Códigos do Amanhã EDA Project

Student: Thayssa Daniele

GitHub

Linkedin

Exploratory Data Analysis

- The base used for analysis was made available by Professor Rafael Roberto.
- Available at: data

Concepts explored

- Linear Correlation.
- Boxplot.
- Use of libraries and functionalities for analysis.
- Linear Regression.
- Model Adjustment - R^2.
- Data Storytelling.

Libraries Used

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.formula.api as sm
```

Correlation of Variables

My first dataset had a column with informations type string, For my analysis I decided to remove this column and apply the correlation function only to my new dataset called "data_limpa" with numeric variables. RESULTING:

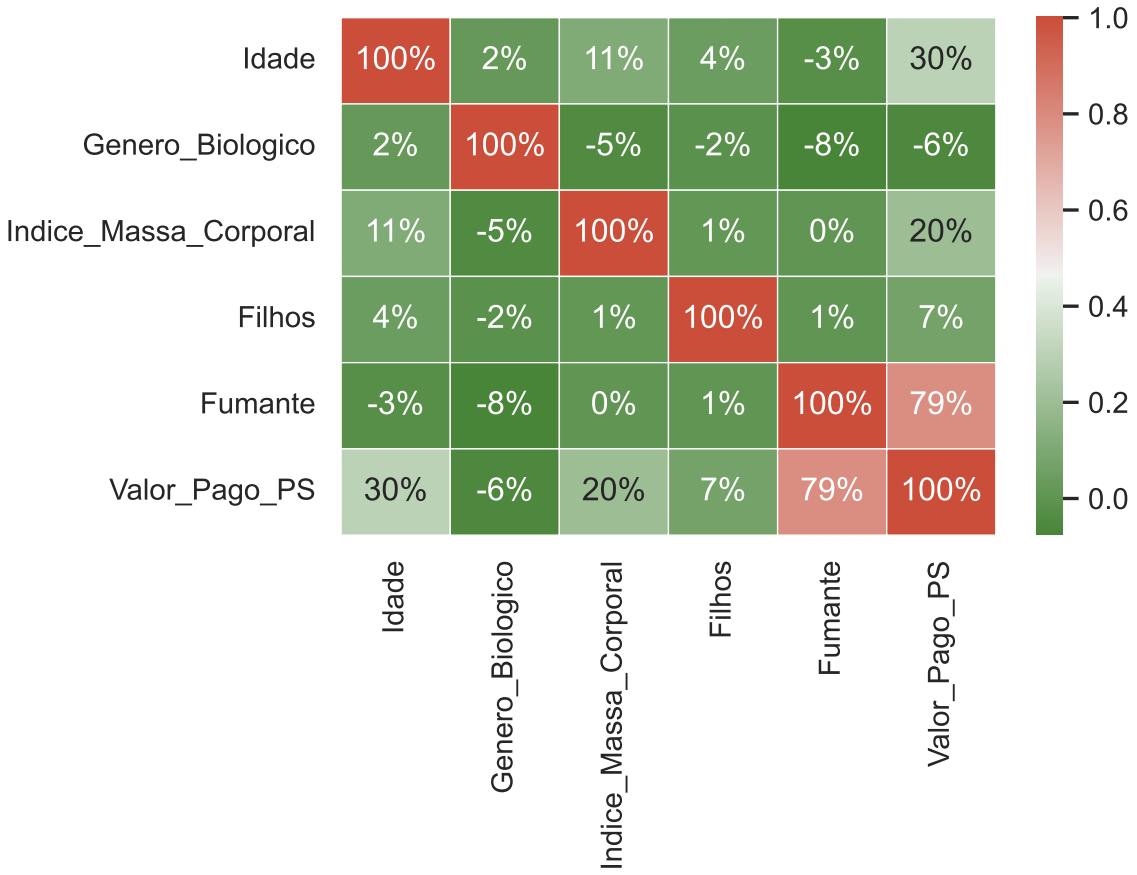
	Idade	Genero_Biologico	Indice_Massa_Corporal	Filhos	Fumante	V
Idade	1.000000	0.020856	0.109272	0.042469	-0.025019	0
Genero_Biologico	0.020856	1.000000	-0.046371	-0.017163	-0.076185	-
Indice_Massa_Corporal	0.109272	-0.046371	1.000000	0.012759	0.003750	0
Filhos	0.042469	-0.017163	0.012759	1.000000	0.007673	0
Fumante	-0.025019	-0.076185	0.003750	0.007673	1.000000	0
Valor_Pago_PS	0.299008	-0.057292	0.198341	0.067998	0.787251	1

It's visible when we make a comparative of correlation between all variables,that our strongest correlation is between 'Indice_Massa_Corporal', 'Fumante', 'Valor_Pago_PS' an 'Idade'.

