Relatório Projeto Interdisciplinar - Ciência de Dados

Empregabilidade na área de dados

Nomes:	RA:
Beatriz de Castilho Ferreira	23024947
Giulia Nogueira Lopes de Sá	23024383
Lara Marina de Oliveira	23024708
Pedro Henrique Dangelo dos Reis de Oliveira	23024777

Sumário

Objetivos do projeto

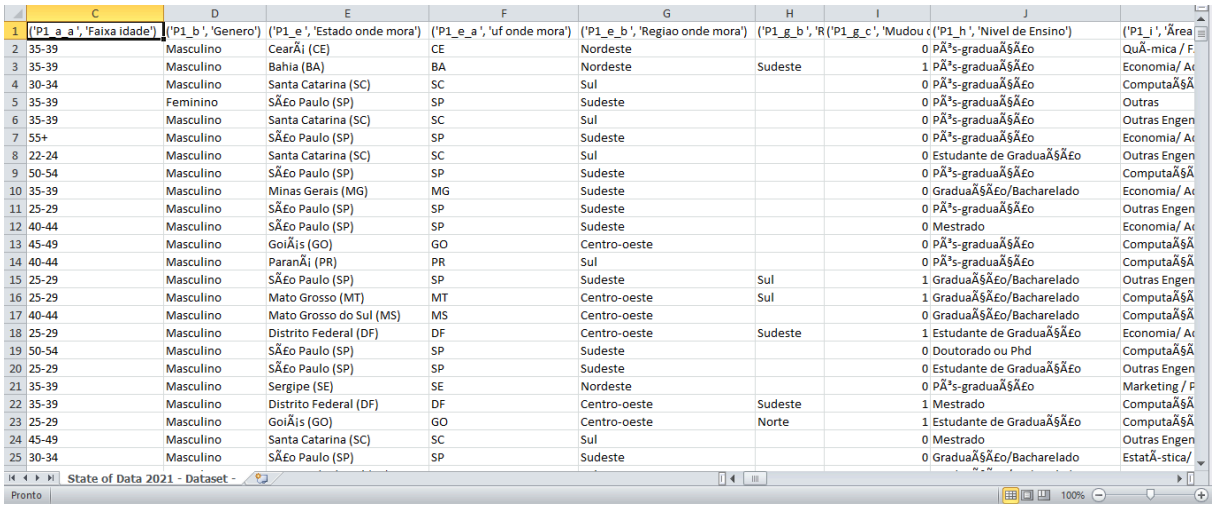
Nos últimos anos, a área de dados tem se consolidado como um dos setores mais dinâmicos e essenciais para o desenvolvimento de diversas indústrias. Profissionais especializados em ciência de dados, engenharia de dados, análise de dados e outras funções correlacionadas estão cada vez mais demandadas, refletindo em uma crescente digitalização e uma busca por soluções baseadas em dados para tomada de decisões. No entanto, além de observar as competências técnicas dos profissionais dessa área, é crucial compreender fatores sociais e demográficos que influenciam a empregabilidade e o crescimento da carreira dentro do setor.
O projeto de Ciência de Dados do grupo consiste em analisar a diversidade e as tendências do mercado de TI, ligado à área de Dados no Brasil, focando em como as vagas de emprego estão sendo distribuídas entre diferentes grupos, como por exemplo, gênero, raça, e deficiência. Além disso, queremos filtrar quais tecnologias estão sendo mais demandadas e quais níveis de senioridade estão sendo mais buscados.
O objetivo geral do nosso projeto é gerar análises eficientes para entender como as empresas filtram candidatos, gerando assim maior entendimento na hora dos processos seletivos. Além disso, queremos saber como esses profissionais desempenham seu trabalho, seja por senioridade, horas trabalhadas e entre outros aspectos.

Massa de dados e API

Nossos dados foram coletados na pesquisa da State of Data Brazil© do site Kaggle, da comunidade Data Hackers em Abril 2024, de cunho público, com o conteúdo podendo ser utilizado para fins não comerciais, desde que a fonte seja devidamente creditada. A metodologia adotada pela State of Data Brazil proporcionou uma análise minuciosa e precisa no mercado de dados no Brasil.
Para a análise nas entregas solicitadas, utilizamos dados referentes aos anos de 2023, 2022 e 2021, porém, a única devidamente tratada para gerar nosso dashboard foi a mais recente, a fim de proporcionar um estudo mais aproximado da atualidade.
Já para a API, construímos uma API própria para analisarmos os dados extraídos da pesquisa e construirmos um dashboard interativo.
Toda a base de dados e API está disponível para visualização no GitHub do grupo, no link https://github.com/2024-2-NCC4/Projeto4.

Estrutura gerada (Banco de Dados)

Dos dados coletados, geramos uma única tabela em formato *.CSV* com aproximadamente 5.200 linhas e 350 colunas. Desse total, foram utilizadas aproximadamente 33 colunas para abordarmos os pontos mais relevantes para nossa análise.



