NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Faculty of Computer Science
Bachelor's Programme "Data Science and Business Analytics"

Project N1

Made by:

Khaibrakhmanov Timur,

Timchenko Daniil,

Malutin Alexander

Moscow 2023

Contents

1 Introduction

In this study we are going to work out the econometric regression model, which would allow to predict, based on several regressors, the Healthcare charges of individual billed by their Insurance provider, which is considered the variable of interest.

Before we began our investigation it was important to familiarize ourselves with the subject. We have selectively read some information on multiple regression analysis in [2].

Furthermore, in our work we have faced the issue with multimulticollinearity of the independent variables. In [2] we did not manage to find proper solution, thus proceeded to read about it in [1].

Moreover, several Internet resources, such as the statmodels [4] Python module, scikit-learn [3] Python library were used for the model implementation.

2 Data description

For the econometric analysis, we chose a dataset with data on patients that could in any way influence their medical insurance charges per year. The original dataset has 9 columns and 1070 rows. The dataset represents montly payment to insurance in a year.

- *Unnamed: 0* - a set of indices that do not reflect useful information.
- *age* - information about a person's age.
- *sex* - information about a person's sex.
- *bmi* - information about a person's BMI. Body Mass Index (BMI) is a person's weight in kilograms divided by the square of height in meters.
- *children* - information about the number of children a person has.
- *smoker* - if the value in this column is 1, the person smokes, if 0, he doesn't.
- *region* - information about a person's region. The column is made up of different numbers, the source does not give information which regions these numbers mean, but the differences will give an indication of whether the region of residence is relevant to hospital charges.
- *BMI group* - information about BMI group a person belongs to.
- *charges* - information about the hospital insurance charges.

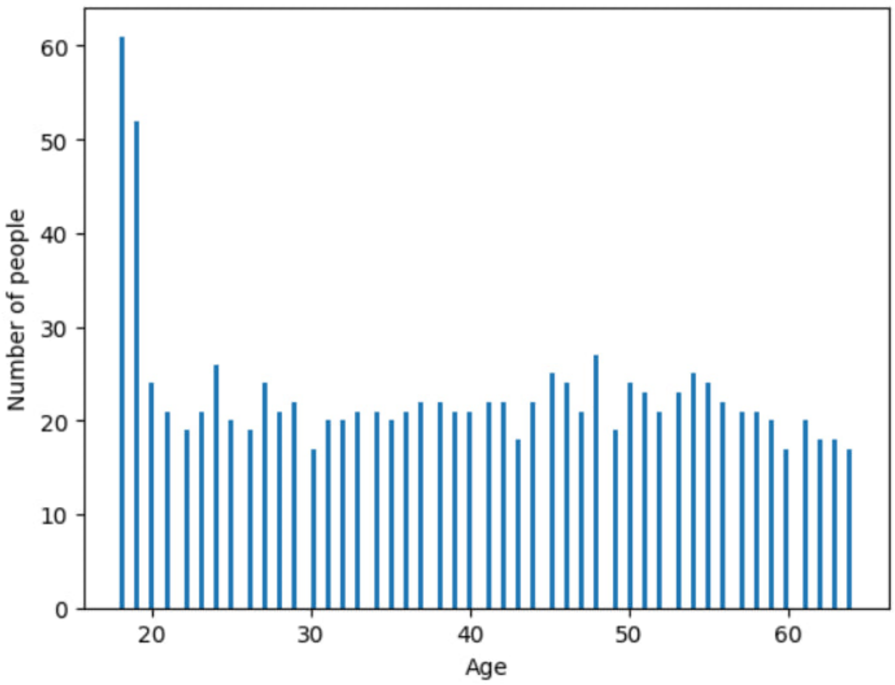

Figure 2.1: Number of people of different ages

This dataset includes people between the ages of 18 and 64. The largest and the smallest numbers of patients are 18 and 19 years old, 61 and 62 respectively. There are approximately equal numbers of people of every other age. The distribution of people with different ages can be seen in Figure 2.1

