AMAZON

Analysing Sales Data with SQL

Presented By:
Vigneshkumar S

AIM OF THE PROJECT

The major aim of this project is to gain insight into the sales data of Amazon to understand the different factors that affect sales of the different branches.

ABOUT DATA

The document contains sales data for various branches in Yangon, Naypyitaw, and Mandalay, including information on customer types, products, prices, quantities, and payment methods. The sales data shows a variety of product lines such as health and beauty, electronic accessories, home and lifestyle, food and beverages, fashion accessories, and sports and travel. The total sales amount for each transaction includes a 5% tax, cost of goods sold (COGS), gross margin percentage, and gross income. The ratings provided for each transaction range from 4.1 to 9.6, indicating the level of customer satisfaction with their purchases. The data contains 17 columns and 1000 rows.

ANALYSIS LIST

Product Analysis

This analysis aims to understand the different product lines, the products lines performing best and the product lines that need to be improved.

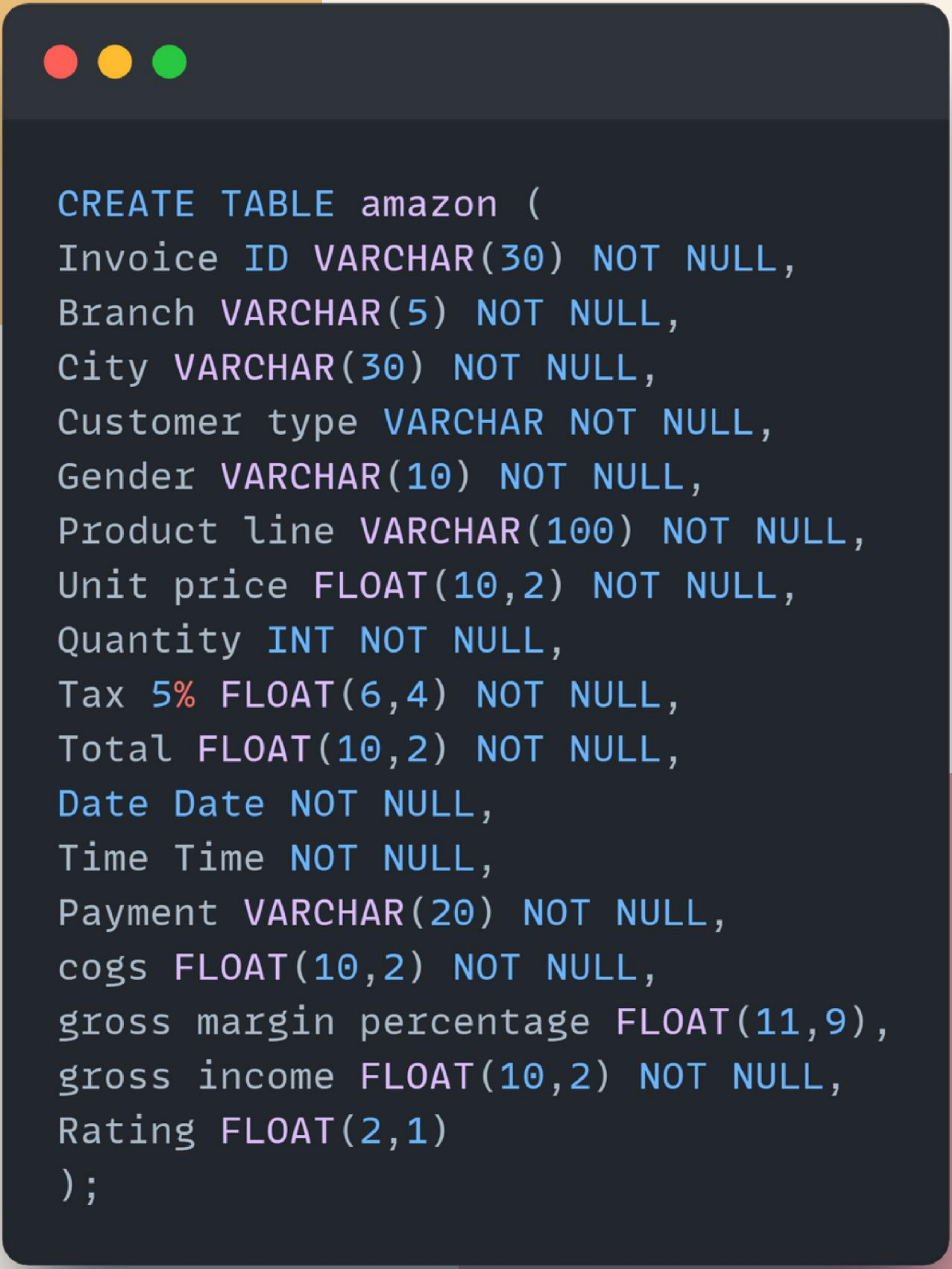
Sales Analysis

This analysis aims to answer the question of the sales trends of product. The result of this can help us measure the effectiveness of each sales strategy the business applies and what modifications are needed to gain more sales.

Customer Analysis

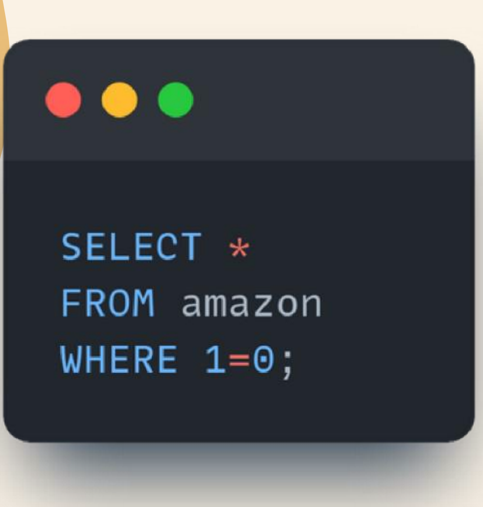
This analysis aims to uncover the different customer segments, purchase trends and the profitability of each customer segment.

CREATE TABLE



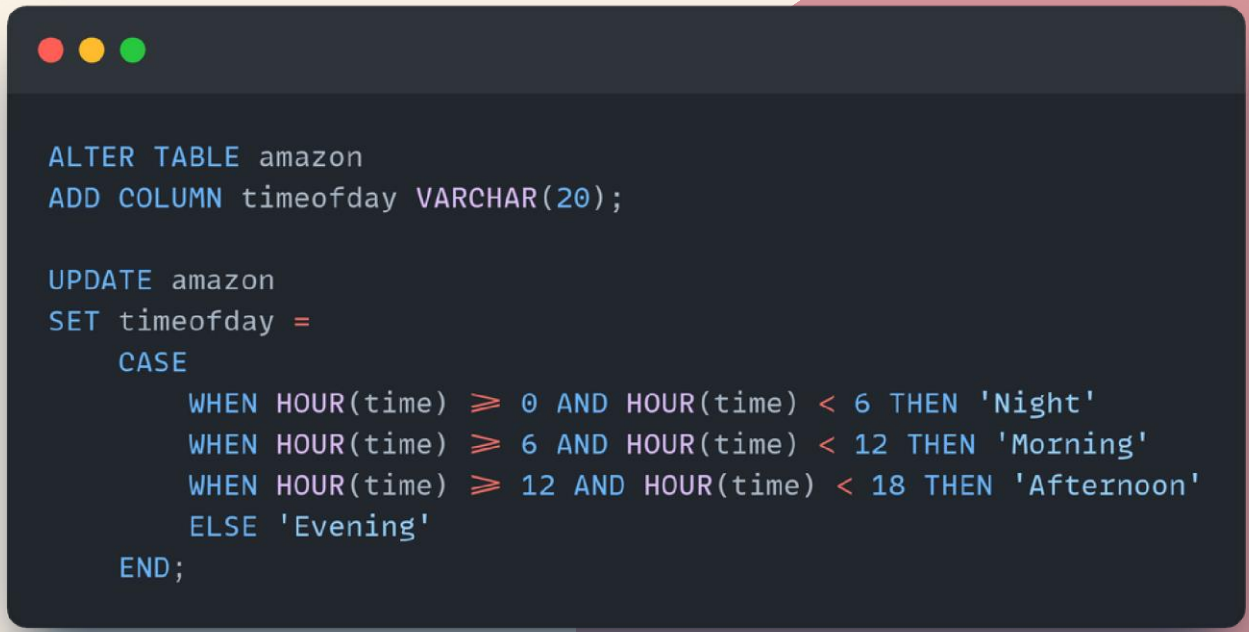
```
CREATE TABLE amazon (
Invoice ID VARCHAR(30) NOT NULL,
Branch VARCHAR(5) NOT NULL,
City VARCHAR(30) NOT NULL,
Customer type VARCHAR NOT NULL,
Gender VARCHAR(10) NOT NULL,
Product line VARCHAR(100) NOT NULL,
Unit price FLOAT(10,2) NOT NULL,
Quantity INT NOT NULL,
Tax 5% FLOAT(6,4) NOT NULL,
Total FLOAT(10,2) NOT NULL,
Date Date NOT NULL,
Time Time NOT NULL,
Payment VARCHAR(20) NOT NULL,
cogs FLOAT(10,2) NOT NULL,
gross margin percentage FLOAT(11,9),
gross income FLOAT(10,2) NOT NULL,
Rating FLOAT(2,1)
);
```

CHECK FOR NULL VALUES



```sql
SELECT *
FROM amazon
WHERE 1=0;
```

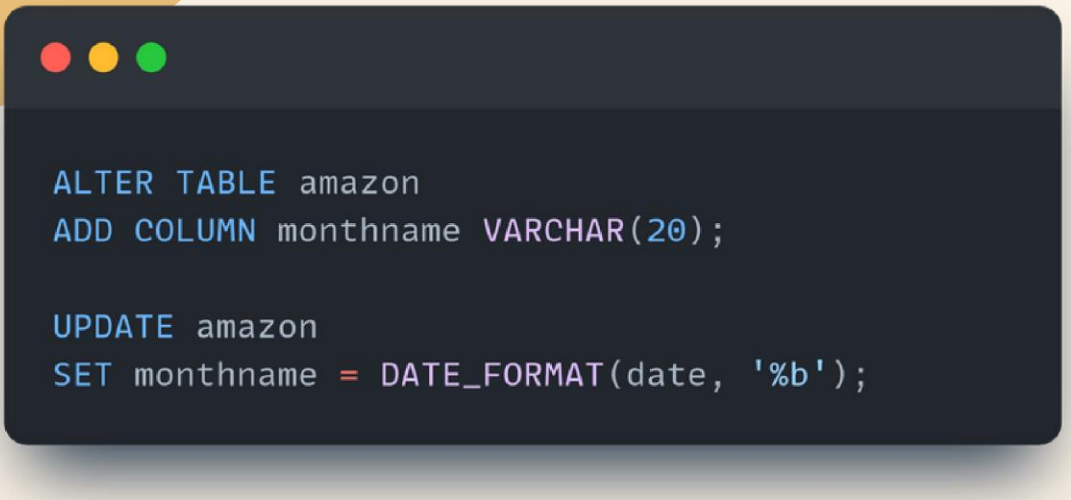
This condition ensures that no rows are returned.



```sql
ALTER TABLE amazon
ADD COLUMN timeofday VARCHAR(20);

UPDATE amazon
SET timeofday =
    CASE
        WHEN HOUR(time) >= 0 AND HOUR(time) < 6 THEN 'Night'
        WHEN HOUR(time) >= 6 AND HOUR(time) < 12 THEN 'Morning'
        WHEN HOUR(time) >= 12 AND HOUR(time) < 18 THEN 'Afternoon'
        ELSE 'Evening'
    END;
```

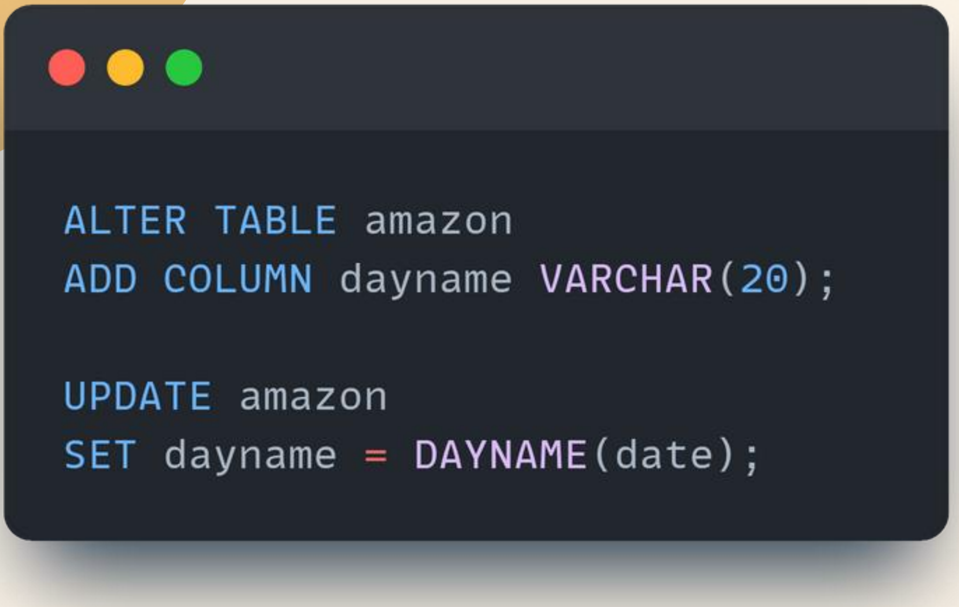
TO ADD NEW COLUMN timeofday TO ADD NEW COLUMN monthname



```
ALTER TABLE amazon
ADD COLUMN monthname VARCHAR(20);

UPDATE amazon
SET monthname = DATE_FORMAT(date, '%b');
```

