Avaliação Final
Python para Ciência de Dados

Instituição: Pontifícia Universidade Católica de Minas Gerais
Alunos: Alessandro Augusto Bezerra
Isabela D'loan
Robson Gomes de Lima
Vitor Fernando de Souza Rodrigues
Oferta: 07 - **Turma:** 1
Disciplina: Python para Ciência de Dados
Docente: Leandro Lessa
Grupo: 04

Sumário

Introdução

No decorrer do tempo, as organizações vêm enfrentando maiores desafios sobre lidar com os seus respectivos volumes massivos de dados diariamente, exigindo tecnologias eficientes para análises. A ciência e a análise de dados são essenciais no cotidiano moderno. Elas permitem transformar vastas quantidades de dados em insights valiosos, ajudando a resolver problemas e a otimizar processos em diversos setores, como saúde, educação e negócios. Com a análise de dados, é possível prever tendências, personalizar experiências e auxiliar na tomada de decisão. No campo da ciência e análise de dados, várias ferramentas e bibliotecas têm sido desenvolvidas para aprimorar essas práticas. Exemplos disso são as bibliotecas Pandas e NumPy, da linguagem Python, que são cruciais para a manipulação e organização de dados. Matplotlib e Seaborn permitem criar visualizações que tornam os dados mais compreensíveis. Além disso, análises preditivas com Machine Learning e inteligência artificial possibilitam previsões precisas e automação de decisões. Portanto, essas tecnologias colaboram a transformar grandes volumes de dados em insights valiosos, capacitando as organizações a melhorar seus processos e decisões estratégicas visando maior precisão e confiança. Com isso, elas não apenas resolvem problemas imediatos, mas também se preparam para um futuro mais eficiente e inovador.

Um fluxo desde a coleta, tratamento, manipulação até a análise descritiva e exploratória de dados

Com base nos datasets *dados_caracteristicas_fisicas.csv, dados_pessoais.csv e na lista dos estados brasileiros*, que simulam informações demográficas e características de clientes, foi realizado um trabalho que tem como objetivo mostrar o processo desde a coleta dos dados, transformações, manipulações, análises, indicadores estatísticos com visuais gráficos relevantes para o processo de compreensão e entendimentos dos dados auxiliando que visa auxiliar em uma tomada de decisão.
Em nossa jornada de exploração, vamos conduzir uma análise minuciosa desses conjuntos de dados. Vamos mergulhar fundo para identificar tendências, padrões e informações valiosas que possam trazer benefícios tangíveis para uma organização. No decorrer do texto, mostraremos insights que possam impulsionar estratégias de acordo com os conjuntos de dados e seus respectivos indicadores.

Bibliotecas Python utilizadas nesse trabalho

- Pandas
- Numpy
- Seaborn
- Matplotlib
- Requests
- Warnings

Análises de dados, gráficos e insights obtidos

- **Processo de coleta de dados**
  O conjunto de dados foi baixado a partir do link. Estes arquivos (**caracteristicas fisicas** e **dados pessoais**) representam dados fictícios que contém informações detalhadas sobre os clientes de uma empresa, abrangendo uma ampla gama de aspectos, desde características demográficas até dados físicos e de saúde.

- **Tratamento de dados**
  - **Investigando os dados**
    Foi realizada uma análise a respeito das colunas dos dataframes passados com o intuito de identificar os pontos de trabalho e limpeza.
  - **Tratamento de dataframes**
    Realizamos o tratamento dos valores nulos, dos tipos de dados e dos dados duplicados. Substituímos as variáveis categóricas pela moda e as variáveis numéricas pelas médias.
  - **União (Joins) de Dataframes**
    Realizamos os joins e selecionamos apenas as colunas necessárias para análise, criando apenas um único dataframe.

Análises dos conjuntos de dados tratados e seus respectivos indicadores

- **Qual é a cor de cabelo mais comum entre os clientes?**

Ruivo (1272 clientes). Considerando que pessoas de cabelos ruivos representam cerca de 2% da população mundial (vide link), é provável que esta coleta foi executada em regiões norte-oeste da Europa, onde este contingente é maior.

- **Existe alguma correlação entre a altura e o peso dos clientes?**

No geral, o peso de uma certa população tem certo vínculo com sua respectiva altura dado que a maioria tende a ter um padrão de corpo saudável, sendo nem extremamente magro ou extremamente obeso. Analisaremos futuramente este ponto com mais afinco.

- **Qual é a distribuição de tatuagens entre os clientes?**

Vemos que a maioria da população possui 3 tatuagens, sendo quase que equiparado com o pessoal que possui duas e logo após, a população sem tatuagens, para no fim as pessoas com apenas uma tatuagem serem minoria.

É possível inferir que uma vez que a pessoa realiza uma tatuagem, a mesma se empolga e deseja fazer mais com o passar do tempo.

- **Quanto clientes tem o tipo sanguíneo AB-?**

646 clientes.

O tipo sanguíneo mais comum na população mundial é o O+, chegando a 42% da população (World Atlas). Em termos percentuais, esta população de equipara bastante, o comum seria ver um percentual maior destinado para os O+. Este tipo de percentual acaba de mostrando comum nas análises feitas.

- **Qual é o tipo de pele mais prevalente entre os clientes?**

Mista (1323 clientes).

Tipos de pele pode ter grande validade para empresas cosméticas que buscam vender seu produto cada vez mais para o público-alvo. Ver que a maioria da população possui pele do tipo Normal ou mista será de grande importância.

- **Qual é o nível de escolaridade predominante entre os clientes?**

Ensino médio (1279 clientes).

Esta escala acaba sendo similar a alguns países emergentes onde já foi vencida a barreira do ensino fundamental provocando uma população que consegue finalizar o ensino médio. No entanto, esta realidade não se difere um pouco com o Brasil onde cerca de metade da população já conseguiu ter formação básica completa. Vale salientar que estes tipos de análises são do tipo quantitativos e não qualitativos.

