

RESUMO CASE TÉCNICO – ANALISTA DE DADOS SR – MAGALU

Isac Carvalho 19/02/2024

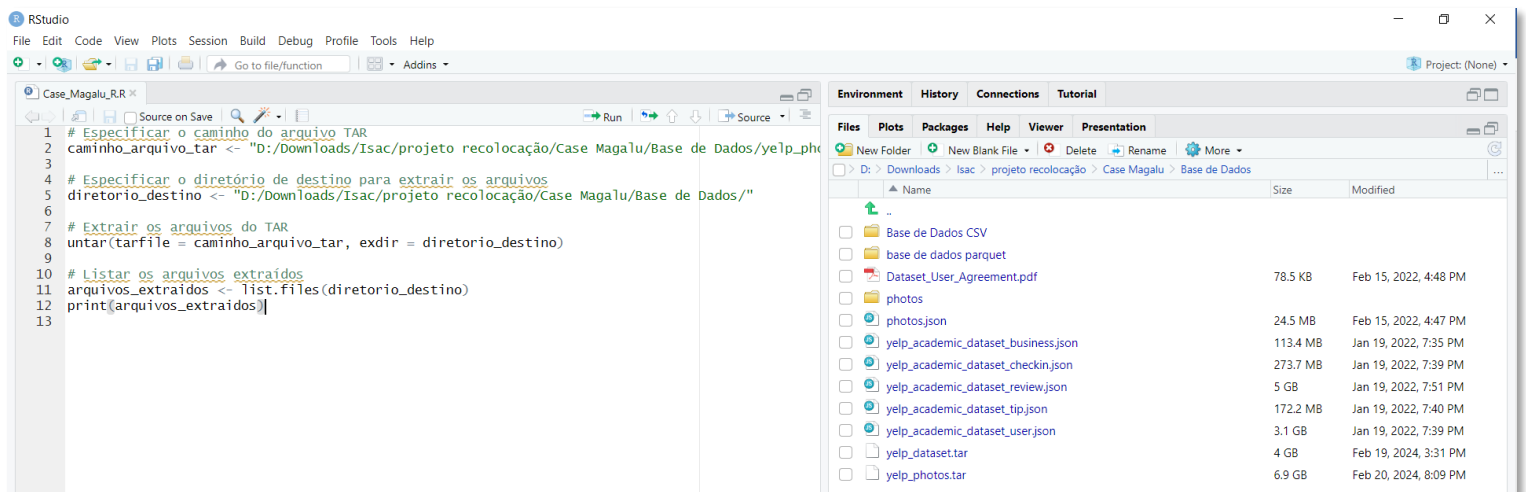
O case técnico consiste nas seguintes etapas:

1. Realizar o download do dataset.
2. Tratar os dados que vem no formato JSON usando a linguagem Python
3. Salvar em um novo dataset
4. Realizar o processo de JOINS usando SQL
5. Conectar a Query no DataStudio (Looker), fique a vontade para escolher qualquer ferramenta de DataViz.
6. Gerar um painel com uma visão analítica
7. Gerar um painel com uma visão gerencial ou estratégica
8. Montar uma apresentação para apresentar os resultados mais insights relevantes do negócio.

Todo o planejamento para realização do case foi pensado em um fator crucial: Tempo. Tenho um prazo de 3 dias para realizar todas as etapas e gerar valor com esses dados. Levando isso em consideração o case será desenvolvido de forma mais simples possível e evitando soluções complexas para que o “cliente” possa ter acesso aos dados dentro do prazo.

Etapa 1 – Extraindo os Datasets

Os datasets tem como origem o site acadêmico [yelp](#) e estavam encapsulados em um arquivo tipo TAR. Para extrair usei o R Studio.