TO ADD NEW COLUMN dayname



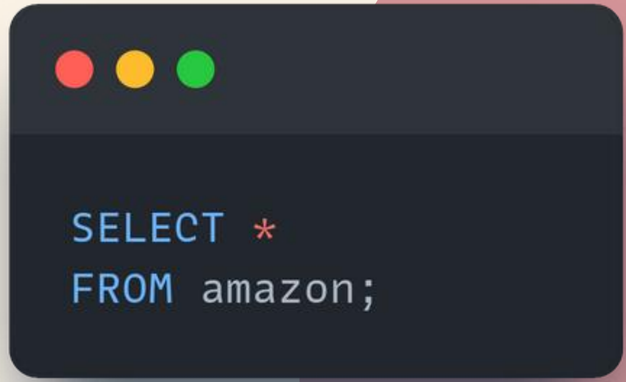
```
ALTER TABLE amazon
ADD COLUMN dayname VARCHAR(20);

UPDATE amazon
SET dayname = DAYNAME(date);
```

BUSINESS PROBLEMS

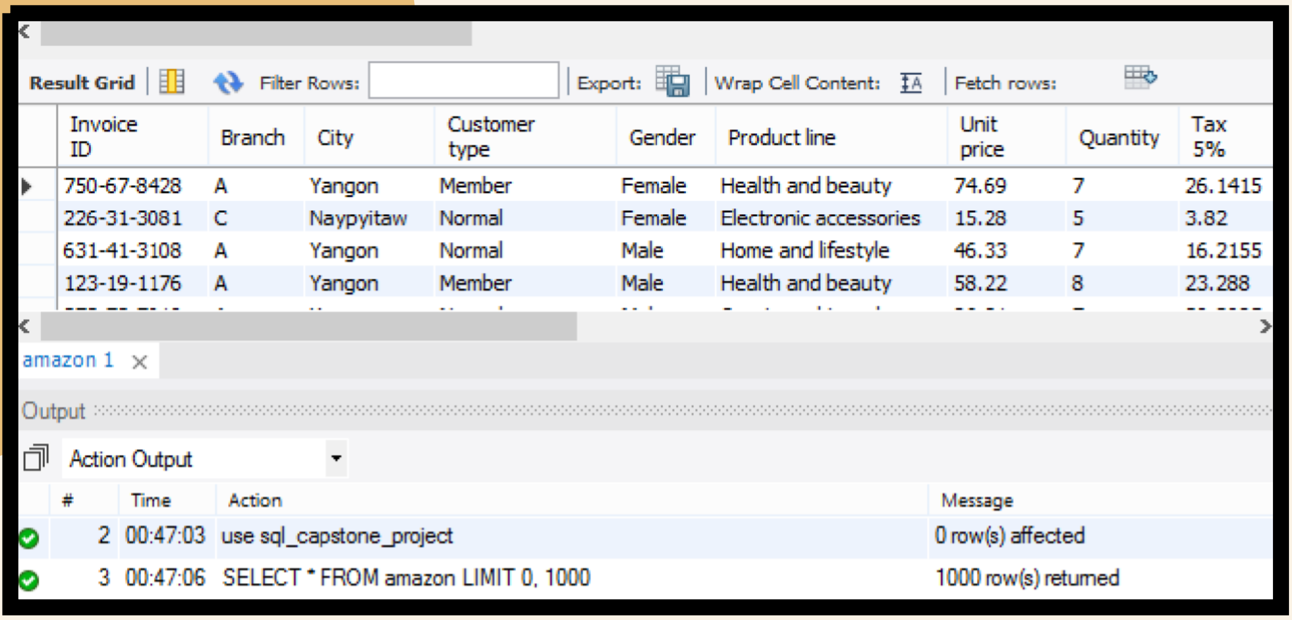
What are the columns in the table.

INPUT



```
SELECT *
FROM amazon;
```

OUTPUT



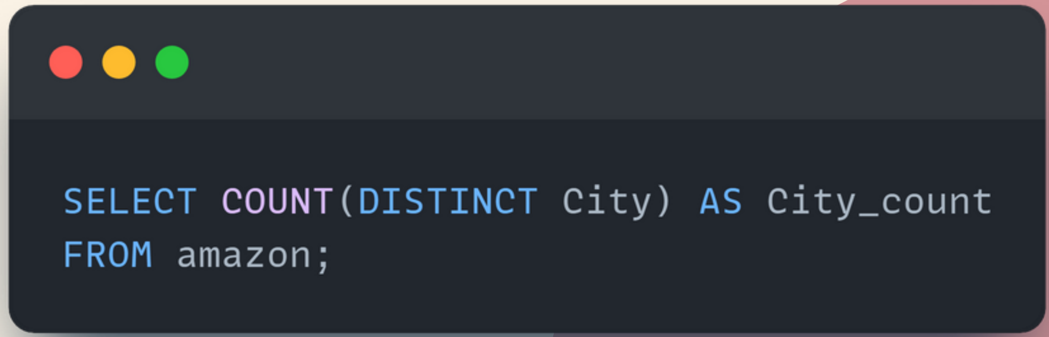
	Invoice ID	Branch	City	Customer type	Gender	Product line	Unit price	Quantity	Tax 5%
▶	750-67-8428	A	Yangon	Member	Female	Health and beauty	74.69	7	26.1415
	226-31-3081	C	Naypyitaw	Normal	Female	Electronic accessories	15.28	5	3.82
	631-41-3108	A	Yangon	Normal	Male	Home and lifestyle	46.33	7	16.2155
	123-19-1176	A	Yangon	Member	Male	Health and beauty	58.22	8	23.288

#	Time	Action	Message
2	00:47:03	use sql_capstone_project	0 row(s) affected
3	00:47:06	SELECT * FROM amazon LIMIT 0, 1000	1000 row(s) returned

The data contains 17 columns and 1000 rows.

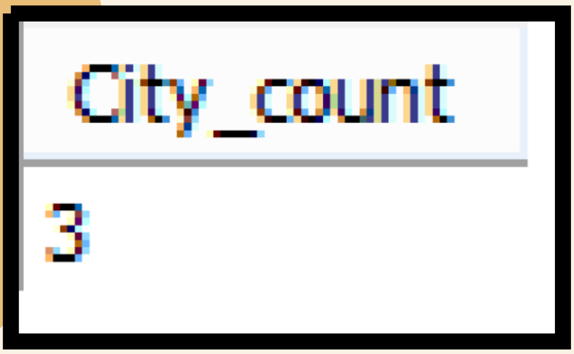
What is the count of distinct cities in the dataset?

INPUT



```
SELECT COUNT(DISTINCT City) AS City_count
FROM amazon;
```

OUTPUT

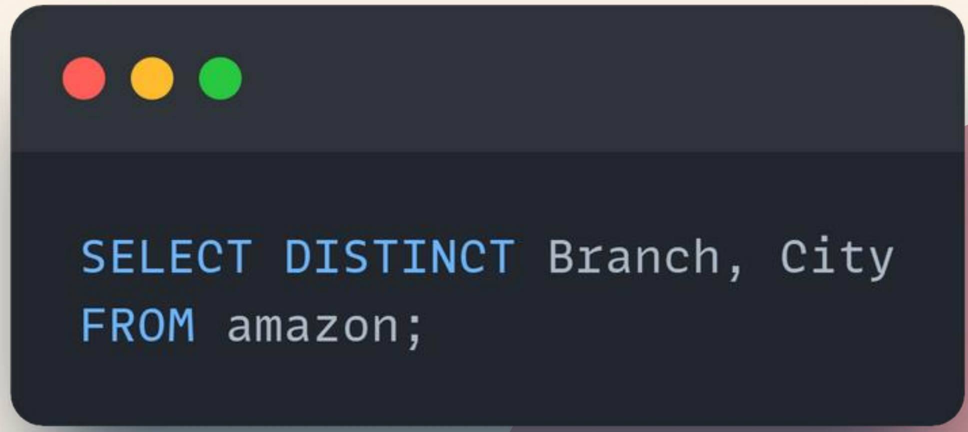


City_count
3

The count of distinct cities in the dataset is 3.

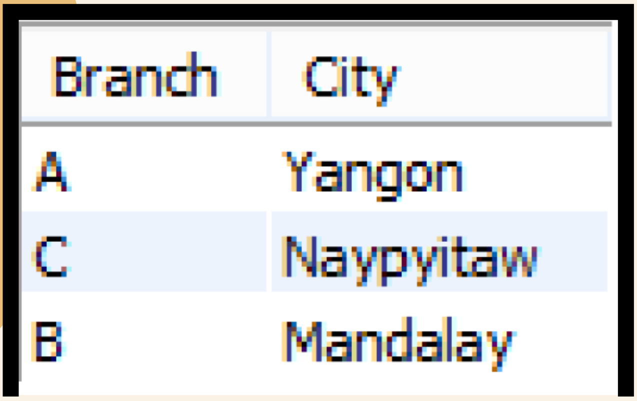
For each branch, what is the corresponding city?

INPUT



```
SELECT DISTINCT Branch, City
FROM amazon;
```

OUTPUT



Branch	City
A	Yangon
C	Naypyitaw
B	Mandalay

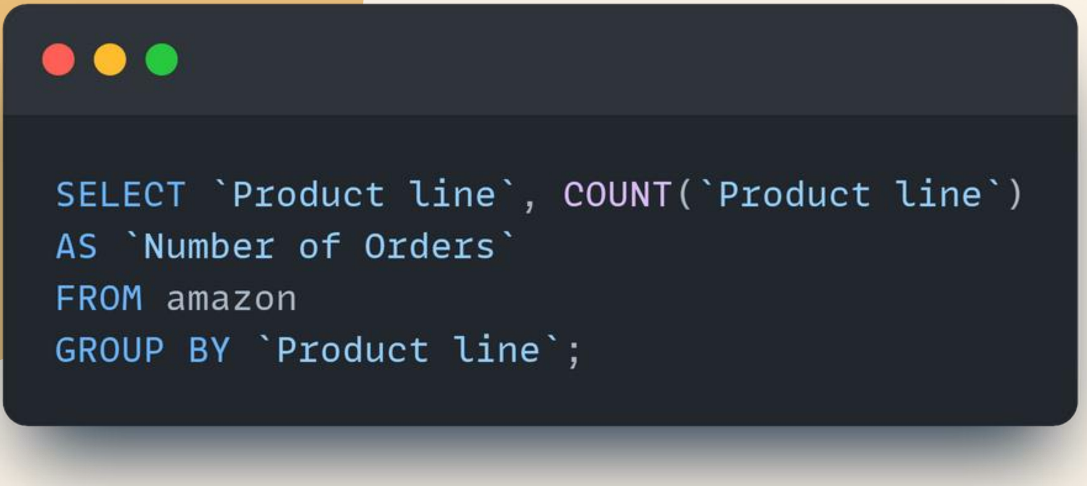
Branch A is located in Yangon.

Branch B is located in Mandalay.

Branch C is located in Naypyitaw.

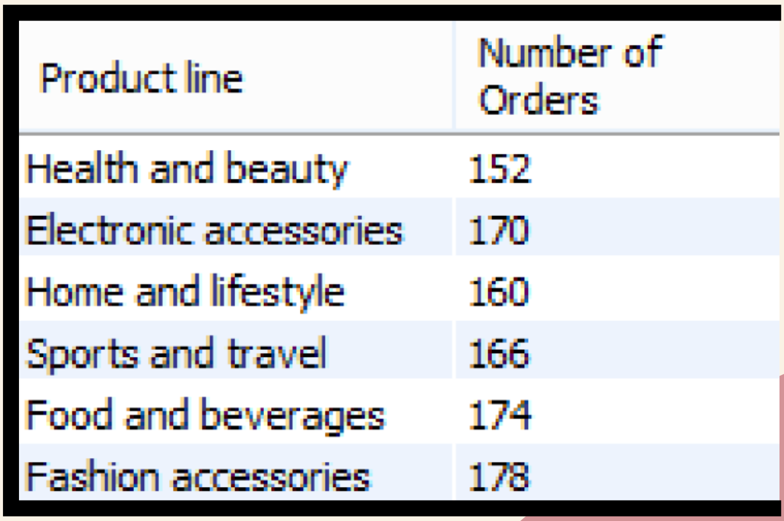
What is the count of distinct product lines in the dataset?

INPUT



```
SELECT `Product line`, COUNT(`Product line`)
AS `Number of Orders`
FROM amazon
GROUP BY `Product line`;
```

OUTPUT

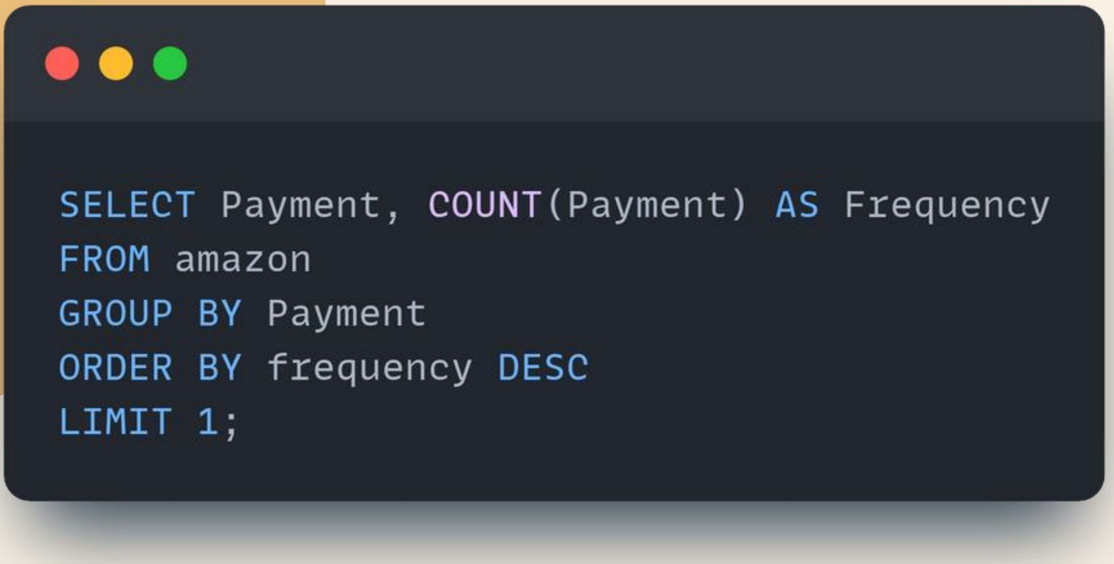


Product line	Number of Orders
Health and beauty	152
Electronic accessories	170
Home and lifestyle	160
Sports and travel	166
Food and beverages	174
Fashion accessories	178

The count of distinct product lines in the dataset is 6.

