Walmart Sales Data Analysis Python and SQL

About This project aims to explore the Walmart Sales data to understand top performing branches and products, sales trend of of different products, customer behaviour. The aims is to study how sales strategies can be improved and optimized. The dataset was obtained from the Kaggle Walmart Sales Forecasting Competition.

"In this recruiting competition, job-seekers are provided with historical sales data for 45 Walmart stores located in different regions. Each store contains many departments, and participants must project the sales for each department in each store. To add to the challenge, selected holiday markdown events are included in the dataset. These markdowns are known to affect sales, but it is challenging to predict which departments are affected and the extent of the impact."

Purposes Of The Project

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

In [1]:

```
#import necessary libraries
import mysql.connector
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import sql.run
import warnings
warnings.filterwarnings("ignore")
```

In [2]:

```
#create a connection and cursor objects
db=mysql.connector.connect(
    host="localhost",
    user="root",
    password="1996",
    database="walmartsales"

)

cursor=db.cursor()
print(db)
```

```
<mysql.connector.connection_cext.CMySQLConnection object at 0x00000201A2688220>
```

In [3]:

```
sql.run.ResultSet.pretty = None
```

In [4]:

```
%load_ext sql
%sql mysql+mysqldb://root:1996@localhost/walmartsales
```

Business Questions To Answer

Generic Questions

How many unique cities does the data have?

In [5]:

```
%%sql
SELECT COUNT(DISTINCT city) AS "Number cities"
FROM sales;
```

```
 * mysql+mysqldb://root:***@localhost/walmartsales
1 rows affected.
```

Out[5]:

Number cities
3

In which city is each branch?

In [6]:

```
%%sql
SELECT
DISTINCT city,
branch
FROM sales;
```

```
 * mysql+mysqldb://root:***@localhost/walmartsales
3 rows affected.
```

Out[6]:

city	branch
Yangon	A
Naypyitaw	C
Mandalay	B

PRODUCTS QUESTIONS

How many unique product lines does the data have?

In [7]:

```
%%sql
SELECT
COUNT(DISTINCT product_line) AS "Number of product lines"
FROM sales;
```

```
 * mysql+mysqldb://root:***@localhost/walmartsales
1 rows affected.
```

Out[7]:

Number of product lines
6

What is the most common payment method?

In [8]:

```
%%sql
SELECT
payment,
COUNT(payment) AS cnt
```

```
FROM sales
GROUP BY payment
ORDER BY cnt DESC;
```

```
 * mysql+mysqldb://root:***@localhost/walmartsales
3 rows affected.
```

Out[8]:

payment	cnt
Cash	344
Ewallet	342
Credit card	309

cash is the most common payment method

What is the most selling product line?

In [9]:

```
%%sql
SELECT
product_line,
COUNT(product_line) as cnt
FROM sales
GROUP BY product_line
ORDER BY cnt DESC;
```

```
 * mysql+mysqldb://root:***@localhost/walmartsales
6 rows affected.
```

Out[9]:

product_line	cnt
Fashion accessories	178
Food and beverages	174
Electronic accessories	169
Sports and travel	163
Home and lifestyle	160
Health and beauty	151

Fashion accessories is the most selling product line

what is the total revenue by month?

In [10]:

```
%%sql
SELECT
month_name AS months,
SUM(total) as revenue
FROM sales
GROUP BY months
ORDER BY revenue DESC  ;
```

```
 * mysql+mysqldb://root:***@localhost/walmartsales
3 rows affected.
```

Out[10]:

months	revenue
January	116291.8680

months	revenue
March	108867.1500
February	95727.3765

What month had the largest COGS?

In [11]:

```
%%sql
SELECT
month_name AS months,
SUM(cogs) AS cogs
FROM sales
GROUP BY months
ORDER BY cogs DESC;
```

```
 * mysql+mysqldb://root:***@localhost/walmartsales
3 rows affected.
```

Out[11]:

months	cogs
January	110754.16
March	103683.00
February	91168.93

January is the month had the largest COGS

What product line had the largest revenue?

