



MINERAÇÃO DE DADOS



Professor: Murilo Varges



Aluno: José Cenci



Aluno: Raul Dantas

CONTEÚDO:

- INTRODUÇÃO
- PRÉ-PROCESSAMENTO
- ANÁLISE EXPLORATÓRIA
- ANÁLISE DESCRITIVA
- RESULTADOS

INTRODUÇÃO



Esta Foto de Autor Desconhecido está licenciado em CC BY-SA

A classificação de preços de celulares é uma aplicação prática da mineração de dados que pode fornecer insights valiosos para consumidores, fabricantes e vendedores, ajudando a entender as tendências do mercado e as preferências dos usuários.

CONJUNTO DE DADOS

- A base de dados escolhida foi a "mobile-price-classification", como o próprio nome diz, o objetivo principal é classificar o preço dos celulares de acordo com características de telefones celulares.
- A classificação do celular entra de 0 a 3, no caso seria a única variável dependente: price_range
- 0 Barato
- 1 Médio
- 2 Caro
- 3 Muito caro.

COLUNAS

battery_power: Potência da bateria do dispositivo (mAh).

blue: Disponibilidade de Bluetooth.

clock_speed: Velocidade de relógio do processador do dispositivo (GHz).

dual_sim: Suporte a dual SIM).

fc: Megapixels da câmera frontal.

four_g: Disponibilidade de 4G.

int_memory: Memória interna do dispositivo (GB).

m_dep: Profundidade do dispositivo móvel em cm.

mobile_wt: Peso do dispositivo móvel (em gramas).

n cores: Número de cores do processador.

pc: Megapixels da câmera principal (traseira).

px_height: Altura da resolução de pixel da tela.

px_width: Largura da resolução de pixel da tela.

ram: Memória RAM do dispositivo (MB).

sc_h: Altura da tela do dispositivo em cm.

sc_w: Largura da tela do dispositivo em cm.

talk_time: Tempo máximo de conversação em uma única carga de bateria (horas).

three_g: Disponibilidade de 3G.

touch_screen: Presença de tela sensível ao toque.

wifi: Disponibilidade de WiFi.

DESCRIÇÃO DOS PROBLEMAS DA BASE E ESTRATÉGIAS ADOTADAS

train.info()

- O dataset de treino+teste tem 21 colunas e 2000 linhas.
- 20 variáveis independentes e 1 variável dependente.
- clock_speed e m_dep são do tipo "float". Todas as outras são "int".
- Não existem valores faltantes.
- Temos 8 variáveis categóricas: n_cores, price_range, blue,
 dual_sim, four_g, three_g, touch_screen, wifi
- Temos 13 variáveis numéricas: battery_power, clock_speed,
 fc, int_memory, m_dep, mobile_wt, pc, px_height, px_width,
 ram, talk_time, sc_h, sc_w

Data	columns (total	21 columns):									
#	Column	Non-Null Count Dtype									
0	battery_power	2000 non-null int64									
1	blue	2000 non-null int64									
2	clock_speed	2000 non-null float64									
3	dual_sim	2000 non-null int64									
4	fc	2000 non-null int64									
5	four_g	2000 non-null int64									
6	int_memory	2000 non-null int64									
7	m_dep	2000 non-null float64									
8	mobile_wt	2000 non-null int64									
9	n_cores	2000 non-null int64									
10	pc	2000 non-null int64									
11	px_height	2000 non-null int64									
12	px_width	2000 non-null int64									
13	ram	2000 non-null int64									
14	sc_h	2000 non-null int64									
15	sc_w	2000 non-null int64									
16	talk_time	2000 non-null int64									
17	three_g	2000 non-null int64									
18	touch_screen	2000 non-null int64									
19	wifi	2000 non-null int64									
20	price_range	2000 non-null int64									
dtype	es: float64(2),	int64(19)									

REDUÇÃO, DISCRETIZAÇÃO E TRANSFORMAÇÃO DE DADOS

- Não possui redundância de dados.
- Não reduzimos a base pois já temos 2000 linhas somando as bases de treino e teste.

ANÁLISE EXPLORATÓRIA

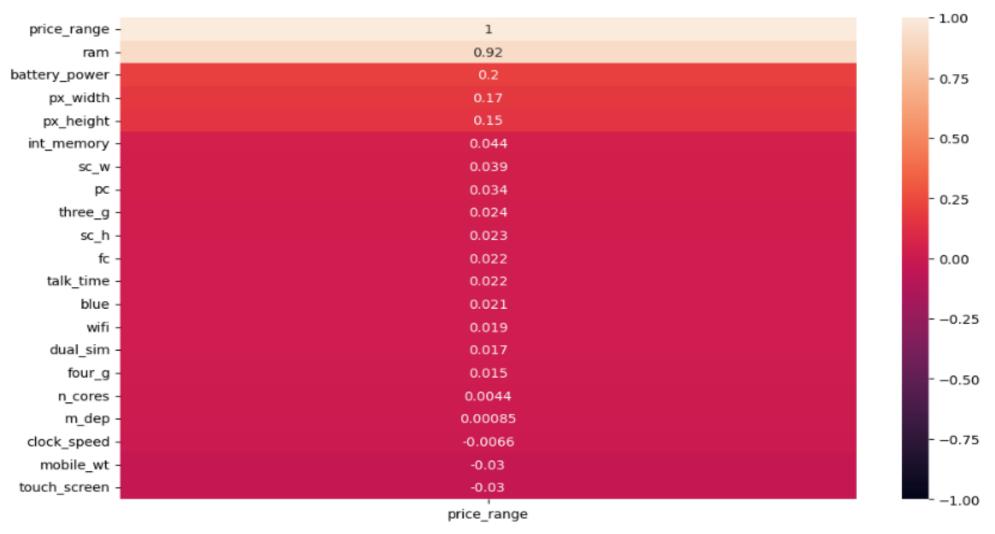
- Matriz de correlação.
- Análise componentes principais PCA.
- Histogramas.
- Distribuições por faixa de preço.

