1) Em relação aos dados disponibilizados, existem dados missing? Descreva o que foi encontrado. Em situações como essa, o que é necessário ser feito?

```
[ ]  1  #Verificando se existe algum valor nulo nas colunas
     2  dados.isna().sum()
```

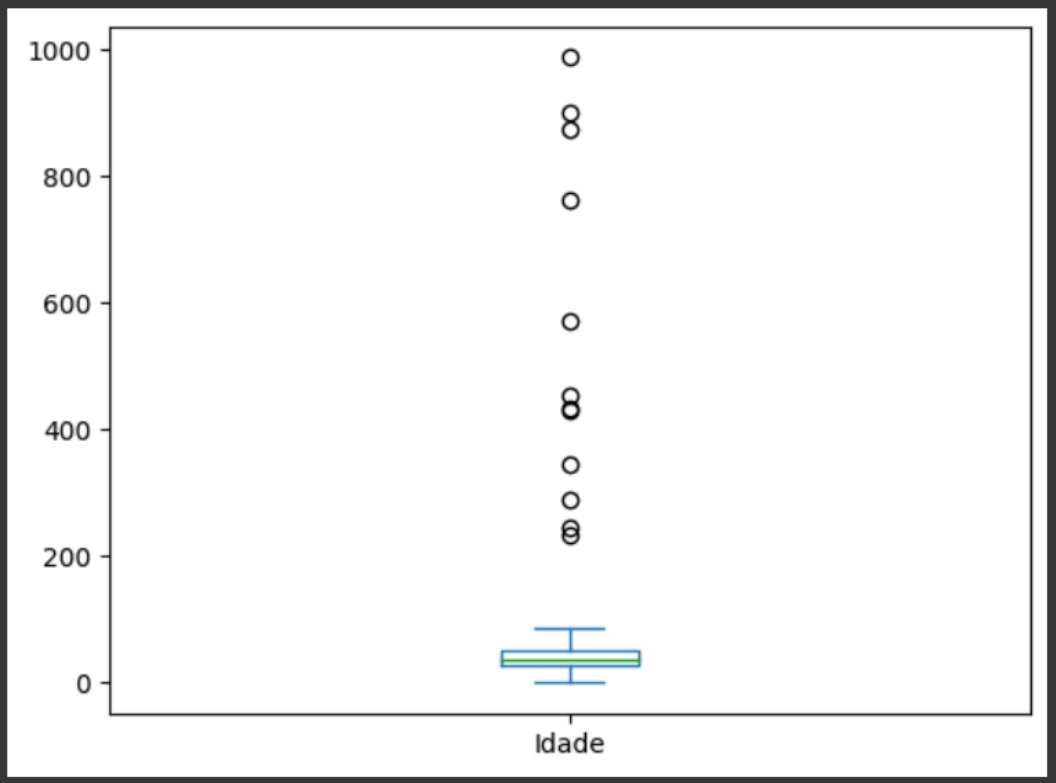
```
Região País              0
Estado                   0
Data                     0
FormaPagto               0
Sexo Informado Cliente   0
Idade                    0
valor ticket médio       0
numero pedido            0
dtype: int64
```

Foi utilizado um método para verificar se existem valores nulos nas colunas da planilha e tivemos como resultado que nenhuma das colunas possuem valores nulos. Então não será necessário tratar valores nulos

2) Analise os dados na perspectiva da coluna idade. Existem Outliers nos dados disponibilizados? É possível identificar algo em relação ao ticket médio de vendas relacionadas a esses Outliers? Justifique sua resposta.

