Axis Insurance Project

This file is meant for personal use by d.blythe@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Objective

Do statistical analysis and extract actionable insights from the data

We will be majorly focusing on these problems -

- Extracting insights using Exploratory Data Analysis.
- Prove (or disprove) that the medical claims made by the people who smoke is greater than those who don't?
- Prove (or disprove) with statistical evidence that the BMI of females is different from that of males.
- Is the proportion of smokers significantly different across different regions?
- Is the mean BMI of women with no children, one child and two children the same? Explain your answer with statistical evidence.

This file is meant for personal use by d.blythe@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Data Information

Variable	Description
Age	This is an integer indicating the age of the primary beneficiary (excluding those above 64 years, since they are generally covered by the government).
Sex	This is the policy holder's gender, either male or female.
BMI	This is the body mass index (BMI), which provides a sense of how over or under-weight a person is relative to their height. BMI is equal to weight (in kilograms) divided by height (in meters) squared. An ideal BMI is within the range of 18.5 to 24.9.
Children	This is an integer indicating the number of children / dependents covered by the insurance plan.
Smoker	This is yes or no depending on whether the insured regularly smokes tobacco.
Region	This is the beneficiary's place of residence in the U.S., divided into four geographic regions - northeast, southeast, southwest, or northwest.
Charges	Individual medical costs billed by health insurance

Observations	Variables
1338	7

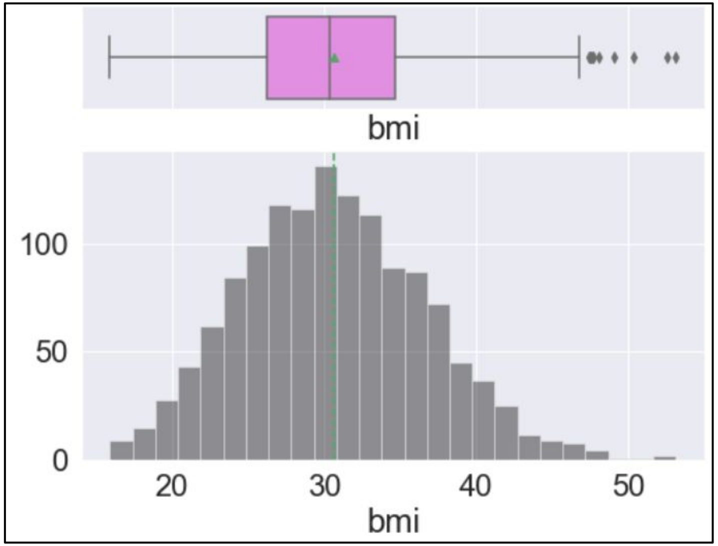
Note:

- There are no missing values in the dataset
- The sex, smoker and region columns have been converted to category

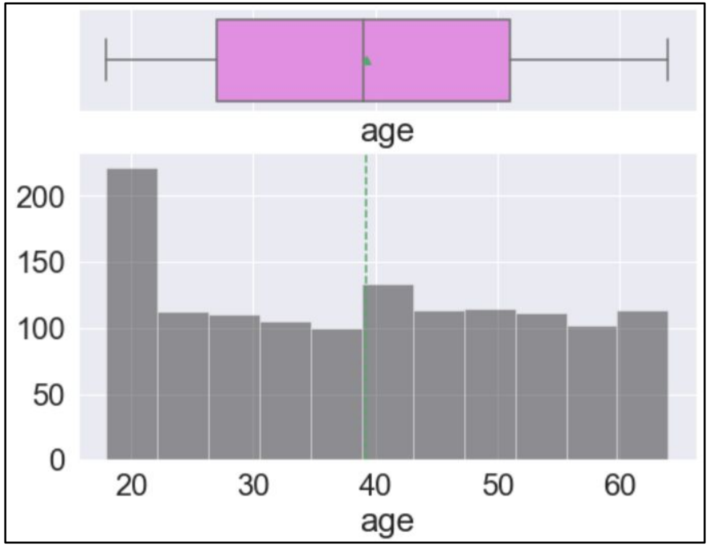
This file is meant for personal use by d.blythe@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Exploratory Data Analysis – Age & BMI

BMI



Age



- BMI looks to have a fairly normal distribution

- Age seems uniformly distributed, with both mean and median around 40 years.

