Q1. Business Case: Apollo Hospitals - Hypothesis Testing

Apollo Hospitals was established in 1983, renowned as the architect of modern healthcare in India. As the nation's first corporate hospital, Apollo Hospitals is acclaimed for pioneering the private healthcare revolution in the country.

As a data scientist working at Apollo 24/7, the ultimate goal is to tease out meaningful and actionable insights from Patient-level collected data.

You can help Apollo hospitals to be more efficient, to influence diagnostic and treatment processes, to map the spread of a pandemic.

One of the best examples of data scientists making a meaningful difference at a global level is in the response to the COVID-19 pandemic, where they have improved information collection, provided ongoing and accurate estimates of infection spread and health system demand, and assessed the effectiveness of government policies.

How can you help here?

The company wants to know:

• Which variables are significant in predicting the reason for hospitalization for different regions

• How well some variables like viral load, smoking, Severity Level describe the hospitalization charges

- Import the dataset and do usual exploratory data analysis steps like checking the structure & characteristics of the dataset
- Try establishing a relation between the dependent and independent variable (Dependent “hospitalization charges” & Independent: Smoker, Severity Level etc)
- Statistical Analysis:
    - Prove (or disprove) that the hospitalization of people who do smoking is greater than those who don't? **(T-test Right tailed)**
    - Prove (or disprove) with statistical evidence that the viral load of females is different from that of males **(T-test Two tailed)**
    - Is the proportion of smoking significantly different across different regions? **(Chi-square)**
    - Is the mean viral load of women with 0 Severity level , 1 Severity level, and 2 Severity level the same? Explain your answer with statistical evidence **(One way Anova)**
- Set up Null Hypothesis (H0)
- State the alternate hypothesis (H1)
- Check assumptions of the test (Normality, Equal Variance). You can check it using Histogram, Q-Q plot or statistical methods like levene’s test, Shapiro-wilk test (optional)
    - Please continue doing the analysis even If some assumptions fail (levene’s test or Shapiro-wilk test) but double check using visual analysis and report wherever necessary
- Set a significance level (alpha)
- Calculate test Statistics.
- Decision to accept or reject null hypothesis.
- Inference from the analysis

Apollo Hospitals - Hypothesis Testing

Apollo Hospitals was established in 1983, renowned as the architect of modern healthcare in India. As the nation's first corporate hospital, Apollo Hospitals is acclaimed for pioneering the private healthcare revolution in the country.

The ultimate goal is to tease out meaningful and actionable insights from Patient-level collected data. As per the business problem, the company wants to know:

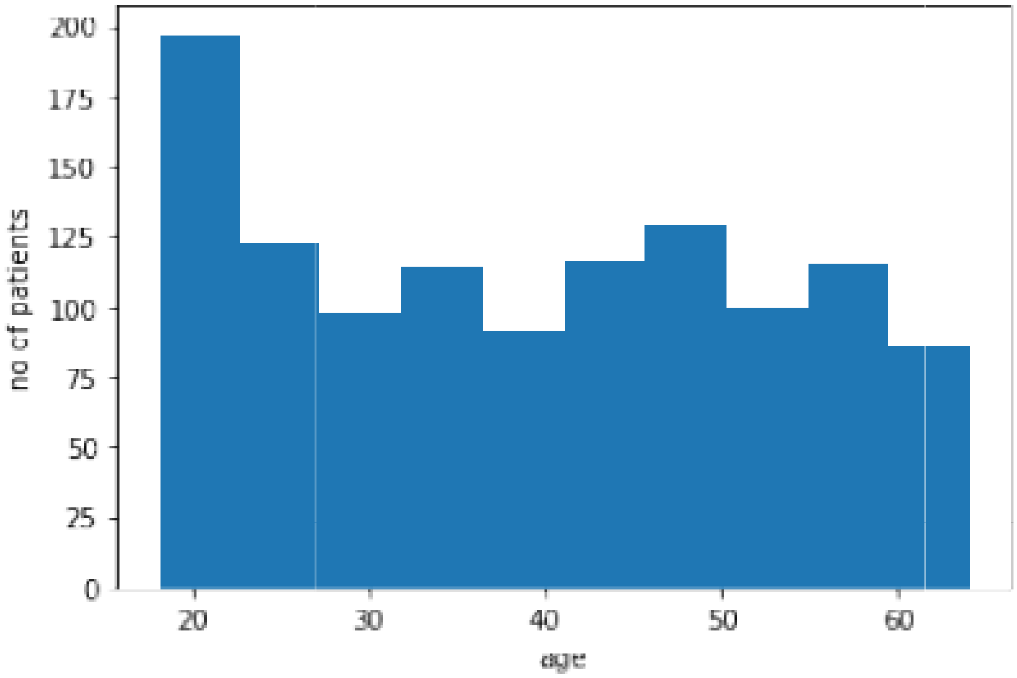
- Which variables are significant in predicting the reason for hospitalization for different regions.