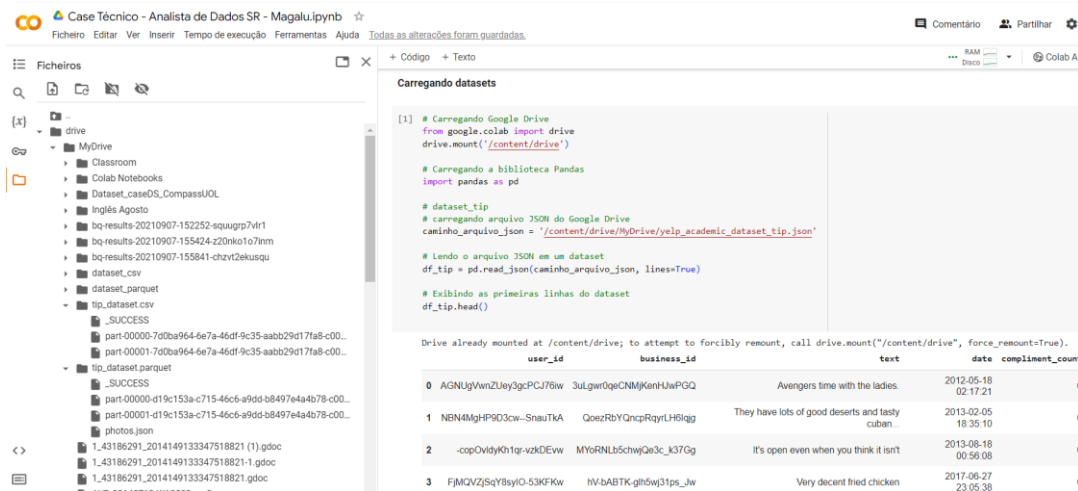


Ambiente de Trabalho R Studio

Vamos ter 6 datasets no formato JSON, sendo eles: Business, Checkin, Review, Tip, User e Photos. Os datasets foram copiados para o meu drive pessoal onde vou poder tratar esses dados no Colab.

Etapas 2 e 3 – Tratando os dados com Python no Colab

Por que no Colab? O Colab é um ambiente pronto para trabalhar nossos dados, não vou precisar gastar tanto tempo configurando o ambiente de trabalho, o que me vai fazer ganhar tempo e também me traz a possibilidade de usar o Google Drive que vai me ajudar bastante.



```
[1] # Carregando Google Drive
from google.colab import drive
drive.mount('/content/drive')

# Carregando a biblioteca Pandas
import pandas as pd

# dataset_tip
# carregando arquivo JSON do Google Drive
caminho_arquivo_json = '/content/drive/MyDrive/yelp_academic_dataset_tip.json'

# Lendo o arquivo JSON em um dataset
df_tip = pd.read_json(caminho_arquivo_json, lines=True)

# Exibindo as primeiras linhas do dataset
df_tip.head()
```

	user_id	business_id	text	date	compliment_count
0	AGNlUgVWmZUey3gcPCJ76iw	3uLgwr0qeCNMjKenHuwPGQ	Avengers time with the ladies	2012-05-18 02:17:21	0
1	NBN4MgHP9D3cw-SnauTKA	QoezRbYQncpRqytLH6lqg	They have lots of good deserts and tasty Cuban...	2013-02-05 18:35:10	0
2	-copOvdyKh1qr-vzkDEWv	MYoRNLb5chwJQe3c_k37Gg	It's open even when you think it isn't	2013-06-18 00:56:08	0
3	FJM2VZjSqY8syIO-53KFkw	IVbABTK-gh6wj31ps_Jw	Very decent fried chicken	2017-06-27 23:05:38	0

Ambiente de trabalho no Colab criado com sucesso.

