AMAZON SALE'S DATA INSIGHTS

amazon

Location: Myanmar(Mandalay,Yangon,Nay Pyi Taw)

OVERVIEW OF AMAZON DATA

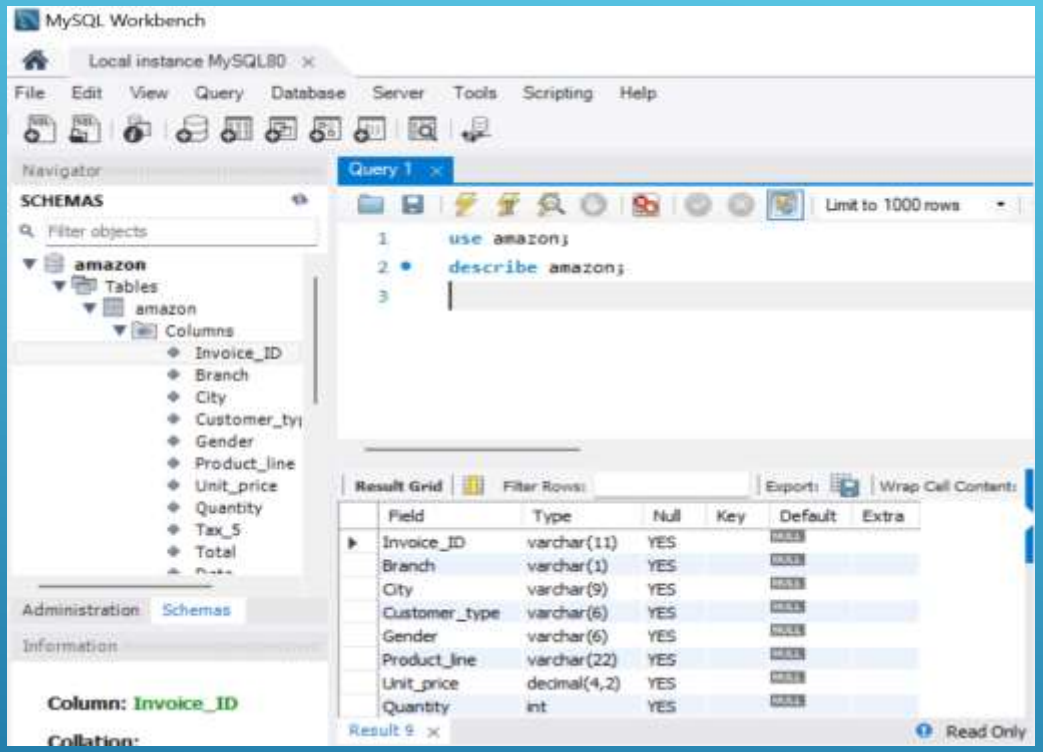
THE DATA CONSISTS OF SALES RECORD OF THREE CITIES IN **MYANMAR**

THE DATA CONSISTS OF 1000 ROWS AND 17 COLUMNS.

AFTER FEATURE ENGINEERING NUMBER OF ROWS WII BE 1000 AND NUMBER COLUMNS WILL BE 20.

OBJECTIVE OF PROJECT

THE AIM OF THE THIS PROJECT IS TO GAIN INSIGHTS INTO THE SALES DATA OF THE AMAZON TO UNDERSTAND THE DIFFERENT FACTORS THAT AFFECTS SALES OF THE DIFFERENT BRANCHES



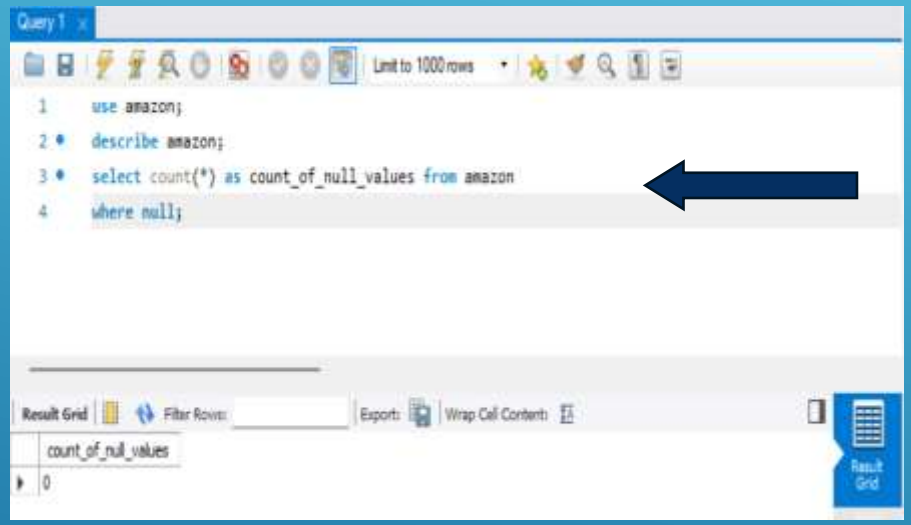
DATA WRANGLING

1) Create a database named amazon in MySQL.
2) Importing data in the form of a table in amazon database.