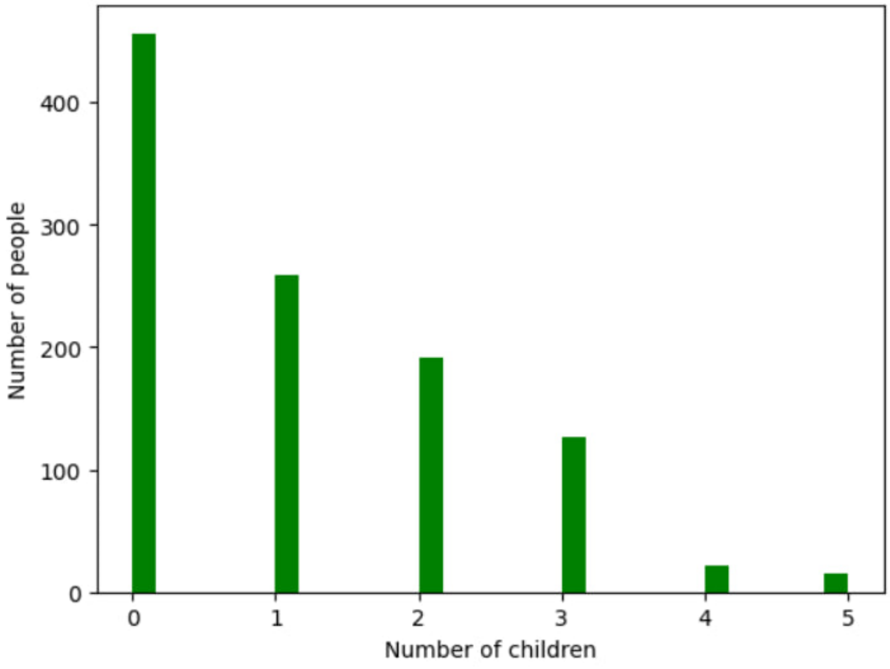

Figure 2.2: Number of people with different number of children

Each of the 4 regions has the same number of people. There are also the same number of patients of each sex. There are 4 times more smokers in the dataset than non-smokers. The distribution of people with different numbers of children can be seen in Figure 2.2