('P1_a_a ', 'Faixa idade')	('P1_b ', 'Genero')	('P1_e ', 'Estado onde mora')	('P1_e_a ', 'uf onde mora')	('P1_e_b ', 'Regiao onde mora')	('P1_g_b ', 'R	('P1_g_c ', 'Mudou c	('P1_h ', 'Nivel de Ensino')	('P1_i ', 'Área
35-39	Masculino	CearÃ¡ (CE)	CE	Nordeste		0	PÃ³s-graduaÃ§Ã£o	QuÃ-mica / F
35-39	Masculino	Bahia (BA)	BA	Nordeste	Sudeste	1	PÃ³s-graduaÃ§Ã£o	Economia/ Ac
30-34	Masculino	Santa Catarina (SC)	SC	Sul		0	PÃ³s-graduaÃ§Ã£o	ComputaÃ§Ã
35-39	Feminino	SÃ£o Paulo (SP)	SP	Sudeste		0	PÃ³s-graduaÃ§Ã£o	Outras
35-39	Masculino	Santa Catarina (SC)	SC	Sul		0	PÃ³s-graduaÃ§Ã£o	Outras Engen
55+	Masculino	SÃ£o Paulo (SP)	SP	Sudeste		0	PÃ³s-graduaÃ§Ã£o	Economia/ Ac
22-24	Masculino	Santa Catarina (SC)	SC	Sul		0	Estudante de GraduaÃ§Ã£o	Outras Engen
50-54	Masculino	SÃ£o Paulo (SP)	SP	Sudeste		0	PÃ³s-graduaÃ§Ã£o	ComputaÃ§Ã
35-39	Masculino	Minas Gerais (MG)	MG	Sudeste		0	GraduaÃ§Ã£o/Bacharelado	Economia/ Ac
25-29	Masculino	SÃ£o Paulo (SP)	SP	Sudeste		0	PÃ³s-graduaÃ§Ã£o	Outras Engen
40-44	Masculino	SÃ£o Paulo (SP)	SP	Sudeste		0	Mestrado	Economia/ Ac
45-49	Masculino	GoiÃ¡s (GO)	GO	Centro-oeste		0	PÃ³s-graduaÃ§Ã£o	ComputaÃ§Ã
40-44	Masculino	ParanÃ¡ (PR)	PR	Sul		0	PÃ³s-graduaÃ§Ã£o	ComputaÃ§Ã
25-29	Masculino	SÃ£o Paulo (SP)	SP	Sudeste	Sul	1	GraduaÃ§Ã£o/Bacharelado	Outras Engen
25-29	Masculino	Mato Grosso (MT)	MT	Centro-oeste	Sul	1	GraduaÃ§Ã£o/Bacharelado	ComputaÃ§Ã
40-44	Masculino	Mato Grosso do Sul (MS)	MS	Centro-oeste		0	GraduaÃ§Ã£o/Bacharelado	ComputaÃ§Ã
25-29	Masculino	Distrito Federal (DF)	DF	Centro-oeste	Sudeste	1	Estudante de GraduaÃ§Ã£o	Economia/ Ac
50-54	Masculino	SÃ£o Paulo (SP)	SP	Sudeste		0	Doutorado ou Phd	ComputaÃ§Ã
25-29	Masculino	SÃ£o Paulo (SP)	SP	Sudeste		0	Estudante de GraduaÃ§Ã£o	Outras Engen
35-39	Masculino	Sergipe (SE)	SE	Nordeste		0	PÃ³s-graduaÃ§Ã£o	Marketing / P
35-39	Masculino	Distrito Federal (DF)	DF	Centro-oeste	Sudeste	1	Mestrado	ComputaÃ§Ã
25-29	Masculino	GoiÃ¡s (GO)	GO	Centro-oeste	Norte	1	Estudante de GraduaÃ§Ã£o	ComputaÃ§Ã
45-49	Masculino	Santa Catarina (SC)	SC	Sul		0	Mestrado	Outras Engen
30-34	Masculino	SÃ£o Paulo (SP)	SP	Sudeste		0	GraduaÃ§Ã£o/Bacharelado	EstatÃ-stica/

State of Data 2021 - Dataset -

Pronto

100%

Para analisarmos esses dados no Colab e gerarmos nossos gráficos, foram utilizadas as bibliotecas do Python, sendo elas, matplotlib, numpy e pandas. Com auxílio delas, geramos gráficos tanto de barra quanto de pizza, como exemplo:

```python
# DITRIBUIÇÃO POR FAIXA SALARIAL (GRÁFICO DE BARRAA)
import matplotlib.pyplot as plt
import numpy as np

def faixa_salarial_barras():
    # Contanto frequência
    num_faixa_salarial = df_dados_2023['dsc_faixa_salarial'].value_counts()
    num_faixa_salarial = num_faixa_salarial[num_faixa_salarial > 0]
    num_faixa_salarial = num_faixa_salarial.sort_values(ascending=False)  # ordenando decrescente
    aspect_labels = num_faixa_salarial.index # (índices)
    # Gráfico
    plt.figure(figsize=(10, 6))
    cmap = plt.cm.get_cmap('plasma')
    colors = cmap(np.linspace(0, 1, len(num_faixa_salarial)))
    plt.bar(aspect_labels, num_faixa_salarial, color=colors[:len(num_faixa_salarial)])
    plt.xticks(rotation=45, ha='right')
    plt.ylabel('Contagem')
    plt.title('Distribuição das Faixas Salariais')
    plt.tight_layout()
    plt.show()

# Testando
if __name__ == "__main__":
    faixa_salarial_barras()
```