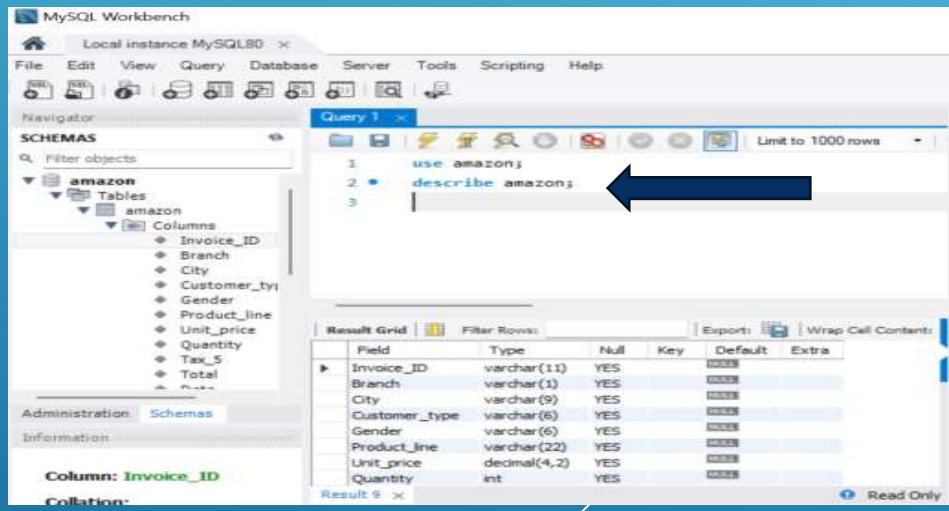
- **After completion of step one and step 2 we can start writing queries, in picture you can see that I have used a query (describe).**

CHECKING NULL VALUES AND DATATYPE.

NULL VALUES



DATA TYPE



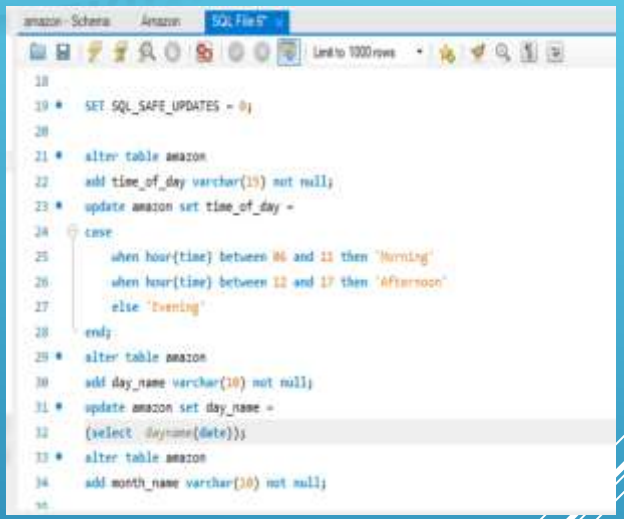
FEATURE ENGINEERING

In feature engineering we are creating some new columns to extract data from DATE and TIME column

This will help us to analyse the data in time format and day format more specifically

For example :1). time-of –day (Morning, Afternoon, Evening).

2). Day-of–week (Monday to Sunday)

