



Pontifícia Universidade Católica do Rio de Janeiro

Pós-Graduação em Ciência de Dados e Analytics

Sprint de Engenharia de Dados

Carlos Eduardo Azevedo Costinhas da Silva

MVP - Criação de um pipeline de dados utilizando a plataforma Google Cloud para carga, transformação, armazenamento e análise de dados provenientes da pesquisa State of Data 2022.

Projeto disponível em <https://github.com/costinhas/puc-rio-data-engineering>

Rio de Janeiro
Setembro de 2023

Resumo

Este projeto foi desenvolvido para aplicar os conhecimentos adquiridos no módulo de engenharia de dados do curso de especialização em Ciência de Dados e Analytics da PUC-Rio. O escopo consiste em um MVP (produto mínimo viável) de um pipeline de dados que inclui a busca, coleta, modelagem, carga e análise de dados utilizando tecnologias da plataforma *Google Cloud*.

Os dados utilizados neste projeto foram obtidos através da pesquisa *State of Data Brazil 2022*, que divulgou um panorama sobre o mercado de trabalho brasileiro na área de dados no ano de 2022. O estudo, de autoria da comunidade *Data Hackers* e da consultoria *Bain & Company*, foi publicado na plataforma *Kaggle* e está disponível para consulta pública em <https://www.kaggle.com/datasets/datahackers/state-of-data-2022>.

Objetivos do projeto

Realizar uma análise sobre o mercado de trabalho brasileiro na área de dados e consolidar informações que possam auxiliar pessoas que desejam iniciar sua carreira ou realizar uma transição de carreira para a área de dados. As seguintes questões foram avaliadas:

- 1) Que tipo de educação formal é necessária para trabalhar na área de dados no Brasil?
- 2) Quais são os cargos ou funções mais comuns no mercado brasileiro na área de dados?
- 3) Qual é a média salarial dos profissionais de dados no Brasil?
- 4) Os profissionais de dados no Brasil estão satisfeitos com seus empregos atuais?
- 5) Quais são as principais tecnologias em uso atualmente na área de dados no Brasil?

Desenvolvimento do projeto

O projeto foi desenvolvido seguindo etapas bem definidas para construção de um pipeline de dados: busca, coleta, modelagem, carga, transformação e análise dos dados. A plataforma em nuvem selecionada para execução deste projeto foi a *Google Cloud*, que fornece soluções adequadas para todas as etapas necessárias. As seguintes ferramentas da plataforma foram utilizadas:

- **Google Cloud Storage**, para armazenamento dos dados brutos, em seu formato original;
- **Google Cloud Data Fusion**, para gerenciar o pipeline de dados e permitir atividades de ETL (*Extract, Transform, and Load*);
- **Google BigQuery**, como banco de dados, para armazenar os dados transformados de forma estruturada;
- **Google Cloud Dataplex**, para criação e atualização do catálogo de dados.

Os detalhes de cada etapa serão descritos a seguir.

Etapa 1: Busca dos dados

Atualmente, existem diversas plataformas que disponibilizam bases de dados gratuitas na Internet. Este projeto utilizou como fonte a plataforma *Kaggle* (<https://www.kaggle.com/>), não só por sua popularidade entre estudantes e profissionais da área de dados, mas também pela quantidade e diversidade de bases de dados disponíveis.

Dentre as diversas bases de dados disponíveis na plataforma, a que mais se adequa ao objetivo deste projeto foi disponibilizada pela pesquisa *State of Data Brazil 2022*, de autoria da comunidade Data Hackers e da consultoria Bain & Company, que divulgou um panorama sobre o mercado de trabalho brasileiro na área de dados no ano de 2022.

A pesquisa, disponível em <https://www.kaggle.com/datasets/datahackers/state-of-data-2022>, foi realizada entre 10 de outubro e 28 de novembro de 2022 através de um questionário online, coletou informações de 4.271 pessoas de todo o Brasil e reuniu indicadores relacionados ao perfil demográfico, formação, atuação no setor, remuneração, rotatividade e fatores de satisfação no ambiente de trabalho.

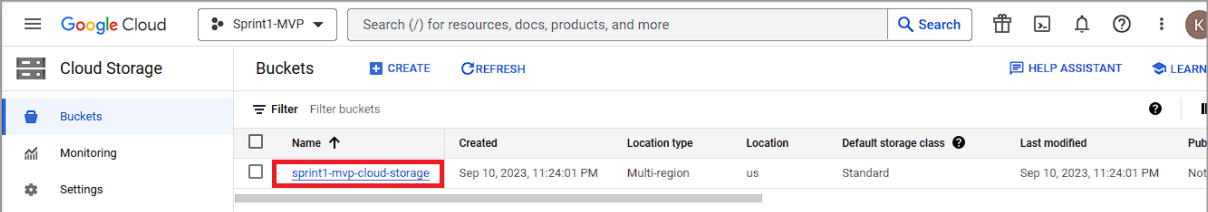
Etapa 2: Coleta dos dados

Os dados foram disponibilizados em um único *dataset*, que foi anonimizado pelos autores da pesquisa para garantir a privacidade dos respondentes. Os dados que se diferenciam drasticamente de todos os outros (*outliers*) foram removidos antes da divulgação dos dados, para evitar qualquer forma de identificar um entrevistado. Nenhuma etapa adicional de anonimização de dados ou remoção de *outliers* foi realizada neste projeto, tendo em vista que os dados já foram disponibilizados anonimizados.

O *dataset* foi disponibilizado em um arquivo CSV de aproximadamente 10 MB, em <https://www.kaggle.com/datasets/datahackers/state-of-data-2022/download?datasetVersionNumber=1> através de um arquivo compactado com extensão “.zip”.

Para simplificar o escopo deste MVP, o *dataset* foi extraído manualmente da plataforma Kaggle e salvo no repositório do GitHub deste projeto, para que seja acessado sem a necessidade de login na plataforma. O arquivo final pode ser acessado através do caminho https://github.com/costinhas/puc-rio-data-engineering/raw/main/State_of_Data_Brazil_2022.zip.

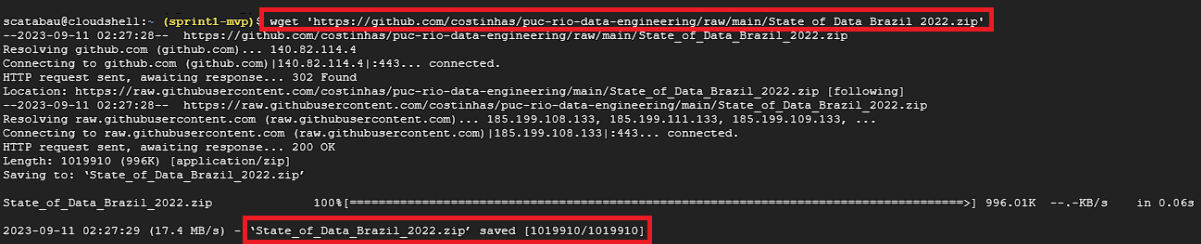
Em seguida, foi criado um bucket denominado *sprint1-mvp-cloud-storage* no Google Cloud Storage, que será o repositório utilizado:



The screenshot shows the Google Cloud Storage interface. On the left, there's a sidebar with 'Cloud Storage' selected. The main area shows a list of buckets. One bucket, 'sprint1-mvp-cloud-storage', is highlighted with a red box. The table has columns: Name, Created, Location type, Location, Default storage class, Last modified, and Pub. The bucket 'sprint1-mvp-cloud-storage' was created on Sep 10, 2023, at 11:24:01 PM, is Multi-region, located in 'us', with a Standard storage class, and was last modified on the same date and time.

Name	Created	Location type	Location	Default storage class	Last modified	Pub
sprint1-mvp-cloud-storage	Sep 10, 2023, 11:24:01 PM	Multi-region	us	Standard	Sep 10, 2023, 11:24:01 PM	Not

Através da interface de shell, o arquivo foi extraído do GitHub, descompactado e copiado para o diretório de input no Cloud Storage:



The screenshot shows a terminal window with the following commands and output:

```
scatsbau@cloudshell: (sprint1-mvp) $ wget 'https://github.com/costinhas/puc-rio-data-engineering/raw/main/State_of_Data_Brazil_2022.zip'
--2023-09-11 02:27:28-- https://github.com/costinhas/puc-rio-data-engineering/raw/main/State_of_Data_Brazil_2022.zip
Resolving github.com (github.com)... 140.82.114.4
Connecting to github.com (github.com)[140.82.114.4]:443... connected.
HTTP request sent, awaiting response... 302 Found
Location: https://raw.githubusercontent.com/costinhas/puc-rio-data-engineering/main/State_of_Data_Brazil_2022.zip [following]
--2023-09-11 02:27:28-- https://raw.githubusercontent.com/costinhas/puc-rio-data-engineering/main/State_of_Data_Brazil_2022.zip
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.108.133, 185.199.111.133, 185.199.109.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)[185.199.108.133]:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 1019910 (996K) [application/zip]
Saving to: 'State_of_Data_Brazil_2022.zip'

State_of_Data_Brazil_2022.zip 100%[=====] 996.01K --.-KB/s in 0.06s

2023-09-11 02:27:29 (17.4 MB/s) - 'State_of_Data_Brazil_2022.zip' saved [1019910/1019910]
```

Download do dataset, do repositório no GitHub para o Cloud Storage

Etapa 3: Modelagem dos dados

O questionário de coleta dos dados que deu origem ao *dataset* utilizado foi dividido em 8 seções, cada uma delas contendo suas perguntas e opções de respostas:

- Parte 1 - Dados demográficos
- Parte 2 - Dados sobre carreira
- Parte 3 - Desafios dos gestores de times de dados
- Parte 4 - Conhecimentos na área de dados
- Parte 5 - Objetivos na área de dados
- Parte 6 - Conhecimentos em Engenharia de Dados/DE
- Parte 7 - Conhecimentos em Análise de Dados/DA
- Parte 8 - Conhecimentos em Ciências de Dados/DS

O *dataset* foi disponibilizado com apenas uma relação (ou “tabela flat”), com todos os atributos consolidados em colunas desta tabela. Por este motivo, durante a modelagem conceitual não foi criado nenhum diagrama de entidades e relacionamentos (DER), mas os detalhes da estrutura, domínio e restrições dos dados foram descritos nas seções a seguir.

Etapa 3.1: Nomenclatura das colunas do dataset

As colunas do *dataset* original são identificadas por uma tupla com dois elementos, para permitir identificar a quais perguntas cada coluna se refere. As perguntas cujas respostas são multi-valoradas ocupam mais de uma coluna na tabela, cada uma contendo uma das opções de resposta.

Para descrever esta estrutura, utilizaremos como referência as colunas (*'P3_b '*, *'Quais desses papéis/cargos fazem parte do time (ou chapter) de dados da sua empresa?'*) e (*'P3_b_1 '*, *'Analytics Engineer'*):

- O primeiro elemento da tupla é o identificador da pergunta, sendo uma combinação de Parte, Letra da pergunta e Número da opção escolhida. No exemplo acima, a coluna **“P3_b_1”**, corresponde à:
 - **Parte 3:** “Desafios dos gestores de times de dados”;
 - **Pergunta B:** “Quais desses papéis/cargos fazem parte do time (ou chapter) de dados da sua empresa?”;
 - **Opção 1:** “Analytics Engineer”
- O segundo elemento da tupla contém a descrição da pergunta à qual esta coluna se refere. Se esta coluna fizer referência a uma pergunta com várias respostas, este elemento fará referência à descrição da alternativa escolhida.

Para facilitar a interpretação e análise dos dados, todas as colunas foram renomeadas para títulos mais próximos à sua descrição. Os detalhes estão descritos na seção “Etapa 3.3: Catálogo de dados”.

Etapa 3.2: Detalhamento do esquema do *dataset*

Seguindo a abordagem *schema-on-read* de um Data Lake, os dados brutos foram armazenados na íntegra, no seu formato original (CSV), esquema original e estrutura descrita na seção anterior. Todas as transformações realizadas neste projeto foram feitas em uma instância dedicada, mantendo a fonte original sem alterações. Isto possibilitará análises futuras neste mesmo *dataset*, com transformações e agrupamentos distintos, conforme novas necessidades.