	Indice_Massa_Corporal	Fumante	Valor_Pago_PS	Idade
Indice_Massa_Corporal	1.000000	0.003750	0.198341	0.109272
Fumante	0.003750	1.000000	0.787251	-0.025019
Valor_Pago_PS	0.198341	0.787251	1.000000	0.299008
Idade	0.109272	-0.025019	0.299008	1.000000

HEATMAP

Visual representation of correlation with every variable of my new dataset "data_limpa" in percent.



SCATTER CHART

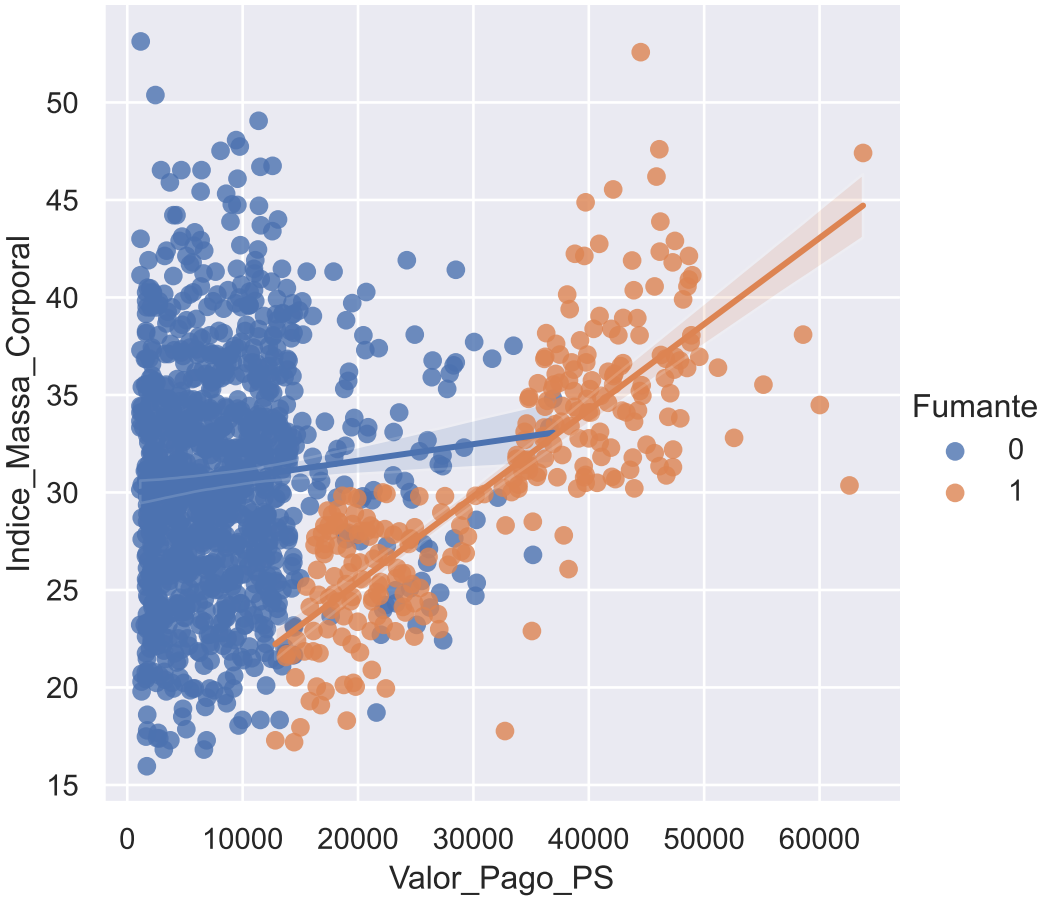
The relationship between Y = 'Indice_Massa_Corporal' and X ='Valor_Pago_PS' can be shown for different subsets of the data using the hue, it was choose differentiate this subset using variable 'Fumante'

LINEAR REGRESSION

R^2 (R-Square) or Determinated Coeficint, is a metric that indicates how well the linear regression model fits the data. It measures the proportion of variation in the dependent variable (the variable we are trying to predict, 'Valor_Pago_PS') that can be explained by the independent variables (the explanatory variables, 'Fumante + Indice_Massa_Corporal') in the model.

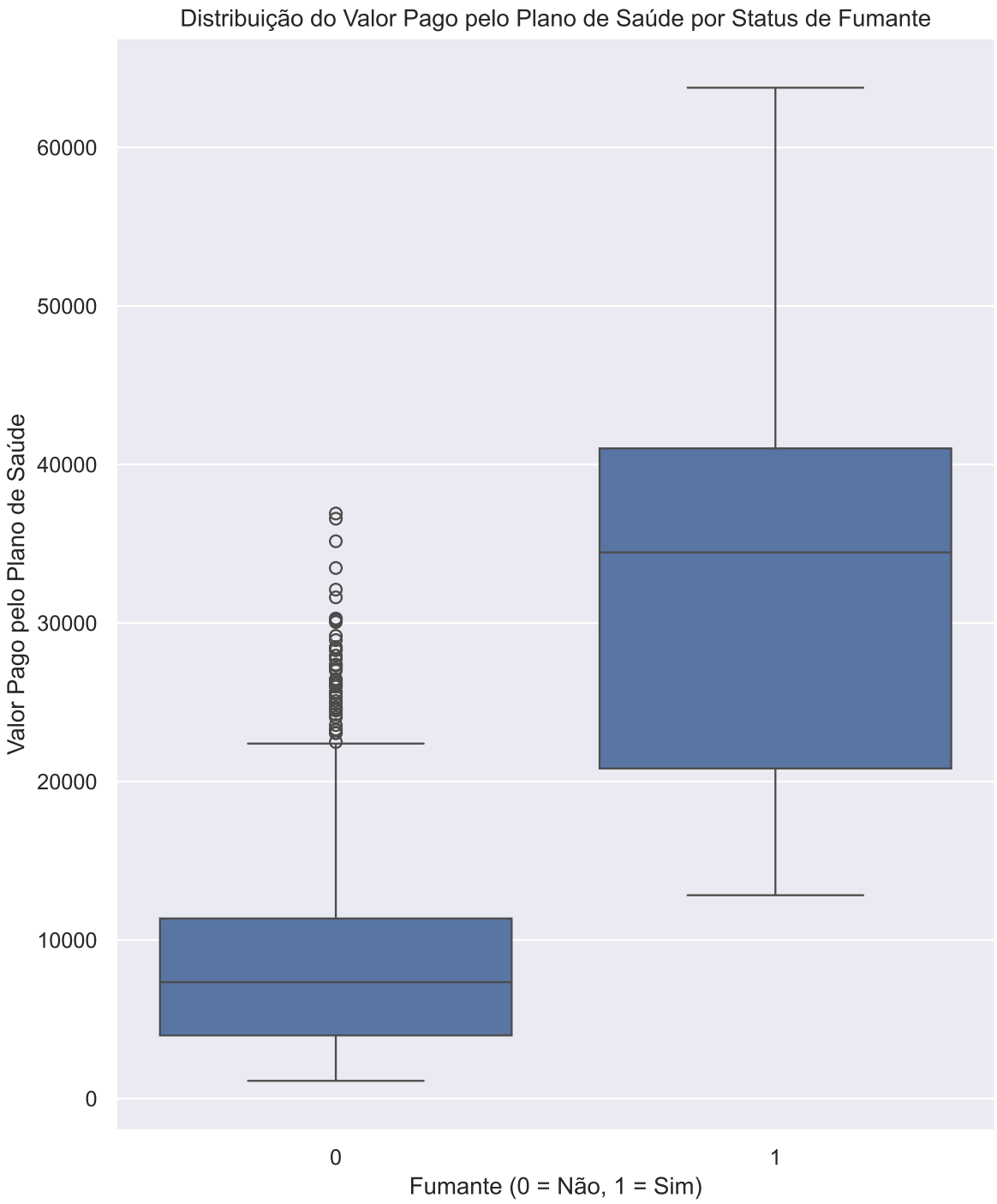
```
                            OLS Regression Results
==============================================================================
Dep. Variable:          Valor_Pago_PS   R-squared:                       0.658
Model:                            OLS   Adj. R-squared:                  0.657
Method:                 Least Squares   F-statistic:                     1284.
Date:                Mon, 21 Oct 2024   Prob (F-statistic):          1.03e-311
Time:                        16:36:27   Log-Likelihood:                -13760.
No. Observations:                1338   AIC:                         2.753e+04
Df Residuals:                    1335   BIC:                         2.754e+04
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept              -3459.0955    998.279     -3.465      0.001   -5417.463   -1500.728
Fumante                 2.359e+04    480.180     49.136      0.000    2.27e+04    2.45e+04
Indice_Massa_Corporal    388.0152     31.787     12.207      0.000     325.656     450.374
==============================================================================
Omnibus:                      153.688   Durbin-Watson:                   2.034
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              239.253
Skew:                           0.805   Prob(JB):                     1.11e-52
Kurtosis:                       4.303   Cond. No.                         161.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Visual Representation of linear regression between 'Valor_Pago_PS' and 'Indice_Massa_Corporal' taking a new dimension whith the attribute that distinguishes different categories, hue='Fumante'