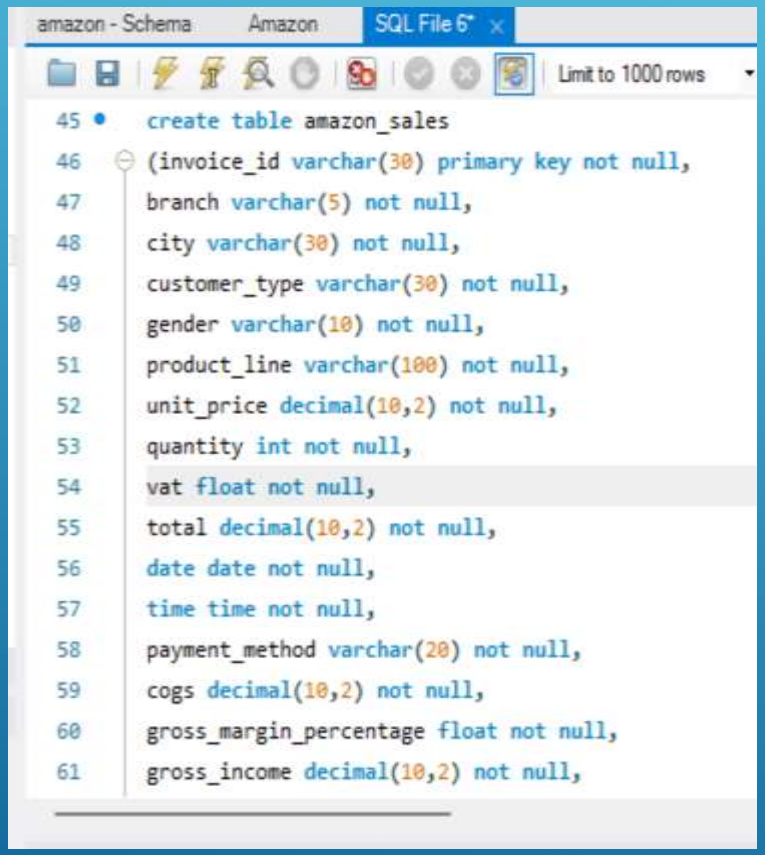


```
SET SQL_SAFE_UPDATES = 0;

alter table amazon
add time_of_day varchar(10) not null;
update amazon set time_of_day =
case
    when hour(time) between 06 and 11 then 'Morning'
    when hour(time) between 12 and 17 then 'Afternoon'
    else 'Evening'
end;
alter table amazon
add day_name varchar(10) not null;
update amazon set day_name =
(select dayname(date));
alter table amazon
add month_name varchar(10) not null;
```

price	Quantity	Tax_5	Total	Date	Time	Payment	cogs	gross_margin_percentage	gross_income	Rating	time_of_day	day_name	month_name
	7	26.1415	548.9715	2019-01-05 00:00:00	13:08:00	Ewallet	522.83	4.761904762	26.1415	9.1	Afternoon	Saturday	January
	5	3.8200	80.2200	2019-03-08 00:00:00	10:29:00	Cash	76.40	4.761904762	3.8200	9.6	Morning	Friday	March
	7	16.2155	340.5255	2019-03-03 00:00:00	13:23:00	Credit card	324.31	4.761904762	16.2155	7.4	Afternoon	Sunday	March
	8	23.2880	489.0480	2019-01-27 00:00:00	20:33:00	Ewallet	465.76	4.761904762	23.2880	8.4	Evening	Sunday	January
	7	30.2085	634.3785	2019-02-08 00:00:00	10:37:00	Ewallet	604.17	4.761904762	30.2085	5.3	Morning	Friday	February
	7	29.8865	627.6165	2019-03-25 00:00:00	18:30:00	Ewallet	597.73	4.761904762	29.8865	4.1	Evening	Monday	March
	6	20.6520	433.6920	2019-02-25 00:00:00	14:36:00	Ewallet	413.04	4.761904762	20.6520	5.8	Afternoon	Monday	February
	10	36.7800	772.3800	2019-02-24 00:00:00	11:38:00	Ewallet	735.60	4.761904762	36.7800	8.0	Morning	Sunday	February

EXPLORATORY DATA ANALYSIS



Creating new table named amazon sales by adding correct column name, datatype, constraints while copying values from demo table amazon

Checking size of table, count of null values, unique values

```
72
73 •  select count(*) as total_columns from information_schema.columns
74    where table_name ='amazon_sales';
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

total_columns
20

TOTAL COLUMNS

TOTAL ROWS