In [12]:

```
%%sql
SELECT
product_line,
SUM(total) AS revenue
FROM sales
GROUP BY  product_line
ORDER BY revenue DESC;
```

```
 * mysql+mysqldb://root:***@localhost/walmartsales
6 rows affected.
```

Out[12]:

product_line	revenue
Food and beverages	56144.8440
Fashion accessories	54305.8950
Sports and travel	53936.1270
Home and lifestyle	53861.9130
Electronic accessories	53783.2365
Health and beauty	48854.3790

Food and beverages had the largest revenue

What is the city with the largest revenue?

In [13]:

```
%%sql
SELECT
```

```
city,
SUM(total) as revenue
FROM sales
GROUP BY city
ORDER BY revenue DESC;
```

```
 * mysql+mysqldb://root:***@localhost/walmartsales
3 rows affected.
```

Out[13]:

city	revenue
Naypyitaw	110490.7755
Yangon	105861.0105
Mandalay	104534.6085

Naypyitaw is the city that had the largest revenue

What product line had the largest VAT?

In [14]:

```
%%sql
SELECT
product_line,
AVG(tax_pct) AS VAT
FROM sales
GROUP BY product_line
ORDER BY VAT DESC;
```

```
 * mysql+mysqldb://root:***@localhost/walmartsales
6 rows affected.
```

Out[14]:

product_line	VAT
Home and lifestyle	16.03033124
Sports and travel	15.75697549
Health and beauty	15.40661591
Food and beverages	15.36531029
Electronic accessories	15.15447632
Fashion accessories	14.52806181

Home and lifestyle is the productline that had the largest VAT

Fetch each product line and add a column to those product line showing "Good", "Bad". Good if its greater than average sales

In [15]:

```
%%sql
SELECT
  product_line,SUM(total) AS total_sales,CASE
    WHEN SUM(total) > (
      SELECT AVG(line_sales)
      FROM (
        SELECT SUM(total) AS line_sales
        FROM sales
        GROUP BY product_line
      ) AS avg_subquery
    )
    THEN 'Good'
```

```
    ELSE 'Bad'
END AS rating
FROM sales
GROUP BY product_line;
```

```
 * mysql+mysqldb://root:***@localhost/walmartsales
6 rows affected.
```

Out[15]:

product_line	total_sales	rating
Food and beverages	56144.8440	Good
Health and beauty	48854.3790	Bad
Sports and travel	53936.1270	Good
Fashion accessories	54305.8950	Good
Home and lifestyle	53861.9130	Good
Electronic accessories	53783.2365	Good

Which branch sold more products than average product sold?

In [16]:

```
%%sql
SELECT
branch,
SUM(quantity) AS qty
FROM sales
GROUP BY branch
HAVING SUM(quantity) > (SELECT AVG(quantity) FROM sales);
```

```
 * mysql+mysqldb://root:***@localhost/walmartsales
3 rows affected.
```

Out[16]:

branch	qty
A	1849
C	1828
B	1795

Branch A sold more products than average products sold

What is the most common product line by gender?

In [17]:

```
%%sql
SELECT
gender,
product_line,
COUNT(gender) AS cnt
FROM sales
GROUP BY gender, product_line
ORDER BY cnt DESC;
```

```
 * mysql+mysqldb://root:***@localhost/walmartsales
12 rows affected.
```

Out[17]:

gender	product_line	cnt
Female	Fashion accessories	96
Female	Food and beverages	90

gender	product_line	cnt
Female	Food and beverages	90
Male	Health and beauty	88
Female	Sports and travel	86
Male	Electronic accessories	86
Male	Food and beverages	84
Female	Electronic accessories	83
Male	Fashion accessories	82
Male	Home and lifestyle	81
Female	Home and lifestyle	79
Male	Sports and travel	77
Female	Health and beauty	63

In females fashion accessories are the most common productline.In males health and beauty is the most common productline

What is the average rating of each product line?

In [18]:

```
%%sql
SELECT
product_line,
ROUND(AVG(rating), 2) AS AVG_rating
FROM sales
GROUP BY product_line
ORDER BY AVG_rating DESC;
```

```
 * mysql+mysqldb://root:***@localhost/walmartsales
6 rows affected.
```

Out[18]:

product_line	AVG_rating
Food and beverages	7.11
Fashion accessories	7.03
Health and beauty	6.98
Electronic accessories	6.91
Sports and travel	6.86
Home and lifestyle	6.84

SALES QUESTIONS

Number of sales made in each time of the day per weekday

In [19]:

```
%%sql
SELECT
time_of_day,
COUNT(*) as total_sales
FROM sales
GROUP BY time_of_day
ORDER BY total_sales DESC;
```

```
 * mysql+mysqldb://root:***@localhost/walmartsales
3 rows affected.
```

Out[19]:

time_of_day	total_sales
Evening	429
Afternoon	376
Morning	190

Which of the customer types brings the most revenue?