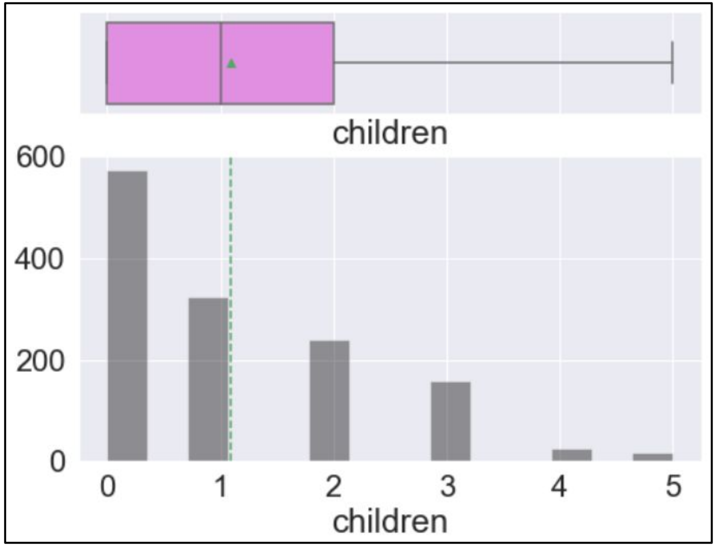
This file is meant for personal use by d.blythe@gmail.com only.

Sharing or publishing the contents in part or full is liable for legal action.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

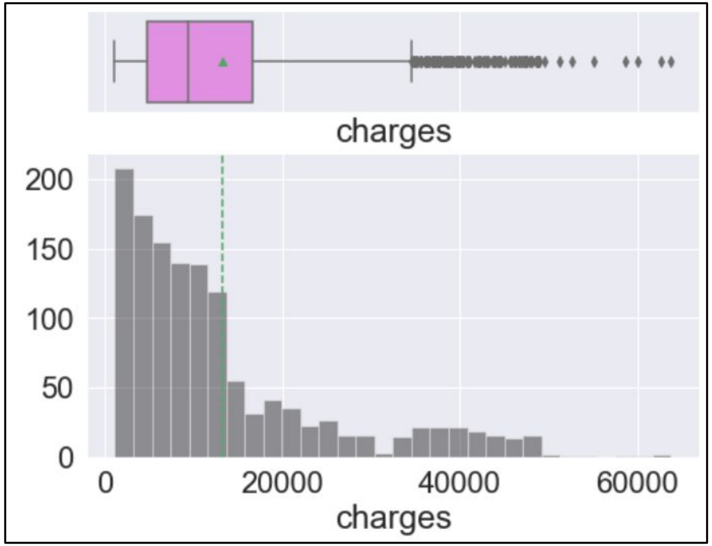
Exploratory Data Analysis – Children & Charges

Children



- The number of children has a left skewed distribution.
- The plot suggests that we should convert the children variable to categorical for further analysis.

