

Excel Tasks:

1. Data Exploration:

Task 1: Create a statistical summary for numerical features

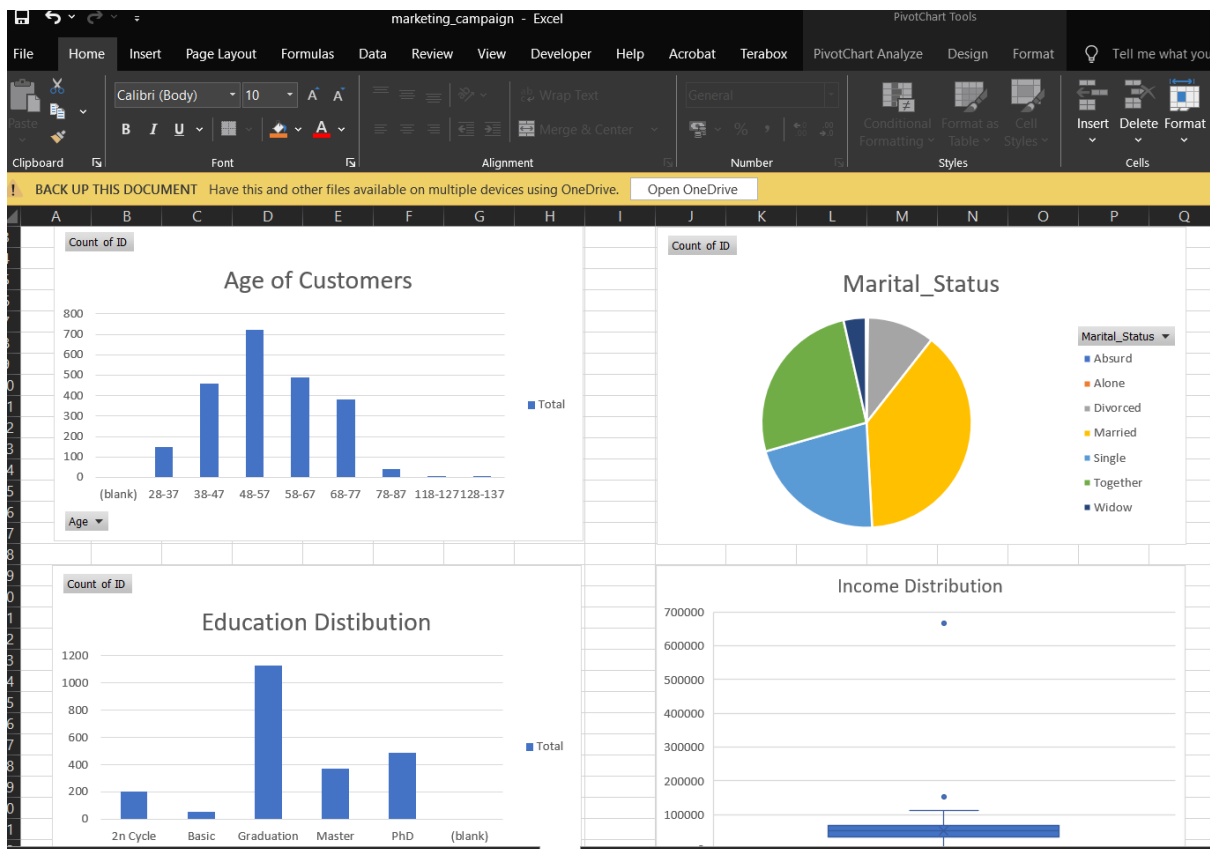
Task 2: Create a line chart for the number of enrolments by year

