Exploring Health Insurance Data

Unraveling Patterns in Insurance Charges

Team Members:
Dhanavikram Sekar, Hariharan Kumar, Indu Varshini Jayapal, Naveen Vinayaga Murthy, Nidhi Choudhary

Agenda

- Problem Statement
- Data Summary
- Univariate Analysis
- Multivariate Analysis
- Hypothesis Testing
- Conclusion
- Challenges and Limitations
- References

PROBLEM STATEMENT

Any interesting distribution patterns in the charges incurred?

What are the factors influencing the variability in insurance costs in the United States?

Data Source: https://github.com/stedy/Machine-Learning-with-R-datasets/blob/master/insurance.csv

DATA SUMMARY

1337

Number of records in the dataset

7

Features uniquely identifying a data point

Data Snapshot

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

Numerical Features

Age: 18 to 64
BMI: 15.96 to 53.13
Charges: $1,121 to $63,770

Categorical Features

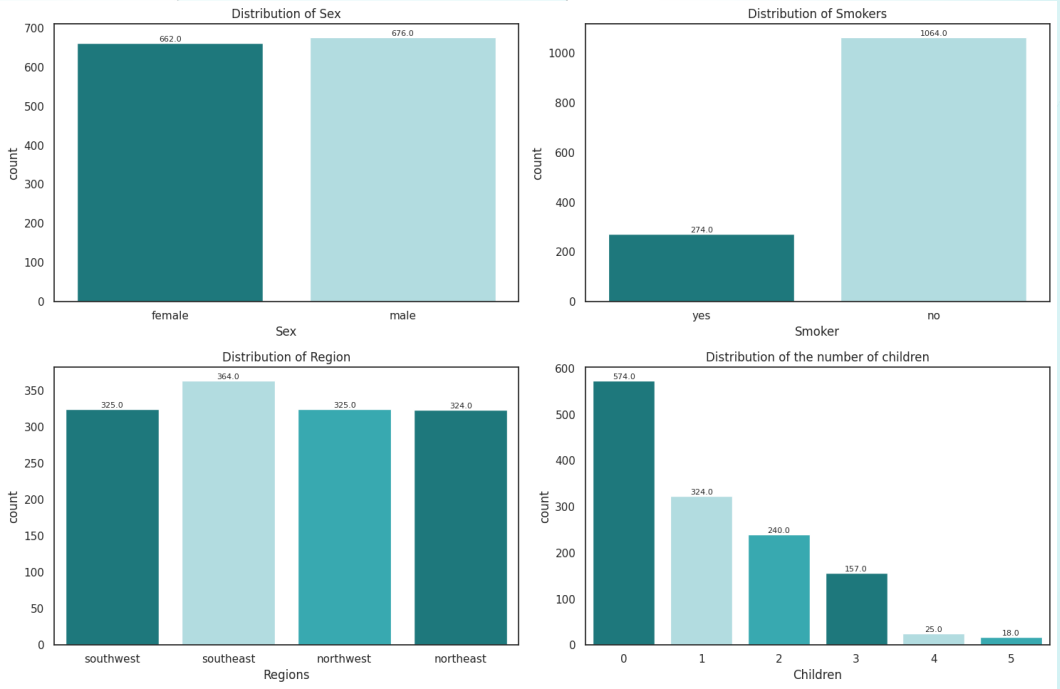
Sex: Male, Female
Children: 0 to 5
Smoker: Yes, No
Region: Northeast, Northwest, Southeast, Southwest

UNIVARIATE ANALYSES

UNIVARIATE ANALYSIS

Categorical Features

- Almost equal representation of male and female observations
- Imbalance between smokers and non-smokers, with non-smokers count being ~ 4x the count of smokers
- Balanced representation across southwest, northwest, and northeast regions with slightly higher number of observations from the southeast region.
- Number of children column is positively skewed