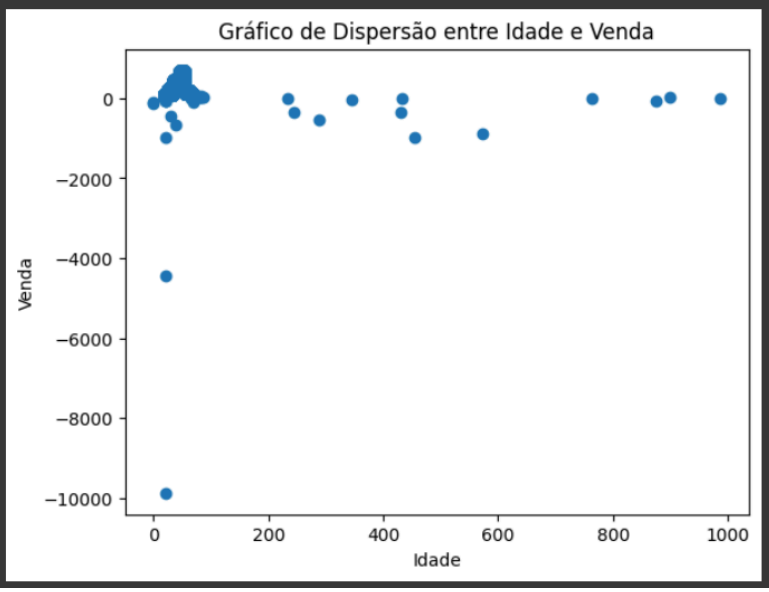
Sim, foram identificados outliers. E sim foram identificadas relações entre os outliers entre idade e ticket médio onde os valores de ticket médio que estão correlacionados também são outliers pois possuem valores negativos abaixo de zero.

```
1  dados['Idade'].plot.box()
2  plt.show()
3
```



Aqui podemos verificar através do gráfico de dispersão que existem idades muito discrepantes como idades acima de 200 e idades 0

```python
# Gráfico de dispersão cruzando as colunas 'idade' e 'valor ticket médio' para verificar se existe alguma relaçã
plt.scatter(dados['Idade'], dados['valor ticket médio'])
plt.title('Gráfico de Dispersão entre Idade e Venda')
plt.xlabel('Idade')
plt.ylabel('ticket médio')

# Mostrar o gráfico
plt.show()
```




```python
# Mostrando relação existente entre as colunas idade e valor ticket médio
relacao = dados[['Idade','valor ticket médio']]
relacao[relacao['Idade'] >= 200]
# Essa relação é possível identificar uma relação de valor ticket médio com os
# outliers da coluna idade que quando ocorre uma idade discrepante ela está relacionado a um outlier da coluna valor ticket médio
```

	Idade	valor ticket médio
9525	244	-345.0
9546	345	-22.0
9805	455	-987.0
10246	289	-542.0
10708	431	-341.0
12422	572	-872.0
13494	763	0.0
14475	987	0.0
15680	233	0.0
16409	433	-2.0
17060	874	-76.0
18054	900	32.0

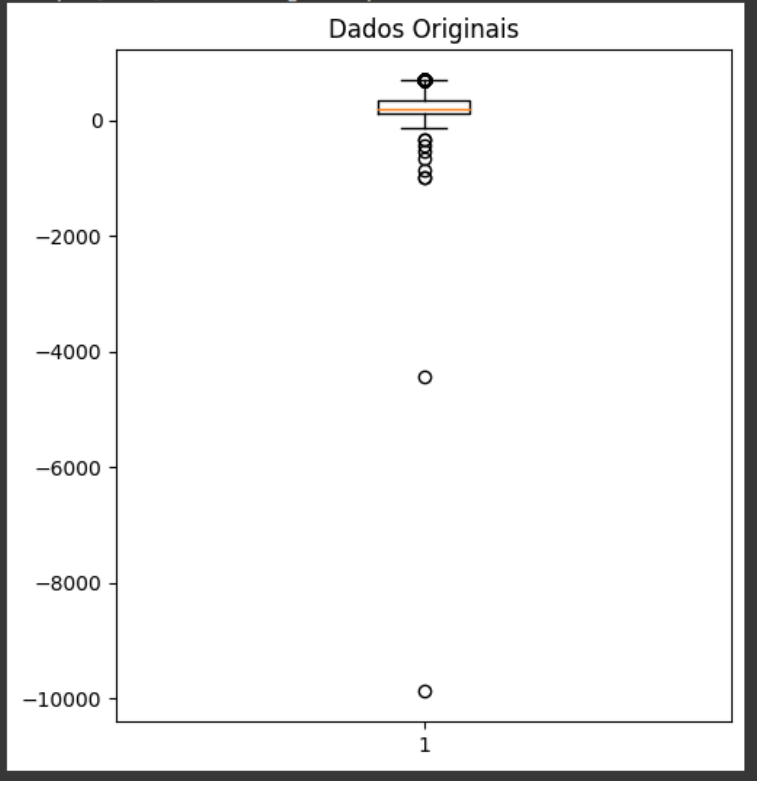
A partir desse gráfico é possível identificar uma relação de valor ticket médio com os outliers da coluna idade que quando ocorre uma idade discrepante ela está relacionado a um outlier da coluna valor ticket médio como mostrado acima

