# ESCUELA TECNICA SUPERIOR DE INGENIEROS DE TELECOMUNICACION

## Master's Degree in Signal Theory and Communications

Statistical Modelling


*UNIVERSIDAD POLITÉCNICA DE MADRID*

---

# STMO : Project

---

*December 28, 2023*

*Authors*
**MARONE Mamadou**

*Professor*
**Pedro J. Zufiria**

## Introduction

In this project, our focus revolves around investigating the impact of various factors, such as *Age*, *number of children*, *BMI*, and *sex*, on individuals' insurance charges. Our ultimate goal is to establish a predictive model, particularly a linear regression model, to estimate insurance charges based on provided data. The dataset employed for this analysis has been sourced from Kaggle, and the corresponding code is accessible on my GitHub repository.

To execute this project, we will use a meticulously crafted methodology based on the topics covered in the ***Statistical Modeling*** courses. This framework equips us with essential statistical tools that will be instrumental throughout the project. The procedural steps are outlined as follows:

- **Exploratory and Descriptive Analysis:** Our initial phase involves a comprehensive exploratory and descriptive analysis of the dataset. This not only grants us a preliminary understanding of the data but also aids in identifying correlations among different features, guiding us in selecting pertinent variables for constructing an effective predictive model.
- **Hypothesis Testing:** Subsequent to the exploratory phase, we conduct hypothesis tests to validate and assess the insights derived earlier. The outcomes of these tests inform our choice of features for inclusion in the predictive model.
- **Model Building and Analysis:** In this stage, we construct and analyze various models based on different combinations of the selected features. This iterative process allows us to ascertain the significance of each feature in predicting the outcome. Our analysis relies on tests for the significance of the regression coefficients.
- **Model Evaluation:** Finally, we choose the most effective model and subject it to a rigorous evaluation. Additionally, we provide prediction intervals for a more comprehensive understanding of the model's predictive capability.

## I. Data Preprocessing

Initially, it is imperative to preprocess our data to ensure its suitability for next steps. This involves loading the data and conducting a thorough check for any missing or irrelevant values, followed by their systematic removal. Subsequently, we scrutinize the variable types to ensure that each variable has the appropriate type and rectified if it is not the case.

We begin by examining the data structure and previewing the initial rows to gain an initial understanding. This visualization of the data structure helps us in identifying the type of each data point. Additionally, inspecting the first few rows provides us with an indication of the potential presence of missing or irrelevant values in the data.

Relying on the Figure **??**, we can observe that there are ***1338 records*** and ***7 variables*** in the data. Regarding the types of the variables, we have noticed that the categorical variables are either ***integer*** or ***character***, consequently, they need to be converted into ***categorical*** or as ***factor***(in R language).

```
      age       sex       bmi  children    smoker    region   charges
"integer"  "factor" "numeric"  "factor"  "factor"  "factor" "numeric"
```

```
'data.frame':   1338 obs. of  7 variables:
 $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
 $ sex     : chr  "female" "male" "male" "male" ...
 $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
 $ children: int  0 1 3 0 0 0 1 3 2 0 ...
 $ smoker  : chr  "yes" "no" "no" "no" ...
 $ region  : chr  "southwest" "southeast" "southeast" "northwest" ...
 $ charges : num  16885 1726 4449 21984 3867 ...
```

(a) The structure of the data before preprocessing

Description: df [6 × 7]

| | age <int> | sex <chr> | bmi <dbl> | children <int> | smoker <chr> | region <chr> | charges <dbl> |
|---|---|---|---|---|---|---|---|
| 1 | 19 | female | 27.900 | 0 | yes | southwest | 16884.9... |
| 2 | 18 | male | 33.770 | 1 | no | southeast | 1725.552 |
| 3 | 28 | male | 33.000 | 3 | no | southeast | 4449.462 |
| 4 | 33 | male | 22.705 | 0 | no | northwest | 21984.4... |
| 5 | 32 | male | 28.880 | 0 | no | northwest | 3866.855 |
| 6 | 31 | female | 25.740 | 0 | no | southeast | 3756.622 |

6 rows

(b) First rows of the data

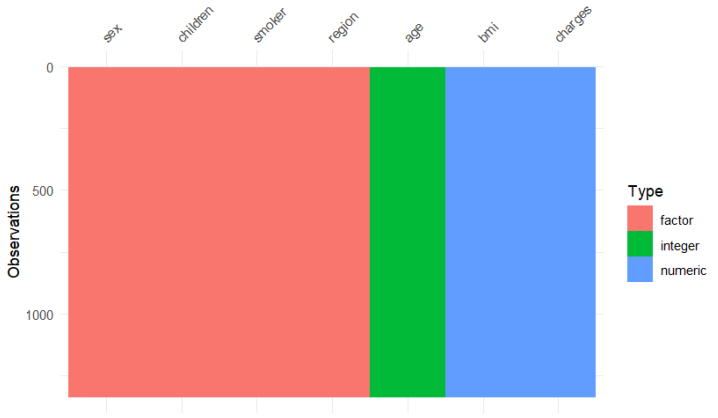



Figure 2: Missing values & Type values visualization

Following this, we verify the presence of missing values in two distinct ways. First, we determine whether there are any missing values in the entire dataset and display the count of missing values for each column. Then, to corroborate the results obtained, we utilize the **visdat** library to visualize both the variable types and the presence of missing values in the data.

```
Any missing values in the dataframe: FALSE
Missing values per column:
     age      sex      bmi children   smoker   region  charges
       0        0        0        0        0        0        0
```

In both cases, we observe that our dataset does not contain any missing values, and all variable types are now correct.

