media
.monks

Processo Seletivo Interno

Data Science

Alinhamentos Gerais

- A etapa do case será dividida em dois projetos, um com foco no SQL e outro mais focado na análise de dados. Essa seleção de projetos e a estruturação deles reflete o nosso dia-a-dia de trabalho;

- É importante que faça a prova com base nos seus conhecimentos. Reforçamos isso pois buscamos entender e conhecer melhor as capacidades técnicas de cada pessoa candidata;

- É **proibido** a consulta em ferramentas como ChatGPT/Bard para a resolução da prova. Se comprovada a consulta, isso poderá acarretar na desclassificação da pessoa candidata.

Projeto em SQL

- Esta é uma prova de SQL baseada no modelo de consultas do Google Big Query;

- Utilize o schema da tabela abaixo para desenvolver os SELECTs;

- Queremos analisar se você consegue elaborar a estrutura do SELECT (não se preocupe com erros de formatação ou de sintaxe **específicos** do Big Query, mas sim, a sintaxe e estrutura de SQL);

- **Não é** necessário entrar no Google BigQuery e executar as consultas, apenas faça a query que achar correta.

Importante: Deixar indicação na prova de qual SQL foi utilizado para responder às questões. Ex: Standard SQL no BigQuery.

Considere a seguinte tabela sobre corridas de táxi em Nova Iorque:

Nome do campo	Tipo	Descrição
vendor_id	STRING	A code indicating the TPEP provider that provided the record. 1= Creative Mobile Technologies, LLC; 2= VeriFone Inc
pickup_datetime	DATETIME	The date and time when the meter was engaged.
dropoff_datetime	DATETIME	The date and time when the meter was disengaged.
passenger_count	INTEGER	The number of passengers in the vehicle. This is a driver-entered value
trip_distance	NUMERIC	The elapsed trip distance in miles reported by the taximeter.
rate_code	STRING	The final rate code in effect at the end of the trip. 1= Standard rate 2=JFK 3=Newark 4=Nassau or Westchester 5=Negotiated fare 6=Group ride
store_and_fwd_flag	STRING	This flag indicates whether the trip record was held in vehicle memory before sending to the vendor, aka "store and forward," because the vehicle did not have a connection to the server. Y= store and forward trip N= not a store and forward trip
payment_type	STRING	A numeric code signifying how the passenger paid for the trip. 1= Credit card 2= Cash 3= No charge 4= Dispute 5= Unknown 6= Voided trip
fare_amount	NUMERIC	The time-and-distance fare calculated by the meter
extra	NUMERIC	Miscellaneous extras and surcharges. Currently, this only includes the $0.50 and $1 rush hour and overnight charges.
mta_tax	NUMERIC	$0.50 MTA tax that is automatically triggered based on the metered rate in use
tip_amount	NUMERIC	Tip amount – This field is automatically populated for credit card tips. Cash tips are not included
tolls_amount	NUMERIC	Total amount of all tolls paid in trip.
imp_surcharge	NUMERIC	$0.30 improvement surcharge assessed trips at the flag drop. The improvement surcharge began being levied in 2015.
total_amount	NUMERIC	The total amount charged to passengers. Does not include cash tips
pickup_location_id	STRING	TLC Taxi Zone in which the taximeter was engaged
dropoff_location_id	STRING	TLC Taxi Zone in which the taximeter was disengaged

Utilizando a documentação de SQL do Google BigQuery e baseando no schema da tabela acima e nos dados presentes na aba "*TABELA*" da planilha fornecida, crie as consultas SQL correspondentes às seguintes questões:

Perguntas:

1) Qual foi a receita de cada tipo de pagamento no dia 15 de Março de 2018?

2) Considere que corridas de táxi válidas tenham de 1 a 5 passageiros. Qual a quantidade de corridas feitas com cada número de passageiros, faturamento médio de cada corrida e faturamento médio por passageiro?

3) Qual a hora que mais começaram corridas?

4) Considerando apenas as corridas que houveram pedágios (tolls) e que transportaram até 3 passageiros, qual a média do valor pago em pedágios por corrida?

Projeto em Análise de Dados

Suponha que você recebeu o seguinte e-mail de um cliente cuja ONG atua definindo e implementando políticas e ações sociais baseadas em dados:

Bom dia, tudo bem?

Temos uma ONG e criamos ações sociais com base em dados governamentais. Com o nosso time de BI nós conseguimos mapear algumas bases do IBGE, focadas nos casamentos homoafetivos do Brasil.

Os dados estão disponíveis aqui.

Contexto do assunto

Em maio de 2011, o plenário do Supremo Tribunal Federal (STF) equiparou as relações entre pessoas do mesmo sexo às uniões estáveis entre homens e mulheres, reconhecendo a união homoafetiva como núcleo familiar. A partir desse entendimento, o Conselho Nacional de Justiça (CNJ) publicou a Resolução 175 em 2013, que proíbe tabeliães e juízes de se recusarem a registrar a união de pessoas do mesmo sexo, determinando que todos os cartórios do país realizem casamentos homoafetivos.

A orientação sexual é uma informação pessoal, então há poucos dados sobre isso, já que os órgãos não costumam coletar essa informação. A base de dados do IBGE, porém, é uma exceção positiva nesse levantamento, pois além de ter dados oficiais do Brasil, permite uma ampla análise exploratória.

Objetivo da análise:

Queremos entender quais insights conseguimos tirar dessa base, e também qual a previsão de casamentos homoafetivos para os próximos 2 anos nos 3 estados com maior número de casamentos no último ano. Seria possível realizar essa análise?

*Por favor, envie um relatório com **gráficos** e **tabelas** que mostre as informações obtidas de maneira organizada, além das conclusões que conseguir tirar e que auxilie nas decisões que tomaremos neste ano a respeito da definição de novas ações sociais da ONG. Esse relatório deverá ser entregue em um arquivo ou em um local em nuvem, constando os códigos, com tabelas, gráficos e o passo a passo do que foi feito.*

Posso contar com você?

Abs,

Seu trabalho nesse projeto será:

1) Construir o relatório solicitado pelo cliente

Os dados necessários para a construção do relatório estão presentes no Google Drive em arquivos CSV.

As ferramentas para uso: Python ou R

Abaixo há algumas abordagens que serão consideradas um diferencial. Não tem problema se você não teve experiências prévias com essas metodologias, o importante é você se desafiar e tentar aplicar na análise. Estes materiais irão te auxiliar e não é necessário aplicar todas as técnicas abaixo:

- Análises com embasamento estatístico, utilizando teste de hipóteses ou correlação;
- Utilizar métodos de aprendizado de máquina, como regressão, séries temporais ou outro que julgar adequado;
- Descreva os pressupostos e as etapas para chegar à sua conclusão, lembre-se de citar as técnicas utilizadas e sob quais condições você poderia utilizá-las ou não.