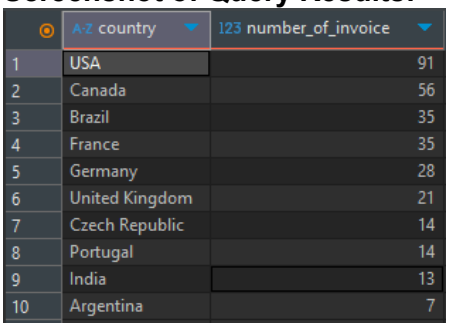
SQL & Pandas – Mentoring Week 2
Shafa Aldiena Meraldin
shafaaldienameraldin@gmail.com

Basic Information about transactional data of Digital Media Store

1. (10 point) Determine which countries have the most number of invoices (top 10). Order them by the number of invoices in descending order and if there are the same number of invoices, sort them by country name in ascending order. Show Country Name and total number of invoices.
   Answer here
   - **SQL Query Syntax:**

```
select "BillingCountry" as Country,
                count("BillingCountry") as number_of_invoice
from    "Invoice"
group by "BillingCountry"
order by number_of_invoice desc, country asc
limit 10;
```

   - **Screenshot of Query Results:**



	country	number_of_invoice
1	USA	91
2	Canada	56
3	Brazil	35
4	France	35
5	Germany	28
6	United Kingdom	21
7	Czech Republic	14
8	Portugal	14
9	India	13
10	Argentina	7

   - **Description of Query Results:**
     In this query, We collected top 10 countries with the most number of invoices consist of :
     1. USA = 91 invoices
     2. Canada = 56 invoices
     3. Brazil = 35 invoices
     4. France = 35 invoices
     5. Germany = 28 invoices
     6. UK = 21 invoices
     7. Czech Republic = 14 invoices
     8. Portugal = 14 invoices
     9. India = 13 invoices
     10. Argentina = 7 invoices.

     USA dominated the sales and these markets are the strongest and promising for exclusive promotions and loyalty programs.
     There are some countries with equal demand such as Brazil and France also Czech Republic and Portugal.
     Argentina has the lowest sales requires further market research to identify barriers.

2. (10 point) The top 10 genres by total sales in the database. The total sales are obtained by multiplying the quantity of items sold by their respective prices. Shows Genre Name and Total Sales

Answer here

- **SQL Query Syntax**

```
select "Genre"."Name" as genre_name,
            sum("Quantity" * "InvoiceLine"."UnitPrice") as total_sales
from "Track"
join "Genre"
using ("GenreId")
join "InvoiceLine"
using ("TrackId")
group by genre_name
order by total_sales desc, genre_name asc
limit 10;
```

- **Screenshot of Query Results:**

	A-Z genre_name	123 total_sales
1	Rock	826.65
2	Latin	382.14
3	Metal	261.36
4	Alternative & Punk	241.56
5	TV Shows	93.53
6	Jazz	79.2
7	Blues	60.39
8	Drama	57.71
9	Classical	40.59
10	R&B/Soul	40.59

- **Description of Query Results:**

  In this query, we collected top 10 genre with the most total sales which consist of :

  1. Rock with $826.65
  2. Latin $382.14
  3. Metal $261.36
  4. Alternative & punk $241.56
  5. TV Shows $93.53
  6. Jazz $79.2
  7. Blues $60.39
  8. Drama $57.71
  9. Classical $40.59
  10. R&B/Soul $40.59

  Rock dominated sales, while Classical and R&B/Soul, at the bottom of the top 10, had equal total sales.

3. (10 point) Who are the top 10 customers by their total spending? Shows Customer Name (consist of first name and last name), Email, and Total Spending

Answer here

- **SQL Query Syntax**

```
select CONCAT("FirstName", ' ', "LastName") as customer_name,
            "Email",
            sum("Total") as total_spending
from "Customer"
join "Invoice"
using ("CustomerId")
group by customer_name, "Email"
order by total_spending desc
limit 10;
```

