Relatório Estatístico Prova Técnica Grupo Racoon

Data 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

Objetivo:

Estágio em Ciência de dados

AMOSTRA BASE DADOS

Base: 639 registro divididos em 5 variáveis
Período de coleta: 01/01/2019 à 30/09/2020
Tempo Dias: 639 (639 dias),
Tempo em meses: 21 meses
Tempo em anos: 1 ano e 9 meses

Estatísticas descritivas dos dados

A **estatística descritiva**, cujo objetivo básico é o de sintetizar uma série de valores de mesma natureza, permitindo dessa forma que se tenha uma visão global da variação desses valores, organiza e descreve os **dados** de três maneiras: por meio de tabelas, de gráficos e de medidas **descritivas**.

	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

Agora que possuímos nossas estatísticas descritivas podemos iniciar nossa análise estatística.

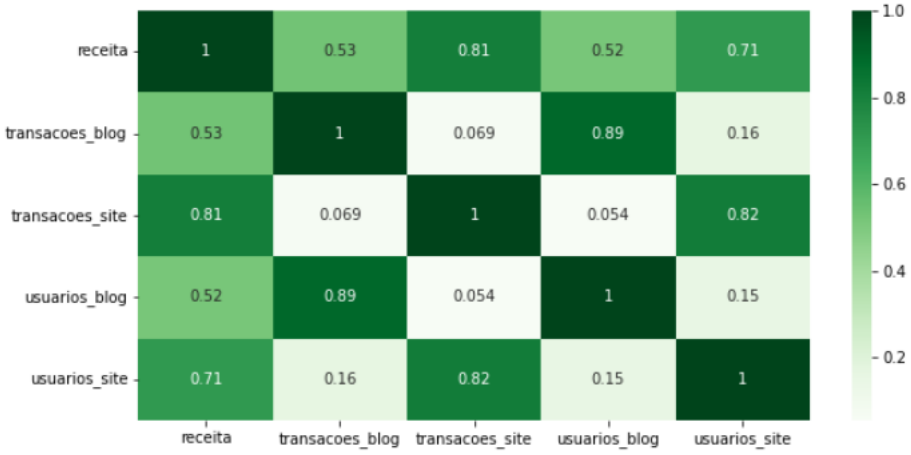
Matriz de correlação

O coeficiente de correlação é uma medida de associação linear entre duas variáveis e situa-se entre -1 e +1 sendo que -1 indica associação negativa perfeita e +1 indica associação positiva perfeita.

A análise correlacional indica a relação entre 2 variáveis lineares e os valores sempre serão entre +1 e -1. O sinal indica a direção, se a correlação é positiva ou negativa, e o tamanho da variável indica a força da correlação.

Cabe observar que, como o coeficiente é concebido a partir do ajuste linear, então a fórmula não contém informações do ajuste, ou seja, é composta apenas dos dados.

	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



Interpretando o coeficiente de correlação de Pearson

0.9 para mais ou para menos indica uma correlação muito forte.

0.7 a 0.9 positivo ou negativo indica uma correlação forte.

0.5 a 0.7 positivo ou negativo indica uma correlação moderada.

0.3 a 0.5 positivo ou negativo indica uma correlação fraca.

0 a 0.3 positivo ou negativo indica uma correlação desprezível.

Boxplot, Distribuição de frequência e Comportamento da variável Dependente Receita

Boxplot

O boxplot ou diagrama de caixa é uma ferramenta gráfica que permite visualizar a distribuição e valores discrepantes (outliers) dos dados, fornecendo assim um meio complementar para desenvolver uma perspectiva sobre o caráter dos dados. Além disso, o boxplot também é uma disposição gráfica comparativa.

As medidas de estatísticas descritivas como o mínimo, máximo, primeiro quartil, segundo quartil ou mediana e o terceiro quartil formam o boxplot.