In [20]:

```
%%sql
SELECT
customer_type,
SUM(total) as revenue
FROM sales
GROUP BY customer_type
ORDER BY revenue DESC;
```

```
 * mysql+mysqldb://root:***@localhost/walmartsales
2 rows affected.
```

Out[20]:

customer_type	revenue
Member	163625.1015
Normal	157261.2930

Member customers are customers that bring most revenues

Which city has the largest tax percent/ VAT (Value Added Tax)?

In [21]:

```
%%sql
SELECT
city,
AVG(tax_pct) as VAT
FROM sales
GROUP BY city
ORDER BY VAT DESC;
```

```
 * mysql+mysqldb://root:***@localhost/walmartsales
3 rows affected.
```

Out[21]:

city	VAT
Naypyitaw	16.0901085
Mandalay	15.13020824
Yangon	14.87020798

Naypyitaw has the largest VAT

Which customer type pays the most in VAT?

In [22]:

```
%%sql
SELECT
customer_type,
avg(tax_pct) as VAT
```

```
FROM sales
GROUP BY customer_type
ORDER BY VAT DESC;
```

```
 * mysql+mysqldb://root:***@localhost/walmartsales
2 rows affected.
```

Out[22]:

customer_type	VAT
Member	15.61457214
Normal	15.0980504

Member cuatomers pays more VAT

CUSTOMER QUESTIONS

How many unique customer types does the data have?

In [23]:

```
%%sql
SELECT DISTINCT
customer_type
FROM sales;
```

```
 * mysql+mysqldb://root:***@localhost/walmartsales
2 rows affected.
```

Out[23]:

customer_type
Normal
Member

the data has two unque customer types , Member and Normal

What is the most common customer type?

In [24]:

```
%%sql
SELECT
customer_type,
COUNT(*) AS cnt
FROM sales
GROUP BY customer_type
ORDER BY cnt DESC;
```

```
 * mysql+mysqldb://root:***@localhost/walmartsales
2 rows affected.
```

Out[24]:

customer_type	cnt
Member	499
Normal	496

Member customers are the most common customer type

Which customer type buys the most?

In [25]:

```
%%sql
SELECT
customer_type,
SUM(quantity) as total_quantity
FROM sales
GROUP BY customer_type
ORDER BY total_quantity DESC;
```

```
 * mysql+mysqldb://root:***@localhost/walmartsales
2 rows affected.
```

Out[25]:

customer_type	total_quantity
Member	2773
Normal	2699

Member customers uys the most

What is the gender of most of the customers?

In [26]:

```
%%sql
SELECT
gender,
COUNT(*) AS gender_cnt
FROM sales
GROUP BY gender
ORDER BY gender_cnt DESC;
```

```
 * mysql+mysqldb://root:***@localhost/walmartsales
2 rows affected.
```

Out[26]:

gender	gender_cnt
Male	498
Female	497

Males are slighty more than females

What is the gender distribution per branch?

In [27]:

```
%%sql
SELECT
gender,
COUNT(*) AS gender_cnt
FROM sales
WHERE branch = 'A'
GROUP BY gender
ORDER BY gender_cnt DESC;
```

```
 * mysql+mysqldb://root:***@localhost/walmartsales
2 rows affected.
```

Out[27]:

gender	gender_cnt
Male	179

gender	gender_cnt

Branch A has more male compared to females

In [28]:

```
%%sql
SELECT
gender,
COUNT(*) AS gender_cnt
FROM sales
WHERE branch = 'B'
GROUP BY gender
ORDER BY gender_cnt DESC;
```

```
 * mysql+mysqldb://root:***@localhost/walmartsales
2 rows affected.
```

Out[28]:

gender	gender_cnt
Male	169
Female	160

Branch B has more male compared to females

In [29]:

```
%%sql
SELECT
gender,
COUNT(*) AS gender_cnt
FROM sales
WHERE branch = 'C'
GROUP BY gender
ORDER BY gender_cnt DESC;
```

```
 * mysql+mysqldb://root:***@localhost/walmartsales
2 rows affected.
```

Out[29]:

gender	gender_cnt
Female	177
Male	150

Branch C has more females compared to males

Which time of the day do customers give most ratings?