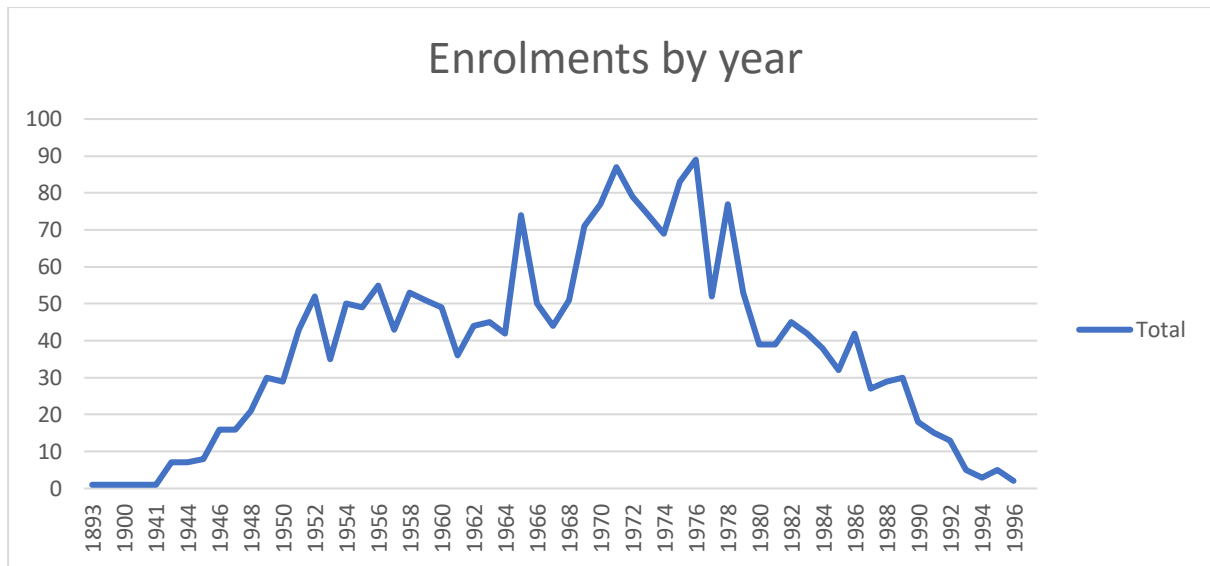
Task 3: Give a cross-tabulated count for response values against education

Task 4: Make a boxplot on income and write your observations

Task 5: Calculate the age of customers and make a histogram of that

Task 6: Visualize the response against Marital_Status





Data Loading: Create a schema named "retail_data"

Set "retail_data" as the default schema

Create tables to store the retail transaction data

Set ct_customer as the datetime field while loading the data and apply the appropriate date format

```

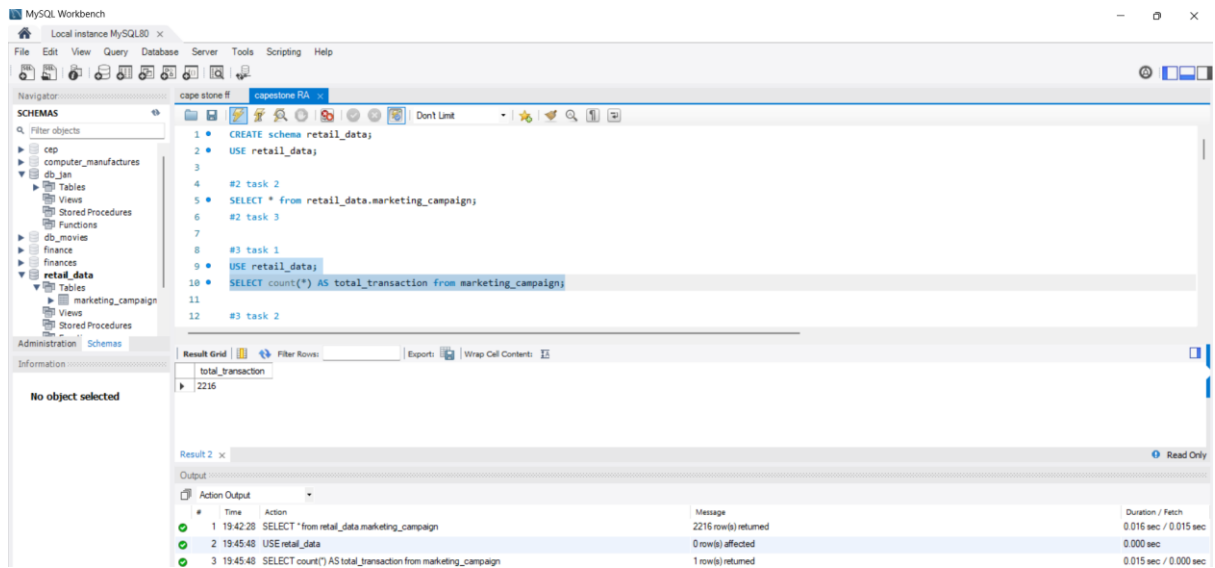
1 CREATE schema retail_data;
2 USE retail_data;
3
4 #2 task 2
5 SELECT * from retail_data.marketing_campaign;
6 #2 task 3
7
8 #3 task 1
9 USE retail_data;
10 SELECT count(*) AS total_transaction from marketing_campaign;
11
12 #3 task 2
  