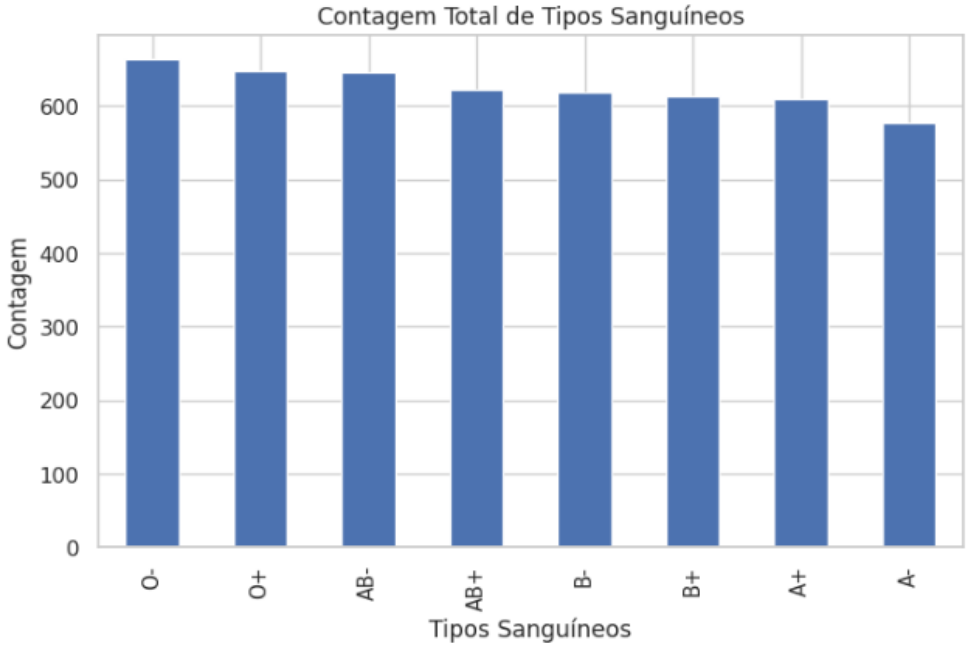
- **Quais as profissões mais comuns entre os clientes?**

Por haver uma maioria da população de estudantes, é provável que esta coleta tenha sido feita próximo a setores universitários ou de estágios profissionais, visto que a quantidade de estudantes se equipara com as quantidades dos profissionais que vem logo após.

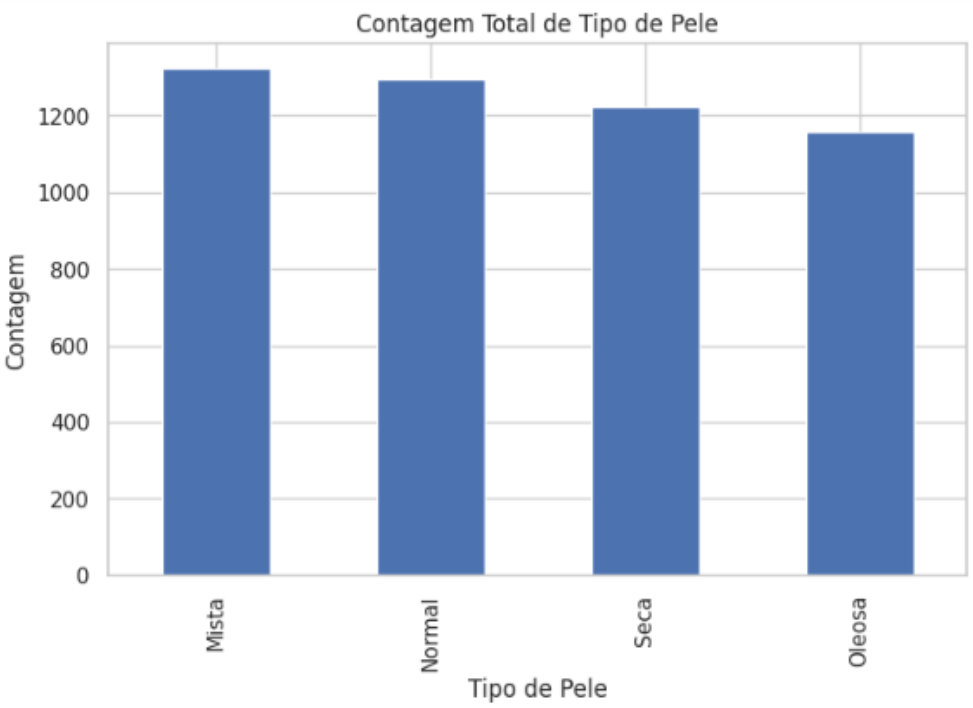
Profissão	
Estudante	669
Médico	648
Artista	623
Engenheiro	621
Advogado	618
Professor	612
Empresário	611
Outros	597

- **Gráfico: Quantidade total de pessoas por tipo sanguíneo.**

É curioso termos uma população onde o tipo sanguíneo O+ não seja maioria. Além disto, as quantidades estão bastante perto umas das outras.