In [30]:

```
%%sql
SELECT
time_of_day,
AVG(rating) as rating
FROM sales
GROUP BY time_of_day
ORDER BY rating DESC;
```

```
 * mysql+mysqldb://root:***@localhost/walmartsales
3 rows affected.
```

Out[30]:

time_of_day	rating

time_of_day	rating
Afternoon	7.0234
Morning	6.94474
Evening	6.90536

Customers give most rating in the afternoon

Which time of the day do customers give most ratings per branch?

In [31]:

```
%%sql
SELECT
time_of_day,
AVG(rating) as rating
FROM sales
WHERE branch = 'A'
GROUP BY time_of_day
ORDER BY rating DESC;
```

```
 * mysql+mysqldb://root:***@localhost/walmartsales
3 rows affected.
```

Out[31]:

time_of_day	rating
Afternoon	7.18889
Morning	7.00548
Evening	6.87143

For branch A customers give most rating in the afternoon

In [32]:

```
%%sql
SELECT
time_of_day,
AVG(rating) as rating
FROM sales
WHERE branch = 'B'
GROUP BY time_of_day
ORDER BY rating DESC;
```

```
 * mysql+mysqldb://root:***@localhost/walmartsales
3 rows affected.
```

Out[32]:

time_of_day	rating
Morning	6.83793
Afternoon	6.81129
Evening	6.75102

For branch B customers give most rating in the morning

In [33]:

```
%%sql
SELECT
time_of_day,
AVG(rating) as rating
FROM sales
WHERE branch = 'C'
```

```
GROUP BY time_of_day
ORDER BY rating DESC;
```

```
 * mysql+mysqldb://root:***@localhost/walmartsales
3 rows affected.
```

Out[33]:

time_of_day	rating
Evening	7.09859
Afternoon	7.06667
Morning	6.97458

For branch C customers give most rating in the evening

Which day fo the week has the best avg ratings?

In [34]:

```
%%sql
SELECT
day_name,
AVG(rating) AS rating
FROM sales
GROUP BY day_name
ORDER BY rating DESC;
```

```
 * mysql+mysqldb://root:***@localhost/walmartsales
7 rows affected.
```

Out[34]:

day_name	rating
Monday	7.13065
Friday	7.05507
Tuesday	7.00316
Sunday	6.98864
Saturday	6.90183
Thursday	6.88986
Wednesday	6.76028

Monday has the best average rating

Which day of the week has the best average ratings per branch?

In [35]:

```
%%sql
SELECT
day_name,
AVG(rating) AS rating
FROM sales
WHERE branch = 'A'
GROUP BY day_name
ORDER BY rating DESC;
```

```
 * mysql+mysqldb://root:***@localhost/walmartsales
7 rows affected.
```

Out[35]:

day_name	rating

day_name	rating
Friday	7.312
Monday	7.09792
Sunday	7.07885
Tuesday	7.05882
Thursday	6.9587
Wednesday	6.84286
Saturday	6.746

For branch A friday has the best average rating

In [36]:

```
%%sql
SELECT
day_name,
AVG(rating) AS rating
FROM sales
WHERE branch = 'B'
GROUP BY day_name
ORDER BY rating DESC;
```

```
 * mysql+mysqldb://root:***@localhost/walmartsales
7 rows affected.
```

Out[36]:

day_name	rating
Monday	7.26579
Tuesday	7.00189
Sunday	6.79706
Thursday	6.75227
Saturday	6.73667
Friday	6.69412
Wednesday	6.37959

For branch B monday has the best average rating

In [37]:

```
%%sql
SELECT
day_name,
AVG(rating) AS rating
FROM sales
WHERE branch = 'C'
GROUP BY day_name
ORDER BY rating DESC;
```

```
 * mysql+mysqldb://root:***@localhost/walmartsales
7 rows affected.
```

Out[37]:

day_name	rating
Saturday	7.22963
Friday	7.20541
Wednesday	7.064
Monday	7.03684
Sunday	7.02826

Tuesday	6.95185

day_name	rating
Thursday	6.95

For branch C saturday has the best average rating

In [ ]: