Statistical Report Technical Test Racoon Group

date 17/06/2022

Klayton Crul Correa

E-mail: klayton12341@live.com

Linkedin: https://www.linkedin.com/in/klayton-crul

Github: https://github.com/klaytoncrul/data-science-projects

Telefone: (45)998023698

Objetive:

Monk Racoon data Science internship program

SAMPLE DATABASE

Base: 639 registers splited in 5 variables
Collection period: 01/01/2019 to 09/30/2020

Time Days: 639 (639 days),

Time in months: 21 months

Time in years: 1 year and 9 months

Descriptive data statistics

Descriptive statistics, which the basic objective is to synthesize series of values of the same nature, thus allowing a global view of the variation of these values, it organizes and describes the **data** in three ways: through tables, graphs and **descriptive measures**.

	receita	transacoes_blog	transacoes_site	usuarios_blog	usuarios_site
count	639.00	639.00	639.00	639.00	639.00
mean	1623891.19	528.35	19039.14	1439.85	101610.49
std	1160581.16	1201.78	13677.73	3369.87	37240.23
min	32085.00	0.00	3557.00	0.00	26298.00
25%	807342.00	0.00	11013.00	0.00	77727.00
50%	1263161.00	0.00	16069.00	0.00	96104.00
75%	2232769.50	0.00	22606.50	0.00	117586.50
max	12266844.00	5586.00	188955.00	13059.00	369989.00

Now that we have our descriptive statistics we can start our statistical analysis.

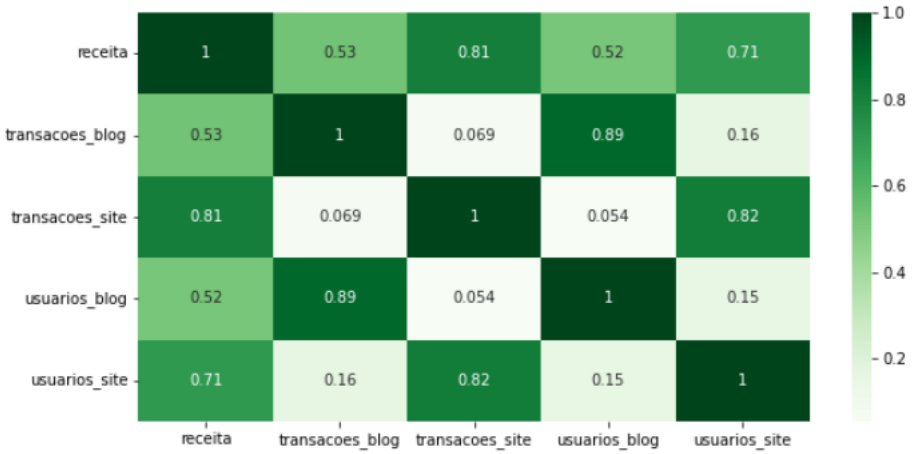
Correlation matrix

The **correlation coefficient** is a measure of linear association between two variables and it´s between -1 and +1, with -1 indicating a perfect negative association and +1 indicating a perfect positive association.

Correlational analysis indicates the relationship between 2 linear variables and the values will always be between +1 and -1. The sign indicates the direction, whether the correlation is positive or negative, and the size of the variable indicates the strength of the correlation.

It should be noted that, as the coefficient is conceived from the linear fit, then the formula does not contain fit information, that is, it is composed only by the data

	receita	transacoes_blog	transacoes_site	usuarios_blog	usuarios_site
receita	1.0000	0.5317	0.8126	0.5180	0.7112
transacoes_blog	0.5317	1.0000	0.0689	0.8933	0.1623
transacoes_site	0.8126	0.0689	1.0000	0.0543	0.8200
usuarios_blog	0.5180	0.8933	0.0543	1.0000	0.1518
usuarios_site	0.7112	0.1623	0.8200	0.1518	1.0000



Interpreting Pearson's correlation coefficient

0.9 plus or minus indicates a very strong correlation.

0.7 to 0.9 positive or negative indicates a strong correlation.

0.5 to 0.7 positive or negative indicates a moderate correlation.

0.3 to 0.5 positive or negative indicates a weak correlation.

0 to 0.3 positive or negative indicates negligible correlation.