O *dataset* original contém 4271 linhas e 353 colunas. No entanto, nem todas as colunas foram necessárias para responder às perguntas descritas no objetivo deste projeto. Após uma análise em todos os atributos, foram selecionadas 119 colunas, detalhadas a seguir na seção “Etapa 3.3: Catálogo de dados”.

Etapa 3.3: Catálogo de dados

A ferramenta utilizada para gerenciar o catálogo de dados deste dataset foi o Dataplex, disponível na plataforma *Google Cloud*. A tabela foi catalogada com uma visão geral, e seu esquema foi detalhado conforme evidências a seguir:

Google Cloud

Sprint1-MVP

dataplex

Search

respostas

STAR

ATTACH TAGS

OPEN IN BIGQUERY

EXPLORE WITH LOOKER STUDIO

SCAN WITH SENSITIVE DATA PROTECTION

EXPLORE WITH SHEETS

Sprint1-MVP

us

state_of_data

DETAILS

SCHEMA

LINEAGE

DATA PROFILE

DATA QUALITY

BigQuery table details

Type

TABLE

System

BIGQUERY

Table type

BigQuery table

Creation time

Sep 11, 2023, 12:18:00 AM

Last modification time

Sep 11, 2023, 10:44:01 AM

Expiration time

Never

Location

us

Queried (Past 30 days)

173

Resource URL

sprint1-mvp.state_of_data.respostas

Labels

dataplex-d...:us-central1

Fully qualified name

bigquery:sprint1-mvp.state_of_data.respostas

Description

Edit in BigQuery

Overview

Dataset que armazena as repostas da pesquisa *State of Data Brazil 2022*, de autoria da comunidade Data Hackers e da consultoria Bain & Company, que divulgou um panorama sobre o mercado de trabalho brasileiro na área de dados no ano de 2022.

A pesquisa, disponível em <https://www.kaggle.com/datasets/datahackers/state-of-data-2022>, foi realizada entre 10 de outubro e 28 de novembro de 2022 através de um questionário online, coletou informações de 4.271 pessoas de todo o Brasil e reuniu indicadores relacionados ao perfil demográfico, formação, atuação no setor, remuneração, rotatividade e fatores de satisfação no ambiente de trabalho.

Os dados foram disponibilizados em um único *dataset*, que foi anonimizado pelos autores da pesquisa para garantir a privacidade dos respondentes. Os dados que se diferenciam drasticamente de todos os outros (*outliers*) foram removidos antes da divulgação dos dados, para evitar qualquer forma de identificar um entrevistado. Nenhuma etapa adicional de anonimização de dados ou remoção de *outliers* foi realizada neste projeto, tendo em vista que os dados já foram disponibilizados anonimizados.

Edit Overview

Tags (0)

You have no tags attached to this data asset.

ATTACH TAGS

Visão geral da tabela resposta, no módulo Dataplex.

Google Cloud Sprint1-MVP

Search (/) for resources, docs, products, and more

respostas

This is a partitioned table. [Learn more](#)

SCHEMA DETAILS PREVIEW LINEAGE DATA PROFILE DATA QUALITY

Filter Enter property name or value

Field name	Type	Mode	Key	Collation	Default Value	Policy Tags	Description
ID_RESPONDENTE	STRING	NULLABLE					Identificador único do respondente.
MORA_NO_BRASIL	BOOLEAN	NULLABLE					Identifica se o respondente mora no Brasil ou não. Valores possíveis: TRUE, FALSE
NIVEL_ENSINO	STRING	NULLABLE					Nível de ensino do respondente. Valores possíveis: Não tenho graduação formal, Estudante de Graduação, Graduação/Bacharelado, Pós-graduação, Mestrado, Doutorado ou PhD, Prefiro não informar.
AREA_FORMACAO	STRING	NULLABLE					Área de formação do respondente. Valores possíveis: Ciências Biológicas/ Farmácia/ Medicina/ Área da Saúde, Ciências Sociais, Computação / Engenharia de Software / Sistemas de Informação/ TI, Economia/ Administração / Contabilidade / Finanças/ Negócios, Estatística/ Matemática / Matemática Computacional/ Ciências Atuariais, Marketing / Publicidade / Comunicação / Jornalismo, Outras Engenharias, Química / Física, Outra opção.
CARGO_ATUAL_COMO_GESTOR	STRING	NULLABLE					Cargo atual do respondente, caso tenha respondido que é um gestor. Valores possíveis incluem uma lista com valores pré-definidos, e também texto livre através da opção "Outros": Chapter Lead , Diretor/VP , Especialista , Gerente/Head , Líder de Squad , Proprietário, Autônomo , Supervisor/Coordenador , Sócio ou C-level (CEO, CDO, CIO, CTO etc) , Team Leader/Tech Leader.
CARGO_ATUAL_COMO_NAO_GESTOR	STRING	NULLABLE					Cargo atual do respondente, caso tenha respondido que não é um gestor. Valores possíveis incluem uma lista com valores pré-definidos, e também texto livre através da opção "Outros": Analista de BI/BI Analyst , Analista de Dados/Data Analyst , Analista de Inteligência de Mercado/Market Intelligence , Analista de Marketing , Analista de Negócios/Business Analyst , Analista de Suporte/Analista Técnico , Analytics Engineer , Cientista de Dados/Data Scientist , DBA/Administrador de Banco de Dados , Desenvolvedor/ Engenheiro de Software/ Analista de Sistemas , Engenheiro de Dados/Arquiteto de Dados/Data Engineer/Data Architect , Engenheiro de Machine Learning/ML Engineer , Estatístico , Outra Opção , Outras Engenharias (não inclui dev) , Product Manager/ Product Owner (PM/APM/DPM/GPM/PO) , Professor.
NIVEL_CARGO_ATUAL	STRING	NULLABLE					Nível de senioridade do cargo atual do respondente. Valores possíveis: Júnior, Pleno, Sênior.
FAIXA_SALARIAL	STRING	NULLABLE					Faixa salarial do respondente. Valores possíveis divididos entre faixas pré-definidas, começa do em "Menos de R\$ 1.000/mês" e terminando em "Acima de R\$ 40.001/mês".

Visão geral do esquema da tabela RESPOSTA.

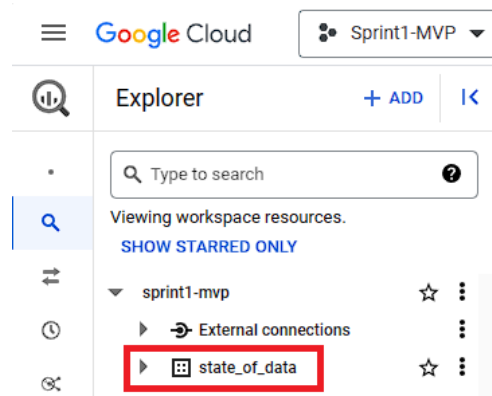
Para permitir uma melhor visualização do catálogo completo de dados, o esquema da tabela foi exportado do *BigQuery* para um documento em PDF, disponível no repositório deste projeto:

https://github.com/costinhas/puc-rio-data-engineering/blob/main/DataCatalog/Catalogo_de_dados_tabela_resposta.pdf

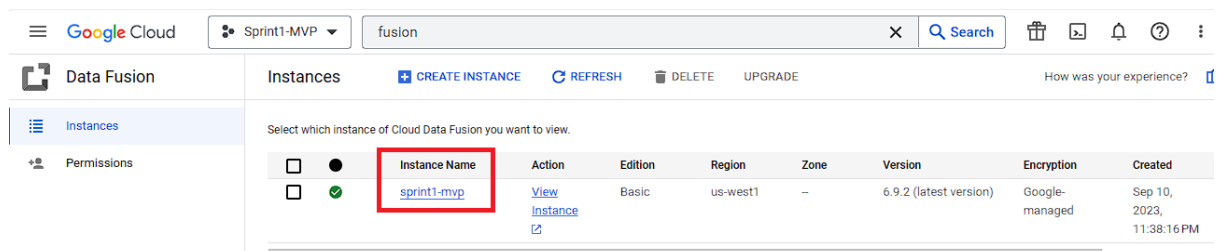
Etapa 4: Carga e transformação dos dados

Etapa 4.1: Preparação do ambiente

Para realização da carga dos dados, foi criada um novo *dataset* no Google BigQuery denominado *state_of_data*:



Todo o processo de ETL foi realizado através do *Google Data Fusion* e, para isto, foi criada uma instância denominada *sprint1-mvp*:



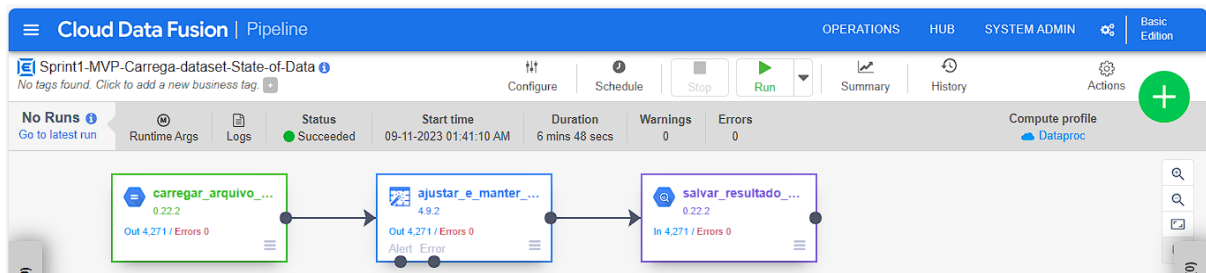
Etapa 4.2: Criação do pipeline de dados

O pipeline de dados foi criado através do *Google Cloud Data Fusion Studio*, para permitir a configuração através de interface gráfica. O fluxo contém as 3 etapas básicas do processo de ETL: carga dos dados, transformação e armazenamento do resultado final.

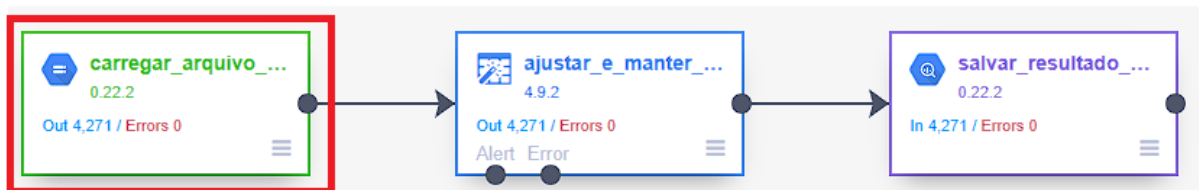
Todas as etapas, schemas e detalhes do pipeline criado estão disponíveis para consulta no repositório deste projeto, em:

<https://github.com/costinhas/puc-rio-data-engineering/tree/main/ETL>

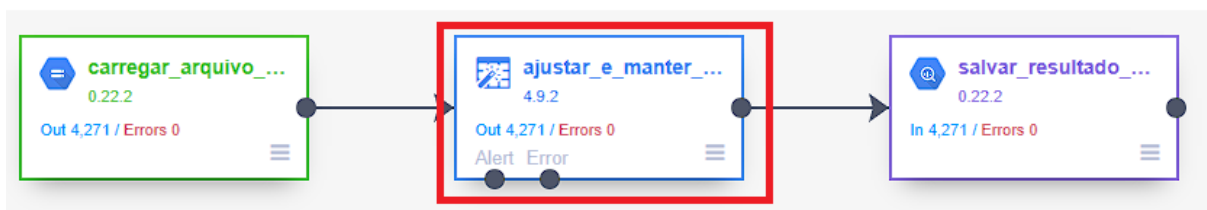
A seguir serão detalhadas todas as etapas realizadas na criação deste pipeline::



O primeiro elemento criado foi o coletor “carregar_arquivo_csv”, que carrega o *dataset* a partir do arquivo CSV salvo no Cloud Storage (detalhado na seção “Etapa 2: Coleta dos dados”):



A etapa seguinte, denominada “ajustar_e_manter_apenas_colunas_desejadas”, foi construída utilizando o módulo *Wrangler* para realizar a transformação dos dados carregados na etapa anterior:

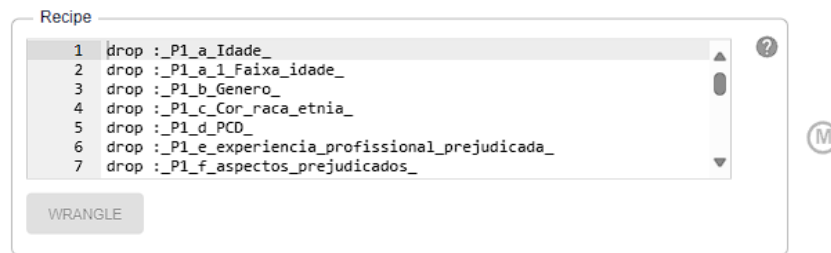


Nesta etapa, foram aplicados três tipos de transformação nos dados:

1) Remoção das colunas não relevantes para o projeto

Das 353 colunas disponíveis no dataset, somente 119 foram relevantes para o estudo. Nesta primeira etapa de transformação foram removidas 234 colunas do esquema através do comando “DROP”:

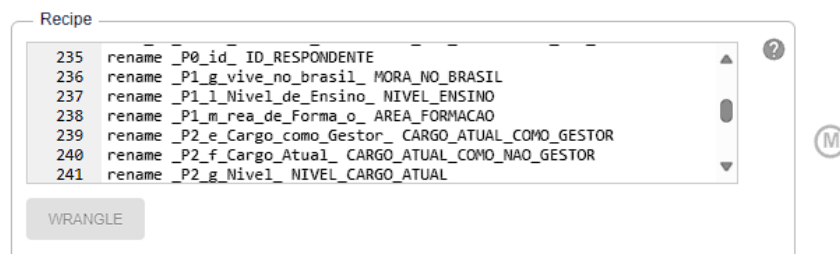
Directives



2) Alteração do nome das colunas

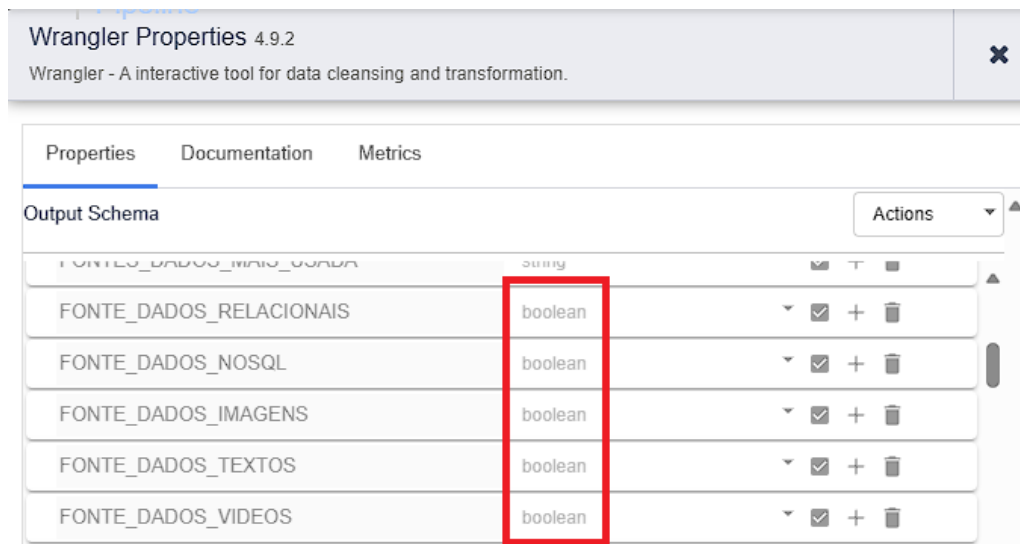
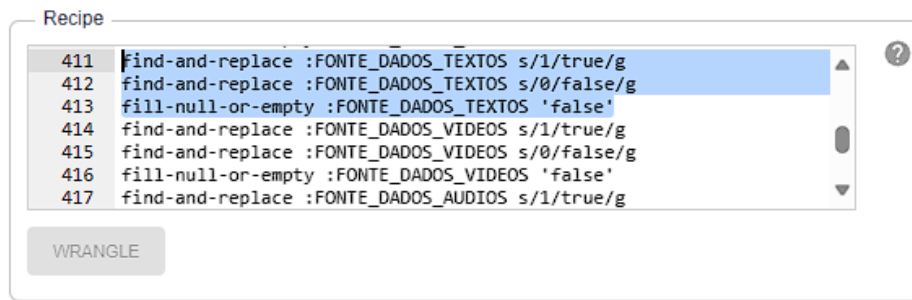
O nome das 119 colunas restantes foi alterado para títulos mais descritivos, para facilitar a análise e manipulação dos dados. Esta atualização do nome foi feita através do comando “RENAME”:

Directives

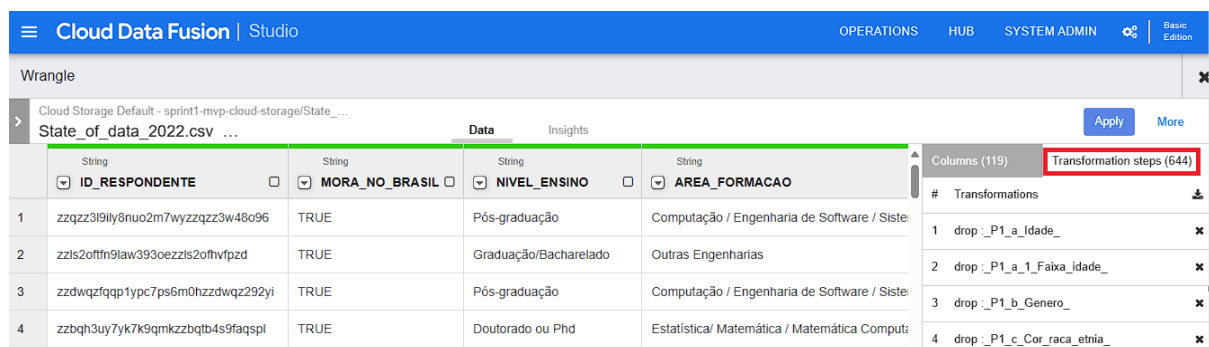


3) Conversão de colunas e seus valores para o tipo BOOLEAN

No dataset original, diversos atributos possuíam natureza booleana e foram coletados com valores “0”, “1” e nulos. Para padronização, todos estes campos foram convertidos do tipo STRING para o tipo BOOLEAN, na definição do esquema, e seus valores foram transformados. Valores nulos ou “0” foram convertidos para FALSE, enquanto valores “1” foram convertidos para TRUE.



Estes três grupos de transformações totalizaram 644 ações que foram aplicadas nos dados, antes de serem armazenados:



Por fim, a última etapa foi denominada “salvar_resultado_dataset_BigQuery” e utiliza o módulo *Sink BigQuery* para armazenar os dados transformados em uma nova tabela “respostas” no *BigQuery*:



Cloud Data Fusion | Studio

BigQuery Properties 0.22.2

This sink writes to a BigQuery table. BigQuery is Google's serverless, highly scalable, enterprise data warehouse. Data is first written to a temporary location on Google Cloud Storage, then loaded into Big...

Properties Documentation

Input Schema

Field	Type	Required	Nullable
ID_RESPONDENTE	string	Yes	No
MORA_NO_BRASIL	boolean	Yes	No
NIVEL_ENSINO	string	Yes	No
AREA_FORMACAO	string	Yes	No
CARGO_ATUAL_COMO_	string	Yes	No
CARGO_ATUAL_COMO_	string	Yes	No
NIVEL_CARGO_ATUAL	string	Yes	No
FAIXA_SALARIAL	string	Yes	No

Basic

Reference Name

03-salvar-dataset-BigQuery

BROWSE

Dataset *

state_of_data

Table *

respostas

Este fluxo completo foi executado em aproximadamente 7 minutos e não apresentou erros:

Cloud Data Fusion | Pipeline

Sprint1-MVP-Carrega-dataset-State-of-Data

No tags found. Click to add a new business tag.

Configure Schedule Stop Run Summary History Actions

No Runs

Go to latest run

Runtime Args Logs

Status

Succeeded

Start time

09-11-2023 01:41:10 AM

Duration

6 mins 48 secs

Warnings

0

Errors

0

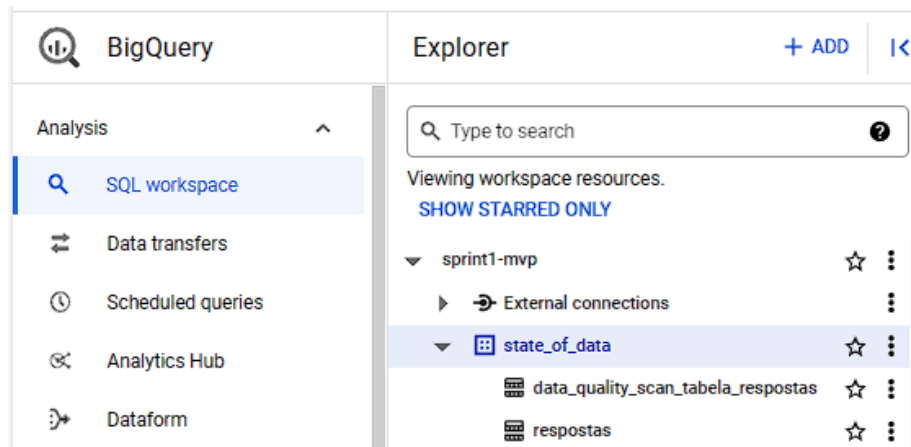
Compute profile

Dataprof

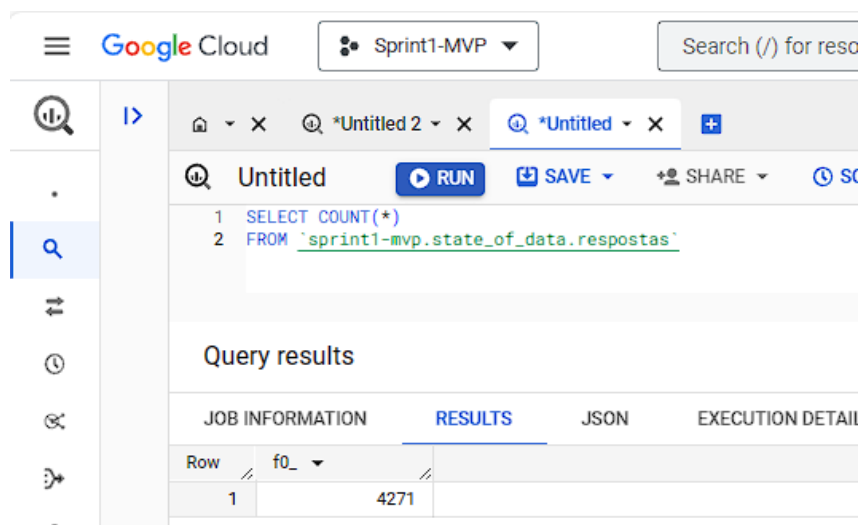
```
graph LR; A[carregar_arquivo_... 0.22.2  
Out 4,271 / Errors 0] --> B[ajustar_e_manter_... 4.9.2  
Out 4,271 / Errors 0  
Alert Error]; B --> C[salvar_resultado_... 0.22.2  
In 4,271 / Errors 0]; style C stroke:#f00,stroke-width:2px
```

Etapa 5: Análise dos dados

Após a execução do processo de ETL no *Cloud Data Fusion*, os dados foram disponibilizados no ambiente *BigQuery*, na tabela “*repostas*”, que foi utilizada como base para todas as etapas de análise de dados que serão descritas nesta seção:



As 4271 respostas do *dataset* original permanecem disponíveis na base, após o processo de ETL:



Etapa 5.1: Identificação de problemas de qualidade nos dados

Durante a etapa de modelagem dos dados, descrita na seção “Etapa 3: Modelagem dos dados”, foram identificados problemas relacionados à nomenclatura de 32 colunas no *dataset* original, que não estavam coerentes com a estrutura pré-definida e divulgada pelos autores:

Nome original da coluna	Problema identificado	Novo nome
('P4_f_4','Amazon Aurora ou RDS')	Identificador incorreto ("P4_f_4"), apesar de estar relacionado ao escopo da coluna "P4_g")	USA_AMAZON_AURORA_RDS
('P4_f_5','DynamoDB')	Identificador incorreto ("P4_f_5"), apesar de estar relacionado ao escopo da coluna "P4_g")	USA_DYNAMODB
('P4_f_6','CoachDB')	Identificador incorreto ("P4_f_6"), apesar de estar relacionado ao escopo da coluna "P4_g")	USA_COACHDB
('P4_f_7','Cassandra')	Identificador incorreto ("P4_f_7"), apesar de estar relacionado ao escopo da coluna "P4_g")	USA_CASSANDRA
('P4_f_8','MongoDB')	Identificador incorreto ("P4_f_8"), apesar de estar relacionado ao escopo da coluna "P4_g")	USA_MONGODB
('P4_f_9','MariaDB')	Identificador incorreto ("P4_f_9"), apesar de estar relacionado ao escopo da coluna "P4_g")	USA_MARIADB
('P4_f_10','Datomic')	Identificador incorreto ("P4_f_10"), apesar de estar relacionado ao escopo da coluna "P4_g")	USA_DATOMIC
('P4_f_11','S3')	Identificador incorreto ("P4_f_11"), apesar de estar relacionado ao escopo da coluna "P4_g")	USA_S3
('P4_f_12','PostgreSQL')	Identificador incorreto ("P4_f_12"), apesar de estar relacionado ao escopo da coluna "P4_g")	USA_POSTGRESQL
('P4_f_13','ElasticSearch')	Identificador incorreto ("P4_f_13"), apesar de estar relacionado ao escopo da coluna "P4_g")	USA_ELASTICSEARCH
('P4_f_14','DB2')	Identificador incorreto ("P4_f_14"), apesar de estar relacionado ao escopo da coluna "P4_g")	USA_DB2
('P4_f_15','Microsoft Access')	Identificador incorreto ("P4_f_15"), apesar de estar relacionado ao escopo da coluna "P4_g")	USA_MICROSOFT_ACCESS
('P4_f_16','SQLite')	Identificador incorreto ("P4_f_16"), apesar de estar relacionado ao escopo da coluna "P4_g")	USA_SQLITE
('P4_f_17','Sybase')	Identificador incorreto ("P4_f_17"), apesar de estar relacionado ao escopo da coluna "P4_g")	USA_SYBASE
('P4_f_18','Firebase')	Identificador incorreto ("P4_f_18"), apesar de estar relacionado ao escopo da coluna "P4_g")	USA_FIREBASE
('P4_f_19','Vertica')	Identificador incorreto ("P4_f_19"), apesar de estar relacionado ao escopo da coluna "P4_g")	USA_VERTICA
('P4_f_20','Redis')	Identificador incorreto ("P4_f_20"), apesar de estar relacionado ao escopo da coluna "P4_g")	USA_REDIS
('P4_f_21','Neo4j')	Identificador incorreto ("P4_f_21"), apesar de estar relacionado ao escopo da coluna "P4_g")	USA_NEO4J
('P4_f_22','Google BigQuery')	Identificador incorreto ("P4_f_22"), apesar de estar relacionado ao escopo da coluna "P4_g")	USA_GOOGLE_BIGQUERY
('P4_f_23','Google Firestore')	Identificador incorreto ("P4_f_23"), apesar de estar relacionado ao escopo da coluna "P4_g")	USA_GOOGLE_FIRESTORE
('P4_f_24','Amazon Redshift')	Identificador incorreto ("P4_f_24"), apesar de estar relacionado ao escopo da coluna "P4_g")	USA_AMAZON_REDSHIFT
('P4_f_25','Amazon Athena')	Identificador incorreto ("P4_f_25"), apesar de estar relacionado ao escopo da coluna "P4_g")	USA_AMAZON_ATHENA
('P4_f_26','Snowflake')	Identificador incorreto ("P4_f_26"), apesar de estar relacionado ao escopo da coluna "P4_g")	USA_SNOWFLAKE
('P4_f_27','Databricks')	Identificador incorreto ("P4_f_27"), apesar de estar relacionado ao escopo da coluna "P4_g")	USA_DATABRICKS
('P4_f_28','HBase')	Identificador incorreto ("P4_f_28"), apesar de estar relacionado ao escopo da coluna "P4_g")	USA_HBASE
('P4_f_29','Presto')	Identificador incorreto ("P4_f_29"), apesar de estar relacionado ao escopo da coluna "P4_g")	USA_PRESTO
('P4_f_30','Splunk')	Identificador incorreto ("P4_f_30"), apesar de estar relacionado ao escopo da coluna "P4_g")	USA_SPLUNK
('P4_f_31','SAP HANA')	Identificador incorreto ("P4_f_31"), apesar de estar relacionado ao escopo da coluna "P4_g")	USA_SAP_HANA
('P4_f_32','Hive')	Identificador incorreto ("P4_f_32"), apesar de estar relacionado ao escopo da coluna "P4_g")	USA_HIVE
('P4_f_33','Firebird')	Identificador incorreto ("P4_f_33"), apesar de estar relacionado ao escopo da coluna "P4_g")	USA_FIREBIRD
('P4_g','Quais das opções de Cloud listadas abaixo você utiliza no trabalho?')	Identificador incorreto ("P4_g"), quando já existia outra com mesmo nome	CLOUD_UTILIZADA_TRABALHO
('P4_i','Microsoft PowerBI')	Coluna estava incorretamente nomeada como "('(P4_i','Microsoft PowerBI'))", porém sua descrição deveria conter uma pergunta sobre ferramenta de visualização	FERRAMENTA_VISUALIZACAO_UTILIZADA

Estes problemas foram resolvidos durante o processo de ETL, mais especificamente na etapa de transformação, quando foram definidos novos nomes para as colunas. Nenhum impacto na qualidade dos dados foi identificado após esta correção.

Após criação da tabela no *BigQuery*, ajuste do esquema e carga dos dados, foi utilizado o módulo de *Data Quality Check* da plataforma para realizar uma análise mais detalhada dos dados. As regras de validação utilizadas foram propostas pela própria ferramenta, com base na estrutura e conteúdo da tabela já populada, e incluíram a checagem de valores nulos, de domínio, valores mínimos e máximos dos dados:

Google Cloud

Sprint1-MVP

bigquery

Search

4

?

:

K

BigQuery

Analysis

SQL workspace

Data transfers

Scheduled queries

Analytics Hub

Dataform

Partner Center

Migration

Assessment

SQL translation

Administration

Monitoring

Release Notes

respostas

QUERY

SHARE

COPY

SNAPSHOT

DELETE

EXPORT

REFRESH

SCHEMA

DETAILS

PREVIEW

LINEAGE

DATA PROFILE

DATA QUALITY

Current rules

MODIFY RULES

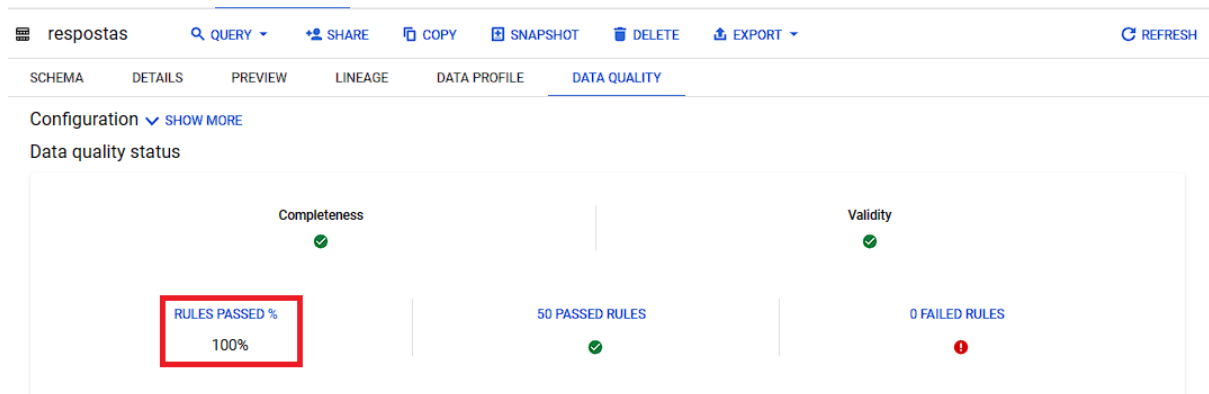
DELETE

Filter

Filter items

<input type="checkbox"/>	Column name	Rule name	Rule type	Evaluation	Dimension	Parameters
<input type="checkbox"/>	AREA_FORMACAO	-	Null Check	Per row	Completeness	
<input type="checkbox"/>	AREA_FORMACAO	-	Value Set Check	Per row	Validity	set of: Computação / Engenharia de Software / Sis...
<input type="checkbox"/>	AREA_FORMACAO	-	Row Condition Check	Per row	Validity	(LENGTH("AREA_FORMACAO")) >= 11 AND LENGT...
<input type="checkbox"/>	BANCO_DADOS_UTILIZADO_TRABALHO	-	Row Condition Check	Per row	Validity	(LENGTH("BANCO_DADOS_UTILIZADO_TRABALH...
<input type="checkbox"/>	CARGO_ATUAL_COMO_GESTOR	-	Row Condition Check	Per row	Validity	(LENGTH("CARGO_ATUAL_COMO_GESTOR")) >= 3 ...
<input type="checkbox"/>	CARGO_ATUAL_COMO_NAO_GESTOR	-	Row Condition	Per row	Validity	(LENGTH("CARGO_ATUAL_COMO_NAO_GESTOR")) ...

Após execução do processo de validação, nenhum problema adicional de qualidade dos dados foi identificado pela ferramenta:



Etapa 5.2: Solução do problema proposto

O objetivo definido para este projeto foi permitir a realização de uma análise sobre o mercado de trabalho brasileiro na área de dados, para consolidar informações que possam auxiliar pessoas que desejam iniciar sua carreira ou realizar uma transição de carreira para a área de dados.

Foram definidas 5 questões principais para guiar esta análise. Para cada uma delas, foram utilizadas consultas SQL na tabela “resultados”, no ambiente *BigQuery* e, em alguns casos, foram criados gráficos utilizando o *Microsoft Excel* com os dados extraídos das consultas, para fornecer uma visão complementar dos dados. Nenhuma outra ferramenta de visualização de dados foi utilizada a fim de simplificar o escopo deste MVP.

Todas as consultas SQL realizadas estão disponíveis na íntegra no repositório deste projeto, através do link:

<https://github.com/costinhas/puc-rio-data-engineering/tree/main/ConsultasRealizadas>

Os detalhes de cada questão serão descritos a seguir:

1) Que tipo de educação formal é necessária para trabalhar na área de dados no Brasil?

Para responder a esta pergunta, foram utilizados os atributos:

- MORA_NO_BRASIL, para filtrar apenas pessoas que moram (e trabalham) no Brasil e dar um panorama sobre o mercado de trabalho brasileiro;
- TEMPO_EXPERIENCIA_DADOS, para filtrar apenas pessoas que já possuem algum tipo de experiência na área de dados;

- NIVEL_ENSINO e AREA_FORMACAO, para listar o tipo de formação que os respondentes do perfil desejado possuem.

Além disso, foram criadas duas variáveis auxiliares para armazenar o total geral de respostas (total_respondentes) e o total de respostas das pessoas que vivem no Brasil e possuem alguma experiência em dados (total_respondentes_brasil).

```

1  /* Variáveis auxiliares */
2  DECLARE total_respondentes, total_respondentes_brasil INT64;
3
4  SET (total_respondentes) = (
5    SELECT AS STRUCT COUNT(*)
6    FROM `sprint1-mvp.state_of_data.respostas`
7  );
8
9  SET (total_respondentes_brasil) = (
10   SELECT AS STRUCT COUNT(*)
11   FROM `sprint1-mvp.state_of_data.respostas`
12   WHERE MORA_NO_BRASIL = true
13   AND TEMPO_EXPERIENCIA_DADOS is not null
14   AND TEMPO_EXPERIENCIA_DADOS <> 'Não tenho experiência na área de dados'
15 );
16
17 SELECT total_respondentes, total_respondentes_brasil, total_respondentes_brasil/total_respondentes as Percentual
18 FROM (SELECT SESSION_USER())

```

← Query results

JOB INFORMATION		RESULTS	JSON	EXECUTION DETAILS	CHART	PREVIEW	EXECUTION GRAPH
Row	total_respondentes	total_respondentes_brasil	Percentual				
1	4271	3430	0.803090611098...				

Nota-se que, dos 4271 respondentes da pesquisa, 3430 (80%) trabalham no Brasil e possuem alguma experiência na área de dados. Ao analisar o nível de ensino destes respondentes, é possível verificar que estes profissionais possuem um alto grau de instrução formal:

```

23 SELECT NIVEL_ENSINO,
24         COUNT(*) as Total,
25         TRUNC((COUNT(*)/total_respondentes_brasil)*100) as Percentual
26 FROM `sprint1-mvp.state_of_data.respostas`
27 WHERE MORA_NO_BRASIL = true
28 AND TEMPO_EXPERIENCIA_DADOS is not null
29 AND TEMPO_EXPERIENCIA_DADOS <> 'Não tenho experiência na área de dados'
30 GROUP BY NIVEL_ENSINO;

```

← Query results

JOB INFORMATION		RESULTS	JSON	EXECUTION DETAILS
Row	NIVEL_ENSINO	Total	Percentual	
1	Graduação/Bacharelado	1239	36.0	
2	Pós-graduação	1125	32.0	
3	Mestrado	416	12.0	
4	Doutorado ou Phd	141	4.0	
5	Estudante de Graduação	433	12.0	
6	Não tenho graduação formal	69	2.0	
7	Prefiro não informar	7	0.0	

Ao agrupar todos os níveis de formação (graduação, pós, mestrado e doutorado) nota-se que 85% dos respondentes possuem educação formal:

```

35 SELECT CASE
36     WHEN NIVEL_ENSINO = 'Graduação/Bacharelado' then 'Graduação / Pós / Mestrado / Doutorado / PHD'
37     WHEN NIVEL_ENSINO = 'Pós-graduação' then 'Graduação / Pós / Mestrado / Doutorado / PHD'
38     WHEN NIVEL_ENSINO = 'Mestrado' then 'Graduação / Pós / Mestrado / Doutorado / PHD'
39     WHEN NIVEL_ENSINO = 'Doutorado ou Phd' then 'Graduação / Pós / Mestrado / Doutorado / PHD'
40 ELSE
41     NIVEL_ENSINO
42 END as Escolaridade,
43 COUNT(*) as Total,
44 TRUNC((COUNT(*)/total_respondentes_brasil)*100) as Percentual
45 FROM
46     'sprint1-mvp.state_of_data.respostas'
47 WHERE MORA_NO_BRASIL = true
48 AND TEMPO_EXPERIENCIA_DADOS is not null
49 AND TEMPO_EXPERIENCIA_DADOS <> 'Não tenho experiência na área de dados'
50 GROUP BY Escolaridade

```

← Query results

JOB INFORMATION				EXECUTION DETAILS		CHART	PREVIEW	EXECUTION GRAF
Row	Escolaridade	Total	Percentual					
1	Graduação / Pós / Mestrado / Doutorado / PHD	2921	85.0					
2	Estudante de Graduação	433	12.0					
3	Não tenho graduação formal	69	2.0					
4	Prefiro não informar	7	0.0					

Ao agrupar estas respostas por área de formação, é possível verificar que 72% das pessoas possuem formação em áreas exatas, sendo que a área mais comum é a de Computação, representando 31% do total:

```

SELECT AREA_FORMACAO,
COUNT(*) as Total,
TRUNC((COUNT(*)/total_respondentes_brasil)*100) as Percentual
FROM
'sprint1-mvp.state_of_data.respostas'
WHERE MORA_NO_BRASIL = true
AND TEMPO_EXPERIENCIA_DADOS is not null
AND TEMPO_EXPERIENCIA_DADOS <> 'Não tenho experiência na área de dados'
AND NIVEL_ENSINO in ('Graduação/Bacharelado', 'Pós-graduação', 'Mestrado', 'Doutorado ou Phd')
GROUP BY AREA_FORMACAO
ORDER BY COUNT(*) DESC

```

Row	AREA_FORMACAO	Total	Percentual	
1	Computação / Engenharia de Software / Sistemas de Informação/ TI	1075	31.0	72%
2	Outras Engenharias	712	20.0	
3	Economia/ Administração / Contabilidade / Finanças/ Negócios	454	13.0	
4	Estatística/ Matemática / Matemática Computacional/ Ciências Atuariais	282	8.0	
5	Outra opção	114	3.0	
6	Marketing / Publicidade / Comunicação / Jornalismo	95	2.0	
7	Química / Física	78	2.0	
8	Ciências Biológicas/ Farmácia/ Medicina/ Área da Saúde	70	2.0	
9	Ciências Sociais	41	1.0	

Conclusão:

Retornando à pergunta inicial “*Que tipo de educação formal é necessária para trabalhar na área de dados no Brasil?*”, foi possível concluir que, dos respondentes que trabalham no Brasil e possuem experiência na área de dados:

- 85% possuem educação formal (no mínimo, Graduação ou Bacharelado):
 - 36% possuem Graduação / Bacharelado
 - 32% possuem Pós-Graduação
 - 12% possuem Mestrado
 - 4% possuem Doutorado ou PHD
- 72% possuem educação formal (no mínimo, Graduação ou Bacharelado) em cursos das áreas exatas.
- 31% possuem educação formal (no mínimo, Graduação ou Bacharelado) em cursos relacionados à área da Computação (Engenharia de Software, Sistemas de Informação, Tecnologia da Informação, etc).

2) Quais são os cargos ou funções mais comuns no mercado brasileiro na área de dados?

Para responder à esta pergunta, foram utilizados os atributos:

- MORA_NO_BRASIL, para filtrar apenas pessoas que moram (e trabalham) no Brasil e dar um panorama sobre o mercado de trabalho brasileiro;
- TEMPO_EXPERIENCIA_DADOS, para filtrar apenas pessoas que já possuem algum tipo de experiência na área de dados;
- CARGO_ATUAL_COMO_NAO_GESTOR e CARGO_ATUAL_COMO_GESTOR, para listar os cargos atuais em ambos os níveis.

Dos 3430 respondentes que trabalham no Brasil e possuem alguma experiência na área de dados, 19% ocupam cargos de gestão e pouco mais de 80% ocupam cargos não gerenciais:

```

SELECT 'NÃO GESTOR' as Nivel_Cargo,
COUNT(*) as Total,
TRUNC((COUNT(*)/total_respondentes_brasil)*100) as Percentual
FROM 'sprint1-mvp.state_of_data.respostas'
WHERE MORA_NO_BRASIL = true
AND TEMPO_EXPERIENCIA_DADOS is not null
AND TEMPO_EXPERIENCIA_DADOS <> 'Não tenho experiência na área de dados'
AND CARGO_ATUAL_COMO_NAO_GESTOR is not null
AND CARGO_ATUAL_COMO_GESTOR is null
UNION ALL
SELECT 'GESTOR' as Nivel_Cargo,
COUNT(*) as Total,
TRUNC((COUNT(*)/total_respondentes_brasil)*100) as Percentual
FROM 'sprint1-mvp.state_of_data.respostas'
WHERE MORA_NO_BRASIL = true
AND TEMPO_EXPERIENCIA_DADOS is not null
AND TEMPO_EXPERIENCIA_DADOS <> 'Não tenho experiência na área de dados'
AND CARGO_ATUAL_COMO_NAO_GESTOR is null
AND CARGO_ATUAL_COMO_GESTOR is not null

```

Row	Nivel_Cargo	Total	Percentual
1	GESTOR	669	19.0
2	NÃO GESTOR	2761	80.0

Os principais cargos ou funções ocupadas pelos gestores foram identificados através da consulta a seguir:

```

SELECT CARGO_ATUAL_COMO_GESTOR,
COUNT(*) as Total,
TRUNC((COUNT(*)/total_respondentes_brasil)*100) as Percentual
FROM 'sprint1-mvp.state_of_data.respostas'
WHERE MORA_NO_BRASIL = true
AND TEMPO_EXPERIENCIA_DADOS is not null
AND TEMPO_EXPERIENCIA_DADOS <> 'Não tenho experiência na área de dados'
AND CARGO_ATUAL_COMO_NAO_GESTOR is null
AND CARGO_ATUAL_COMO_GESTOR is not null
GROUP BY CARGO_ATUAL_COMO_GESTOR
ORDER BY COUNT(*) desc

```

Row	CARGO_ATUAL_COMO_GESTOR	Total	Percentual
1	Gerente/Head	236	6.0
2	Supervisor/Coordenador	224	6.0
3	Team Leader/Tech Leader	105	3.0
4	Sócio ou C-level (CEO, CDO, CIO...	56	1.0
5	Diretor/VP	38	1.0
6	nao tem	1	0.0
7	Analista	1	0.0
8	Chefe de Secretaria	1	0.0
9	Lider	1	0.0
10	Especialista	1	0.0
11	Agente de transformação	1	0.0
12	Chapter Lead	1	0.0

Já para os cargos que não são de gestão, as principais funções foram identificadas pela consulta a seguir::

```
SELECT CARGO_ATUAL_COMO_NAO_GESTOR,
COUNT(*) as Total,
TRUNC((COUNT(*)/total_respondentes_brasil)*100) as Percentual
FROM `sprint1-mvp.state_of_data.respostas`
WHERE MORA_NO_BRASIL = true
AND TEMPO_EXPERIENCIA_DADOS is not null
AND TEMPO_EXPERIENCIA_DADOS <> 'Não tenho experiência na área de dados'
AND CARGO_ATUAL_COMO_NAO_GESTOR is not null
AND CARGO_ATUAL_COMO_GESTOR is null
GROUP BY CARGO_ATUAL_COMO_NAO_GESTOR
ORDER BY COUNT(*) desc
```

Row	CARGO_ATUAL_COMO_NAO_GESTOR	Total	Percentual
1	Analista de Dados/Data Analyst	626	18.0
2	Cientista de Dados/Data Scientist	545	15.0
3	Engenheiro de Dados/Arquiteto de Dados/Data Engineer/Data Architect	472	13.0
4	Analista de BI/BI Analyst	371	10.0
5	Outra Opção	174	5.0
6	Analista de Negócios/Business Analyst	113	3.0
7	Desenvolvedor/ Engenheiro de Software/ Analista de Sistemas	95	2.0
8	Analytics Engineer	70	2.0
9	Engenheiro de Machine Learning/ML Engineer	65	1.0
10	Analista de Suporte/Analista Técnico	58	1.0
11	Product Manager/ Product Owner (PM/APM/DPM/GPM/PO)	50	1.0

Conclusão:

Os papéis de Analista de Dados, Cientista de Dados, Engenheiro/Arquiteto de Dados e Analista de BI são os mais comuns entre as pessoas que não ocupam cargos de gestão.

Entre os cargos de gestão, os papéis mais comuns são os de Gerente, Supervisor ou Team Leader.

3) Qual é a média salarial dos profissionais de dados no Brasil?

Para responder à esta pergunta, foram utilizados os atributos:

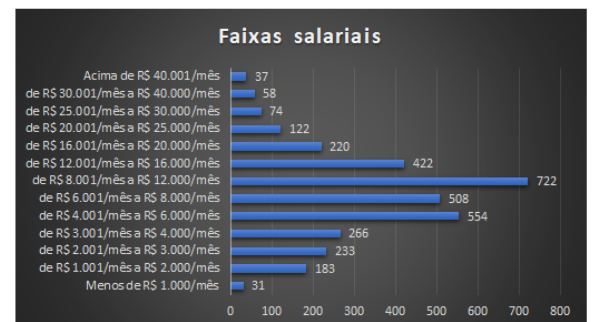
- MORA_NO_BRASIL, para filtrar apenas pessoas que moram (e trabalham) no Brasil e dar um panorama sobre o mercado de trabalho brasileiro;
- TEMPO_EXPERIENCIA_DADOS, para filtrar apenas pessoas que já possuem algum tipo de experiência na área de dados;
- FAIXA_SALARIAL, para análise da faixa salarial dos respondentes;

- NIVEL_CARGO_ATUAL, para análise do nível do cargo atual dos respondentes (júnior, pleno, sênior ou gestor).

A primeira análise incluiu o agrupamento das faixas salariais dos respondentes que moram no Brasil e possuem experiência na área de dados:

```
/* Agrupamento por faixa salarial */
/* Criada subquery com Case para incluir uma ordenação forçada, por faixa salarial */
/* Este sequencial usado na ordenação é excluído na query principal, com a função SUBSTRING*/
SELECT SUBSTRING(FAIXA_SALARIAL, 6) AS FAIXA_SALARIAL,
       Total,
       Percentual
FROM (
  SELECT CASE FAIXA_SALARIAL
    WHEN 'Menos de R$ 1.000/mês' THEN '01 - Menos de R$ 1.000/mês'
    WHEN 'de R$ 1.001/mês a R$ 2.000/mês' THEN '02 - de R$ 1.001/mês a R$ 2.000/mês'
    WHEN 'de R$ 2.001/mês a R$ 3.000/mês' THEN '03 - de R$ 2.001/mês a R$ 3.000/mês'
    WHEN 'de R$ 3.001/mês a R$ 4.000/mês' THEN '04 - de R$ 3.001/mês a R$ 4.000/mês'
    WHEN 'de R$ 4.001/mês a R$ 6.000/mês' THEN '05 - de R$ 4.001/mês a R$ 6.000/mês'
    WHEN 'de R$ 6.001/mês a R$ 8.000/mês' THEN '06 - de R$ 6.001/mês a R$ 8.000/mês'
    WHEN 'de R$ 8.001/mês a R$ 12.000/mês' THEN '07 - de R$ 8.001/mês a R$ 12.000/mês'
    WHEN 'de R$ 12.001/mês a R$ 16.000/mês' THEN '08 - de R$ 12.001/mês a R$ 16.000/mês'
    WHEN 'de R$ 16.001/mês a R$ 20.000/mês' THEN '09 - de R$ 16.001/mês a R$ 20.000/mês'
    WHEN 'de R$ 20.001/mês a R$ 25.000/mês' THEN '10 - de R$ 20.001/mês a R$ 25.000/mês'
    WHEN 'de R$ 25.001/mês a R$ 30.000/mês' THEN '11 - de R$ 25.001/mês a R$ 30.000/mês'
    WHEN 'de R$ 30.001/mês a R$ 40.000/mês' THEN '12 - de R$ 30.001/mês a R$ 40.000/mês'
    WHEN 'Acima de R$ 40.001/mês' THEN '13 - Acima de R$ 40.001/mês'
    ELSE FAIXA_SALARIAL
  END AS FAIXA_SALARIAL,
  COUNT(*) as Total,
  TRUNC((COUNT(*)/total_respondentes_brasil)*100) as Percentual
FROM 'sprint1-mvp.state_of_data.respostas'
WHERE MORA_NO_BRASIL = true
AND TEMPO_EXPERIENCIA_DADOS is not null
AND TEMPO_EXPERIENCIA_DADOS <> 'Não tenho experiência na área de dados'
GROUP BY FAIXA_SALARIAL
ORDER BY FAIXA_SALARIAL);
```

Row	FAIXA_SALARIAL	Total	Percentual
1	Menos de R\$ 1.000/mês	31	0.0
2	de R\$ 1.001/mês a R\$ 2.000/mês	183	5.0
3	de R\$ 2.001/mês a R\$ 3.000/mês	233	6.0
4	de R\$ 3.001/mês a R\$ 4.000/mês	266	7.0
5	de R\$ 4.001/mês a R\$ 6.000/mês	554	16.0
6	de R\$ 6.001/mês a R\$ 8.000/mês	508	14.0
7	de R\$ 8.001/mês a R\$ 12.000/mês	722	21.0
8	de R\$ 12.001/mês a R\$ 16.000/mês	422	12.0
9	de R\$ 16.001/mês a R\$ 20.000/mês	220	6.0
10	de R\$ 20.001/mês a R\$ 25.000/mês	122	3.0
11	de R\$ 25.001/mês a R\$ 30.000/mês	74	2.0
12	de R\$ 30.001/mês a R\$ 40.000/mês	58	1.0
13	Acima de R\$ 40.001/mês	37	1.0



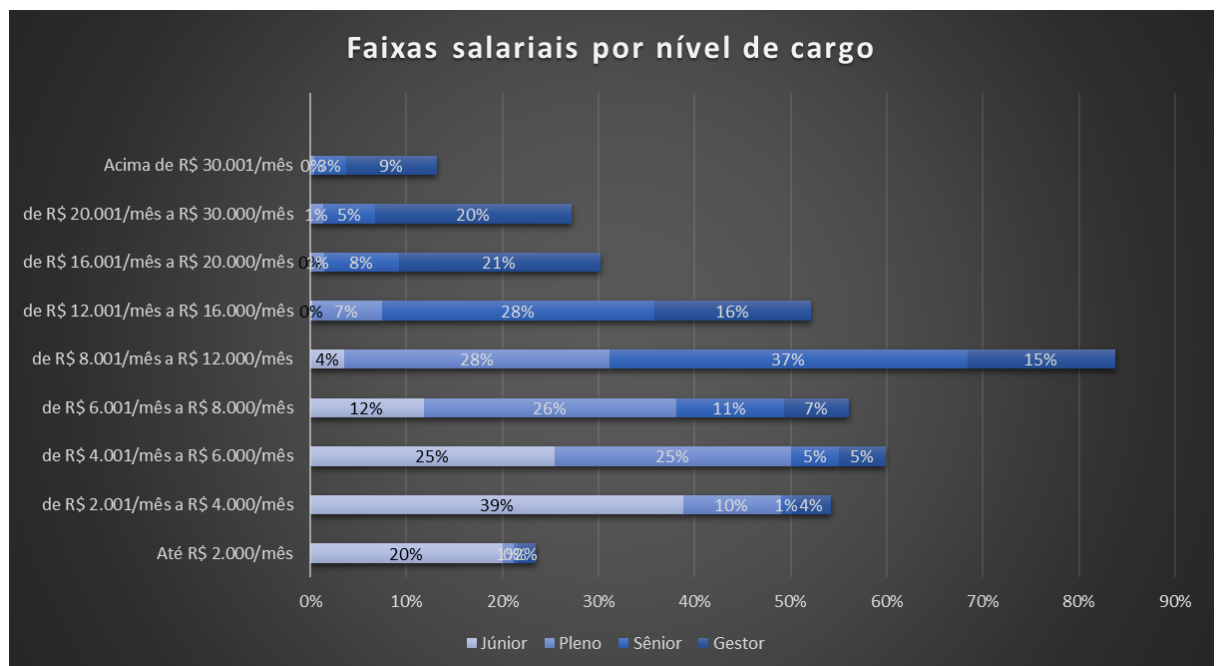
A partir desta visão, é possível verificar que a faixa salarial mais frequente entre os respondentes deste perfil é a de R\$8.000 a R\$12.000 por mês, representando 21% do total. 48% dos respondentes deste perfil possuem salários entre R\$6.000 e R\$16.000 por mês.

Para refinar ainda mais a análise, o nível do cargo foi incluído na consulta. As faixas de valores foram agrupadas em valores mais abrangentes, para reduzir a quantidade de séries distintas:

```

/* Agrupamento por nível de cargo atual e faixa salarial */
/* Criada subquery com Case para incluir uma ordenação forçada, por faixa salarial */
/* Este sequencial usado na ordenação é excluído na query principal, com a função SUBSTRING*/
SELECT NIVEL_CARGO_ATUAL,
SUBSTRING(FAIXA_SALARIAL, 6) AS FAIXA_SALARIAL,
Total,
Percentual
FROM (
SELECT IFNULL(NIVEL_CARGO_ATUAL, "Gestor") as NIVEL_CARGO_ATUAL,
CASE FAIXA_SALARIAL
WHEN 'Menos de R$ 1.000/mês' THEN '01 - Até R$ 2.000/mês'
WHEN 'de R$ 1.001/mês a R$ 2.000/mês' THEN '01 - Até R$ 2.000/mês'
WHEN 'de R$ 2.001/mês a R$ 3.000/mês' THEN '02 - de R$ 2.001/mês a R$ 4.000/mês'
WHEN 'de R$ 3.001/mês a R$ 4.000/mês' THEN '02 - de R$ 2.001/mês a R$ 4.000/mês'
WHEN 'de R$ 4.001/mês a R$ 6.000/mês' THEN '03 - de R$ 4.001/mês a R$ 6.000/mês'
WHEN 'de R$ 6.001/mês a R$ 8.000/mês' THEN '04 - de R$ 6.001/mês a R$ 8.000/mês'
WHEN 'de R$ 8.001/mês a R$ 12.000/mês' THEN '05 - de R$ 8.001/mês a R$ 12.000/mês'
WHEN 'de R$ 12.001/mês a R$ 16.000/mês' THEN '06 - de R$ 12.001/mês a R$ 16.000/mês'
WHEN 'de R$ 16.001/mês a R$ 20.000/mês' THEN '07 - de R$ 16.001/mês a R$ 20.000/mês'
WHEN 'de R$ 20.001/mês a R$ 25.000/mês' THEN '08 - de R$ 20.001/mês a R$ 30.000/mês'
WHEN 'de R$ 25.001/mês a R$ 30.000/mês' THEN '08 - de R$ 20.001/mês a R$ 30.000/mês'
WHEN 'de R$ 30.001/mês a R$ 40.000/mês' THEN '09 - Acima de R$ 30.001/mês'
WHEN 'Acima de R$ 40.001/mês' THEN '09 - Acima de R$ 30.001/mês'
ELSE FAIXA_SALARIAL
END AS FAIXA_SALARIAL,
COUNT(*) as Total,
TRUNC((COUNT(*)/total_respondentes_brasil)*100) as Percentual
FROM `sprint1-mvp.state_of_data.respostas`
WHERE MORA_NO_BRASIL = true
AND TEMPO_EXPERIENCIA_DADOS is not null
AND TEMPO_EXPERIENCIA_DADOS <> 'Não tenho experiência na área de dados'
GROUP BY NIVEL_CARGO_ATUAL, FAIXA_SALARIAL
ORDER BY NIVEL_CARGO_ATUAL, FAIXA_SALARIAL);

```



Conclusão:

Não foi possível obter um valor salarial médio, pois o estudo coletou apenas as faixas salariais dos respondentes. No entanto, é possível identificar as faixas salariais mais comuns para os diferentes níveis de cargos:

- Para os profissionais com nível Júnior:
 - 20% possuem salários de até R\$2.000 por mês;
 - 39% possuem salários entre R\$2.000 e R\$4.000 por mês;
 - 25% possuem salários entre R\$4.000 e R\$6.000 por mês.
- Para os profissionais com nível Pleno:
 - 25% possuem salários entre R\$4.000 e R\$6.000 por mês;
 - 26% possuem salários entre R\$6.000 e R\$8.000 por mês;
 - 28% possuem salários entre R\$8.000 e R\$12.000 por mês.
- Para os profissionais com nível Sênior:
 - 11% possuem salários entre R\$6.000 e R\$8.000 por mês;
 - 37% possuem salários entre R\$8.000 e R\$12.000 por mês;
 - 28% possuem salários entre R\$12.000 e R\$16.000 por mês.
- Para os profissionais com nível de Gestão:
 - 16% possuem salários entre R\$12.000 e R\$16.000 por mês;
 - 21% possuem salários entre R\$16.000 e R\$20.000 por mês;
 - 29% possuem salários acima de R\$20.000 por mês.

Nota-se uma grande variação salarial entre os diferentes níveis de cargo. Em estudos futuros, pode-se considerar outros critérios como tempo de experiência na função e tipo de educação formal em cada um dos níveis de cargo, para aprofundar a análise e possibilitar um melhor entendimento da relação entre estes diferentes atributos.

4) Os profissionais de dados no Brasil estão satisfeitos com seus empregos atuais?

Para responder à esta pergunta, foram utilizados os atributos:

- MORA_NO_BRASIL, para filtrar apenas pessoas que moram (e trabalham) no Brasil e dar um panorama sobre o mercado de trabalho brasileiro;
- TEMPO_EXPERIENCIA_DADOS, para filtrar apenas pessoas que já possuem algum tipo de experiência na área de dados;
- SATISFACAO_EMPRESA_ATUAL, para analisar se respondentes estão satisfeitos ou não, em seus trabalhos atuais;
- Atributos que indicam os diferentes motivos da insatisfação no ambiente de trabalho:

- INSATISFACAO_FALTA_OPORTUNIDADE_CRESCIMENTO
- INSATISFACAO_SALARIO_BAIXO
- INSATISFACAO_RELACAO_GESTOR
- INSATISFACAO_DESEJO_OUTRA_AREA
- INSATISFACAO_POUCOS_BENEFICIOS
- INSATISFACAO_CLIMA_TRABALHO
- INSATISFACAO_FALTA_MATURIDADE_ANALITICA

Inicialmente, foi avaliado o grau de satisfação dos respondentes que moram no Brasil e possuem alguma experiência no mercado de dados:

```

17  /* Visão geral da satisfação com a empresa atual */
18  SELECT  SATISFACAO_EMPRESA_ATUAL,
19          COUNT(*) as Total,
20          ROUND((COUNT(*)/total_respondentes_brasil)*100, 2) as Percentual
21  FROM    `sprint1-mvp.state_of_data.respostas`
22  WHERE   MORA_NO_BRASIL = true
23  AND     TEMPO_EXPERIENCIA_DADOS is not null
24  AND     TEMPO_EXPERIENCIA_DADOS <> 'Não tenho experiência na área de dados'
25  GROUP BY SATISFACAO_EMPRESA_ATUAL;

```

← Query results

JOB INFORMATION		RESULTS	JSON	EXECUTION DETAILS	CHART
Row	SATISFACA	Total	Percentual		
1	false	824	24.02		
2	true	2606	75.98		

É possível notar que cerca de 76% dos respondentes estão satisfeitos com seus empregos atuais, enquanto 24% não estão.

Para estes 24% (824 respondentes) que não estão satisfeitos, foram analisados os principais motivos da insatisfação:

```

/* Visão dos motivos de insatisfação com a empresa atual */
WITH INSATISFEITOS AS (
    SELECT
        COUNTIF(INSATISFACAO_FALTA_OPORTUNIDADE_CRESCIMENTO = TRUE) as FALTA_OPORTUNIDADE_CRESCIMENTO,
        COUNTIF(INSATISFACAO_SALARIO_BAIIXO = TRUE) as SALARIO_BAIIXO,
        COUNTIF(INSATISFACAO_RELACAO_GESTOR = TRUE) as RELACAO_GESTOR,
        COUNTIF(INSATISFACAO_DESEJO_OUTRA_AREA = TRUE) as DESEJO_OUTRA_AREA,
        COUNTIF(INSATISFACAO_POUCOS_BENEFICIOS = TRUE) as POUCOS_BENEFICIOS,
        COUNTIF(INSATISFACAO_CLIMA_TRABALHO = TRUE) as CLIMA_TRABALHO,
        COUNTIF(INSATISFACAO_FALTA_MATURIDADE_ANALITICA = TRUE) as FALTA_MATURIDADE_ANALITICA
    FROM
        `sprint1-mvp.state_of_data.respostas`
    WHERE
        MORA_NO_BRASIL = true
    AND
        TEMPO_EXPERIENCIA_DADOS is not null
    AND
        TEMPO_EXPERIENCIA_DADOS <> 'Não tenho experiência na área de dados'
    AND
        SATISFACAO_EMPRESA_ATUAL = false
)
SELECT
    MOTIVO_INSATISFACAO,
    RESPONDENTES,
    ROUND((RESPONDENTES/total_respondentes_brasil)*100, 2) as PERCENTUAL
FROM
    INSATISFEITOS
UNPIVOT(
    RESPONDENTES
    FOR MOTIVO_INSATISFACAO IN
        (FALTA_OPORTUNIDADE_CRESCIMENTO,
        SALARIO_BAIIXO,
        RELACAO_GESTOR,
        DESEJO_OUTRA_AREA,
        POUCOS_BENEFICIOS,
        CLIMA_TRABALHO,
        FALTA_MATURIDADE_ANALITICA)
)
ORDER BY
    RESPONDENTES DESC;

```

Row	MOTIVO_INSATISFACAO	RESPONDENTES	PERCENTUAL
1	FALTA_OPORTUNIDADE_CRESCIMENTO	358	10.44
2	FALTA_MATURIDADE_ANALITICA	345	10.06
3	SALARIO_BAIIXO	337	9.83
4	DESEJO_OUTRA_AREA	225	6.56
5	POUCOS_BENEFICIOS	165	4.81
6	CLIMA_TRABALHO	124	3.62
7	RELACAO_GESTOR	59	1.72

Conclusão:

Foi possível identificar que, dos respondentes que moram no Brasil e possuem alguma experiência no mercado de dados, 76% estão satisfeitos com seus empregos atuais e 24% não estão satisfeitos.

Os principais motivos de insatisfação citados foram a falta de oportunidade de crescimento, falta de maturidade analítica na empresa, salário baixo e o desejo por outra área de atuação.

5) Quais são as principais tecnologias em uso atualmente na área de dados no Brasil?

Para responder à esta pergunta, foram utilizados os atributos:

- MORA_NO_BRASIL, para filtrar apenas pessoas que moram (e trabalham) no Brasil e dar um panorama sobre o mercado de trabalho brasileiro;
- Atributos que indicam o uso das linguagens utilizadas pelos respondentes:
 - USA_SQL
 - USA_R
 - USA_PYTHON
 - USA_C_CPP_CSHARP
 - USA_DOT_NET
 - USA_JAVA
 - USA_JULIA
 - USA_SAS_STATA
 - USA_VISUAL_BASIC_VBA
 - USA_SCALA
 - USA_MATLAB
 - USA_PHP
 - USA_JAVASCRIPT
- Atributos que indicam o uso das plataformas de nuvem utilizadas pelos respondentes:
 - USA_AZURE
 - USA_AWS
 - USA_GOOGLE_CLOUD
- Atributos que indicam o uso dos bancos de dados utilizados pelos respondentes:
 - USA_MYSQL
 - USA_ORACLE
 - USA_SQL_SERVER
 - USA_AMAZON_AURORA_RDS
 - USA_DYNAMODB
 - USA_COACHDB
 - USA_CASSANDRA
 - USA_MONGODB
 - USA_MARIADB
 - USA_DATOMIC
 - USA_S3
 - USA_POSTGRESQL
 - USA_ELASTICSEARCH
 - USA_DB2
 - USA_MICROSOFT_ACCESS
 - USA_SQLITE
 - USA_SYBASE
 - USA_FIREBASE
 - USA_VERTICA
 - USA_REDIS
 - USA_NEO4J
 - USA_GOOGLE_BIGQUERY
 - USA_GOOGLE_FIRESTORE
 - USA_AMAZON_REDSHIFT

- USA_AMAZON_ATHENA
 - USA_SNOWFLAKE
 - USA_DATABRICKS
 - USA_HBASE
 - USA_PRESTO
 - USA_SPLUNK
 - USA_SAP_HANA
 - USA_HIVE
 - USA_FIREBIRD
- Atributos que indicam o uso das ferramentas de visualização de dados utilizadas pelos respondentes:
 - USA_POWERBI
 - USA_qlik_view_sense
 - USA_TABLEAU
 - USA_METABASE
 - USA_SUPERSET
 - USA_REDASH
 - USA_MICROSTRATEGY
 - USA_IBM_ANALYTICS_COGNOS
 - USA_SAP_BO
 - USA_ORACLE_BUSINESS
 - USA_AMAZON_QUICKSIGHT
 - USA_SALESFORCE_EINSTEIN
 - USA_MODE
 - USA_ALTERYX
 - USA_BIRST
 - USA_LOOKER
 - USA_GOOGLE_DATA_STUDIO
 - USA_SAS_VISUAL_ANALYTICS
 - USA_GRAFANA
 - USA_SPOTFIRE
 - USA_PENTAHO
 - USA_EXCEL_PLANILHA_GOOGLE

A análise será feita agrupando os atributos referentes a cada tipo de ferramenta ou tecnologia, descritas nas seções a seguir:

Linguagens utilizadas:

```

/* Variáveis auxiliares */
DECLARE total_respondentes_brasil INT64;

SET (total_respondentes_brasil) = (
  SELECT AS STRUCT COUNT(*)
  FROM `sprint1-mvp.state_of_data.respostas`
  WHERE MORA_NO_BRASIL = true
);

```

```

/* Visão dos tipo de linguagem utilizadas pelos respondentes */
WITH LINGUAGENS AS (
    SELECT
        COUNTIF(USA_SQL = TRUE) as SQL,
        COUNTIF(USA_R = TRUE) as R,
        COUNTIF(USA_PYTHON = TRUE) as PYTHON,
        COUNTIF(USA_C_CPP_CSHARP = TRUE) as C_CPP_CSHARP,
        COUNTIF(USA_DOT_NET = TRUE) as DOT_NET,
        COUNTIF(USA_JAVA = TRUE) as JAVA,
        COUNTIF(USA_JULIA = TRUE) as JULIA,
        COUNTIF(USA_SAS_STATA = TRUE) as SAS_STATA,
        COUNTIF(USA_VISUAL_BASIC_VBA = TRUE) as VISUAL_BASIC_VBA,
        COUNTIF(USA_SCALA = TRUE) as SCALA,
        COUNTIF(USA_MATLAB = TRUE) as MATLAB,
        COUNTIF(USA_PHP = TRUE) as PHP,
        COUNTIF(USA_JAVASCRIPT = TRUE) as JAVASCRIPT
    FROM `sprint1-mvp.state_of_data.respostas`
    WHERE MORA_NO_BRASIL = true
)
SELECT * FROM LINGUAGENS
UNPIVOT(RESPONDENTES
    FOR LANGUAGE IN
    ( SQL, R, PYTHON, C_CPP_CSHARP,
      DOT_NET, JAVA, JULIA, SAS_STATA,
      VISUAL_BASIC_VBA, SCALA, MATLAB,
      PHP, JAVASCRIPT))
ORDER BY RESPONDENTES DESC;

```

É possível identificar que as principais linguagens utilizadas são SQL e Python:

Row	LINGUAGEM	RESPONDENTES	Percentual
1	SQL	2315	55.5
2	PYTHON	2040	48.91
3	R	373	8.94
4	JAVA	257	6.16
5	VISUAL_BASIC_VBA	209	5.01
6	JAVASCRIPT	188	4.51
7	SAS_STATA	122	2.92
8	SCALA	112	2.69
9	C_CPP_CSHARP	52	1.25
10	PHP	39	0.94
11	DOT_NET	33	0.79
12	MATLAB	23	0.55
13	JULIA	5	0.12

Plataformas de nuvem utilizadas:

```

/* Visão plataformas de nuvem utilizadas pelos respondentes */
WITH PLATAFORMAS AS (
    SELECT
        COUNTIF(USA_AZURE = TRUE) as AZURE,
        COUNTIF(USA_AWS = TRUE) as AWS,
        COUNTIF(USA_GOOGLE_CLOUD = TRUE) as GOOGLE_CLOUD
    FROM `sprint1-mvp.state_of_data.respostas`
    WHERE MORA_NO_BRASIL = true
)
SELECT PLATAFORMA,
    RESPONDENTES,
    ROUND((RESPONDENTES/total_respondentes_brasil)*100, 2) as Percentual
FROM PLATAFORMAS
UNPIVOT(RESPONDENTES
    FOR PLATAFORMA IN
    ( AZURE, AWS, GOOGLE_CLOUD))
ORDER BY RESPONDENTES DESC;

```

É possível identificar que as plataforma de nuvem mais utilizada é a AWS (*Amazon Web Services*), seguida de *Google Cloud* e *Microsoft Azure*.

