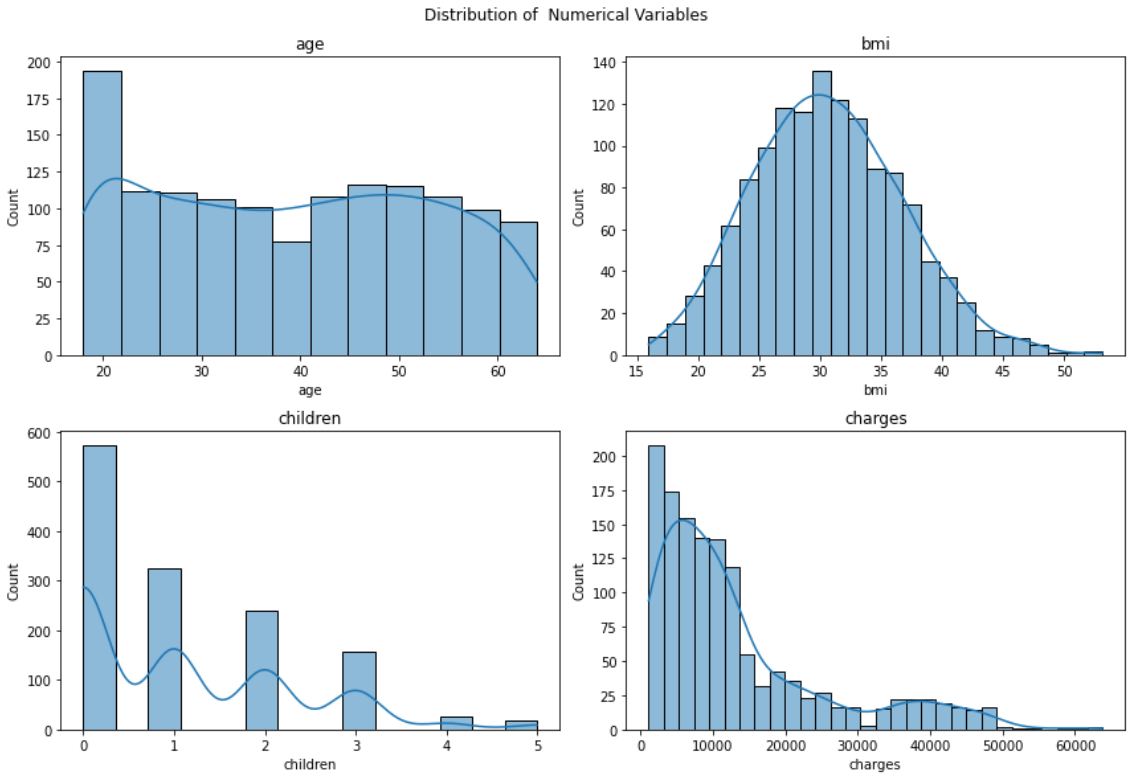
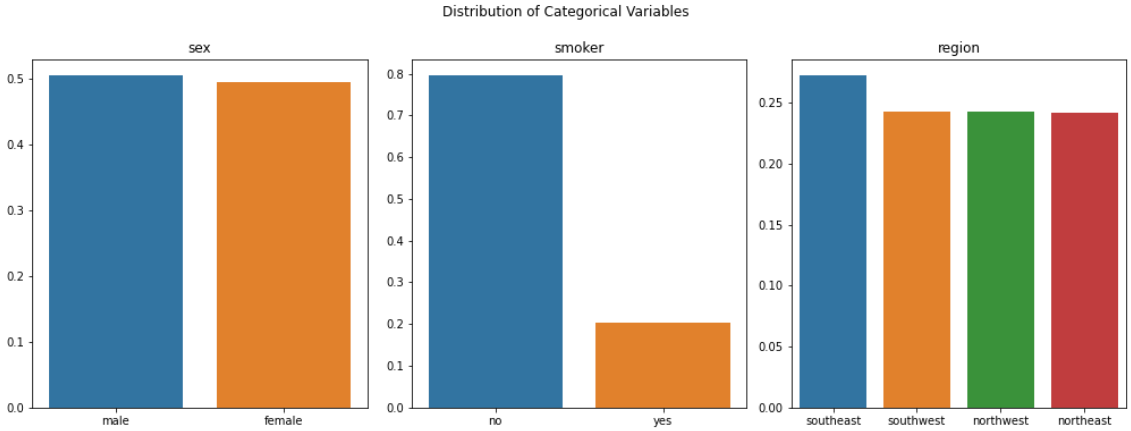
For the course project from Supervised Machine Learning: Regression, I choose to work with Kaggle dataset Medical Cost Personal Datasets. The main objective of the analysis is use Linear Regression to predict the amount spend on medical insurance give some characteristics from the contractor and discuss which factors contribute to a higher or lower spend on insurance. The dataset is composed by 1338 observations and 7 features (including the target) which are:

- age: age of primary beneficiary
- sex: insurance contractor gender, female, male
- bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight.
- children: Number of children covered by health insurance / Number of dependents
- smoker: Smoking
- region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- charges: Individual medical costs billed by health insurance

## Univariate Analysis



- Age has a uniform distribution except for the peak in 20 years.
- bmi has a normal pattern centered around 30. This is an alarming information because a bmi of 30 means obesity and the majority of the sample are in this situation.
- Charges seems to be log-normal distributed with a very long right tail. Many people spend few (in this case from 0 to 10.00) and few people spend a lot (> 10k). We can try some transformations on that to let it more normal.




- Sex is well balanced.
- No smoker is the majority class (more than 80% of observations).
- Region is balanced but southeast region has a little more observations than others.

## Dealing with outliers

For outliers detection I wrote a custom function witch returns all observations outside the range between upper limit (3nd quartile + 1.5 * IQR) and lower limit (1nd quartile – 1.5*IQR) and iterate that for each numerical feature from dataset.

As I know that charges are log-normal, I also tried the log transformation to see the impact on outliers. The results are above:

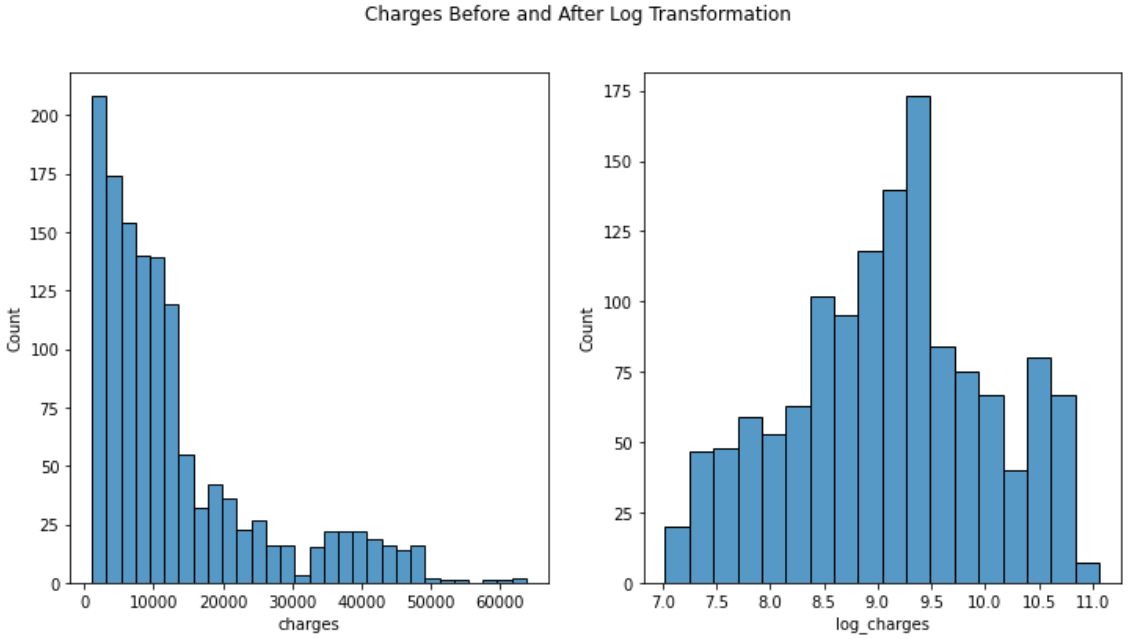
```
age: 0 Outliers (0.00% from total observations)

bmi: 9 Outliers (0.67% from total observations)

children: 0 Outliers (0.00% from total observations)

charges: 139 Outliers (10.39% from total observations)

log_charges: 0 Outliers (0.00% from total observations)
```




There are a significant number of outliers in charges features (10% from all observations). So, I cannot just remove it due to the loss of data. Log transformation do a great job limiting outliers and giving a more normal shape to charges. Let's leave this information to the modeling stage.

## Creating Hypothesis from Data

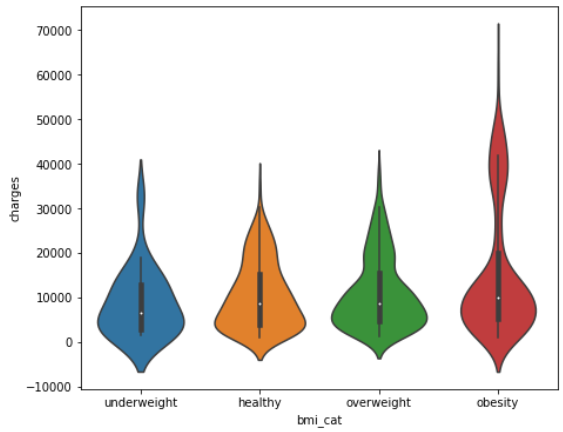
Here I make some assumptions to guide the exploratory data analysis and try to validate them with the data.

