
Faculdade
XPe
RELATÓRIO
PROJETO
APLICADO
PÓS-GRADUAÇÃO

FELIPE FERREIRA TEODORO

XP EDUCAÇÃO

RELATÓRIO DO PROJETO APLICADO

ANÁLISE DE VAGAS DE TRABALHOS RELACIONADAS À DADOS

Relatório de Projeto Aplicado desenvolvido para fins de conclusão do curso de Cientista de Dados.

Orientador (a):

Professor Marcos Prochnow

Brasília-DF

30-08-2024

Sumário

1. CANVAS do Projeto Aplicado

Abaixo segue a Figura 1, que representa todas as etapas do propostas pelo curso de Ciência de Dados.

Figura 1 – Canvas do Projeto Aplicado.

1.1 Desafio

1.1.1 Análise de Contexto

A área de dados (cientista e analista), dentro da Tecnologia da Informação, tem se mostrado uma poderosa ferramenta para as empresas que visam buscar conhecimento e insights das informações geradas por elas, proporcionando decisões estratégicas de negócio. E junto a isso, a demanda dentro da área tem aumentado exponencialmente. Contudo, muitas das pessoas que estão iniciando na área, tem uma certa dificuldade em encontrar o primeiro emprego, devido a falta de conhecimento e informações sobre o mercado de trabalho.

O objetivo principal deste estudo visa auxiliar essas pessoas a obterem informações essenciais sobre o mercado de trabalho da área de dados (cientista, administrador de banco de dados, analista, engenheiro, etc.). A base possui dados referentes à salários por região, tempo de experiência desejado, informações das empresas, sites de recrutamento, requisitos desejados, entre outros.

Uma das maiores dificuldades de quem está ingressando na área é de se localizar dentro de tantas vagas e áreas diferentes. Saber onde procurar, quais requisitos e habilidades são solicitados, as diferenças de uma região para outra e os salários, são fatores importantes para que o candidato consiga saber onde focar seus estudos, e como se preparar para uma entrevista de emprego.

A abordagem desse problema será focada em uma análise dos dados disponibilizados, a fim de identificar tendências, padrões e insights que possibilitam uma orientação aos candidatos no momento de buscar um emprego. O dashboard que será elaborado para isso, conterá as seguintes informações:

- Quais são os salários médios para diferentes cargos e regiões;
- Quais empresas estão contratando e quais são seus requisitos;
- Quais sites de recrutamento são mais utilizados e oferecem as melhores oportunidades;
- Quais são as tendências de demanda no mercado e quais habilidades estão em alta.

O dashboard permitirá uma análise visual e prática sobre o tema em si, permitindo a interpretação das informações e auxiliando os(as) candidatos(as) a encontrarem as melhores oportunidades para suas carreiras.

O desenvolvimento do dashboard será realizado com um tratamento inicial dos dados diretamente na base (.csv) e tratamento secundário na linguagem Python. Então, as informações serão exportadas para um banco de dados no MySQL, e posteriormente será realizada uma conexão desse banco de dados com o Power BI, onde serão elaborados os modelos semânticos, as visualizações e os gráficos.

A Figura 2 apresenta o modelo de matriz CSD (certezas, suposições e dúvidas).

A Figura 3 apresenta a Análise do Contexto do Problema – POEMS.

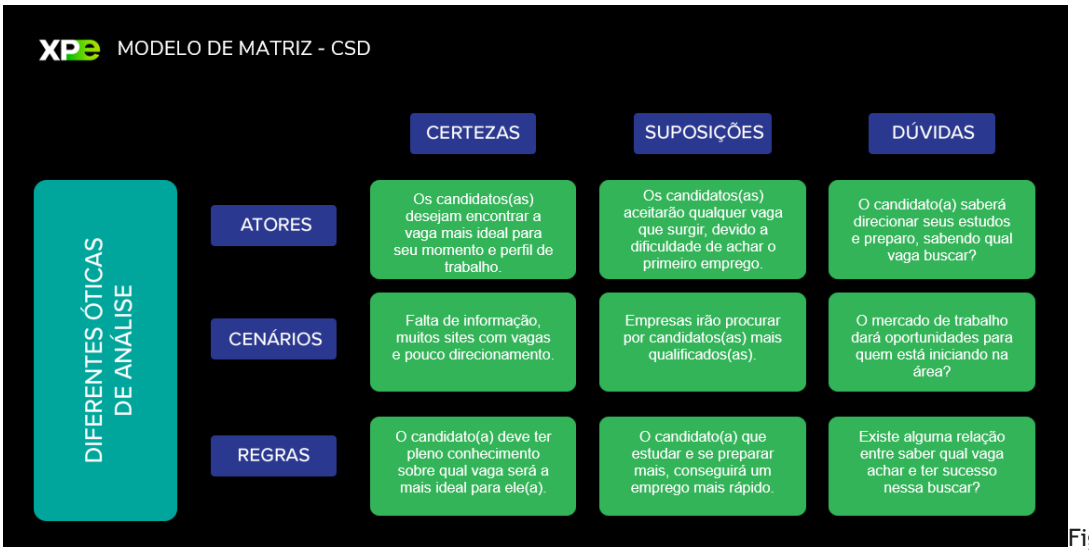


Figura 2 – Matriz CSD

Figura 3 – POEMS

1.1.2 Personas

A persona a qual será referenciada neste trabalho se chama Ana Júlia, uma profissional formada em Administração de Empresas pela USP (Universidade de São Paulo) e que possui pós-graduação em Marketing Digital pela FGV (Fundação Getúlio Vargas). No último 08 (oito) anos, Ana trabalhou na área de Marketing em uma empresa de tecnologia, onde desenvolveu um grande interesse pela área de análise e ciência de dados. O interesse foi tanto, que ela já completou um *bootcamp* de Ciência de Dados na IGTI (Instituto de Gestão e Tecnologia da Informação). Nesse curso e com sua

experiência de trabalho, Ana viu muito potencial nesta área, resultando em todo esse interesse para migrar de carreira.

Ana enfrenta vários desafios para quem está iniciando nesta área, assim como qualquer outra pessoa enfrentaria. Ao aprofundar seus conhecimentos na área, ela percebeu que há uma grande diferença entre os conhecimentos e habilidades que ela adquiriu, com o que o mercado pede hoje nas vagas anunciadas. Mesmo já tendo conhecimento em fundamentos de Python, SQL e visualização de dados, Ana percebeu que as vagas pedem conhecimentos em tecnologias muito específicas e experiência práticas as quais ela ainda não possui. Isso acaba dificultando um pouco no momento de direcionar os estudos, a fim de se especializar mais. Ana também sente dificuldade em entender quais empresas estão mais abertas em contratar pessoas que estão no início de carreira e com pouca experiência.

A Figura 4 representa o Mapa de Empatia, o qual faz uso de 04 (quatro) situações para melhor compreender a persona referenciada neste trabalho: Ouvir, Pensar e Sentir, Ver, Falar e Fazer.

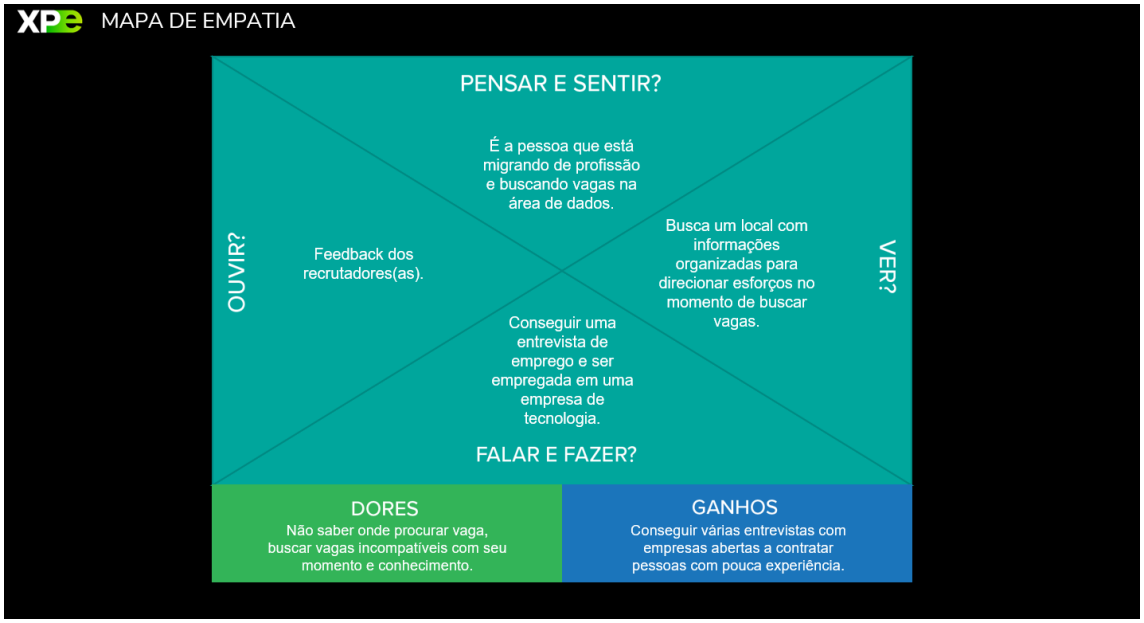


Figura 4 – Mapa de Empatia

1.1.3 Benefícios e Justificativas

Conforme já demonstrado anteriormente, pessoas que estão migrando de profissão ou começando carreira na área de dados, costumam enfrentar vários desafios, alguns deles até significativos, em sua busca pelo primeiro emprego. Seja pela falta de direcionamento ou de uma plataforma que demonstre de maneira

evidente e clara as informações do mercado de trabalho, dificultando a plena compreensão e entendimento sobre as exigências reais das empresas que estão contratando. Tudo isso faz com que os candidatos sejam dependentes de informações incompletas ou possuam um entendimento limitado sobre as experiências e habilidades exigidas no início da carreira.