3) Em relação à consistência do dado valor ticket médio, o que é possível refletir sobre seus conteúdos? Existem dados inconsistentes? Justifique como é possível corrigi-los e realize essa importante atividade, deixando esses dados prontos para análise:

a) podemos refletir que existem valores que não estão consistentes pois há registros de ticket médio com valores negativos.
b) Sim, existem. Para corrigi-los, precisamos primeiro visualizar onde estão esses Outliers, e ao criar essa visualização nesse caso de uso, conseguimos identificar valores de ticket médio negativos, o que impossibilita uma análise acurada.

```
plt.figure(figsize=(12, 6))
plt.subplot(1, 2, 1)
plt.boxplot(dados['valor ticket médio'])
plt.title('Dados Originais')
```

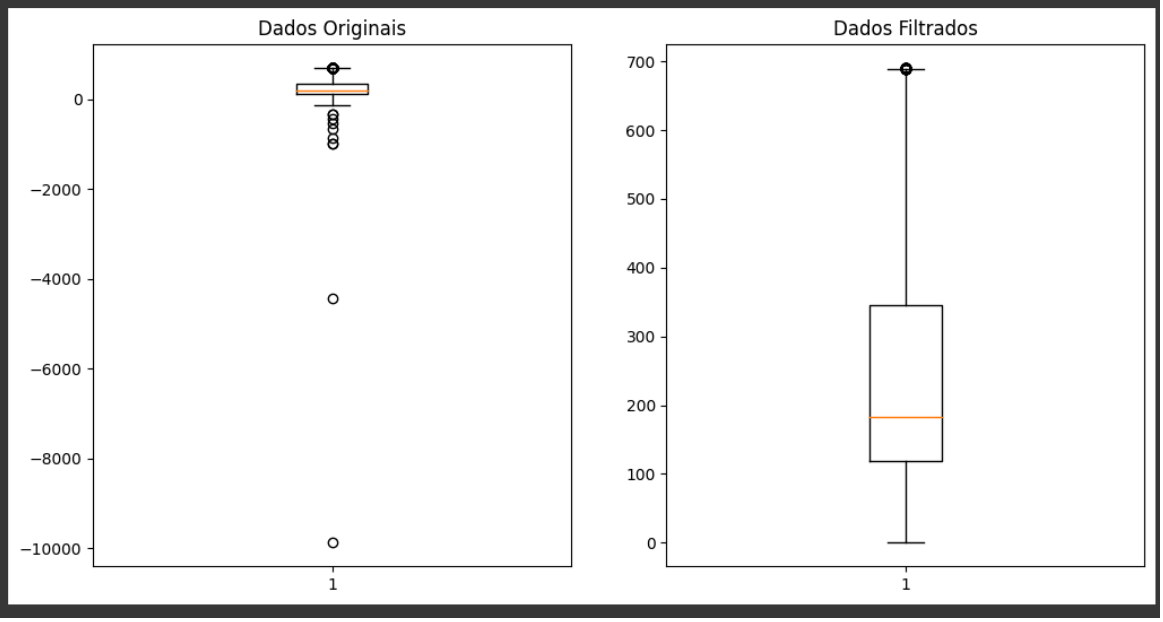
```
Text(0.5, 1.0, 'Dados Originais')
```