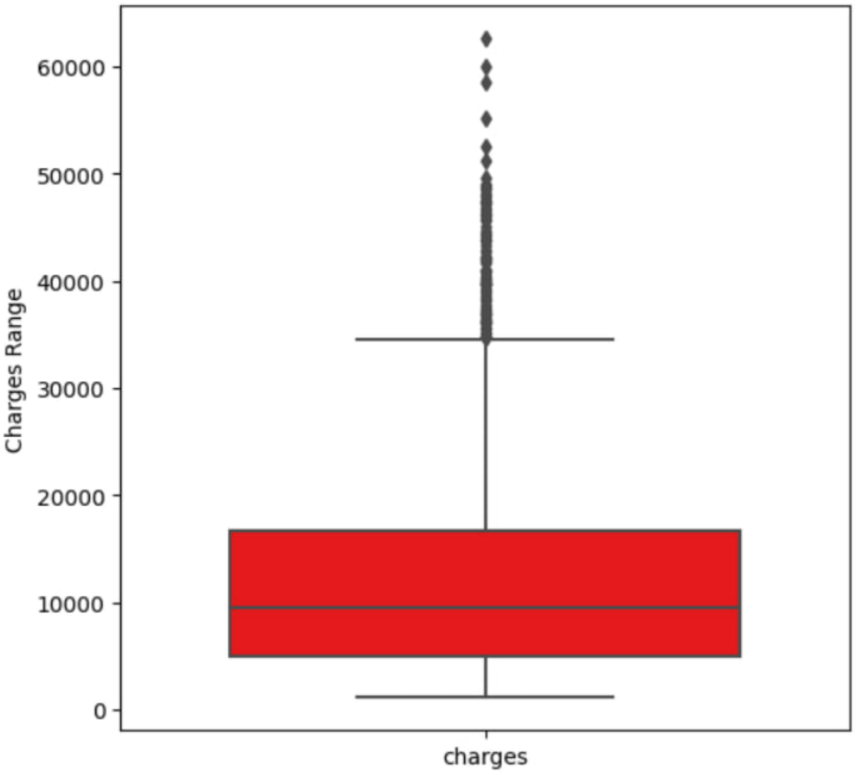

Figure 2.3: Hospital charges boxplot

Figure 2.3 shows that the median of the charges column is 10000, but there are a large number of abnormally large values.

3 Economic model

3.1 Feature Analysis

More deep analysis on data was conducted. Especially statistical dependency of features were estimated by plotting and testing techniques. All tests are conducted with $a = 0.1$ significance level. It is evident ,that insurance charges correlate with age.