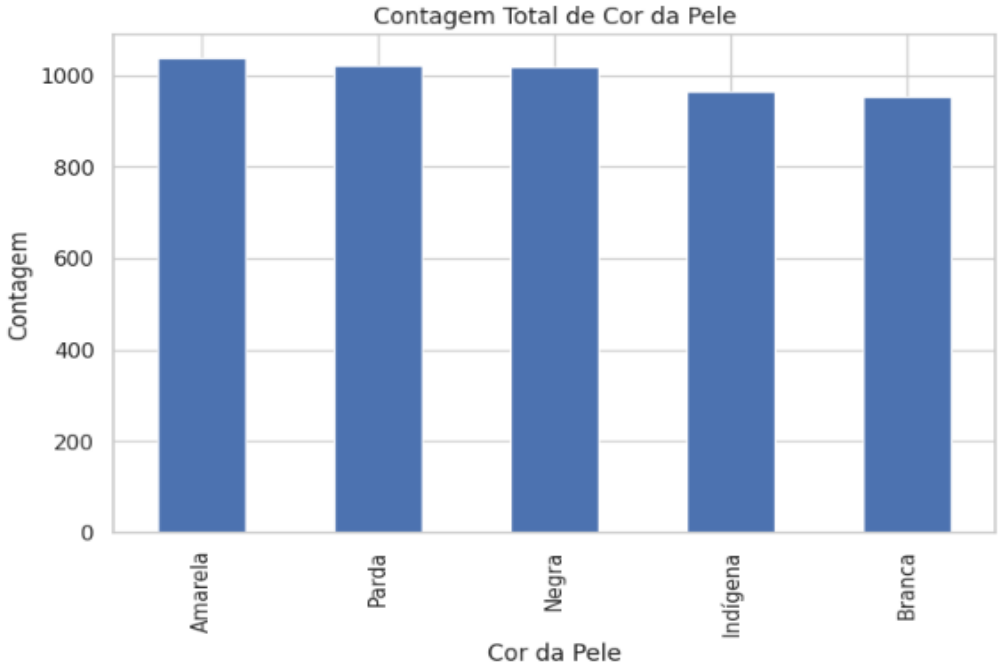


- **Gráfico: Quantidade total de pessoas por tipo de pele.**



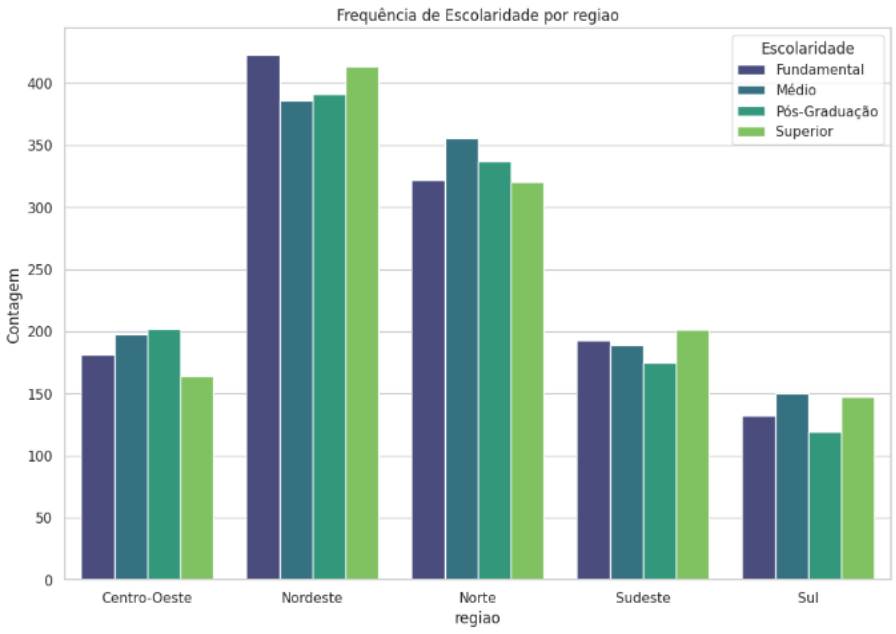
- **Gráfico: Quantidade total de pessoas pela cor de pele.**

Apesar de ser comum dizer que o Brasil é um país bastante miscigenado, na realidade este percentual possui suas maiorias. Segundo o IBGE, cerca de 89% da população se declara branca ou parda, seguindo de negra e uma parcela mínima indígena. O *dataset* mostrado pode conter uma coleta muito específica da população ou dados equivocados.



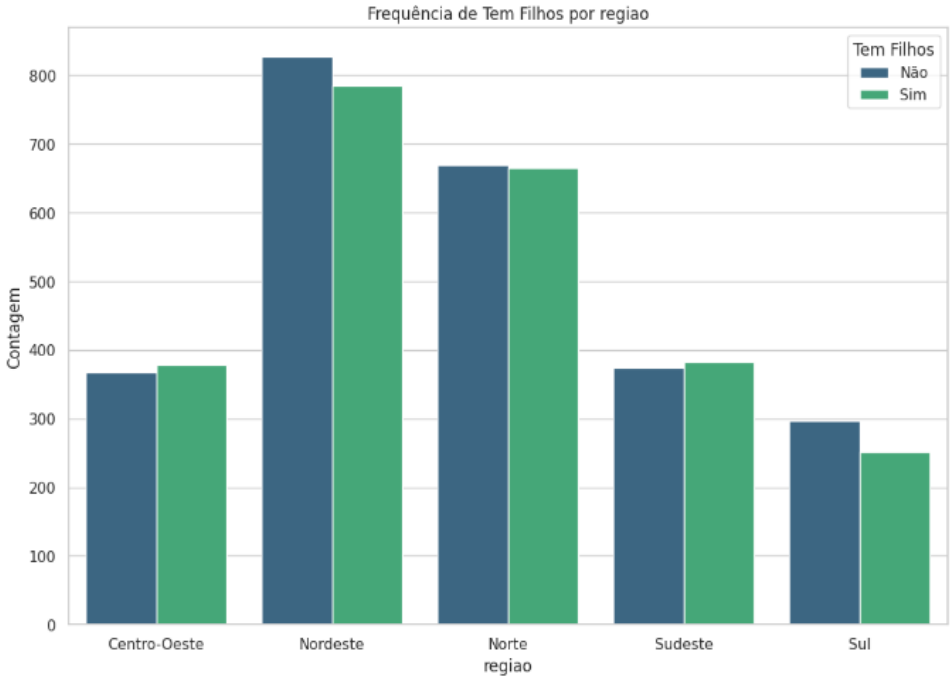
- **Gráfico: Quantidade de pessoas que possuem escolaridade que vai desde o fundamental, ensino médio, pós-graduação e superior, por região.**

  Ao separar a escolaridade por região, fica claro a diferencia populacional entre regiões. No entanto, segundo o IBGE, o Sudeste supera o Nordeste em população o que difere do mostrado no gráfico. Além disso, o índice de escolaridade se encontra bastante equiparado nas regiões.



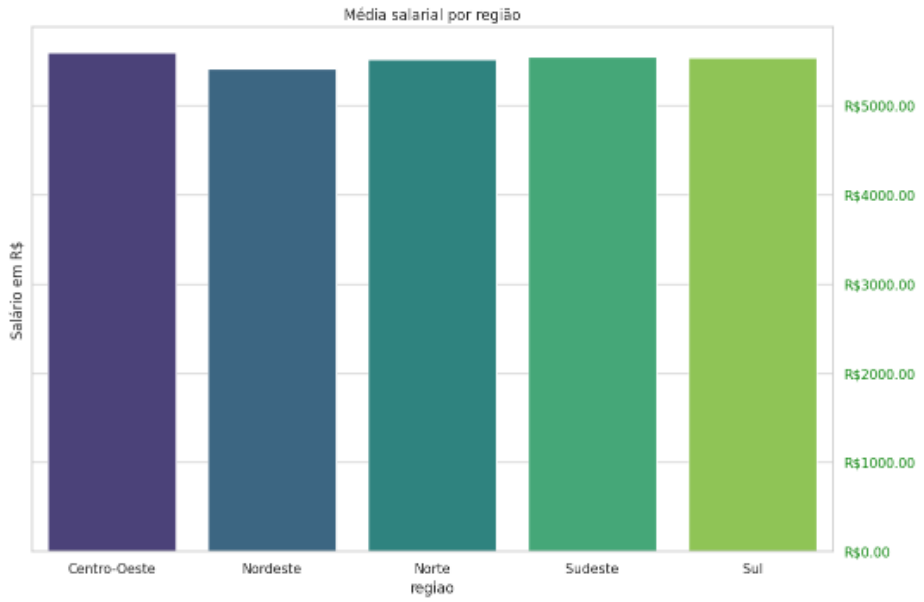
- **Gráfico: Quantidade de pessoas que possuem filhos por região.**

  Os quantitativos de pessoas que têm filhos e que não têm está bastante equiparado o que pode diferir da realidade.