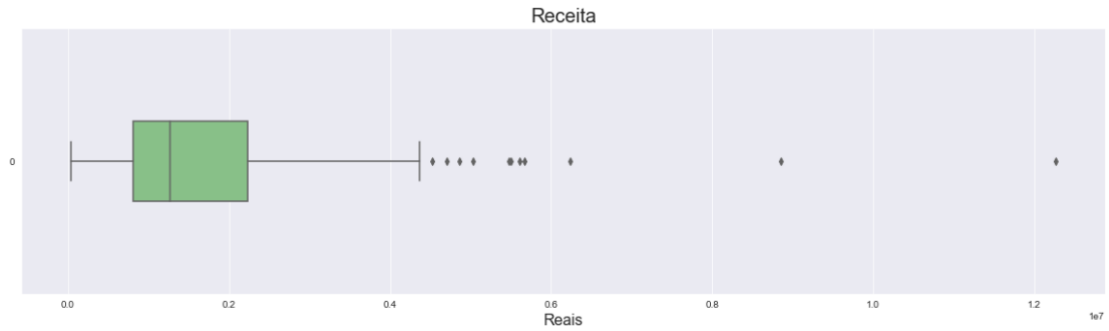
Distribuição de frequência

A **distribuição de frequência** é um arranjo de valores que uma ou mais variáveis tomam em uma amostra. Cada entrada na tabela contém a **frequência** ou a contagem de ocorrências de valores dentro de um grupo ou intervalo específico, e deste modo, a tabela resume a **distribuição** dos valores da amostra.

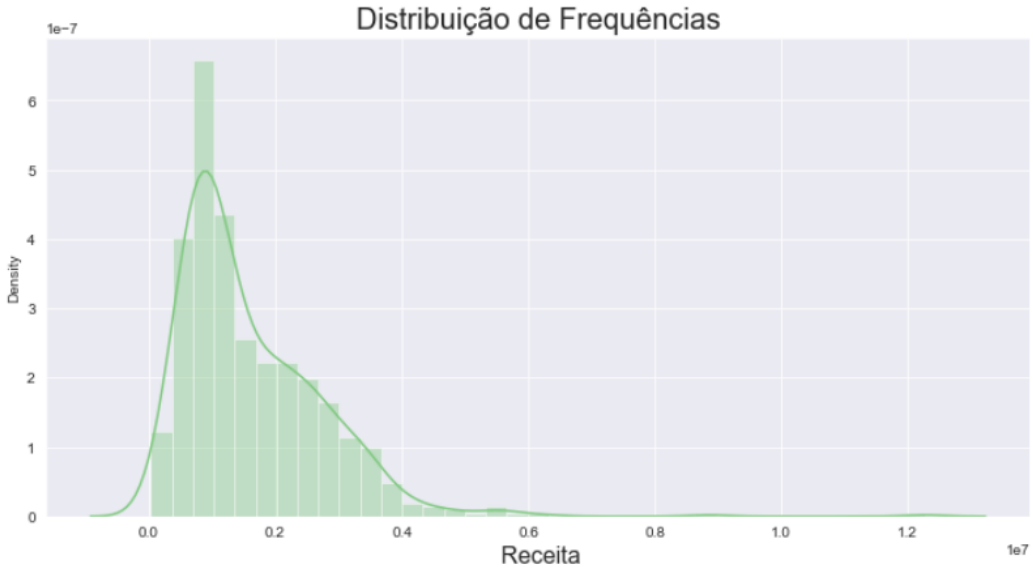
Distribuição de frequências da variável dependente (y)

Com auxílio do boxplot e da distribuição de frequência, foi identificado a assimetria a direita do dado analisando, pois existe uma maior concentração de valores na zona de valores mais reduzidos da amostra.

Análise da distribuição de frequência e outlyers da variável Receita utilizando boxplot e histograma



No boxplot analisado, há um grupo de pessoas que efetuam compras muito acima do comportamento geral da base de dados, e isso aumenta de certa forma a variabilidade, porém, para esse estudo em especial, não achei condizente retirar os dados e analisa-los a parte. Para diminuir o impacto, antes de rodar o modelo normalizei com transformação logarítmica e no final da modelagem reverti a transformação para me dar o valor real.



Distribuição de frequências da variável dependente

Com auxílio do boxplot e do histograma foi identificado a assimetria a direita do dado analisando, pois existe uma maior concentração de valores na zona de valores mais reduzidos da amostra.

Os outlayers não foram levados em consideração pois em variáveis com estas características sempre mostra muitos outlayers gerados por compras com valores mais elevados.