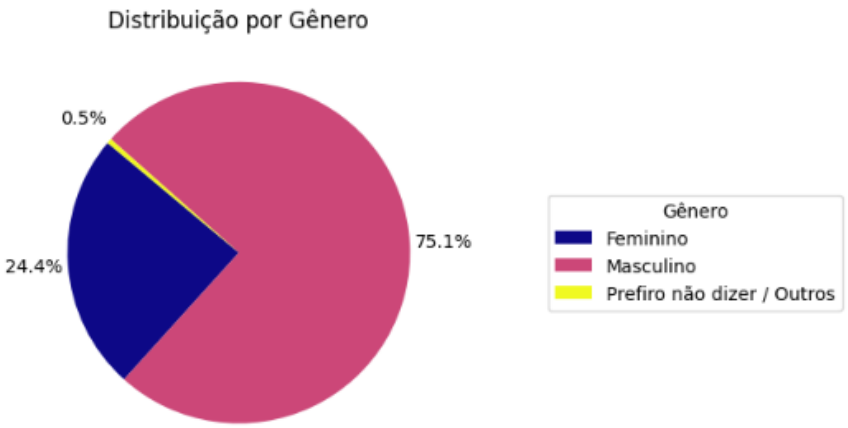
```python
# DISTRIBUIÇÃO POR FAIXA ETÁRIA
import matplotlib.pyplot as plt
def distribuicao_etaria():
  contagem_idade = df_dados_2023['dsc_faixa_idade'].value_counts()
  contagem_idade = contagem_idade.groupby(lambda x: x ).sum()
  plt.figure(figsize=(7, 7))
  cmap = plt.cm.get_cmap('plasma')
  colors = cmap(np.linspace(0, 1, len(contagem_idade)))
  plt.pie(contagem_idade, labels=None, autopct='%1.1f%%', startangle=140, colors=colors, pctdistance=1.2)
  plt.legend(contagem_idade.index, title="Faixa", loc="center left", bbox_to_anchor=(1,  0.5))
  plt.title('Distribuição Etária')
  plt.tight_layout()
  plt.show()

if __name__ == "__main__":
  distribuicao_etaria()
```

Interpretação das análises e gráficos

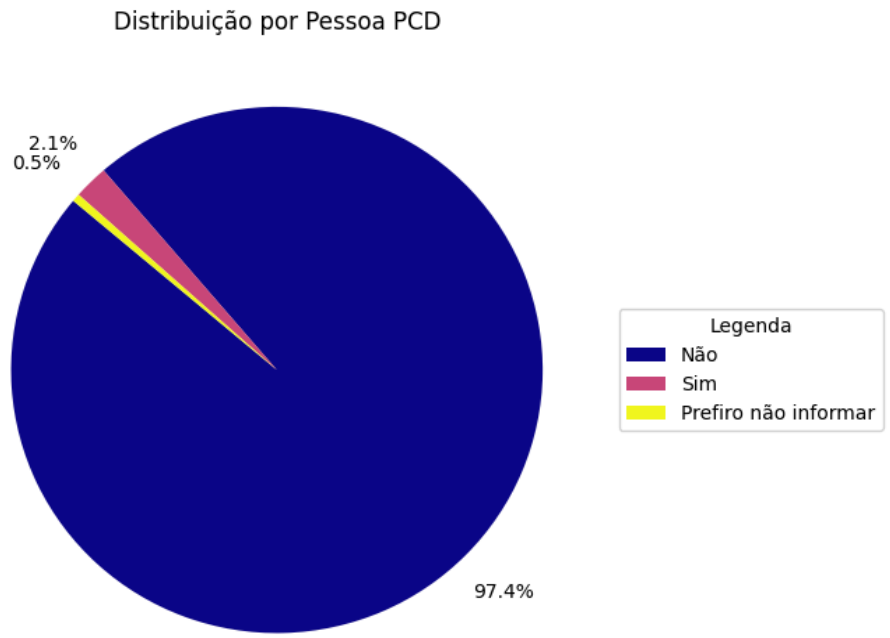
Inicialmente, precisamos de uma análise para entender a diversidade de pessoas que trabalham na área de dados.

1. Distribuição por gênero



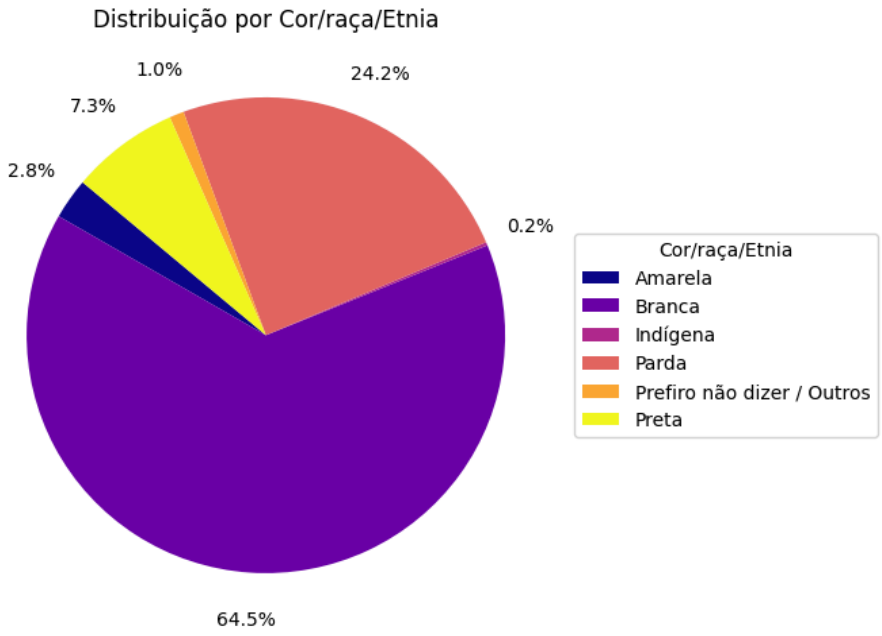
No gráfico gerado podemos notar que 75,1% dos entrevistados são do gênero masculino, o que mostra a realidade onde homens têm mais destaque na carreira da tecnologia, sendo eles ¾ dos empregados. Essa análise reflete em contexto histórico e estrutural, enraizado em problemas culturais, religiosos e sociais ainda existentes na sociedade.