MATRIZ DE CORRELAÇÃO

		Correlation Matrix																			
attery_power -	1.00	0.01	0.01	-0.04	0.03	0.02	-0.00	0.03	0.00	-0.03	0.03	0.01	-0.01	-0.00	-0.03	-0.02	0.05	0.01	-0.01	-0.01	0.20
blue -	0.01	1.00	0.02	0.04	0.00	0.01	0.04	0.00	-0.01	0.04	-0.01	-0.01	-0.04	0.03	-0.00	0.00	0.01	-0.03	0.01	-0.02	0.02
dock_speed -	0.01	0.02	1.00	-0.00	-0.00	-0.04	0.01	-0.01	0.01	-0.01	-0.01	-0.01	-0.01	0.00	-0.03	-0.01	-0.01	-0.05	0.02	-0.02	-0.01
dual_sim -	-0.04	0.04	-0.00	1.00	-0.03	0.00	-0.02	-0.02	-0.01	-0.02	-0.02	-0.02	0.01	0.04	-0.01	-0.02	-0.04	-0.01	-0.02	0.02	0.02
fc -	0.03	0.00	-0.00	-0.03	1.00	-0.02	-0.03	-0.00	0.02	-0.01	0.64	-0.01	-0.01	0.02	-0.01	-0.01	-0.01	0.00	-0.01	0.02	0.02
four_g -	0.02	0.01	-0.04	0.00	-0.02	1.00	0.01	-0.00	-0.02	-0.03	-0.01	-0.02	0.01	0.01	0.03	0.04	-0.05	0.58	0.02	-0.02	0.01
int_memory -	-0.00	0.04	0.01	-0.02	-0.03	0.01	1.00	0.01	-0.03	-0.03	-0.03	0.01	-0.01	0.03	0.04	0.01	-0.00	-0.01	-0.03	0.01	0.04
m_dep -	0.03	0.00	-0.01	-0.02	-0.00	-0.00	0.01	1.00	0.02	-0.00	0.03	0.03	0.02	-0.01	-0.03	-0.02	0.02	-0.01	-0.00	-0.03	0.00
mobile_wt -	0.00	-0.01	0.01	-0.01	0.02	-0.02	-0.03	0.02	1.00	-0.02	0.02	0.00	0.00	-0.00	-0.03	-0.02	0.01	0.00	-0.01	-0.00	-0.03
n_cores -	-0.03	0.04	-0.01	-0.02	-0.01	-0.03	-0.03	-0.00	-0.02	1.00	-0.00	-0.01	0.02	0.00	-0.00	0.03	0.01	-0.01	0.02	-0.01	0.00
pc -	0.03	-0.01	-0.01	-0.02	0.64	-0.01	-0.03	0.03	0.02	-0.00	1.00	-0.02	0.00	0.03	0.00	-0.02	0.01	-0.00	-0.01	0.01	0.03
px_height -	0.01	-0.01	-0.01	-0.02	-0.01	-0.02	0.01	0.03	0.00	-0.01	-0.02	1.00	0.51	-0.02	0.06	0.04	-0.01	-0.03	0.02	0.05	0.15
px_width -	-0.01	-0.04	-0.01	0.01	-0.01	0.01	-0.01	0.02	0.00	0.02	0.00	0.51	1.00	0.00	0.02	0.03	0.01	0.00	-0.00	0.03	0.17
ram -	-0.00	0.03	0.00	0.04	0.02	0.01	0.03	-0.01	-0.00	0.00	0.03	-0.02	0.00	1.00	0.02	0.04	0.01	0.02	-0.03	0.02	0.92
sc_h -	-0.03	-0.00	-0.03	-0.01	-0.01	0.03	0.04	-0.03	-0.03	-0.00	0.00	0.06	0.02	0.02	1.00	0.51	-0.02	0.01	-0.02	0.03	0.02
sc_w -	-0.02	0.00	-0.01	-0.02	-0.01	0.04	0.01	-0.02	-0.02	0.03	-0.02	0.04	0.03	0.04	0.51	1.00	-0.02	0.03	0.01	0.04	0.04
talk_time -	0.05	0.01	-0.01	-0.04	-0.01	-0.05	-0.00	0.02	0.01	0.01	0.01	-0.01	0.01	0.01	-0.02	-0.02	1.00	-0.04	0.02	-0.03	0.02
three_g -	0.01	-0.03	-0.05	-0.01	0.00	0.58	-0.01	-0.01	0.00	-0.01	-0.00	-0.03	0.00	0.02	0.01	0.03	-0.04	1.00	0.01	0.00	0.02
buch_screen -	-0.01	0.01	0.02	-0.02	-0.01	0.02	-0.03	-0.00	-0.01	0.02	-0.01	0.02	-0.00	-0.03	-0.02	0.01	0.02	0.01	1.00	0.01	-0.03
wifi -	-0.01	-0.02	-0.02	0.02	0.02	-0.02	0.01	-0.03	-0.00	-0.01	0.01	0.05	0.03	0.02	0.03	0.04	-0.03	0.00	0.01	1.00	0.02
price_range -	0.20	0.02	-0.01	0.02	0.02	0.01	0.04	0.00	-0.03	0.00	0.03	0.15	0.17	0.92	0.02	0.04	0.02	0.02	-0.03	0.02	1.00
	battery_power -	- pine	dock_speed -	- dual_sim	ψ.	four_g -	int_memory -	m_dep -	mobile wt -	- cores	Ŗ	px_height -	px_width -	- MBI	å.	- M _ 75	talk_time -	three g -	- touch_screen	- wifi	price_range -