```

ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	DT_Customer	Recency	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds	NumDealsPurchases	NumWebPur
5524	1957	Graduation	Single	59138	0	0	04-09-2012	58	635	88	546	172	88	88	3	8
2174	1954	Graduation	Single	46344	1	1	08-03-2014	38	11	1	6	2	1	6	2	1
4141	1965	Graduation	Together	71613	0	0	21-08-2013	26	426	49	127	111	21	42	1	8
6182	1984	Graduation	Together	26646	1	0	10-02-2014	26	11	4	20	10	3	5	2	2

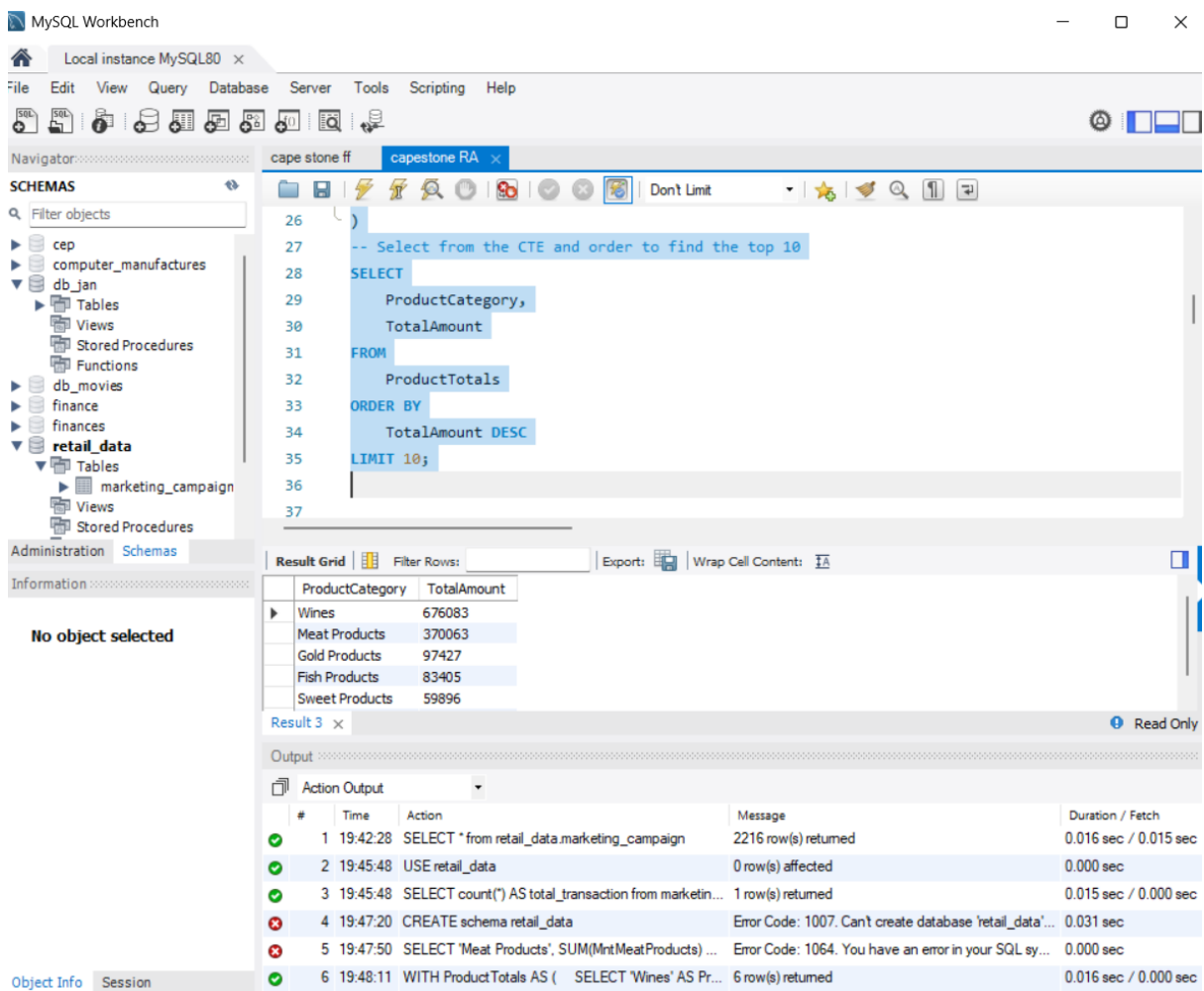
Output: 2216 row(s) returned

1 Data Preprocessing:

Calculate the total number of customer encounters in the marketing campaign dataset



Identify the top 10 most purchased products in the dataset, such as Wines, Meat Products, etc.



Find the count of response values

MySQL Workbench

Local instance MySQL80 x

File Edit View Query Database Server Tools Scripting Help

Navigator: cape stone ff capestone RA x

SCHEMAS

Filter objects

- cep
- computer_manufactures
- db_jan
 - Tables
 - Views
 - Stored Procedures
 - Functions
- db_movies
- finance
- finances
- retail_data
 - Tables
 - marketing_campaign
 - Views
 - Stored Procedures

Administration Schemas

Information

No object selected

Object Info Session

33 ORDER BY

34 Execute the selected portion of the script or everything, if there is no selection

35 LIMIT 10;

36

37

38

39 #3 task 3

40 • USE retail_data;

41 • SELECT count(*) AS response from marketing_campaign;

42

43

44 #3 task 4

Result Grid

Filter Rows: Export: Wrap Cell Content: I

response

2216

Result 4 x Read Only

Output

Action Output