Para resolver isso, fizemos um script que filtra as informações da base que não são Outliers e salvamos em um novo Data Frame dessa forma conseguimos utilizar a base sem as inconsistências antes existentes:

```
[228]
# Removendo outliers das colunas idade e valor ticket médio
dados = dados[
    ((dados['Idade'] >= lower_boundIdade) & (dados['Idade'] <= upper_boundIdade) & (dados['Idade'] != 0)) &
    (dados['valor ticket médio'] >= 0)
]
```

```
[264] # Dados originais
      plt.figure(figsize=(12, 6))
      plt.subplot(1, 2, 1)
      plt.boxplot(dados['valor ticket médio'])
      plt.title('Dados Originais')

      # Dados filtrados
      plt.subplot(1, 2, 2)
      plt.boxplot(dados_filtered['valor ticket médio'])
      plt.title('Dados Filtrados')

      plt.show()
```



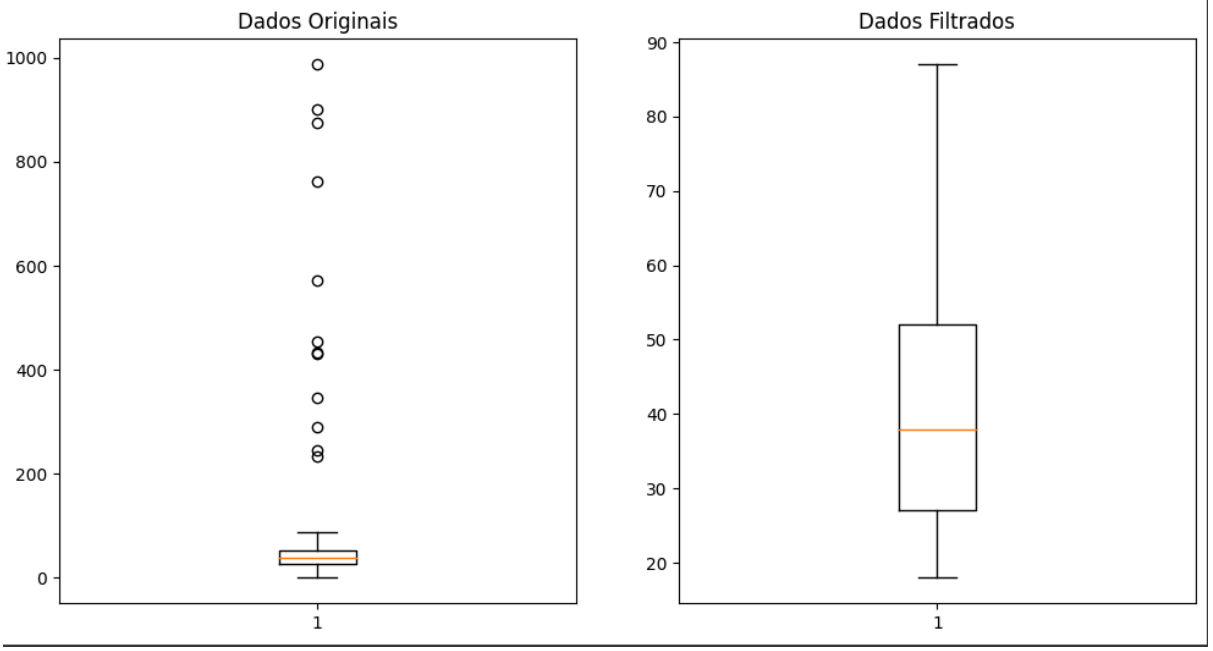
4) A área comercial da Melhores Compras criou um conjunto de faixa etária para tentar compreender melhor o perfil do cliente, mas não conseguiu até o momento chegar a lugar algum. Veja as faixas determinadas: entre 18 e 24 anos; entre 25 e 34 anos; entre 35 e 44 anos; entre 45 e 54 anos; entre 55 e 64 anos; com mais de 65 anos..
Sendo assim, após aplicar a limpeza e tratamento nos dados, tente contribuir com o departamento comercial gerando informações que auxiliem a tomada de decisão, como valor do ticket médio por faixa etária, idade média dos clientes selecionados, variância da idade, desvio padrão da idade, valor médio e mediana por idade ou faixa etária e ranking das vendas por faixa etária são alguns exemplos de contribuição. Por fim, faça uma análise sobre o resultado alcançado e apresenta recomendações para o departamento comercial sobre possíveis ações que podem ser feitas sobre o que foi identificado.