CORRELAÇÃO DAS COLUNAS COM PRICE_RANGE

Features Correlating with Price Range

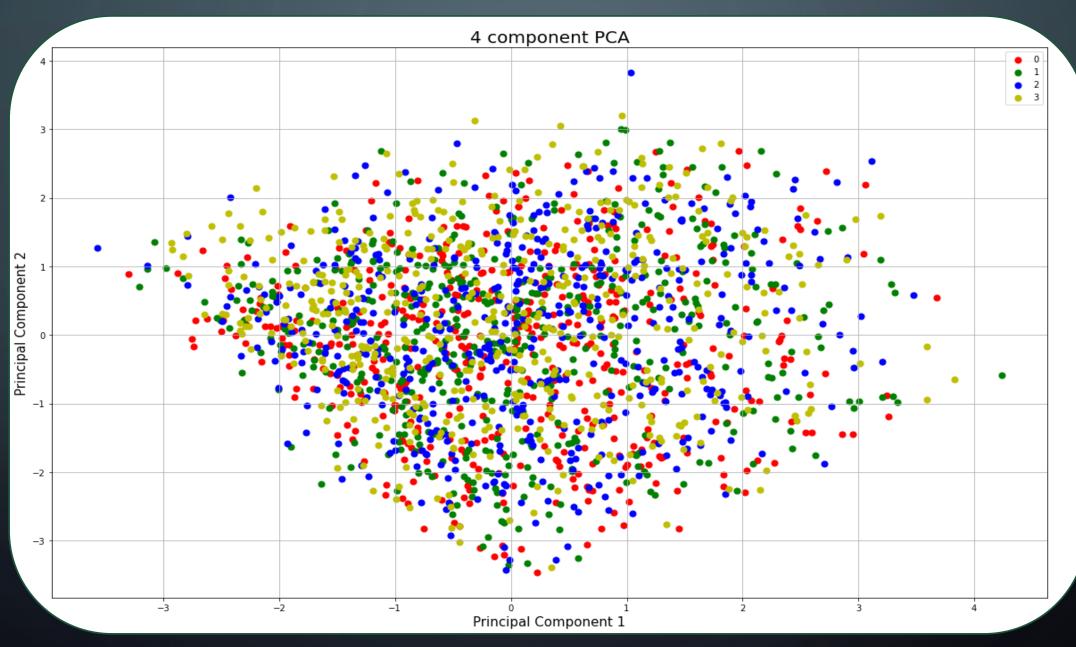


ANÁLISE EXPLORATÓRIA

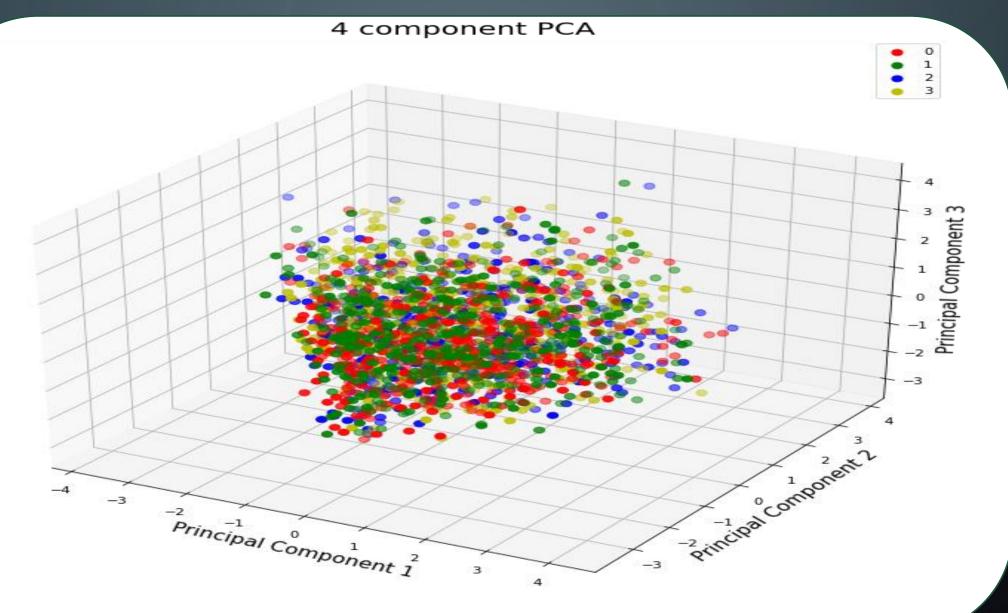
Ao analisarmos a matriz de correlação, foi possível concluir:

- Forte correlação entre ram e price_range.
- price_range tem um valor de correlação baixo com o restante dos recursos, mas isso não pode ser usado como critério para remover esses recursos, pois a correlação de Pearson expressa apenas a relação linear entre duas variáveis.
- Podemos ver uma correlação moderada entre 4G e 3G, fc e pc, px_height e px_width, sc_h e sc_w.

