

Name: ARPIT DHOTE

UMID : UMID29062547419

Project Title	Colorado
Tools	Python, SQL, Power BI
Domain	Data Analysis

### About Dataset

This dataset contains detailed information about motor vehicle sales in Colorado, broken down by year, quarter, and county. It is commonly used for analyzing regional sales trends, seasonal demand, and market performance.

The dataset consists of 501 records with 4 columns, covering:

Time-related features (year, quarter)

Geographic information (county)

Sales data (sales – total sales in dollars)

The data has no missing values and is already structured for further analysis.

The purpose of this dataset is to practice data cleaning, exploratory data analysis, and visualization techniques. It can also be used to build predictive models for vehicle sales forecasting, regional comparisons, and market insights.

### Data Cleaning Steps

We are going to:

Treat nulls (if any appear after transformations)

Handle duplicates

Populate missing rows (e.g., ensure each county has all quarters represented)

Drop unneeded columns (if additional fields are introduced later)

Split columns (if combining year/quarter into a single time variable is required)

### Step1:ImportRequiredLibraries

```
import pandas as pd
import numpy as np
import seaborn as sns
```

### Step2:LoadtheDataset

```
import pandas as pd
```

```
# Load CSV
```

```
df = pd.read_csv("C:/Users/nihal/OneDrive/Documents/jupyter/jupyter  
notebook/unified mentor/colorado motor/colorado_motor_vehicle_sales.csv")
```

```
# Preview
```

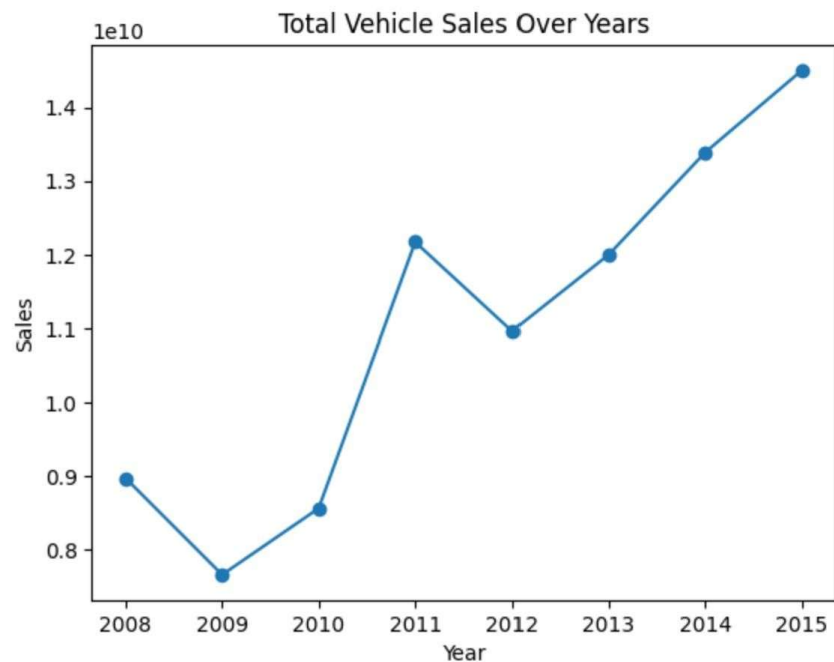
```
print(df.head())
print(df.info())
print(df.describe())
```