## II. Exploratory and Descriptive Data Analysis

As outlined in the introduction, we will conduct a comprehensive analysis of each variable using the statistical tools provided in the Statistical Modeling classes. This includes visualization tools such as **Histograms**, **Box plots**, **Violin plots**, **Cumulative frequency distribution**, and **Scatter plots**. Additionally, we will use numerical tools such as **Parameter estimation**, **Confidence intervals**, and **Correlation matrices**. These tools will assist us in understanding the distribution of each variable, their interrelationships, and most importantly, their impact on the output, which in our case, is the charges.

## 1. Numerical variables

We will start by analyzing the numerical variables, for which we can generate plots for further analysis. To avoid redundancy in code while computing various statistical parameters and plotting different figures for each variable, we have created a function. This function takes the name of a variable as input and outputs the numerical summaries (mean, variance, minimum, maximum, first quartile, median, third quartile, range, mode), the confidence intervals of the mean and the variance, and finally, the plots (histogram, box plot, violin plot, and cumulative frequency distribution).

### a. Age

The age contains the ages of the people whose information are recorded in the data.

| Mean | Mean_CI | Variance | Variance_CI | Q1 | Q3 | Mode | Max | Min |
|---|---|---|---|---|---|---|---|---|
| 39.20703 | [38.45352, 39.96053] | 197.4014 | [183.2522, 213.2636] | 27 | 51 | 18 | 64 | 18 |

Table 1: Summary of the variable Age

The statistical summary of the variable 'Age' provides several key insights into the dataset. The **age range** is wide, going from 18 to 64 years, indicating a **diversity in the population in term of age**. Despite this, the most common age is 18 years, suggesting a **relative majority of young individuals**. However, the **average age** is approximately 39.27 years, indicating a balance between young, middle-aged, and older individuals in the dataset.
The **variance** of approximately 197.40 and the wide **interquartile range** from 27 to 51 years show a significant spread in the ages.
The **confidence intervals** for both the mean age and variance provide a range within which the true population parameters likely fall, adding reliability to these statistics. These insights offer a comprehensive understanding of the age distribution in the dataset.

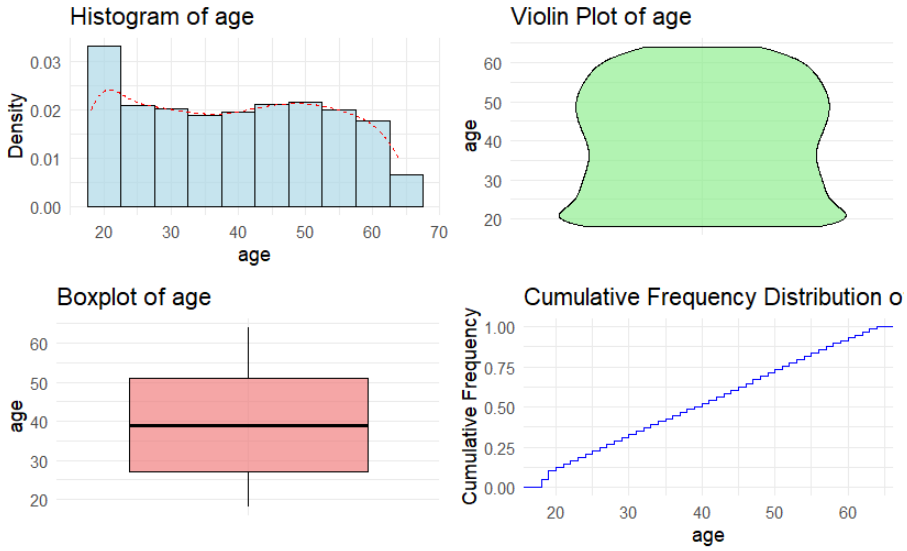



Figure 3: Histogram & Box plot & Violon plot & Cumulative frequency distribution of Age

The Histogram and Violin Plot of Age both illustrate a concentration of individuals between the ages of 30 and 50, consistent with the earlier analysis revealing a mean age of approximately 39.27 years and an interquartile range spanning from 27 to 51 years. Notably, the age distribution appears uniform.
The Boxplot of Age reinforces these findings, showcasing a median age close to 40, aligning with the mean age identified in the previous analysis. Additionally, the boxplot suggests a symmetrical distribution of age.
Examining the Cumulative Frequency Distribution of Age, there is a consistent upward trend in frequency with increasing age, indicating a uniform rise in the cumulative frequency of individuals up to the age of 64. This observation further substantiates and reinforces the notion of a uniform distribution in the age variable, as previously noted.

### b. BMI

| Mean | Mean_CI | Variance | Variance_CI | Q1 | Q3 | Mode | Max | Min |
|---|---|---|---|---|---|---|---|---|
| 30.6634 | [30.33635, 30.99045] | 37.18788 | [34.52236, 40.17611] | 26.29625 | 34.69375 | 32.3 | 53.13 | 15.96 |

Table 2: Summary of the variable BMI

The analysis of this table reveals that the average BMI in the dataset is approximately 30.66. The confidence interval for the mean BMI is between 30.34 and 30.99, which means we can be confident that the true population mean lies within this range. This suggests presence of overweightness in the studied population since the normal BMI is between 18.5 and 25.
The variance is approximately 37.19 with a confidence interval between 34.52 and 40.18.
Regarding the first quartile (Q1) and the third quartile (Q3), we have respectively 26.30 and 34.69. This means that 50% of the BMI values in the dataset fall between 26.30 and 34.69 confirming a high presence of overweight people in the studied population.In addition, the most common BMI in the dataset is 32.3.
Finally, the maximum BMI and the minimum BMIin the dataset are 53.13, and 15.96.