- **Gráfico: Média salarial por região.**

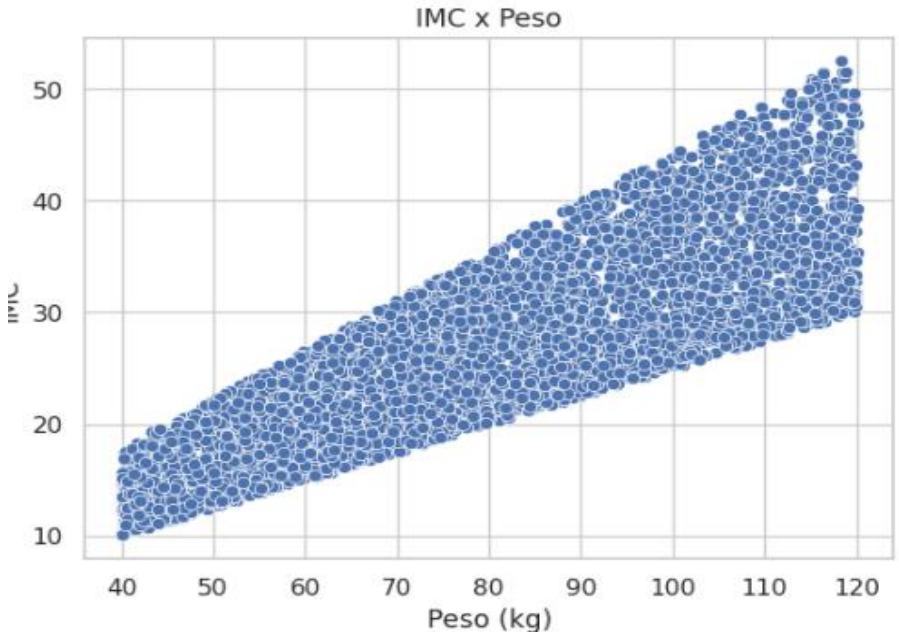
  Dadas informações quanto ao custo de vida, local de moradia e densidade populacional, os salários nas regiões podem se diferenciar especialmente se a profissão for muito qualificada. Para profissões mais comuns a média inicial gira em torno de R$ 1.700,00 a R$ 2.000,00 conforme link do g1.



- **Gráfico: Relação de IMC (Índice de Massa Corpórea).**

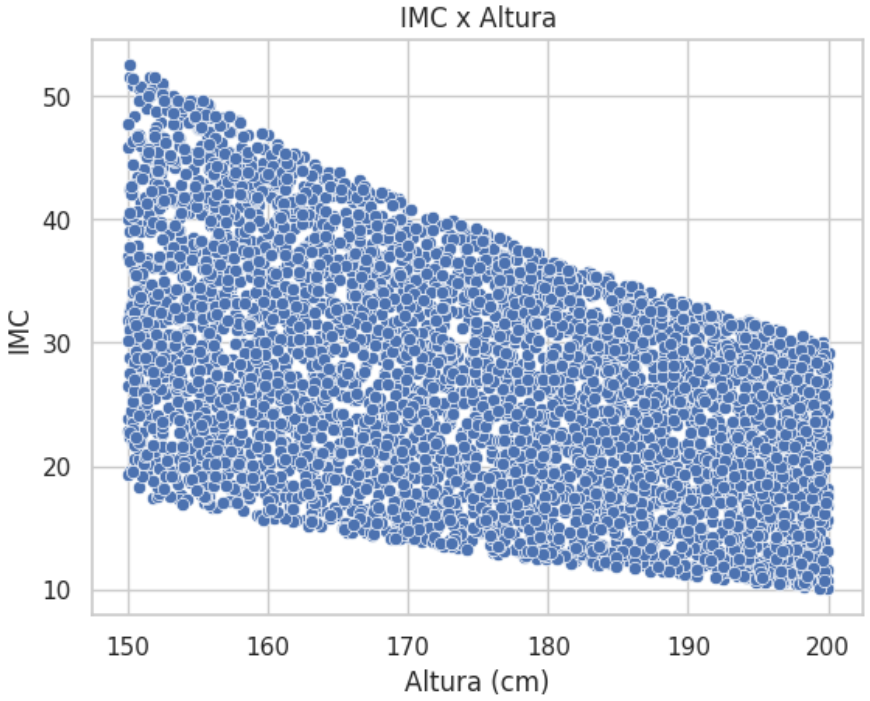
Como o IMC é linearmente dependente do peso, supõe-se que o gráfico acabe seguindo uma tendencia, neste caso é de subida. No entanto, o índice sofre uma

variação muito grande mostrando pessoas altamente subnutridas e extremamente obesas.



- **Gráfico: Relação do IMC por altura.**

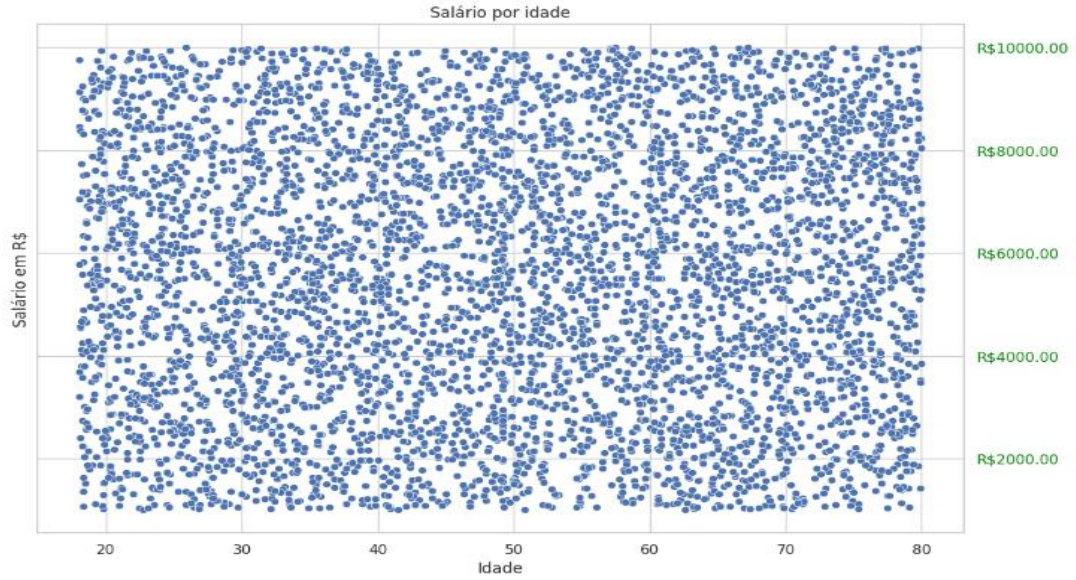
O gráfico da altura se comporta de maneira inversamente proporcional e continua a mostrar uma variação de dados extrema.



- **Gráfico: Média salarial por idade.**

O gráfico mostra que não há relação calculável entre idade e salário, qualquer idade pode receber o mesmo range de salários, o que sabemos na realidade há uma

certa tendência de que quanto maior a idade, maior o salário, dado que o profissional ganho experiência com o passar dos anos.