- **Screenshot of Query Results:**

	customer_name	Email	total_spending
1	Helena Hol|	hholy@gmail.com	49.62
2	Richard Cunningham	ricunningham@hotmail.com	47.62
3	Luis Rojas	luisrojas@yahoo.cl	46.62
4	Ladislav Kov cs	ladislav_kovacs@apple.hu	45.62
5	Hugh O'Reilly	hughoreilly@apple.ie	45.62
6	Fynn Zimmermann	fzimmermann@yahoo.de	43.62
7	Julia Barnett	jubarnett@gmail.com	43.62
8	Frank Ralston	fralston@gmail.com	43.62
9	Victor Stevens	vstevens@yahoo.com	42.62
10	Astrid Gruber	astrid.gruber@apple.at	42.62

- **Description of Query Results:**

  In this query, we collected the top 10 customers by total spending, which consist of:

  1. Helena Holá - $49.62
  2. Richard Cunningham - $47.62
  3. Luis Rojas - $46.62
  4. Ladislav Kovács - $45.62
  5. Hugh O'Reilly - $45.62
  6. Fynn Zimmermann - $43.62
  7. Julia Barnett - $43.62
  8. Frank Ralston - $43.62
  9. Victor Stevens - $42.62
  10. Astrid Gruber - $42.62

  Helena Holá is the highest-spending customer, suggesting she might be a valuable repeat buyer worth targeting for loyalty programs or exclusive offers.
  Spending differences among the top 10 are relatively small, meaning multiple customers have similar purchasing behaviors, which could indicate a price-sensitive customer base.
  Customers like Ladislav Kovács and Hugh O'Reilly have the same spending amount, potentially highlighting a purchasing trend based on product pricing.
  Retention strategies such as personalized recommendations or exclusive discounts may help convert these high spenders into long-term loyal customers.

4. (10 point) In the results list of countries in number 1, which city has the most number of invoices? Show Country Name, City Name and total number of invoices.

Answer here

- **SQL Query Syntax :**

```
select "BillingCountry" as Country,
            "BillingCity" as City,
            count(*) as total_number_of_invoice
from"Invoice"
where "BillingCountry" in (
            select "BillingCountry"
            from    "Invoice"
            group by "BillingCountry"
            order by  count(*) desc
            limit 10
)
group by "BillingCountry" , "BillingCity"
order by total_number_of_invoice desc;
```

- **Screenshot of Query Results:**

	country	city	total_number_of_invoice
1	United Kingdom	London	14
2	Czech Republic	Prague	14
3	Brazil	S�o Paulo	14
4	France	Paris	14
5	Germany	Berlin	14
6	USA	Mountain View	14
7	Canada	Winnipeg	7
8	USA	Cupertino	7
9	France	Bordeaux	7
10	USA	Chicago	7
11	Brazil	Rio de Janeiro	7
12	France	Dijon	7
13	USA	Fort Worth	7
14	Canada	Halifax	7
15	Portugal	Porto	7
16	USA	New York	7
17	Canada	Montr�al	7
18	Belgium	Brussels	7
19	Canada	Edmonton	7
20	France	Lyon	7
21	USA	Boston	7
22	USA	Orlando	7
23	Germany	Frankfurt	7
24	USA	Madison	7
25	USA	Reno	7
26	Canada	Yellowknife	7
27	Canada	Ottawa	7
28	India	Delhi	7
29	Brazil	S�o Jos� dos C	7
30	USA	Tucson	7
31	Brazil	Bras�lia	7
32	Germany	Stuttgart	7
33	United Kingdom	Edinburgh	7
34	Canada	Toronto	7
35	Canada	Vancouver	7
36	USA	Redmond	7
37	USA	Salt Lake City	7
38	Portugal	Lisbon	7
39	India	Bangalore	6

- **Description of Query Results:**
  In this query, we collect list of cities with the highest number of invoices from top 10 countries with the highest number of invoices.
  London, Prague, São Paulo, Paris, Berlin, Mountain View = Each city recorded 14 invoices, making them the top contributors in this dataset.
  Winnipeg, Cupertino, Bordeaux, Chicago, Rio de Janeiro, Dijon, Fort Worth, Halifax, Porto, New York, Montréal, Brussels, Edmonton, Lyon, Boston, Orlando, Frankfurt, Madison, Reno, Yellowknife, Ottawa, Delhi, São José dos Campos, Tucson, Brasília, Stuttgart, Edinburgh, Toronto, Vancouver, Redmond, Salt Lake City, Lisbon = These cities all recorded 7 invoices each.
  Bangalore = The lowest in the list, contributing 6 invoices.