**Step3:Data cleaning** print("Null values before  
cleaning:\n", df.isnull().sum())

**Step 4:EDA** import matplotlib.pyplot as plt

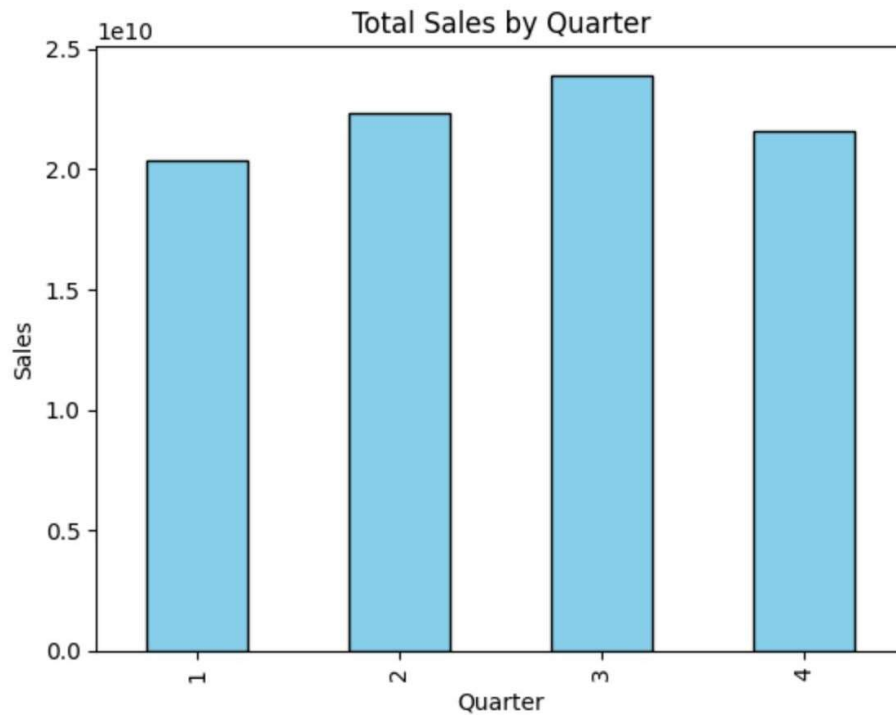
```
sales_by_year = df.groupby("year")["sales"].sum()
```

```
sales_by_year.plot(kind="line", marker="o")
plt.title("Total Vehicle Sales Over Years")
plt.xlabel("Year")
plt.ylabel("Sales")
plt.show()
```



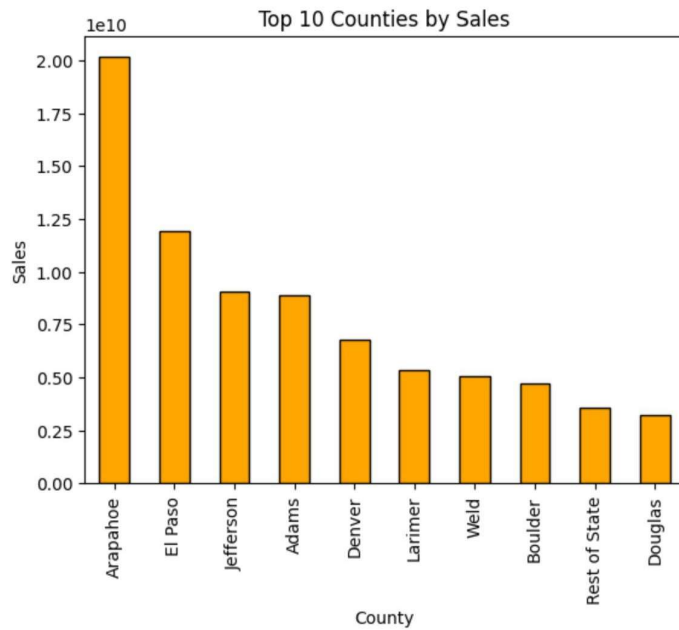
```
sales_by_quarter = df.groupby("quarter")["sales"].sum()
```

```
sales_by_quarter.plot(kind="bar", color="skyblue", edgecolor="black")  
plt.title("Total Sales by Quarter") plt.xlabel("Quarter")  
plt.ylabel("Sales") plt.show()
```