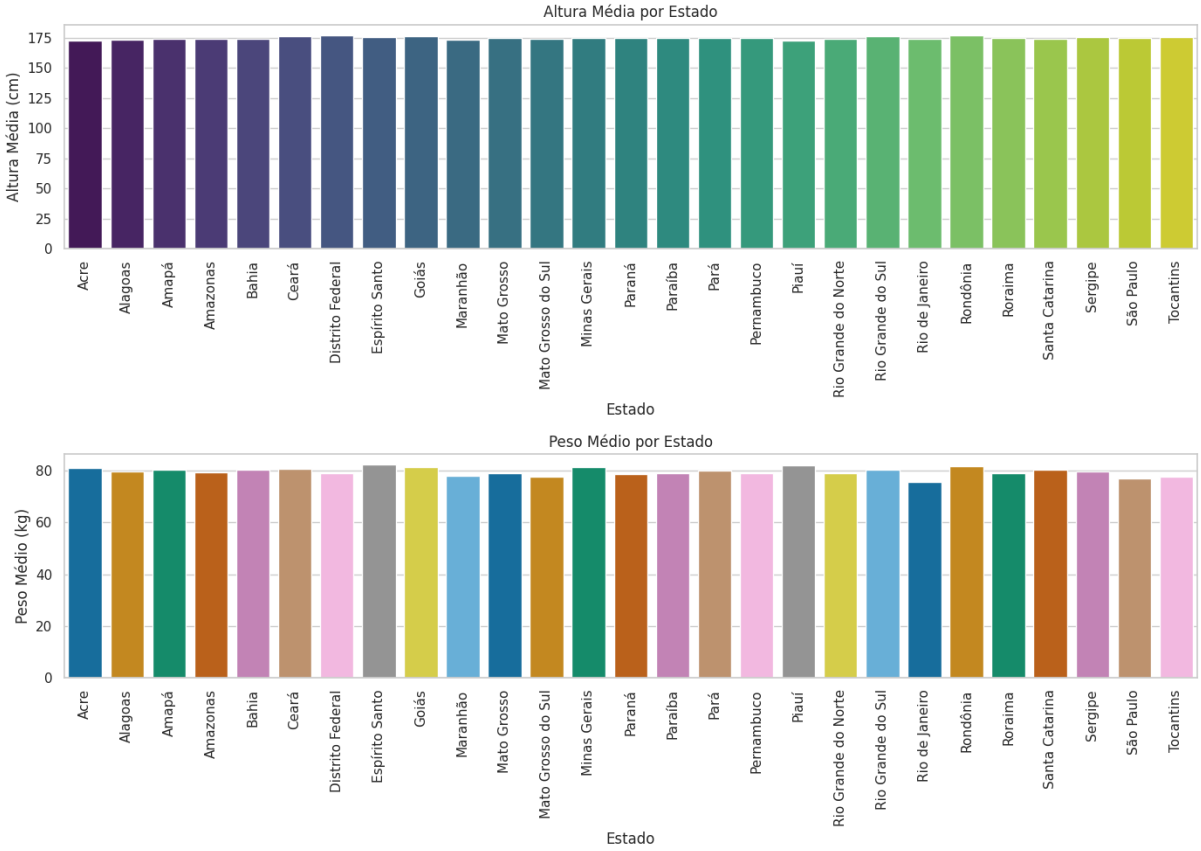


- **Agrupar por estado e calcular a média de altura e o respectivo peso.**

Vemos que a altura e peso se diferem pouco entre os estados do Brasil. Entretanto, o esperado seria que esta variação fosse maior dado que o Brasil é um país continental e foi colonizado por diversas etnias ao longo do tempo. Por exemplo, a região norte possui uma quantidade maior de nativos brasileiros, enquanto, que a região sul/sudeste foi bastante colonizada por culturas europeias e asiáticas. Apenas este ponto já seria o suficiente para visualizar diferenças na altura da média da população.

	estado	Altura (cm)	Peso (kg)
0	Acre	172.679341	81.033772
1	Alagoas	173.657791	79.504110
2	Amapá	174.385507	80.415990
3	Amazonas	174.562840	79.414793
4	Bahia	174.543354	80.324907

- **Gráfico: Calcular a altura e peso médio por estado.**

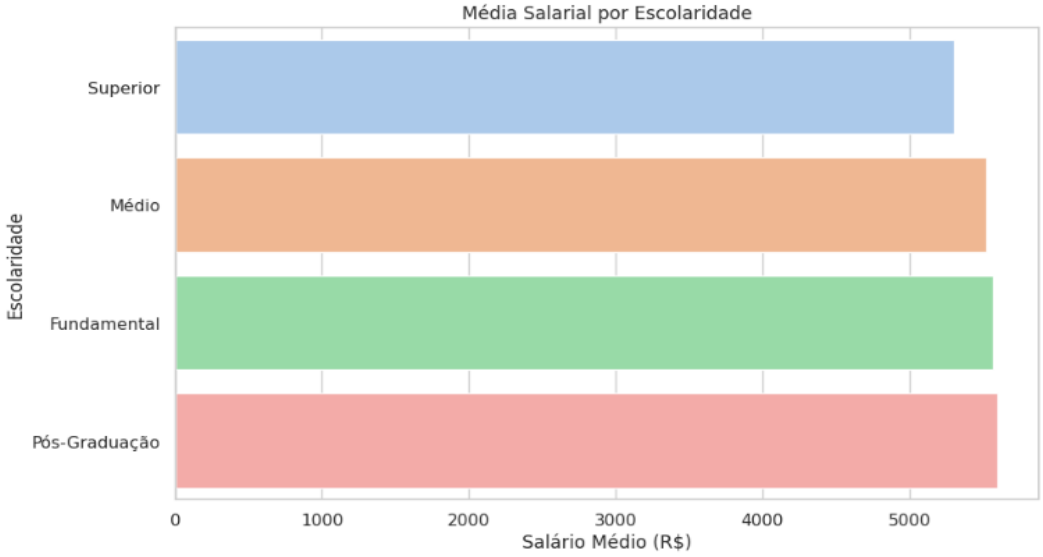


- **Qual é a relação entre escolaridade e salário médio dos clientes?**

A média salarial não segue o senso comum de que o aumento salarial segue o aumento hierárquico da escolaridade do indivíduo. Além disso, os salários estão muito próximos independentemente do nível de escolaridade.

	Escolaridade	Salário
3	Superior	R$ 5.304,07
1	Médio	R$ 5.522,66
0	Fundamental	R$ 5.574,34
2	Pós-Graduação	R$ 5.598,43

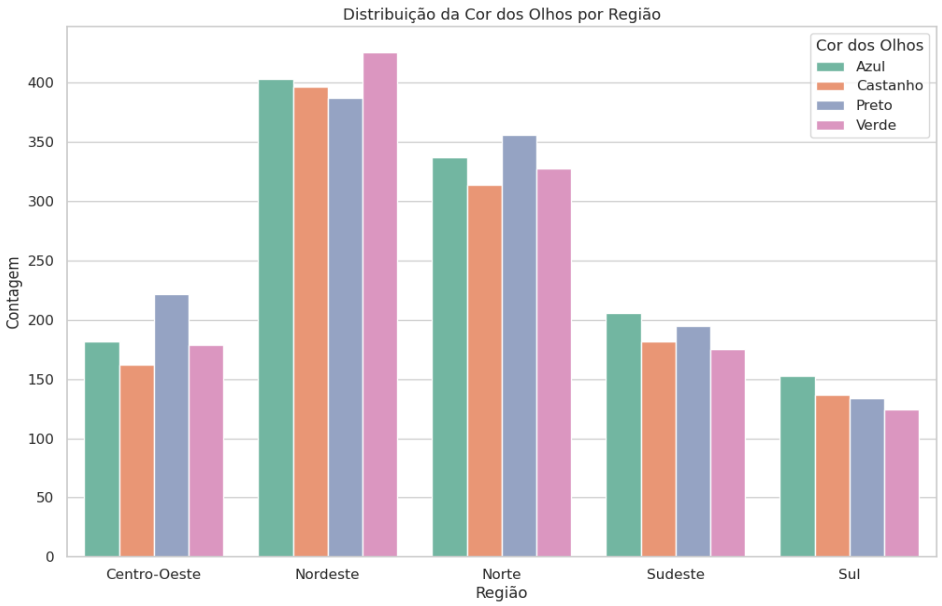
- **Gráfico: Calcular média salarial por escolaridade.**