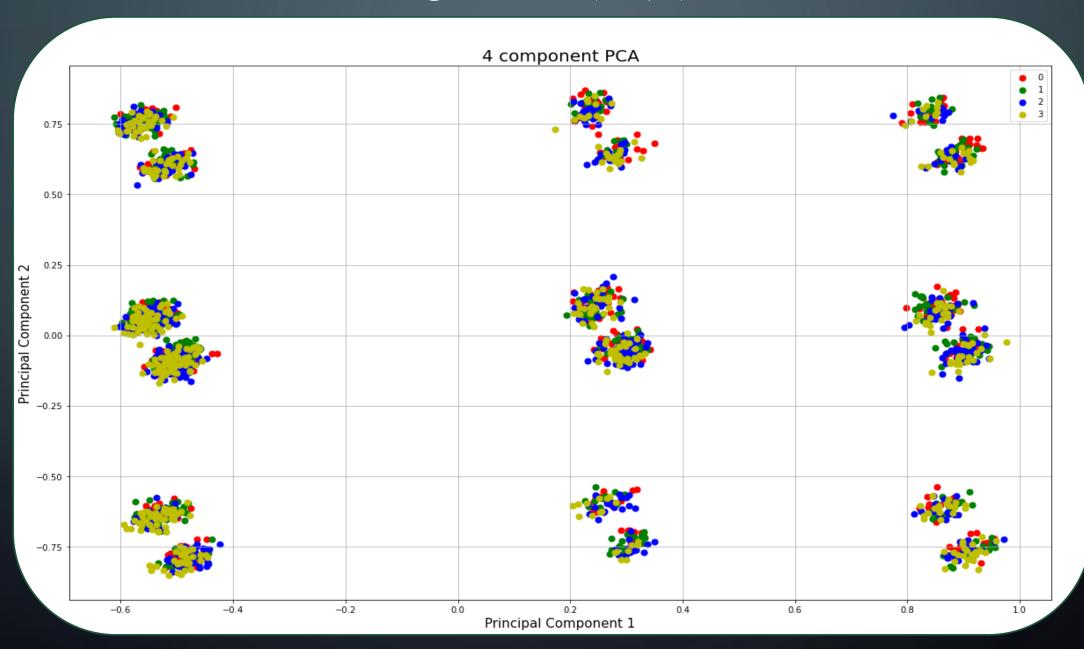
PCA 2D – Z SCORE



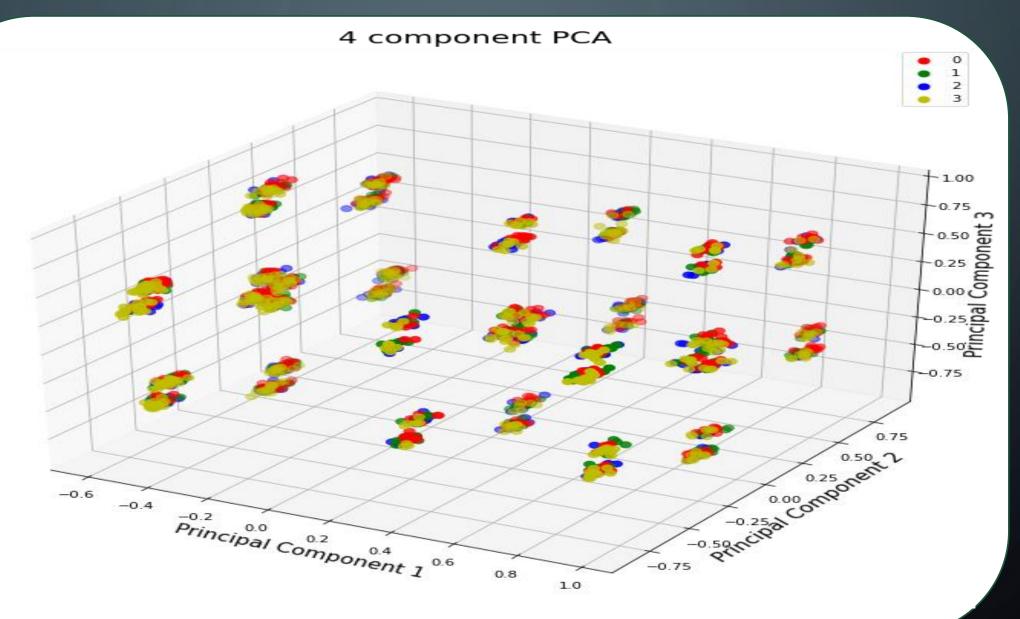
PCA 3D – Z SCORE



PCA 2D – MIN MAX



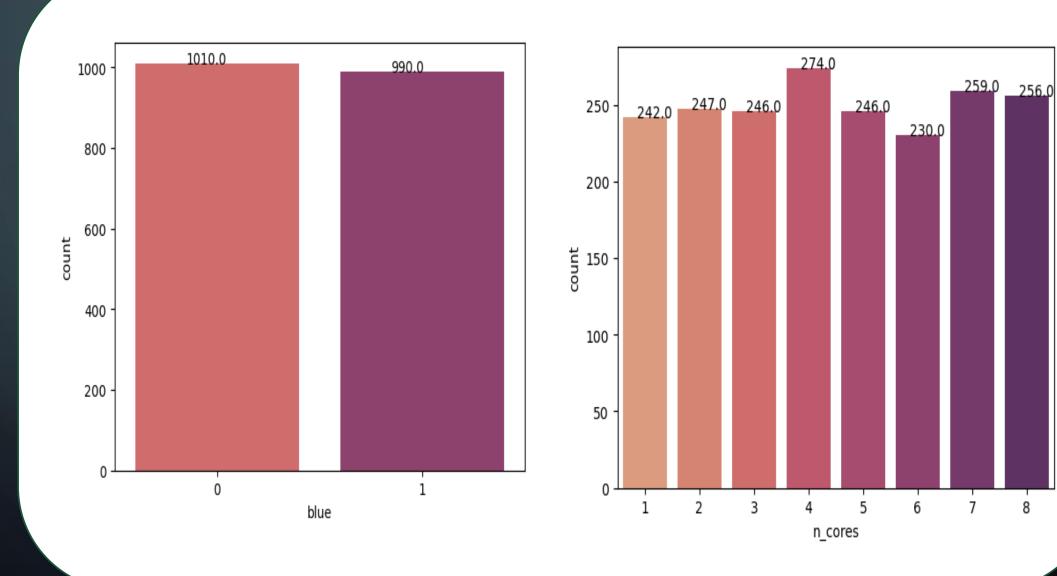
PCA 3D - MIN MAX

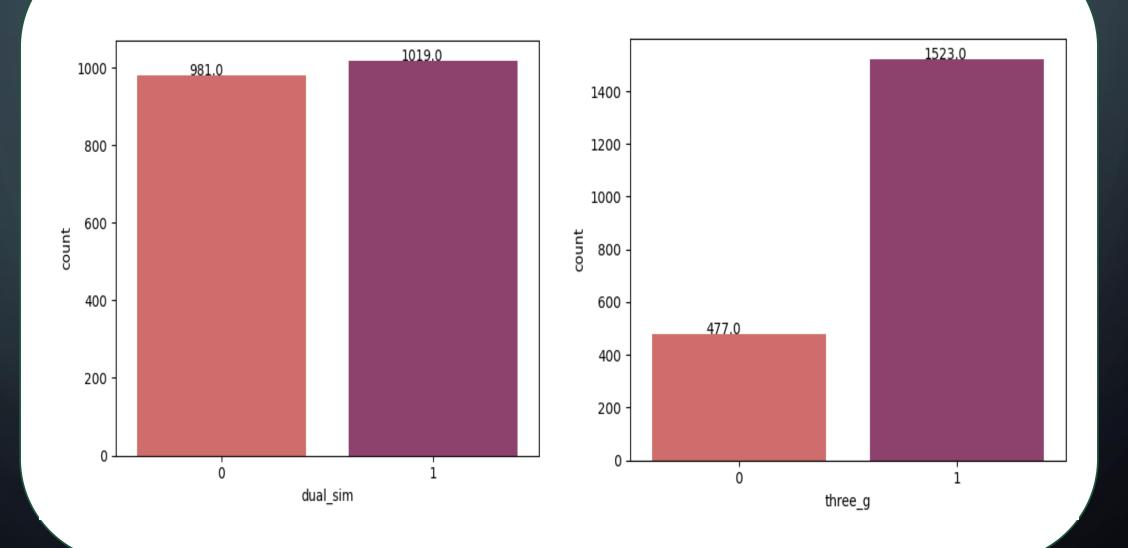


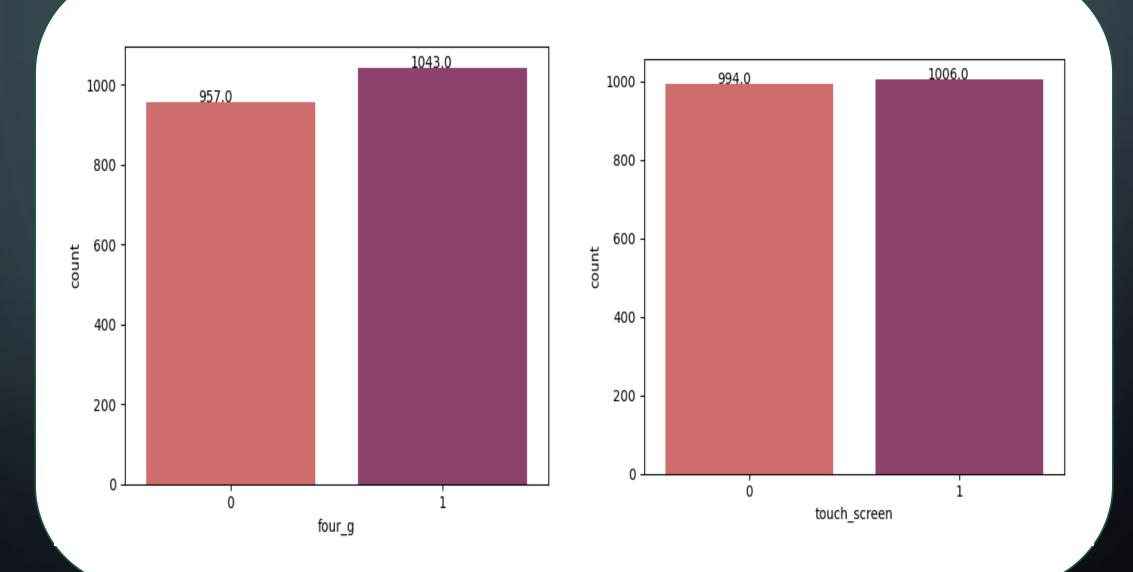
ANÁLISE EXPLORATÓRIA