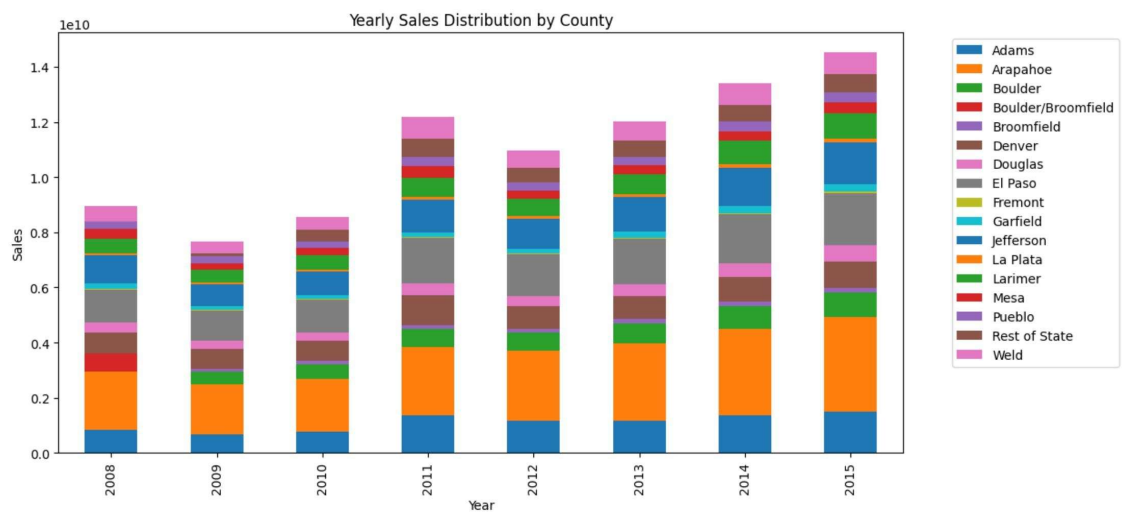
```
top_counties =  
df.groupby("county")["sales"].sum().sort_values(ascending=False).head(10)
```

```
top_counties.plot(kind="bar", color="orange", edgecolor="black") plt.title("Top  
10 Counties by Sales")  
plt.xlabel("County")  
plt.ylabel("Sales")  
plt.show()
```



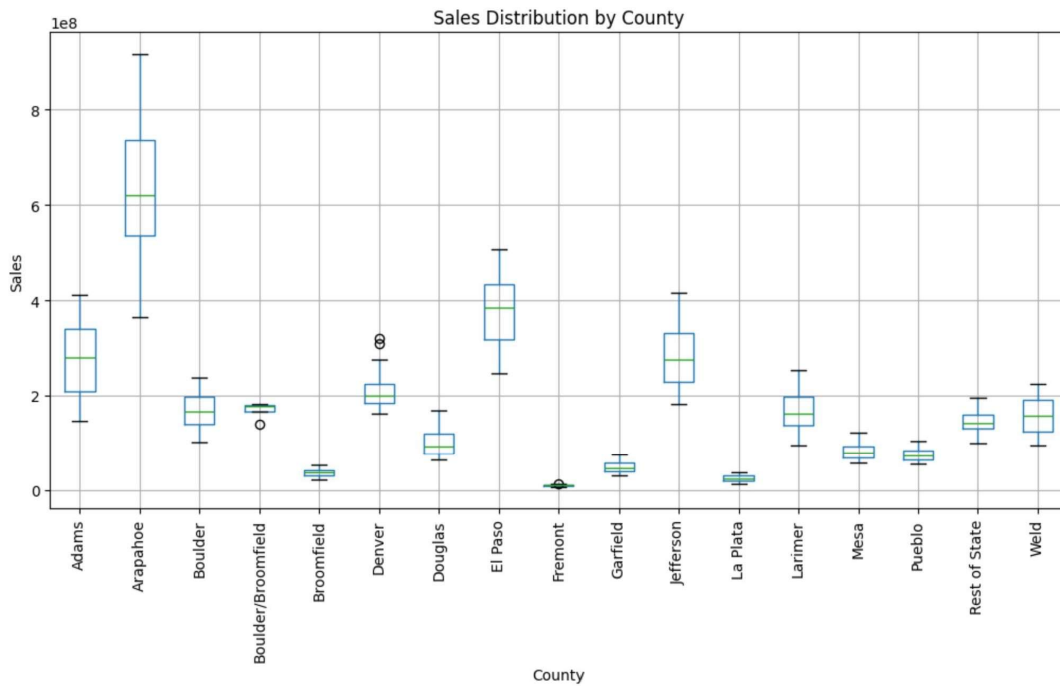
```
sales_by_county_year = df.pivot_table(values="sales", index="year", columns="county",
aggfunc="sum").fillna(0)
```

```
sales_by_county_year.plot(kind="bar", stacked=True, figsize=(12,6)) plt.title("Yearly
Sales Distribution by County")
plt.xlabel("Year")
plt.ylabel("Sales")
plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left')
plt.show()
```