Dispersão Entre as Variáveis do Dataset

No **Diagrama de Dispersão**, podemos ainda **analisar** se a correlação é forte ou fraca:

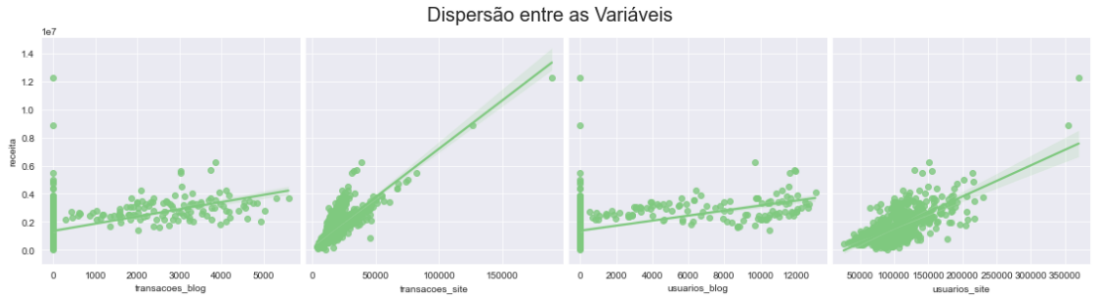
Forte: quanto maior a correlação entre as variáveis, maior será a proximidade dos pontos, ou seja, estarão menos dispersos.

Fraca: quanto menor a correlação entre as variáveis, mais dispersos estarão os pontos.

Então nos gráficos abaixo identificamos se a variável dependente e as variáveis explicativas possuem alguma relação linear.

Reta de regressão

Em estudos de distribuições bidimensionais, quando há correlação entre as variáveis, muitas vezes interessa prever o valor de uma das variáveis quando se conhece o valor correspondente da outra variável.
O processo que se vai utilizar consiste em traçar uma reta que "melhor" se ajuste (aproxime) aos pontos do diagrama de dispersão.



Após traçar a reta de regressão entendemos que como os pontos do diagrama de dispersão se formam ao longo da reta possuímos uma correlação linear.

Regressão Linear

A análise de regressão diz respeito ao estudo da dependência de uma variável (a variável dependente) em relação a uma ou mais variáveis, as variáveis explanatórias, visando estimar e/ou prever o valor médio da primeira em termos dos valores conhecidos ou fixados das segundas.

Obtendo os coeficientes de regressão

Os coeficientes de regressão $\beta 2$ e $\beta 3$ são conhecidos como coeficientes parciais de regressão ou coeficientes parciais angulares.

Um aspecto interessante do modelo log-linear, que o tornou muito utilizado nos trabalhos aplicados, é que os coeficientes angulares $\beta 2$ e $\beta 3$, medem as elasticidades de Y em relação a $X2$, $X3$, $X4$ e $X5$, isto é, a variação percentual

de Y correspondente a uma dada variação percentual (pequena) em $X2$, $X3$, $X4$ e $X5$.

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

De acordo com os coeficientes estimados no nosso modelo de Regressão Linear nas nossas variáveis nossos resultados ficariam assim:

Transações do blog → Mantendo-se o valor constante, um decréscimo de 1% no número de Transações do blog gera, em média, um decréscimo de 0.18%na receita.

Transações do site → Mantendo-se o valor constante, um decréscimo de 1% no número de transações do site gera, em média, um decréscimo de 0.06% na receita.

Usuários site → Mantendo-se o valor constante, um acréscimo de 1% no número de usuários do site gera, em média, um acréscimo de 0.13% na receita.

Usuários do blog →Mantendo-se o valor constante, um acréscimo de 1% no número de usuários do blog gera, em média, um acréscimo de 1.03% na receita.

Transformando os Dados para tentar corrigir a assimetria das nossas variáveis

Por quê?

Testes paramétricos assumem que os dados amostrais foram coletados de uma população com distribuição de probabilidade conhecida. Boa parte dos testes estatísticos assumem que os dados seguem uma distribuição normal (t de Student, intervalos de confiança etc.).

Descobrimos que a variável dependente que estamos analisando é assimétrica à direita então utilizarei uma técnica de transformação de variável para tentar corrigir esse problema e, quem sabe, estimar um modelo de regressão linear com essa base de dados, após a transformação.