Charges



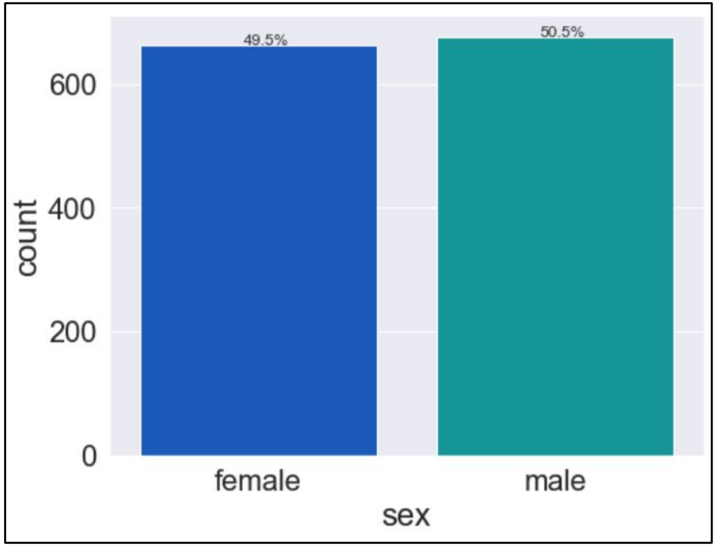
- Charges have a right skewed distribution. The mean charges is higher than the median charges
- This variable has a lot of outliers towards the higher end indicating that some people spend very high on their medicals.

This file is meant for personal use by d.blythe@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

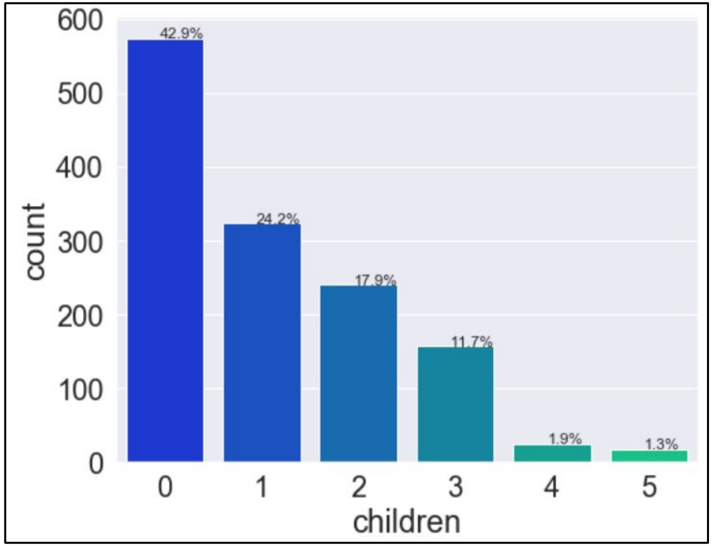
Exploratory Data Analysis – Sex & Children

Sex



- The distribution of observations across genders is fairly similar as we saw earlier as well.

Children



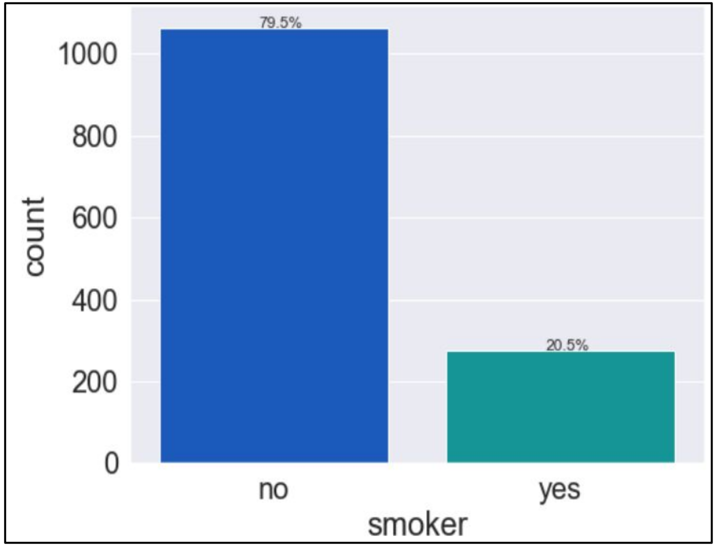
- Nearly 42% insurance holders do not have a child.
- Nearly 42% insurance holders have 1 or 2 children.