#	Time	Action	Message	Duration / Fetch
3	19:45:48	SELECT count(*) AS total_transaction from market...	1 row(s) returned	0.015 sec / 0.000 sec
4	19:47:20	CREATE schema retail_data	Error Code: 1007. Can't create database 'retail_da...	0.031 sec
5	19:47:50	SELECT 'Meat Products', SUM(MntMeatProducts)...	Error Code: 1064. You have an error in your SQL s...	0.000 sec
6	19:48:11	WITH ProductTotals AS (SELECT 'Wines' AS ...	6 row(s) returned	0.016 sec / 0.000 sec
7	19:49:41	USE retail_data	0 row(s) affected	0.000 sec
8	19:49:41	SELECT count(*) AS response from marketing_ca...	1 row(s) returned	0.015 sec / 0.000 sec

Determine the distribution of customers based on their education level and marital status

MySQL Workbench

Local instance MySQL80

File Edit View Query Database Server Tools Scripting Help

capstone ff capstone RA

Open Inspector for the selected object

SCHEMAS

Filter objects

- cep
- computer_manufactures
- db_jan
 - Tables
 - Views
 - Stored Procedures
- Functions
- db_movies
- finance
- finances
- retail_data
 - Tables
 - marketing_campaign
 - Views
 - Stored Procedures

```
45 SELECT
46     Education,
47     Marital_Status,
48     COUNT(*) AS NumberOfCustomers
49 FROM
50     marketing_campaign
51 GROUP BY
52     Education,
53     Marital_Status
54 ORDER BY
55     NumberOfCustomers DESC;
56
```

Administration Schemas

Information

No object selected

Result Grid

Education	Marital_Status	NumberOfCustomers
Graduation	Married	429
Graduation	Together	285
Graduation	Single	246
PhD	Married	190
Master	Married	138

Result 5

Output

Read Only

Identify the average income of customers who participated in the marketing campaign

