

# Programming for Big Data exam

## Brazilian E-Commerce Public Dataset by Olist

Source: Kaggle ([Click here](#))

This is a Brazilian ecommerce public dataset of customer purchases made through the [Olist Store](#). The dataset contains data on 100,000 purchases (orders) from 2016 to 2018 made over multiple marketplaces in Brazil. Its features allow viewing an order from multiple dimensions: from order status, price, payment and freight performance to customer location, product attributes and finally reviews written by customers. There is also a geolocation dataset that relates Brazilian zip codes to lat/lng coordinates.

Brazilian ecommerce is a real commercial dataset.

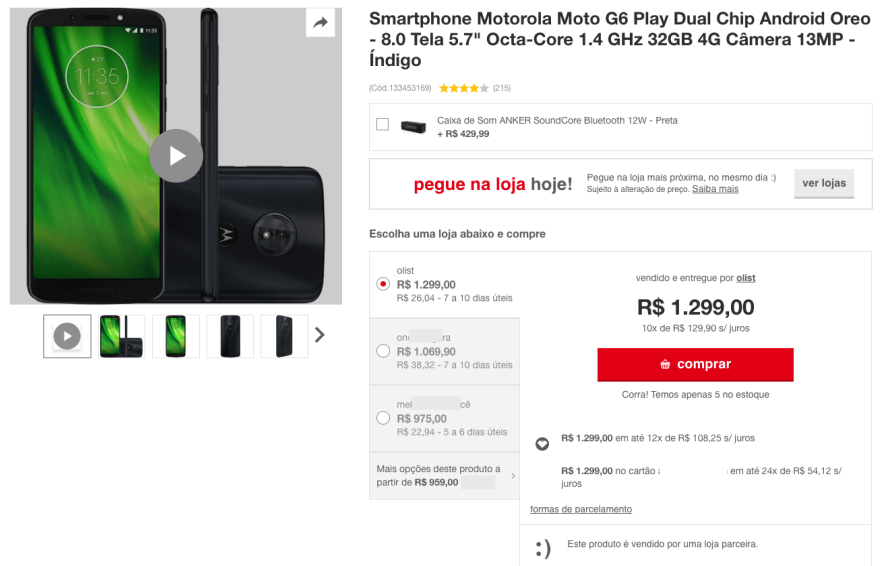
### Context

This dataset was provided by Olist, a department store in Brazilian marketplaces. Olist connects small businesses from all over Brazil to channels. Those merchants are able to sell their products through the Olist Store and ship them directly to the customers using Olist logistics partners.

After a customer purchases a product from Olist Store a seller gets notified to fulfill that order. Once the customer receives the product, or the estimated delivery date is due, the customer gets a satisfaction survey by email where he can give a note for the purchase experience and write down some comments.

### Attention

1. an order might have multiple items
2. each item might be fulfilled by a distinct seller



**Smartphone Motorola Moto G6 Play Dual Chip Android Oreo - 8.0 Tela 5.7" Octa-Core 1.4 GHz 32GB 4G Câmera 13MP - Índigo**

(Cod.133453169) ★★★★★ (215)

☐ Caixa de Som ANKER SoundCore Bluetooth 12W - Preta + R\$ 429,99

**pegue na loja hoje!** Pegue na loja mais próxima, no mesmo dia :) Sujeito à alteração de preço. [Saiba mais](#) [ver lojas](#)

Escolha uma loja abaixo e compre

olista	vendido e entregue por olista
<input checked="" type="radio"/> R\$ 1.299,00 R\$ 26,04 - 7 a 10 dias úteis	<b>R\$ 1.299,00</b> 10x de R\$ 129,90 s/ juros
<input type="radio"/> on/...ra R\$ 1.069,90 R\$ 38,32 - 7 a 10 dias úteis	<b>comprar</b>
<input type="radio"/> mel...cê R\$ 975,00 R\$ 22,94 - 5 a 6 dias úteis	Corral! Temos apenas 5 no estoque