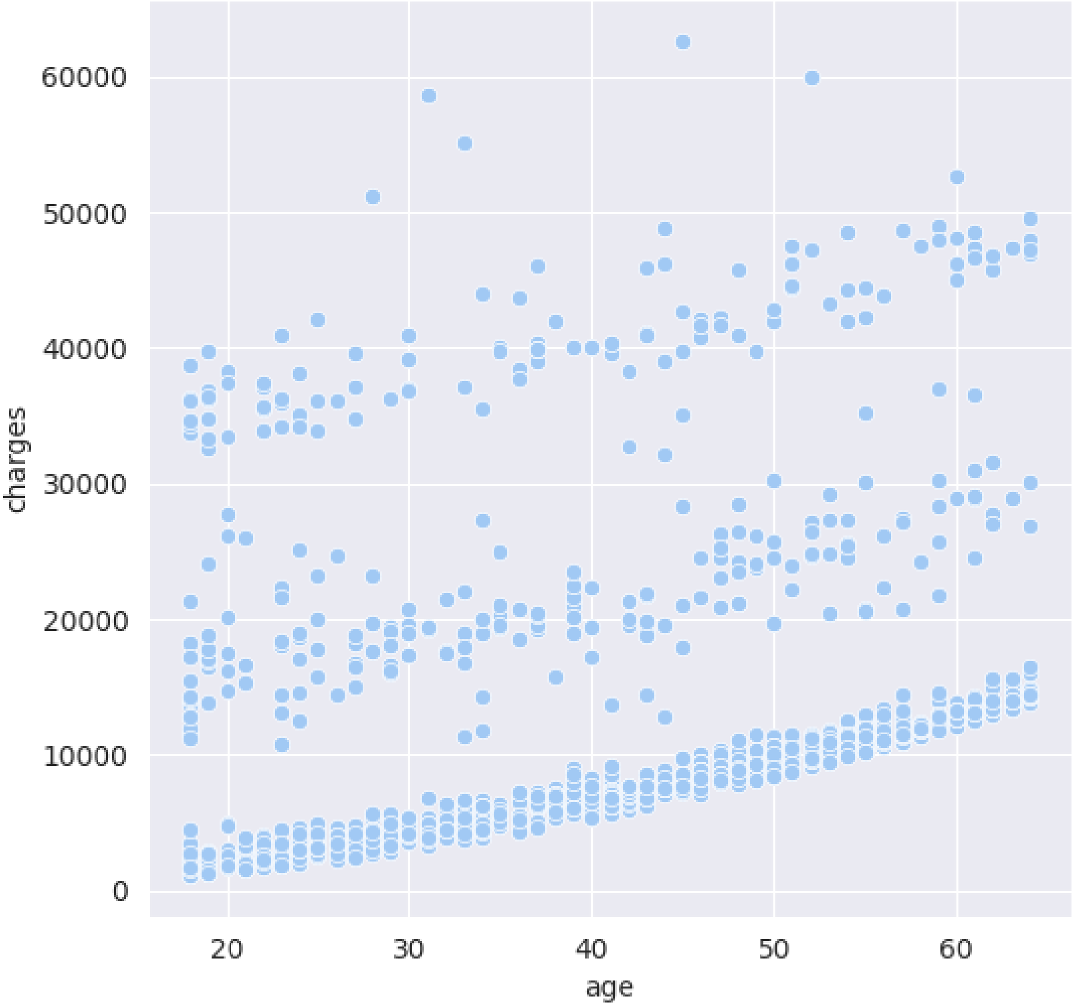

Figure 3.1: Dependency between age and charges

It seems like data may be split on 3 subsets, which represent 3 groups. These groups are separated by charging. Lower class pay 0-10000, middle class pays 10000-25000, higher class pays 30000 - 45000 per year approximately.

Examining the data on sex, upper 25% percent quantile of opposite sex shows slight important shift upwards by 5000 in charges. Difference in sex may affect charges.