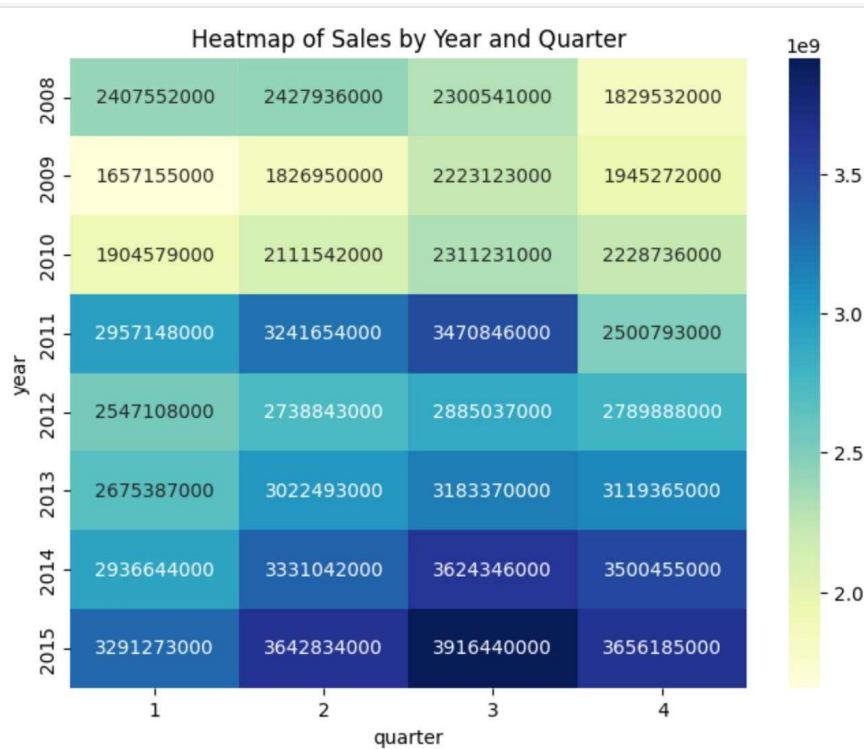
```
df.boxplot(column="sales", by="county", figsize=(12,6))
plt.title("Sales Distribution by County") plt.suptitle("")
# remove automatic title
```

```
plt.xlabel("County")
plt.ylabel("Sales")
plt.xticks(rotation=90)
plt.show()
```



```
import seaborn as sns
pivot = df.pivot_table(values="sales", index="year",
                        columns="quarter", aggfunc="sum")
```

```
plt.figure(figsize=(8,6))
sns.heatmap(pivot, annot=True, fmt=".0f", cmap="YlGnBu")
plt.title("Heatmap of Sales by Year and Quarter")
plt.show
```



## SQL Queries

```
create database colorados;
use colorados;
```

```
SELECT year, SUM(sales) AS total_sales
FROM colorado.colorado_motor_vehicle_sales
GROUP BY year
ORDER BY year;
```

Result Grid		
	year	total_sales
	2008	8965561000
	2009	7652500000
	2010	8556088000
	2011	12170441000
	2012	10960876000
	2013	12000615000
	2014	13392487000
	2015	14506732000

```
SELECT county, SUM(sales) AS total_sales
FROM colorado.colorado_motor_vehicle_sales
GROUP BY county
ORDER BY total_sales DESC
LIMIT 5;
```

Result Grid			Filter Rows:	Export:
	county	total_sales		
▶	Arapahoe	20142323000		
	El Paso	11926044000		
	Jefferson	9058407000		
	Adams	8902115000		
	Denver	6763613000		

```

SELECT year, SUM(sales) AS denver_sales
FROM colorado.colorado_motor_vehicle_sales
WHERE county = 'Denver'
GROUP BY year
ORDER BY year;

```

Result Grid			Filter Rows:	Export:	Wrap Cell
	year	denver_sales			
▶	2008	753352000			
	2009	714075000			
	2010	730723000			
	2011	1089330000			
	2012	817469000			
	2013	818590000			
	2014	896414000			
	2015	943660000			

```

SELECT quarter, SUM(sales) AS total_sales
FROM colorado.colorado_motor_vehicle_sales
GROUP BY quarter
ORDER BY total_sales DESC;

```

Result Grid			Filter Rows:	Export:
	quarter	total_sales		
▶	3	23914934000		
	2	22343294000		
	4	21570226000		
	1	20376846000		

# Power BI