2. Distribuição por pessoa PCD



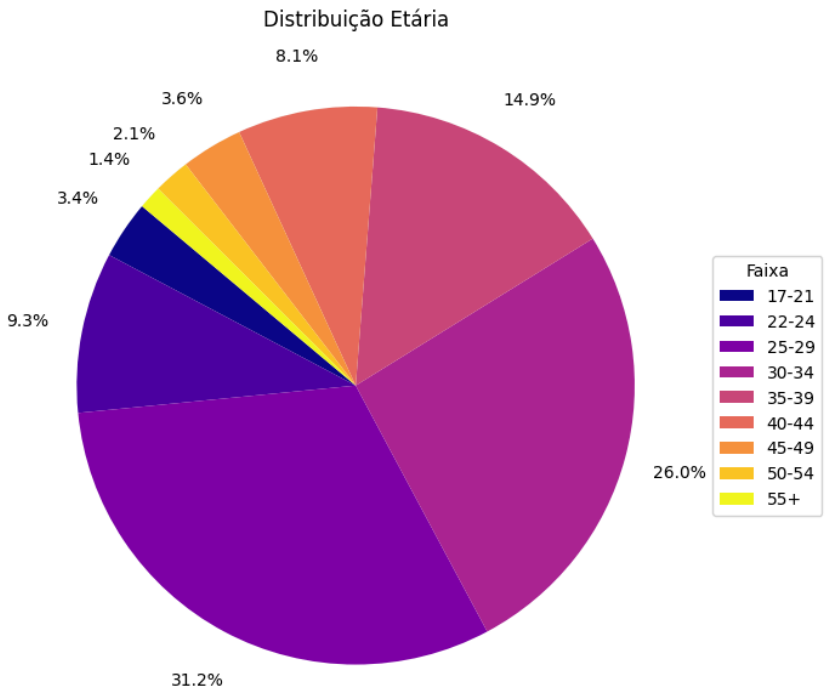
Observando a distribuição de pessoas PCD dentro da área de dados, podemos analisar uma ilusão quase imperceptível, onde apenas 2,1% dos entrevistados possuem algum tipo de deficiência, seja ela visual, mental, física e entre outras. Isso mostra para nós que ainda existe muito esforço a ser feito para que essas pessoas possam ser mais incluídas dentro do mercado da tecnologia.

3. Distribuição por Cor/ Raça e Etnia



Ainda que existam grandes esforços gerados pelas empresas para incluir todas as etnias dentro do mercado de trabalho, o número de pessoas pretas, pardas, amarelas e indígenas não soma o total de pessoas brancas na área da tecnologia. Pessoas brancas ainda protagonizam mais da metade dos participantes, sendo eles 64,5% do total. Esses números refletem não só um contexto atual, mas sim um contexto histórico.

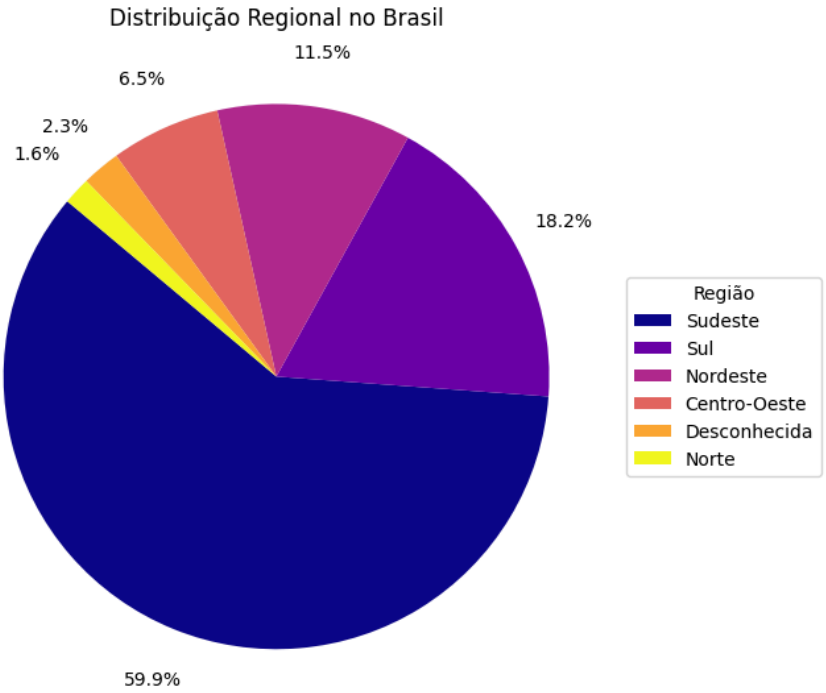
4. Distribuição Etária



A tecnologia é um fato muito atual na nossa sociedade, o que reflete em uma faixa etária mais jovem dentro do mercado de trabalho. Além disso, as empresas buscam talentos que tenham muita disposição no trabalho, o que incentiva os jovens a enfrentarem desafios diários e motivação. No gráfico gerado pela pesquisa, conseguimos observar uma idade predominantemente entre 25 e 35 anos.