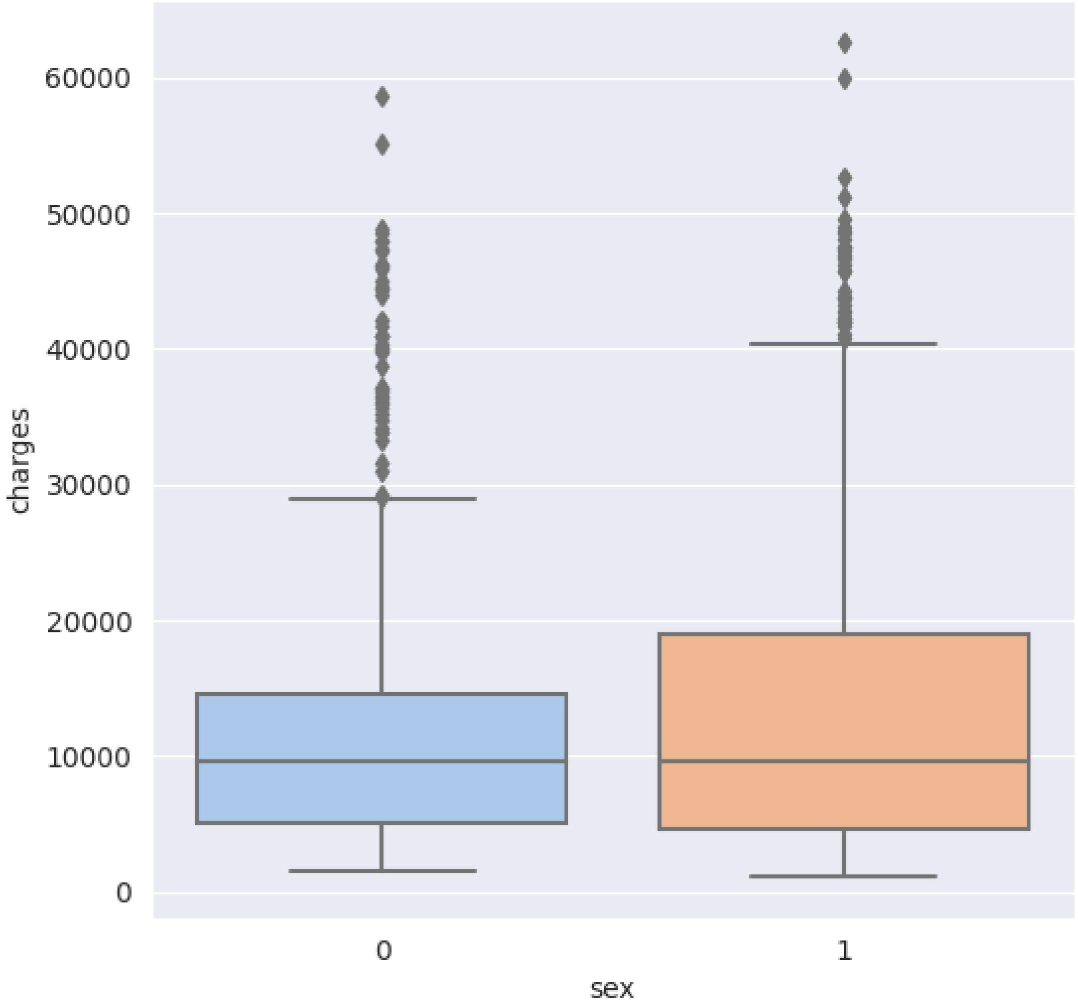

Figure 3.2: Dependency between sex and charges.

For children number dependency represents bell curve. It seems like if this variable is being examined as numerical, it may be approximated by some negative polynomial of degree 2. But better to transform it into dummy variable.

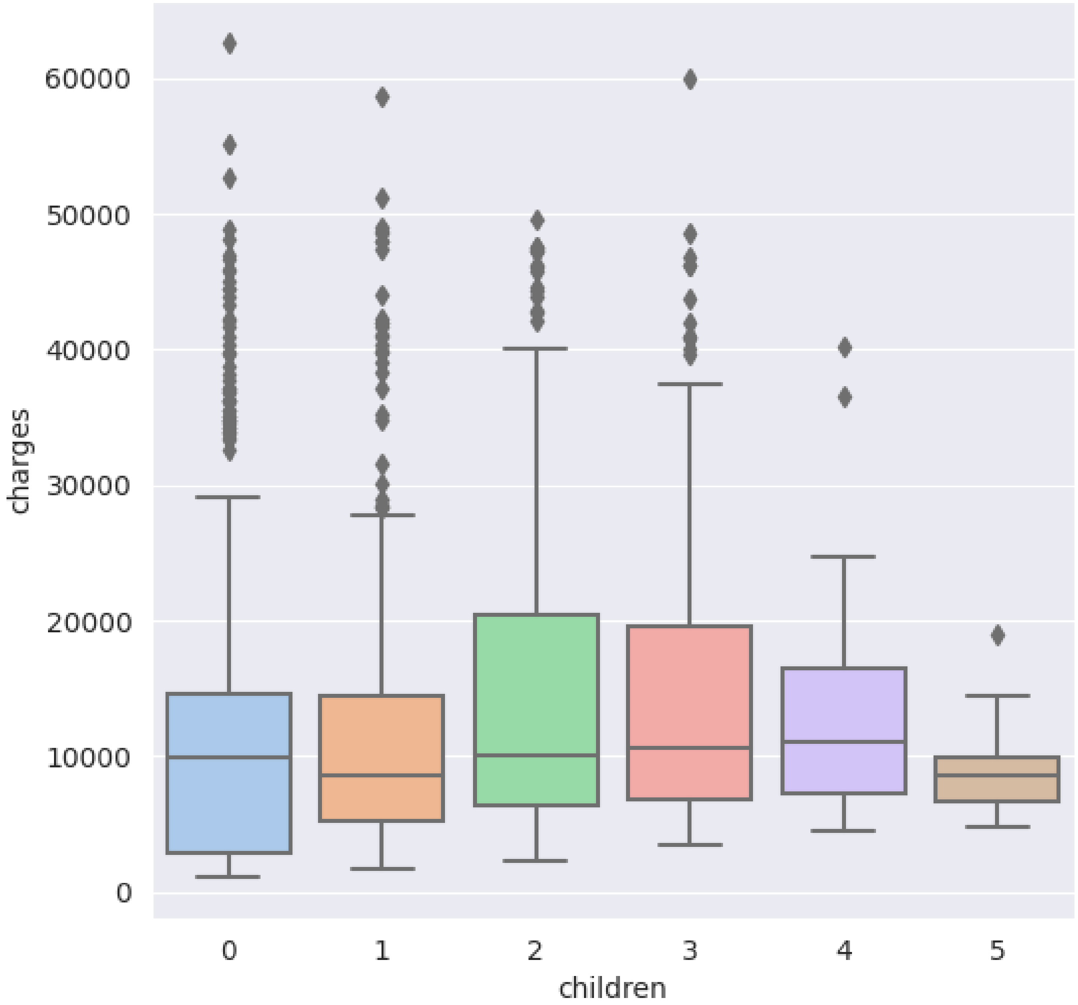

Figure 3.3: Dependency between number of children and charges.