This file is meant for personal use by d.blythe@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

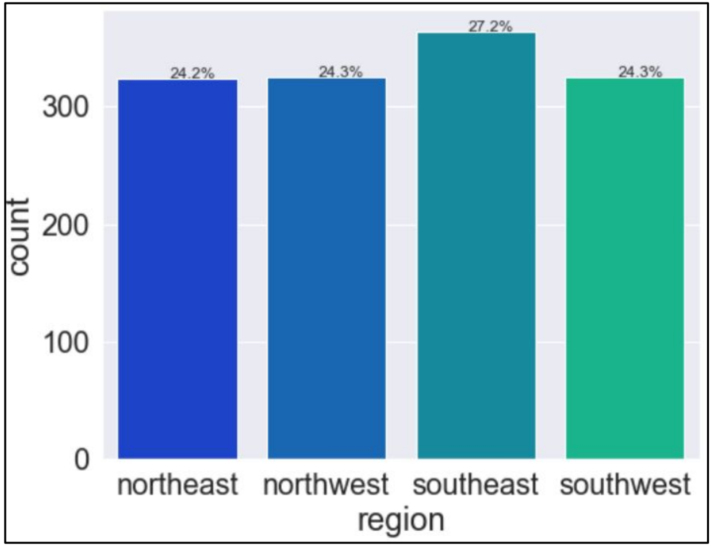
Exploratory Data Analysis – Smoker & Region

Smoker



- 20% of the insurance holders are smokers. It will be interesting to see how smoking affects the insurance claims.

Region



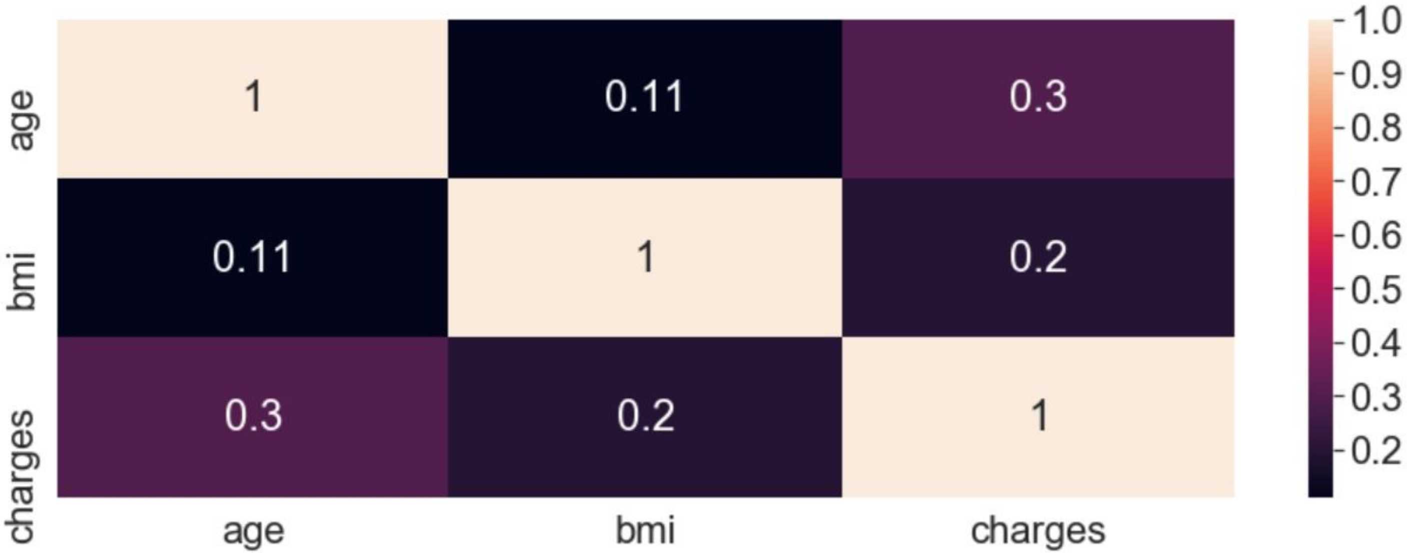
- The distribution of insurance holders across various regions of US is fairly uniform. South east region does have ~3% more observations as compared to others but we will have to test if this difference is statistically significant

This file is meant for personal use by d.blythe@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Exploratory Data Analysis - Correlation matrix

Correlation matrix



- The correlation between between all the continuous variables is positive but not very high.