Com isso, analisamos a necessidade de pessoas mais velhas se manterem atualizadas em relação ao mercado de trabalho no ramo da tecnologia, para assim conseguirem mais oportunidades de emprego e ainda mais conexão com o mundo atual.

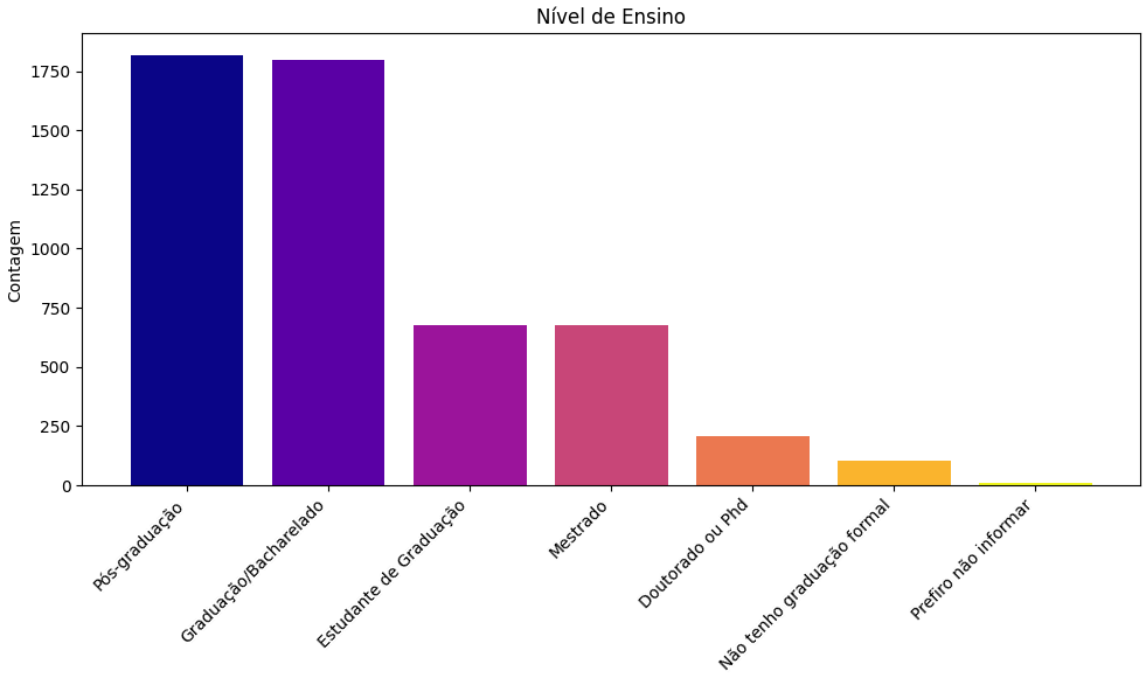
5. Distribuição Regional



Observa-se com os dados coletados que quase 60% dos entrevistados trabalham na região sudeste do país. Isso nos mostra um fato do Brasil, onde os grandes centros econômicos e Big Techs se encontram em cidades como São Paulo e Curitiba, por exemplo, conhecidas como regiões mais desenvolvidas.

Isso infelizmente reflete uma alta concentração em uma única região, fazendo com que muitas pessoas tenham que se deslocar para esses lugares em busca de oportunidades de trabalho.