Para isso, usaremos técnicas de Ciência de Dados, a fim de extrair informações relevantes que possam auxiliar essas pessoas a se posicionarem de maneira mais eficaz no mercado de trabalho. Ao realizar uma análise mais aprofundada nas informações brutas sobre as vagas, poderemos extrair tendências e padrões importantes, os quais possibilitarão os(as) candidatos(as) direcionam seus esforços com mais precisão, aumentando assim as chances de encontrar algo com seu perfil.

A análise desses dados também possibilitará encontrar quais as dificuldades enfrentadas por quem está iniciando na área, como quais qualificações permitirão encontrar empregos com salários melhores, quais locais costumam contratar pessoas com mais experiência, preparar o candidato(a) a elaborar um currículo mais voltado para cada situação. A análise permitirá que os candidatos(as) tomem decisões baseadas em dados e informações concretas.

A Tabela 1 apresentará o mapeamento das ações da Ana, conforme a metodologia *Blueprint*, permitindo entender melhor o problema.

Itens	Detalhamento
Objetivos	Localizar padrões e tendências nas vagas para iniciantes na área.
Atividades	Encontrar vagas compartíveis com o perfil da persona.
Questões	Funcionará também para quem está iniciando, e não migrando de carreira?
Barreiras	Algumas informações essenciais não estão presentes na base.
	Ações do Cliente
Funcionalidades	Analisar de forma mais clara os indicadores, padrões e tendências das vagas de emprego.
Interação	Dashboard interativo e com filtros para selecionar.
Mensagem	Saber direcionar os esforços nas regiões, vagas e empresas certas.
Onde Ocorre	Computador, notebook, tablet, celular.
Tarefas Pendentes	Como aplicar o conhecimento adquirido no momento de buscar as vagas.
Tarefas Escondidas	Como olhar e interpretar os dados do dashboard.
Processos de Suporte	Power BI.
Saída Desejável	Conhecimento e insights para buscar vagas ideais.

Tabela 1 - *Blueprint*

A análise proposta para essa base de dados permitirá dar valor para o momento de busca do primeiro emprego, permitindo que o(a) candidato(a) seja mais assertivo quando for procurar por um trabalho na área de dados. Sendo assim, a Figura 5 irá demonstrar os principais benefícios deste estudo.

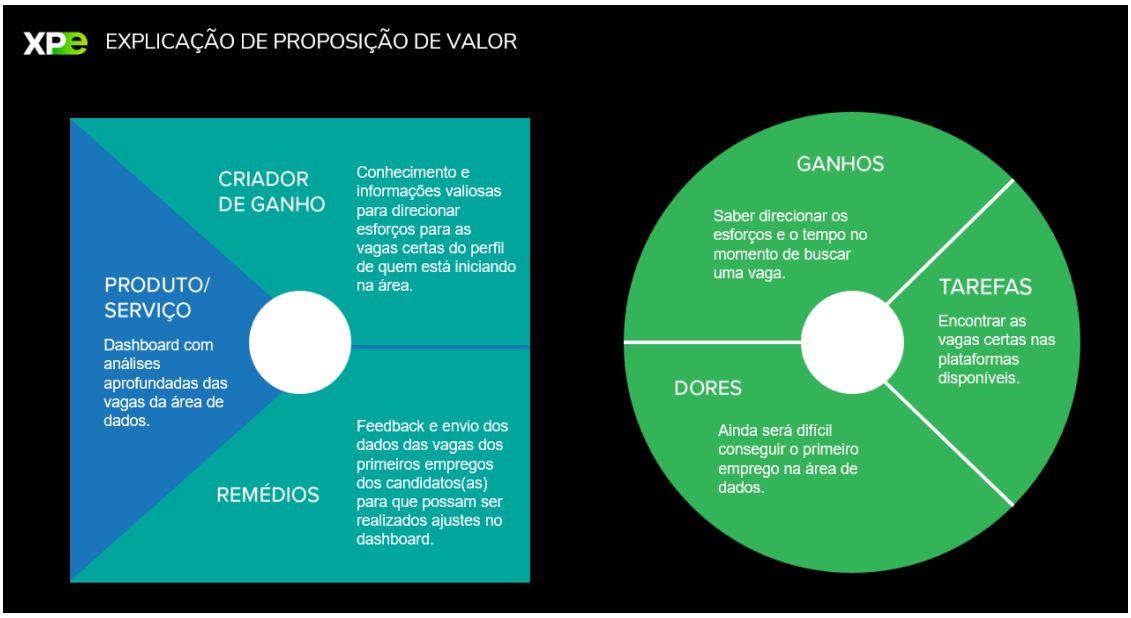


Figura 5 – Explicação de Proposição de Valor

1.1.4 Hipóteses

Já vimos que uma das principais dificuldades para quem está buscando primeiro emprego na área de dados é a falta de experiência prática e conhecimentos e habilidades específicas que são exigidas pelas empresas contratantes. Muitas dessas vagas pedem que o(a) candidato(a) já possuam alguma experiência prévia em projetos práticos do meio corporativo, algo que acaba sendo um impeditivo para quem está buscando o primeiro emprego.

Existe também uma escassez em capacitações e mentorias voltadas para profissionais que estão iniciando na área. Muitos cursos as vezes não focam no que realmente importa, e acabam apenas despejando coisas que nem serão usadas no dia-a-dia de um trabalho prático. Pelo fato de que as empresas preferem pessoas com experiência prévia, os(as) candidatos(as) que estão iniciando acabam apresentando muita insegurança e ficam desmotivados.

Outra hipótese válida para este estudo é a identificação de padrões e tendências nas vagas oferecidas nas plataformas, de forma a permitir um maior preparo e melhor direcionamento dos esforços por parte dos candidatos, fornecendo um conhecimento prévio necessário no momento da busca do primeiro emprego. A Tabela 2 abaixo irá resumir de forma mais clara essas hipóteses.

Observações	Hipóteses
Candidatos(as) gastam tempo e esforços nas vagas erradas	Falta de conhecimento no momento da busca do primeiro emprego. Não saber quais filtros aplicar, quais regiões buscar, quais empresas estão mais abertas as contratar alguem sem experiência prévia.
Insegurança no momento da entrevista	Devido à falta de experiência, os candidatos(as) apresentam muita insegurança e nervosismo no momento das entrevistas.
Não saber em que se especializar	Com tantas vagas pedindo habilidades e conhecimentos específicos, o(a) candidato(a) fica confuso com o que deve se especializar.
Desconhecimento sobre quais salários são justos e quais não são	Sem ter conhecimento prévio sobre as vagas, o(a) candidato(a) não saberá se o salário oferecido é justo ou se está sendo explorado.

Tabela 2 – Observações e Hipóteses

Assim como mencionado anteriormente no desenvolvimento deste trabalho, os(as) candidatos(as) não possuem um conhecimento/direcionamento prévio ideal na busca do primeiro emprego. As ofertas de cursos de hoje estão mais voltadas para o conhecimento e habilidades da área do que para a busca de trabalho para pessoas inexperientes. Após uma análise detalhada da visão que a persona, Ana, possui disso, foi elaborada a Matriz de Priorização de Ideias (Tabela 4). Esta matriz foi elaborada a partir do método BASICO, que analise os seguintes pontos:

- Benefícios;
- Abrangência;
- Satisfação;
- Investimento;
- Cliente;
- Operacionalidade.