Mais opções deste produto a partir de R\$ 959,00 >

☒ R\$ 1.299,00 em até 12x de R\$ 108,25 s/ juros

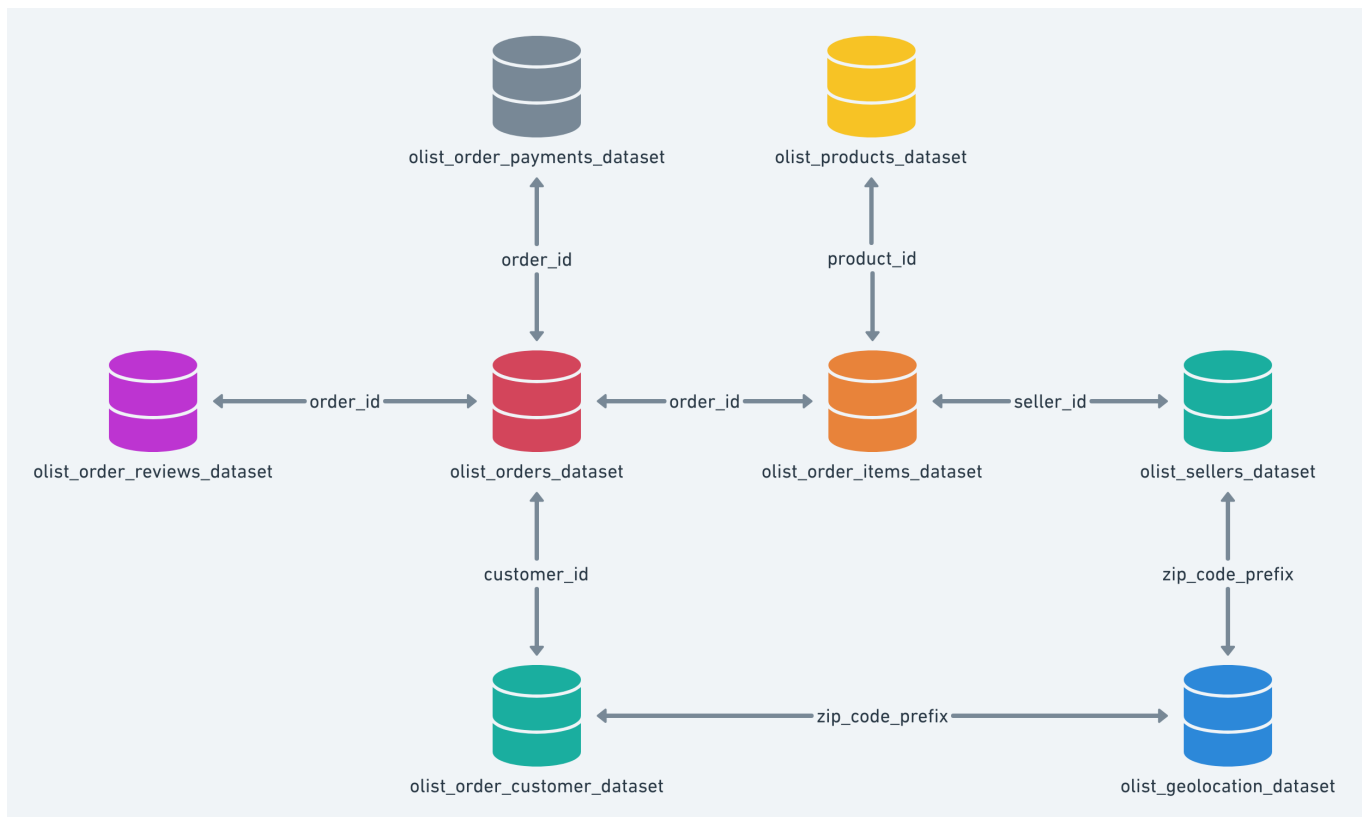
☐ R\$ 1.299,00 no cartão : em até 24x de R\$ 54,12 s/ juros

formas de parcelamento

:) Este produto é vendido por uma loja parceira.

## Data Schema (🧬)

The data is divided into multiple datasets for better understanding and organization. Please refer to the following data schema when working with it:



# Exam

You will have to analyze the Brazilian ecommerce public dataset, and to answer business questions using SQL.

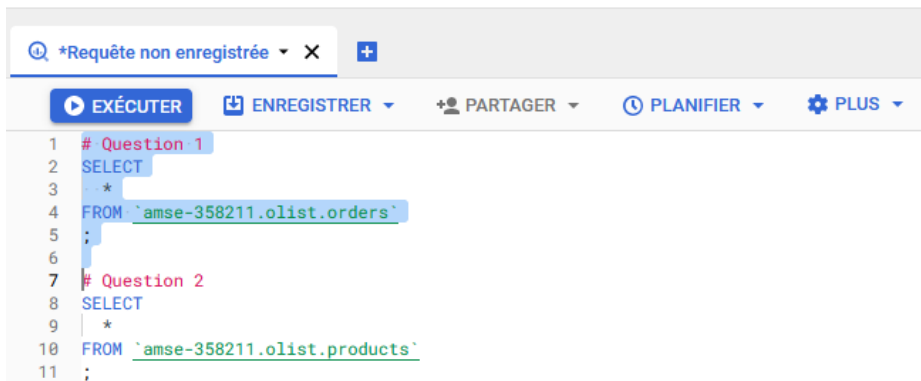
You can form groups of up to 2 people.