Boxplot, Frequency Distribution and Behavior of the Dependent Variable Revenue

Boxplot

The boxplot or box diagram is a graphical tool that allows you to visualize the distribution and outliers of the data, thus providing a complementary mean to develop a perspective on the character of the data. In addition, the boxplot is also a comparative graphic layout.

Descriptive statistics measures such as the minimum, maximum, first quartile, second quartile or median and third quartile form the boxplot.

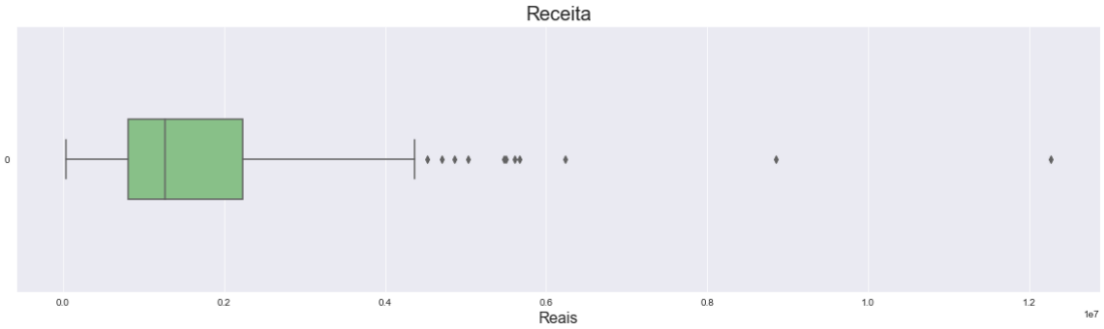
Frequency Distribution

The frequency distribution is an arrangement of values that one or more variables taken in a sample. Each entry in the table contains the frequency or count of occurrences of values within a specific group or range, and thus the table summarizes the distribution of sample values.

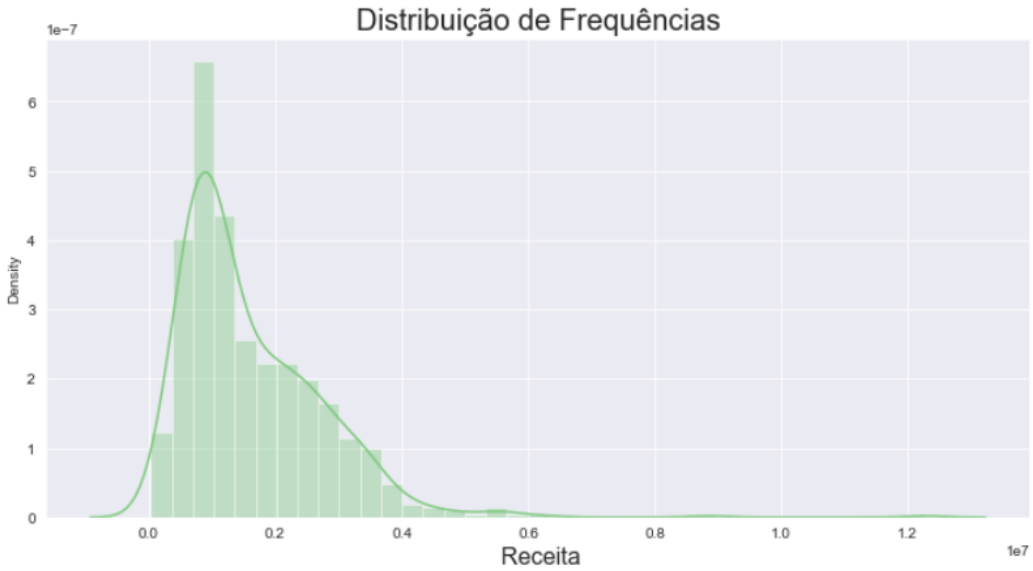
Frequency Distribution of the dependent variable (y)

With the aid of the boxplot and the frequency distribution , the asymmetry to the right of the data being analyzed was identified, as there is a greater concentration of values in the area of lower values of the sample.

Analysis of the frequency distribution and outliers of the Revenue variable using boxplot and histogram



In the boxplot to consider especially, there is a group of people who manufacture purchases far above the general base of analysis, however, to study this variability, however, to study the data and analyze them a part. To reduce the impact, before running the model I normalized it with logarithmic transformation and at the end of the modeling I reversed the transformation to give me the real value.