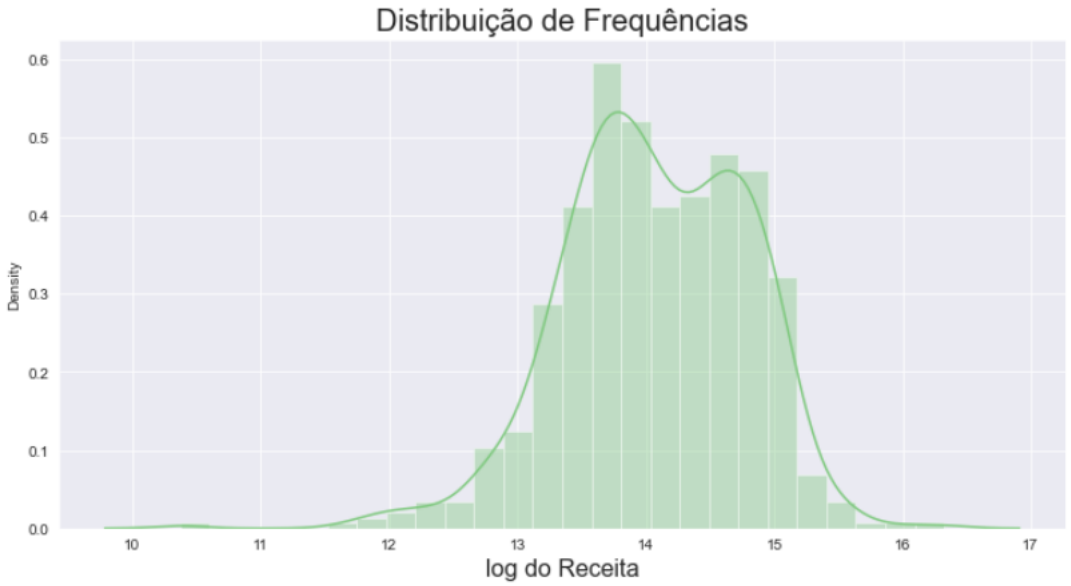
Aplicando a transformação logarítmica

A **transformação logarítmica** é frequentemente usada quando os dados têm uma distribuição distorcida positivamente e existem alguns valores grandes.

Então irei usar a transformação de log para tornar as variações mais constantes e normalizar os dados.

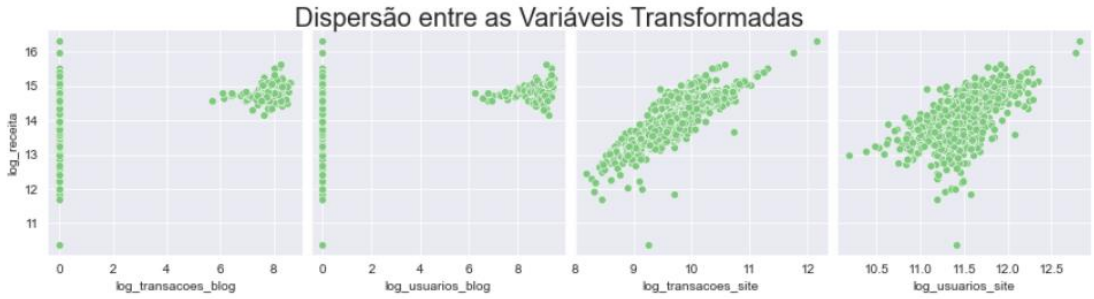
Após a transformação logarítmica a assimetria da variável receita assumiu uma forma mais normalizada.

Distribuição de frequências da variável dependente transformada Receita



Após a transformação logarítmica assimetria da variável receita assumiu uma forma mais normalizada.