- How well some variables like viral load, smoking, Severity Level describe the hospitalization charges.

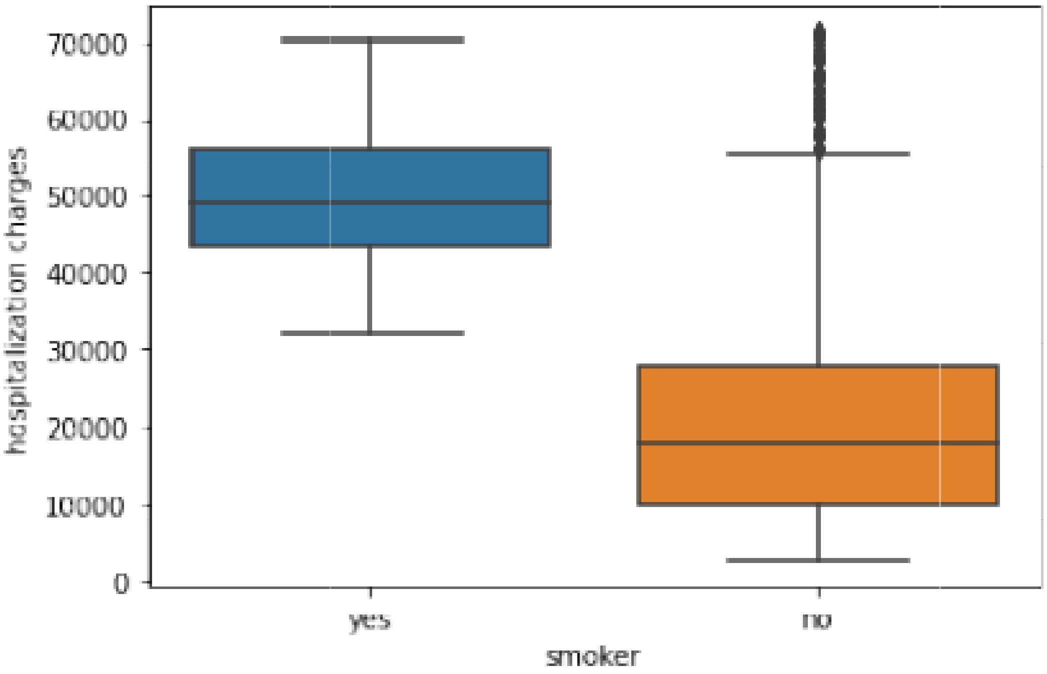
It was a simple data frame which consisted of Apollo patients data. I found the data very clean and properly structured. There were no missing values found. There were outliers found in **Hospitalization charges** feature. I treated them with the IQR 1.5 rule

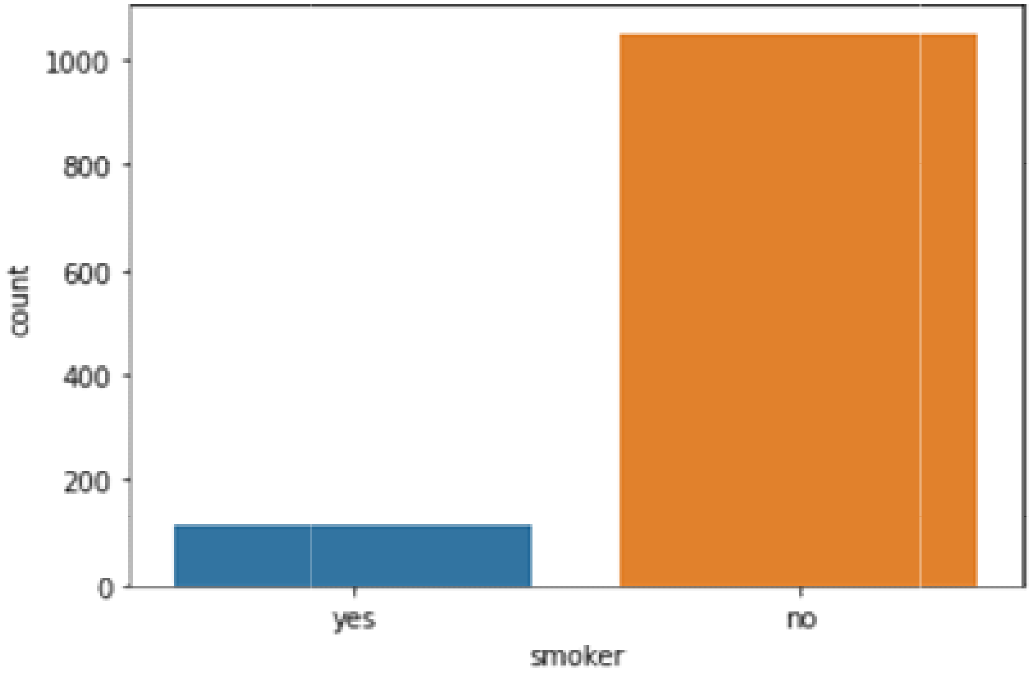
With the given data, I was able to draft certain observations which can help Apollo hospitals to be more efficient, to influence diagnostic and treatment processes, to map the spread of a pandemic. Please find the observations given below:

- Firstly, I used Pearson Correlation coefficient and Spearman rank correlation to check if there is any linear or monotonic relationship between ordinal categorical (**Age** and **Severity Level**) or continuous independent (**Viral Load**) variables and dependent (**Hospitalization charges**) variables. The only correlation was found in Age feature. It was a moderate positive correlation with monotonic relationship being comparatively higher as compared to linear. Hospitalisation charges increase w.r.t age. Keeping same profit margin, we have to start decreasing hospitalisation costs or charges of aged patients as the frequency of the number of patients are less as compared to early age groups.

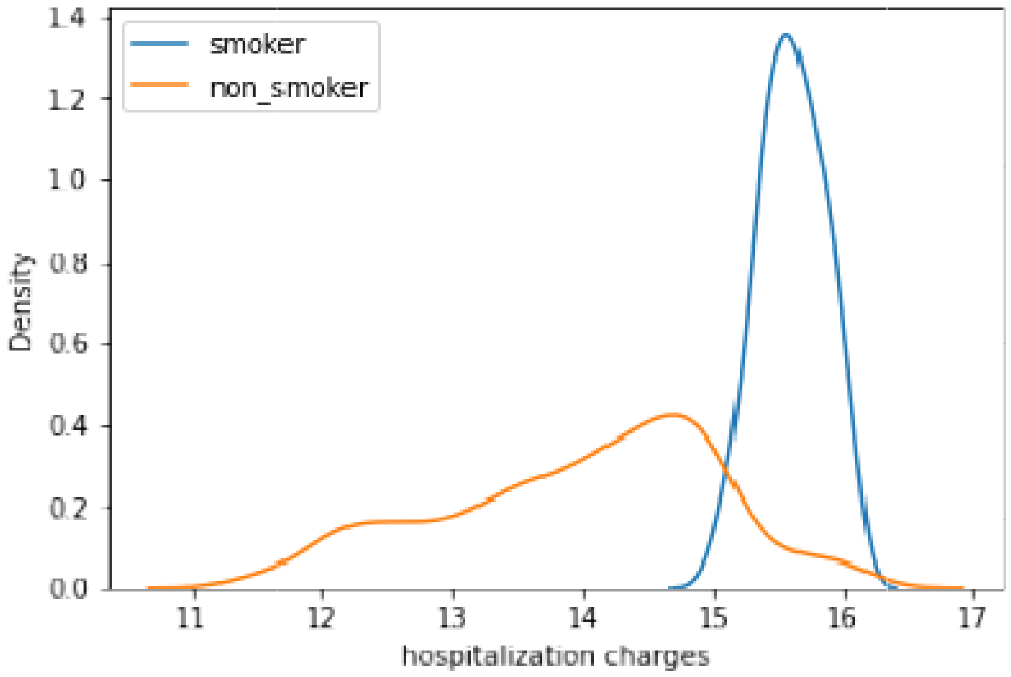


- Secondly, I used box-plots to compare nominal categorical (**Sex, Smoker** and **Region**) dependent and continuous independent (**Hospitalization charges**) variables. I only found a relation in smoker feature. People who smoke pay way higher costs as compared to non smokers. In fact, first quartile of smoker charges is higher than third quartile (75%) of non smoker charges. There are very less patients who smoke are being admitted to the hospital. This is a serious concern, it might be due to high hospitalisation charges. We can give a concession in the charges to bring in more patients who smoke.