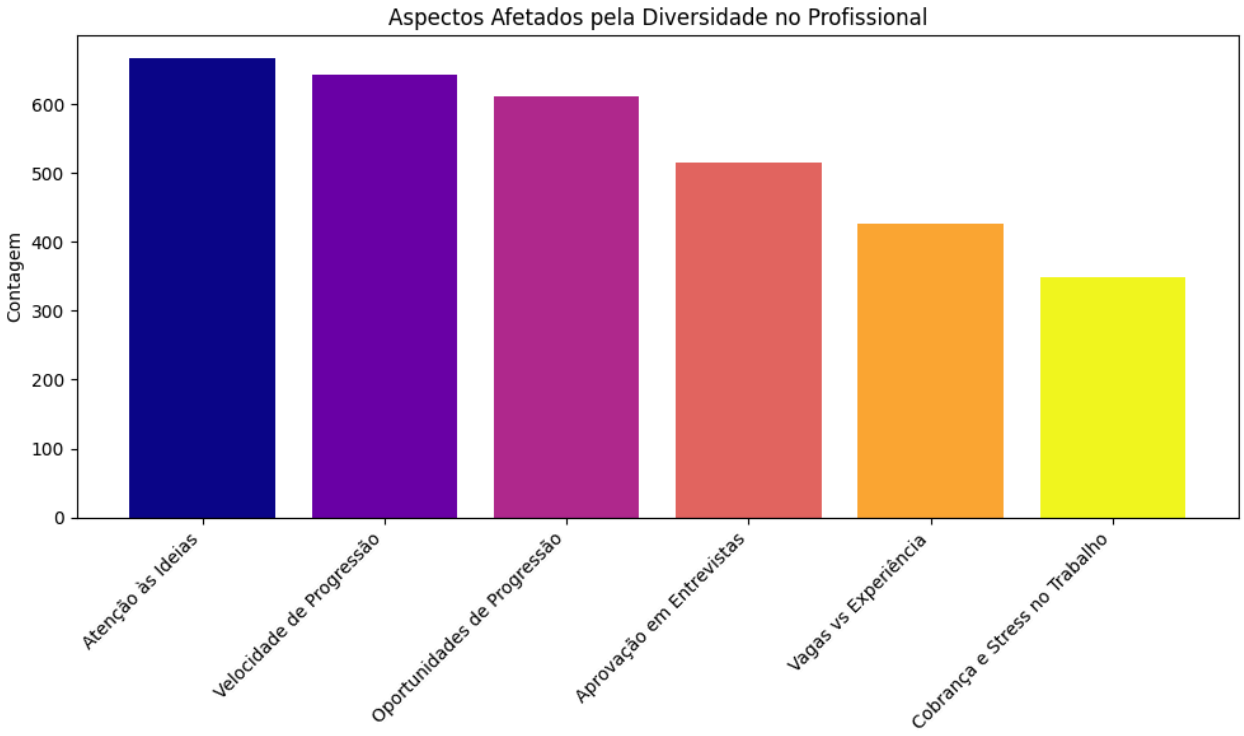
6. Nível de Escolaridade



O nível de escolaridade dos entrevistados, não mostra apenas algo social, mas sim uma grande oportunidade para pessoas que querem futuramente entrar na área de dados. Nossos gráficos mostram que, tendo uma graduação ou pós-graduação, já é uma grande oportunidade para entrar no mercado da tecnologia.

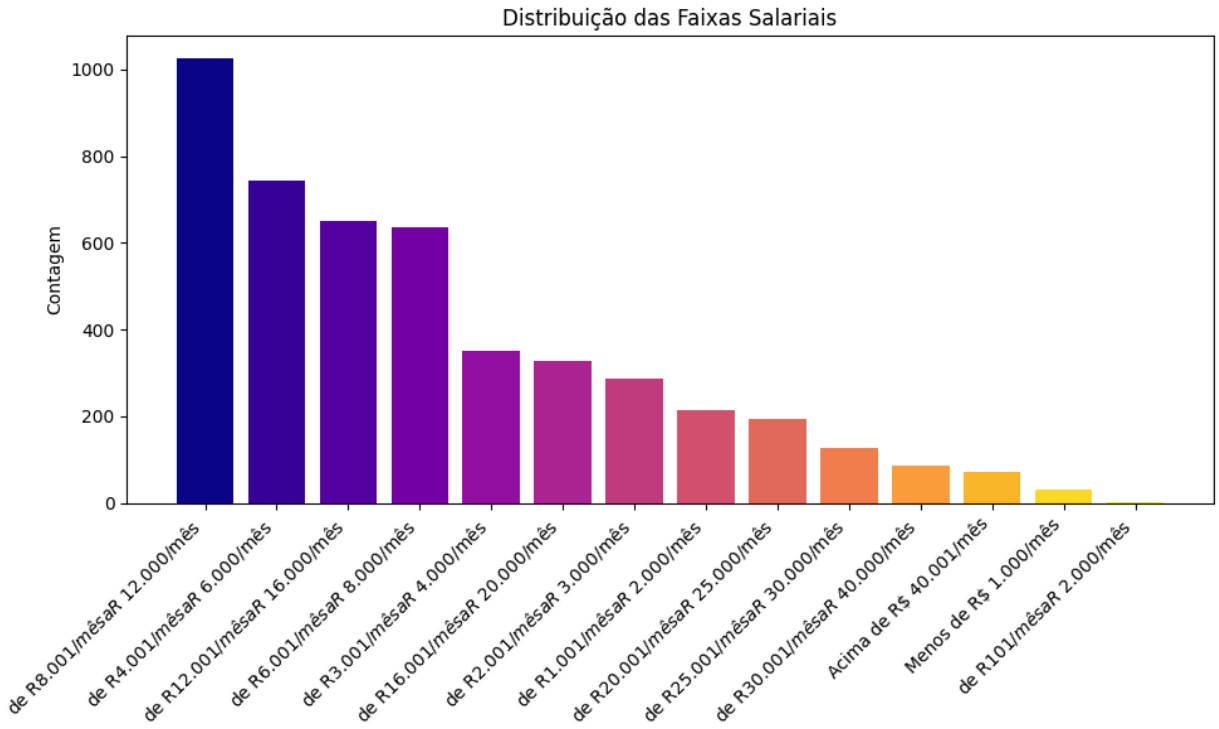
Dando sequência às análises, realizamos análises mais técnicas sobre o mercado e seus desdobramentos teóricos.