Gráficos de dispersão entre as variáveis transformadas do dataset

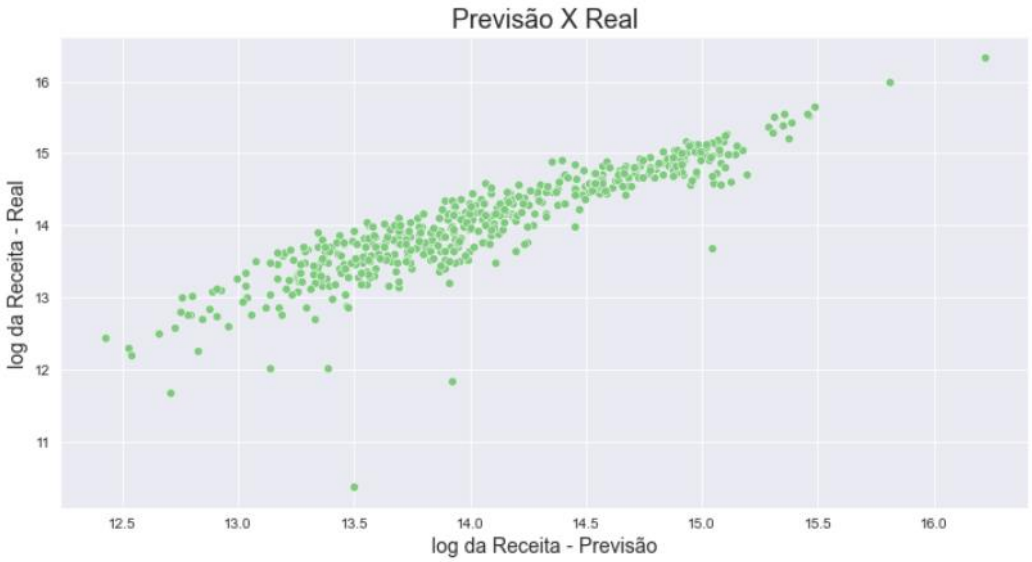


Verificando Relação Linear

Como os dados relativos ao blog são somente dos últimos 3 meses por esse motivo no gráfico de dispersão das variáveis transformadas eles estão em clusters no canto superior direito e podemos perceber que os pontos de dispersão mantem uma correlação linear.

Análises Gráficas dos Resultados do Modelo de regressão linear após a transformação logarítmica

Gráfico de dispersão entre valor estimado e valor real.



Como não temos uma grande dispersão em nosso gráfico, isso indica o quanto nossa estimativa está de certa maneira precisa.

Obtendo os resíduos

Os **resíduos** indicam a variação natural dos dados, um fator aleatório (ou não) que o modelo não capturou. Se as pressuposições do modelo **são** violadas, a análise será levada a resultados duvidosos e não confiáveis para inferência

Plot de distribuição de frequências dos resíduos

Como resultado, temos uma curva agradável que indica dados bem comportados, o que favorece um modelo bem estimado.

Observações

É importante frisar que nosso modelo de regressão não é um modelo a ser usado comercialmente, pois precisaria de mais variáveis e elementos para operar de maneira funcional. Contudo, conseguimos antever qual seria o comportamento adequado para que um modelo estimativo nos fornecesse informações seguras, e para este fim, nosso projeto cabe perfeitamente.

Teste de hipóteses

Testes estatísticos são regras de decisão que permitem avaliar a razoabilidade das hipóteses feitas sobre os parâmetros populacionais e aceitá-las ou rejeitá-las como provavelmente verdadeiras ou falsas tendo como base uma amostra

Testes realizados:

Teste de normalidade

O teste de normalidade testa a hipótese nula $H0$H0 de que a amostra é proveniente de uma distribuição normal.

Função do python utilizada foi a **normaltest** que testa a hipótese nula $H0$ de que a amostra é proveniente de uma distribuição normal.

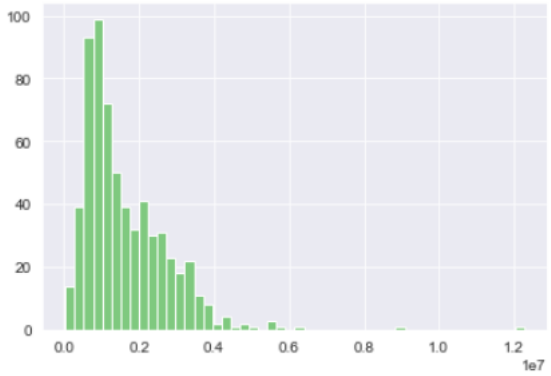
Nível de significância

O nível de significância utilizado em todos os testes foi de **0.05%.**

Analisando o teste e os gráficos após o teste de normalidade.