Next, we can find deeper information to help Product Team

5. (10 point) The product team is looking to add some tracks from new artists to the store and market them in the United Kingdom. Due to budget constraints for marketing, the product team needs to select 4 out of 6 songs to include in the store. The product team assumes that

they should choose songs with genres that are popular in the United Kingdom. Here are the tracks and their respective genres that **will be added** to the store:

- "Lalaland": R&B/Soul
- "Soul Sister": Pop
- "Good to See You": Rock
- "Nothing On You": Jazz
- "Get Ya Before Sunrise": Reggae
- "Before The Coffee Gets Cold": Hip Hop/Rap

Assist the product team in selecting the songs to be included in the store.
(Hint: Find the genres that are popular in the United Kingdom. Popularity is determined by the number of purchases of tracks (quantity) in that genre.)
Answer here

- **SQL Query Syntax**

```
select g."Name" as genre_name,
            sum(il."Quantity") as sold_tracks
from "Invoice" as i
join "InvoiceLine" as il
on i."InvoiceId" = il."InvoiceId"
join "Track" as t
on il."TrackId" = t."TrackId"
join "Genre" as g
on t."GenreId" = g."GenreId"
where i."BillingCountry" = 'United Kingdom'
group by genre_name
order by sold_tracks desc, genre_name asc
```

- **Screenshot of Query Results:**

	genre_name	sold_tracks
1	Rock	37
2	Latin	31
3	Metal	20
4	Alternative & Punk	9
5	Reggae	5
6	Jazz	4
7	Hip Hop/Rap	3
8	Pop	2
9	R&B/Soul	2
10	World	1

- **Description of Query Results:**

Based on the query results of popular genres rank in the UK, we will select 4 out of the 6 songs below to be added to the store and its database:

1. Since Rock is the top genre, we will choose "Good to See You."
2. Since Reggae ranks 5th, we will choose "Get Ya Before Sunrise."
3. Since Jazz ranks 6th, we will choose "Nothin' On You."
4. Since Hip Hop/Rap ranks 7th, we will choose "Before The Coffee Gets Cold."

Jika ditambahkan ke data base bisa menggunakan query berikut untuk membentuk dummy table :

```
SELECT *
FROM (VALUES
    ('Good to See You', 'Rock'),
    ('Get Ya Before Sunrise', 'Reggae'),
    ('Nothin On You', 'Jazz'),
    ('Before The Coffee Gets Cold', 'Hip Hop/Rap')
) AS NewSongs (Title, Genre);
```

Dengan result :

	title	genre
1	Good to See You	Rock
2	Get Ya Before Sunrise	Reggae
3	Nothin On You	Jazz
4	Before The Coffee Gets Cold	Hip Hop/Rap

6. (10 point) The Product Team wants to market albums that are popular in the USA to be marketed in other countries. Help the product team by searching for the 10 most popular albums in the USA based on album units sold

Answer here

- **SQL Query Syntax**

```
select a."Title" as album_title,
            sum(il."Quantity") as album_units_sold
from "Track" as t
join "Album" as a
using ("AlbumId")
join "InvoiceLine" as il
using ("TrackId")
join "Invoice" as i
using ("InvoiceId")
where i."BillingCountry" = 'USA'
group by album_title
order by album_units_sold desc
limit 10;
```

- **Screenshot of Query Results:**

	album_title	album_units_sold
1	The Office, Season 3	14
2	Unplugged	11
3	Prenda Minha	11
4	Chill: Brazil (Disc 2)	10
5	Back to Black	9
6	International Superhits	8
7	B-Sides 1980-1990	7
8	A-Sides	7
9	Vin◆cius De Moraes - Sem Limite	7
10	Serie Sem Limite (Disc 2)	6

- **Description of Query Results:**

  Based on the query results, we collected the top-selling albums along with the number of tracks sold:

  1. The Office, Season 3 – 14 tracks sold (Most popular)
  2. Unplugged – 11 tracks sold
  3. Prenda Minha – 11 tracks sold
  4. Chill: Brazil (Disc 2) – 10 tracks sold
  5. Back to Black – 9 tracks sold
  6. International Superhits – 8 tracks sold
  7. B-Sides 1980-1990 – 7 tracks sold
  8. A-Sides – 7 tracks sold
  9. Vinícius De Moraes - Sem Limite – 7 tracks sold
  10. Serie Sem Limite (Disc 2) – 6 tracks sold

  The Office, Season 3 dominates the sales, showing strong demand for TV show soundtracks.

  Unplugged and *Prenda Minha* have equal track sales, indicating their popularity.

Brazilian music (*Chill: Brazil*, *Vinícius De Moraes*) appears twice, highlighting its appeal.
A mix of rock (*International Superhits*), alternative (*Back to Black*), and compilation albums (*B-Sides 1980-1990*, *A-Sides*) shows varied audience preferences.

7. (10 point) Provide a table that aggregates purchase data by country. In cases where a country has only one customer, group these countries as 'Other.' The results should be sorted by total sales in descending order.

Information to calculate:

- Total Number of Customers: Calculate the count of unique customers within each country.
- Total Value of Sales: Sum the total sales value for each country.
- Average Value of Sales per Customer: Divide the total sales value by the number of unique customers in each country
- Average Order Value: Divide the total sales value by the number of orders (invoices) placed in each country to calculate the average order value.

Answer here

- **SQL Query Syntax:**

```
with CustomerSales as (
    select
            c."Country",
            count(distinct c."CustomerId") as TotalCustomers,
            count(distinct i."InvoiceId") AS TotalOrders,
            sum(il."Quantity" * il."UnitPrice") as TotalSales
    from "Customer" as c
    join "Invoice" as i
    on c."CustomerId" = i."CustomerId"
    join "InvoiceLine" as il
    on i."InvoiceId" = il."InvoiceId"
    group by c."Country"
)
select
    case
            when TotalCustomers = 1
            then 'Other'
            else "Country"
    end as Country,
    sum(TotalCustomers) as TotalCustomers,
    sum(TotalSales) as TotalSales,
    sum(TotalSales) / sum(TotalCustomers) as AverageSalesPerCustomer,
    sum(TotalSales) / sum(TotalOrders) as AverageOrderValue
from CustomerSales
group by Country
order by TotalSales desc;
```

- **Screenshot of Query Results:**

	country	totalcustomers	totalsales	averagesalespercustomer	averageordervalue
1	Other	15	604.3	40.2866666667	5.7552380952
2	USA	13	523.06	40.2353846154	5.7479120879
3	Canada	8	303.96	37.995	5.4278571429
4	France	5	195.1	39.02	5.5742857143
5	Brazil	5	190.1	38.02	5.4314285714
6	Germany	4	156.48	39.12	5.5885714286
7	United Kingdom	3	112.86	37.62	5.3742857143
8	Czech Republic	2	90.24	45.12	6.4457142857
9	Portugal	2	77.24	38.62	5.5171428571
10	India	2	75.26	37.63	5.7892307692

- **Description of Query Results:**
  In this query, we collected an overview of purchase data across different countries, including information such as Total Customers**,** Total Sales**,** Average Sales Per Customer**, and** Average Order Value**.**
  The "Other" category (countries with a single customer) leads in total sales (604.30) and customers (15), indicating untapped potential. USA (523.06) and Canada (303.96) are top revenue contributors, while France, Germany, and the UK maintain strong sales. Czech Republic (AOV 6.45) and India (AOV 5.79) show high spending potential.

8. (10 point) Some genres have low sales, the product team wants to analyze which genres need to be boosted by carrying out additional promotion or other strategies. Because each country has different behavior, the product team started by analyzing sales in USA
   (The total sales are obtained by multiplying the quantity of items sold by their respective prices)
   Answer here
   - **SQL Query Syntax**

```sql
select g."Name" as genre_name,
            count(il."Quantity") as qty,
            sum(il."Quantity" * il."UnitPrice") as total_sales
from "Track" as t
join "Genre" as g
using ("GenreId")
join "InvoiceLine" as il
using ("TrackId")
join "Invoice" as i
using ("InvoiceId")
where i."BillingCountry" = 'USA'
group by genre_name
order by total_sales asc, genre_name asc;
```

   - **Screenshot of Query Results:**

	genre_name	qty	total_sales
1	Science Fiction	1	1.99
2	Easy Listening	3	2.97
3	Rock And Roll	3	2.97
4	Heavy Metal	4	3.96
5	Hip Hop/Rap	4	3.96
6	Soundtrack	4	3.96
7	Alternative	5	4.95
8	Pop	5	4.95
9	Reggae	6	5.94
10	Bossa Nova	7	6.93
11	Classical	8	7.92
12	Sci Fi & Fantasy	5	9.95
13	R&B/Soul	12	11.88
14	Drama	6	11.94
15	Blues	15	14.85
16	Comedy	8	15.92
17	Jazz	22	21.78
18	TV Shows	14	27.86
19	Alternative & Punk	50	49.5
20	Metal	64	63.36
21	Latin	91	90.09
22	Rock	157	155.43

   - **Description of Query Results:**
     In this query, we collected some genres that need to be boosted by carrying out additional promotion or other strategies. The genres are :

1. Science Fiction = Lowest sales at only $1.99 with 1 track sold, indicating minimal demand.
2. Easy Listening & Rock and Roll = Both generated $2.97 with 3 tracks sold, showing weak performance.
3. Heavy Metal, Hip Hop/Rap & Soundtrack = Both generated $3.96 with 4 tracks sold, suggesting limited interest.
4. Alternative & Pop = Both generated $4.95 with 5 tracks sold, but remain underperforming compared to other mainstream genres.
5. Reggae & Bossa Nova – With sales of $5.94 and $6.93, these genres might benefit from niche-targeted promotion

Now, let's deep dive into the behavior of our customers

9. (10 point) We want to advertise songs to the customer based on how much each customers spent per genre. Help Marketing Team to find Top genre for each customers with the most spent

Answer here

- **SQL Query Syntax:**

```
with CustomerGenreSales as (
    -- (a) Menghitung total sales setiap customer untuk tiap genre
    select
        c."CustomerId",
        concat(c."FirstName", ' ', c."LastName") AS "CustomerName",
        g."Name" AS "Genre",
        SUM(il."Quantity" * il."UnitPrice") AS "TotalSpend"
    from "Customer" as c
    join "Invoice" as i on c."CustomerId" = i."CustomerId"
    join "InvoiceLine" as il ON i."InvoiceId" = il."InvoiceId"
    join "Track" as t on il."TrackId" = t."TrackId"
    join "Genre" as g on t."GenreId" = g."GenreId"
    group by c."CustomerId", c."FirstName", c."LastName", g."Name"
),
RankedSales as (
    -- (b) Memberi ranking berdasarkan total sales (descending)
    select
        "CustomerId",
        "CustomerName",
        "Genre",
        "TotalSpend",
        rank() over (partition by "CustomerId" order by "TotalSpend" desc) as "Rank"
    from CustomerGenreSales
)
-- (c) Filter hanya genre dengan ranking 1 untuk tiap customer dan tampilkan rank
select "CustomerId", "CustomerName", "Genre", "TotalSpend", "Rank"
from RankedSales
where "Rank" = 1;
```