As for smoking dummy variable, it is apparent, that there is a significant increase in price for the insurance for smoking users.

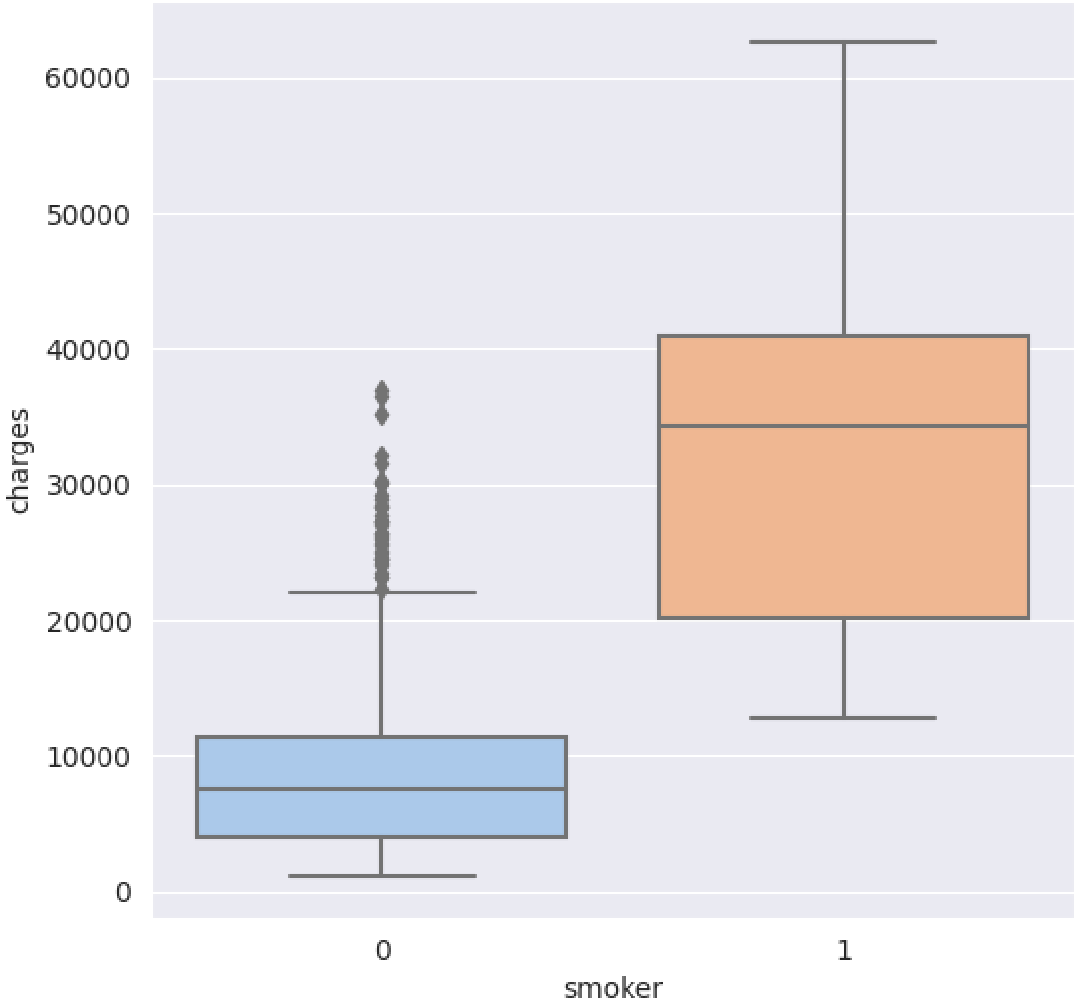

Figure 3.4: Dependency between smoking and charges.