```
73 •  select count(*) as total_columns from information_schema.columns
74    where table_name ='amazon_sales';
75 •  select count(*) as total_rows from amazon_sales;
76 •  select count(*) as null_values from amazon_sales where null;
77
78 •  create view count_unique_values as
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

total_rows
1000

```
85 •  select * from count_unique_values;
86
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

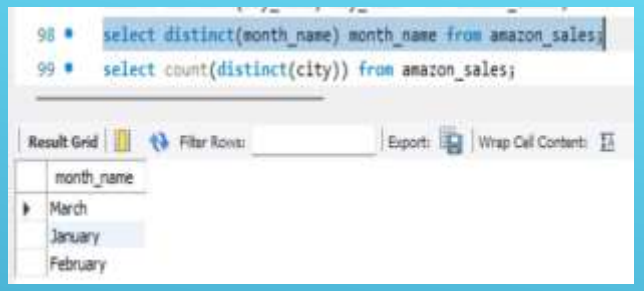
invoice_id	branch	city	customertype	gender	product_line	unit_price	quantity	vat	total	date	time	payment_me
1000	3	3	2	2	6	943	10	990	990	89	506	3

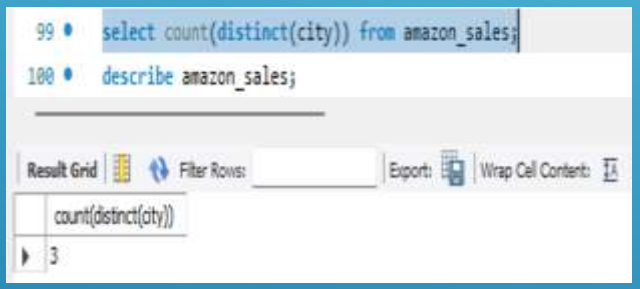
UNIQUE VALUES

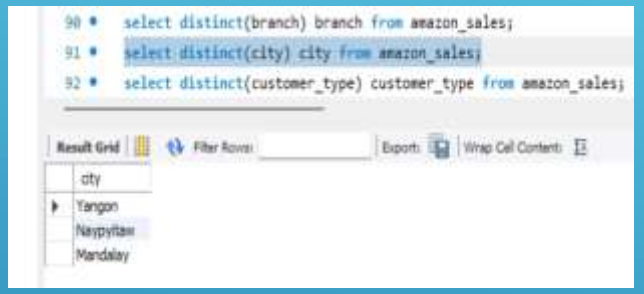
Checking the unique values in each categorical column. There are 10 categorical columns .

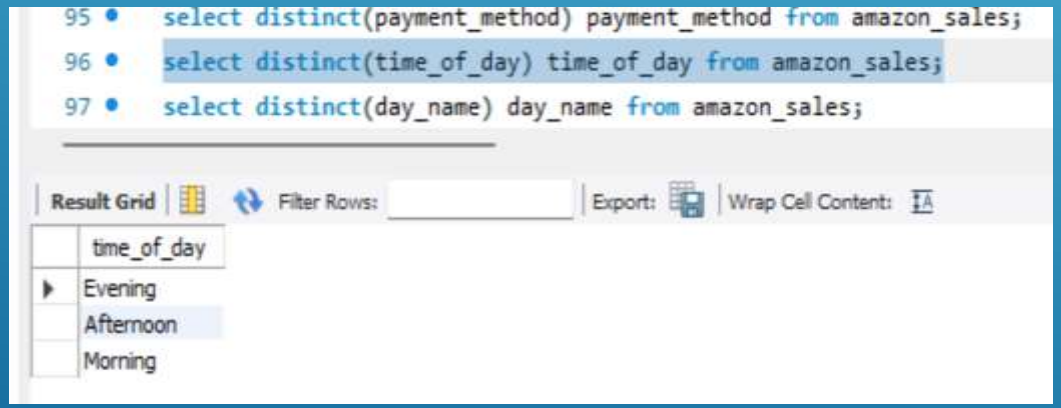
1.) invoice_id
2.)branch
3.)city
4.)customer_type
5.)gender
6.)product_line
7.)payment_method
8.)time_of_day
9.)day_name
10.)month_name

Some of the unique values









QUESTIONS RELATED TO DATASET

1.) What is the count of distinct cities in the dataset?

select count(distinct(city)) from amazon_sales;

2.) For each branch, what is the corresponding city?

select distinct city, branch from amazon_sales;

3.) What is the count of distinct product lines in the dataset?

select count(distinct(product_line)) from amazon_sales;

4.) Which payment method occurs most frequently?

```
select payment_method, count(*) as occurance from amazon_sales
group by payment_method
order by occurance desc;
```

5.) Which product line has the highest sales?

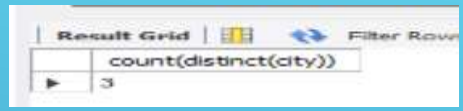
```
select product_line, sum(quantity) as total_sales from amazon_sales
group by product_line
order by total_sales desc;
```

6.) How much revenues is generated each month?