Deliverables:

- A zipped folder that contains :
  - An explanation of your answers in part B (in a .pdf file)  
Also, outline what each member of the group was responsible for, the specific tasks they completed, and any challenges you encountered.
  - A single SQL file (.txt or .sql) that contains all SQL queries (with mention of the question you are answering to)
- Oral presentation (15 minutes) :
  - Explanation of your answers in part B
  - Explanation of some SQL queries
  - Presentation of your dashboard on LookerStudio

Tip: In BigQuery, while elaborating your queries, you can write your different queries in the same single editor. To this end, separate your queries with semicolons (;), as shown in the screenshot below. To execute queries separately, select with the mouse the block of code to execute, and click on the "Execute" button.

*Warning: if you do not select a block, all queries will be executed.*



```
1 # Question 1
2 SELECT
3   *
4 FROM `amse-358211.olist.orders`
5 ;
6
7 # Question 2
8 SELECT
9   *
10 FROM `amse-358211.olist.products`
11 ;
```

The rating is divided into different parts.

	points	oral exam explanation
upload data in BigQuery	0	-
data understanding	2	-
SQL queries	12	Your SQL queries results should fit the results in the screenshots.
SQL indentation	-1 to 1	a not or badly indented SQL query will be penalized, well indented queries rewarded
Looker Studio	3	<p><u>Dashboard content</u></p> <p>Your dashboard should contain at least two pages.</p> <p>It should include visualizations, such as charts, graphs, and tables, that convey meaningful insights from the e-commerce data.</p> <p><i>Examples of components you can include:</i></p> <ul style="list-style-type: none"> <li>• Sales trends over time (daily, monthly, seasonality, etc).</li> <li>• Customer segmentation analysis.</li> <li>• Product performance metrics. (best-sellers, worst-sellers, KPIs, etc.)</li> <li>• Geographic sales distribution.</li> <li>• Customer feedback analysis.</li> </ul> <p><u>Dashboard interactivity</u></p> <p>Ensure your dashboard is interactive, allowing users to explore the data, <i>apply filters</i> to get deeper insights. Use interactive elements such as drop-down menus, sliders, or clickable charts.</p> <p><u>Presentation</u></p> <p>During the presentation, emphasize the most important findings and insights.</p>
oral presentation	3	<p>You must distribute your presentation among the group members and explain how you shared the work.</p> <p>During the oral presentation, I will pay close attention to the roles each student takes, and take notes on who is speaking, and how you answer questions.</p> <p>Present part B (data understanding), some SQL queries you found challenging, and the LookerStudio dashboard.</p>
<b>total</b>	<b>20</b>	

## A.Upload data in BigQuery

Create a Kaggle account ([here](#)) and download the zipped dataset ([here](#)).

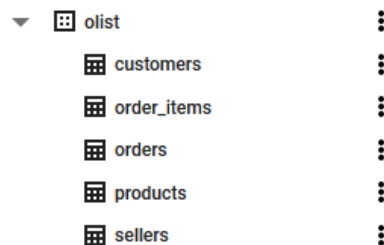
Follow the indications in **'step1.png'** & **'step2.png'** for the following.

Create a **dataset** in Google BigQuery named 'olist'.

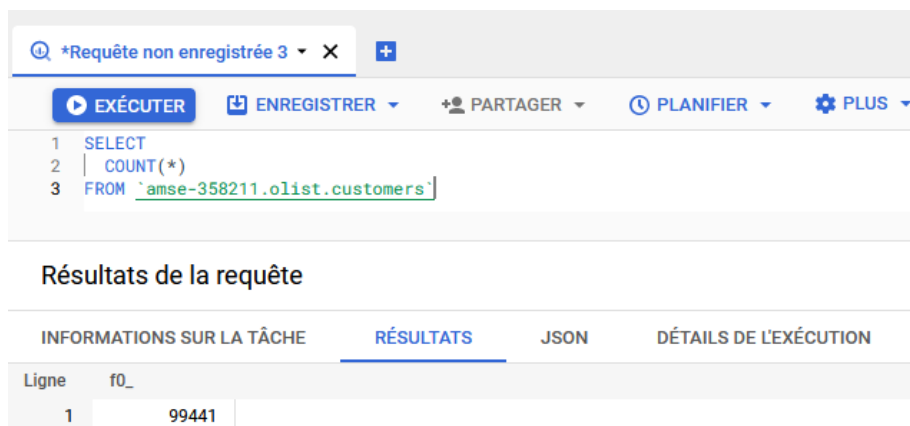
Use the **.csv** files in the zipped folder to create the following **tables** in Google BigQuery