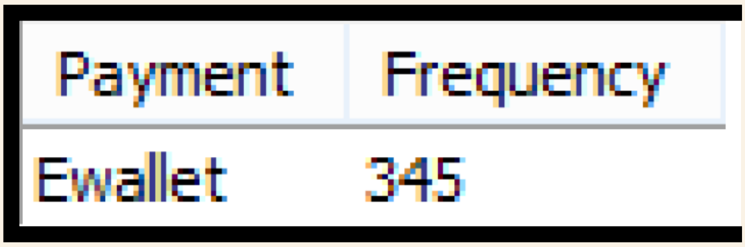
Which payment method occurs most frequently?

INPUT



```
SELECT Payment, COUNT(Payment) AS Frequency
FROM amazon
GROUP BY Payment
ORDER BY frequency DESC
LIMIT 1;
```

OUTPUT

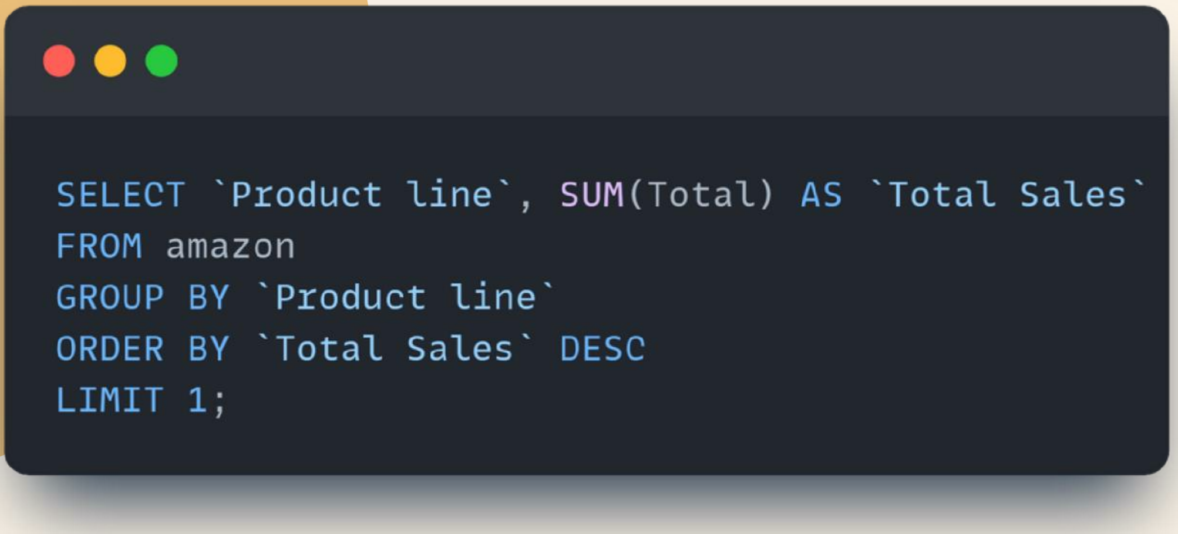


Payment	Frequency
Ewallet	345

The payment method that occurs most frequently is Ewallet with 345 occurrences.

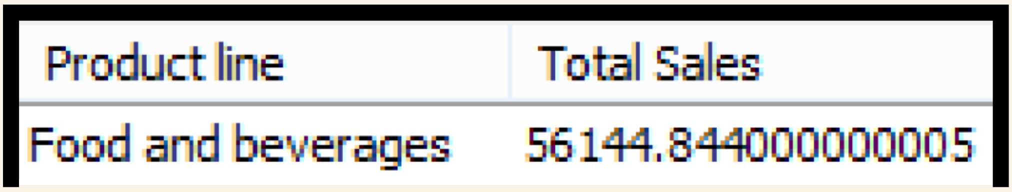
Which product line has the highest sales?

INPUT



```
SELECT `Product line`, SUM(Total) AS `Total Sales`
FROM amazon
GROUP BY `Product line`
ORDER BY `Total Sales` DESC
LIMIT 1;
```

OUTPUT

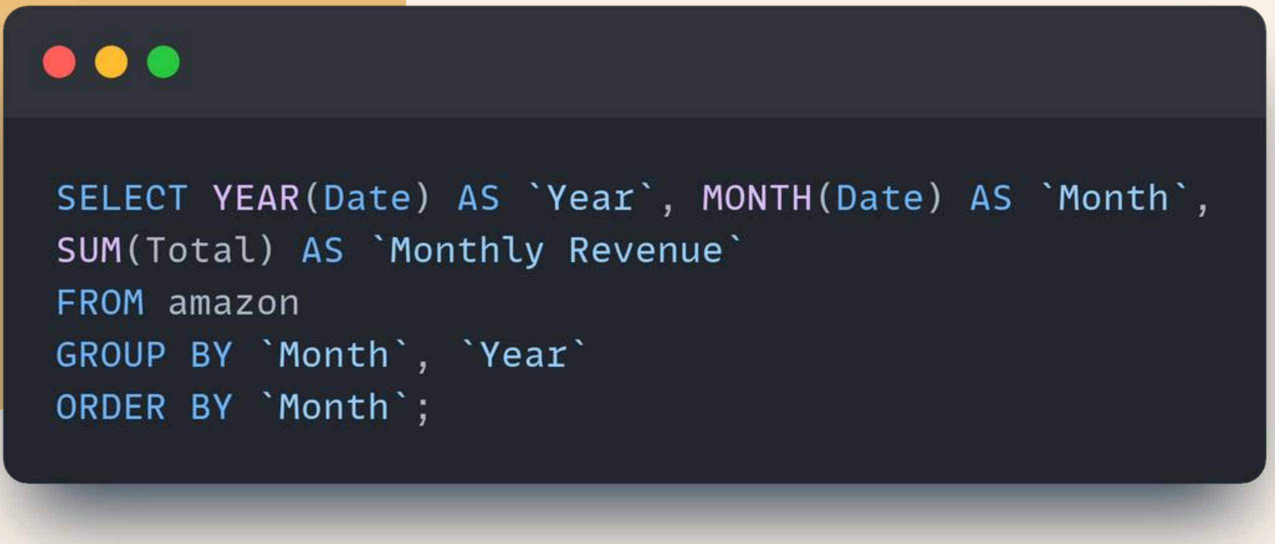


Product line	Total Sales
Food and beverages	56144.844000000005

The product line with the highest sales is 'Food and beverages' with total sales amounting to $56,144.84.

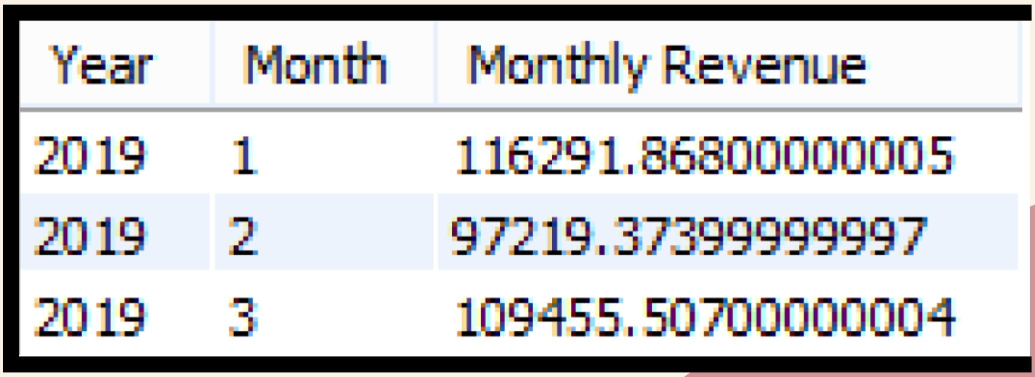
How much revenue is generated each month?

INPUT



```
SELECT YEAR(Date) AS `Year`, MONTH(Date) AS `Month`,
SUM(Total) AS `Monthly Revenue`
FROM amazon
GROUP BY `Month`, `Year`
ORDER BY `Month`;
```

OUTPUT



Year	Month	Monthly Revenue
2019	1	116291.86800000005
2019	2	97219.37399999997
2019	3	109455.50700000004

The revenue generated for each month is as follows:

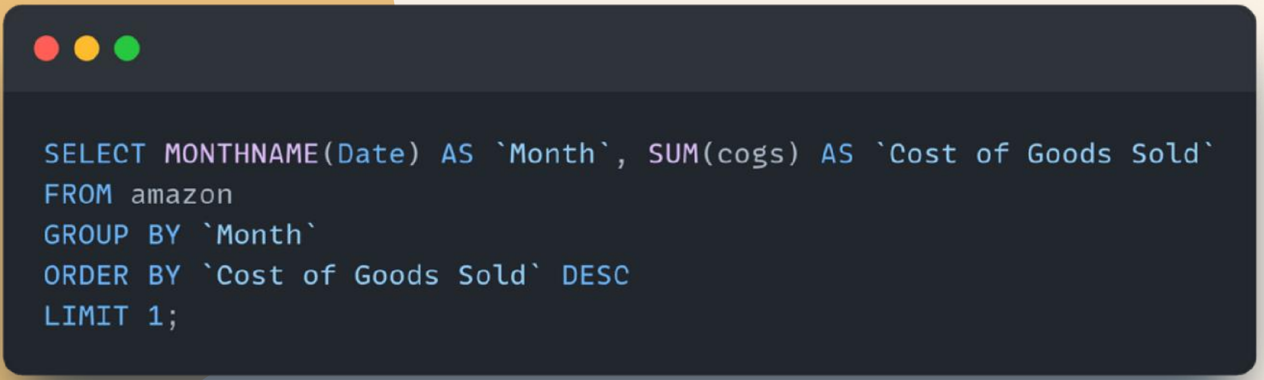
January 2019: $116,291.87

February 2019: $97,219.37

March 2019: $109,455.51

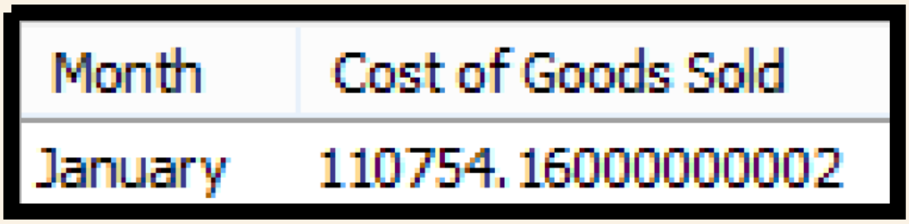
In which month did the cost of goods sold reach its peak?

INPUT



```sql
SELECT MONTHNAME(Date) AS `Month`, SUM(cogs) AS `Cost of Goods Sold`
FROM amazon
GROUP BY `Month`
ORDER BY `Cost of Goods Sold` DESC
LIMIT 1;
```

OUTPUT

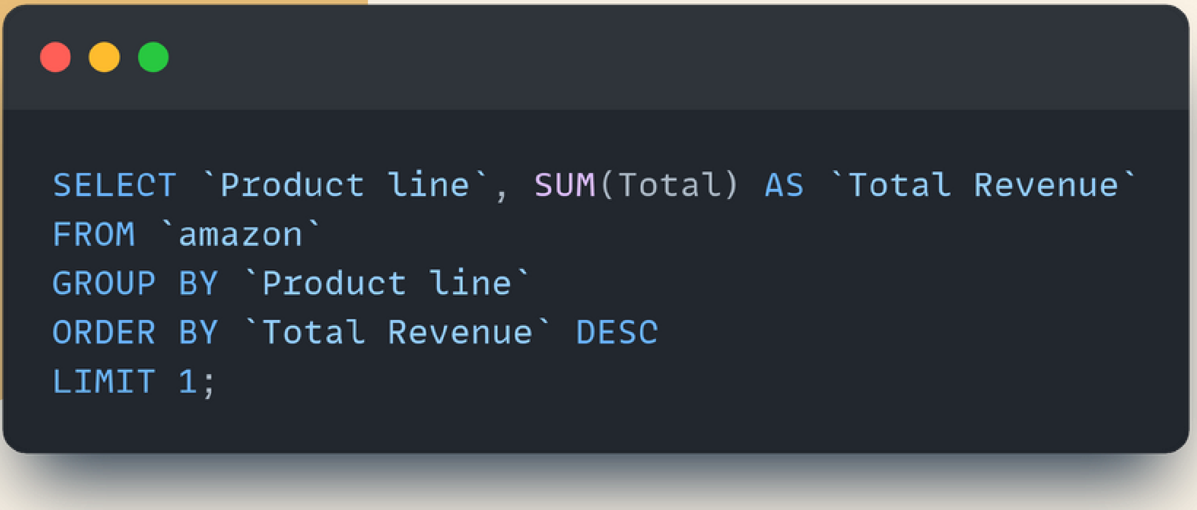


Month	Cost of Goods Sold
January	110754.16000000002

The month when the cost of goods sold reached its peak was January 2019 with a total cost of goods sold amounting to $110,754.16.