Frequency distribution of the dependent variable

With the aid of the boxplot and the histogram, the asymmetry to the right of the data being analyzed was identified, as there is a greater concentration of values in the zone of lower values of the sample.

The outlayers were not taken into account because in variables with these characteristics it always shows many outlayers generated by purchases with higher values.

Dispersion Between Dataset Variables

In the **Scatter Diagram**, we can also **analyze** whether the correlation is strong or weak:

Strong: the greater the correlation between the variables, the greater the proximity of the points, that is, they will be less dispersed.

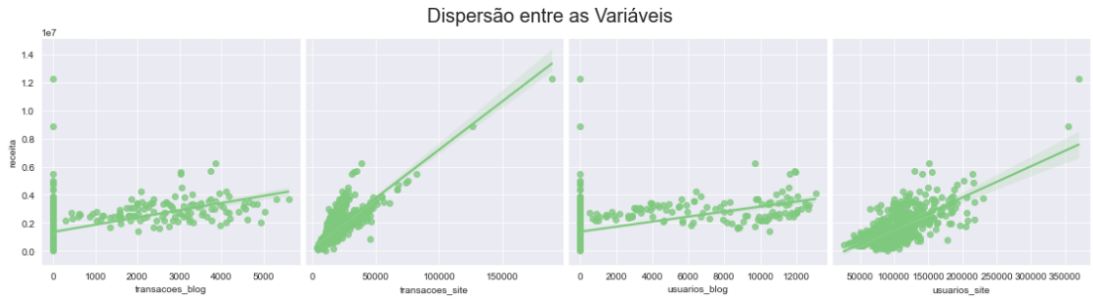
Weak: the lower the correlation between the variables, the more dispersed the points will be.

So in the graphs below we identify whether the dependent variable and the explanatory variables have any linear relationship.

Regression line

In studies of two-dimensional distributions, when there is a correlation between the variables, it is often interesting to predict the value of one of the variables when the corresponding value of the other variable is known.

The process to be used consists of drawing a line that "best" fits (approximates) the points of the scatterplot.



After tracing the regression line, we understand that as the points of the scatter diagram are formed along the line, we have a linear correlation.

Linear Regression

Regression analysis concerns the study of the dependence of a variable (the dependent variable) on one or more variables, the explanatory variables, in order to estimate and/or predict the average value of the first in terms of the known or fixed values of the second. .

Getting the regression coefficients

The regression coefficients $\beta 2$ and $\beta 3$ are known as regression partial factors or angular partial factors.

An interesting aspect of the log-linear model, which has made it widely used in applied work, is that the slopes $\beta 2$ and $\beta 3$, measure the elasticities of Y with respect to $X2$, $X3$, $X4$ and $X5$, that is, the percentage change of Y corresponding to a given (small) percentage change in $X2$, $X3$, $X4$ and $X5$.

	Parâmetros
Intercepto	6.084591
log_transacoes_site	-0.059825
log_usuarios_site	0.134347
log_usuarios_blog	1.030719
log_transacoes_blog	-0.186352

According to the estimated coefficients in our Linear Regression model in our variables, our results would look like this:

Blog Transactions → Holding the value constant, a 1% decrease in the number of Blog Transactions generates, on average, a 0.18% decrease in revenue.

Website Transactions → Holding the value constant, a 1% decrease in the number of website transactions generates, on average, a 0.06% decrease in revenue.

Site users → Keeping the value constant, a 1% increase in the number of site users generates, on average, an increase of 0.13% in revenue.

Blog Users → Holding constant, a 1% increase in the number of blog users generates, on average, a 1.03% increase in revenue.

Transforming the Data to try to correct the asymmetry of our variables

Why?

Parametric tests assume that the sample data were collected from a population with a known probability distribution. Most statistical tests assume that the data follow a normal distribution (Student's t, confidence intervals, etc.).