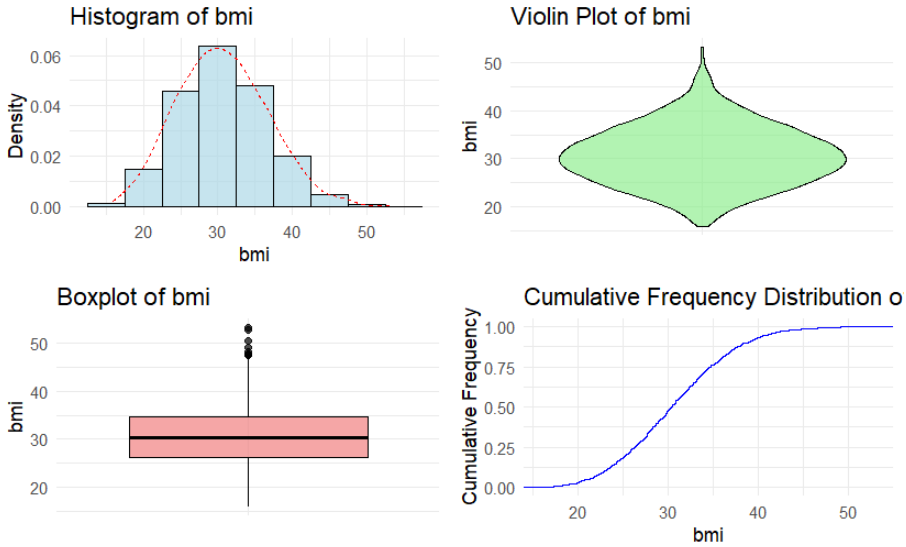



Figure 4: Histogram & Box plot & Violon plot & Cumulative frequency distribution of BMI

The Histogram of BMI and the Violin Plot of BMI both show that most individuals have a BMI around 30, which aligns with the previous analysis where the mean BMI was found to

be approximately 30.66. We can also observe that the shape of the histogram is similar to a normal distribution density function.
The Boxplot of BMI reveals that the median BMI is around 30, which is close to the mean BMI from the previous analysis. The interquartile range in the boxplot spans from approximately 26.30 to nearly 34.69, which is exactly the same as the previously analyzed interquartile range. This suggests a symmetry in the BMI distribution, supporting the assumption of normal distribution of the bmi.
The Cumulative Frequency Distribution of BMI leads to the same conclusion as the analysis of the histogram.

### c. Charges

| Mean | Mean_CI | Variance | Variance_CI | Q1 | Q3 | Mode | Max | Min |
|---|---|---|---|---|---|---|---|---|
| 13270.42 | [12620.95, 13919.89] | 146652372 | [136140722, 158436617] | 4740.287 | 16639.91 | 1639.563 | 63770.43 | 1121.874 |

Table 3: Summary of the variable Charges

Relying on the results in the table above, the mean of the charges in the population is approximately 13270.42. The confidence interval for the mean charge is between 12620.95 and 13919.89, which means we can be confident that the true population mean lies within this range.
Regarding the variance, it is approximately 146652372 with a confidence interval between 136140722 and 158436617.
The first(Q1) and third(Q3) quartiles suggests that 50% of the charges in the dataset are between 4740.287 and 16639.91. The most common charge in the dataset is 1639.563.
The maximum charge in the dataset is 63770.43, and the minimum is 1121.874.

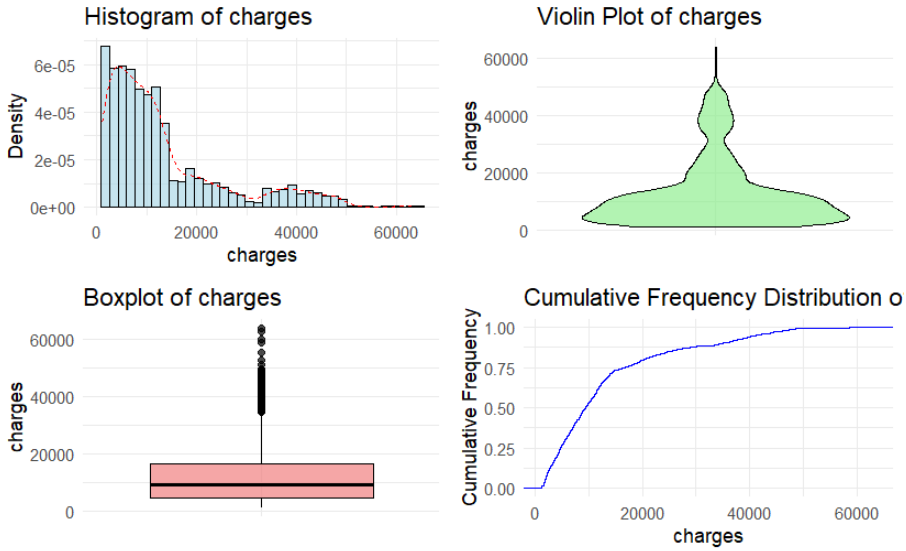



Figure 5: Histogram & Box plot & Violon plot & Cumulative frequency distribution of Charges

The Histogram of Charges and the Violin Plot of Charges both show that most charges are clustered between approximately 0 and 10,000, which aligns with the previous analysis where the first quartile (Q1) was found to be 4740.287. Then, we have to other clusters the first one

between 10.000 to 30.0000 and the other starting from 30.000 to the maximum. That might correspond to different categories of people.
The Boxplot of Charges reveals that there are some outliers above the upper whisker, indicating that there are some charges that are significantly higher than the rest as suggested previously.
The Cumulative Frequency Distribution of Charges reveals a uniform increase of the frequency when the charges increase until 10.000, and adopt a different behaviour from there.