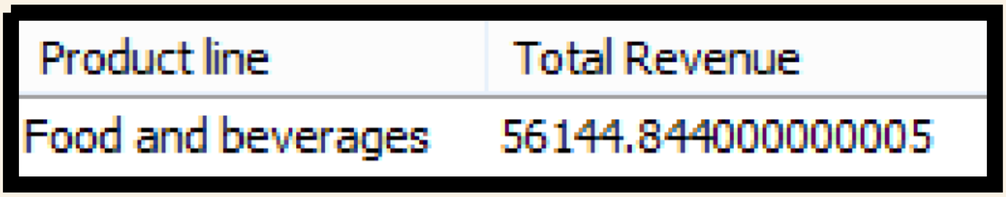
Which product line generated the highest revenue?

INPUT



```
SELECT `Product line`, SUM(Total) AS `Total Revenue`
FROM `amazon`
GROUP BY `Product line`
ORDER BY `Total Revenue` DESC
LIMIT 1;
```

OUTPUT

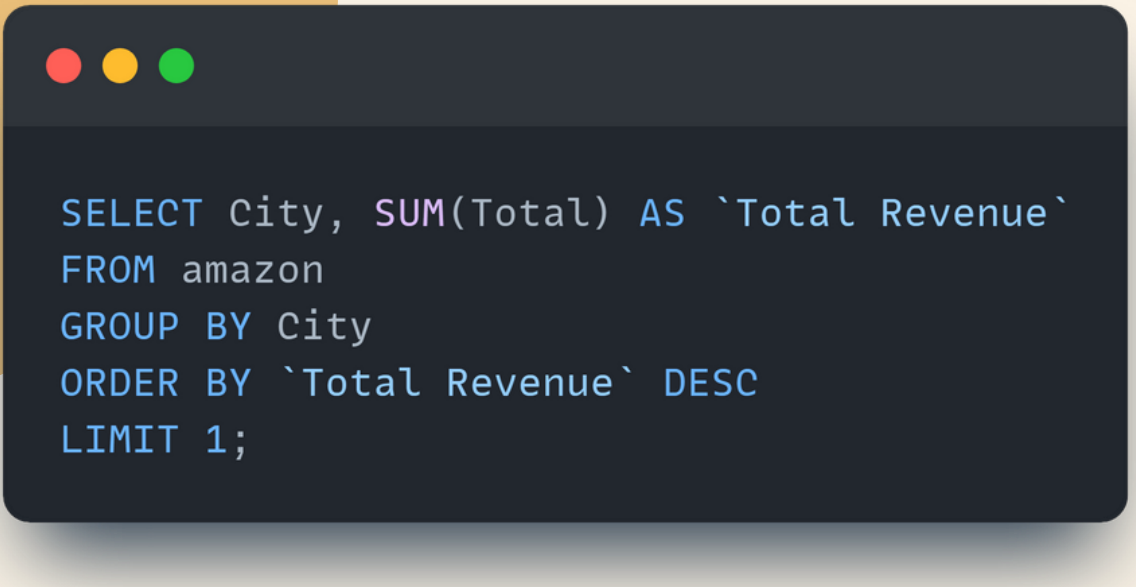


Product line	Total Revenue
Food and beverages	56144.844000000005

The product line that generated the highest revenue is "Food and beverages".

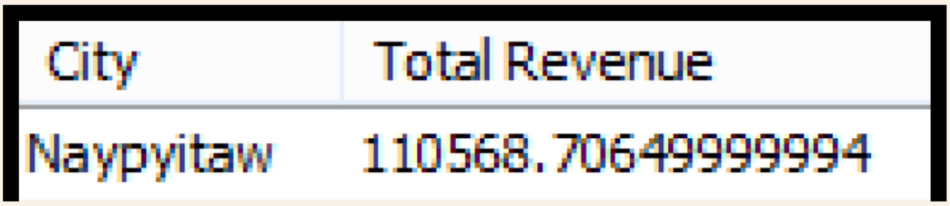
In which city was the highest revenue recorded?

INPUT



```
SELECT City, SUM(Total) AS `Total Revenue`
FROM amazon
GROUP BY City
ORDER BY `Total Revenue` DESC
LIMIT 1;
```

OUTPUT

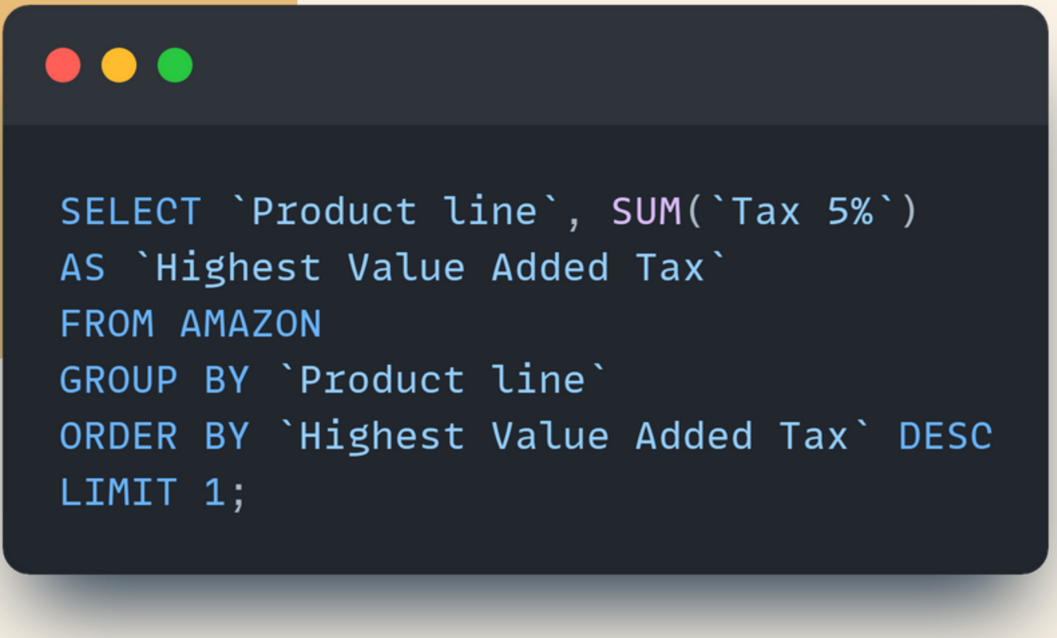


City	Total Revenue
Naypyitaw	110568.70649999994

The city with the highest revenue recorded is Naypyitaw.

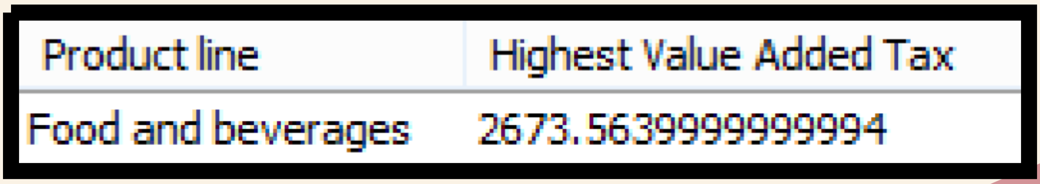
Which product line incurred the highest Value Added Tax?

INPUT



```sql
SELECT `Product line`, SUM(`Tax 5%`)
AS `Highest Value Added Tax`
FROM AMAZON
GROUP BY `Product line`
ORDER BY `Highest Value Added Tax` DESC
LIMIT 1;
```

OUTPUT

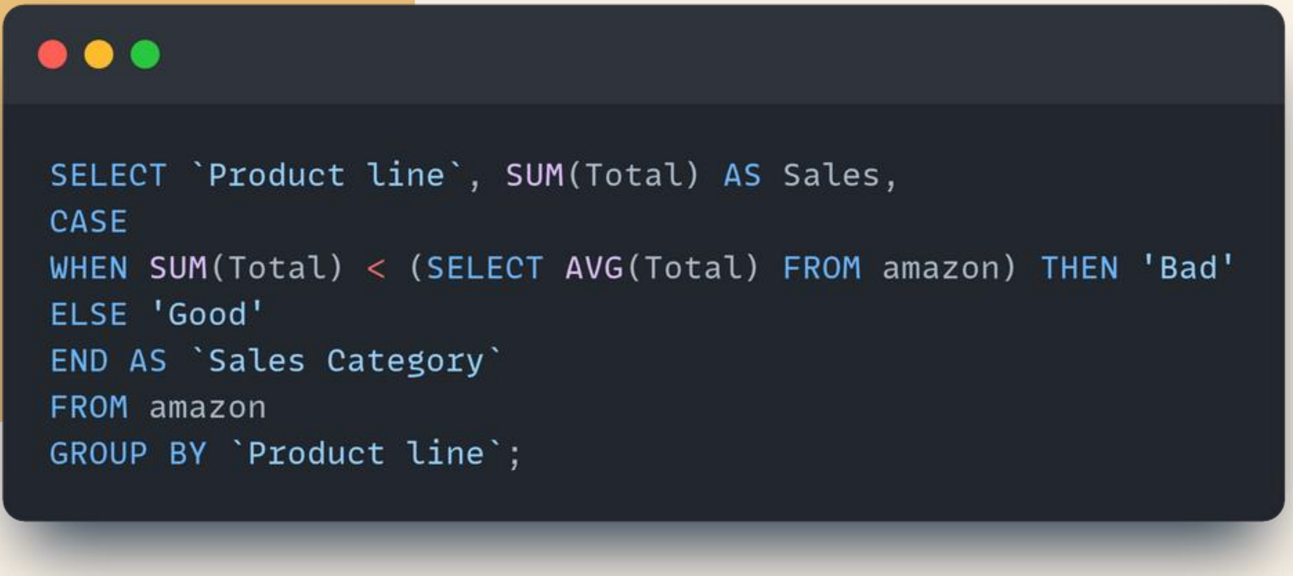


Product line	Highest Value Added Tax
Food and beverages	2673.5639999999994

The product line that incurred the highest Value Added Tax is "Food and beverages".

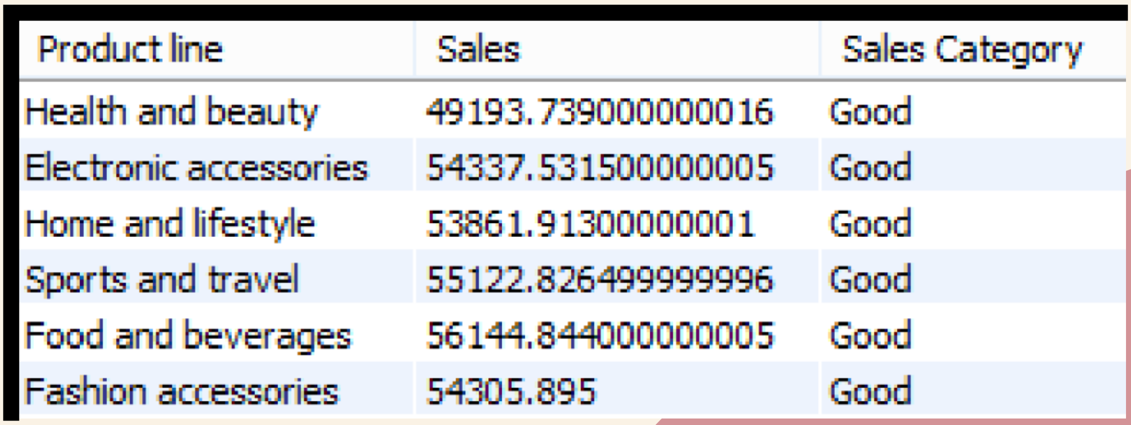
For each product line, add a column indicating "Good" if its sales are above average, otherwise "Bad."

INPUT



```
SELECT `Product line`, SUM(Total) AS Sales,
CASE
WHEN SUM(Total) < (SELECT AVG(Total) FROM amazon) THEN 'Bad'
ELSE 'Good'
END AS `Sales Category`
FROM amazon
GROUP BY `Product line`;
```

OUTPUT

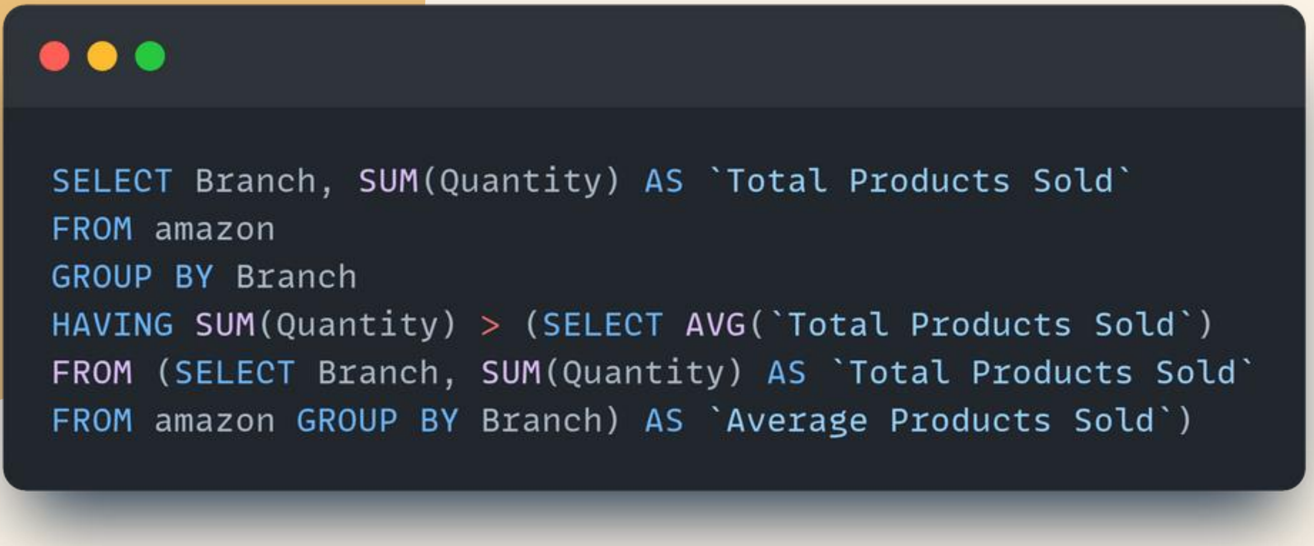


Product line	Sales	Sales Category
Health and beauty	49193.739000000016	Good
Electronic accessories	54337.531500000005	Good
Home and lifestyle	53861.91300000001	Good
Sports and travel	55122.826499999996	Good
Food and beverages	56144.844000000005	Good
Fashion accessories	54305.895	Good

All the product lines are in the good category as they all have sales above average.