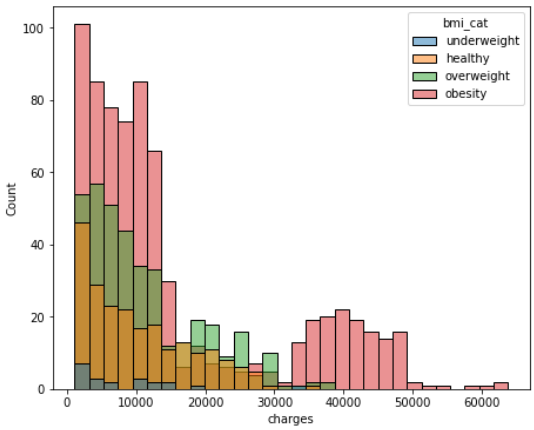
**H1: People with poor physical conditioning (high bmi) spend more on insurance. Maybe because they are more worried about health.**

To average that I decided to discretize bmi in groups following the CDC classification:

- below 18.5 – underweight

- 18.5 – 24.9 – healthy
- 25.0 – 29.9 – overweight
- 30.0 and above – obesity



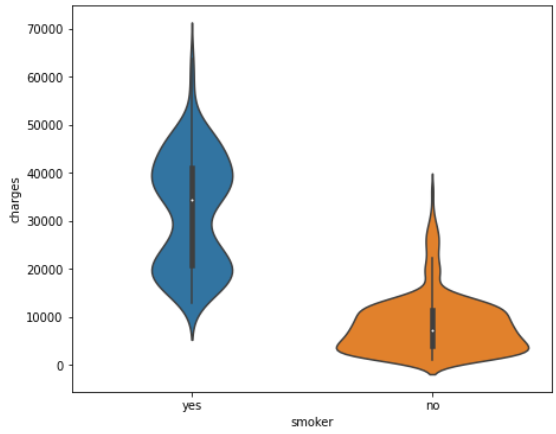



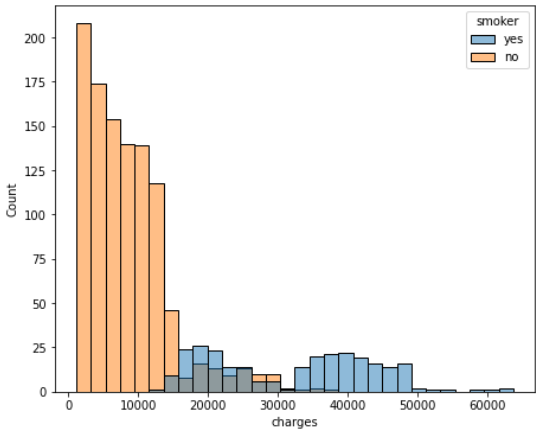

- Obesity (bmi > 29.9) prevails over other bmi categories in all regions.

- Obesity dominates charges right tail and had more outliers than other groups. This mean that obesity group spend more higher values than other groups.

We can conclude that obesity group spend higher values with insurance more than any other group.

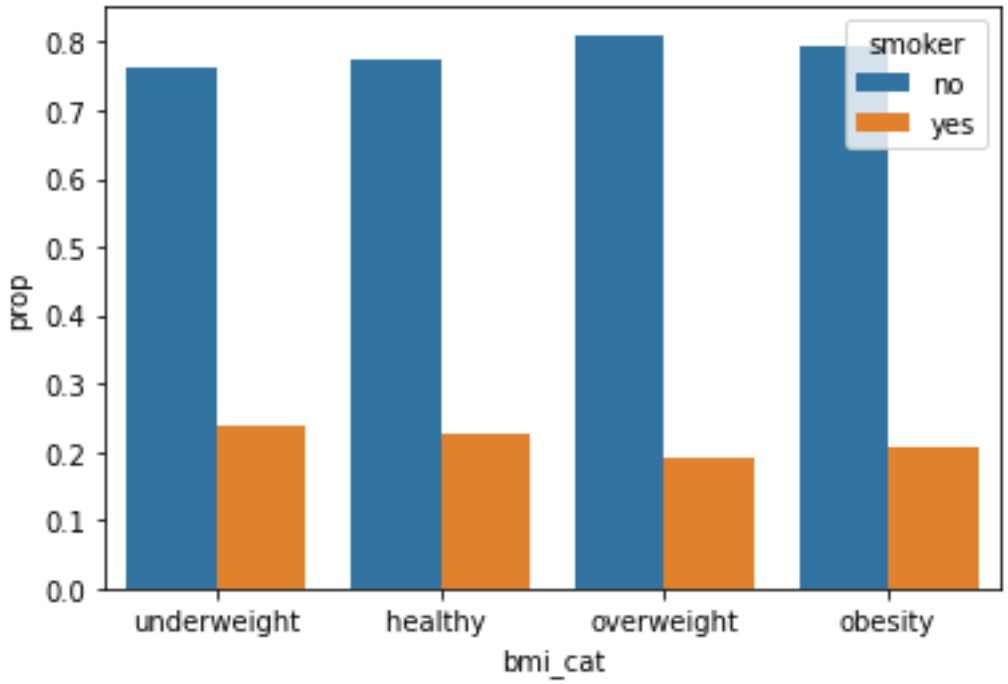
**H2: Smokers spend more on insurance than no smokers.**