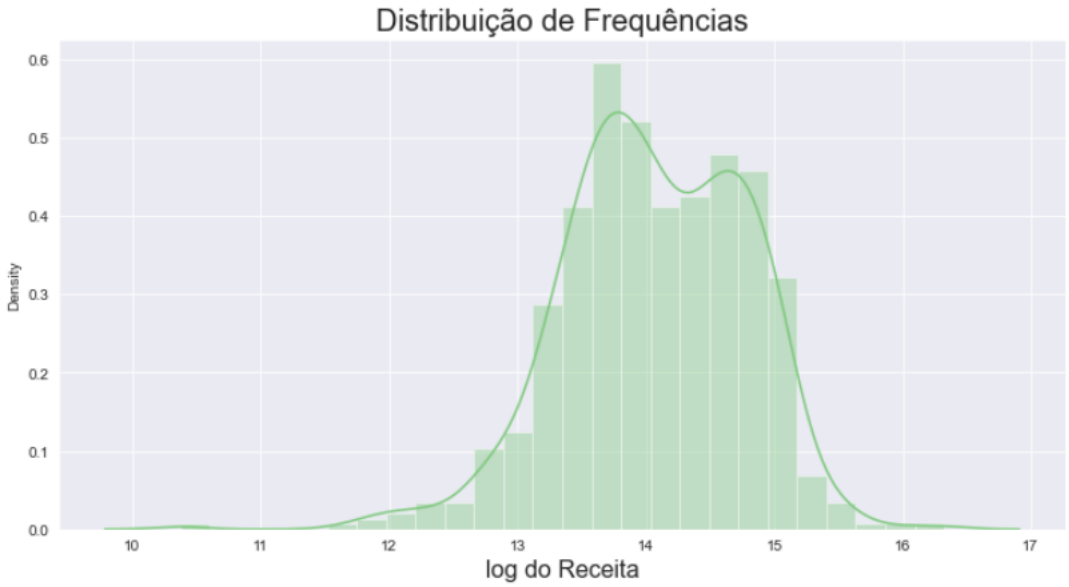
We found that the dependent variable we are analyzing is right skewed so I will use a variable transformation technique to try to correct this problem and, who knows, estimate a linear regression model with this database after the transformation.

Applying the logarithmic transformation

Logarithmic transformation is often used when the data have a positively skewed distribution and there are some large values. So I'll use the log transformation to make the variances more constant and normalize the data.

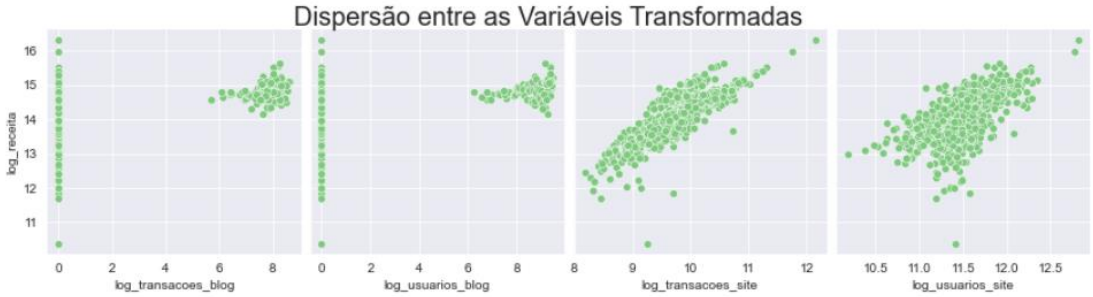
After the logarithmic transformation, the asymmetry of the revenue variable assumed a more normalized form.

Frequency distribution of the transformed dependent variable Revenue



After the logarithmic transformation asymmetry of the revenue variable assumed a more normalized form.