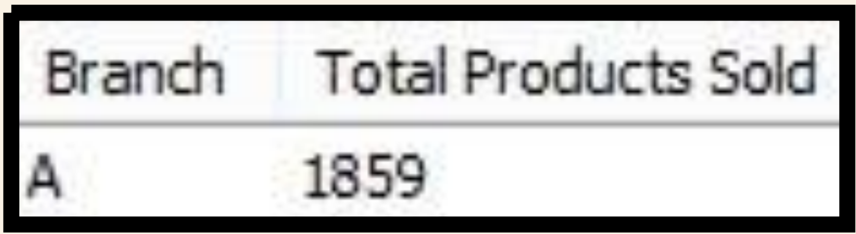
Identify the branch that exceeded the average number of products sold.

INPUT



```
SELECT Branch, SUM(Quantity) AS `Total Products Sold`
FROM amazon
GROUP BY Branch
HAVING SUM(Quantity) > (SELECT AVG(`Total Products Sold`)
FROM (SELECT Branch, SUM(Quantity) AS `Total Products Sold`
FROM amazon GROUP BY Branch) AS `Average Products Sold`)
```

OUTPUT

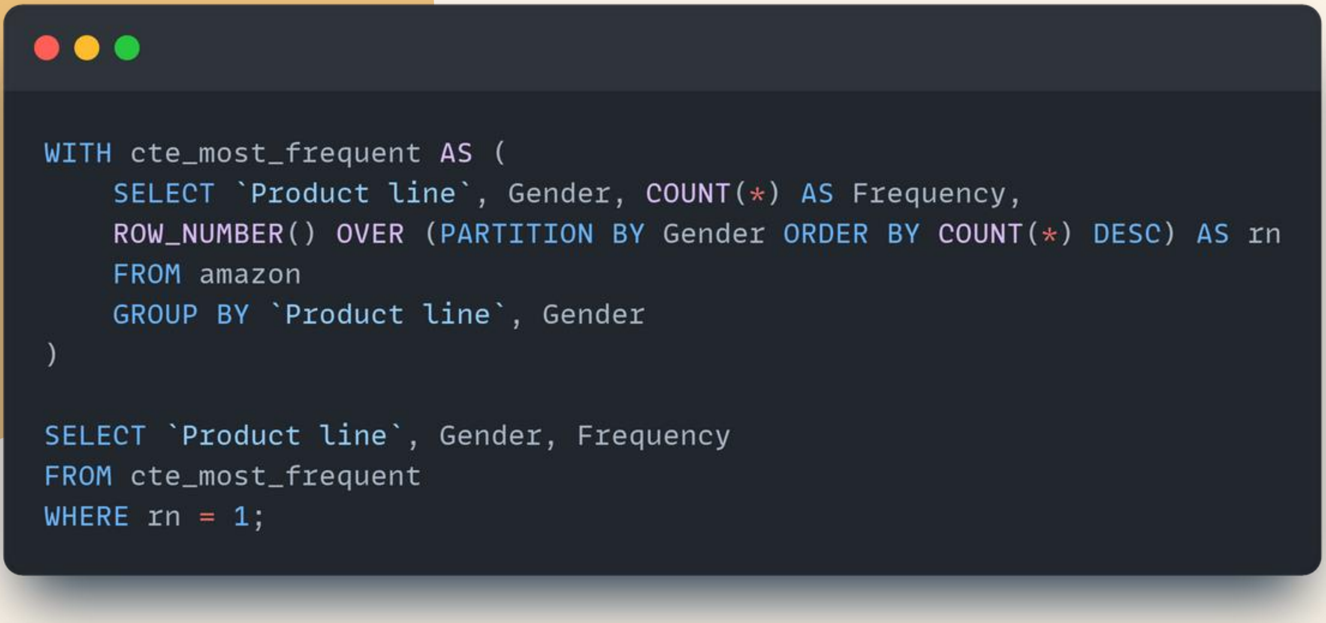


Branch	Total Products Sold
A	1859

The branch that exceeded the average number of products sold is Branch A with the total number of products sold are 1859 units.

Which product line is most frequently associated with each gender?

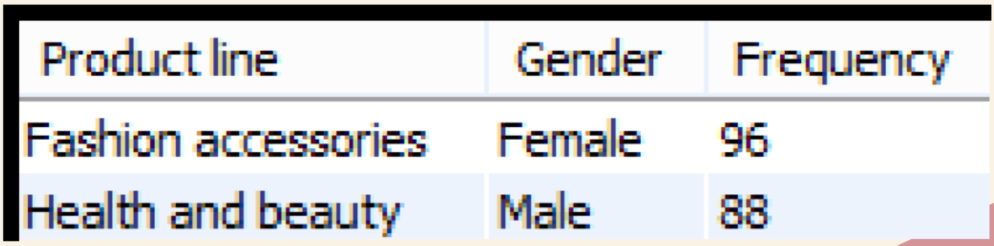
INPUT



```
WITH cte_most_frequent AS (
    SELECT `Product line`, Gender, COUNT(*) AS Frequency,
    ROW_NUMBER() OVER (PARTITION BY Gender ORDER BY COUNT(*) DESC) AS rn
    FROM amazon
    GROUP BY `Product line`, Gender
)

SELECT `Product line`, Gender, Frequency
FROM cte_most_frequent
WHERE rn = 1;
```

OUTPUT

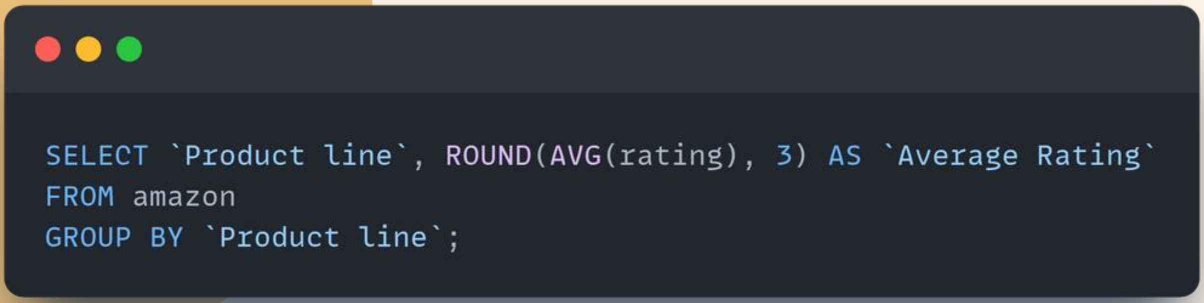


Product line	Gender	Frequency
Fashion accessories	Female	96
Health and beauty	Male	88

The product line most frequently associated with each gender is as follows: Female: Fashion accessories
Male: Health and beauty

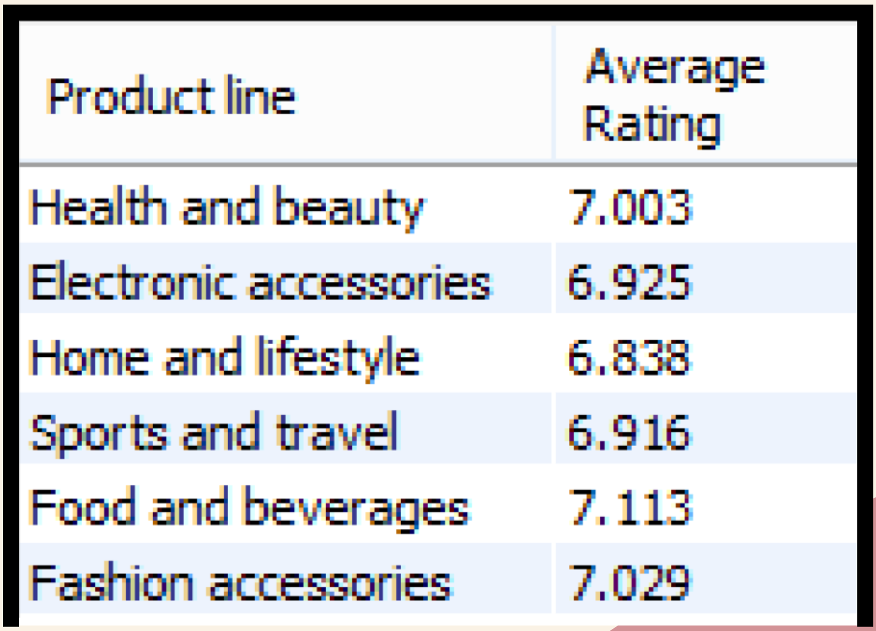
Calculate the average rating for each product line.

INPUT



```
SELECT `Product line`, ROUND(AVG(rating), 3) AS `Average Rating`
FROM amazon
GROUP BY `Product line`;
```

OUTPUT

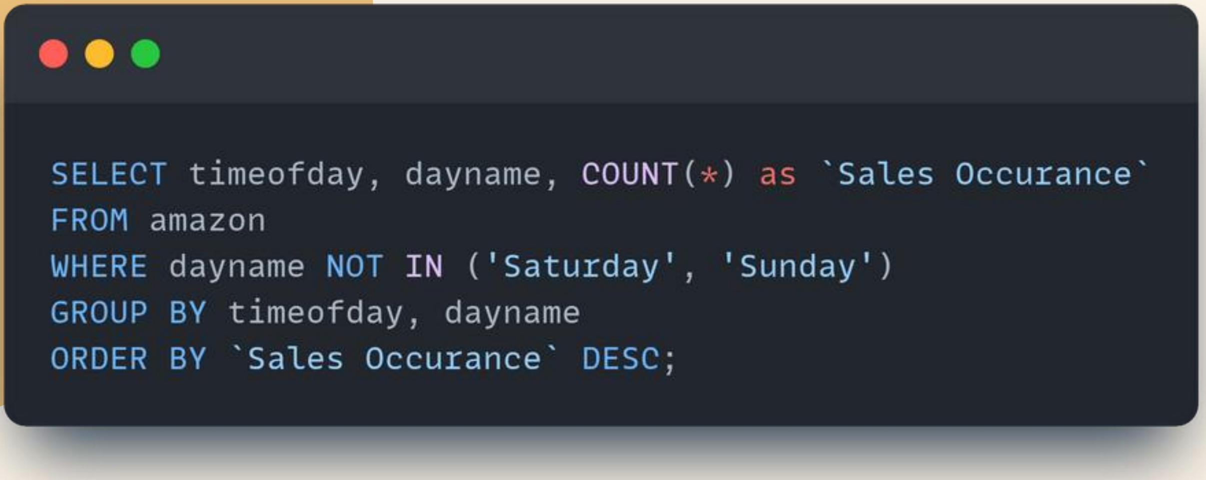


Product line	Average Rating
Health and beauty	7.003
Electronic accessories	6.925
Home and lifestyle	6.838
Sports and travel	6.916
Food and beverages	7.113
Fashion accessories	7.029

The average rating for each product line is shown in the above table.

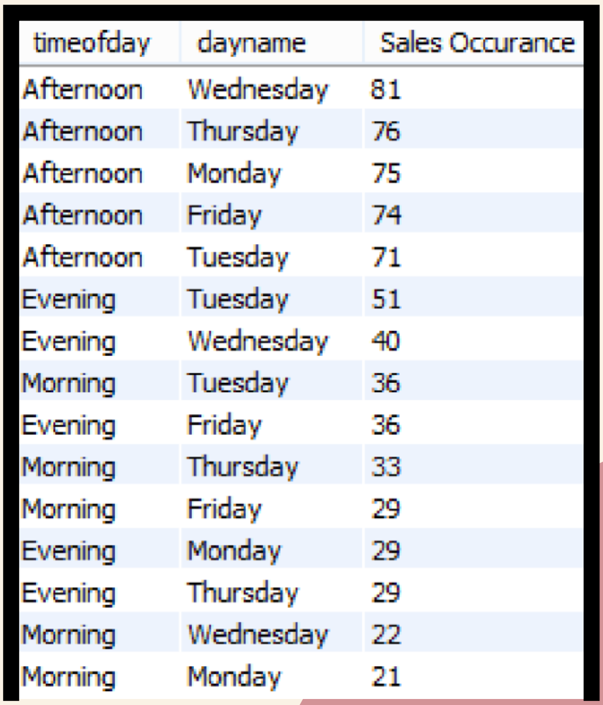
Count the sales occurrences for each time of day on every weekday.