## 2. Categorical Variables

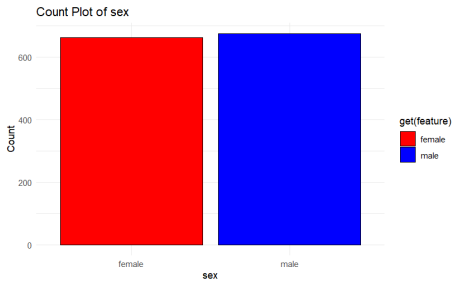
For the numerical variables, we have developed functions to extract information about these variables. There are three functions in total. The first function accepts the name of a categorical variable as input and outputs the proportion of each category along with the corresponding confidence interval, and plots the count plot of the variable.
The second function takes the names of two categorical features as input and plots the count plots of the first variable against the categories of the second variable. This provides us with an understanding of the proportions of each category of the first variable, taking into account the second one.
The final function is similar to the previous one but accepts a categorical and a numerical variable instead. It aids in analyzing the numerical variables in relation to the categorical one. Its goal is to check if the distribution of the numerical variables changes from one category to another. This is particularly important when it comes to the charges, as we can determine if a categorical variable has an impact on the charges and if it is worthwhile to select it as a feature for our model. Considering the other numerical variables can also help to better understand the relationship between the variables.

### a. Sex

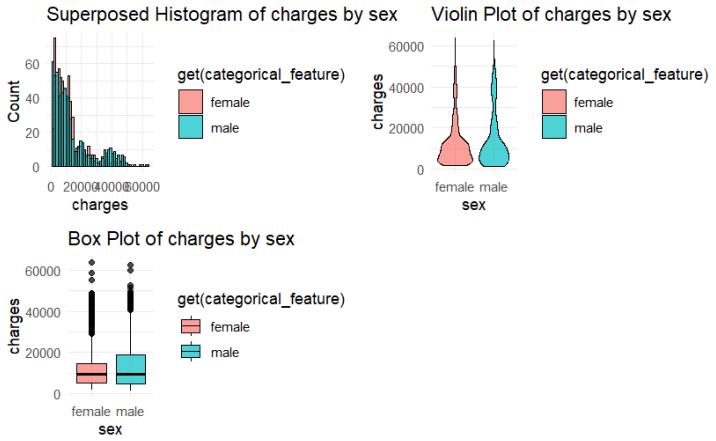
| Category | Proportion | Proportion_CI |
|---|---|---|
| Male | 0.5052317 | $[0.4780927, 0.5323403]$ |
| Female | 0.4947683 | $[0.4676597, 0.5219073]$ |

Table 4: Summary of the categorical variable Sex

This table shows that the proportion of males in the dataset is approximately 0.5052317, or about 50.52%. The confidence interval for this proportion is between 0.4780927 and 0.5323403, making us confident that the true population proportion lies within this range.
Similarly, the proportion of females in the dataset is approximately 0.4947683, or about 49.48% with a confidence between 0.4676597 and 0.5219073. We can conclude that there is no imbalance between the two categories of sex.




(a) Count plot of Sex




(b) Distribution of charges among different Sex

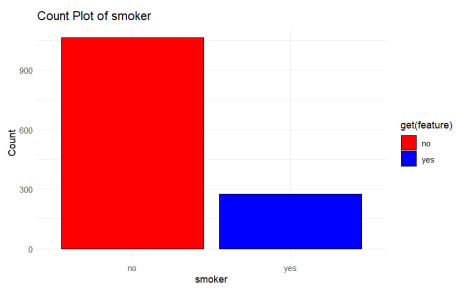
The counterplot confirms the analysis conducted previously.
The Superposed Histogram and the Violin Plot of Charges by Sex seems to be similar with the same shape as when we considered all the sample.
The Box plots also shows similarities with a the third quartile that is slightly higher for the men than for the women.
In conclusion, we will consider that there no difference between male and female in term of insurance charges and therefore, the sex is not a relevant variables to predict the charges. This assumption will be verified during the hypothesis test.

### b. Smoker

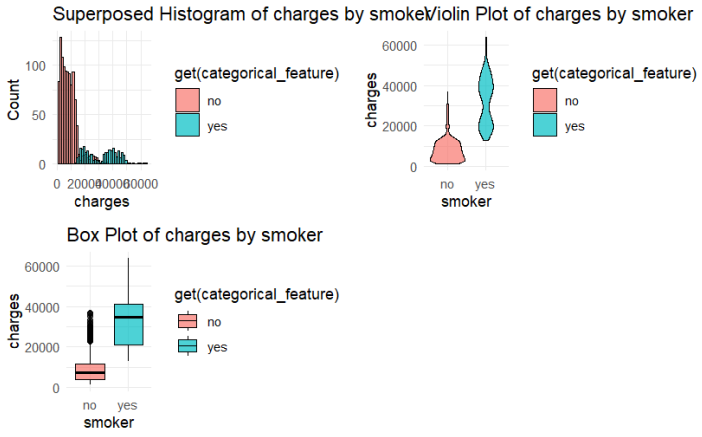
| Category | Proportion | Proportion_CI |
|---|---|---|
| no | 0.7952167 | [0.7723761, 0.8163379] |
| yes | 0.2047833 | [0.1836621, 0.2276239] |

Table 5: Summary of the categorical variable Smokers

This results shows that the proportion of non-smokers in the dataset is approximately 0.7952167, or about 79.52% with a confidence interval between 0.7723761 and 0.8163379, which means we can be confident that the true population proportion lies within this range.
Unsurprisingly, the proportion of smokers in the dataset is approximately 0.2047833, or about 20.48% and lower than the non-smoker's proportion. Besides, we can be confident that the true population proportion lies between 0.1836621 and 0.2276239.
The main information is that, the majority of the true population are non-smokers and our dataset is imbalanced considering the variable smoker.