- **Distribuição da cor dos olhos entre diferentes regiões.**

As cores dos olhos possuem uma boa variação nas populações regionais, especialmente no Centro-Oeste. A quantidade de pessoas segue mesma curva dos gráficos anteriores onde mostra uma densidade populacional maior no Nordeste.

	regiao	Cor dos Olhos	Quantidade
9	Norte	Castanho	314
15	Sudeste	Verde	175
17	Sul	Castanho	137
18	Sul	Preto	134
16	Sul	Azul	153

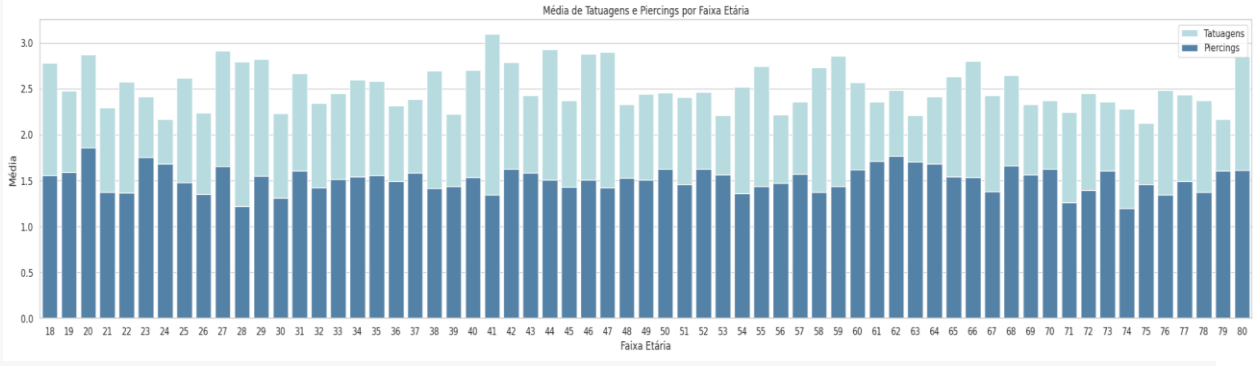


- **Verificar se existe correlação entre a quantidade de tatuagens e piercings com a idade dos clientes?**

A média de tatuagens e piercings segue dentro de uma faixa mínima e máxima, não importando a idade do indivíduo.

	Idade	Tatuagens	Piercings
0	18	2.777778	1.555556
1	19	2.474359	1.589744
2	20	2.868421	1.855263
3	21	2.293333	1.373333
4	22	2.576471	1.364706

- **Gráfico: Mostrar a média de tatuagens e piercings por faixa etária.**

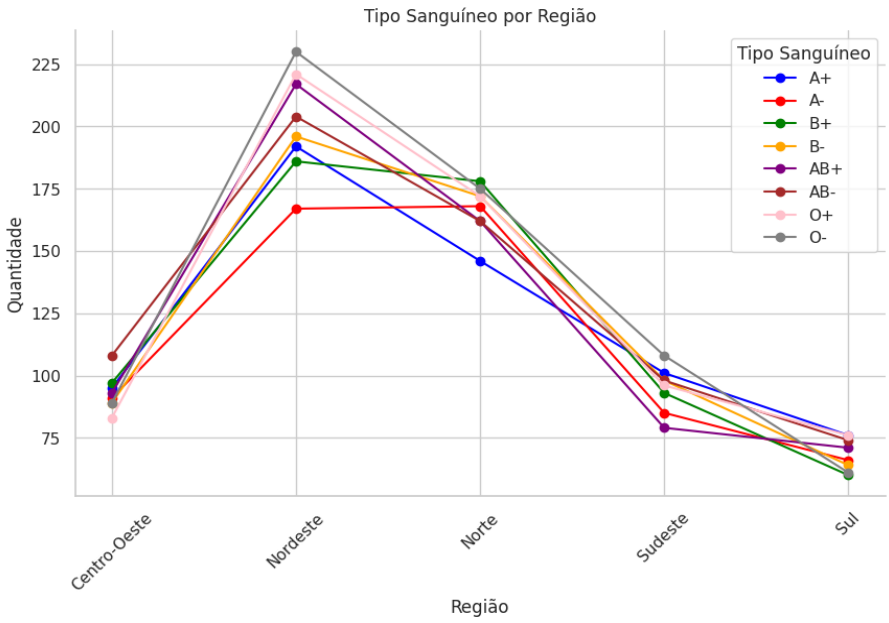


Um indicador que nos mostra a variação etária e respectivas médias de possuir tatuagens e piercings.

- **Qual é a distribuição do tipo sanguíneo por região**?
  - Agrupando por região e tipo sanguíneo.

	regiao	Tipo Sanguíneo	Quantidade
2	Centro-Oeste	AB+	93
16	Norte	A+	146
33	Sul	A-	66
12	Nordeste	B+	186
14	Nordeste	O+	221

- **Gráfico: Mostrar a quantidade de tipo sanguíneo por região.**

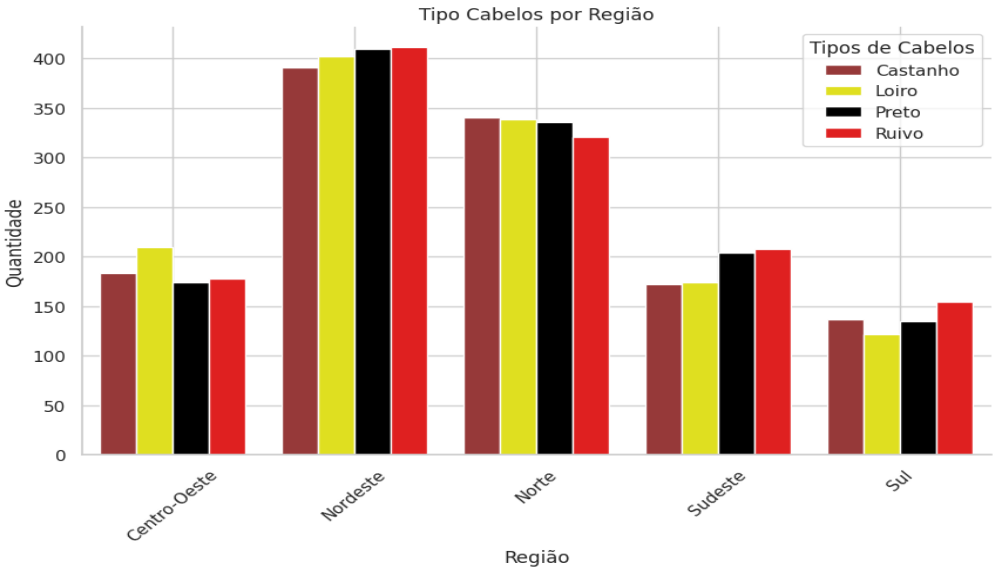


- **Qual é a distribuição da cor do cabelo dos clientes por região?**

	regiao	Cor do Cabelo	Quantidade
19	Sul	Ruivo	154
11	Norte	Ruivo	321
3	Centro-Oeste	Ruivo	178
18	Sul	Preto	135
0	Centro-Oeste	Castanho	183