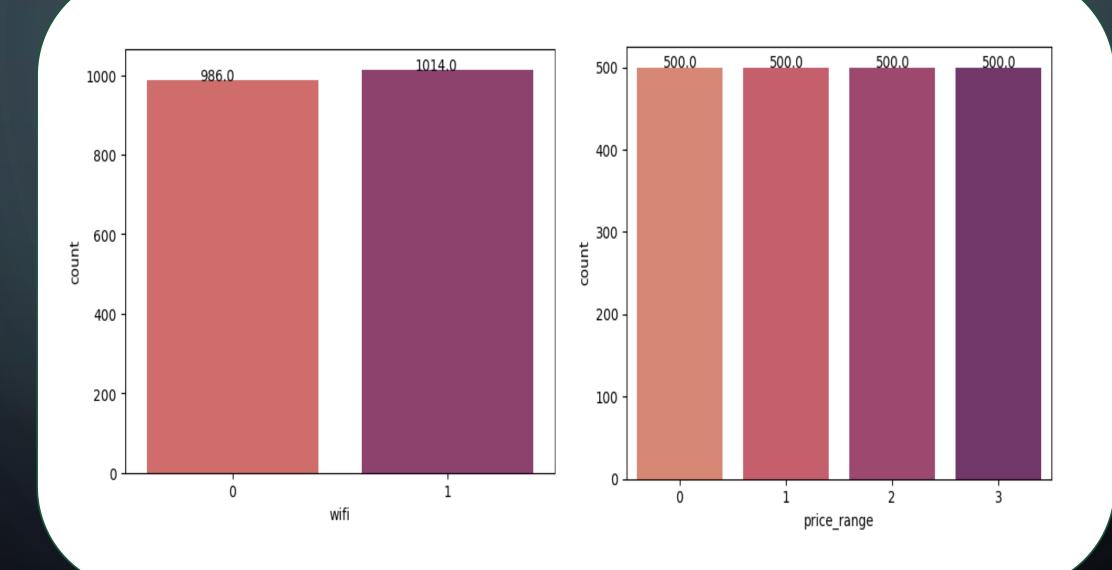
Ao analisarmos os principais componentes, foi possível concluir:

 Concluímos que a normalização z-score seria mais apropriada para a análise de componentes principais (PCA). Devido que a normalização MinMax fazem os dados ficarem muito agrupados.