(a) Count plot of Smoker




(b) Distribution of charges among different Smoker

The counter plot leads to the same conclusion as the the previous analysis.
The Superposed Histogram of Charges by Smoker shows that non-smokers have more counts but lower charges, while smokers, although fewer in number, tend to have higher charges.
The Violin Plot of Charges by Smoker provides a similar view, with a wider distribution of charges for smokers, indicating variability in the charges for smokers.
The Box Plot of Charges by Smoker reveals that smokers have a higher median charge than non-smokers, indicating that on average, smokers tend to have higher charges.

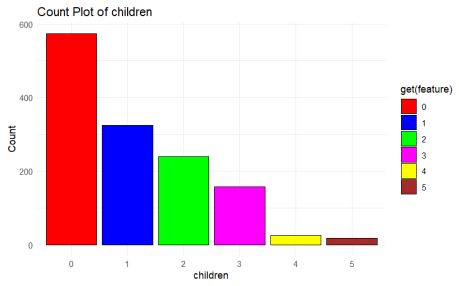
In conclusion, the plots provide a visual representation of the charges distribution for different categories of smokers. They show that although smokers are fewer in number, they tend to have higher and more variable charges compared to non-smokers. This reveals that the variable smoker is discriminative considering the charges and can be very useful for the prediction of this latter.

### c. Children

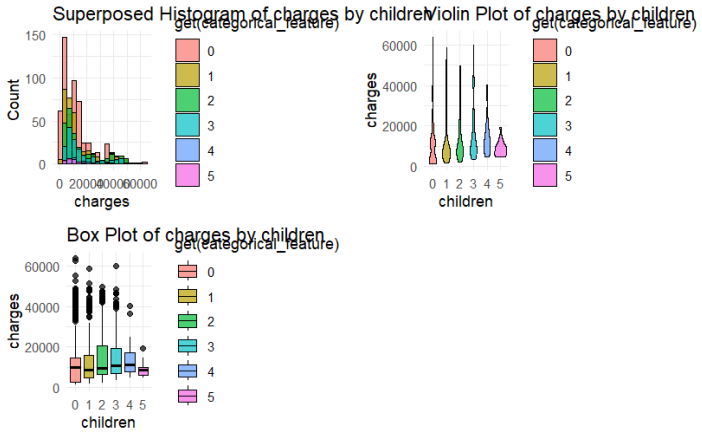
| Category | Proportion | Proportion_CI |
|---|---|---|
| 0 | 0.42899851 | [0.40234965, 0.45605961] |
| 1 | 0.24215247 | [0.21959718, 0.26620809] |
| 2 | 0.17937220 | [0.15938555, 0.20122793] |
| 3 | 0.11733931 | [0.10083173, 0.13608519] |
| 4 | 0.01868460 | [0.01238493, 0.02787937] |
| 5 | 0.01345291 | [0.00823645, 0.02161946] |

Table 6: Summary of the categorical variable Children

Briefly, this results suggest a prevalence of individuals without or with only one child compared to those with multiple children. The data indicates that very few individuals in the true population have more than three children.




(a) Count plot of Children




(b) Distribution of charges among different Children

The Superposed Histogram of Charges by Children shows that the distribution of charges varies slightly based on the number of children. However, this observation is not correctly done because we cannot see the complete histograms. To fix this we could use transparent histograms, but we will rely on the box plots and violon plots.
Both the Violin Plot and the Box plot of Charges by Children provides a similar view which is that when the number of children is higher than 3, the charges start being very different compared to those with less number of children.

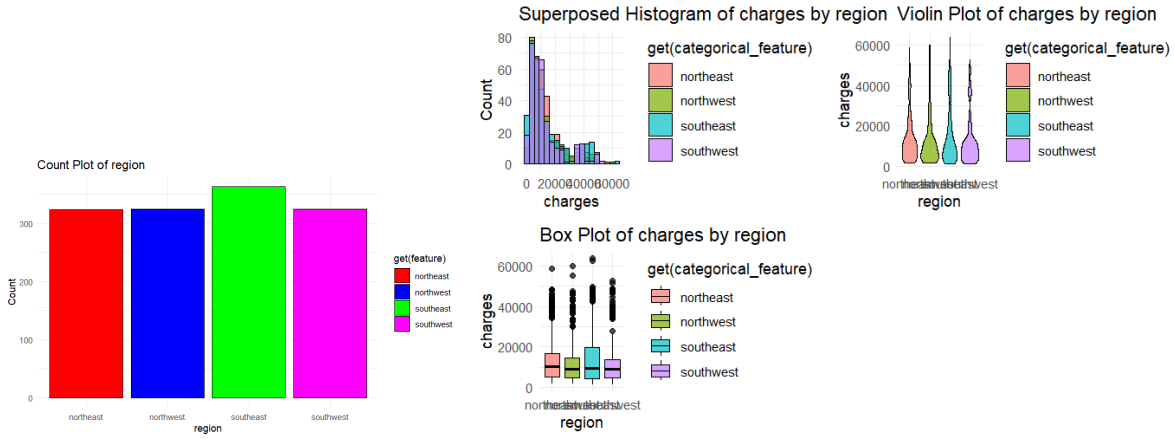
In conclusion, the plots show that the number of children can affect the charges, with different distributions observed for each category. However this difference is noticeable only when the number of children is higher than 3. Therefore, we can create a **new categorical variable** that contains the information whether or not the person has more than 3 three children. We will call it ***many_children***.