```
select month_name, sum(total) as monthly_revenue$ from amazon_sales
group by month_name
order by monthly_revenue$ desc;
```

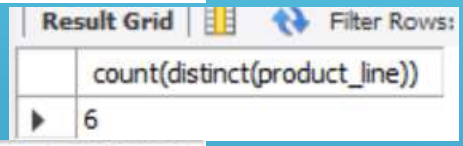
7.) In which month did the cost of goods sold reach its peak?

```
select month_name, sum(cogs) as cost_of_goods_sold from amazon_sales
group by month_name
order by cost_of_goods_sold desc;
```



count(distinct(city))
3

city	branch
Yangon	A
Naypyitaw	C
Mandalay	B



Result Grid | Filter Rows:

count(distinct(product_line))
6

payment_method	occurance
Ewallet	345
Cash	344
Credit card	311

product_line	total_sales
Electronic accessories	971
Food and beverages	952
Sports and travel	920
Home and lifestyle	911
Fashion accessories	902
Health and beauty	854

month_name	monthly_revenue$
January	116292.11
March	109455.74
February	97219.58

month_name	cost_of_goods_sold
January	110754.16
March	104243.34
February	92589.88

8.) Which product line generated the highest revenue?

```
select product_line, sum(total) as total_revenue$ from amazon_sales
group by product_line
order by total_revenue$ desc;
```

product_line	total_revenue$
Food and beverages	56144.96
Sports and travel	55123.00
Electronic accessories	54337.64
Fashion accessories	54306.03
Home and lifestyle	53861.96
Health and beauty	49193.84

9.) In which city was the highest revenue recorded?

```
select city, sum(total) as revenue$ from amazon_sales
group by city
order by revenue$ desc;
```

city	revenue$
Naypyitaw	110568.86
Yangon	106200.57
Mandalay	106198.00

10.) Which product line incurred the highest value added tax ?

```
select product_line, max(vat) highest_vat  from amazon_sales
group by product_line
order by highest_vat desc;
```

product_line	highest_vat
Fashion accessories	49.65
Food and beverages	49.26
Home and lifestyle	48.75
Sports and travel	47.72
Health and beauty	45.25
Electronic accessories	44.8785

11.) For each product line, add a column indicating "Good" if its sales are above average, otherwise "Bad".

```
case
      when sum(total) > (select sum(total)/count(distinct(product_line)) from amazon_sales) then 'Good'
  else 'Bad'
end performance
from amazon_sales
group by product_line;
```

product_line	revenue	performance
Food and beverages	56144.96	Good
Health and beauty	49193.84	Bad
Sports and travel	55123.00	Good
Fashion accessories	54306.03	Good
Home and lifestyle	53861.96	Good
Electronic accessories	54337.64	Good

12.) Identify the branch that exceeded the average number of products sold.

```
select branch, sum(quantity) as product_sold from amazon_sales
group by branch
having product_sold > (select sum(quantity)/count(distinct branch) as avg_quantity from amazon_sales);
```

branch	product_sold
A	1859

13.) Which product line is most frequently associated with each gender?

```
(select gender, product_line, count(*) as count from amazon_sales
group by gender, product_line),
Max_count as
(select max(count) from new group by gender)
select * from new
where count in (select * from max_count) limit 2;
```

gender	product_line	count
Male	Health and beauty	88
Female	Fashion accessories	96

14.) Calculate the average rating for each product line .

```
select product_line, avg(rating) as avg_rating from amazon_sales
group by product_line;
```

product_line	avg_rating
Food and beverages	7.11322
Health and beauty	7.00329
Sports and travel	6.91627
Fashion accessories	7.02921
Home and lifestyle	6.83750
Electronic accessories	6.92471

15.) Count sales occurrences for each time of day on every weekday.

```
select day_name, time_of_day, count(*) sales from amazon_sales
group by day_name, time_of_day
order by field(day_name, 'Sunday', 'Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday', 'Saturday'),
field(time_of_day, 'Morning', 'Afternoon', 'Evening');
```

day_name	time_of_day	sales
Sunday	Morning	22
Sunday	Afternoon	70
Sunday	Evening	41
Monday	Morning	21
Monday	Afternoon	75
Monday	Evening	29
Tuesday	Morning	36
Tuesday	Afternoon	71

16.) Identify the customer type contributing the highest revenue.

```
select customer_type, sum(total) as revenue from amazon_sales
group by customer_type
order by revenue desc;
```

customer_type	revenue
Member	164223.81
Normal	158743.62

17.) Determine the city with the highest VAT percentage.