1. Aspectos Afetados pela Diversidade no âmbito profissional.



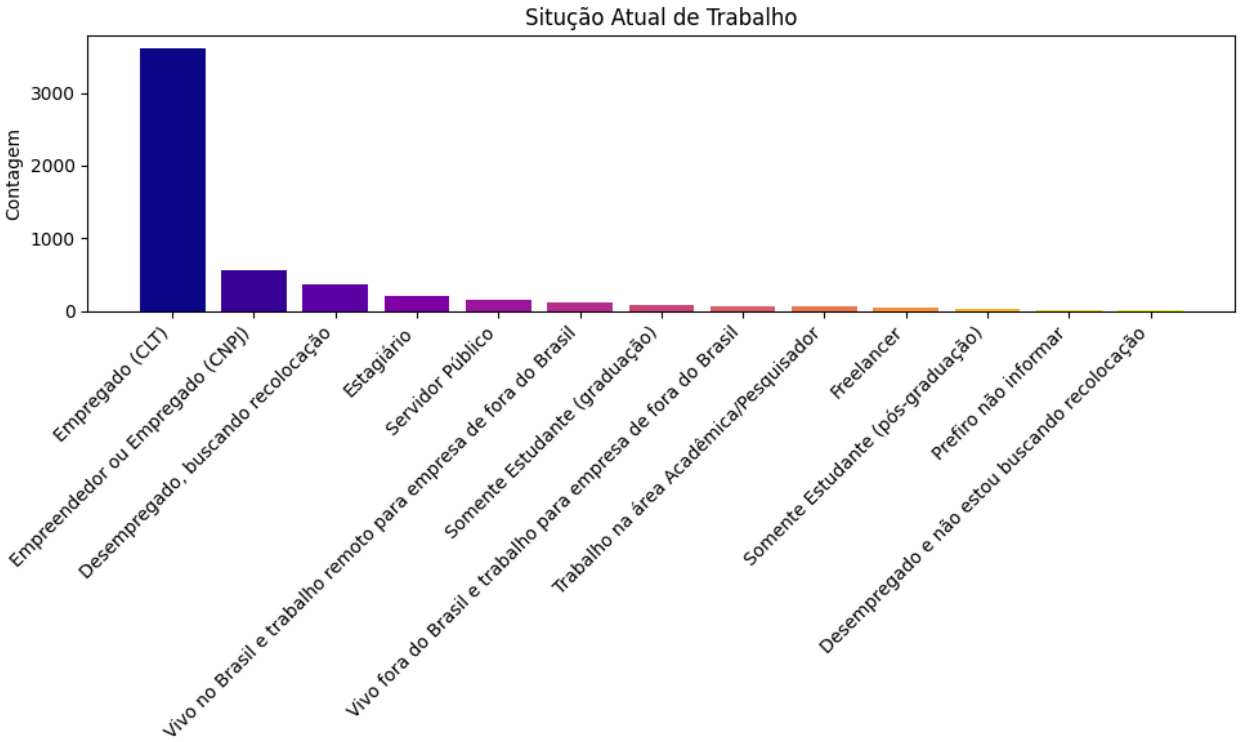
Os aspectos afetados pela diversidade são fatores que pessoas pertencentes aos grupos de diversidade citados acima (pessoas portadoras de deficiência, mulheres ou negros) sentem falta no ambiente de trabalho. Uma análise clara nos mostra que existem muitos aspectos onde os funcionários que pertencem à minoria reclamam, sendo como maiores ofensores a falta de atenção às ideias dadas, a velocidade de progressão de carreira e oportunidades de progressão de carreira.

2. Distribuição das Faixas Salariais



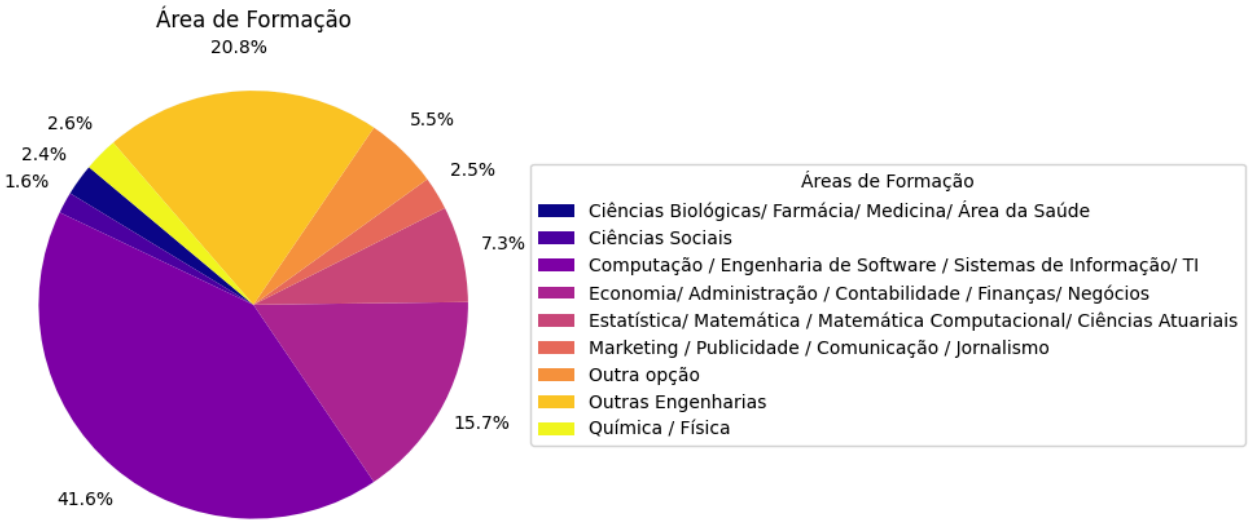
De acordo com os nossos dados, podemos observar que aproximadamente 1.000 das 5.200 pessoas entrevistadas, possuem uma faixa salarial entre 8.000,00 a 12.000,00 reais por mês, o que nos mostra que a área de dados é muito valorizada dentro do mercado de trabalho.