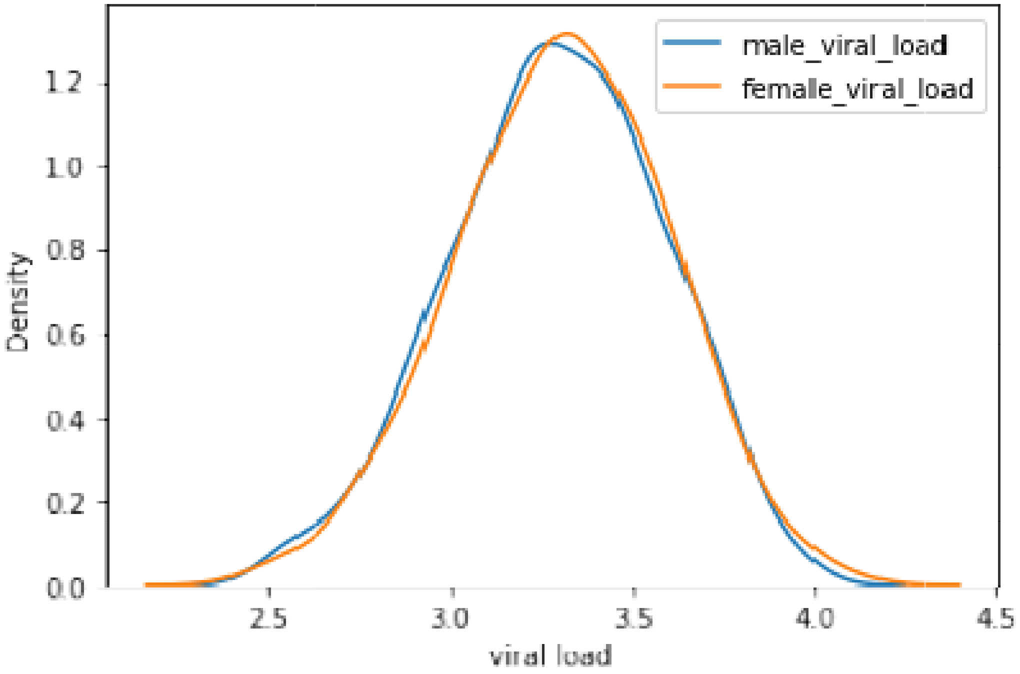




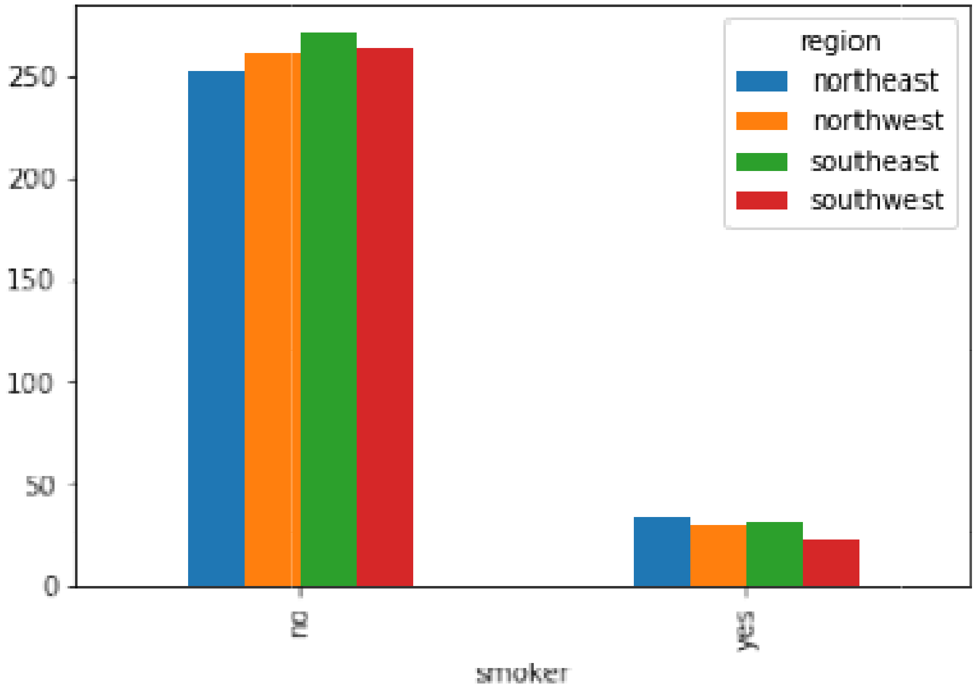
- I conducted a one tailed T-Test to Prove (or disprove) that the hospitalization of people who do smoking is greater than those who don't. I could see that the hospitalisation charges feature was not able to follow a normal distribution as per the qq plot. Even the log or box-cox transforms were not helpful. I still went ahead with the t test with the log transformed data. Other assumptions were satisfied though. After conducting the test, I had to reject Null hypothesis proving that the hospitalization charges of smokers is greater than those who don't.