MySQL Workbench

Local instance MySQL80 x

File Edit View Query Database Server Tools Scripting Help

Navigator: cape stone ff capestone RA* x

SCHEMAS

Filter objects

- cep
- computer_manufactures
- db_jan
 - Tables
 - Views
 - Stored Procedures
 - Functions
- db_movies
- finance
- finances
- retail_data
 - Tables
 - marketing_campaign
 - Views
 - Stored Procedures

Administration Schemas

Information

Columns:

- ID int
- Year_Birth int
- Education te
- Marital_Status te
- Income int
- Kidhome int
- Teenhome int
- Dt_Customer te
- Recency int
- MntWines int
- MntFruits int
- MntMeatProducts int
- MntFishProducts int
- MntSweetProducts int
- MntGoldProds int
- NumDealsPurchases int
- NumWebPurchases int
- NumCatalogPurchases int

Object Info Session

Result Grid

Filter Rows: Export: Wrap Cell Content: I

avg(Income)

52247.2514

Result 9 x Read Only

Output

Action Output

#	Time	Action	Message	Duration / Fetch
14	20:02:42	ALTER TABLE marketing_campaign ADD COLU...	0 row(s) affected Records: 0 Duplicates: 0 Wami...	0.125 sec
15	20:02:42	UPDATE marketing_campaign SET Age1 = (EXT...	2216 row(s) affected Rows matched: 2216 Chan...	0.406 sec
16	20:03:40	ALTER TABLE marketing_campaign ADD COLU...	Error Code: 1060. Duplicate column name 'Age_gr...	0.000 sec
17	20:03:54	SELECT Age_group, AVG(NumWebVisitsMo...	4 row(s) returned	0.015 sec / 0.000 sec
18	20:08:43	SELECT COUNT(*) FROM marketing_campa...	2 row(s) returned	0.000 sec / 0.000 sec
19	20:48:46	SELECT avg(Income) FROM marketing_ca...	1 row(s) returned	0.000 sec / 0.000 sec

Query Completed

Identify the distribution of customers' responses to the last campaign

The screenshot shows the MySQL Workbench interface. The left sidebar displays the 'SCHEMAS' tree with 'retail_data' expanded, showing the 'marketing_campaign' table. The main editor window contains a SQL query:

```
#3 task 7
SELECT
  COUNT(*)
FROM
  marketing_campaign
GROUP BY
  AcceptedCmp5;
```

The query results are displayed in the 'Result Grid' tab, showing two rows:

COUNT(*)
2054
162

The bottom panel shows the 'Output' tab with a table of actions and their results:

#	Time	Action	Message	Duration / Fetch
13	20:00:20	SELECT AVG(Kidhome) AS avg_children, ...	1 row(s) returned	0.000 sec / 0.000 sec
14	20:02:42	ALTER TABLE marketing_campaign ADD COLU...	0 row(s) affected Records: 0 Duplicates: 0 Wami...	0.125 sec
15	20:02:42	UPDATE marketing_campaign SET Age1 = (EXT...	2216 row(s) affected Rows matched: 2216 Chan...	0.406 sec
16	20:03:40	ALTER TABLE marketing_campaign ADD COLU...	Error Code: 1060. Duplicate column name 'Age_gr...	0.000 sec
17	20:03:54	SELECT Age_group, AVG(NumWebVisitsMo...	4 row(s) returned	0.015 sec / 0.000 sec
18	20:08:43	SELECT COUNT(*) FROM marketing_campa...	2 row(s) returned	0.000 sec / 0.000 sec

Calculate the average number of children and teenagers in customers' households

The screenshot shows the MySQL Workbench interface. The left sidebar displays the 'SCHEMAS' tree with 'retail_data' expanded, showing 'marketing_campaign' as a table. The main editor window shows a SQL query for 'marketing_campaign' with columns 'avg_children' and 'avg_teens'. The query is as follows:

```

63 marketing_campaign
64 GROUP BY
65 Response
66 ORDER BY
67 NumberOfCustomers DESC;
68
69 #3 task 8
70 SELECT
71     AVG(Kidhome) AS avg_children,
72     AVG(Teenhome) AS avg_teens
73 FROM marketing_campaign;
74

