Python or R to run the nonparametric test and provide a screenshot of the output and the results of all calculations from the nonparametric statistical test you performed.

```
#Ensure all charge values are numeric and drop all empty cells relevant to our test
df['charges'] = pd.to_numeric(df['charges'])
df=df.dropna(subset=['charges','region'])

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

#Run the Kruskal Wallis Test
h_value, p_value = stats.kruskal(*grouped_charges)

#Print outputs
print("H-Value:",h_value)
print("P-Value:",p_value)
```

```
H-Value: 4.734181215658743
P-Value: 0.19232908072121002
```

I. Evaluate nonparametric test results by doing the following:

1. Discuss the test results, including the decision to reject or fail to reject the null hypothesis from part H4.

The p-value from the test is greater than 0.05. This tells us that we do not have enough statistical evidence to reject the null hypothesis. Therefore we fail to reject our null hypothesis, which means there is not a significant difference in the distribution of charges across the different regions.

2. Create an answer to your research question from part G1 based on the decision to reject or fail to reject the null hypothesis.

Based on statistical evidence, there is no significant difference in the charges across the four US regions.

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

Stakeholders benefit from results of the Kruskal Wallis test by knowing that region does not have a significant impact on the charges the policyholders bring in. Knowing this, they can avoid putting too much weight into region when working on pricing strategy. This means they can devote the resources to other other factors which may have a greater impact on charges when deciding pricing strategy.

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

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

Based on these findings, I recommend that region not be overweighed when analyzing costs and deciding on pricing strategy. I also recommend putting resources into analyzing other factors (smoker/non-smoker, BMI, age, etc.) which might have a greater impact on costs.

2. Discuss the limitations of your data analysis.

This analysis is limited in some significant ways, first off, it analyzes differences across all four regions as a whole, but doesn't provide insight into differences between 2 specific

regions (i.e. Southeast v. Northeast). Another limitation is that regions are very broad, so you don't get clean insight into smaller areas, such as urban v. suburban v. rural areas within those regions.