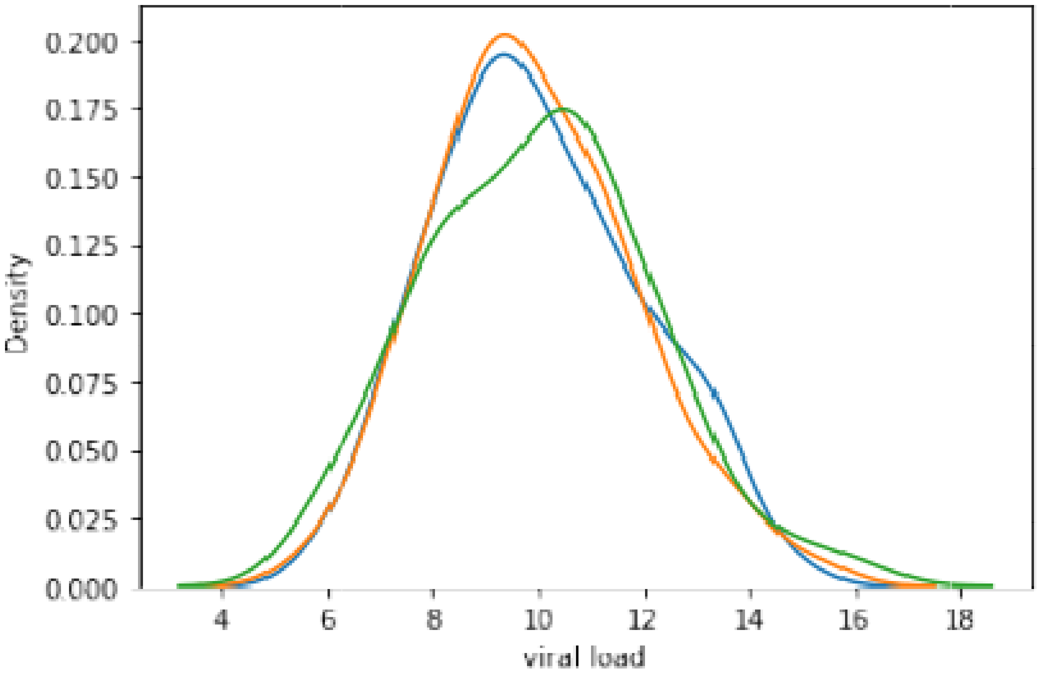
- I conducted a two tailed T-Test to Prove (or disprove) with statistical evidence that the viral load of females is different from that of males. I could see that the viral load feature was not able to follow a normal distribution as per the qq plot. Even the log or box-cox transforms were not helpful. I still went ahead with the t test with the log transformed data. Other assumptions were satisfied though. After conducting the test, I failed to reject Null hypothesis disproving with statistical evidence that the viral load of females is different from that of males.



- To check if proportion of smoking significantly different across different regions, I performed a Chi-square test. It is a non parametric test with only one assumption which is the data in the cells should be frequencies, or counts rather than percentages or some other transformation of the data. I was able to get this using pandas cross tab. After the test, I was able to come to a conclusion that there is no association or dependency between smoker and regions. Hence, proportion of smoking is not different across different regions.



- I also conducted a 1 way ANNOVA to check if mean viral load of women with 0 Severity level , 1 Severity level, and 2 Severity level the same. I could see that the viral load feature was not able to follow a normal distribution as per the qq plot. Even the log or box-cox transforms were not helpful. I still went ahead with the t test with the log transformed data. After conducting the test, I failed to reject Null hypothesis concluding with statistical evidence that mean viral load of women with 0 Severity level , 1 Severity level and 2 Severity level is the same.



To whoever reads this, I hope my insights from this case study were meaningful.

Thank you,
Krishna