Escala	B - Benefícios	A - Abrangência	S - Satisfação	I - Investimentos	C - Cliente	O - Operacionalidade
5	De vital importância	Total (de 70 a 100%)	Muito grande	Pouquíssimo investimento	Nenhum impacto	Muito fácil
4	Significativo	Muito grande (de 40 a 70%)	Grande	Algum investimento	Impacto pequeno	Fácil
3	Razoável	Razoável (de 20 a 40%)	Média	Médio investimento	Médio impacto	Média facilidade
2	Poucos benefícios	Pequena (de 5 a 20%)	Pequena	Alto investimento	Impacto grande	Difícil
1	Algum benefício	Muito pequena	Quase não é notada	Altíssimo investimento	Impacto muito grande no cliente	Muito difícil

Tabela 3 – Notas da Matriz BASICO

Fonte: https://engenhariaexercicios.com.br/gestao-de-qualidade/matriz-gutbasicoconceito-

aplicacao-das-matrizes-priorizacao/

Ideias	B	A	S	I	C	O	Total	Priorização
Elaboração de dashboard apresentando de forma visual e organizada as informações das vagas	5	5	5	3	5	3	**26**	3º
Localizar tendências e padrões das informações contidas na base	4	4	4	2	4	2	**20**	6º
Fornecer conhecimento prévio para auxiliar os(as) candidatos(as) a buscarem o primeiro emprego	5	5	4	5	5	3	**27**	2º
Identificar as empresas e regiões do mundo em que estão mais abertas a contratar pessoas com pouca experiência	4	4	3	5	3	5	**24**	4º
Fornecer uma noção de salários para pessoas menos experientes	3	3	2	5	3	5	**21**	5º
Fornecer informações sobre quais habilidades as empresas buscam nos candidatos com menos experiência	5	4	5	5	4	5	**28**	1º

Tabela 4 – Matriz BASICO

1.2 Solução

1.2.1 Objetivo SMART

Este estudo terá o objetivo de trazer informação e conhecimento para as pessoas que estão migrando de profissão ou iniciando na área de dados. As informações sobre as vagas que estão na base de dados serão organizadas de forma a trazer um olhar diferente para os padrões e tendências apresentadas. O(a) candidato(a) terá um norte no momento de direcionar seus esforços, sabendo onde e como procurar o primeiro emprego.

A intenção é que a pessoa consiga aumentar em pelo menos 50% o número de entrevistas e retornos dos(as) recrutadores(as), pois ao analisar as informações de forma mais analítica e estratégica, o(a) candidato saberá onde e como buscar um trabalho, e não perderá tempo em vagas, empresas ou regiões que estão menos dispostas a contratar alguém sem experiência prévia.

Os resultados deverão surgir de forma rápida, mas dependerão de como o(a) candidato(a) irá se dedicar a buscar vagas, como cada empresa opera e qual seu tempo de retorno das ofertas de vagas.

1.2.2 Premissas e Restrições

Este estudo apresentou uma solução já considerando algumas restrições que foram identificadas no início do levantamento de necessidades dos profissionais que estão em transição de carreira ou iniciando na área de dados.

Primeiramente, algumas informações essenciais não estão na base de dados. Informações que trariam maior valor agregado ao produto final que o presente estudo desenvolverá, tais como: o tempo que a pessoa está ou ficou empregada na empresa, quantas etapas foram necessárias para a contratação e como foram cada uma dessas etapas.

Uma segunda restrição é que quem está iniciando na área se preocupa mais em se preparar tecnicamente, com os conhecimentos da área, do que se preparar para as entrevistas em si, ou aprender mais sobre quais empresas são mais propícias em contratar iniciantes.

Outra restrição é a falta de histórico ou mais bases de dados referentes ao assunto. Não há muitos estudos ou relatórios disponíveis para basear um estudo aprofundado no assunto, tornando um maior desafio o escopo deste trabalho.

Dessa forma, será apresentada abaixo a Tabela 5 com a matriz de riscos do projeto, a qual demonstra as consequências geradas pelas restrições apresentadas no presente tópico.

Risco Identificado	Impacto Potencial	Ações Preventivas	Ações Corretivas
Falta de informação na base de dados	O dashboard deixará de apresentar informações essenciais	Incentivar a divulgação por parte das empresas das informações faltantes	Fazer busca minuciosa de mais fontes de dados
Candidatos(as) não buscam conhecimento sobre entrevistas e/ou empresas	Gastam tempo e energia em vagas que não irão contratar eles(as)	Demonstrar nas redes sociais as vantagens da busca por este tipo de conhecimento	Recrutadores incentivar os candidatos nas entrevistas a se informarem mais sobre os processos e vagas em si
Ausência de relatórios e estudos na área	Ausência de material para realizar os estudos	Incentivar as empresas a divulgarem este tipo de informação	Postar nas redes sociais (Linkedin) sobre as vantagens de possuir este tipo de conhecimento

Tabela 5 - Matriz de Riscos

1.2.3 Backlog de Produto

Nesta etapa, será descrito detalhadamente cada passo que deve ser realizado para a elaboração da proposta final deste trabalho. Utilizamos o *Trello* (Figura 6) para demonstrar os passos do backlog do produto, o que está em andamento e o que já está concluído.

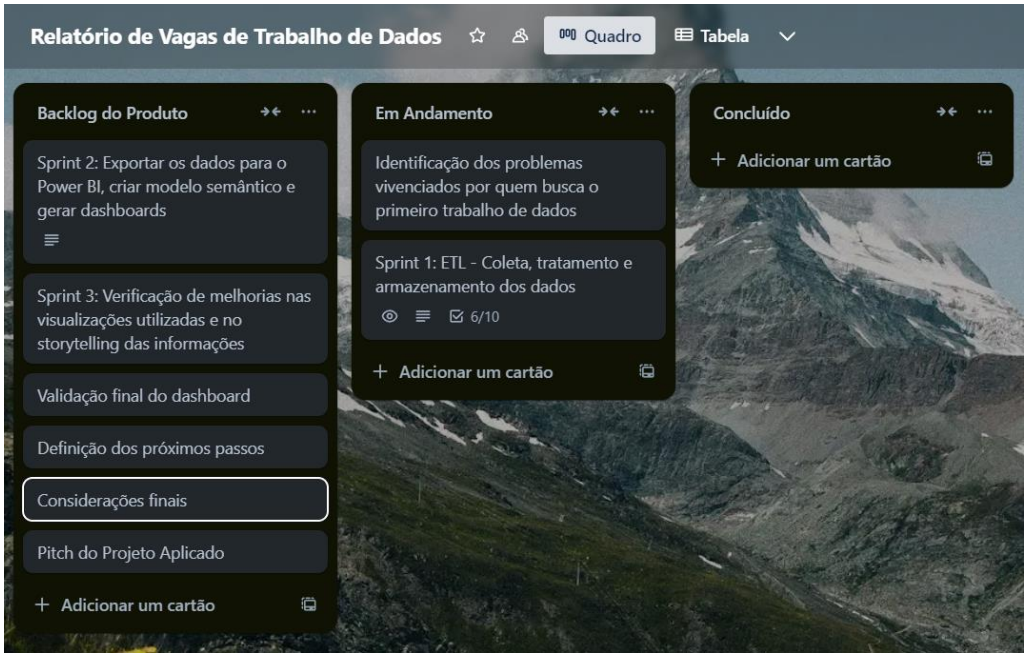


Figura 6 - Trello

2. Área de Experimentação

Esta seção tem a finalidade de mostrar a execução das atividades propostas no Backlog do Produto. Serão mostrados em detalhes todas as etapas, dividas em 03 (três) Sprints. Cada um deles representa um conjunto de tarefas a serem executadas, de forma que, juntos, representam a solução do problema proposto.

A ferramenta escolhida para isso foi o *Trello*, onde através dele, todas as tarefas foram divididas. Cada cartão representa uma etapa ou *Sprint*. Dentro de cada um, estão as anotações e *checklist* das atividades.

2.1 Sprint 1

2.1.1 Solução

- Evidência do planejamento:

  Todo planejamento do Sprint 1 foi catalogado dentro do Trello. Nele foram criados vários *Checklists* para organizar cada etapa. Isso permite uma maior organização das etapas, de forma a visualizar de forma detalhada tudo o que tem que ser feito, além de registrar todas as tarefas. A Figura 7 mostra em detalhes o resultado deste planejamento.