### d. Region

| Category | Proportion | Proportion_CI |
|---|---|---|
| northeast | 0.2421525 | [0.2195972, 0.2662081] |
| northwest | 0.2428999 | [0.2203184, 0.2669772] |
| southeast | 0.2720478 | [0.2485184, 0.2969028] |
| southwest | 0.2428999 | [0.2203184, 0.2669772] |

Table 7: Summary of the categorical variable Region

In summary, the various regions are seemingly equally represented in our dataset, and the confidence interval implies that this equal representation extends to the true population. However, it is worth noting a slight overrepresentation of individuals from the southeast compared to other regions. Nevertheless, this minor discrepancy will not be considered in the sequel.




(a) Count plot of Region (b) Distribution of charges among different Region

Similarly, the plots show that the region has no impact on the charges, since all the different plots are similar form a region to another.
Therefore, we will not consider the region in the variables to predict the charges in the sequel.

## 3. Scatter plot & Correlation matrix

The scatter plot and the correlation matrix provide both visual and numerical insights into the relationships between the different variables. Given that the insurance charges for smokers and non-smokers are significantly different, I have decided to use distinct colors for smokers and non-smokers in the plots. As previously mentioned, we will not consider the variables sex, region, and children in our analysis.

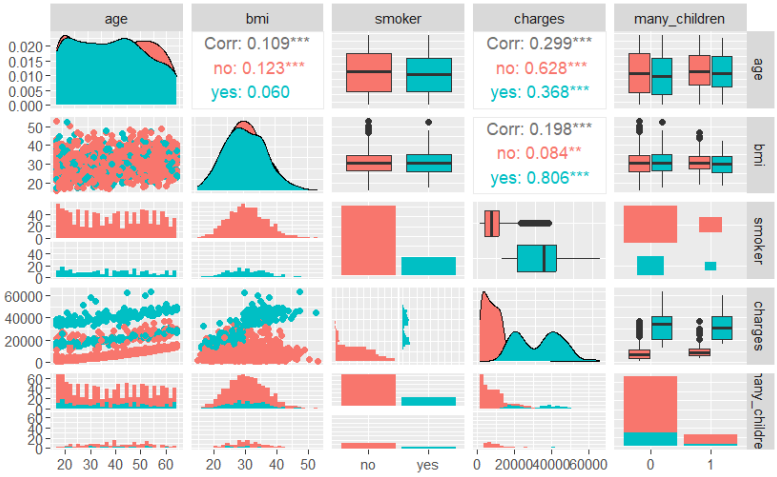



Figure 10: Scatter plot

The scatter plots show clear distinctions in charges between smokers and non-smokers. Each category is represented by a different color, and the distribution of charges varies based on the category.

Age seems to have a positive correlation with charges across both categories(smokers and non-smokers); as age increases, the medical charges also tend to increase. This is consistent with the previous analysis where it was found that age is a significant predictor of charges.

BMI does not exhibit a strong correlation with age but exerts a notable influence on medical charges across both categories. It becomes apparent that BMI significantly affects charges only when the individual is a smoker. Indeed, the blue points are elevated and form a distinct trend, whereas for non-smokers, there is no apparent correlation.

Having many children does not show a significant pattern in relation to age or BMI but shows some variation in medical charges across the two categories. This could suggest, as we mentioned it, that the number of children could be a significant predictor of charges.

## III. Hypothesis Tests

In this section, we will employ hypothesis testing to evaluate the observations made in the preceding steps. Additionally, we will ascertain whether the conditions of Multivariate Normality, Autocorrelation, and Homoscedasticity, which are prerequisites for a linear regression model, are satisfied when considering various combinations of variables.

To perform a hypothesis test, we will follow the next steps:

- **Formulation of the null hypothesis:** The null hypothesis is a statement about the population that will be tested. The null hypothesis is often an initial claim that is based on previous analyses or specialized knowledge.
- **Choice of the significance level:** Next, we choose a significance level, also known as alpha. The significance level is a threshold that determines when we reject the null hypothesis. It is the probability of rejecting the null hypothesis when it is true. For all the tests in this analysis, we will choose an alpha level of 0.05.
- **Conducting the test and obtaining a p-value:** After conducting the test, we will obtain a p-value. The p-value is the probability that the results of your test occurred at random. If the p-value is less than alpha, we reject the null hypothesis. If the p-value is greater than alpha, we fail to reject the null hypothesis.
- **Interpretation of the results:** Finally, we interpret the results in the context of the problem. We have to be careful to not make claims of causation based on the results of the hypothesis test. The results only suggest that there is evidence to reject or fail to reject the null hypothesis.

However, the R language provides functions that can automatically conduct the test and provide the results of the test.

### 2. Same distribution test among categories of the different categorical variables

The objective of this section is to evaluate whether the distribution of charges varies among different categories for each categorical variable. To accomplish this, we utilize the ***Kruskal-Wallis test***, a hypothesis test that asserts the null hypothesis of equal charge distributions across various categories. This test is applied to all categorical variables. If the resulting p-value is below 0.05, we reject the null hypothesis, indicating dissimilar distributions. Conversely, if the p-value exceeds 0.05, we do not reject the null hypothesis, signifying that the charge distributions in the different categories are deemed similar.

### a. Sex

| Category | n | statistic | df | p-value |
|---|---|---|---|---|
| Male | 676 | 0.5789309 | 1 | 0.447 |
| Female | 662 | 0.5789309 | 1 | 0.447 |

Table 8: Kruskal-Wallis test result for charges distribution among sexes

The p-value for both categories is 0.447, exceeding the threshold of 0.05. Consequently, we do not reject the null hypothesis, indicating that the charge distributions for males and females are deemed similar.
In simpler terms, according to this test, there is no statistically significant difference in the distribution of charges between males and females.
In conclusion, this reaffirms the observation made during the descriptive analysis of the variable sex. We will decisively exclude this variable as a predictor for our model.

