FIAP

CLOUD & COGNITIVE ENVIRONMENTS

MBA EM DATA SCIENCE & IA

Datasource e dicionário

vendor_id	STRING	A code indicating the LPEP provider that provided the record. 1= Creative Mobile Technologies, LLC; 2= VeriFone Inc.
pickup_datetime	TIMESTAMP	The date and time when the meter was engaged
dropoff_datetime	TIMESTAMP	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 on hailed trips at the flag drop. The improvement surcharge began being levied in 2015.
airport_fee	NUMERIC	
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
data_file_year	INTEGER	Datafile timestamp year value
data_file_month	INTEGER	Datafile timestamp month value

tabela	Registros
bigquery-public-data.new_york_taxi_trips.tlc_yellow_trips_2011	176.887.248
	...
bigquery-public-data.new_york_taxi_trips.tlc_yellow_trips_2021	30.904.297

Parte 01 Análises Big Data na nuvem

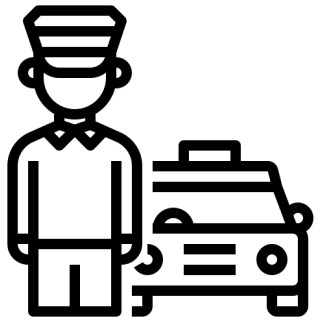
FIAP MBA+

Total: 4,0

- Utilizando apenas ferramentas na nuvem, faça análise de dados "GIGANTES" em seu ambiente utilizando um Notebook Python. Não serão aceitas entregas que não sejam via Notebook.
- Avalie as corridas dos táxis amarelos de NY e responda:
  - Em 2011, como foram as quantidades de corridas de acordo com o dia da semana? (0,5)
  - Em 2011, como foram os valores das corridas de acordo com o dia da semana? (0,5)
  - Em 2011, como foram os valores das corridas de acordo com a quantidade de passageiros? (0,25)
  - Em 2011, como foram as gorjetas de acordo com o dia da semana? (0,25)

- Ainda neste conjunto de dados, avalie a evolução ano a ano até 2021
  - Total absoluto de corridas (0,5)
  - Total relativo de corridas por dia da semana (0,5)
  - Total relativo de corridas por período do dia (1,5)
    - (pesquisar sobre extração de hora)

- Todas as análises devem ser em formato tabular e gráfico.



Sim	42%
Não	58%

Parte 02 - Serviços de ML na nuvem para servir modelos

FIAP MBA+

Total: 4,0

- Criar um modelo de classificação em um notebook Python que, de acordo com os dados da corrida, consiga prever se o motorista irá ganhar alguma gorjeta (1,0);
  - Será avaliado o uso de plataforma em nuvem no apoio do processo, logo, o modelo não será avaliado em sua qualidade.
- Servir esse modelo na Azure Cloud, como visto em aula (3,0);
  - Será avaliado o uso de plataforma em nuvem no apoio ao processo de deploy de modelo. Como existem inúmeras formas de deploy, será avaliado apenas o deploy exatamente como feito em aula;
  - Colocar em um doc Word ou PPT os passos utilizados para o deploy;

Parte 03 Integração com modelos LLM

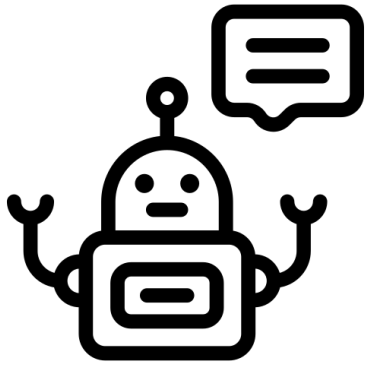
FIAP MBA+

Total: 2,0

- Criar no notebook uma integração com uma API de LLM (em aula vimos o PaLM da GCP) que consiga fazer interpretações analíticas, apoiando o cientista de dados.
- Para isso, será necessário montar um prompt dinamicamente com os dados buscados no Big Query, e então, integrar com esse serviço.
  - Obtenção de dados e montagem de prompt (1,5)
  - Chamada de API e exibição do retorno (0,5)

Exemplo fictício de prompt:

Faça uma análise interpretativa sobre as corridas de táxi de NY considerando:
- Em 1998 foram 44565464 de viagens e em 1999 foram 64565464 de viagens;
- O tempo médio de corridas em 1998 era de 48 minutos e em 1999 era de 69 minutos;
- As viagens em 1998 se concentravam nos dias da semana, enquanto que em 1999 a distribuição era mais uniforme.




FIAP