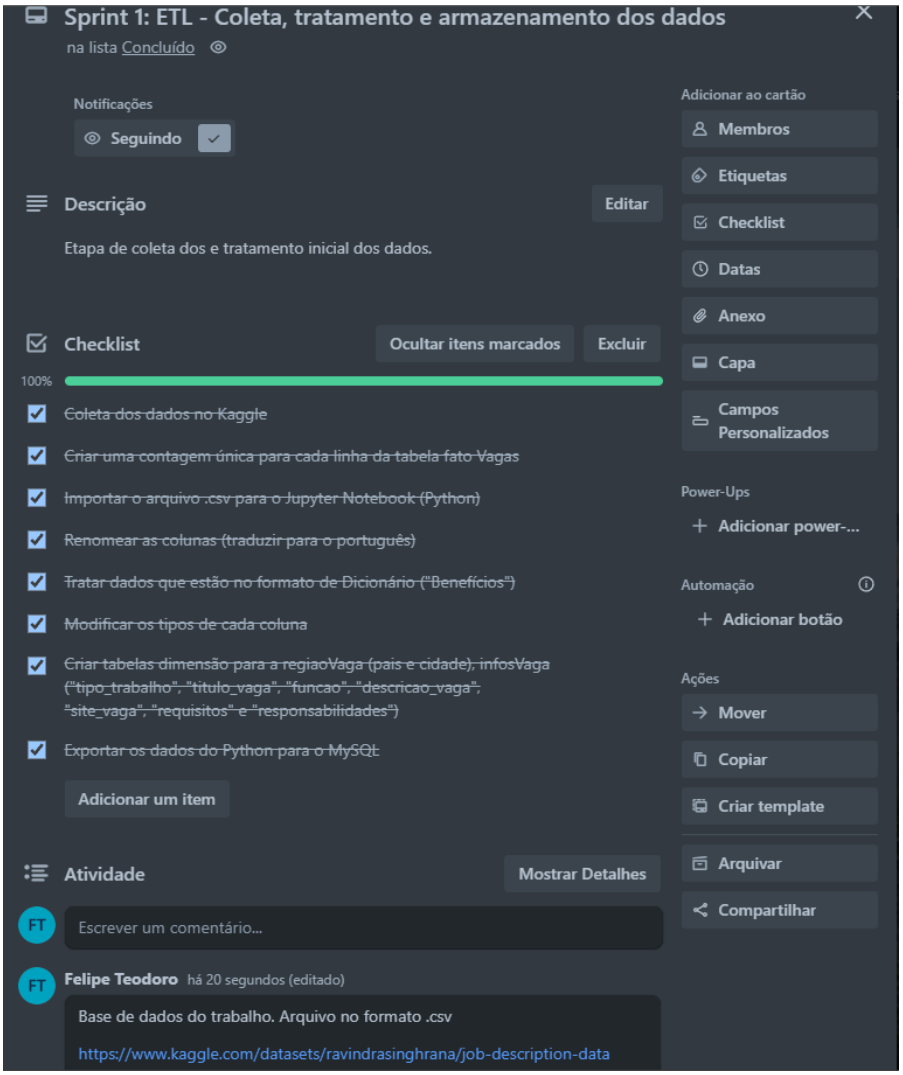


Figura 7 – Trello (Sprint 1)

A base de dados utilizada foi obtida no site Kaggle, através do link 'https://www.kaggle.com/datasets/ravindrasinghrana/job-description-dataset'. O arquivo baixado está no formato "csv".

Foram realizados alguns tratamentos diretamente na base, tais como filtros na coluna "Role" para retornar apenas vagas relacionadas a profissões de dados e criação de uma coluna com o ID de cada vaga.

- Evidência da execução de cada requisito:

Após o tratamento inicial na base do Excel, foi realizada a importação do arquivo .csv dentro do Jupyter Notebook, para que outros tratamentos e preparo para exportação para a base de dados do MySQL.

As Figuras 8, 9 e 10 mostram um pouco mais sobre a importação e quais tratamentos foram realizados na linguagem Python.

```
[1]: import pandas as pd

     df = pd.read_csv(r"C:\Users\felip\OneDrive\Área de Trabalho\Pós Graduação - Ciência de Dados\Projeto Aplicado\Base de Dados\base_dados.csv", sep=';', enc

[2]: df = df.rename(columns={"Job Id": "id_vaga", "Experience":"tempo_experiencia", "Qualifications": "formacao", "Salary Range (Min)": "salario_min",
                            "Salary Range (Max)": "salario_max", "location": "cidade", "Country":"pais", "Work Type": "tipo_trabalho", "Company Size": "tamanh
                            "Job Posting Date":"data_vaga", "Preference":"sexo", "Contact Person":"nome_recrutador","Contact":"contato_recrutador",
                            "Job Title":"titulo_vaga", "Role":"funcao", "Job Portal":"site_vaga","Job Description":"descricao_vaga", "skills":"requisitos",
                            "Responsibilities": "responsabilidades", "Company":"empresa", "Benefits":"beneficios", "Company Profile":"perfil_empresa"})

[3]: df['beneficios'] = df['beneficios'].str.replace("{", "", regex=False).str.replace("}", "", regex=False)
     df['beneficios']

[3]: 0        'Tuition Reimbursement, Stock Options or Equit...
     1        'Childcare Assistance, Paid Time Off (PTO), Re...
     2        'Employee Assistance Programs (EAP), Tuition R...
     3        'Employee Referral Programs, Financial Counsel...
     4        'Childcare Assistance, Paid Time Off (PTO), Re...
                                    ...
     73000    'Childcare Assistance, Paid Time Off (PTO), Re...
     73001    'Employee Referral Programs, Financial Counsel...
     73002    'Employee Referral Programs, Financial Counsel...
     73003    'Life and Disability Insurance, Stock Options ...
     73004    'Transportation Benefits, Professional Develop...
     Name: beneficios, Length: 73005, dtype: object

[4]: df = df.drop('perfil_empresa', axis = 1)
```

Figura 8 – Tratamento dos dados na linguagem Python (Parte 1)

```
[5]: df['salario_min'] = df['salario_min'].str.replace('$', '', regex=False)
     df['salario_max'] = df['salario_max'].str.replace('$', '', regex=False)

     df['salario_min'] = df['salario_min'].str.replace('.', '', regex=False)
     df['salario_max'] = df['salario_max'].str.replace('.', '', regex=False)

     df['salario_min'] = df['salario_min'].str.replace(',00', '', regex=False)
     df['salario_max'] = df['salario_max'].str.replace(',00', '', regex=False)

[6]: df['salario_min'] = df['salario_min'].astype(float)
     df['salario_max'] = df['salario_max'].astype(float)

     df['data_vaga'] = pd.to_datetime(df['data_vaga'], format='%d/%m/%Y')

[7]: df.isna().sum()

[7]: id_vaga              0
     tempo_experiencia    0
     formacao             0
     salario_min          0
     salario_max          0
     cidade               0
     pais                 0
     tipo_trabalho        0
     tamanho_empresa      0
     data_vaga            0
     sexo                 0
     nome_recrutador      0
     contato_recrutador   0
     titulo_vaga          0
     funcao               0
     site_vaga            0
     descricao_vaga       0
     beneficios           0
     requisitos           0
     responsabilidades    0
     empresa              0
     dtype: int64
```

Figura 9 – Tratamento dos dados na linguagem Python (Parte 2)

```
[7]: df.isna().sum()

[7]: id_vaga              0
     tempo_experiencia    0
     formacao             0
     salario_min          0
     salario_max          0
     cidade               0
     pais                 0
     tipo_trabalho        0
     tamanho_empresa      0
     data_vaga            0
     sexo                 0
     nome_recrutador      0
     contato_recrutador   0
     titulo_vaga          0
     funcao               0
     site_vaga            0
     descricao_vaga       0
     beneficios           0
     requisitos           0
     responsabilidades    0
     empresa              0
     dtype: int64

[8]: df.duplicated().sum()

[8]: 0

[9]: df.info()

     <class 'pandas.core.frame.DataFrame'>
     RangeIndex: 73005 entries, 0 to 73004
     Data columns (total 21 columns):
      #   Column              Non-Null Count  Dtype
     ---  ------              --------------  -----
      0   id_vaga             73005 non-null  int64
      1   tempo_experiencia   73005 non-null  object
      2   formacao            73005 non-null  object
      3   salario_min         73005 non-null  float64
      4   salario_max         73005 non-null  float64
      5   cidade              73005 non-null  object
      6   pais                73005 non-null  object
      7   tipo_trabalho       73005 non-null  object
      8   tamanho_empresa     73005 non-null  int64
      9   data_vaga           73005 non-null  datetime64[ns]
      10  sexo                73005 non-null  object
```

Figura 10 – Tratamento dos dados na linguagem Python (Parte 3)

Percebe-se que os tratamentos iniciais são referentes a tratamentos básicos, como renomeação das colunas, substituição de *strings*, *drop* de coluna, conversão do tipo de coluna e verificação de valores repetidos e nulos. Ressalto a importância dessa ultima etapa, pois valores repetidos e nulos mascaram muitos resultados durante as análises e criação dos indicadores.

É importante ressaltar a codependência das etapas de tratamento de dados. Por exemplo: a conversão das colunas de salário mínimo e máximo não seria possível sem antes substituir as *strings* "$", "." e ",00", pois elas estavam caracterizando a coluna como formatadas em *string*, e precisamos que essa coluna esteja no formato *float*.

Após estes processos citados acima, vamos preparar para realizar a exportação da base de dados já tratada para o MySQL. A Figura 11 mostra com detalhes como foi iniciada essa conexão, onde foram importadas as bibliotecas necessárias, passados parâmetros para a conexão (*user*, *password* e *host*), criação da partição de conexão com o MySQL e criação da base de dados "vagas_emprego" dentro do MySQL.

```
[25]: user = 'root'
      password = 'Fft270192!'
      host = 'localhost'
      string_conexao = f'mysql://{user}:{password}@{host}'

      engine = create_engine(string_conexao)
      conn = engine.connect()

[23]: database = 'vagas_emprego'
      query = text(f'CREATE SCHEMA {database}')
      conn.execute(query)
```

Figura 11 – Conexão Python com MySQL

Após todo esse processo, agora podemos exportar as tabelas para o banco de dados MySQL. Teremos 03 (três) tabelas no MySQL no fim deste processo: fato_vagas, dim_infosVagas e dim_regiaoVaga. A base que está tratada em Python possui todas as informações necessárias para elaboração de cada uma das tabelas que serão utilizadas. Para isso, precisaremos realizar algumas etapas.

Inicialmente, enviaremos a base principal diretamente para o MySQL. Ela servirá como tabela fato, após passar por algumas modificações dentro do banco de dados. A Figura 12 mostra como foi feita a exportação da base.

```
df.to_sql('stg_vagas', con=conn, schema='vagas_emprego', if_exists='replace', index=False)
```

```
73005
```

Figura 12 – Exportando a base para o MySQL

Para as tabelas dimensões, precisaremos realizar transformações na nossa base principal. Essas transformações consistem em selecionar as colunas referentes a cada dimensão, retirar as duplicadas e criar uma coluna de index, a qual utilizaremos para realizar a conexão com a tabela fato, através das chaves primárias e secundárias. As Figuras 13 e 14 detalham como foi realizado este tratamento.

```
[47]: regiaoVaga = df[['cidade', 'pais']]

      duplicatas = regiaoVaga.duplicated()

      # Exibir as linhas duplicadas
      duplicatas_linhas = regiaoVaga[duplicatas]
      print(duplicatas_linhas)
```

```
                cidade        pais
21             Nairobi       Kenya
24            Chisinau     Moldova
27              London          UK
46     Port-au-Prince       Haiti
60          Georgetown      Guyana
...                ...         ...
73000       Stockholm      Sweden
73001            Apia       Samoa
73002          Mbabane    Eswatini
73003     Antananarivo  Madagascar
73004             Lima        Peru

[72789 rows x 2 columns]
```

```
[49]: num_duplicatas = regiaoVaga.duplicated().sum()
      print(f"Número de linhas duplicadas: {num_duplicatas}")
```

```
Número de linhas duplicadas: 72789
```

```
[55]: regiaoVaga = regiaoVaga.drop_duplicates()
      regiaoVaga['id_regiaoVaga'] = regiaoVaga.index
      regiaoVaga
```

[55]:

	cidade	pais	id_regiaoVaga
0	Yaounde	Cameroon	0
1	London	UK	1
2	Papeete	French Polynesia	2
3	George Town	Cayman Islands	3
4	Quito	Ecuador	4
...	...	...	...
1043	Bandar Seri Begawan	Brunei	1043
1080	Pristina	Kosovo	1080

Figura 13 – Elaboração da dim_regiaoVaga

```
infosVaga = df[['tipo_trabalho', 'titulo_vaga', 'funcao', 'descricao_vaga', 'site_vaga', 'requisitos', 'responsabilidades']]

duplicatas = infosVaga.duplicated()

# Exibir as linhas duplicadas
duplicatas_linhas = infosVaga[duplicatas]
print(duplicatas_linhas)

num_duplicatas = infosVaga.duplicated().sum()
print(f"Número de linhas duplicadas: {num_duplicatas}")

infosVaga = infosVaga.drop_duplicates()
infosVaga['id_infosVaga'] = infosVaga.index
infosVaga
```

Figura 14 – Elaboração da dim_infosVaga

Após o todo tratamento para elaboração das da “dim_infoVagas” e “dim_regiaoVaga”, o mesmo comando realizado para a base “df”, que renomeamos como “stg_vendas”, será feito também com as tabelas dimensões, a fim de enviar elas para o MySQL. A Figura 15 ilustra como foi feito isso:

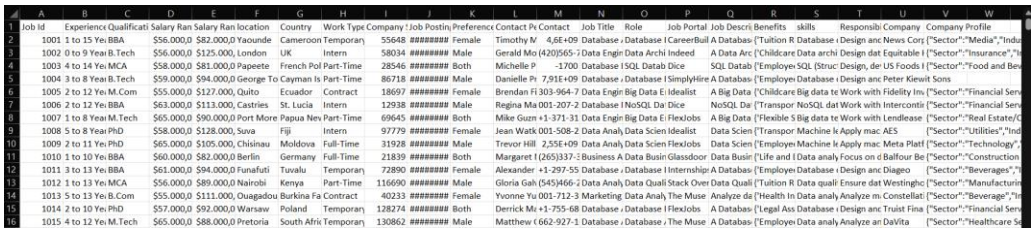
```
[67]: infosVaga.to_sql('dim_infosVaga', con=conn, schema='vagas_emprego', if_exists='replace', index=False)

[67]: 1200

[69]: regiaoVaga.to_sql('dim_regiaovaga', con=conn, schema = 'vagas_emprego', if_exists='replace', index=False)

[69]: 216
```

Figura 15 – Comando de envio das tabelas para o MySQL

- Evidência dos resultados:

Conforme mostrado na Figura 7, nosso principal objetivo no Sprint 1 foi exportar os dados do Python para o MySQL. Inicialmente tínhamos uma base no formato “.csv”, baixada do site Kaggle. Essa base não continha os tratamentos necessários para envio ao MySQL, e por isso, tivemos que realizar vários deles antes.

Como evidência principal dos resultados do Sprint 1, podemos demonstrar nas Figuras 16 e 17, a base como era no Excel, e a mesma base tratada, dividida em uma tabela que servirá para elaboração da tabela fato no Sprint 2, e as duas tabelas dimensões.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	Job Id	Experience	Qualificati	Salary Ran	Salary Ran	location	Country	Work Type	Company !	Job Postin	Preference	Contact Pe	Contact	Job Title	Role	Job Portal	Job Descri	Benefits	skills	Responsibi	Company	Company Profile	
2	1001	1 to 15 Yea	BBA	$56.000,0	$82.000,0	Yaounde	Cameroon	Temporary	55648	########	Female	Timothy M	4,6E+09	Database A	Database I	CareerBuil	A Database	{'Tuition R	Database A	Design and	News Corp	{"Sector":"Media","Indu	
3	1002	0 to 9 Year	B.Tech	$56.000,0	$125.000,	London	UK	Intern	58034	########	Male	Gerald Mo	(420)565-7	Data Engin	Data Archi	Indeed	A Data Arc	{'Childcare	Data archi	Design dat	Equitable I	{"Sector":"Insurance","I	
4	1003	4 to 14 Yea	MCA	$58.000,0	$81.000,0	Papeete	French Pol	Part-Time	28546	########	Both	Michelle P	-1700	Database I	SQL Datab	Dice	SQL Datab	{'Employee	SQL (Struc	Design, de	US Foods I	{"Sector":"Food and Bev	
5	1004	3 to 8 Year	B.Tech	$59.000,0	$94.000,0	George To	Cayman Is	Part-Time	86718	########	Male	Danielle Pr	7,91E+09	Database A	Database I	SimplyHire	A Database	{'Employee	Database a	Design and	Peter Kiewit Sons		
6	1005	2 to 12 Yea	M.Com	$55.000,0	$127.000,	Quito	Ecuador	Contract	18697	########	Female	Brendan Fi	303-964-7	Data Engin	Big Data E	Idealist	A Big Data	{'Childcare	Big data te	Work with	Fidelity Inv	{"Sector":"Financial Serv	
7	1006	2 to 12 Yea	BBA	$63.000,0	$113.000,	Castries	St. Lucia	Intern	12938	########	Male	Regina Ma	001-207-2	Database I	NoSQL Da	Dice	NoSQL Da	{'Transpor	NoSQL dat	Work with	Intercontir	{"Sector":"Financial Serv	
8	1007	1 to 8 Year	M.Tech	$65.000,0	$90.000,0	Port More	Papua Nev	Part-Time	69645	########	Both	Mike Guzn	+1-371-31	Data Engin	Big Data E	FlexJobs	A Big Data	{'Flexible S	Big data te	Work with	Lendlease	{"Sector":"Real Estate/C		
9	1008	5 to 8 Year	PhD	$58.000,0	$128.000,	Suva	Fiji	Intern	97779	########	Female	Jean Watk	001-508-2	Data Analy	Data Scien	Idealist	Data Scien	{'Transpor	Machine le	Apply mac	AES	{"Sector":"Utilities","Ind		
10	1009	2 to 11 Yea	PhD	$65.000,0	$105.000,	Chisinau	Moldova	Full-Time	31928	########	Male	Trevor Hill	2,55E+09	Data Analy	Data Scien	FlexJobs	Data Scien	{'Employee	Machine le	Apply mac	Meta Platf	{"Sector":"Technology",		
11	1010	1 to 10 Yea	BBA	$60.000,0	$82.000,0	Berlin	Germany	Full-Time	21839	########	Both	Margaret I	(265)337-3	Business A	Data Busin	Glassdoor	Data Busin	{'Life and I	Data analy	Focus on d	Balfour Be	{"Sector":"Construction		
12	1011	3 to 13 Yea	BBA	$61.000,0	$94.000,0	Funafuti	Tuvalu	Temporary	72890	########	Female	Alexander	+1-297-55	Database A	Database I	Internship	A Database	{'Employee	Database a	Design and	Diageo	{"Sector":"Beverages","I		
13	1012	1 to 13 Yea	MCA	$56.000,0	$89.000,0	Nairobi	Kenya	Part-Time	116690	########	Male	Gloria Gah	(545)466-2	Data Analy	Data Quali	Stack Over	Data Quali	{'Tuition R	Data quali	Ensure dat	Westinghc	{"Sector":"Manufacturin		
14	1013	5 to 13 Yea	B.Com	$55.000,0	$111.000,	Ouagadou	Burkina Fa	Contract	40233	########	Female	Yvonne Yu	001-712-3	Marketing	Data Analy	The Muse	Analyze da	{'Health In	Data analy	Analyze m	Constellati	{"Sector":"Beverage","In		
15	1014	2 to 10 Yea	PhD	$57.000,0	$92.000,0	Warsaw	Poland	Temporary	128274	########	Both	Derrick Mc	+1-755-68	Database A	Database I	FlexJobs	A Database	{'Legal Ass	Database a	Design and	Truist Fina	{"Sector":"Financial Serv		
16	1015	4 to 12 Yea	M.Tech	$65.000,0	$88.000,0	Pretoria	South Afric	Temporary	130862	########	Male	Matthew (	662-927-1	Database A	Database I	The Muse	A Database	{'Employee	Data analy	Analyze an	DaVita	{"Sector":"Healthcare Se		