As cores de cabelo possuem baixa variação de quantidade, além da população de ruivos no Brasil ser muito alta para a quantidade real.

- **Gráfico: Mostrar a quantidade de tipos de cabelos por região.**

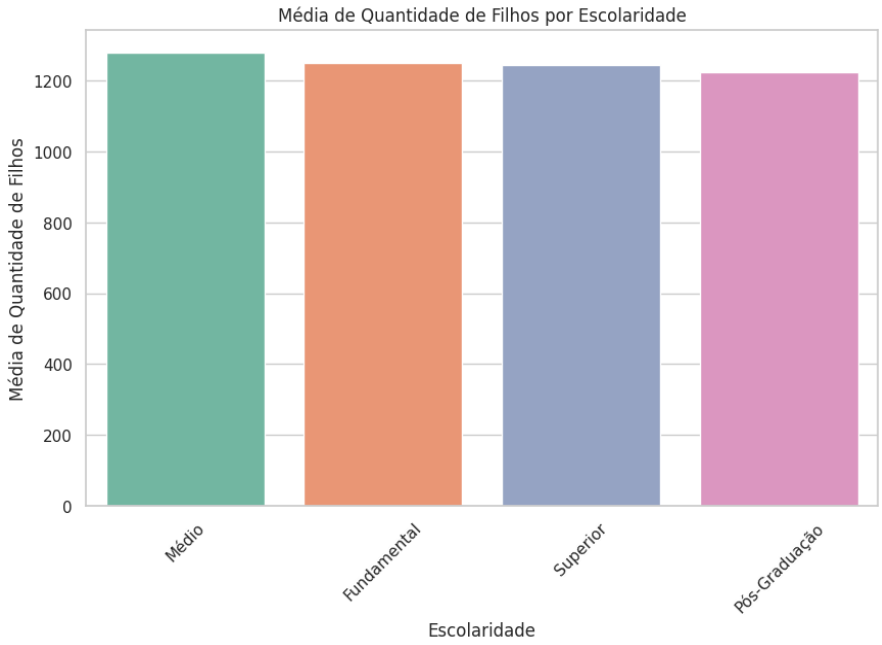


- **Há alguma relação entre a escolaridade e a quantidade de filhos?**

A quantidade de filhos tem baixa variação entre as escolaridades. Entretanto, a ordem descendente em que aparece, faz sentido onde hierarquias escolares maiores possuem menos filhos.

	Escolaridade	Tem Filhos
1	Médio	1279
0	Fundamental	1251
3	Superior	1245
2	Pós-Graduação	1224

- **Gráfico: Quantidade Total de filhos por escolaridade (ordem decrescente).**

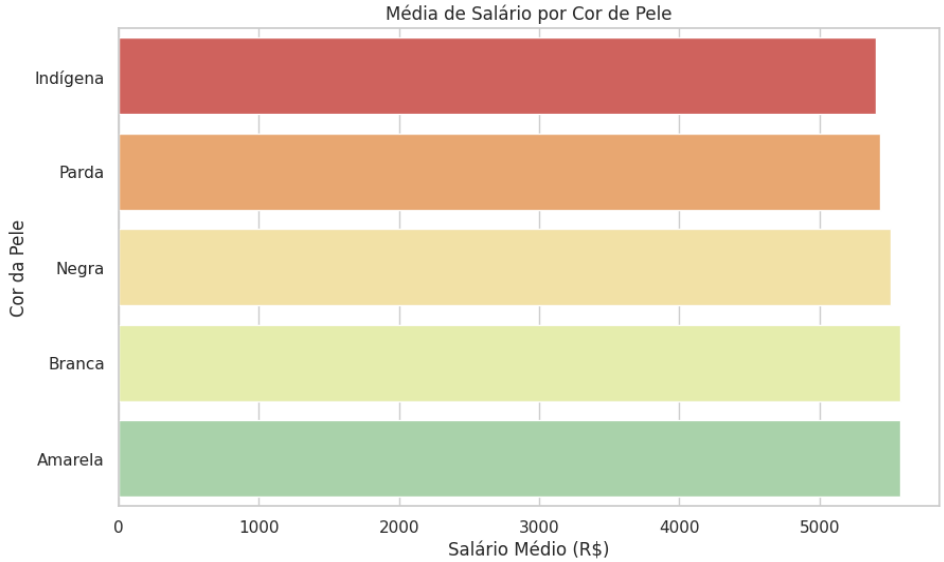


- **Como o salário dos clientes varia entre diferentes cores de peles?**

Indivíduos com cor de pele amarela possuem maior média salarial, seguido por pessoas brancas, negras e pardas, finalizando com indígenas.

	Cor da Pele	Salário
2	Indígena	5401.76
4	Parda	5432.74
3	Negra	5510.71
1	Branca	5573.94
0	Amarela	5577.73

- **Gráfico: Calcular média salarial por cores de peles.**

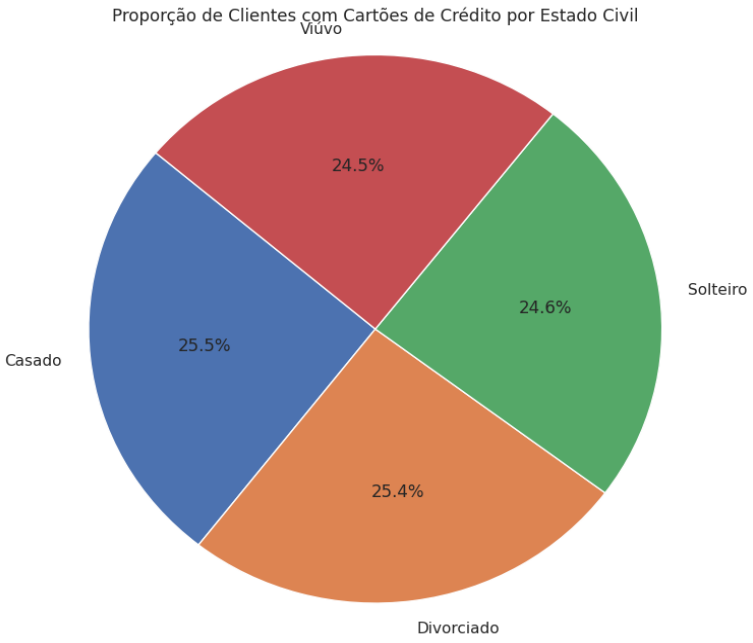


- **Qual é a proporção de clientes com cartões de crédito por estado civil?**

A proporção se encontra bastante igualitária da quantidade de cartão por estado civil.