### b. Smoker

| Category | n | statistic | df | p-value |
|---|---|---|---|---|
| no | 1064 | 588.5197 | 1 | 5.26e-130 |
| yes | 274 | 588.5197 | 1 | 5.26e-130 |

Table 9: Kruskal-Wallis test result for charges distribution among smokers

The p-value for both categories is approximately 5.26e-130, which is less than the threshold of 0.05 meaning that we reject the null hypothesis, signifying that the charge distributions for smokers and non-smokers are considered different.
In other words, based on this test, there is a statistically significant difference in the distribution of charges between smokers and non-smokers. This supports our assumption that the smoker variable has a strong impact on the charges.

### c. many_children

| Category | n | statistic | df | p-value |
|---|---|---|---|---|
| 0(no) | 1138 | 12.17762 | 1 | 0.000484 |
| 1(yes) | 200 | 12.17762 | 1 | 0.000484 |

Table 10: Kruskal-Wallis test result for charges distribution among number of children

The p-value for both categories is approximately 0.000484, which is below the threshold of 0.05. Therefore, we reject the null hypothesis, indicating that the charge distributions for individuals with many children and those with fewer or no children are considered different.
While the p-value is not as low as in the smoker case, according to this test, there is still a statistically significant difference in the distribution of charges between individuals with many children and those with fewer or no children.

| Category | n | statistic | df | p-value |
|---|---|---|---|---|
| northeast | 324 | 4.734181 | 3 | 0.192 |
| northwest | 325 | 4.734181 | 3 | 0.192 |
| southeast | 364 | 4.734181 | 3 | 0.192 |
| southwest | 325 | 4.734181 | 3 | 0.192 |

Table 11: Kruskal-Wallis test result for charges distribution among number of Region

#### d. Region

Regarding the region, the p-value for all categories is approximately 0.192, which is greater than the threshold of 0.05. Then, we do not reject the null hypothesis, signifying that the charge distributions for different regions are considered similar.
In conclusion, there is no statistically significant difference in the distribution of charges between different regions. Therefore, we keep rejecting the region variable as predictor for our model.

## Test to check the validity condition of the variables for a linear regression & Model analysis

In this section, our goal is to validate the essential assumptions for a linear regression model, which encompass **Multivariate normality**, **Autocorrelation**, and **Homoscedasticity**. For each of these tests, the null hypothesis posits the non-validity of the corresponding condition. To accomplish this, we build a linear regression model trained on the entire dataset and execute various tests. To enhance efficiency and eliminate code redundancy, we have implemented a function for automated testing.

Moreover, this step involves the analysis of different models, leveraging the coefficients to comprehend the impact of each selected variable on the output. The relevance of these variables is assessed by examining the p-values associated with the T-tests of these coefficients. Simultaneously, we scrutinize the R-squared value to gauge how effectively the different variables explain the variability of the model. The model summary provides crucial information such as the coefficients of different variables and their associated p-values from T-tests, along with the R-squared value, offering a measure of how well the selected variables account for the variability in charges.

#### a. Considering all the selected variables: Age, BMI, smoker, many_children

| Test | p-value | validated |
|---|---|---|
| Multivariate normality | 1.063972e-28 | no |
| Autocorrelation | 1.280000e-01 | yes |
| Homoscedasticity | 9.390859e-24 | no |

Table 12: Condition's validity checking considering all selected variables.

The multivariate normality and homoscedasticity conditions are not validated, indicating that the data does not follow a multivariate normal distribution and the variances are not equal across the groups. The autocorrelation condition is validated, suggesting that there is no autocorrelation in the dataset.

```
Residuals:
     Min       1Q   Median       3Q      Max
-12290.1  -2954.8   -981.7   1422.2  29101.3

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -11763.01     938.32 -12.536   <2e-16 ***
age               258.64      11.94  21.664   <2e-16 ***
bmi               322.81      27.47  11.752   <2e-16 ***
smokeryes       23815.50     412.61  57.719   <2e-16 ***
many_children1    787.30     467.30   1.685   0.0923 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6088 on 1333 degrees of freedom
Multiple R-squared:  0.748,  Adjusted R-squared:  0.7473
F-statistic: 989.2 on 4 and 1333 DF,  p-value: < 2.2e-16
```

The coefficients table shows the estimated effects of the variables age, bmi, smoker status, and having many children on the charges. All variables except for having many children are statistically significant at a 0.001 level, as indicated by the ‘***’ next to their p-values. This suggests that age, bmi, and smoker status have a significant impact on charges. The variable many_children1 is significant at the 0.1 level, as indicated by the ‘.’ next to its p-value, suggesting a weaker evidence of its effect on charges.
The R-squared value is 0.748, suggesting that approximately 74.8% of the variability in charges can be explained by the model.
The F-statistic and its associated p-value test the overall significance of the model. The F-statistic is 989.2 and the p-value is less than 2.2e-16, indicating that the model is statistically significant.

In conclusion, despite not meeting all the required conditions, the model summary reveals a consistent model that accounts for 74.8% of the variability in the charges. Moreover, apart from the variable many_children, all other selected variables are strong predictors of the charges. Consequently, we may consider excluding the many_children variable or constructing two separate models that account for the two categories within many_children.

### b. Considering the same variables execpt the variables many_children

| Test | p-value | validated |
|---|---|---|
| Multivariate normality | 2.662086e-28 | no |
| Autocorrelation | 1.520000e-01 | yes |
| Homoscedasticity | 2.827322e-24 | no |

Table 13: Condition’s validity checking considering all selected variables.