Figura 16 – Base de dados inicial

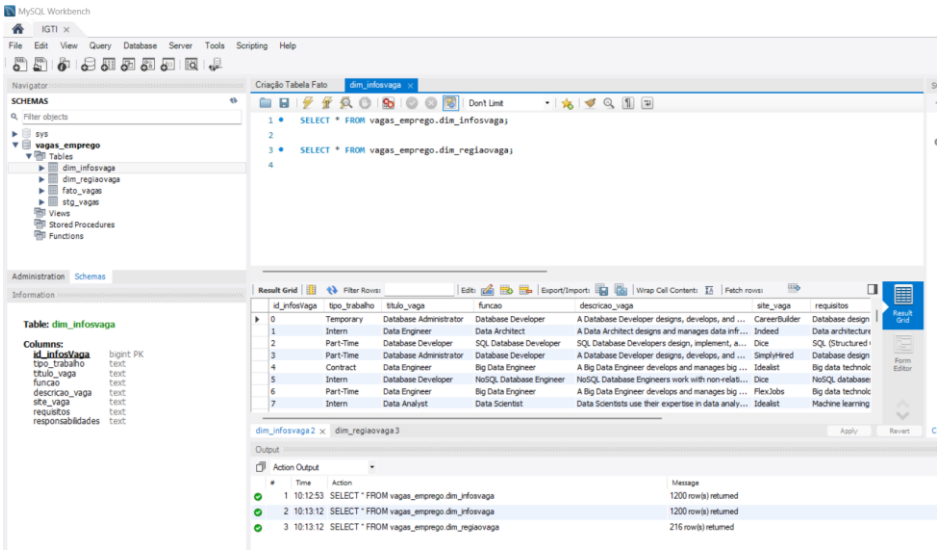


Figura 17 – Base no banco de dados MySQL

2.1.2 Lições Aprendidas

No Sprint 1, foi notada a importância do tratamento dos dados antes que de fato começássemos a trabalhar com eles. Todo processo possui uma codependência desta etapa de ETL. Um exemplo é colocar as colunas de salários mínimo e máximo, para que então fosse possível a conversão das colunas para o formato *float*.

Outro fato que devemos destacar, é que o tratamento inicial dos dados permite o modelo possuir uma performance maior na outra ponta, que no presente caso, será na construção do Dashboard. Como utilizaremos o Power BI, é de suma importância que o tratamento dentro do Power Query seja o mínimo possível, para que então nosso painel possa performar de maneira eficiente.

Notou-se também a necessidade de seguir outro caminho no momento de gerar as tabelas dimensões dentro do MySQL. Durante as aulas, foi ensinada uma maneira muito eficiente, que percorria em loop a nossa base de dados, retirava as duplicadas, gerada o index e fazia o envio para o MySQL, assim como mostrado na Figura 18. No entanto, essa metodologia servia para tabelas dimensão com 01 (uma) coluna, e não foi possível utilizá-la para gerar tabelas com mais de uma coluna. As Figuras 13 e 14 mostram em detalhes como foram elaboradas as tabelas, e durante o Sprint 02 (dois), entraremos no conceito de linguagem SQL e o que foi feito para gerar as chaves primárias e secundárias entre as tabelas fato e dimensões.

```
[47]: def insert_registro_tabela_simples(lista_dados, tabela):
          coluna_tabela = tabela[3:]

          for registro in lista_dados:
              try:
                  query = f"""insert into {tabela} ({coluna_tabela})
                              values('{registro}')"""
                  conn.execute(query)

                  print(f"Registro inserido com sucesso: {registro}")
              except Exception as e:
                  print(f"Não foi possível inserir o registro {registro}. O erro encontrado foi: {e}")

[48]: lista_estado = dados_licenca_medicos['estado_colaborador'].drop_duplicates().reset_index(drop=True)
      tabela_banco = 'tb_estado'
      insert_registro_tabela_simples(lista_dados = lista_estado,
                                     tabela = tabela_banco)

      Registro inserido com sucesso: Rio Grande do Sul
      Registro inserido com sucesso: Ceará
      Registro inserido com sucesso: Rio Grande do Norte
      Registro inserido com sucesso: Tocantins
      Registro inserido com sucesso: Roraima
      Registro inserido com sucesso: Santa Catarina
      Registro inserido com sucesso: Pernambuco
      Registro inserido com sucesso: Alagoas
      Registro inserido com sucesso: Piauí
      Registro inserido com sucesso: Paraíba
      Registro inserido com sucesso: Minas Gerais
```

Figura 18 – Envio dos dados para MySQL (Aula Pós Graduação XP Educação)

2.2 Sprint 2

2.2.1 Solução

- Evidência do planejamento:

Nesta etapa, o objetivo será realizar o tratamento da tabela fato (“fato_vagas”), para então realizarmos a conexão do MySQL com o Power BI, onde serão realizadas as etapas finais do Sprint 2 e a maioria das etapas do Sprint 3.

O planejamento foi executado da mesma forma que o Sprint 1, onde temos um cartão no Trello, apresentando as etapas do Sprint 2, conforme mostra a Figura 19.

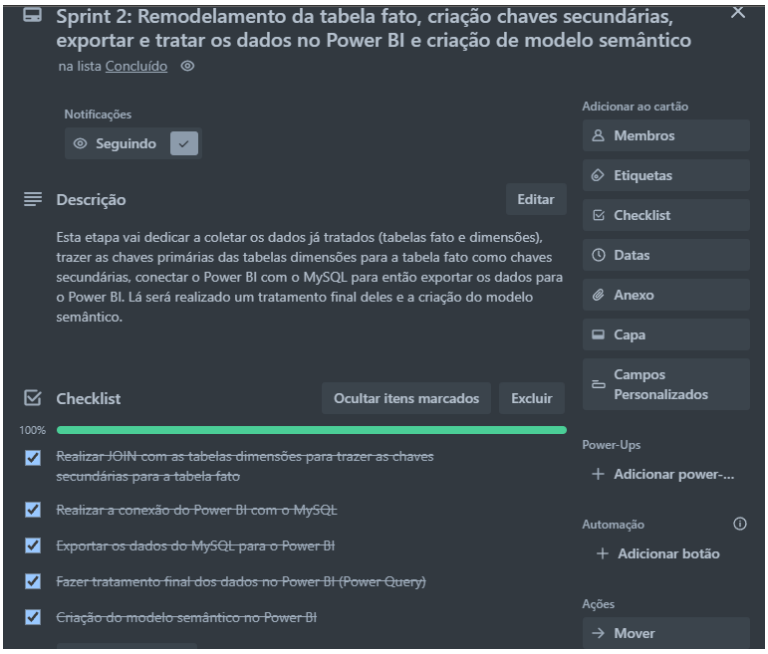


Figura 19 – Planejamento do Sprint 2 (Cartão do Trello)

- Evidência da execução de cada requisito:

A principal ideia do Sprint 2 será realizar a conexão do MySQL com o Power BI. Mas para isso, será necessário criar a tabela fato (“fato_vagas”). Não haverá nenhum tratamento, mas sim, preparar a tabela para criação do modelo semântico dentro do Power BI. Será necessário substituir algumas colunas da tabela fato, que até então é a “stg_vagas”, pelas chaves secundárias (chaves primárias das tabelas dimensões).

A forma escolhida para realizar essa operação foi através da linguagem SQL, onde não só serão retornadas apenas as colunas específicas da tabela fato, como também a implementação do recurso “Inner Join”, para que será possível trazer as colunas de “ID” de cada uma das tabelas dimensão. A Figura 20 mostra como foi realizado esse processo.

Criação Tabela Fato* | dim_infosvaga

Don't Limit

```sql
CREATE TABLE fato_vagas AS

SELECT id_vaga, B.id_regiaoVaga, C.id_infosvaga, tempo_experiencia, formacao, salario_min, salario_max,
       tamanho_empresa, data_vaga, sexo, nome_recrutador, contato_recrutador,
       beneficios, empresa
FROM stg_vagas as A

INNER JOIN dim_regiaovaga B
ON A.cidade = B.cidade
AND A.pais = B.pais

INNER JOIN dim_infosvaga C
ON A.tipo_trabalho = C.tipo_trabalho
AND A.titulo_vaga = C.titulo_vaga
AND A.funcao = C.funcao
AND A.descricao_vaga = C.descricao_vaga
AND A.site_vaga = C.site_vaga
AND A.requisitos = C.requisitos
AND A.responsabilidades = C.responsabilidades;
```

Figura 20 – Criação da tabela “fato_vagas”

Conforme citado no parágrafo acima, colunas específicas foram selecionadas, e foi feito um “Inner Join”, para retornar as colunas ID das tabelas dimensões. Foi realizado também o comando “CREATE TABLE”, o qual irá criar uma nova tabela a partir da consulta realizada na Figura 20.