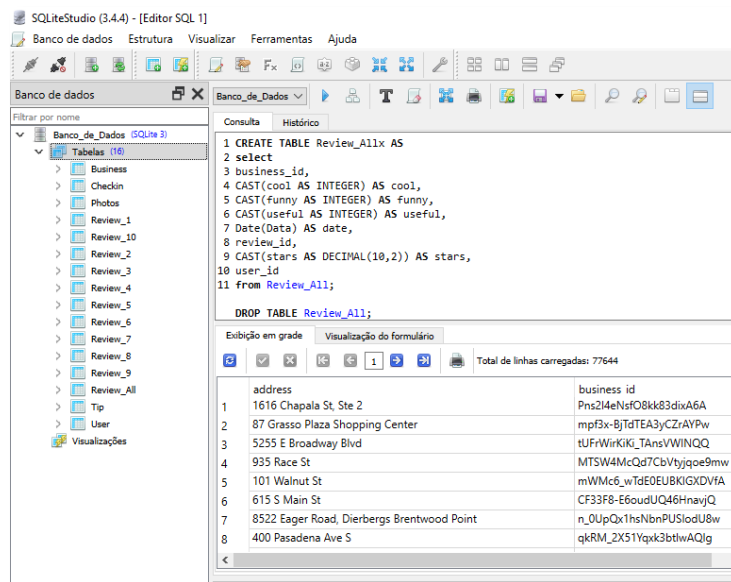
Explorei os datasets para conhecer os dados um pouco melhor e entender se precisam de algum tratamento como dados nulos por exemplo. A boa notícia é que todos os datasets estão em ordem em questão de qualidades de dados, podemos trabalhar em algumas outras coisas porém vamos deixar para etapa do SQL.

Temos duas bases de dados bem grandes e que vão dar um certo trabalho. Seriam elas: Usuários (dataset_user) e Avaliações de Usuários (dataset_review). Para lidar com essas tabelas vamos usar o pyspark. O pyspark trabalha com processamento distribuído o que é ideal para dados em grande escala.

Vamos converter os 6 datasets em um formato mais fácil de consumir, deixaremos de usar o formato JSON e vamos usar o formato CSV. Porque CSV? Vamos ganhar tempo para analisar esses dados com SQL, importar dados em csv é bem mais simples e rápido. E as base de Usuários e Avaliações? O pyspark vai parcionar estes datasets, isso facilita o armazenamento tornando viável o formato em CSV.

Etapa 4 - Realizar o processo de JOINS usando SQL

Nesta etapa vamos usar o Dbeaver como administrador de banco de dados. O Dbeaver não atendeu para o que eu precisava, vou precisar mudar a estratégia e usar o SQLite. Dois pontos importantes: o SQLite tem suporte para o Power BI (o que é ideal para o meu caso), porém vou ter que gastar mais tempo preparando o ambiente de dados para poder conseguir trabalhar com o SQL.

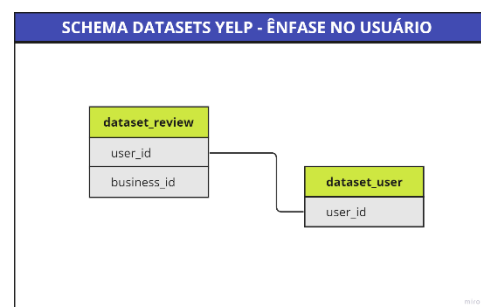
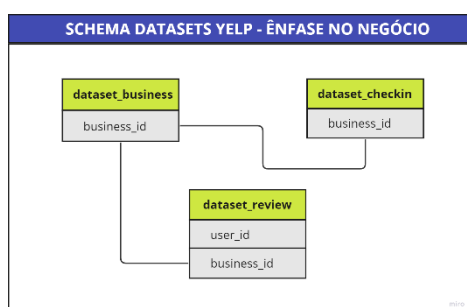


Ambiente de dados no SQLite criado com sucesso.

O objetivo é criar duas tabelas principais. A primeira tabela quero dar ênfase aos pontos comerciais(Business), olhando para os dados de Reviews trazendo os dados de Business de forma granular e resumizando os dados em relação aos usuários. A segunda tabela será o inverso, quero dar ênfase aos usuários(User), olhando para os dados de Reviews trazendo os dados de User de forma granular e resumizando os dados em relação aos pontos comerciais.

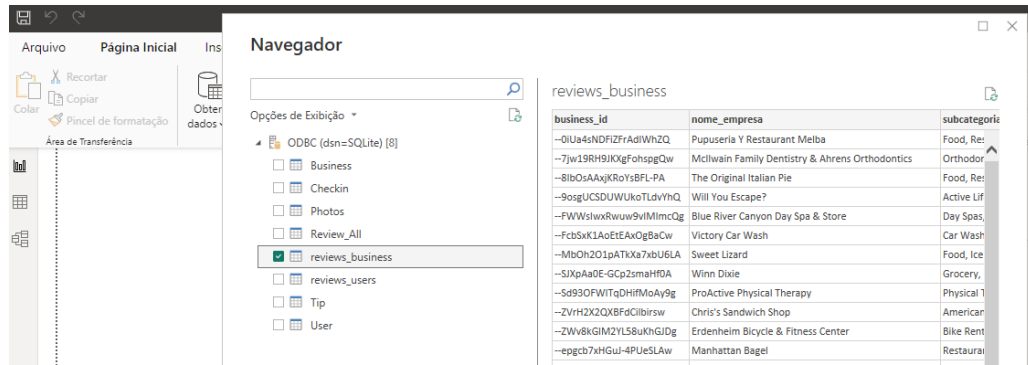
Por que essa decisão? Primeiro que o case deixou a análise em aberto, o que me faz ter a autonomia em escolher como quero trazer esses dados e trazendo dados do ponto de vista do comércio e do usuário vai nos trazer insights interessantes. O segundo motivo é evitando de trazer o dado bruto, resumizando os dados ou de comércio ou de usuário, eu tenho uma base de dados menor e mais leve, o que vai me dar um ganho em desempenho e processamento destes dados.

Lendo a documentação da base de dados da Yelp podemos perceber que temos dois campos principais que se relacionam entre as tabelas, são elas: 'user_id' e 'business_id'. Usando como base a documentação, o objetivo da análise e a estrutura que as tabelas devem ter, cheguei na conclusão aos dois Schemas abaixo:



Etapas 5, 6 e 7 – Criar relatório no Power BI.

Com as duas queries prontas e criadas em tabelas no SQLite, agora eu vou instalar e configurar o SQLite ODBC Driver para conectar diretamente no Power BI. Vai demandar mais tempo nesse processo, mas vou ganhar em mais confiabilidade dos dados e evitar o risco que o dado quebre tentando usar outros formatos.



Conector ODBC SQLite configurado com sucesso.