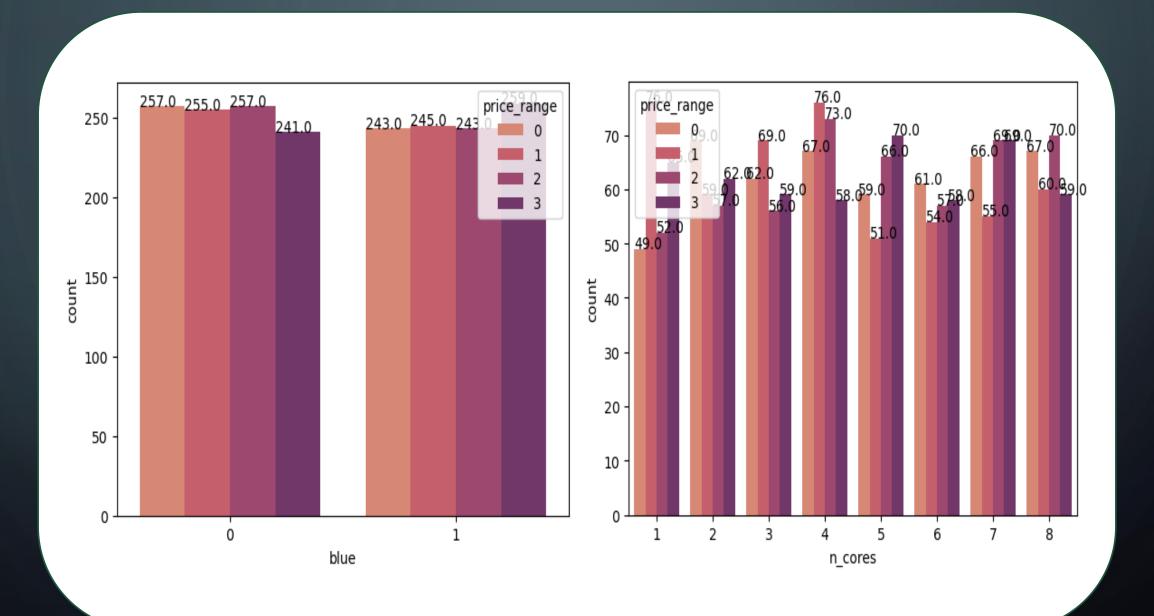




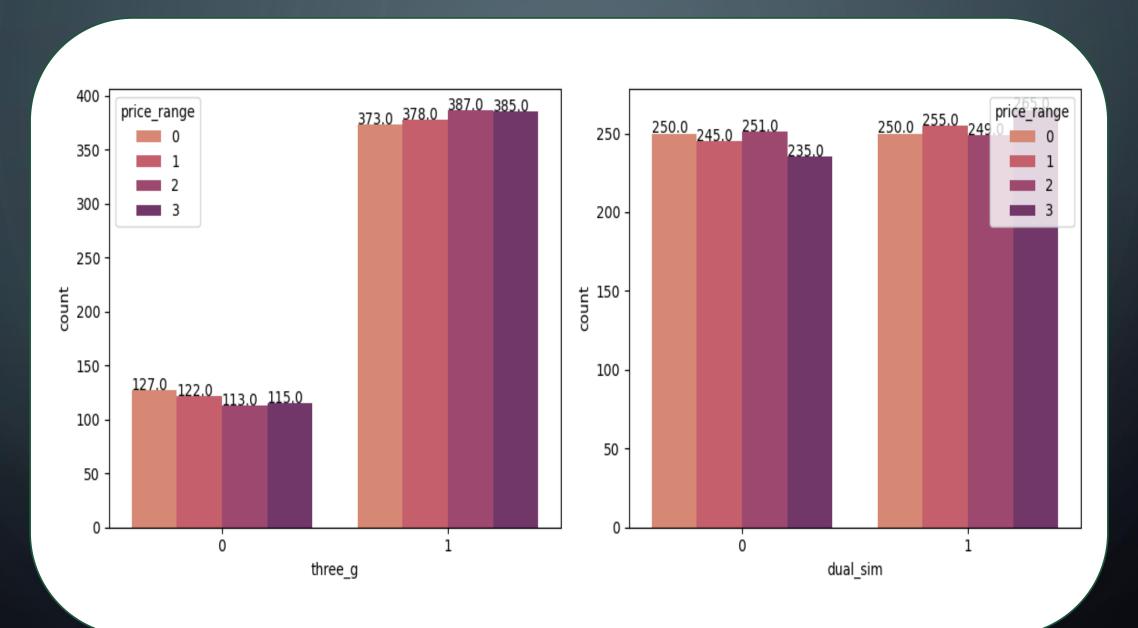




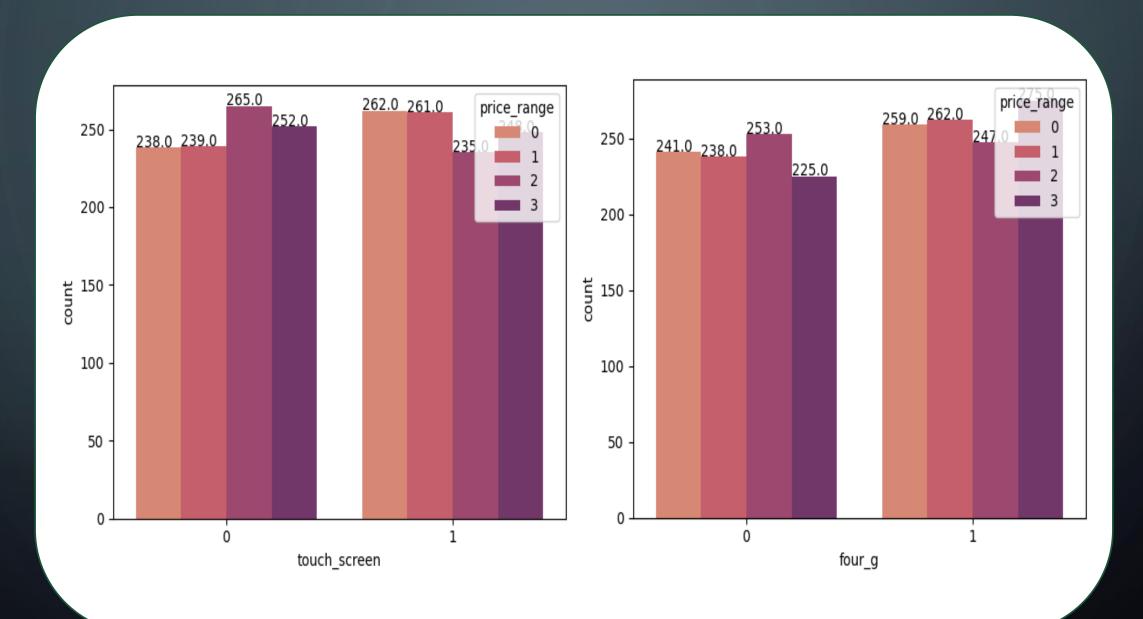
ANÁLISE DESCRITIVA – DADOS CATEGÓRICOS – RELACIONADO COM PRICE_RANGE



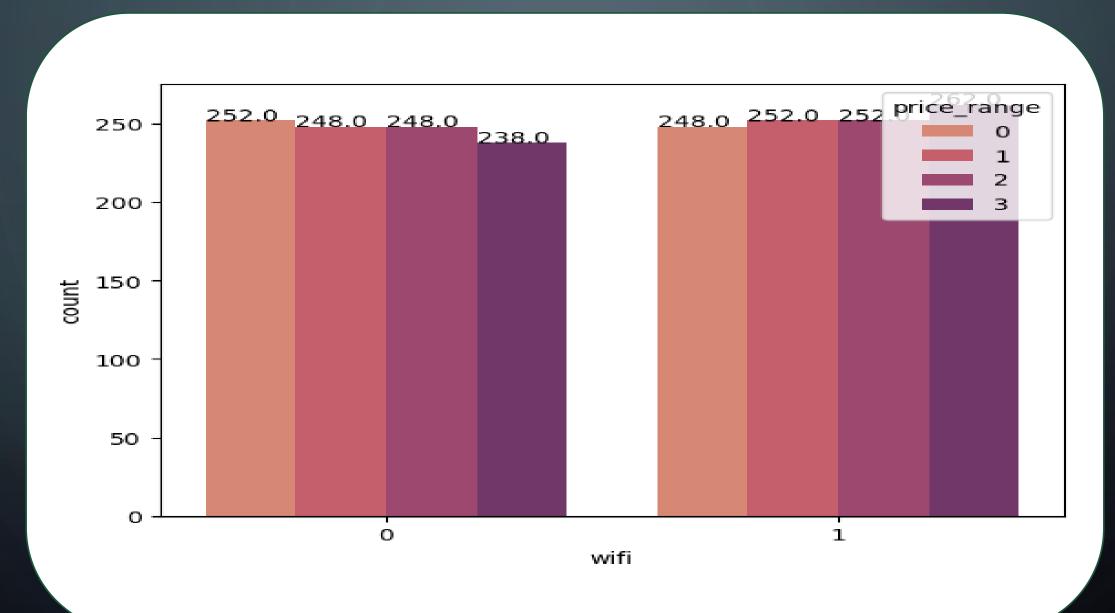
ANÁLISE DESCRITIVA – DADOS CATEGÓRICOS – RELACIONADO COM PRICE_RANGE



ANÁLISE DESCRITIVA — DADOS CATEGÓRICOS — RELACIONADO COM PRICE_RANGE

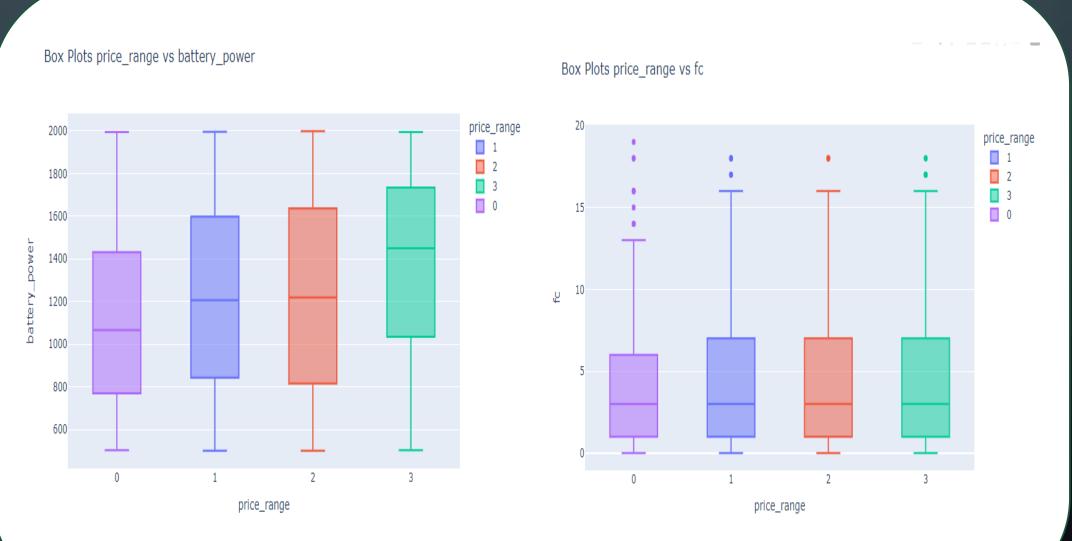