P_value após o teste = 1.1361263738507747e-60

variável receita



Critério do valor p

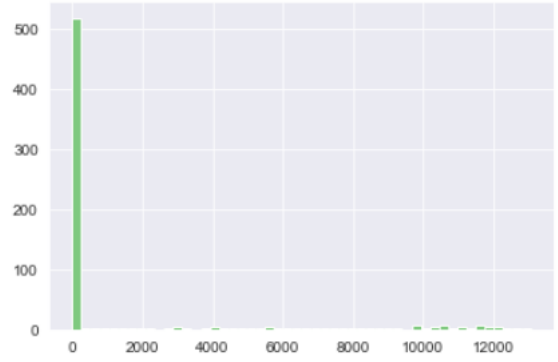
Rejeitar $H0$ se o valor p≤0,05

Analisando o gráfico

Em relação à assimetria, temos vários valores positivos extremos, o que sugeriria assimetria positiva.

Após o **normaltest** foi retornado um **P_value** de **1.1361263738507747e-60** e como nosso P_value é maior que nosso nível significância de **0.5%** o H0 é rejeitado e a amostra não é proveniente de uma distribuição normal.

Variável Usuários do blog



Critério do valor p

Rejeitar $H0$ se o valor p≤0,05

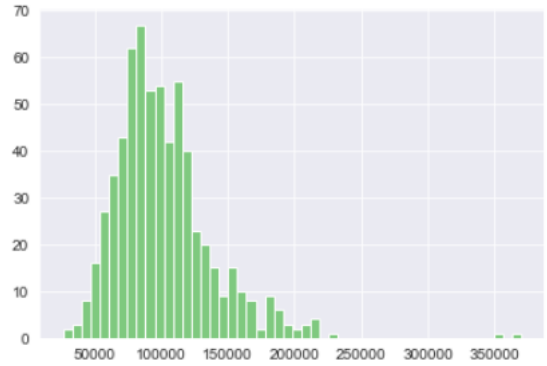
Analisando o gráfico

Em relação à assimetria, temos vários valores positivos extremos, o que sugeriria assimetria positiva.

Após o **normaltest** foi retornado um **P_value** de **1.1361263738507747e-60**

e como nosso P_value é maior que nosso nível significância de **0.5%** o H0 é rejeitado e a amostra não é proveniente de uma distribuição normal.

Variável Usuários do site



Critério do valor p

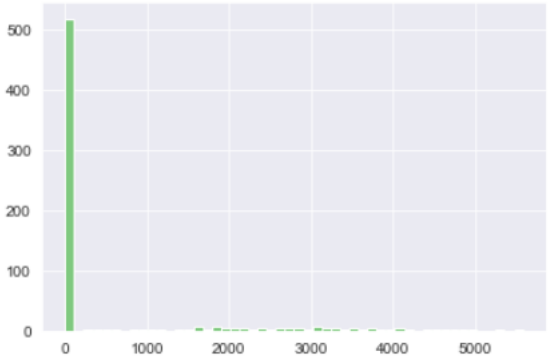
Rejeitar $H0$ se o valor p≤0,05

Analisando o gráfico

Em relação à assimetria, temos vários valores positivos extremos, o que sugeriria assimetria positiva.

Após o **normaltest** foi retornado um **P_value** de **2.3947832487712726e-59**
e como nosso P_value é maior que nosso nível significância de **0.5%** o H0 é rejeitado e a amostra não é proveniente de uma distribuição normal.

Variável transações do blog



Critério do valor p

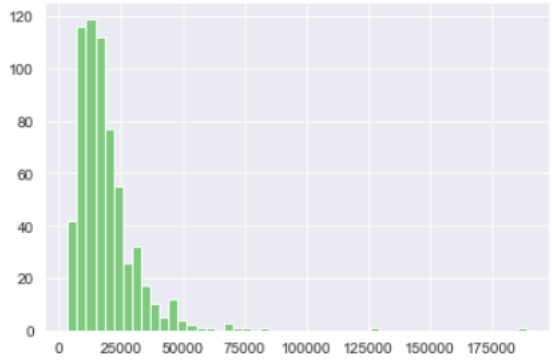
Rejeitar $H0$ se o valor p≤0,05

Analisando o gráfico

Em relação à assimetria, temos vários valores positivos extremos, o que sugeriria assimetria positiva.