As próximas etapas serão de realizar a conexão entre o MySQL e o Power BI e exportar as tabelas, para que assim seja possível trabalhar com os dados tratados e criar as visualizações e análises. As Figuras 21 e 22 mostram em detalhes como foi realizada essa conexão.

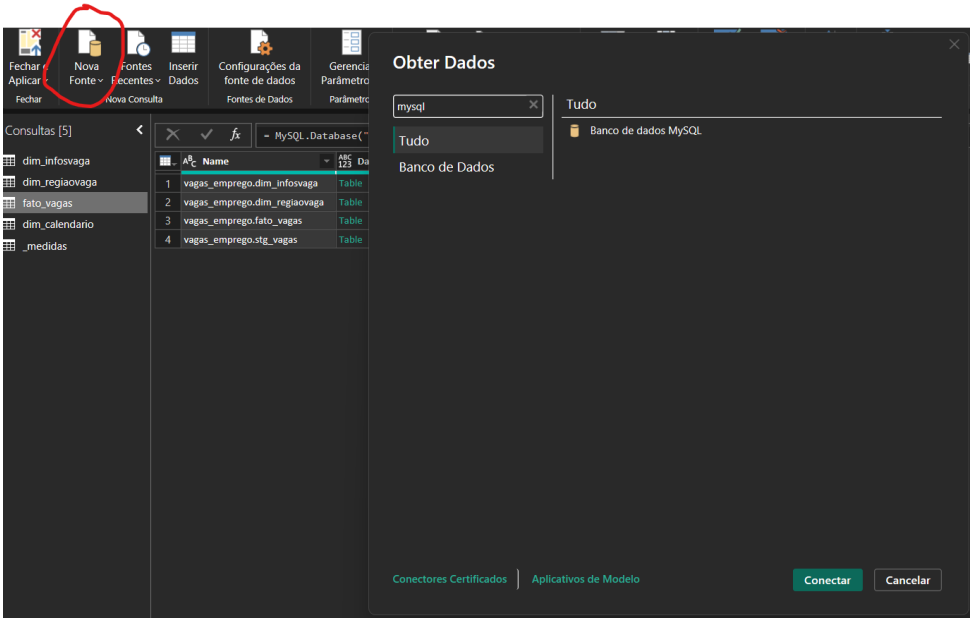


Figura 21 – Nova fonte de dados (Power Query)

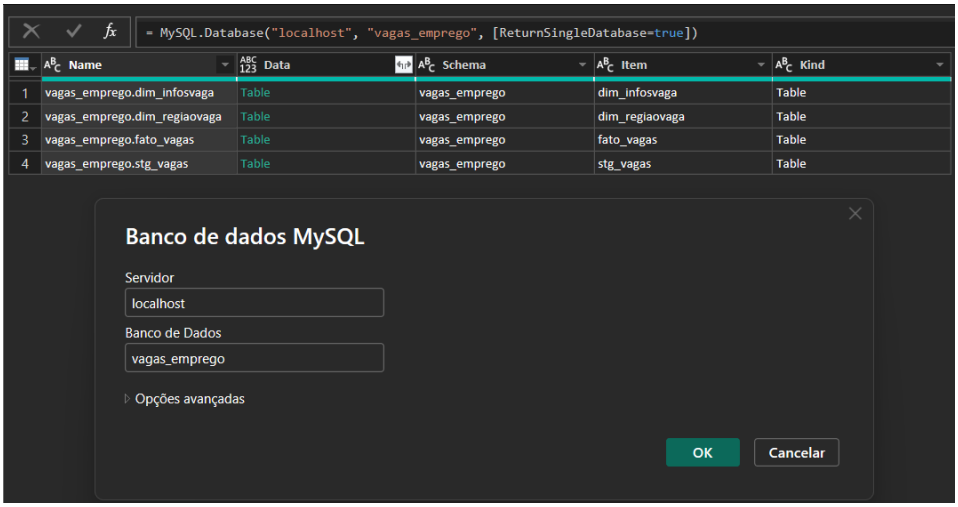


Figura 22 – Conexão do MySQL com o Power Query

Dessa maneira, qualquer modificação realizada na fonte de dados (MySQL), irá refletir dentro do Power BI.

Após isso, foi realizado um tratamento final dos dados. Vale ressaltar que todo processo de tratamento já foi realizado na linguagem Python, e aqui, será realizada apenas uma adequação, para fins de performance, das colunas ID's das tabelas fato e dimensão. Essas tabelas foram consideras como números inteiros, mas o ideal é que elas sejam classificadas como *string*, para que assim o Power BI não realize nenhum tipo de agrupamento (soma, multiplicação, média) desnecessário nelas. A Figura 23 mostra como esta etapa foi executada (o mesmo procedimento foi realizado para todas as tabelas).

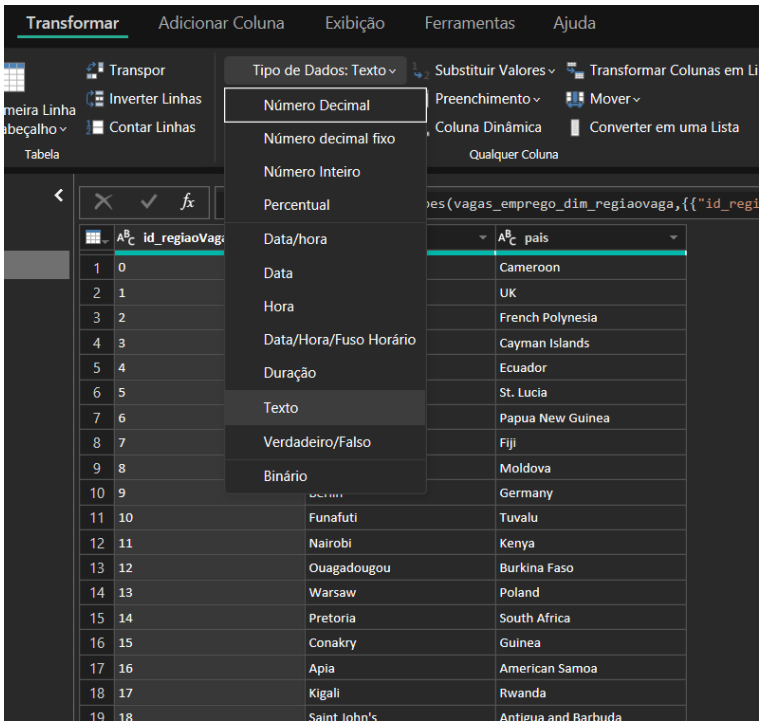


Figura 23 – Tratamento final dos dados

Para a última etapa do *Sprint* 2, será elaborado o modelo semântico do projeto. Este modelo representará as relações entre as tabelas, permitindo que informações sejam filtradas de uma para a outra.

Dentro da etapa de ETL (*Extract, Transform and Load*), foram criadas chaves primárias e secundárias em cada uma das tabelas (primária em todas as tabelas e secundária apenas na tabela fato). Serão essas chaves que permitirão um relacionamento entre tabelas.

Deve-se destacar a importância desta etapa, pois o modelo semântico permite que a tabela fato, a qual normalmente é a que possui uma maior quantidade de linhas, consiga performar de maneira mais eficiente, já que poupamos ela de um excesso de informações (colunas), pois o modelo buscará tudo através das chaves das tabelas dimensão. A Figura 24 mostra como foi feito o modelo semântico.