We get the same result as in the previous case.

```
Residuals:
     Min       1Q   Median       3Q      Max
-12415.4  -2970.9   -980.5   1480.0  28971.8

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -11676.83     937.57  -12.45   <2e-16 ***
age            259.55      11.93   21.75   <2e-16 ***
bmi            322.62      27.49   11.74   <2e-16 ***
smokeryes    23823.68     412.87   57.70   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6092 on 1334 degrees of freedom
Multiple R-squared:  0.7475,  Adjusted R-squared:  0.7469
F-statistic:  1316 on 3 and 1334 DF,  p-value: < 2.2e-16
```

In this case, it is noteworthy that all variables have a strong impact on the charges, as indicated by the low p-values. Furthermore, the R-squared value remains nearly constant, suggesting that the model's ability to explain charges does not significantly change even when the variable many_children is excluded.

However, the non-respect of the required conditions persists, and addressing this issue could potentially lead to improved results. This discrepancy might be attributed to the presence of the categorical variable 'smoker.' Therefore, we will bifurcate the model, considering the distinct categories of the variable 'smoker.'

**b. Considering the same variables but separating smokers and non-smokers and**

| Category | Test | p-value | validated |
|---|---|---|---|
| **Smokers** | Multivariate normality | 1.450170e-08 | no |
| | Autocorrelation | 3.800000e-01 | yes |
| | Homoscedasticity | 8.061665e-01 | yes |
| **Non-Smokers** | Multivariate normality | 1.310657e-47 | no |
| | Autocorrelation | 3.600000e-01 | yes |
| | Homoscedasticity | 9.216818e-01 | yes |

Table 14: Condition's validity checking taking smokers.

For both smokers and non-smokers, the multivariate normality condition is not validated, indicating that the data does not follow a multivariate normal distribution. However, the autocorrelation and homoscedasticity conditions are validated, suggesting that there is no autocorrelation and the variances are equal across the groups. This can be explained by the uniform distribution of the age. A solution to addresse this issue could be to apply a transformation to the age to make it normally distributed.

***Model summary for smokers :***

```
Residuals:
     Min       1Q   Median       3Q      Max
-14604.4  -4315.1   -240.5   3638.0  29316.7

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -22367.45    1931.86  -11.58   <2e-16 ***
age            266.29      25.06   10.63   <2e-16 ***
bmi           1438.09      55.22   26.05   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5754 on 271 degrees of freedom
Multiple R-squared:  0.7532,  Adjusted R-squared:  0.7514
F-statistic: 413.6 on 2 and 271 DF,  p-value: < 2.2e-16
```

***Model summary for non-smokers:***

```
Residuals:
    Min      1Q  Median      3Q     Max
-3237.7 -1947.5 -1354.7  -650.6 24430.6

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2293.632    804.061  -2.853  0.00442 **
age           266.877     10.245  26.048  < 2e-16 ***
bmi             7.075     23.877   0.296  0.76703
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4669 on 1061 degrees of freedom
Multiple R-squared:  0.3944,  Adjusted R-squared:  0.3932
F-statistic: 345.4 on 2 and 1061 DF,  p-value: < 2.2e-16
```

The summaries of both models reveal that age exerts a strong impact on charges regardless of the smoker category. However, BMI significantly influences charges only for individuals who smoke. In the case of smokers, both age and BMI have very low p-values associated with the T-test. Conversely, for non-smokers, BMI exhibits a notably high p-value, indicating its limited relevance as a predictor of charges.
Moreover, it is noteworthy that the R-squared error is higher in the case of smokers but substantially lower in the other case.

In conclusion, the model constructed for smokers demonstrates high consistency and explains 75.3% of the variability in charges. Conversely, for non-smokers, the model struggles to explain much variability, and BMI appears to be an ineffective predictor.

## 1. Model evaluation

In this section, we split our dataset into a training set and a testing set. The model is trained on the training set and its performance is evaluated on the testing set. The evaluation metrics used are Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R-squared (R2).

Description: df [1 × 3]

| RMSE <dbl> | MAE <dbl> | R_squared <dbl> |
|---|---|---|
| 6346.556 | 4234.608 | 0.7061747 |

1 row

Figure 11: Evaluation metrics

The evaluation metrics of the linear regression model, which was built to predict insurance charges, provide valuable insights into the model's performance. The Root Mean Square Error (RMSE) is 6346.556, and the Mean Absolute Error (MAE) is 4234.608. Given the scale of the charges, these errors might be considered high, indicating that the model's predictions can deviate from the actual charges by these amounts. However, considering the complexity and variability of healthcare charges, these errors might be acceptable.

The R-squared value is 0.7061747, suggesting that approximately 70.6% of the variability in charges can be explained by the model. This is a relatively high value, indicating that the model captures a substantial portion of the information in the data.

### 2. Prediction interval

The prediction interval provides a range for the predicted value of the dependent variable for a given significance level. It gives an interval within which we can expect the actual value to lie with a certain level of confidence.

The predictions were made using a random sample from the test dataset. The aim is to ensure that the evaluation metrics are unbiased estimates of the model's performance on new, unseen data.

Here is an example of a prediction and its prediction interval:

```
Actual Value: 11488.32
Prediction: 11959.83
Prediction Interval: [ 111.6027 ; 23808.05 ]
```

This outcome signifies that we can be confident in the prediction's reliability, as the actual value is expected to fall within the predicted range (prediction interval). Considering the scale of the charges (in the range of 60,000), this prediction is acceptable.

## Conclusion

In conclusion, although there is room for improvement, the model offers a solid starting point for predicting insurance charges. Future work could explore alternative modeling approaches, incorporate additional variables, or employ advanced techniques for handling outliers and non-linear relationships. Despite its limitations, the model provides valuable insights and serves as a useful tool for understanding and predicting healthcare charges.

Furthermore, this project provided us with information about the true population through the analysis of various variables and their relationships.