This file is meant for personal use by d.blythe@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Hypothesis Testing - Medical cost

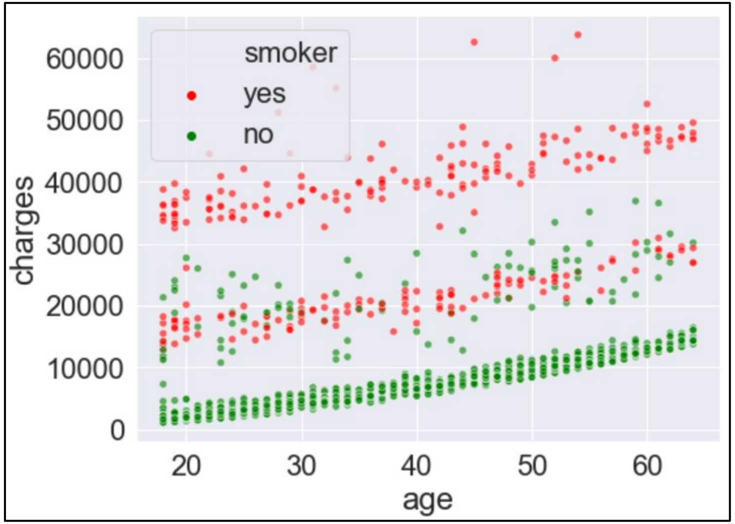
Problem: Prove(or disprove) that the medical claims made by the people who smoke is greater than those who don't?

- Null Hypothesis = Ho = "Mean charges of smokers is less than or equal to non-smokers."

- Alternate Hypothesis = Ha = "Mean charges of smokers is greater than non-smokers."

By using Independent t-test, we get the p-value is 4.13e-283 that is <0.05.

Therefore, we reject the null hypothesis that the mean charges of smokers is less than or equal to non-smokers.



- Visually the difference between charges of smokers and charges of non-smokers is apparent.
- The non-smokers have much lower medical bill claims compared to the smokers.
- We will have to perform a two sample t-test to test to check if the mean charges of smokers and non-smokers is indeed different.

This file is meant for personal use by d.blythe@gmail.com only.

Sharing or publishing the contents in part or full is liable for legal action.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

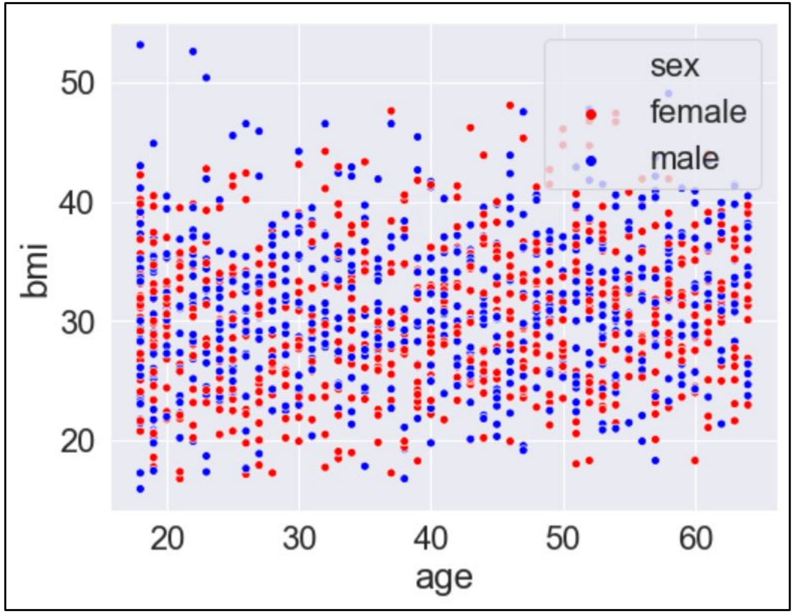
Hypothesis Testing - BMI

Problem: Prove (or disprove) with statistical evidence that BMI of females is different from that of males.