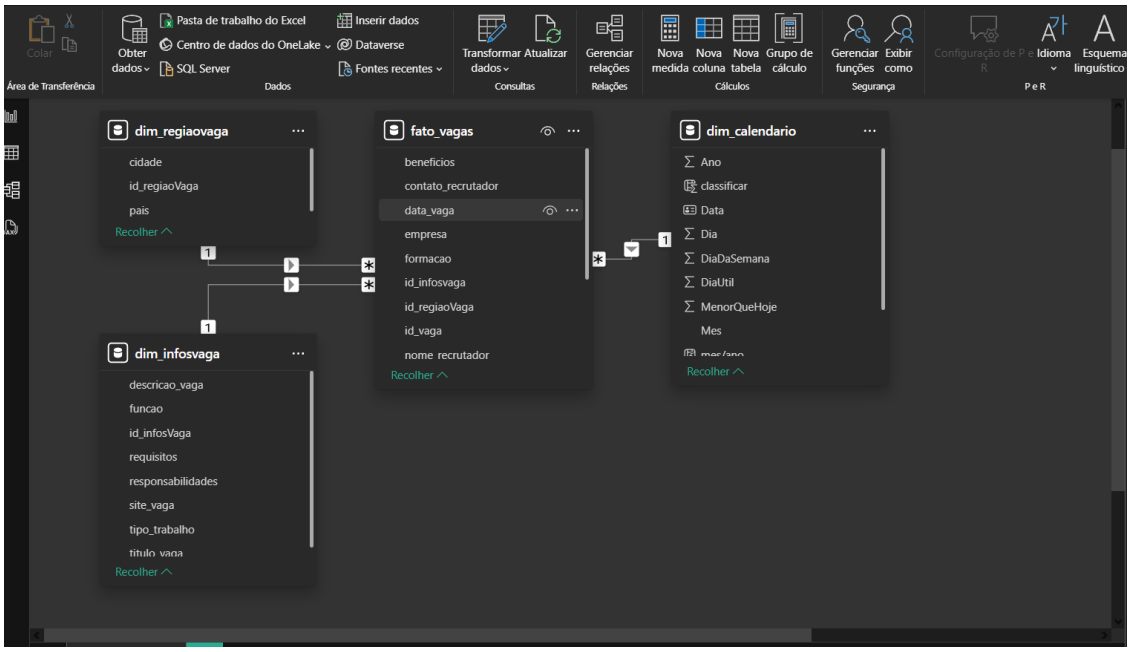


Figura 24 – Modelo Semântico no Power BI

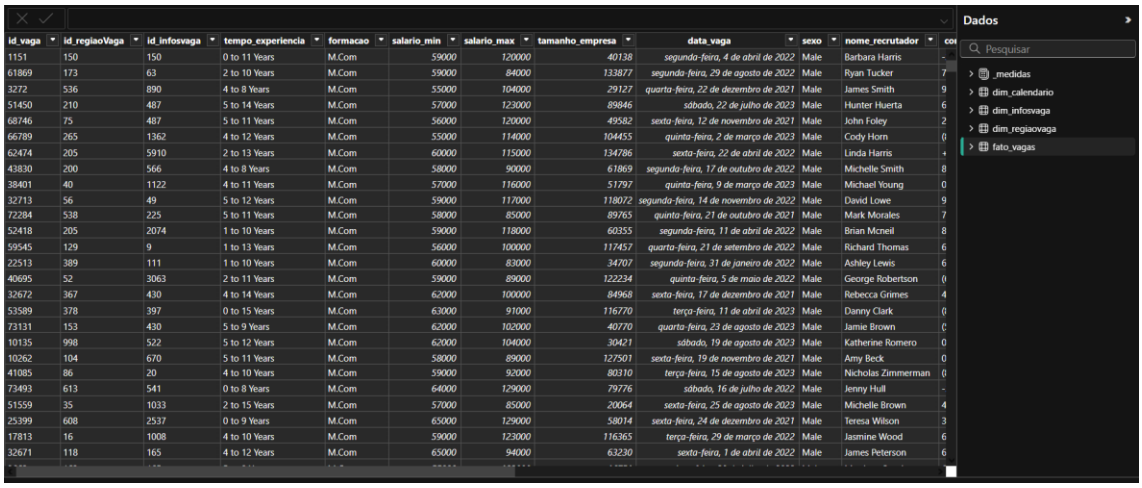
Todo este processo foi realizado dentro da área de “Exibição de Modelos” do Power BI.

- Evidência dos resultados:

Com todas as etapas do *Sprint* 2 realizadas, agora o modelo está pronto para iniciar a elaboração dos visuais e análises as quais foram propostas para resolver o problema da persona deste trabalho, Ana.

Como evidência, as tabelas estão disponíveis para que agora seja possível trabalhar de maneira mais analítica, onde serão elaborados indicadores, visuais, tabelas e tudo disponível que possa contribuir em transformar a informação em conhecimento.

A Figura 25 mostra como ficaram as tabelas (tabela fato no exemplo) após todos os processos já realizados.



id_vaga	id_regiaoVaga	id_infosvaga	tempo_experiencia	formacao	salario_min	salario_max	tamanho_empresa	data_vaga	sexo	nome_recrutador
1151	150	150	0 to 11 Years	M.Com	59000	120000	40138	segunda-feira, 4 de abril de 2022	Male	Barbara Harris
61869	173	63	2 to 10 Years	M.Com	59000	84000	133877	segunda-feira, 29 de agosto de 2022	Male	Ryan Tucker
3272	536	890	4 to 8 Years	M.Com	55000	104000	29127	quarta-feira, 22 de dezembro de 2021	Male	James Smith
51450	210	487	5 to 14 Years	M.Com	57000	123000	89846	sábado, 22 de julho de 2023	Male	Hunter Huerta
68746	75	487	5 to 11 Years	M.Com	56000	120000	49582	sexta-feira, 12 de novembro de 2021	Male	John Foley
66789	265	1362	4 to 12 Years	M.Com	55000	114000	104455	quinta-feira, 2 de março de 2023	Male	Cody Horn
62474	205	5910	2 to 13 Years	M.Com	60000	115000	134786	sexta-feira, 22 de abril de 2022	Male	Linda Harris
43830	200	566	4 to 8 Years	M.Com	58000	90000	61869	segunda-feira, 17 de outubro de 2022	Male	Michelle Smith
38401	40	1122	4 to 11 Years	M.Com	57000	116000	51797	quinta-feira, 9 de março de 2023	Male	Michael Young
32713	56	49	5 to 12 Years	M.Com	59000	117000	118072	segunda-feira, 14 de novembro de 2022	Male	David Lowe
72284	538	225	5 to 11 Years	M.Com	58000	85000	89765	quinta-feira, 21 de outubro de 2021	Male	Mark Morales
52418	205	2074	1 to 10 Years	M.Com	59000	118000	60355	segunda-feira, 11 de abril de 2022	Male	Brian Mcneil
59545	129	9	1 to 13 Years	M.Com	56000	100000	117457	quarta-feira, 21 de setembro de 2022	Male	Richard Thomas
22513	389	111	1 to 10 Years	M.Com	60000	83000	34707	segunda-feira, 31 de janeiro de 2022	Male	Ashley Lewis
40695	52	3063	2 to 11 Years	M.Com	59000	89000	122234	quinta-feira, 5 de maio de 2022	Male	George Robertson
32672	367	430	4 to 14 Years	M.Com	62000	100000	84968	sexta-feira, 17 de dezembro de 2021	Male	Rebecca Grimes
53589	378	397	0 to 15 Years	M.Com	63000	91000	116770	terça-feira, 11 de abril de 2023	Male	Danny Clark
73131	153	430	5 to 9 Years	M.Com	62000	102000	40770	quarta-feira, 23 de agosto de 2023	Male	Jamie Brown
10135	998	522	5 to 12 Years	M.Com	62000	104000	30421	sábado, 19 de agosto de 2023	Male	Katherine Romero
10262	104	670	5 to 11 Years	M.Com	58000	89000	127501	sexta-feira, 19 de novembro de 2021	Male	Amy Beck
41085	86	20	4 to 10 Years	M.Com	59000	92000	80310	terça-feira, 15 de agosto de 2023	Male	Nicholas Zimmerman
73493	613	541	0 to 8 Years	M.Com	64000	129000	79776	sábado, 16 de julho de 2022	Male	Jenny Hull
51559	35	1033	2 to 15 Years	M.Com	57000	85000	20064	sexta-feira, 25 de agosto de 2023	Male	Michelle Brown
25399	608	2537	0 to 9 Years	M.Com	65000	129000	58014	sexta-feira, 24 de dezembro de 2021	Male	Teresa Wilson
17813	16	1008	4 to 10 Years	M.Com	59000	123000	116365	terça-feira, 29 de março de 2022	Male	Jasmine Wood
32671	118	165	4 to 12 Years	M.Com	65000	94000	63230	sexta-feira, 1 de abril de 2022	Male	James Peterson

Figura 25 – Modelo final dos dados brutos

2.2.2 Lições Aprendidas

No *Sprint* 2 foi necessário realizar uma etapa de adaptação, conforme citado no tópico 2.1.2 deste trabalho. Para criação das chaves secundárias da tabela fato, o processo que normalmente seria realizado no tratamento dos dados antes da exportar para o MySQL, foi feito com a linguagem SQL, dentro do banco de dados.

Visualizou-se também a necessidade de realizar um tratamento final dentro do Power BI, transformando as colunas de ID's de todas as tabelas em *strings*. Isso resultará em um aumento de performance, já que o Power BI não realizará procedimentos desnecessários com essas colunas, tais como agrupamentos.

2.3 Sprint 3

2.3.1 Solução

- Evidência do planejamento:
- Evidência da execução de cada requisito:
- Evidência dos resultados:

2.3.2 Lições Aprendidas

3. Considerações Finais

3.1 Resultados

Por meio de um texto detalhado, apresente os principais resultados alcançados pelo seu Projeto Aplicado.

Cite os pontos positivos e negativos, as dificuldades enfrentadas e as experiências vivenciadas durante todo o processo.

3.2 Contribuições

Apresente quais foram as contribuições que o seu Projeto Aplicado trouxe para que o Desafio proposto fosse solucionado.

Cite, por exemplo, as inovações, as vantagens sobre os similares, as melhorias alcançadas, entre outros.

3.3 Próximos passos

Descreva quais são os próximos passos que poderão contribuir com o aprimoramento da solução apresentada pelo seu Projeto Aplicado.