Scatter plots between the transformed variables of the dataset

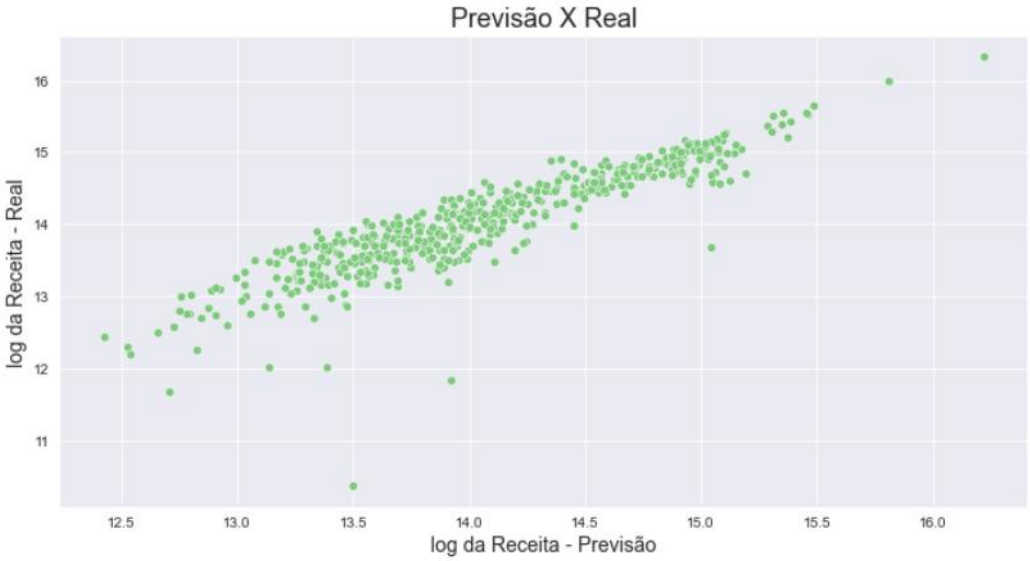


Checking Linear Relationship

As the data related to the blog are only from the last 3 months, for this reason in the scatter plot of the transformed variables they are in clusters in the upper right corner and we can see that the scatter points maintain a linear correlation.

Graphical Analysis of the Results of the Linear Regression Model after Logarithmic Transformation

Scatter plot between estimated value and actual value.

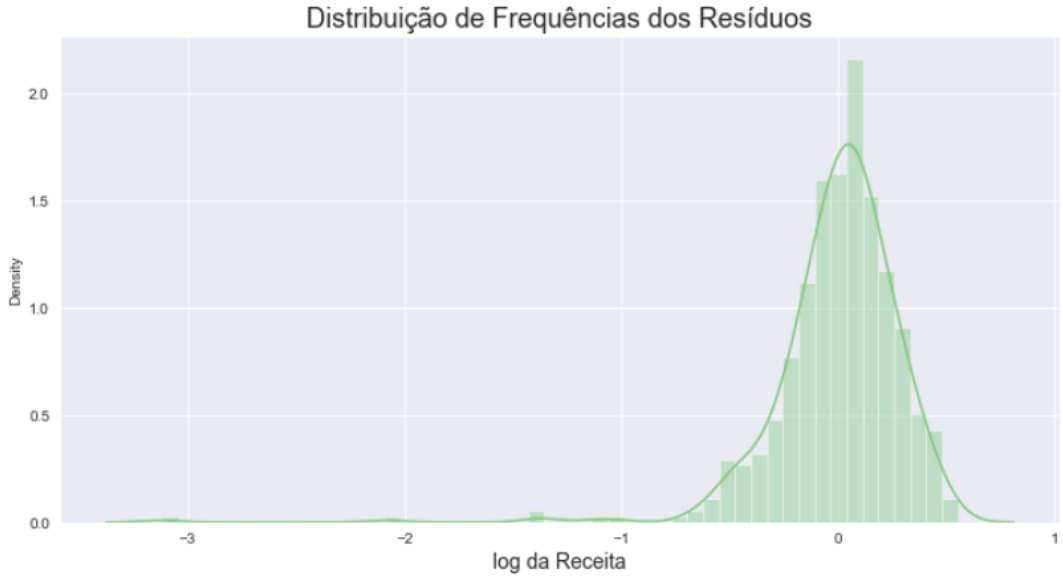


As we don't have a large scatter in our graph, this indicates how accurate our estimate is in some way.

getting the waste

Residuals indicate the natural variation of the data, a random factor (or not) that the model did not capture. If the model assumptions are violated, the analysis will lead to dubious and unreliable results for inference.

Residual frequency distribution plot



As a result, we have a nice curve that indicates well-behaved data, which favors a well-estimated model.

Comments

It is important to note that our regression model is not a model to be used commercially, as it would need more variables and elements to operate in a functional way. However, we were able to foresee what would be the appropriate behavior for an estimative model to provide us with reliable information, and for this purpose, our project fits perfectly.

Hypothesis test

Statistical tests are decision rules that allow assessing the reasonableness of hypotheses made about population parameters and accepting or rejecting them as probably true or false based on a sample.

Tests:

Normality test

The normality test tests the null hypothesis $H0$H0 that the sample comes from a normal distribution.