Limpeza da coluna idade:

```python
# Criando os limites dos quantils da coluna idade
QIdade1 = dados['Idade'].quantile(0.25)
QIdade3 = dados['Idade'].quantile(0.75)
#Calculando o IRQ
IQRIdade = QIdade3 - QIdade1

# Intervalo inferior do quartil
lower_boundIdade = QIdade1 - 1.5 * IQRIdade

# Intervalo superior do quartil
upper_boundIdade = QIdade3 + 1.5 * IQRIdade

```

```python
dados_filtered = dados[
    ((dados['Idade'] >= lower_boundIdade) & (dados['Idade'] <= upper_boundIdade) & (dados['Idade'] != 0)) &
    (dados['valor ticket médio'] >= 0)
]
```

```python
# Dados originais
plt.figure(figsize=(12, 6))
plt.subplot(1, 2, 1)
plt.boxplot(dados['Idade'])
plt.title('Dados Originais')

# Dados filtrados
plt.subplot(1, 2, 2)
plt.boxplot(dados_filtered['Idade'])
plt.title('Dados Filtrados')

plt.show()
```



1. **Remoção dos outliers:** Utilizamos o critério de identificação do boxplot foi feita a separação dos quartis Q1 e Q3 dos dados da idade. Q1 é o valor

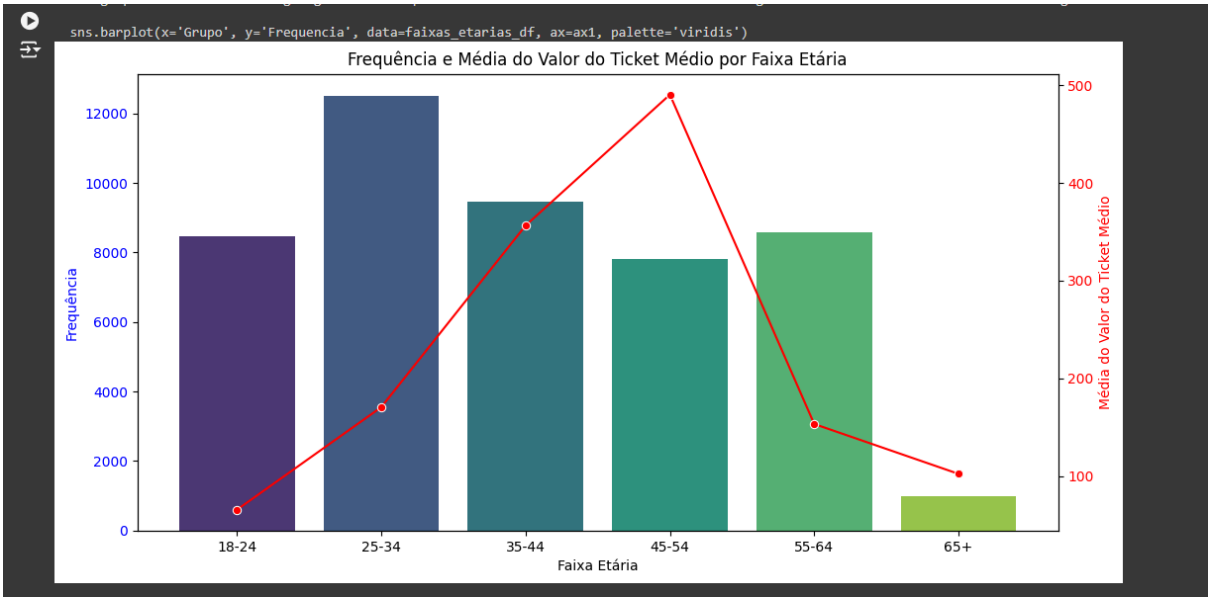
abaixo do qual está o 25% inferior dos dados, enquanto Q3 é o valor abaixo do qual está os 75% inferior dos dados.