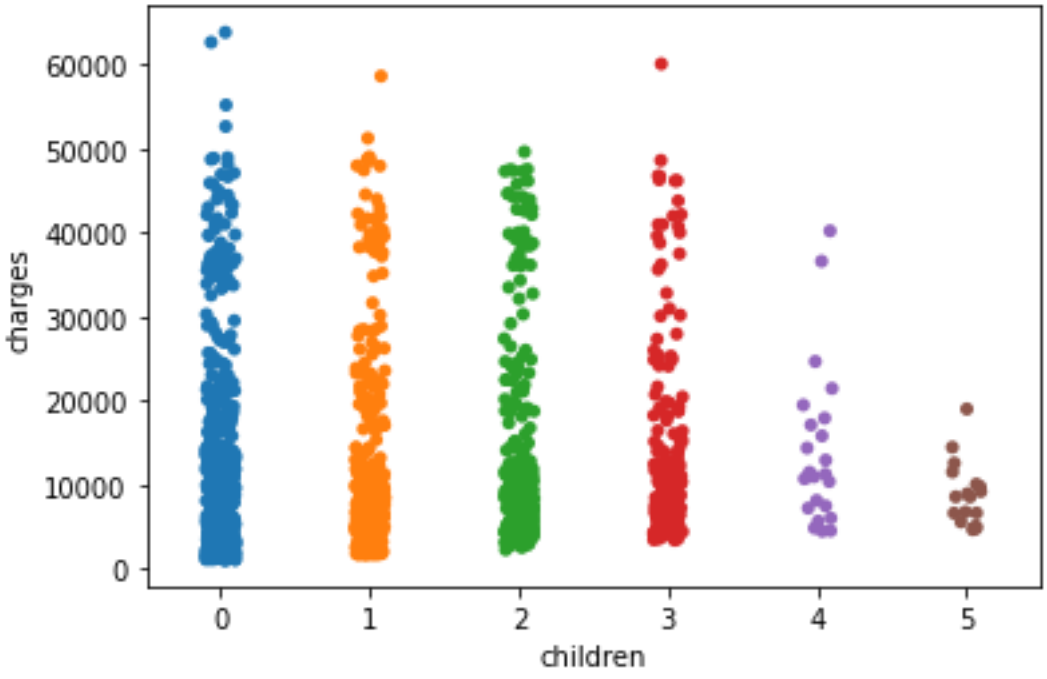

The difference between smoker and no smoker are huge. Smokers spend on average 280% more than no smokers. This reflects on the very small p-value ($2e^{-20}$) on t-test which mean we can reject the null hypothesis and assume the samples means are significantly different.

**H3: There is a relationship between smoke and high bmi index.**




Around 20% from the obesity smokes. This proportion don't vary between other groups. Smoke don't seem to be related to obesity.

**H4: Have more children lead to more spends on insurance.**




Increase children's numbers don't seem to be related to higher charges as we can see for 4 and 5 children.

## Machine Learning

To encode categorical features I use One Hot Enconder using pandas.get_dummies() function. I use 30% from dataset as test set and the remainder as train. For model selection the strategy chosen was a cross validation with 4 folds on training set with the followings models:

- A simple Linear Model as baseline
- A Linear Model with the target log transformed
- A Lasso Regression
- A Polynomial Regression with Lasso Regularization

For the chosen of alpha in regularized models, I iterate over different values and choose that returns the lower RMSE.

The RMSE for each model tried is listed above:

| | Model | RMSE(Cross Val) |
|---|---|---|
| 0 | Baseline (Linear Reg) | 6216.574884 |
| 1 | Linear Reg + Target Transf | 8687.101151 |
| 2 | Lasso Regression | 6211.067000 |
| 3 | Polynomial Reg + Lasso | 5001.815000 |

Because Polynomial Regression with Lasso reached the lowest RMSE score in cross validation It was chosen as the final version from model. The final score on test was RMSE = 4511.7690065103725.

The code for this analysis can be accessed on this repository