The python function used was the **normaltest** which tests the null hypothesis $H0$ that the sample comes from a normal distribution.

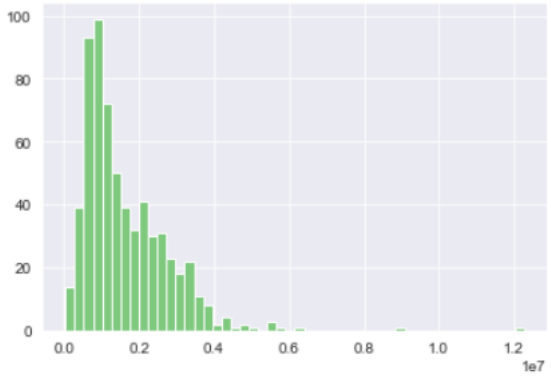
Significance level

The significance level used in all tests was **0.05%.**

Analyzing the test and the graphs after the normality test.

P_value after testing = 1.1361263738507747e-60

revenue variable



p-value criterion

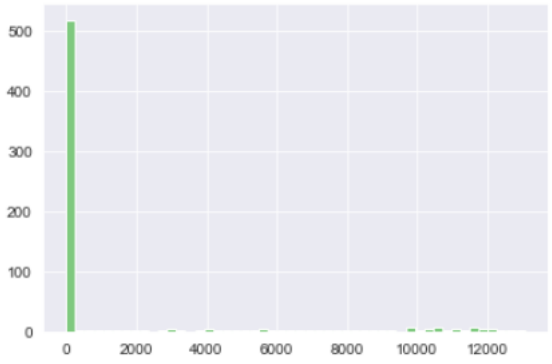
Reject $H0$ if the value p≤0.05

Analyzing the chart

Regarding asymmetry, we have several extreme positive values, which would suggest positive asymmetry.

After the **normaltest**, a P_value of **1.1361263738507747e-60** was returned and as our P_value is greater than our significance level of 0.5%, H0 is rejected and the sample **does not** come from a **normal distribution**.

Variable Blog users



p-value criterion

Reject $H0$ if the value $p \leq 0.05$

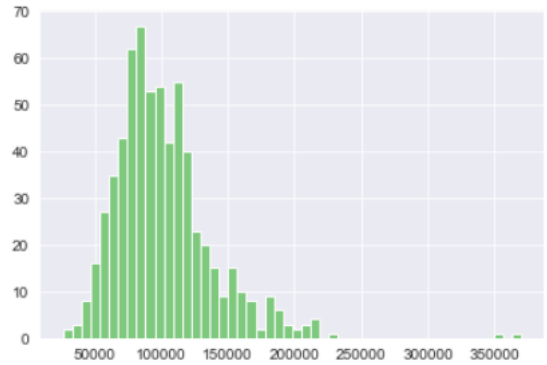
Analyzing the chart

Regarding asymmetry, we have several extreme positive values, which would suggest positive asymmetry.

After the **normaltest**, a P_value of **1.1361263738507747e-60** was returned

and as our P_value is greater than our 0.5% significance level, the H0 is rejected and the sample **does not** come from a **normal distribution**.

Site Users Variable



p-value criterion

Reject $H0$ if the value $p \leq 0.05$

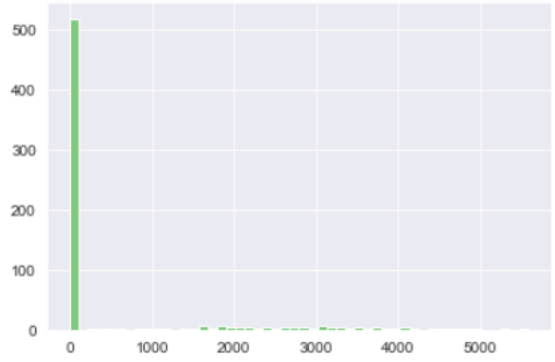
Analyzing the chart

Regarding asymmetry, we have several extreme positive values, which would suggest positive asymmetry.

After the **normaltest**, a P_value of **2.3947832487712726e-59** was returned

and as our P_value is greater than our 0.5% significance level, the H0 is rejected and the sample **does not** come from a **normal distribution**.