Region variable shows almost no signs of dependency:

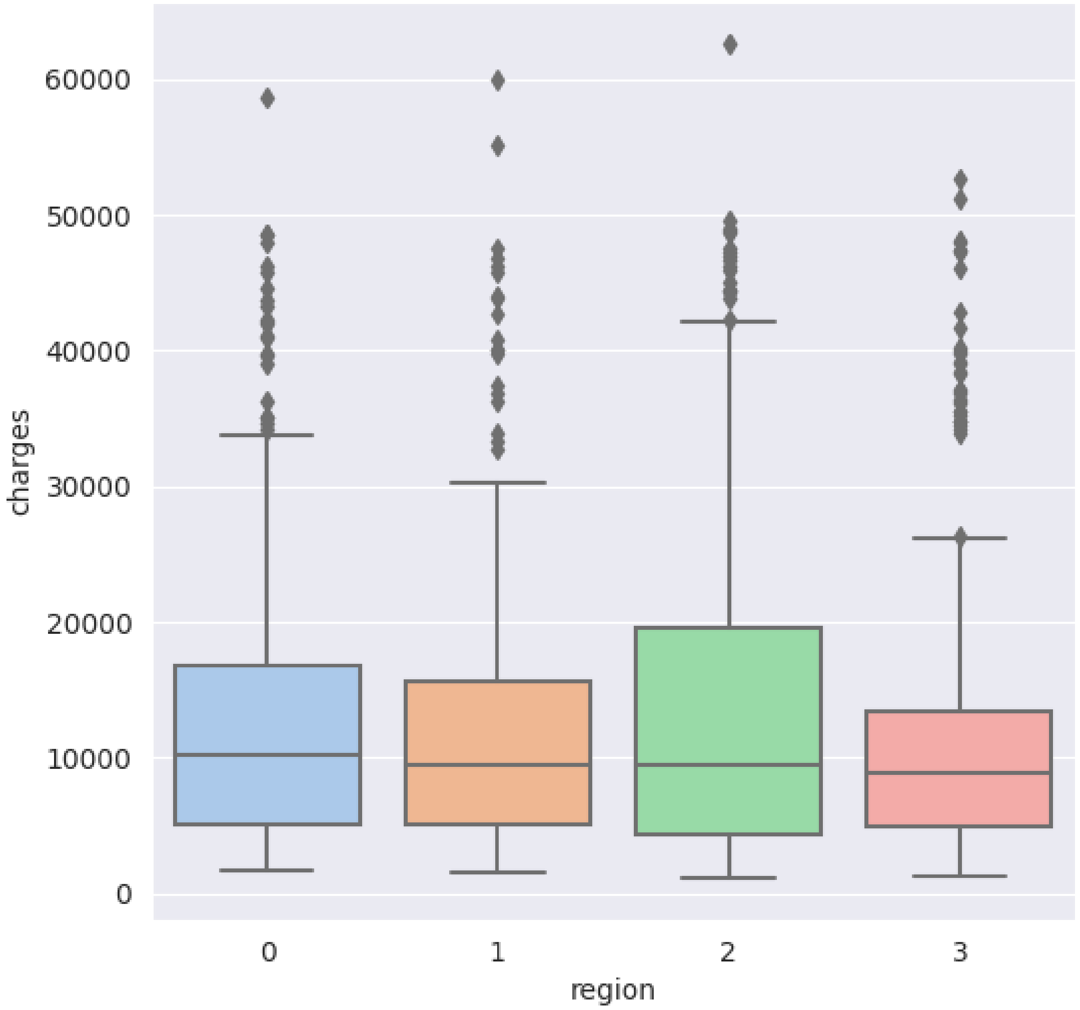

Figure 3.5: Dependency between smoking and charges.

3.2 Statistical tests

For categorical features, one-way ANOVA was used, all categorical features except "region" shown enough evidence to believe that they are truly dependent with target variable. These features are: sex, smoker, children. For numerical variable age, pearson correlation test was applied which also resulted in rejection of null hypothesis, and approval of significant dependence. P values can be examined in the table below:

Variable	P-Value
Sex	0.063
Smoker	1.03^{-211}
Children	0.012
Region	0.138
Age	$5.69 * 10^{-21}$

Therefore region variable was omitted.

3.3 Model Choice

Such features as sex, smoker, children and age does not seem like they should have multiplicative effect on charge, so it is probably not well generalized by a logarithmic model. It is more like a linear dependency. We expect constant additions to charge, influenced by our features. We also can consider polynomial models, which capture non-linear dependency, but obtained data seems highly linear, except children feature, but as this feature can be considered categorical (as we have low number of categories and they are whole numbers, logically they also represent different sub-groups of people), we can perform One-Hot encoding and approximate well with a linear model. Therefore the best strategy would be to use linear regression model. Several types of model building was attempted, but none of them achieved surpassing performance compared to base model ($Y \sim sex + smoker + children + age$, formula without coefficients). Region feature addition and lasso regularization with hyper-parameter tuning did not surpass base model.

4 Model estimation results

Model, which was obtained after the coding implementation:

$$
\begin{aligned}
Y = 2827.0183 &+ 269.3562 * Age - 64.4194 * Sex + 2.317e + 04 * Smoker + \\
&+ 562.3537 * No\_children + 1984.1002 * One\_child + 578.2710 * Two\_children + \\
&+ 2123.8547 * Three\_children + 563.3992 * Four\_children
\end{aligned}
$$

, **where** $No\_children$, One_child, $Two\_children$, $Three\_children$, $Four\_children$ and $Smoker$ are dummy variables.

4.1 Variables interpretation

After the t-test conduction in order to acquire the significance of variables, the only significant variables turned out to be Age, Smoker and One_child, thus the final model(see final model in our notebook):

$$Y = -2598.9032 + 270.7177 * Age + 2.316e + 04 * Smoker + 1670.1845 * One\_child, R^2 = 0.716,$$

which means, that our model explains Hospital charges of an individual pretty good.
As all dependencies are linear, they can be interpreted as follows: $\triangle Y = \hat{\beta} \triangle X$, that is, with a rise in either Age or Smoker or One_child there follows a linear increase in Y. Moreover, whether an individual is a smoker or not seems to have the greatest effect on the price of hospital charges.

5 Discussion

Our team selected a dataset with data on hospital charges for individuals. We worked with the data, examined all the columns and the relationship of data in each of them. In the following paragraphs the statistical dependency of features was examined and patterns were found. Through statistical tests it was found out that all categorical features are truly dependent with target variable. Then we explained why this model was chosen and showed the model estimation results.
The results could be improved with better filtering and cleaning of the data. Perhaps also isotonic regression could be used as it has good interpretability and can approximate on a non-linear data..

References

[1] Jiehong Cheng, Jun Sun, Kunshan Yao, Min Xu, and Yan Cao. “A variable selection method based on mutual information and variance inflation factor”. In: *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 268 (2022), p. 120652. ISSN: 1386-1425. DOI: https://doi.org/10.1016/j.saa.2021.120652. URL: https://www.sciencedirect.com/science/article/pii/S1386142521012294.

[2] Christopher Dougherty. *Introduction to econometrics*. English. Fifth. Oxford University Press, 2021. ISBN: 0199676828;9780199676828;

[3] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[4] Skipper Seabold and Josef Perktold. “statsmodels: Econometric and statistical modeling with python”. In: *9th Python in Science Conference*. 2010.