	Estado Civil	proporcao_cartao
0	Casado	0.762740
1	Divorciado	0.757851
2	Solteiro	0.733483
3	Viúvo	0.732540

- **Gráfico: Calcular a proporção de clientes com cartões de crédito por estado civil (Pizza).**

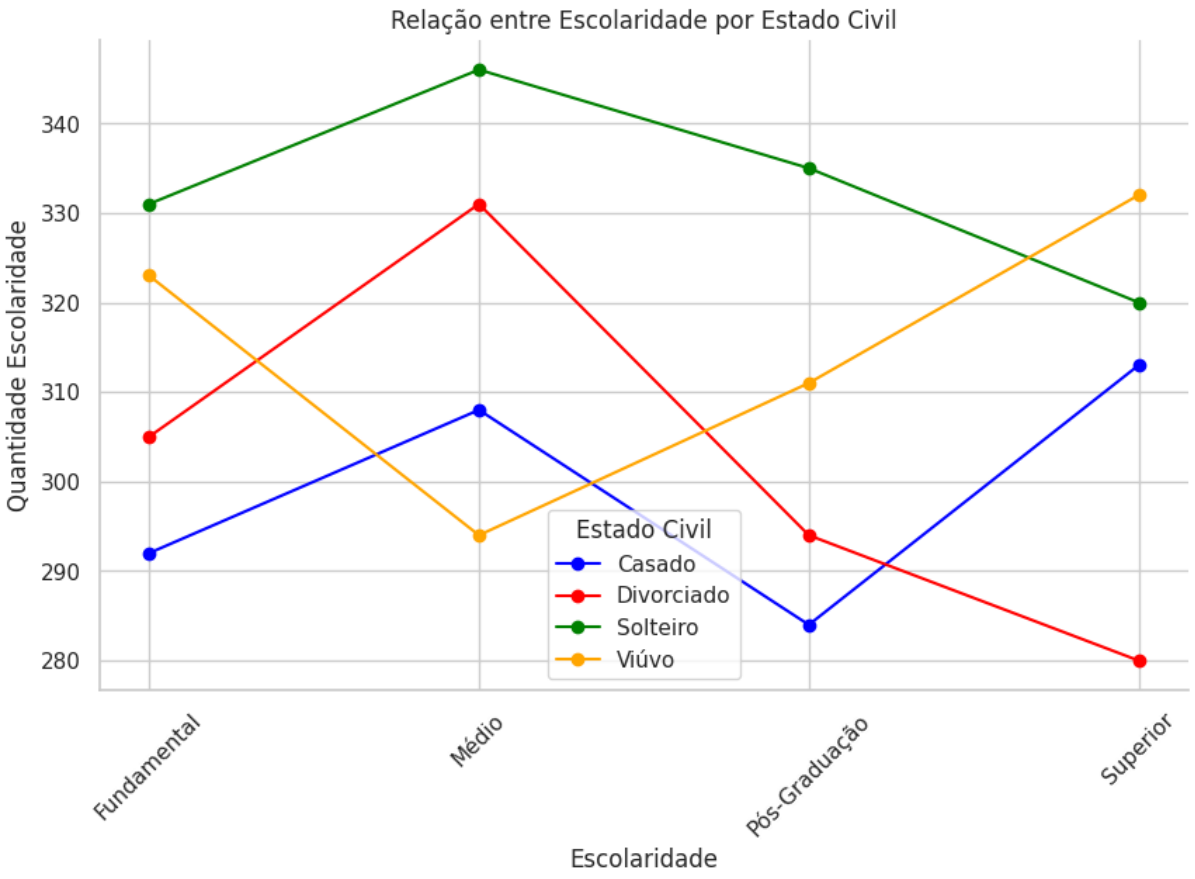


- **Qual é a quantidade de pessoas por escolaridade e seus respectivo estado civil?**

A quantidade de escolaridade é basicamente liderada por pessoas solteiras. Isto é esperado dado que a maioria das pessoas iniciam a vida acadêmica antes de engajar em relacionamentos sérios.

	Escolaridade	Estado Civil	Quantidade_EES
2	Fundamental	Solteiro	331
10	Pós-Graduação	Solteiro	335
13	Superior	Divorciado	280
6	Médio	Solteiro	346
1	Fundamental	Divorciado	305

- **Gráfico: Mostrar a relação da quantidade de pessoas pela sua escolaridade e estados civil.**



Esse indicador visa mostrar as variações entre a quantidade de pessoas que possuem escolaridade e que pertencem a qual estado civil. Um indicador social relevante e comportamental.

Conclusão

Com base na análise detalhada dos dados, é possível concluir os seguintes pontos:

- **Distribuição Populacional Incongruente**: A análise dos dados revela que certas características populacionais, como a distribuição de cores de pele e tipos de cabelo, não correspondem às expectativas regionais. Por exemplo, há uma quantidade desproporcionalmente alta de pessoas ruivas em regiões onde isso é raro.
- **Baixa Variação em Médias Salariais**: As médias salariais apresentadas no dataset mostram pouca variação, independentemente de faixa etária, nível de escolaridade ou região de moradia. Esta baixa variação é estatisticamente improvável e sugere problemas na coleta ou geração dos dados.
- **Correlação Entre Variáveis**: O mapa de correlação confirma que há baixíssima correlação entre as variáveis analisadas, exceto pela relação esperada entre IMC, peso e altura. A falta de correlação significativa entre as outras variáveis reforça a ideia de que os dados podem não ser confiáveis.
- **Qualidade dos Dados**: As inconsistências e incongruências observadas indicam que os dados podem ter sido coletados de forma inadequada ou que podem ser fictícios. Isso compromete a validade das análises e qualquer inferência baseada nesses dados.

Dito isto, temos as seguintes recomendações:

- **Revisão do Processo de Coleta de Dados**: Recomenda-se uma revisão completa do processo de coleta de dados para garantir que os dados sejam representativos e precisos.
- **Verificação da Integridade dos Dados**: É crucial verificar a integridade dos dados e, se necessário, realizar uma limpeza para remover ou corrigir entradas errôneas.
- **Reanálise com Dados Confiáveis**: Uma vez garantida a qualidade dos dados, novas análises devem ser conduzidas para obter insights precisos e acionáveis.

Essas ações são essenciais para assegurar que as análises futuras sejam baseadas em dados sólidos e representativos, proporcionando uma base confiável para a tomada de decisões estratégicas.