2. **Foi feito o intervalo interquartil(IQR):** Ele é calculado subtraindo Q1 de Q3. Ele fornece uma medida da dispersão dos dados, considerando apenas os valores centrais.
3. **Calculamos o limite inferior e superior:** Os limites inferior e superior são calculados usando o IQR. O limite inferior é calculado subtraindo 1.5 vezes o IQR de Q1, enquanto o limite superior é calculado adicionando 1.5 vezes o IQR a Q3.
4. **Identificação de Outliers:** Valores abaixo do limite inferior ou acima do limite superior são considerados outliers e podem ser tratados de acordo com as necessidades da análise. Eles podem ser removidos do conjunto de dados ou tratados de outra maneira, dependendo do contexto da análise.
5. Removendo idades zeradas: O critério removeu 99% dos outliers, porém não removeu as idades zeradas. Portanto no momento de filtrar só colocamos uma condição de filtrar somente idades diferentes de 0

```
[267]
dados = pd.DataFrame(dados)
coluna_idade = 'Idade'
faixas_etarias = {
    'Grupo': ['18-24', '25-34', '35-44', '45-54', '55-64', '65+'],
    'Frequencia': [0, 0, 0, 0, 0, 0]
}
faixas_etarias = [18, 25, 35, 45, 55, 65, float('inf')]

# Criando as faixas etárias automaticamente
faixas_etarias_rotulos = ['18-24', '25-34', '35-44', '45-54', '55-64', '65+']
dados['faixa_etaria'] = pd.cut(dados[coluna_idade], bins=faixas_etarias, labels=faixas_etarias_rotulos, right=False)

# Contando a frequência de cada faixa etária
frequencia_faixas_etarias = dados['faixa_etaria'].value_counts().sort_index()

# Criando DataFrame a partir da frequência das faixas etárias
faixas_etarias_df = pd.DataFrame({'Grupo': frequencia_faixas_etarias.index, 'Frequencia': frequencia_faixas_etarias.values})

# Exibindo o DataFrame
faixas_etarias_df
```

	Grupo	Frequencia
0	18-24	8456
1	25-34	12512
2	35-44	9458
3	45-54	7816
4	55-64	8588
5	65+	978

a) **Valor do ticket médio por faixa etária e frequência:**



Conclusão: em termos de frequência de compra concluímos que o grupo de 25-34 anos possui uma maior frequência dentre os grupos, contudo o seu ticket médio fica entre 100 e 200 reais(dentre os mais baixos). E ao analisarmos o ticket médio vemos que o grupo de 45-54 anos possuem a maior média de ticket médio dos grupos, mesmo estando entre as mais baixas frequências médias de compra dos grupos analisados.

b) **Idade média e mediana dos clientes:**

```
[273]  1  # CALCULANDO A IDADE MEDIA DOS CLIENTES
       2
       3  dados ['Idade'].mean()
       4
       5  # A idade média é aproximadamente 39 anos
```

```
39.36464608433735
```

```
[ ]  1  # CALCULANDO A IDADE MEDIANA DOS CLIENTES
     2
     3  dados ['Idade'].median()
     4
     5  # A idade mediana é aproximadamente 38 anos
```

```
38.0
```