INPUT



```
SELECT timeofday, dayname, COUNT(*) as `Sales Occurance`
FROM amazon
WHERE dayname NOT IN ('Saturday', 'Sunday')
GROUP BY timeofday, dayname
ORDER BY `Sales Occurance` DESC;
```

OUTPUT

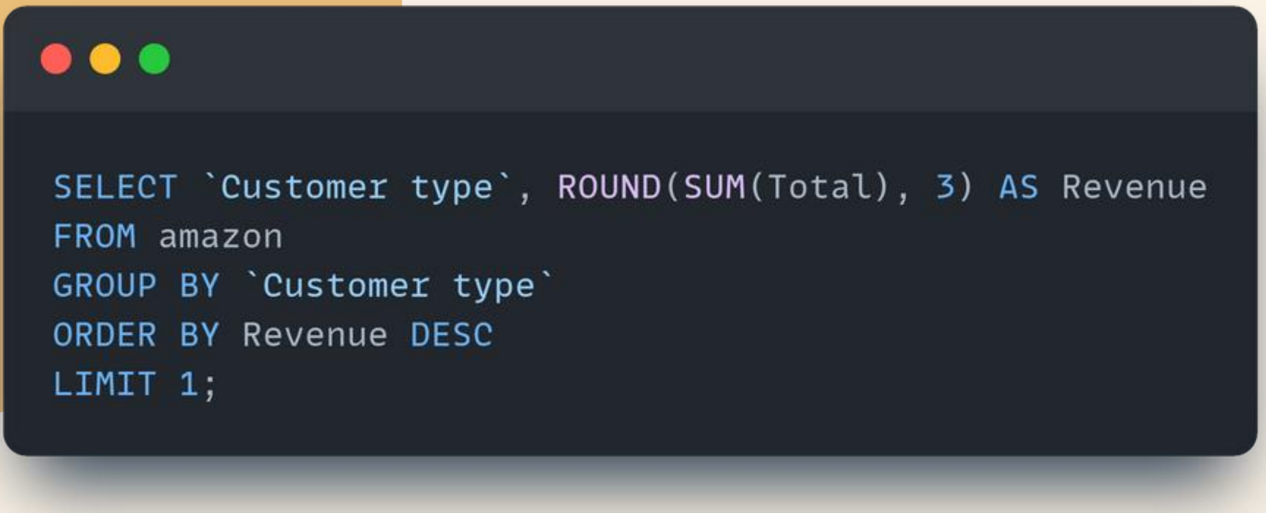


timeofday	dayname	Sales Occurance
Afternoon	Wednesday	81
Afternoon	Thursday	76
Afternoon	Monday	75
Afternoon	Friday	74
Afternoon	Tuesday	71
Evening	Tuesday	51
Evening	Wednesday	40
Morning	Tuesday	36
Evening	Friday	36
Morning	Thursday	33
Morning	Friday	29
Evening	Monday	29
Evening	Thursday	29
Morning	Wednesday	22
Morning	Monday	21

This table indicates the number of sales occurrences during the Morning, Afternoon and
Evening for each day of the week

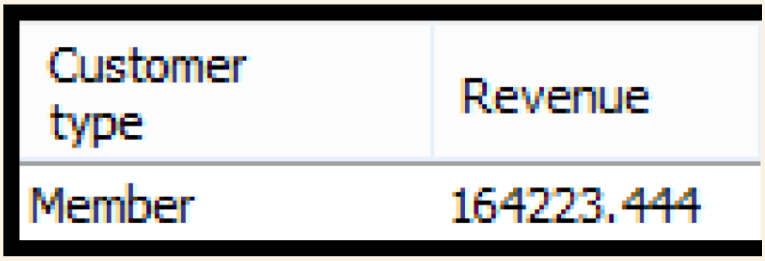
Identify the customer type contributing the highest revenue.

INPUT



```sql
SELECT `Customer type`, ROUND(SUM(Total), 3) AS Revenue
FROM amazon
GROUP BY `Customer type`
ORDER BY Revenue DESC
LIMIT 1;
```

OUTPUT

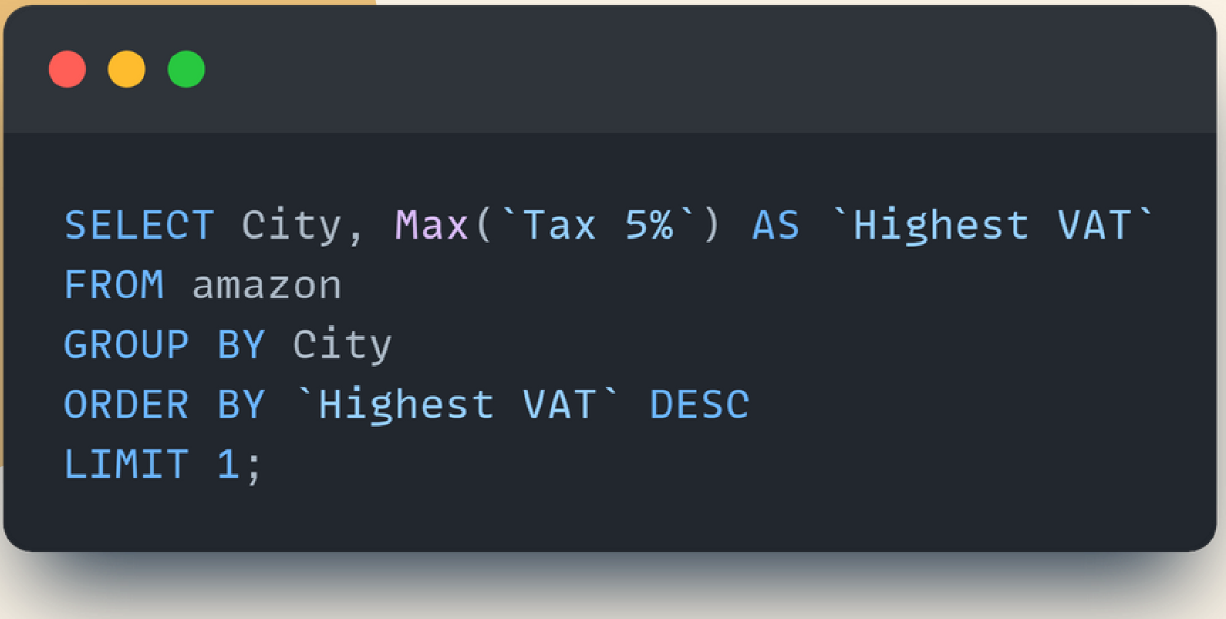


Customer type	Revenue
Member	164223.444

The customer type contributing the highest revenue is 'Member' with a total revenue of 164,223.44.

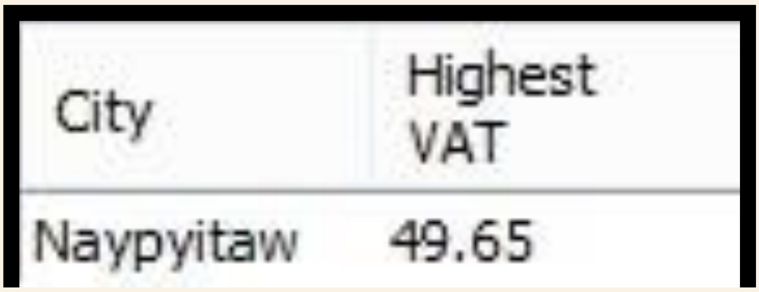
Determine the city with the highest VAT percentage.

INPUT



```
SELECT City, Max(`Tax 5%`) AS `Highest VAT`
FROM amazon
GROUP BY City
ORDER BY `Highest VAT` DESC
LIMIT 1;
```

OUTPUT

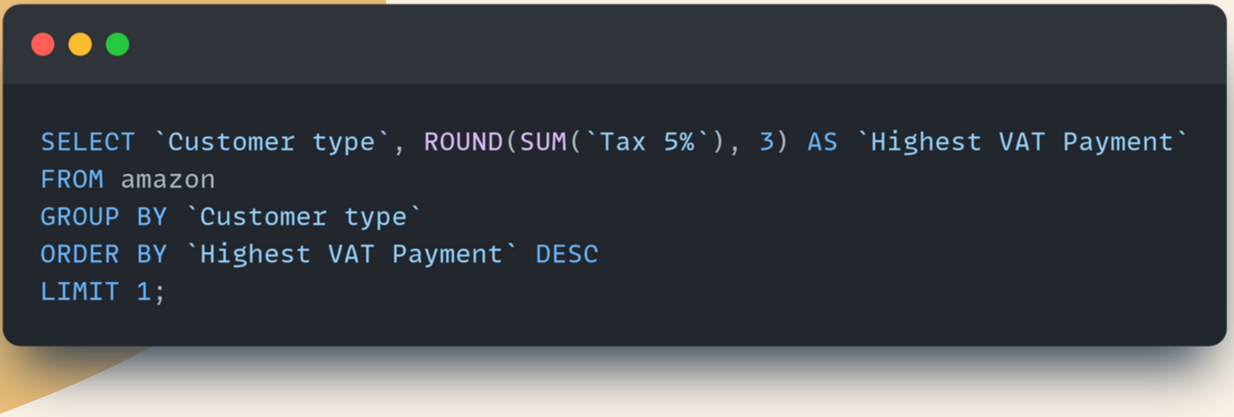


City	Highest VAT
Naypyitaw	49.65

The city with the highest VAT percentage is Naypyitaw.

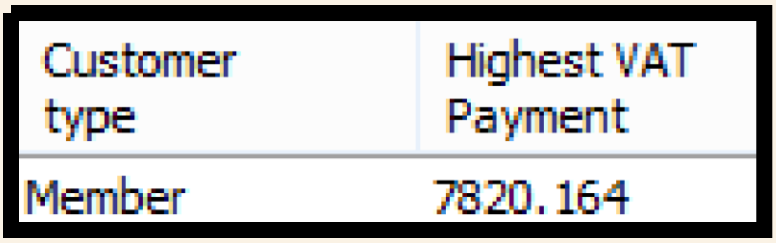
Identify the customer type with the highest VAT payments.

INPUT



```
SELECT `Customer type`, ROUND(SUM(`Tax 5%`), 3) AS `Highest VAT Payment`
FROM amazon
GROUP BY `Customer type`
ORDER BY `Highest VAT Payment` DESC
LIMIT 1;
```

OUTPUT

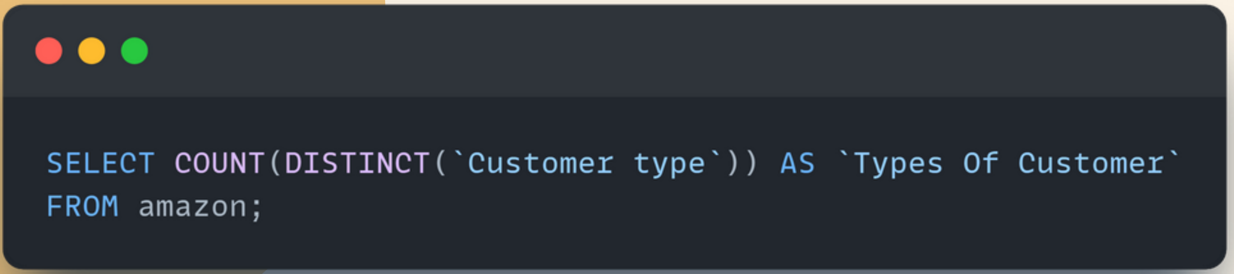


Customer type	Highest VAT Payment
Member	7820.164

The customer type with the highest VAT payments is 'Member', with a total VAT collected amounting to $7820.164.

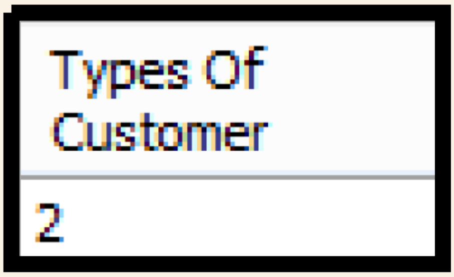
What is the count of distinct customer types in the dataset?

INPUT



```
SELECT COUNT(DISTINCT(`Customer type`)) AS `Types Of Customer`
FROM amazon;
```

OUTPUT

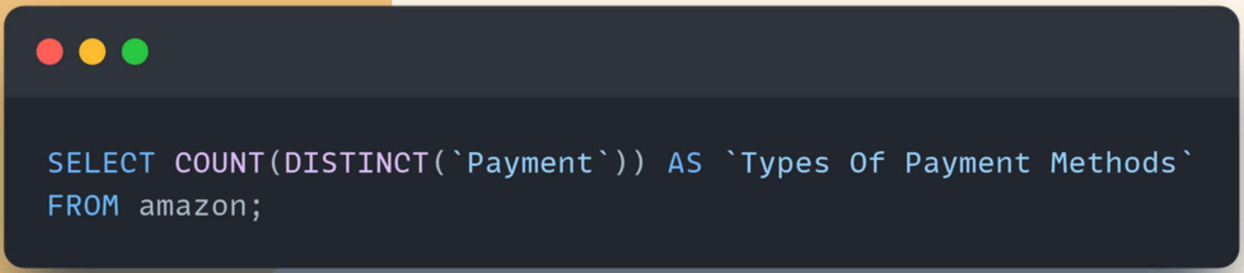


Types Of Customer
2

The count of distinct customer types in the dataset is 2.

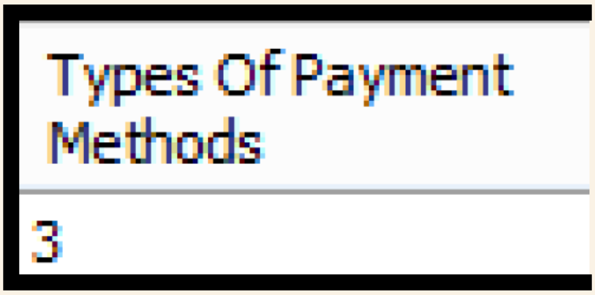
What is the count of distinct payment methods in the dataset?

INPUT



```
SELECT COUNT(DISTINCT(`Payment`)) AS `Types Of Payment Methods`
FROM amazon;
```

OUTPUT

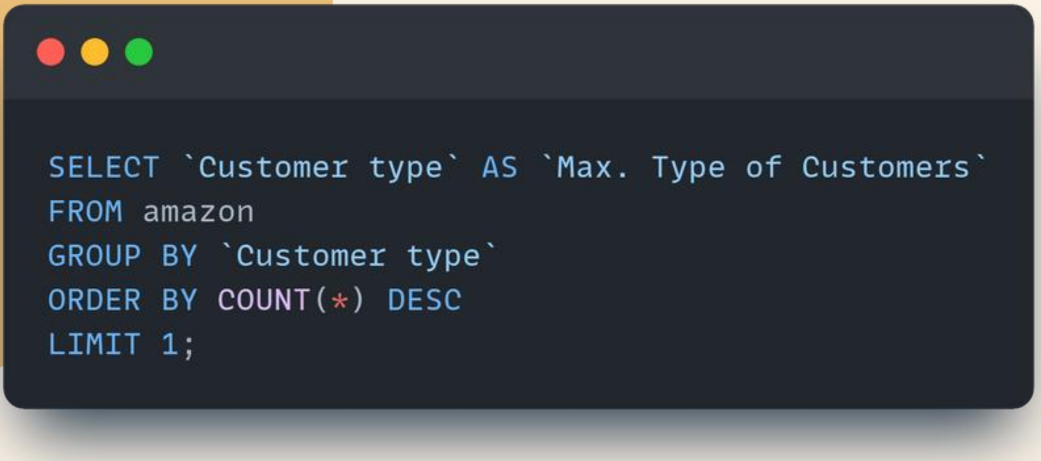


Types Of Payment Methods
3

The count of distinct payment methods in the dataset is 3.

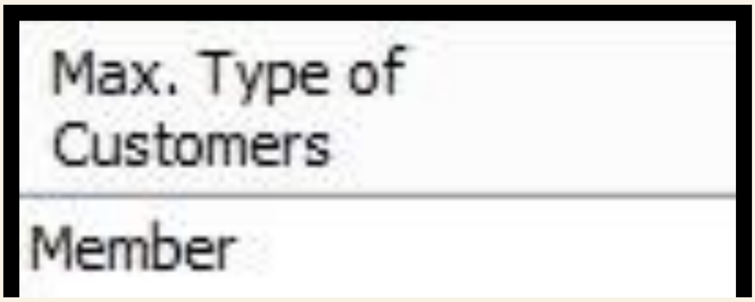
Which customer type occurs most frequently?

INPUT



```
SELECT `Customer type` AS `Max. Type of Customers`
FROM amazon
GROUP BY `Customer type`
ORDER BY COUNT(*) DESC
LIMIT 1;
```

OUTPUT

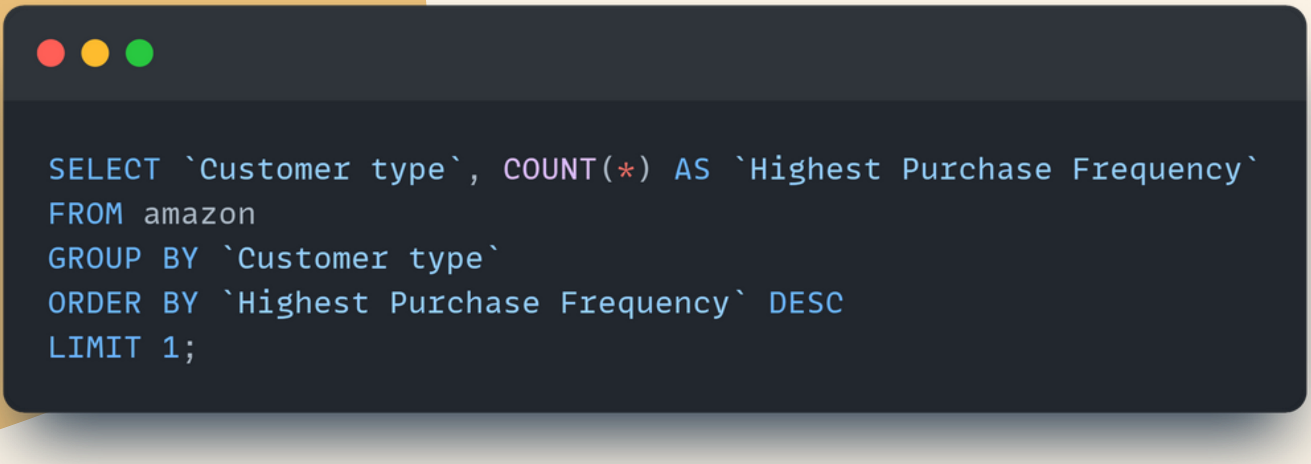


Max. Type of Customers
Member

The customer type that occurs most frequently in the dataset is "Member".

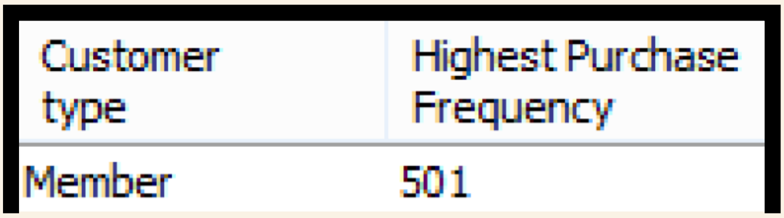
Identify the customer type with the highest purchase frequency.

INPUT



```
SELECT `Customer type`, COUNT(*) AS `Highest Purchase Frequency`
FROM amazon
GROUP BY `Customer type`
ORDER BY `Highest Purchase Frequency` DESC
LIMIT 1;
```

OUTPUT

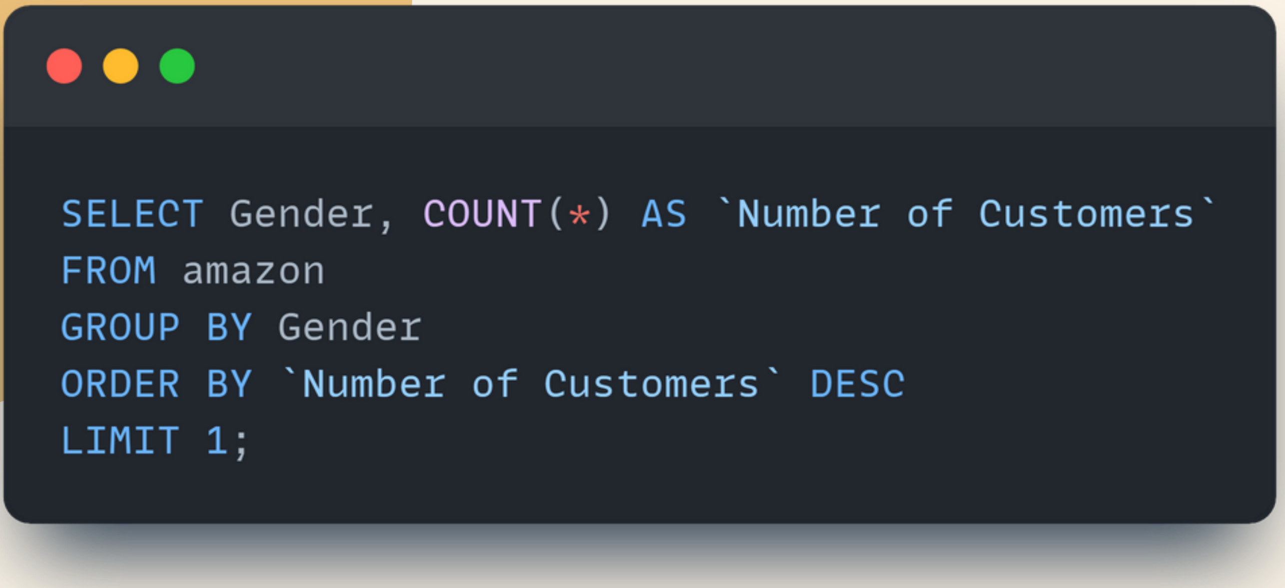


Customer type	Highest Purchase Frequency
Member	501

The customer type with the highest purchase frequency in the dataset is "Member" with 501 occurrences.

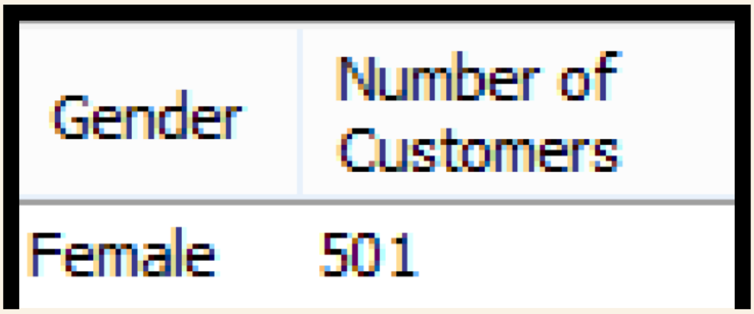
Determine the predominant gender among customers.

INPUT



```
SELECT Gender, COUNT(*) AS `Number of Customers`
FROM amazon
GROUP BY Gender
ORDER BY `Number of Customers` DESC
LIMIT 1;
```

OUTPUT

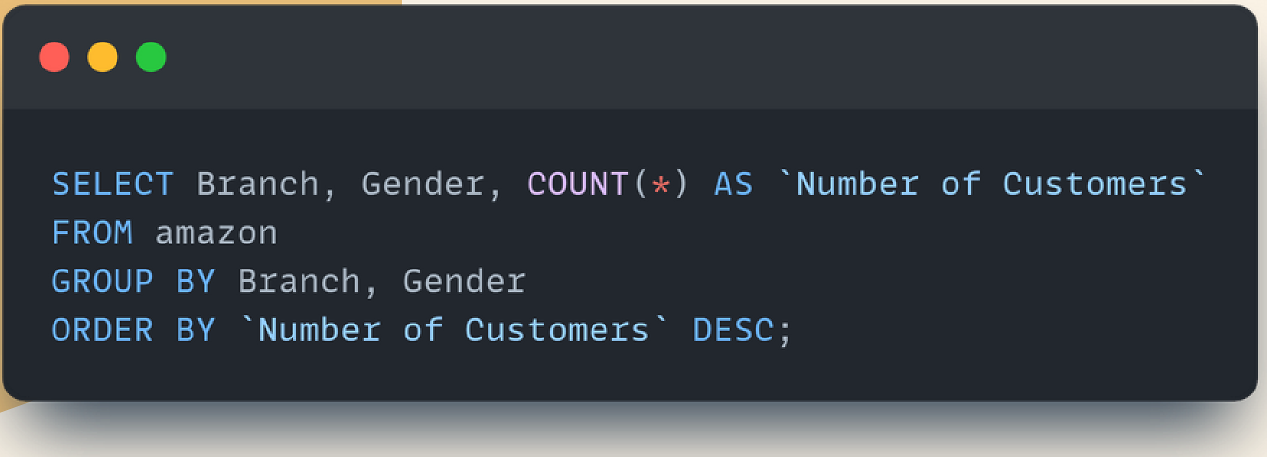


Gender	Number of Customers
Female	501

The predominant gender among customers is Female with a total count of 501 occurrences.

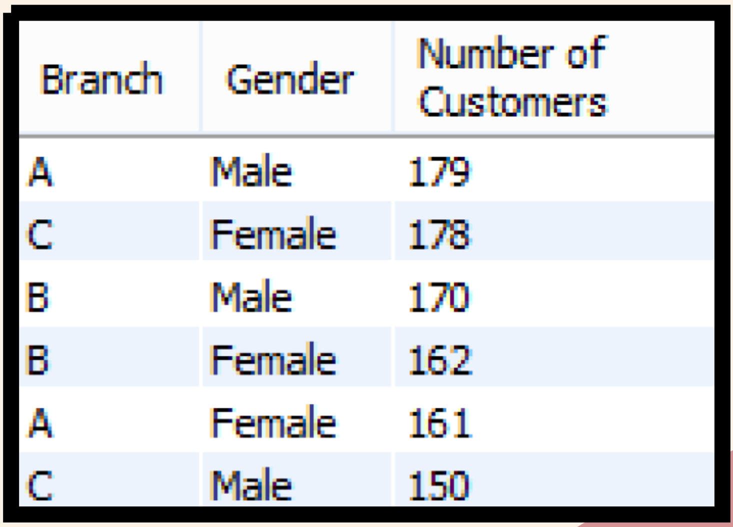
Examine the distribution of genders within each branch.

INPUT



```
SELECT Branch, Gender, COUNT(*) AS `Number of Customers`
FROM amazon
GROUP BY Branch, Gender
ORDER BY `Number of Customers` DESC;
```

OUTPUT

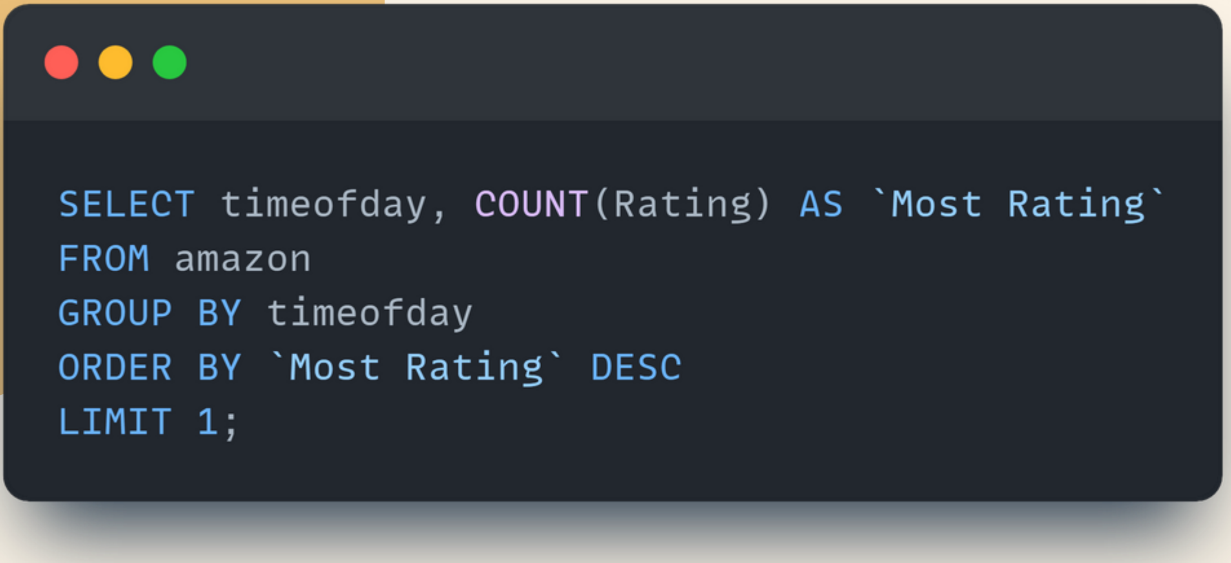


Branch	Gender	Number of Customers
A	Male	179
C	Female	178
B	Male	170
B	Female	162
A	Female	161
C	Male	150

Branch A: 161 Female customers and 179 Male customers.
Branch B: 162 Female customers and 170 Male customers.
Branch C: 178 Female customers and 150 Male customers.

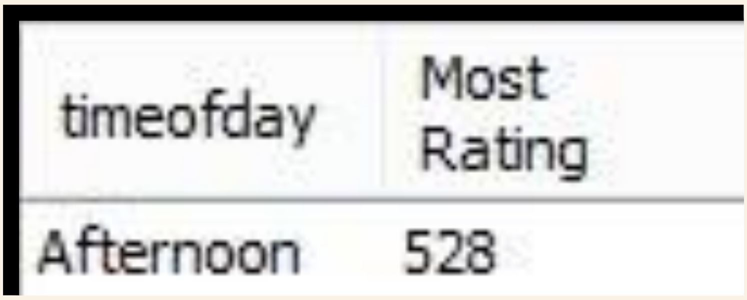
Identify the time of day when customers provide the most ratings.

INPUT



```
SELECT timeofday, COUNT(Rating) AS `Most Rating`
FROM amazon
GROUP BY timeofday
ORDER BY `Most Rating` DESC
LIMIT 1;
```

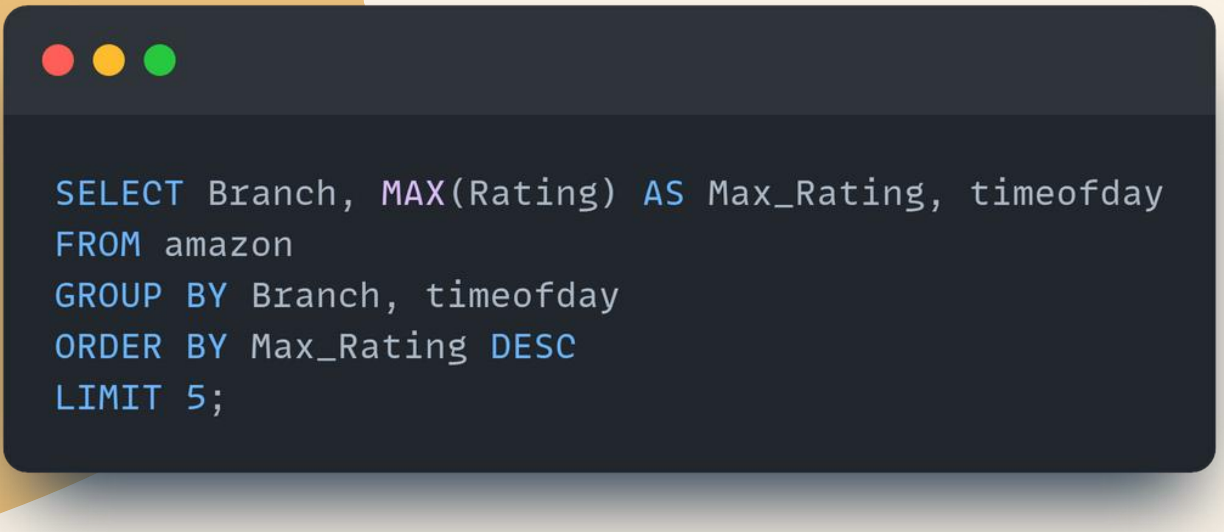
OUTPUT



timeofday	Most Rating
Afternoon	528

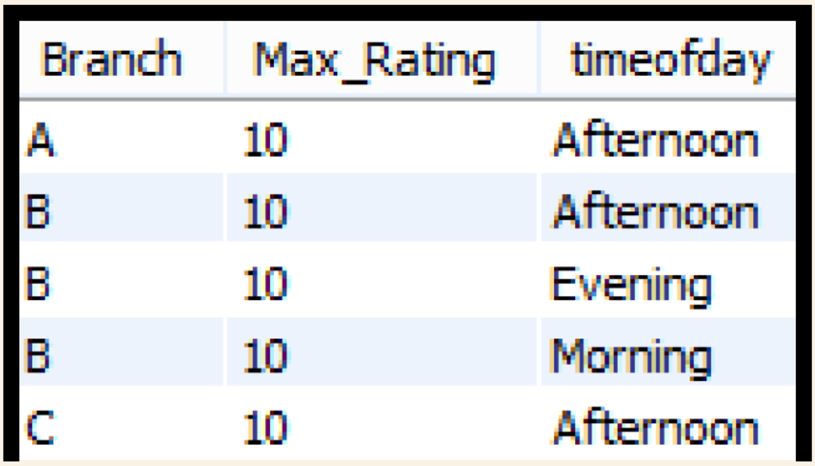
There are 528 ratings provided by customers during the afternoon hours.

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



```
SELECT Branch, MAX(Rating) AS Max_Rating, timeofday
FROM amazon
GROUP BY Branch, timeofday
ORDER BY Max_Rating DESC
LIMIT 5;
```

INPUT OUTPUT

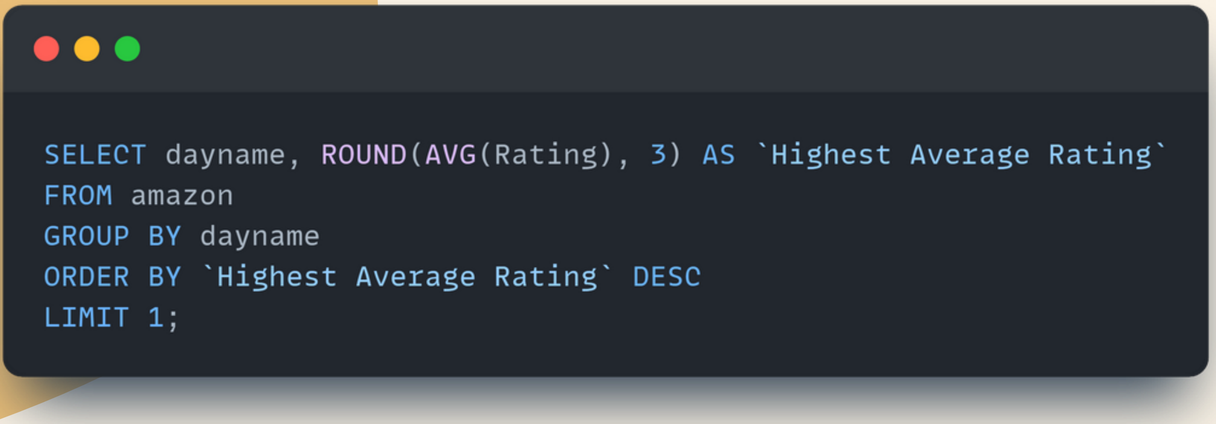


Branch	Max_Rating	timeofday
A	10	Afternoon
B	10	Afternoon
B	10	Evening
B	10	Morning
C	10	Afternoon

For Branch A, the highest rating occurs in Afternoon.
For Branch B, the highest rating occurs in Morning, Afternoon and Evening.
For Branch C, the highest rating occurs in Afternoon.

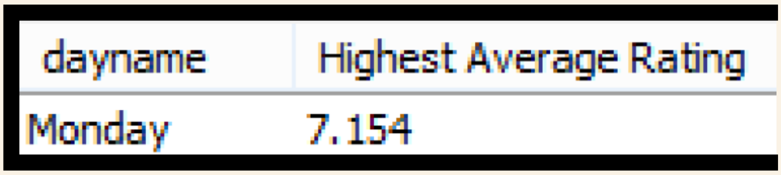
Identify the day of the week with the highest average ratings.

INPUT



```
SELECT dayname, ROUND(AVG(Rating), 3) AS `Highest Average Rating`
FROM amazon
GROUP BY dayname
ORDER BY `Highest Average Rating` DESC
LIMIT 1;
```

OUTPUT

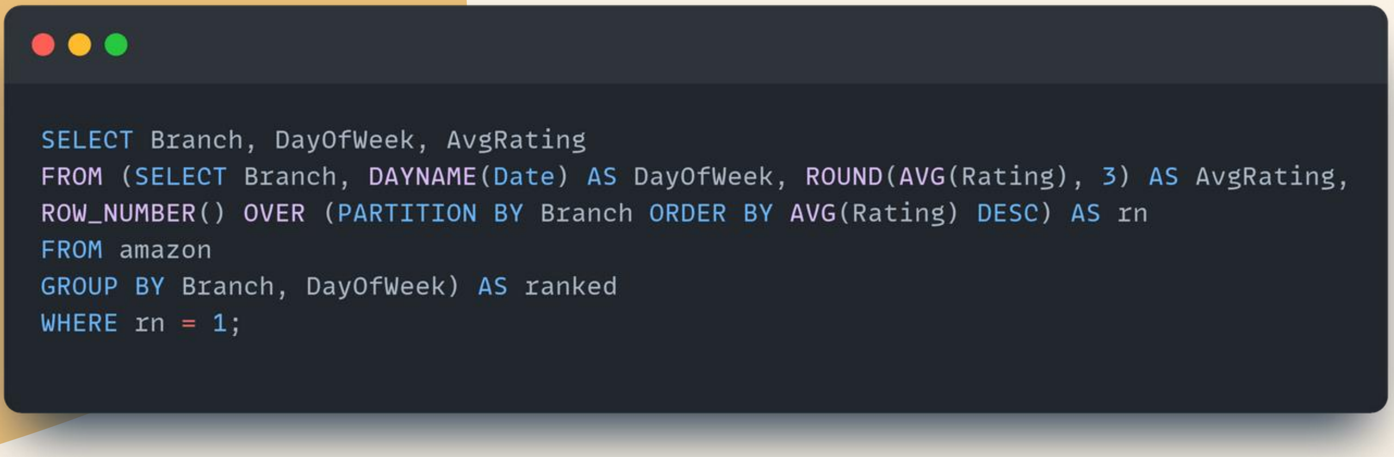


dayname	Highest Average Rating
Monday	7.154

The day of the week with the highest average rating is Monday.

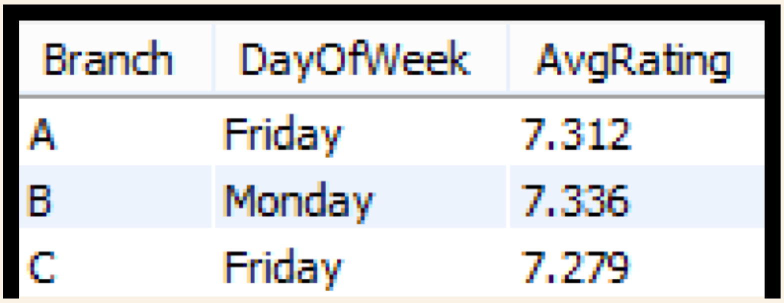
Determine the day of the week with the highest average rating for each branch.

INPUT



```
SELECT Branch, DayOfWeek, AvgRating
FROM (SELECT Branch, DAYNAME(Date) AS DayOfWeek, ROUND(AVG(Rating), 3) AS AvgRating,
ROW_NUMBER() OVER (PARTITION BY Branch ORDER BY AVG(Rating) DESC) AS rn
FROM amazon
GROUP BY Branch, DayOfWeek) AS ranked
WHERE rn = 1;
```

OUTPUT



Branch	DayOfWeek	AvgRating
A	Friday	7.312
B	Monday	7.336
C	Friday	7.279

The days of the week with the highest average ratings for each branch are as follows:

Branch A: Friday

Branch B: Monday

Branch C: Friday

KEY INSIGHTS

- The product line that is the top-performing based on total sales is "***Food and beverages***" with the highest average customer rating

7.11.

- The product line that needs improvement as it has the lowest total sales is "***Health and beauty***."
- "***Electronic accessories***" has the highest total quantity sold (***971 units***) with an average unit price of approximately ***$53.55*** and an average customer rating of ***6.92***.
- "***Health and beauty***" has the lowest total quantity sold (***854 units***) with an average unit price of approximately ***$54.85***, and an average customer rating of ***7.00***.
- ***Female customers*** who are ***members*** have the highest total sales, followed by male members. Non-member females and males have similar total sales.

RECOMMONDATIONS

- Increasing membership benefits could boost even more loyalty and sales. This can be done through personalized discounts, member-only sales events or reward points systems.
- Invest more in the marketing and stock of the 'Food and beverages' as it is the top performer. This could include themed promotions.

- Examine the non-member customers through marketing campaigns or incentive programs. • Encourage more customers to leave ratings and reviews. This can improve customer trust and increase sales.
- In categories that are lagging in sales utilize customer feedback and sales data to inform the development of new products.

Thank you!