- ☐ olist\_customers\_dataset.csv → customers
- ☐ olist\_orders\_dataset.csv → orders
- ☐ olist\_order\_items\_dataset.csv → order\_items
- ☐ olist\_products\_dataset.csv → products
- ☐ olist\_sellers\_dataset.csv → sellers

you should end up with  
something similar to  
this screenshot



Check that the tables have been successfully imported using the SQL command `SELECT COUNT(*) FROM `your_schema.olist.customers`` in the BigQuery Editor:




The screenshot shows the Google BigQuery Editor interface. At the top, there's a tab labeled '\*Requête non enregistrée 3'. Below the tab, there are buttons: 'EXÉCUTER', 'ENREGISTRER', 'PARTAGER', 'PLANIFIER', and 'PLUS'. The SQL query is displayed in the editor:

```
1 SELECT
2 | COUNT(*)
3 FROM `amse-358211.olist.customers`
```

Below the query, there's a section titled 'Résultats de la requête'. Under this section, there are tabs: 'INFORMATIONS SUR LA TÂCHE', 'RÉSULTATS', 'JSON', and 'DÉTAILS DE L'EXÉCUTION'. The 'RÉSULTATS' tab is selected. The results are displayed in a table with two columns: 'Ligne' and 'f0\_'. The first row shows the value '99441'.

Ligne	f0_
1	99441

## B.Data preview and understanding

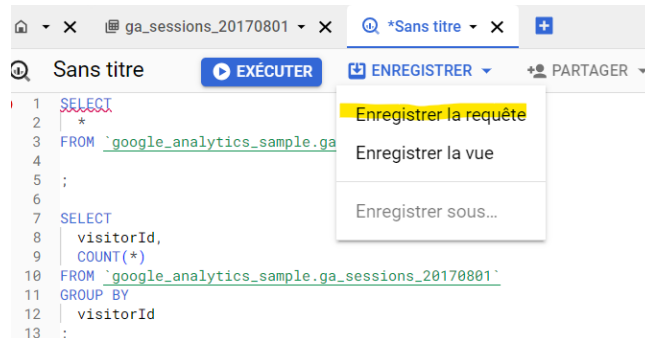
Using the data schema (, see page 2) and by observing the overview in BigQuery ([APERÇU](#)) of the different tables:

- ☐ indicate which columns are primary keys and foreign keys
- ☐ in the "customers" table, it is possible to go further in the data normalization. Explain why and how.
- ☐ explain whether or not the other tables are normalized (first, second and third normal forms)

## C. Answer business questions using SQL

Answer those business questions using SQL queries.

*Don't forget to frequently **save** your queries.*



### CUSTOMERS

- ☐ Q1 How many unique customers does the Olist store have in these datasets?

Desired output

Ligne	nb_unique_c...
1	96096

- ☐ Q2 Count the number of unique customers living in a postal code (customer\_zip\_code\_prefix) beginning with '13' (Use a [string function](#))

Desired output

Ligne	f0_
1	6464

- ☐ Q3 Select regions with more than 10,000 clients. Order the result by descending number of customers.

Desired output

Ligne	customer_state	nb_custome...
1	SP	40302
2	RJ	12384
3	MG	11259

### PRODUCTS

- ☐ Q4 How many unique products does Olist Store have ?

Desired output

Ligne	f0_
1	32951

- ☐ Q5 Count the number of NULL values in each column of the `products` table

Desired output

Ligne	product_id	product_cat...	product_na...	product_des...	product_ph...	product_we...	product_len...	product_hei...	product_wid...
1	0	610	610	610	610	2	2	2	2

- ☐ Q6 Select the 10 (*only 10*) most present product categories in the product table

Desired output:

Ligne	product_category_name	f0_
1	cama_mesa_banho	3029
2	esporte_lazer	2867
3	moveis_decoracao	2657
4	beleza_saude	2444
5	utilidades_domesticas	2335
6	automotivo	1900
7	informatica_acessorios	1639
8	brinquedos	1411
9	relogios_presentes	1329
10	telefonica	1134

- ☐ Q7 Select the 10 product categories with the highest average weight.