- **Screenshot of Query Results:**

	CustomerId	CustomerName	Genre	TotalSpend	Rank
1	1	Lu\u{fffd}s Gon\u{fffd}alves	Rock	13.86	1
2	2	Leonie K\u{fffd}hler	Rock	16.83	1
3	3	Fran\u{fffd}ois Tremblay	Metal	9.9	1
4	4	Bj\u{fffd}rn Hansen	Rock	16.83	1
5	5	Franti\u{fffd}ek Wichterlov\u{fffd}	Rock	14.85	1
6	6	Helena Hol\u{fffd}	TV Shows	11.94	1
7	7	Astrid Gruber	Rock	14.85	1
8	8	Daan Peeters	Rock	20.79	1
9	9	Kara Nielsen	Rock	20.79	1
10	10	Eduardo Martins	Rock	28.71	1
11	11	Alexandre Rocha	Latin	15.84	1
12	12	Roberto Almeida	Latin	15.84	1
13	12	Roberto Almeida	Rock	15.84	1
14	13	Fernanda Ramos	Rock	10.89	1
15	14	Mark Philips	Rock	12.87	1
16	15	Jennifer Peterson	Rock	11.88	1
17	16	Frank Harris	Metal	11.88	1
18	17	Jack Smith	Rock	7.92	1
19	18	Michelle Brooks	Rock	18.81	1
20	19	Tim Goyer	Rock	14.85	1
21	20	Dan Miller	Latin	13.86	1
22	21	Kathy Chase	Rock	8.91	1
23	22	Heather Leacock	Metal	11.88	1
24	23	John Gordon	Latin	11.88	1
25	24	Frank Ralston	Rock	11.88	1
26	25	Victor Stevens	Rock	13.86	1
27	26	Richard Cunningham	Rock	13.86	1
28	27	Patrick Gray	Rock	15.84	1
29	28	Julia Barnett	Rock	16.83	1
30	29	Robert Brown	Rock	24.75	1
31	30	Edward Francis	Rock	18.81	1
32	31	Martha Silk	Latin	12.87	1
33	32	Aaron Mitchell	Latin	13.86	1
34	33	Ellie Sullivan	Rock	18.81	1
35	34	Jo\u{fffd}o Fernandes	Rock	14.85	1
36	35	Madalena Sampaio	Rock	15.84	1
37	36	Hannah Schneider	Metal	17.82	1
38	37	Fynn Zimmermann	Rock	10.89	1
39	38	Niklas Schr\u{fffd}der	Rock	20.79	1
40	39	Camille Bernard	Rock	12.87	1
41	40	Dominique Lefebvre	Rock	16.83	1
42	41	Marc Dubois	Rock	9.9	1
43	42	Wyatt Girard	Metal	13.86	1
44	43	Isabelle Mercier	Rock	17.82	1
45	44	Terhi H\u{fffd}m\u{fffd}l\u{fffd}inen	Rock	17.82	1
46	45	Ladislav Kov\u{fffd}cs	Rock	10.89	1
47	46	Hugh O'Reilly	TV Shows	13.93	1
48	47	Lucas Mancini	Rock	17.82	1
49	48	Johannes Van der Berg	Rock	17.82	1
50	49	Stanislaw W\u{fffd}jcik	Rock	21.78	1
51	50	Enrique Mu\u{fffd}oz	Rock	21.78	1
52	51	Joakim Johansson	Latin	11.88	1
53	52	Emma Jones	Latin	14.85	1
54	53	Phil Hughes	Rock	17.82	1
55	54	Steve Murray	Rock	10.89	1
56	55	Mark Taylor	Rock	21.78	1
57	56	Diego Guti\u{fffd}rrez	Alternative & Pu\u{fffd}	8.91	1
58	56	Diego Guti\u{fffd}rrez	Rock	8.91	1
59	57	Luis Rojas	Rock	8.91	1
60	58	Manoj Pareek	Rock	12.87	1
61	59	Puja Srivastava	Rock	11.88	1

- **Description of Query Results:**
  The query identifies the top genre for each customer based on total spending. Most customers prefer Rock, followed by Latin and Metal, with a few favoring TV Shows and Alternative & Punk. This insight helps the marketing team target customers with personalized song recommendations based on their highest-spending genre.

10. (10 point) The Marketing team wants to increase advertising in countries with customers who have spent the most money. Help the Marketing team find the top 10 countries with the highest-spending customers.

Answer here

- **SQL Query Syntax :**

```
select i."BillingCountry" as country,
       sum(i."Total") as total_spending
from "Invoice" as i
group by country
order by total_spending desc
limit 10;
```

- **Screenshot of Query Results:**

country	total_spending
USA	523.06
Canada	303.96
France	195.1
Brazil	190.1
Germany	156.48
United Kingdom	112.86
Czech Republic	90.24
Portugal	77.24
India	75.26
Chile	46.62

- **Description of Query Results:**
  In this query we collected the top 10 countries where customers have spent the most on purchases.

1. USA = Customers spent 523.06
2. Canada = Customers spent 303.96
3. France = Customers spent 195.10
4. Brazil = Customers spent 190.10
5. Germany = Customers spent 156.48
6. United Kingdom = Customers spent 112.86
7. Czech Republic = Customers spent 90.24
8. Portugal = Customers spent 77.24
9. India = Customers spent 75.26
10. Chile = Customers spent 46.62

The USA leads with total sales of 523.06, followed by Canada (303.96) and France (195.10). Other high-spending countries include Brazil, Germany, and the United Kingdom. The insights help the Marketing team focus advertising efforts on countries with the highest revenue potential.