3. Situação Atual de Trabalho



A análise de situação de trabalho é importante para representar aonde e qual é o estilo mais procurado pelo mercado. De acordo com os dados obtidos, foi possível observar que a contratação CLT ainda encontra-se em predominância. Sendo seguida bem distante por Empreendedor ou Empregado(CNPJ) e Desempregados. Pode-se observar uma baixa concentração nas modalidades Freelancer e Somente Estudante (Pós-Graduação).

4. Área de Formação



Os dados coletados apontam uma grande amplitude de cursos para profissionais da área, isso pode se dar por muitos fatores diferentes como a importância do uso de dados em outras áreas. Mas, é notória a hegemonia dos cursos focados em tecnologia, como Computação, Engenharia de Software, entre outros.

Técnicas de Inferência Aplicadas

Para fazermos uma predição usamos nossos dados de faixa salarial. Utilizamos o R para calcularmos as predições a seguir:

```
> #ENTREGA 03
>
> # Usar a mesma planilha e colunas
> # Calcular a média, desvio padrão, e tamanho da amostra para "Salario_Convertido"
> media_salario <- mean(dados_pesquisa_2023$`Salario_Convertido`, na.rm = TRUE)
> desvio_padrao_salario <- sd(dados_pesquisa_2023$`Salario_Convertido`, na.rm = TRUE)
> n <- length(na.omit(dados_pesquisa_2023$`Salario_Convertido`))
>
> # Nível de confiança (por exemplo, 95%)
> nivel_confianca <- 0.95
> z <- qnorm(1 - (1 - nivel_confianca) / 2)
>
> # Calcular o erro padrão
> erro_padrao <- desvio_padrao_salario / sqrt(n)
>
> # Calcular o erro (Epsilon)
> epsilon <- z * erro_padrao
>
> # Intervalo de confiança
> limite_inferior <- media_salario - epsilon
> limite_superior <- media_salario + epsilon
>
> # Calcular o erro usando a largura do intervalo
> epsilon_calculado <- (limite_superior - limite_inferior) / 2
>
> # Exibir resultados
> cat("Intervalo de Confiança para Salario_Convertido:", limite_inferior, "a", limite_superior, "\n")
Intervalo de Confiança para Salario_Convertido: 8617.899 a 9011.442
> cat("Erro (Epsilon) calculado pela largura do intervalo:", epsilon_calculado, "\n")
Erro (Epsilon) calculado pela largura do intervalo: 196.7715
>
```

Com os dados amostrais coletados pelo grupo, obtemos o intervalo de confiança e podemos confirmar que, pegando uma quantidade de dados consideráveis, os valores estarão entre uma média de R$8.617 e R$9.011. Junto a análise do intervalo de confiança, também obtivemos um erro amostral de R$196, ou seja, a média salarial (R$8.815) pode variar para mais ou para menos deste valor.

Outra técnica de inferência que utilizamos no projeto foi o teste de hipótese, para compararmos os dados que temos e vermos se eles possuem relação entre si. Para essa análise, cruzamos nossos dados de gênero e nível de ensino.

```
> # Contar as frequências de cada categoria por gênero
> frequencias <- table(dados_pesquisa_2023$dsc_genero, dados_pesquisa_2023$dsc_nivel_ensino)
>
> # Exibir a tabela de frequências
> print(frequencias)

                    Doutorado ou Phd Estudante de Graduação Graduação/Bacharelado Mestrado
  Feminino                        72                    132                   411      196
  Masculino                      137                    542                  1376      476
  Outro                            0                      3                     5        0
  Prefiro não informar             1                      1                     6        4

                    Não tenho graduação formal Pós-graduação Prefiro não informar
  Feminino                                  20           462                     0
  Masculino                                 85          1351                     8
  Outro                                      0             1                     0
  Prefiro não informar                       0             4                     0
>
```

```
>
> # Realizar o teste qui-quadrado de independência usando a tabela de frequências
> teste_qui2 <- chisq.test(frequencias)
Warning message:
In chisq.test(frequencias) :
  Aproximação do qui-quadrado pode estar incorreta
>
> # Exibir os resultados
> cat("Estatística Qui-Quadrado:", teste_qui2$statistic, "\n")
Estatística Qui-Quadrado: 46.00697
> cat("Valor-p:", teste_qui2$p.value, "\n")
Valor-p: 0.000295979
>
```

```
> # Interpretação do resultado
> nivel_significancia <- 0.05
> if (teste_qui2$p.value < nivel_significancia) {
+   cat("Resultado: Existe uma relação significativa entre o gênero e o nível de formação.\n")
+ } else {
+   cat("Resultado: Não existe uma relação significativa entre o gênero e o nível de formação de exata
s.\n")
+ }
Resultado: Existe uma relação significativa entre o gênero e o nível de formação.
>
```

De acordo com os dados obtidos, podemos concluir que os homens possuem um maior nível de escolaridade em relação às mulheres, ressaltando uma relação significativa entre os gêneros e nível de formação. Isso provavelmente se reflete em diversos fatores sociais já existentes na sociedade atual.

Conclusão

Retornando ao objetivo inicial do projeto que era analisar o atual mercado de trabalho na área de dados para fins tanto dos funcionários quanto das empresas, obtivemos conclusões relevantes como, por exemplo, a predominância de homens brancos dentro do mercado. Durante o desenvolvimento do projeto utilizamos técnicas aplicadas em sala e aprofundamos diversos conhecimentos.