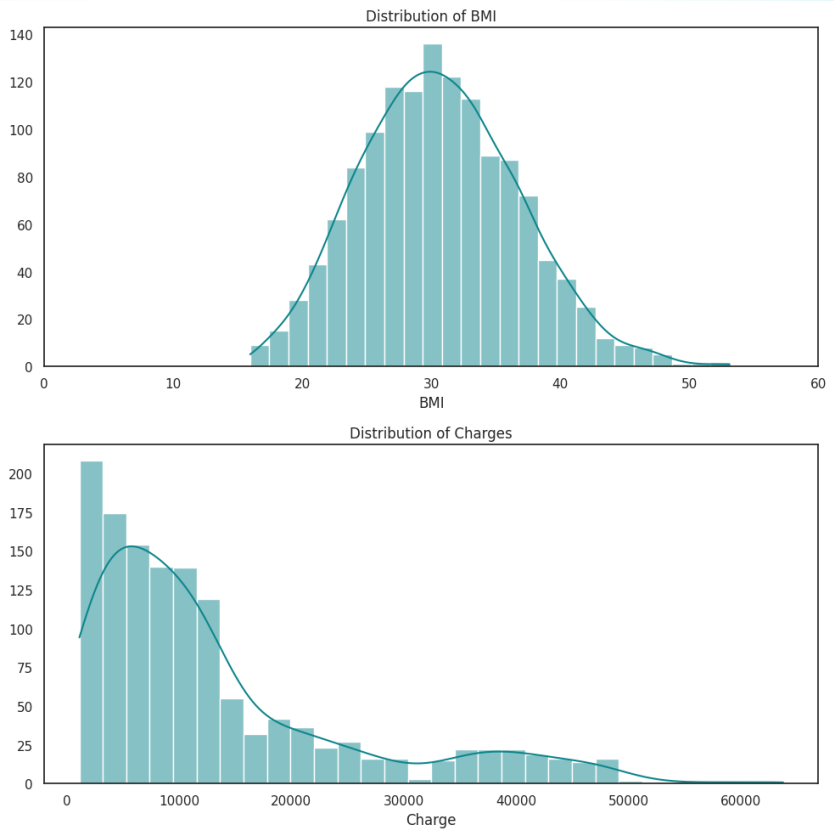


Countplots of Sex, Smokers, Region and Children Features

UNIVARIATE ANALYSIS

Numerical Features

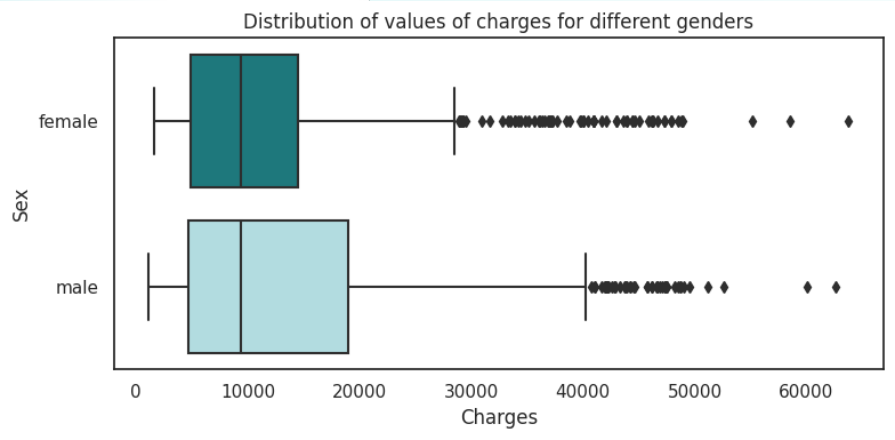
- Distribution of BMI is approximately ‘Normal’
- This symmetrical distribution indicates that a substantial portion of individuals have BMI around the mean which is ~30
- The charges histogram shows that majority of the insurance charges are clustered around the left tail, indicating that it is exhibits right-skewness
- This implies that a significant proportion of the individuals incur lower insurance costs with fewer individuals incurring higher costs

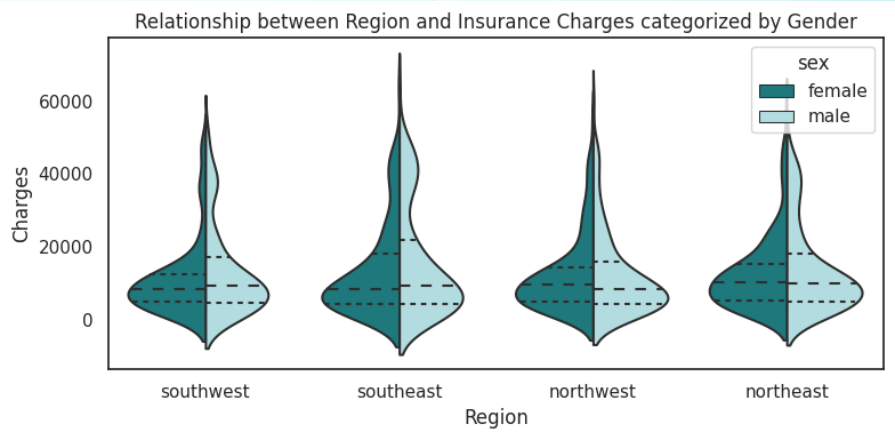


Histograms of BMI and Charges Features

MULTIVARIATE ANALYSES

Males tend to have higher charges than Females





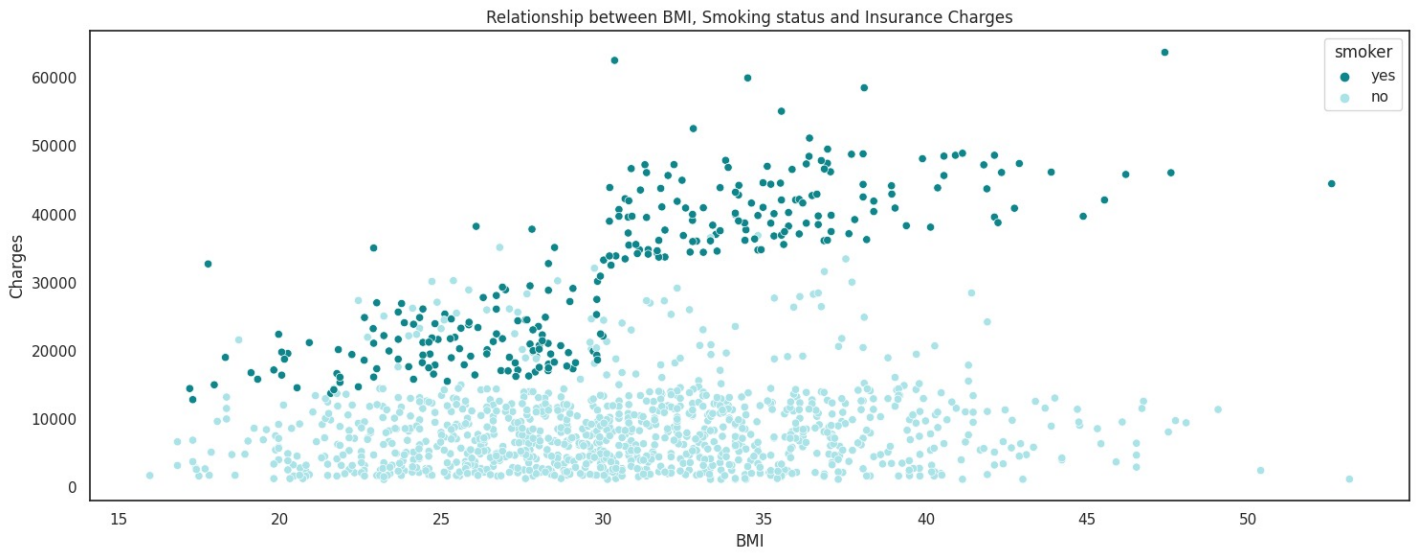
Gender and Charges:

- The median charge is similar for both genders
- However, a notable difference in the 75th percentile and maximum values suggests that males tend to have higher charges than females

Gender and Region Analysis:

- No significant difference in the distribution of charges is observed across genders within different regions
- Consistently across all regions, males tend to incur higher charges compared to females, extending the observation from the initial bar plot

Is there any relationship between BMI, Smoking and Insurance Charges?



Insights:

- Non-smokers generally exhibit lower charges despite having comparable BMI values to smokers
- No noticeable relationship is observed between BMI and charges
- The correlation matrix of BMI and Insurance Charges shows a low correlation value of approximately 0.2, indicating independence between the two

HYPOTHESES TESTING

HYPOTHESES TO TEST

01

Charges and Sex

"The insurance charges for men are greater than those for women."

02

Charges and Smoking

"The charges differ for both smokers and non-smokers."