Desired output:

Ligne	product_category_name	product_wei...
1	moveis_colchao_e_estofado	13190.0
2	moveis_escritorio	12740.8673...
3	moveis_cozinha_area_de_servi...	11598.5638...
4	moveis_quarto	9997.22222...
5	eletrodomesticos_2	9913.33333...
6	moveis_sala	8934.84615...
7	pcs	7995.33333...
8	industria_comercio_e_negocios	5929.19117...
9	agro_industria_e_comercio	5263.40540...
10	climatizacao	4459.95967...

- ☐ Q8 Select the product categories with an average description length between 400 and 450 or between 1000 and 1050.

Desired output:

Ligne	product_category_name	product_des...
1	bebidas	1047.72839...
2	portateis_casa_forno_e_cafe	1046.38709...
3	telefonica_fixa	1017.38793...
4	fashion_underwear_e_moda_pr...	1004.43396...
5	artigos_de_festas	446.230769...
6	papelaria	439.128386...
7	la_cuisine	422.299999...
8	artigos_de_natal	412.2
9	eletrodomesticos	404.597297...



## ORDERS

- ☐ Q9 What has been Olist's revenue over years?

Ligne	f0_	f1_
1	2016	49785.9200...
2	2017	6155806.98...
3	2018	7386050.80...

- ☐ Q10 How many orders occurred on a Monday? (Use [date functions](#))

Ligne	f0_
1	16196

- ☐ Q11 How many orders occurred in December or June? (Use [date functions](#))

Ligne	f0_
1	15086

- ☐ Q12 How many orders occurred on a Sunday between [9am, 23pm[ in 2018?

Ligne	f0_
1	5656

- ☐ Q13 Count the number of orders for each name of day (Monday, Tuesday, ..., Sunday). Must be in a single query. Order by day of week.

Ligne	day	dayname	nb_orders
1	1	Sunday	11960
2	2	Monday	16196
3	3	Tuesday	15963
4	4	Wednesday	15552
5	5	Thursday	14761
6	6	Friday	14122
7	7	Saturday	10887

- ☐ Q14 Divide order\_purchase\_timestamp into four categories

- Morning: [06 am to 12 am[
- Lunch: [12 to 14 pm[
- Afternoon: [14 to 18 pm[
- Evening: [18 pm to 23 pm[
- Night: [23 pm to 06 am[

and count the number of orders in each of those categories.

Ligne	daypart	f0_
1	Morning	22240
2	Night	8863
3	Afternoon	25848
4	Lunch	12513
5	Evening	29977

- ☐ Q15 Calculate the average delivery time in terms of days

Desired output

Ligne	time_to_deli...
1	12.4973361...

- ☐ Q16 Calculate the number of orders and the average delivery time by destination (customer's region)

Desired output:

Ligne	customer_state	nb_orders	time_to_del...
1	SP	41746	8.70053092...
2	PR	5045	11.9380459...
3	MG	11635	11.9465433...
4	DF	2140	12.8990384...
5	SC	3637	14.9075274...
6	RJ	12852	15.2376750...
7	RS	5466	15.2485029...
8	GO	2020	15.5360245...
9	MS	715	15.5449358...
10	ES	2033	15.7238095...

- ☐ Q17 Select the top 10 customers of 2018 in terms of revenue

Desired output:

Ligne	customer_id	nb_orders	total_revenue
1	ec5b2ba62e574342386871631...	1	7160.0
2	f48d464a0baaea338cb25f816...	1	6729.0
3	e0a2412720e9ea4f26c1ac985...	1	4599.9
4	3d979689f636322c62418b634...	1	4590.0
5	cc803a2c412833101651d3f90...	1	4400.0
6	1afc82cd60e303ef09b4ef9837...	1	4399.87
7	35a413c7ca3c69756cb75867d...	1	4099.99
8	e9b0d0eb3015ef1c9ce6cf5b9...	1	4059.0
9	31e83c01fce824d0ff786fcd48...	1	3930.0
10	eb7a157e8da9c488cd4ddc487...	1	3899.0

- ☐ Q18 Select the most purchased product categories in the PB and RO regions in 2018 that had at least 5+ purchases

Desired output:

Ligne	customer_state	product_category_name	nb_orders
1	PB	beleza_saude	42
2	PB	relogios_presentes	31
3	PB	informatica_acessorios	27
4	PB	esporte_lazer	16
5	PB	telefonica	16
6	PB	automotivo	12
7	PB	moveis_decoracao	12
8	PB	cama_mesa_banho	11
9	PB	papelaria	10
10	PB	utilidades_domesticas	10
11	PB	eletronicos	9
12	PB	cool_stuff	7
13	PB	brinquedos	7
14	PB	bebes	6
15	RO	beleza_saude	12
16	RO	informatica_acessorios	9
17	RO	telefonica	8
18	RO	cool_stuff	8
19	RO	esporte_lazer	6
20	RO	relogios_presentes	6
21	RO	bebes	6

### MORE COMPLEX QUESTIONS

- ☐ Q19 Select customers who have purchased multiple times the same product at different times. (Use a [CTE](#))

Desired output:

Ligne	f0_
1	332

- ☐ Q20 Identify the top 20% of products that contribute to 80% of the total revenue over 2017 and derive their respective sales trends in 2018 (growth from 2017 to 2018)
- ☐ Q21 Find a way to identify seasonality over product categories.

## D. Looker Studio

SQL queries are used to manipulate data (define, create, update, delete, query, transform) in databases, but also to answer business questions.

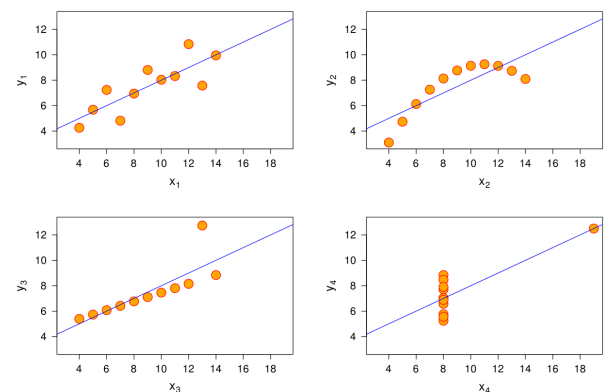
Queries return relations (tables) as results (as shown in the previous screenshots for instance).

But graphs are often better than tables to summarize, understand and present data. Data visualization is a must-have skill.

The Anscombe's quartet is a set of four data sets that have nearly identical simple descriptive statistics but appear very different when graphed. It illustrates the importance of data visualization in the understanding and interpretation of data.

Each of the four data sets in Anscombe's quartet consists of 11 data points. The data sets have the same mean, variance, correlation, and linear regression line, making them indistinguishable when examining only basic statistical properties. However, when graphed, they reveal striking differences. The key takeaway from Anscombe's quartet is that summary statistics alone may not provide a complete understanding of data, and visualization is essential for uncovering patterns, relationships, and anomalies.

I		II		III		IV	
x	y	x	y	x	y	x	y
10,0	8,04	10,0	9,14	10,0	7,46	8,0	6,58
8,0	6,95	8,0	8,14	8,0	6,77	8,0	5,76
13,0	7,58	13,0	8,74	13,0	12,74	8,0	7,71
9,0	8,81	9,0	8,77	9,0	7,11	8,0	8,84
11,0	8,33	11,0	9,26	11,0	7,81	8,0	8,47
14,0	9,96	14,0	8,10	14,0	8,84	8,0	7,04
6,0	7,24	6,0	6,13	6,0	6,08	8,0	5,25



To present and understand data, graphs are better than tables

Google Looker Studio is a free data visualization tool, part of the Google Cloud Platform that allows the creation of automated, customized and interactive dashboards.

Most of the features of Google Looker Studio are easy to use.

In this part, you will create a denormalized table in BigQuery, and explore it using Looker Studio. You will create a dashboard of at least two pages, to present the insights you found in the dataset.

Follow those instructions.

1. Execute the following query ([lookerstudio\\_step1.png](#))

```
SELECT
  *
FROM `amse-358211.olist.order_items` oi
  LEFT JOIN `amse-358211.olist.products` p ON (oi.product_id = p.product_id)
  LEFT JOIN `amse-358211.olist.sellers` s ON (oi.seller_id = s.seller_id)
  LEFT JOIN `amse-358211.olist.orders` o ON (oi.order_id = o.order_id)
  LEFT JOIN `amse-358211.olist.customers` c ON (o.customer_id =
c.customer_id)
```

2. Create the resulting denormalized relation as a BigQuery table ([lookerstudio\\_step2.png](#))
3. Open Looker Studio using the denormalized table ([lookerstudio\\_step3.png](#))
4. Free analysis. Create a dashboard of at least two pages

The end.