- Null Hypothesis = Ho = "Mean BMI of females is same as that of males"
- Alternate Hypothesis = Ha = "Mean BMI of females is different from males"

By using Independent t-test, we get the p-value is 0.0899 that is >0.05.

Therefore, we fail to reject the null hypothesis that the mean BMI of females is same as that of males.



- Visually, there is no apparent relation between gender and BMI

This file is meant for personal use by d.blythe@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

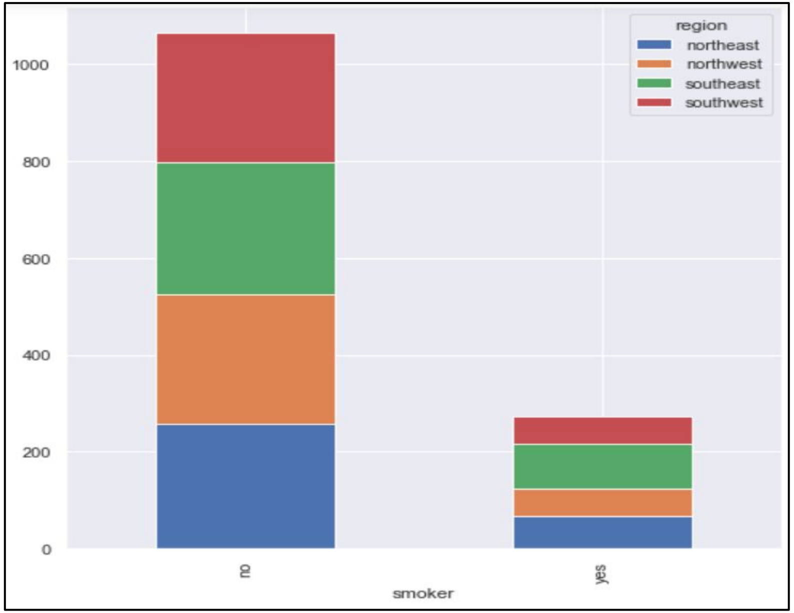
Hypothesis Testing - Smokers across region

Problem: Is the proportion of smokers significantly different across different regions?

- Null Hypothesis = Ho = "Region has no effect on smoking habits"
- Alternate Hypothesis = Ha = "Region has an effect on smoking habits"

By using chi-square test, we get the p-value is 0.062that is >0.05.

Therefore, we fail to reject the null hypothesis that the region has no effect on smoking habits.



- The proportion of smokers in southeast region is higher than others.

This file is meant for personal use by d.blythe@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Hypothesis Testing - BMI of women

Problem: Is the mean BMI of women with no children, one child and two children the same? Explain your answer with statistical evidence?

- Null Hypothesis = Ho = "No. of children has no effect on bmi"
- Alternate Hypothesis = Ha = "No. of children has an effect on bmi"

By using anova test, we get the p-value is 0.716 that is >0.05.

Therefore, we fail to reject the null hypothesis that the no. of children has no effect on bmi.

This file is meant for personal use by d.blythe@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Conclusion

Based on our previous analysis, we can conclude that:

- The claims made by smoker are higher as compared to the non-smokers. We should create personalised policies for these customer categories.
- Very few people have more than 2 children. 75% of the people have 2 or less children. However number of children has no effect on BMI of the women insurance holders.
- BMI has a slight positive correlation with the medical claims.

This file is meant for personal use by d.blythe@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

greatlearning

Power Ahead

This file is meant for personal use by d.blythe@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.