Blog transactions variable



p-value criterion

Reject $H0$ if the value $p \leq 0.05$

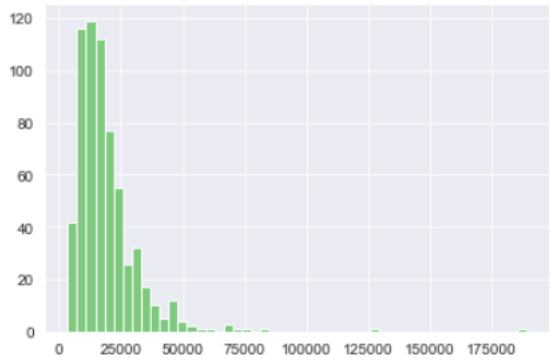
Analyzing the chart

Regarding asymmetry, we have several extreme positive values, which would suggest positive asymmetry.

After the **normaltest**, a P_value of **2.896413130631297e-60** was returned

and as our P_value is greater than our 0.5% significance level, the H0 is rejected and the sample **does not** come from a **normal distribution**.

Website transactions variable



p-value criterion

Reject $H0$ if the value $p \leq 0.05$

Analyzing the chart

Regarding asymmetry, we have several extreme positive values, which would suggest positive asymmetry.

After the **normaltest,** a P_value **of 4.8317020805360967e-147** was returned

and as our P_value is greater than our 0.5% significance level, the H0 is rejected and the sample does **not come** from a **normal distribution.**

NON-PARAMETRIC TESTS

Nonparametric tests, also known as free distribution tests, are those based on certain hypotheses, but which do not have a normal organization. They usually contain statistical results from their sorts.

Nonparametric tests have some limitations, among them they are not strong enough when a normal hypothesis is fulfilled. This can cause it not to be rejected, even if it is fake. Another of their limitations is that they require the hypothesis to be changed when the test does not correspond to the procedural question if the sample is not proportional.

In our dataset we have data about our sales site and we created the blog with content about 3 months ago, and we want to understand what influence a blog of our brand has.

Then I will select two samples from our dataset. In order to prove such influence on the company's revenue **I will test the equality of the averages** between these two samples with a **significance level of 5%**.

Tests:

Mann-Whitney test

Mann-Whitney is a nonparametric test used to verify whether two independent samples were selected from populations that have the same mean. As it is a non-parametric test, Mann-Whitney becomes an alternative to the parametric test of means comparison.

The python function used was the **mannwhitneyu** which tests the null hypothesis $H0$ that the sample comes from a normal distribution.

Significance level

The significance level used in the tests was **0.05%.**

p-value criterion

Reject $H0$ if the value $p \leq \alpha$

Testing the blog users and website users variable

p-value criterion

Reject $H0$ if the value $p \leq \alpha$

After the **mannwhitneyu**, a P_value of **2.391193425606197e-225** was returned, and as our P_value is greater than our significance level of 0.5%, the H0 is rejected. According to the results, we accept the hypothesis that **there is an influence** of the blog on our brand, that is , we conclude that the average of blog users and website users have influence over our brand.

Testing the blog transactions vs website transactions variable

p-value criterion

Reject $H0$ if the value $p \leq \alpha$

After the **mannwhitneyu**, a P_value of **4.689855769706814e-225** was returned, and as our P_value is greater than our significance level of 0.5%, the H0 is rejected. According to the results, we accept the hypothesis that **there is an influence** of the blog on our brand, that is, we concluded that the average of blog transactions and website transactions influence our brand.

Testing the blog users and revenue variable

p-value criterion

Reject $H0$ if the value $p \leq \alpha$

After the **mannwhitneyu**, a P_value of **2.391204760924378e-225** was returned, and as our P_value is greater than our significance level of 0.5%, the H0 is rejected. According to the results, we accept the hypothesis that **there is an influence** of the blog on our brand, that is, we concluded that average blog users and revenue have an influence on our brand.

Testing the blog transactions and revenue variable

p-value criterion

Reject $H0$ if the value $p \leq \alpha$

After the **mannwhitneyu**, a P_value of **2.391200982478832e-225** was returned, and as our P_value is greater than our significance level of 0.5%, the H0 is rejected. According to the results, we accept the hypothesis **that there is an influence** of the blog on our brand, that is, we concluded that average blog users and revenue have an influence on our brand.