```
select city, max(vat) as vat_percentage from amazon_sales
group by city
order by vat_percentage desc;
```

city	vat_percentage
Naypyitaw	49.65
Yangon	49.49
Mandalay	48.69

18.) Identify the customer type with the highest VAT payments.

```
select customer_type, max(vat) as vat_percentage from amazon_sales
group by customer_type
order by vat_percentage desc;
```

customer_type	vat_percentage
Member	49.65
Normal	49.49

19.) What is the count of the distinct customer types in the dataset?

```
select count(distinct(customer_type)) as count_distinct_customer_type from amazon_sales;
```

count_distinct_customer_type
2

20.)What is the count of distinct payment methods in the dataset?

```
select count(distinct(payment_method)) as count_distinct_payment from amazon_sales;
```

count_distinct_payment
3

21.) Which customer type occurs most frequently?

```
select customer_type, count(*) as count from amazon_sales
group by customer_type
order by count desc;
```

customer_type	count
Member	501
Normal	499

22.) Identify the customer type with the highest purchase frequency.

```
select customer_type, count(*) as count from amazon_sales
group by customer_type
order by count desc;
```

customer_type	purchase_frequency
Member	164223.81
Normal	158743.62

23.) Determine the predominant gender with the highest purchase frequency.

```
select gender, count(*) as count from amazon_sales
Group by gender
order by count desc;
```

gender	count
Female	501
Male	499

24.) Examine the distribution of genders within each branch.

```
select branch, gender, count(*) as count from amazon_sales
group by branch, gender
order by branch, gender;
```

branch	gender	count
A	Female	161
A	Male	179
B	Female	162
B	Male	170
C	Female	178
C	Male	150

179

25.) Identify the time of day when customers provide the most ratings.

```
select time_of_day, count(rating) as rating_count from amazon_sales
group by time_of_day
order by rating_count desc;
```

time_of_day	rating_count
Afternoon	528
Evening	281
Morning	191

26.) Determine the time of day with the highest customer rating for each branch.

```
select branch, time_of_day, max(rating) highest_rating from amazon_sales
group by branch, time_of_day
having highest_rating = (select max(x.max) from (select branch, time_of_day, max(rating) max from amazon_sales
group by branch, time_of_day) as x where x.branch= amazon_sales.branch)
order by branch
```

branch	time_of_day	highest_rating
A	Afternoon	10.0
B	Afternoon	10.0
B	Evening	10.0
B	Morning	10.0
C	Afternoon	10.0

27.) Identify the day of the week with the highest average ratings

```
select day_name, avg(rating) as avg_rating from amazon_sales
Group by day_name
order by avg_rating desc;
```

branch	day_name	highest_avg_rat
A	Friday	7.31200
B	Monday	7.33590
C	Friday	7.27895

28.) Determine the day of the week with the highest average ratings for each branch.

with avg_rating as

(select branch, day_name, avg(rating) avg_rat from amazon_sales group by branch, day_name),

max_rating as

(select max(avg_rat) from avg_rating group by branch)

select branch, day_name, avg_rat as highest_avg_rat from avg_rating where avg_rat in (select * from max_rating);

branch	day_name	highest_avg_rat
A	Friday	7.31200
B	Monday	7.33590
C	Friday	7.27895

INSIGHTS

CUSTOMER ANALYSIS

HIGHEST REVENUE GENDER : FEMALE[$167883.26]

DISTRIBUTION OF MEMEBERS BASED ON GENDER : MALE(240) FEMALE(261)

HIGHEST REVENUE BY CUSTOMER TYPE : MEMBER [$164223.81]

PRODUCT ANALYSIS

PRODUCT LINE WHICH HAS MADE HIGHEST SALE : ELECTRONIC ACCESSORIES (971)

PRODUCT LINE WHICH HAS MADE LOWEST SALE : HEALTH AND BEAUTY (854)

PRODUCT LINE WITH HIGHEST REVENUE : FOOD AND BEVERAGES [$56144.96]

PRODUCT LINE WITH LOWEST REVENUE : HEALTH AND BEAUTY [$49193.84]

SALES ANALYSIS

MONTH WITH HIGHEST REVENUE : JAN[$116292.11]

MONTH WITH LOWEST REVENUE : FEB[$97219.52]

CITY WITH HIGHEST REVENUE : NAY PYI TAW ,BANCH CODE(C),[$110568.86]

CITY WITH LOWEST REVENUE : MANDALAY, BRANCH CODE(B),[$106198.00]