BOXPLOT

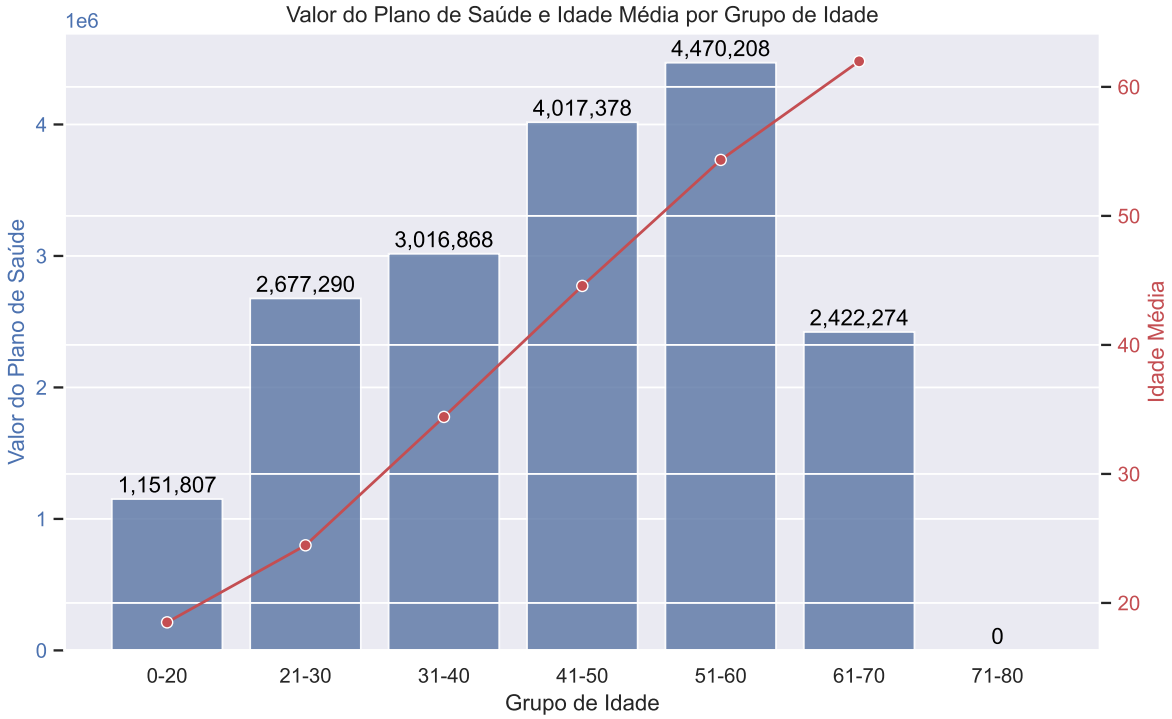
Distribution and Outliers. For the non-smoking group (0), there are several outliers. These are represented by the circles above the upper limit of the box. They indicate that there are non-smokers who pay much more for health insurance compared to most.



The visualization of the Boxplot was not very intuitive for the user, so, in a more explanatory way, a bar/line graph was generated for better visualization. When analyzing the graph, it is

observed that the age groups that spend the most on health insurance are between 41-60 years old, and it can be assumed that in this age group more medical care usually arises.

```
C:\Users\admin1\AppData\Local\Temp\ipykernel_7600\4198989146.py:7: FutureWarning: The default
  grouped_data = data.groupby('Grupo_Idade').agg({'Valor_Pago_PS': 'sum', 'Idade': 'mean'}).
```



CONCLUSION, GREETINGS

First all, i want to thank my mentor that help me to get a better insight about my first analysis with python in dataset that was challenged to me and my friends who helped as external viewers.

About the analysis, with all insights generated trought libraries functions used, it is observed that the age groups that spend the most on health insurance are between 41-60 years old, and it can be assumed that in this age group more medical care usually arises. categorating we can conclue that The graph suggests that smokers tend to pay higher amounts for health insurance, and there is a stronger correlation between BMI and the amount paid in this group. That is, for smokers, people with higher BMI seem to pay more. For non-smokers, the amount paid by the health plan seems to vary less in relation to BMI, that is, the impact of BMI on the amount paid is lower in this group, containing more outliers.