Row	PLATAFORMA	RESPONDENTES	Percentual
1	AWS	1128	27.04
2	GOOGLE_CLOUD	708	16.97
3	AZURE	535	12.83

Bancos de dados utilizados:

```
/* Visão dos bancos de dados utilizados pelos respondentes */
WITH SGBDS AS (
    SELECT
        COUNTIF(USA_MYSQL = TRUE) as MYSQL,
        COUNTIF(USA_ORACLE = TRUE) as ORACLE,
        COUNTIF(USA_SQL_SERVER = TRUE) as SQL_SERVER,
        COUNTIF(USA_AMAZON_AURORA_RDS = TRUE) as AMAZON_AURORA_RDS,
        COUNTIF(USA_DYNAMODB = TRUE) as DYNAMODB,
        COUNTIF(USA_COACHDB = TRUE) as COACHDB,
        COUNTIF(USA_CASSANDRA = TRUE) as CASSANDRA,
        COUNTIF(USA_MONGODB = TRUE) as MONGODB,
        COUNTIF(USA_MARIADB = TRUE) as MARIADB,
        COUNTIF(USA_DATOMIC = TRUE) as DATOMIC,
        COUNTIF(USA_S3 = TRUE) as S3,
        COUNTIF(USA_POSTGRESQL = TRUE) as POSTGRESQL,
        COUNTIF(USA_ELASTICSEARCH = TRUE) as ELASTICSEARCH,
        COUNTIF(USA_DB2 = TRUE) as DB2,
        COUNTIF(USA_MICROSOFT_ACCESS = TRUE) as MICROSOFT_ACCESS,
        COUNTIF(USA_SQLITE = TRUE) as SQLITE,
        COUNTIF(USA_SYBASE = TRUE) as SYBASE,
        COUNTIF(USA_FIREBASE = TRUE) as FIREBASE,
        COUNTIF(USA_VERTICA = TRUE) as VERTICA,
        COUNTIF(USA_REDIS = TRUE) as REDIS,
        COUNTIF(USA_NEO4J = TRUE) as NEO4J,
        COUNTIF(USA_GOOGLE_BIGQUERY = TRUE) as GOOGLE_BIGQUERY,
        COUNTIF(USA_GOOGLE_FIRESTORE = TRUE) as GOOGLE_FIRESTORE,
        COUNTIF(USA_AMAZON_REDSHIFT = TRUE) as AMAZON_REDSHIFT,
        COUNTIF(USA_AMAZON_ATHENA = TRUE) as AMAZON_ATHENA,
        COUNTIF(USA_SNOWFLAKE = TRUE) as SNOWFLAKE,
        COUNTIF(USA_DATABRICKS = TRUE) as DATABRICKS,
        COUNTIF(USA_HBASE = TRUE) as HBASE,
        COUNTIF(USA_PRESTO = TRUE) as PRESTO,
        COUNTIF(USA_SPLUNK = TRUE) as SPLUNK,
        COUNTIF(USA_SAP_HANA = TRUE) as SAP_HANA,
        COUNTIF(USA_HIVE = TRUE) as HIVE,
        COUNTIF(USA_FIREBIRD = TRUE) as FIREBIRD
    FROM `sprint1-mvp.state_of_data.respostas`
    WHERE MORA_NO_BRASIL = true
)

SELECT SGBD,
    RESPONDENTES,
    ROUND((RESPONDENTES/total_respondentes_brasil)*100, 2) as Percentual
FROM SGBDS
UNPIVOT(RESPONDENTES
    FOR SGBD IN
        ( MYSQL, ORACLE, SQL_SERVER,
          AMAZON_AURORA_RDS, DYNAMODB, COACHDB,
          CASSANDRA, MONGODB, MARIADB, DATOMIC,
          S3, POSTGRESQL, ELASTICSEARCH, DB2,
          MICROSOFT_ACCESS, SQLITE, SYBASE,
          FIREBASE, VERTICA, REDIS,
          NEO4J, GOOGLE_BIGQUERY, GOOGLE_FIRESTORE,
          AMAZON_REDSHIFT, AMAZON_ATHENA,
          SNOWFLAKE, DATABRICKS, HBASE,
          PRESTO, SPLUNK, SAP_HANA,
          HIVE, FIREBIRD))
ORDER BY RESPONDENTES DESC;
```

Nesta categoria, é possível identificar um uso mais diversificado entre as opções. As principais soluções utilizadas são SQL Server, MySQL, PostgreSQL, Google BigQuery, Amazon S3, Databricks e Oracle.

Row	SGBD	RESPONDENTES	Percentual
1	SQL_SERVER	934	22.39
2	MYSQL	781	18.72
3	POSTGRESQL	738	17.69
4	GOOGLE_BIGQUERY	719	17.24
5	S3	583	13.98
6	DATABRICKS	521	12.49
7	ORACLE	470	11.27
8	AMAZON_ATHENA	317	7.6
9	AMAZON_REDSHIFT	308	7.38
10	MONGODB	302	7.24
11	HIVE	241	5.78
12	SAP_HANA	163	3.91
13	SQLITE	162	3.88
14	SNOWFLAKE	155	3.72
15	AMAZON_AURORA_RDS	147	3.52
16	DYNAMODB	127	3.04

Row	SGBD	RESPONDENTES	Percentual
17	MICROSOFT_ACCESS	122	2.92
18	PRESTO	115	2.76
19	ELASTICSEARCH	114	2.73
20	MARIADB	83	1.99
21	DB2	83	1.99
22	REDIS	77	1.85
23	FIREBASE	68	1.63
24	SPLUNK	45	1.08
25	CASSANDRA	35	0.84
26	HBASE	32	0.77
27	FIREBIRD	32	0.77
28	GOOGLE_FIRESTORE	28	0.67
29	NEO4J	27	0.65
30	SYBASE	22	0.53
31	DATOMIC	12	0.29
32	COACHDB	6	0.14
33	VERTICA	3	0.07

Ferramentas de visualização de dados utilizadas:

```

/* Visão das ferramentas de visualização de dados utilizadas pelos respondentes
WITH FERRAMENTAS AS (
    SELECT
        COUNTIF(USA_POWERBI = TRUE) as POWERBI,
        COUNTIF(USA_QLIK_VIEW_SENSE = TRUE) as QLIK_VIEW_SENSE,
        COUNTIF(USA_TABLEAU = TRUE) as TABLEAU,
        COUNTIF(USA_METABASE = TRUE) as METABASE,
        COUNTIF(USA_SUPERSET = TRUE) as SUPERSET,
        COUNTIF(USA_REDASH = TRUE) as REDASH,
        COUNTIF(USA_MICROSTRATEGY = TRUE) as MICROSTRATEGY,
        COUNTIF(USA_IBM_ANALYTICS_COGNOS = TRUE) as IBM_ANALYTICS_COGNOS,
        COUNTIF(USA_SAP_BO = TRUE) as SAP_BO,
        COUNTIF(USA_ORACLE_BUSINESS = TRUE) as ORACLE_BUSINESS,
        COUNTIF(USA_AMAZON_QUICKSIGHT = TRUE) as AMAZON_QUICKSIGHT,
        COUNTIF(USA_SALESFORCE_EINSTEIN = TRUE) as SALESFORCE_EINSTEIN,
        COUNTIF(USA_MODE = TRUE) as MODE,
        COUNTIF(USA_ALTERYX = TRUE) as ALTERYX,
        COUNTIF(USA_BIRST = TRUE) as BIRST,
        COUNTIF(USA_LOOKER = TRUE) as LOOKER,
        COUNTIF(USA_GOOGLE_DATA_STUDIO = TRUE) as GOOGLE_DATA_STUDIO,
        COUNTIF(USA_SAS_VISUAL_ANALYTICS = TRUE) as SAS_VISUAL_ANALYTICS,
        COUNTIF(USA_GRAFANA = TRUE) as GRAFANA,
        COUNTIF(USA_SPOTFIRE = TRUE) as SPOTFIRE,
        COUNTIF(USA_PENTAHO = TRUE) as PENTAHO,
        COUNTIF(USA_EXCEL_PLANILHA_GOOGLE = TRUE) as EXCEL_PLANILHA_GOOGLE,
        COUNTIF(NAO_USA_FERRAMENTA_VISUALIZACAO = TRUE) as NAO_USA_FERRAMENTA_VISUALIZACAO
    FROM `sprint1-mvp.state_of_data.respostas`
    WHERE MORA_NO_BRASIL = true
)

SELECT FERRAMENTA,
    RESPONDENTES,
    ROUND((RESPONDENTES/total_respondentes_brasil)*100, 2) as Percentual
FROM FERRAMENTAS
UNPIVOT(RESPONDENTES
    FOR FERRAMENTA IN
        ( POWERBI, QLIK_VIEW_SENSE,
          TABLEAU, METABASE,
          SUPERSET, REDASH,
          MICROSTRATEGY, IBM_ANALYTICS_COGNOS,
          SAP_BO, ORACLE_BUSINESS,
          AMAZON_QUICKSIGHT, SALESFORCE_EINSTEIN,
          MODE, ALTERYX, BIRST,
          LOOKER, GOOGLE_DATA_STUDIO,
          SAS_VISUAL_ANALYTICS,
          GRAFANA, SPOTFIRE, PENTAHO,
          EXCEL_PLANILHA_GOOGLE, NAO_USA_FERRAMENTA_VISUALIZACAO ))
ORDER BY RESPONDENTES DESC;

```

Nesta categoria, é possível identificar que poucos respondentes utilizam ferramentas de visualização de dados. A ferramenta mais utilizada é a PowerBI, que é utilizada por apenas 2,3% dos respondentes, seguido de Google Data Studio, Tableau e Metabase.

Row	FERRAMENTA	RESPONDENTES	Percentual
1	POWERBI	96	2.3
2	GOOGLE_DATA_STUDIO	38	0.91
3	TABLEAU	33	0.79
4	METABASE	26	0.62
5	NAO_USA_FERRAMENTA	26	0.62
6	LOOKER	19	0.46
7	EXCEL_PLANILHA_GOOGLE	19	0.46
8	GRAFANA	18	0.43
9	QLIK_VIEW_SENSE	8	0.19
10	ALTERYX	7	0.17
11	PENTAHO	7	0.17
12	SUPERSET	4	0.1

Row	FERRAMENTA	RESPONDENTES	Percentual
13	SALESFORCE_EINSTEIN	3	0.07
14	SAS_VISUAL_ANALYTICS	3	0.07
15	REDASH	2	0.05
16	IBM_ANALYTICS_COGNOS	2	0.05
17	SAP_BO	2	0.05
18	AMAZON_QUICKSIGHT	2	0.05
19	MICROSTRATEGY	1	0.02
20	SPOTFIRE	1	0.02
21	ORACLE_BUSINESS	0	0.0
22	MODE	0	0.0
23	BIRST	0	0.0

Conclusão:

Foi possível identificar as principais tecnologias ou ferramentas utilizadas pelos respondentes, de acordo com seu tipo:

- As principais linguagens utilizadas são SQL e Python;
- A plataforma de nuvem mais utilizada é a AWS (Amazon Web Services), seguida de Google Cloud e Microsoft Azure;
- As principais soluções de banco de dados utilizadas são SQL Server, MySQL, PostgreSQL, Google BigQuery, Amazon S3, Databricks e Oracle;
- A ferramenta de visualização mais utilizada é a Microsoft PowerBI, seguido de Google Data Studio, Tableau e Metabase.

Etapa 6: Conclusão

Através da criação de um pipeline de dados utilizando ferramentas da plataforma *Google Cloud* e dados provenientes da pesquisa "State of Data Brazil 2022", disponibilizados na plataforma *Kaggle*, foi possível realizar a análise dos dados e obter respostas relevantes para as questões propostas no objetivo deste projeto.

Uma das principais descobertas com a análise dos dados foi a relevância da educação formal na área de dados no Brasil. Os dados mostraram que a maioria dos profissionais da área possui formação superior, e que programas de pós-graduação, mestrado ou doutorado são relevantes no mercado.

Além disso, foi possível identificar os cargos e funções mais comuns no mercado brasileiro de dados e ter um panorama das faixas salariais por nível de cargo, o que pode ser valioso para pessoas que buscam entender melhor as possibilidades de carreira na área ou negociar seus salários atuais.

A satisfação no trabalho é um fator crucial em qualquer carreira e a análise dos dados revelou que a maioria dos profissionais de dados no Brasil se encontra satisfeita com seus empregos atuais. Isso pode ser um estímulo adicional para aqueles que consideram uma mudança de carreira ou um investimento em educação na área de dados.

Por fim, foi possível identificar as tecnologias mais comuns em uso na área de dados no Brasil, trazendo informações relevantes para pessoas que desejam se manter atualizadas e relevantes no mercado de trabalho.

Em resumo, este projeto ofereceu uma visão abrangente e atualizada do mercado de trabalho brasileiro na área de dados, com a expectativa de que as informações obtidas possam ser úteis para pessoas que busquem ingressar ou evoluir sua carreira nesse segmento dinâmico e em constante evolução.