ANÁLISE DESCRITIVA – DADOS CATEGÓRICOS – RELACIONADO COM PRICE_RANGE



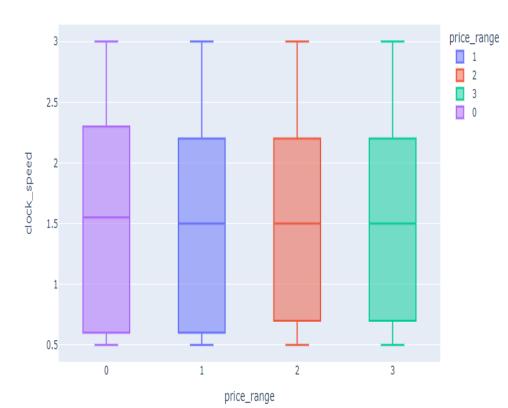
ANÁLISE DESCRITIVA

- Vemos quase a mesma frequência em termos de ter ou não Bluetooth, 4G, dois cartões SIM, tela "touchscreen" e quantidade de núcleos de processamento utilizados.
- Vemos uma grande diferença em termos de ter ou não 3G.

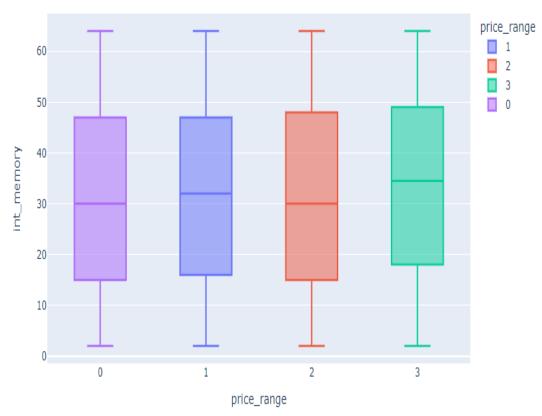


Celulares com maiores baterias, são mais caros.