Conclusão: a idade média dos clientes da base analisada é de aproximadamente 39 anos, e a mediana é exatos 38 anos.

c) Variância na idade:

```
[275]  1  # CALCULANDO A VARIANCIA NA IDADE DOS CLIENTES
       2
       3  dados ['Idade'].var()
       4
       5  # A variância idade dos clientes é de 275
```

```
195.28571390825084
```

Conclusão: a variância de idade dos clientes na base analisada é de 195.28.

d) Desvio padrão da idade:

```
[276]  1  # CALCULANDO O DESVIO PADRÃO NA IDADE DOS CLIENTES USANDO NUMPY
       2
       3  idade=np.std(dados['Idade'])
       4
       5  idade
       6  # O desvio padrão na idade dos clientes é de 16.61
```

```
13.974320345430456
```

Conclusão: a idade média dos clientes analisados na base é de aproximadamente 39 anos, com um desvio-padrão aproximado de 13.9 anos para mais e para menos.

e) Valor médio e mediana por faixa etária:

```python
# CALCULANDO O VALOR MÉDIO POR FAIXA ETÁRIA

media_faixa_etaria = dados.groupby('faixa_etaria')['Idade'].mean().sort_index()
faixas_etarias_df = pd.DataFrame({
    'Grupo': frequencia_faixas_etarias.index,
    'Media_Idade_Faixa_Etaria': media_faixa_etaria.values,
})

# 9. Exibindo o DataFrame
faixas_etarias_df
```

	Grupo	Media_Idade_Faixa_Etaria
0	18-24	21.217952
1	25-34	29.418558
2	35-44	39.310002
3	45-54	49.634340
4	55-64	59.459013
5	65+	65.512270

```python
# CALCULANDO A MEDIANA POR FAIXA ETÁRIA

mediana_faixa_etaria = dados.groupby('faixa_etaria')['Idade'].median().sort_index()
faixas_etarias_df = pd.DataFrame({
    'Grupo': frequencia_faixas_etarias.index,
    'Mediana_Idade_Faixa_Etaria': mediana_faixa_etaria.values,
})

# EXIBINDO
faixas_etarias_df
```

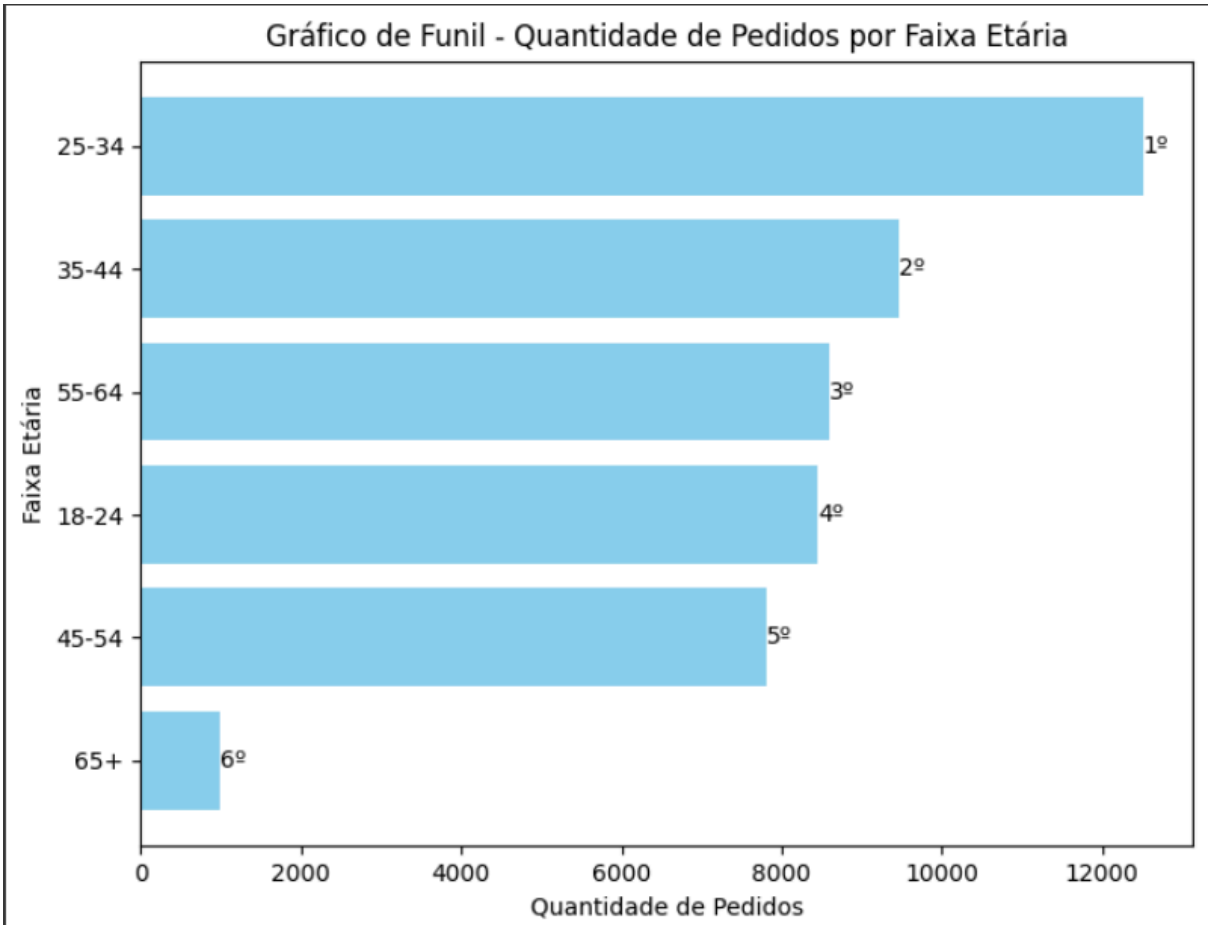
	Grupo	Mediana_Idade_Faixa_Etaria
0	18-24	21.0
1	25-34	29.0
2	35-44	39.0
3	45-54	50.0
4	55-64	59.0
5	65+	65.0

Conclusão: a faixa etária de 65+ possui uma média de 65.5 e a mediana da mesma possui 65.0.

f) Ranking de vendas por faixa etária:

```python
# RANKING DE VENDAS POR FAIXA ETÁRIA

pedidos_faixa_etaria = dados.groupby('faixa_etaria')['numero pedido'].count()
ranking_faixa_etaria = pd.DataFrame({
    'Grupo': frequencia_faixas_etarias.index,
    'Ranking_faixa_etaria': pedidos_faixa_etaria.values
})

# ORDENANDO DO MAIOR PARA O MENOR
ranking_faixa_etaria = ranking_faixa_etaria.sort_values(by='Ranking_faixa_etaria', ascending=False)

# EXIBINDO O DF
ranking_faixa_etaria
```

	Grupo	Ranking_faixa_etaria
1	25-34	12512
2	35-44	9458
4	55-64	8588
0	18-24	8456
3	45-54	7816
5	65+	978



Conclusão: o nº1 no ranking de vendas está representado pela faixa etária de 25-34 anos.

Considerações finais: ao analisarmos a base e os perfis dos clientes do Melhores Compras, concluímos que temos diferentes perfis de consumidores: os que compram mais com valores de tickets médios menores, e os que compram consideravelmente menos e que investem em tickets maiores.
Dessa forma podemos fazer bases de ações segmentadas caso a estratégia seja aumentar o ticket desses que já possuem uma boa frequência de compra, ou aumentar a frequência de quem já possui um ticket médio consideravelmente bom na plataforma, tudo vai depender da estratégia de negócio no momento.