Painel Visão Estratégica

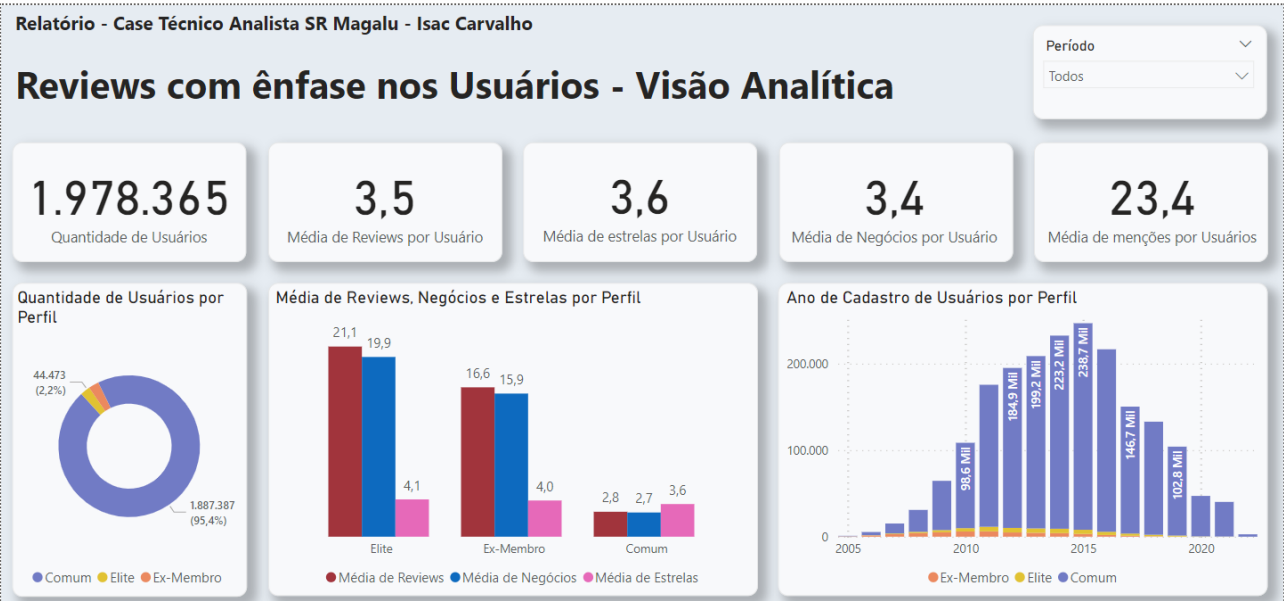


Eu construí o painel estratégico pensando em oportunidades que poderiam ser encontradas. Isto é, encontrar pontos positivos e negativos cruciais para o negócio.

Vamos alguns destaques. Starbucks e McDonalds disparados em Reviews, porém a grande maioria em reviews negativos, isso porque Starbucks tem uma média de estrelas de 2,9 e o McDonalds com uma média de estrelas de miséros 1,6. Isso prova que ter muitos Reviews seja sinal de sucesso.

Um outro ponto interessante é o Café Du Monde. Esse estabelecimento tem uma média de Check-ins muito alta porém a quantidade de usuários é super baixa. Isso quer dizer que o Café Du Monde tem uma base de usuários muito fiel, só pra termos noção, em todo esse período os frequentadores visitaram o mesmo estabelecimento mais de 20 vezes.

Painel Visão Analítica



Eu construí o painel analítico pensando em entender o comportamento do usuário. E encontramos algumas coisas interessantes.

Usuários do perfil Elite e ex-membros são muito mais engajados, positivos e mais propícios a visitar algum estabelecimento. Isto acontece porque o usuário do perfil Elite tem mais Reviews e a média de estrelas superior ao usuário Comum. Isso é essencial para o negócio porque o usuário Elite tende a comprar mais.

Um outro ponto interessante é olhar por ano de cadastro por perfis de usuário. A partir do ano de 2016 os cadastros começam a diminuir e o pior em 2020 os usuários Elite praticamente desapareceram. Isso é um ponto crítico que precisa ser trabalhado.

O restante das coisas vou deixar para uma próxima conversa. Ficarei muito em poder dividir com vocês como foi todo o processo de cada etapa, como cheguei nas soluções, como resolvi as dificuldades no caminho e etc. Muito obrigado pela oportunidade foi um desafio muito bacana de fazer.

Cronograma

Segunda-feira	Terça-feira	Quarta-feira	Quinta-feira	Sexta-feira
- Case recebido na segunda a tarde, porém não consegui trabalhar no case por motivos pessoais.	- Passei o dia fora no hospital com a minha filha. Trabalhei somente na parte da noite planejando o que iria ser feito e preparando o ambiente de trabalho.	- Desenvolvi todo o trabalho com o Python no Colab tratando os dados para levar no para o SQL. Em paralelo documentando o que eu vinha desenvolvendo.	- Desenvolvi todo o trabalho no SQL. Tive um imprevisto com o Dbeaver e tive que usar o SQLite. Em paralelo documentando o que eu vinha desenvolvendo.	- Desenvolvi todo o trabalho no Power BI criando o relatório. Tive um imprevisto com o conector odbc. Em paralelo documentando o que eu vinha desenvolvendo.