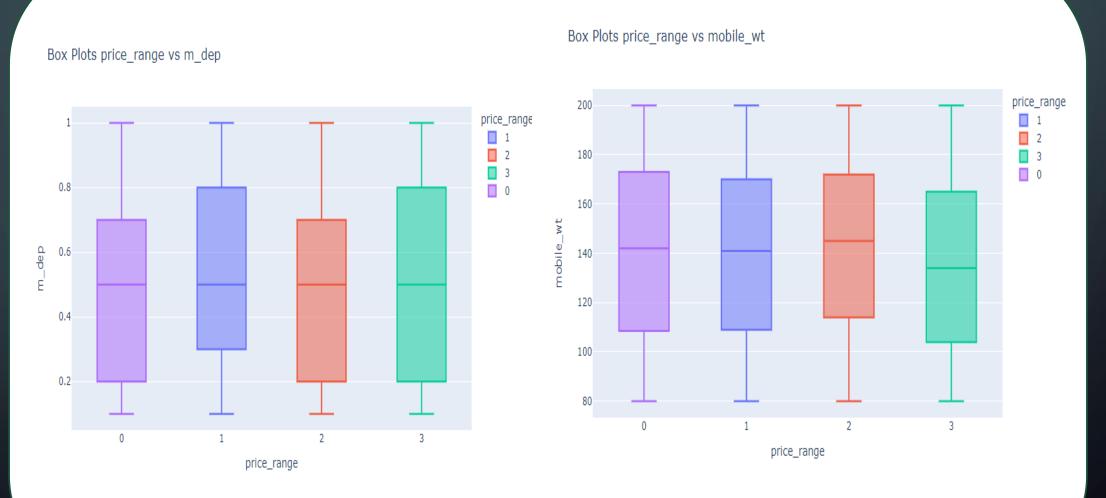




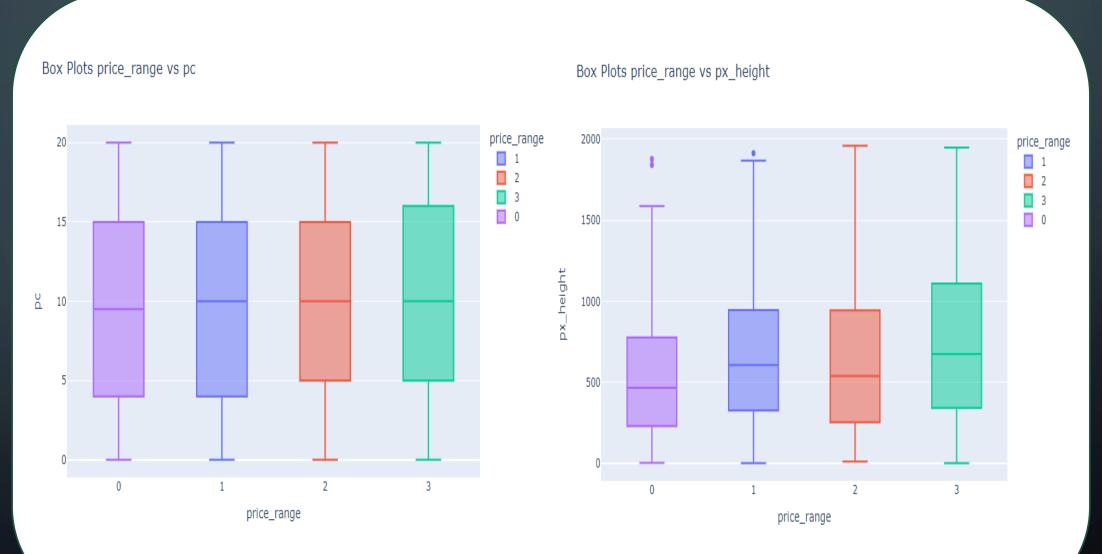
Box Plots price_range vs int_memory



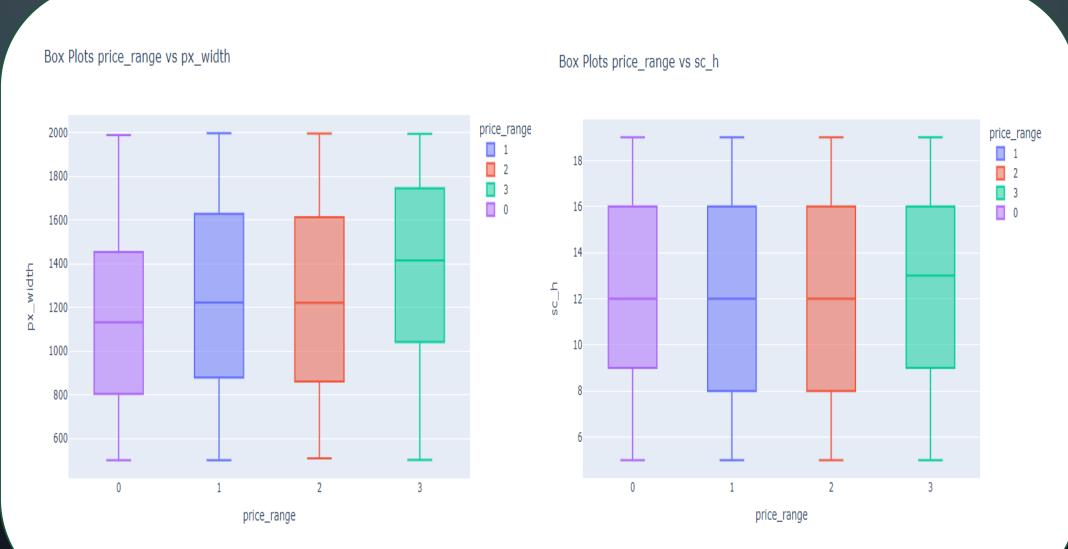
Memória interna tem uma leve relação com o preço.



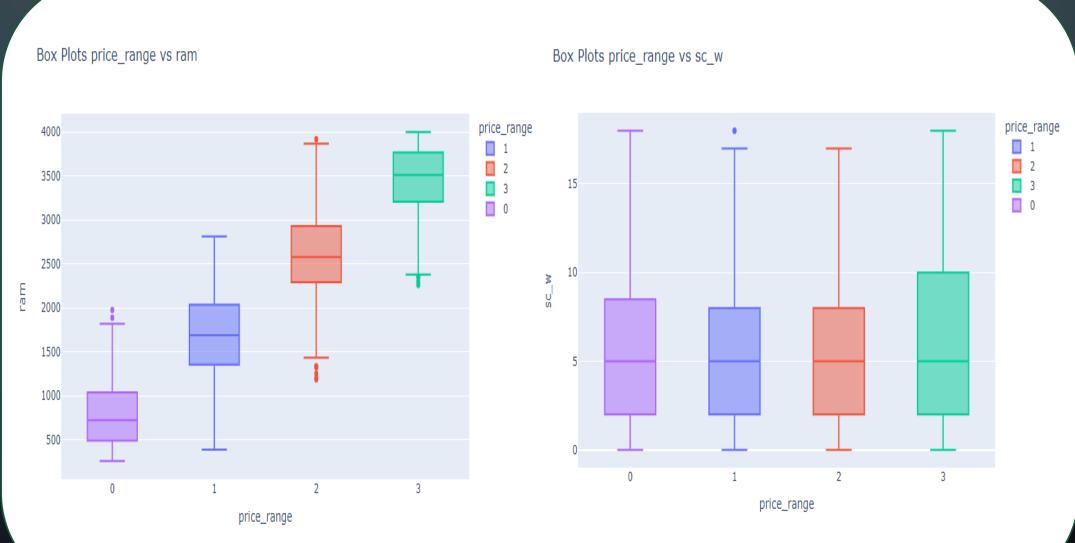
Celulares mais caros, pela análise, são mais leves.



Quanto maior altura da resolução de pixel da tela maior o preço.



Quanto maior largura da resolução de pixel da tela maior o preço.

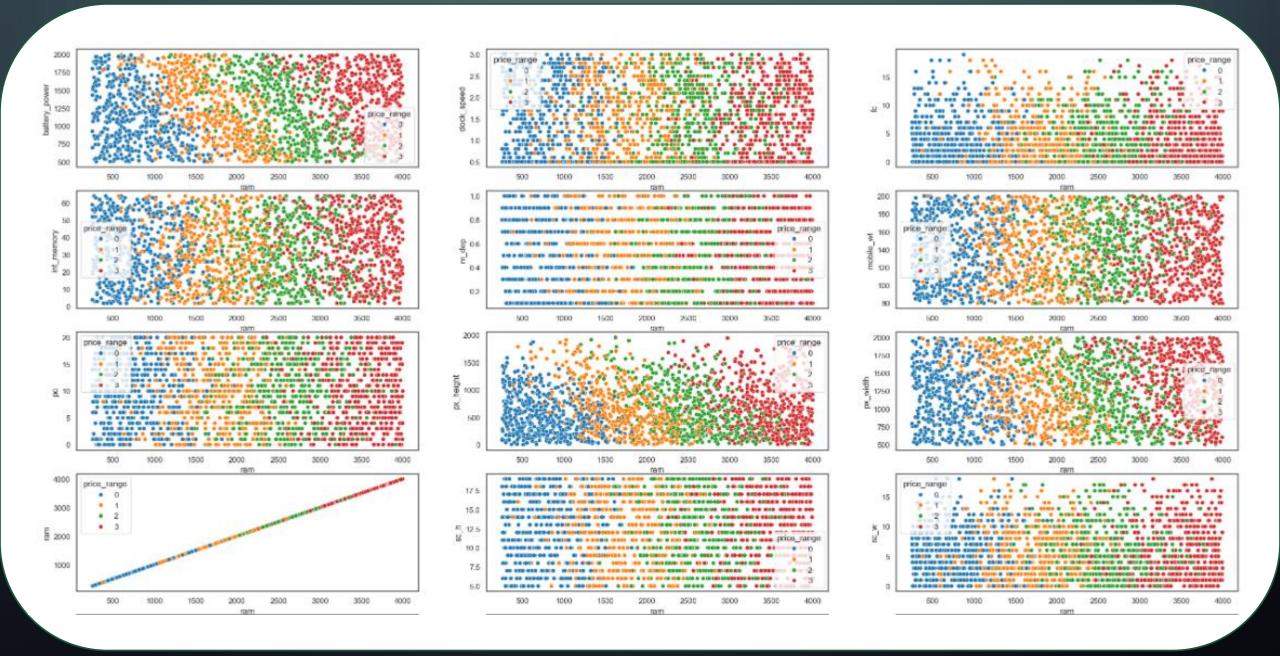


Quanto maior largura da resolução de pixel da tela maior o preço.

Box Plots price_range vs talk_time



ANÁLISE DESCRITIVA - DADOS NUMÉRICOS - ATRIBUTOS NUMÉRICOS X RAM.



FIM... MUITO OBRIGADO!