Após o **normaltest** foi retornado um **P_value** de **2.896413130631297e-60**
e como nosso P_value é maior que nosso nível significância de **0.5%** o H0 é rejeitado e a amostra não é proveniente de uma distribuição normal.

Variável transações do site



Critério do valor p

Rejeitar $H0$ se o valor p≤0,05

Analisando o gráfico

Em relação à assimetria, temos vários valores positivos extremos, o que sugeriria assimetria positiva.

Após o **normaltest** foi retornado um **P_value** de **4.8317020805360967e-147**
e como nosso P_value é maior que nosso nível significância de **0.5%** o H0 é rejeitado e a amostra não é proveniente de uma distribuição normal.

TESTES NÃO PARAMÉTRICOS

Os testes não paramétricos, também conhecidos como testes de distribuição gratuita, são aqueles baseados em certas hipóteses, mas que não possuem uma organização normal. Geralmente, contêm resultados estatísticos provenientes de suas ordenações.

Testes não paramétricos têm algumas limitações, entre eles, está que não são fortes suficientemente fortes quando uma hipótese normal é preenchida. Isso pode fazer com que não seja rejeitado, mesmo que seja falso. Outra de suas limitações é que necessitam que a hipótese seja alterada quando o teste não corresponde à questão do procedimento se a amostra não for proporcional.

Em nosso dataset temos dados sobre o nosso site de vendas e criamos o blog com conteúdo há aproximadamente 3 meses atrás, e queremos entender qual influência de um blog da nossa marca.

Então selecionarei duas amostras em nosso dataset. Com o objetivo de comprovar tal influência sobre a receita da empresa **testarei a igualdade das médias** entra estas duas amostras com um nível de **significância de 5%**.

Testes realizados:

Teste de Mann-Whitney

Mann-Whitney é um teste não paramétrico utilizado para verificar se duas amostras independentes foram selecionadas a partir de populações que têm a mesma média. Por ser um teste não paramétrico, Mann-Whitney torna-se uma alternativa ao teste paramétrico de comparação de médias.

Função do python utilizada foi a **mannwhitneyu** que testa a hipótese nula $H0$ de que a amostra é proveniente de uma distribuição normal.

Nível de significância

O nível de significância utilizada no testes foi de **0.05%.**

Critério do valor p

Rejeitar $H0$ se o valor $p \leq \alpha$

Testando a variável usuários do blog e usuários do site

Critério do valor p

Rejeitar $H0$ se o valor $p \leq \alpha$

Após o **mannwhitneyu** foi retornado um P_value de **2.391193425606197e-225**, e como nosso P_value é maior que nosso nível significância de 0.5% o H0 é rej eitado De acordo com os resultados aceitamos hipótese de que existe uma influ ência do blog em nossa marca, isto é, concluímos que a média dos usuários do blog e usuários do site possuem influência sobre nossa marca.

Testando a variável transações do blog x transações do site

Critério do valor p

Rejeitar $H0$ se o valor $p \leq \alpha$

Após o **mannwhitneyu** foi retornado um P_value de **4.689855769706814e-225**, e como nosso P_value é maior que nosso nível significância de 0.5% o H0 é rejeitado De acordo com os resultados aceitamos hipótese de que existe uma influência do blog em nossa marca, isto é, concluímos que a média das transações do blog e transações do site possuem influência sobre nossa marca.

Testando a variável usuários do blog e receita

Critério do valor p

Rejeitar $H0$ se o valor $p \leq \alpha$

Após o **mannwhitneyu** foi retornado um P_value de **2.391204760924378e-225,** e como nosso P_value é maior que nosso nível significância de 0.5% o H0 é rej eitado De acordo com os resultados aceitamos hipótese de que existe uma influ ência do blog em nossa marca, isto é, concluímos que a média de usuários do b log e receita possuem influência sobre nossa marca.

Testando a variável transações do blog e receita

Critério do valor p

Rejeitar $H0$ se o valor $p \leq \alpha$

Após o **mannwhitneyu** foi retornado um P_value de **2.391200982478832e-225** , e como nosso P_value é maior que nosso nível significância de 0.5% o H0 é r ejeitado De acordo com os resultados aceitamos hipótese de que existe uma inf luência do blog em nossa marca, isto é, concluímos que a média de usuários do blog e receita possuem influência sobre nossa marca.