03

Charges and Region

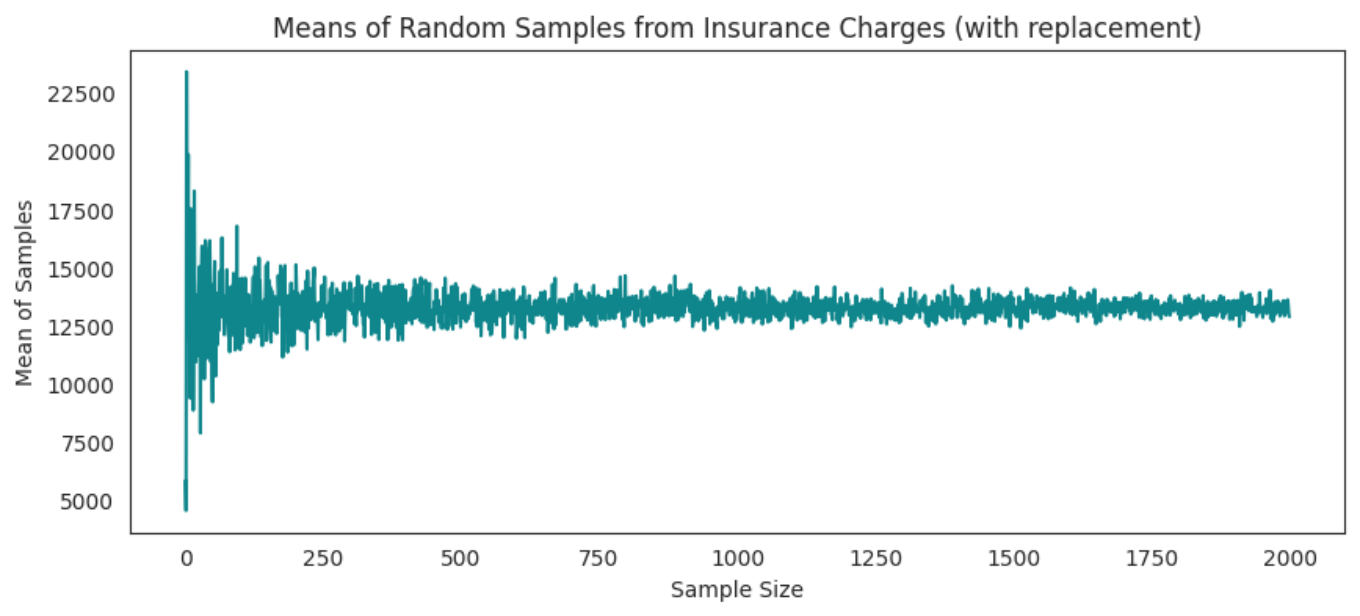
"Charges are equal for all regions."

04

Smokers and Sex

"The proportion of smokers is the same for both males and females."

How are we selecting a test for our target variable?



- In the insurance charges column, as the sample size increases, the distribution of sample means approaches normality
- T-test is appropriate here as we are dealing with a sufficiently large sample size, and the sample mean aligns with the principles of the central limit theorem

Hypothesis 1: The insurance charges for men are greater than those for women

H_0 : Charges of both men and women are equal
H_1 : Charges of men is greater than women

One tailed (Right tailed) Two Sample T-Test

Why did we choose this test?

Test Results

Critical Value: 0.05
t Value: 2.1275
P Value: 0.0169

Outcome

Reject the null hypothesis

Hypothesis 2: The insurance charges differ for both smokers and non-smokers

H_0 : Charges for both smokers and non-smokers are equal
H_1 : Charges for both smokers and non-smokers are not equal

Two tailed Two Sample T-Test

Why did we choose this test?

Test Results

Critical Value: 0.05
t Value: 46.6447
P Value: 1.4067e-282

Outcome

Reject the null hypothesis

Hypothesis 3: Insurance charges are equal for all regions

H_0 : Charges for all regions are equal
H_1 : Charges for all regions are not equal

ANOVA Test

Why did we choose this test?

Test Results

Critical Value: 0.05
f value: 2.9696
p value: 0.03089

Outcome

Reject the null hypothesis

Hypothesis 4: The proportion of smokers is the same for both males and females

H_0 : Proportion of smokers in male and female are same
H_1 : Proportion of smokers in male and female are not same

Chi-Square Test

Why did we choose this test?

Test Results

Critical Value: 0.05
Chi-square coefficient value: 7.39291081459996
p Value: 0.0065481435035806 96

Outcome

Reject the null hypothesis

CONCLUSION

Can we conclude that this holds true for the entire population in the United States?

01

From both the multivariate analysis and hypothesis testing, we found that males tend to have higher charges than females.

02

From the hypothesis testing, we found that the charges across the regions are not the same.

03

Our analysis reveals that gender, smoking status, and geographic location within the United States significantly influence the incurred insurance charges.

CHALLENGES

Data Collection

Procuring the insurance dataset with “Independent and Identically Distributed” datapoints

Sample to Population

Dataset is a small sample of the entire population, and the hypothesis testing results are only on this sample and not the population

REFERENCES

1. Data Source: https://github.com/stedy/Machine-Learning-with-R-datasets/blob/master/insurance.csv
2. https://www.kaggle.com/code/jordanrich/hypothesis-testing-of-health-insurance-data
3. https://www.kaggle.com/code/mayank2896/insurance-eda-hypothesis-testing
4. https://www.kaggle.com/code/yogidsba/insurance-claims-eda-hypothesis-testing/notebook
5. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html
6. https://www.reneshbedre.com/blog/anova.html
7. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chisquare.html
8. https://machinelearningmastery.com/chi-squared-test-for-machine-learning/
9. https://intapi.sciendo.com/pdf/10.2478/eoik-2019-0024
10. https://statisticsbyjim.com/basics/central-limit-theorem/
11. https://allendowney.blogspot.com/2013/08/are-my-data-normal.html

THANK YOU!