```

The 'Result Grid' shows the following data:

avg_children	avg_teens
0.4418	0.4418 054

The 'Information' panel on the left shows the table structure for 'marketing_campaign':

Columns:	int
ID	int
Year_Birth	te
Education	te
Marital_Status	te
Income	int
Kidhome	int
Teenhome	int
Dt_Customer	te
Recency	int
MntWines	int
MntFruits	int
MntMeatProducts	int
MntFishProducts	int
MntSweetProducts	int
MntGoldPrnds	int

The 'Output' panel shows the execution log with the following actions:

#	Time	Action	Message	Duration / Fetch
8	19:49:41	SELECT count(*) AS response from marketing_ca...	1 row(s) returned	0.015 sec / 0.000 sec
9	19:50:42	SELECT Education, Marital_Status, COU...	31 row(s) returned	0.031 sec / 0.000 sec
10	19:58:51	SELECT AVG(Kidhome) AS avg_children, AVG(...	Error Code: 1064. You have an error in your SQL s...	0.000 sec
11	19:59:11	AVG(Kidhome) AS avg_children, AVG(Teenho...	Error Code: 1064. You have an error in your SQL s...	0.000 sec
12	19:59:53	SELECT AVG(Kidhome) AS avg_children, ...	Error Code: 1064. You have an error in your SQL s...	0.000 sec
13	20:00:20	SELECT AVG(Kidhome) AS avg_children, ...	1 row(s) returned	0.000 sec / 0.000 sec

Create an Age column by subtracting year_birth from the current year

- Create Age_group columns based on the below condition:
 WHEN Age BETWEEN 18 AND 25 THEN '18-25'
 WHEN Age BETWEEN 26 AND 35 THEN '26-35'
 WHEN Age BETWEEN 36 AND 45 THEN '36-45'
 WHEN Age BETWEEN 46 AND 55 THEN '46-55'
 ELSE '56+'
- Determine the average number of visits per month for customers in each age group

MySQL Workbench

Local instance MySQL80 x

File Edit View Query Database Server Tools Scripting Help

Navigator: capestone ff capestone RA*

SCHEMAS

Filter objects

cep
computer_manufactures
db_jan
Tables
Views
Stored Procedures
Functions
db_movies
finance
finances
retail_data
Tables
marketing_campaign
Views
Stored Procedures

Administration Schemas

Information

Table: marketing_campaign

Columns:

ID int
Year_Birth int
Education te
Marital_Status te
Income int
Kidhome int
Teenhome int
Dt_Customer te
Recency int
MntWines int
MntFruits int
MntMeatProducts int
MntFishProducts int
MntSweetProducts int
MntGnrlPrds int

```

90 ELSE '56+'
91 END;
92 SELECT
93   Age_group,
94   AVG(NumWebVisitsMonth) AS AvgMonthlyVisits
95 FROM
96   marketing_campaign
97 GROUP BY
98   Age_group
99 ORDER BY
100  Age_group;
101

```

Result Grid

Age_group	AvgMonthlyVisits
26-35	4.9016
36-45	5.5559
46-55	5.6603
56+	5.0282

Result 7 x

Output

Action Output

#	Time	Action	Message	Duration / Fetch
12	19:59:53	SELECT AVG(Kidhome) AS avg_children	Error Code: 1064. You have an error in your SQL s...	0.000 sec
13	20:00:20	SELECT AVG(Kidhome) AS avg_children	1 row(s) returned	0.000 sec / 0.000 sec
14	20:02:42	ALTER TABLE marketing_campaign ADD COLU...	0 row(s) affected Records: 0 Duplicates: 0 Warni...	0.125 sec
15	20:02:42	UPDATE marketing_campaign SET Age1 = (EXT...	2216 row(s) affected Rows matched: 2216 Chan...	0.406 sec
16	20:03:40	ALTER TABLE marketing_campaign ADD COLU...	Error Code: 1060. Duplicate column name 'Age_gr...	0.000 sec
17	20:03:54	SELECT Age_group, AVG(NumWebVisitsMo...	4 row(s) returned	0.015 sec / 0.000 sec

Query Completed

Task 1: General Data Overview:

Exploratory Data Analysis Check the first few rows of the dataset to understand its structure

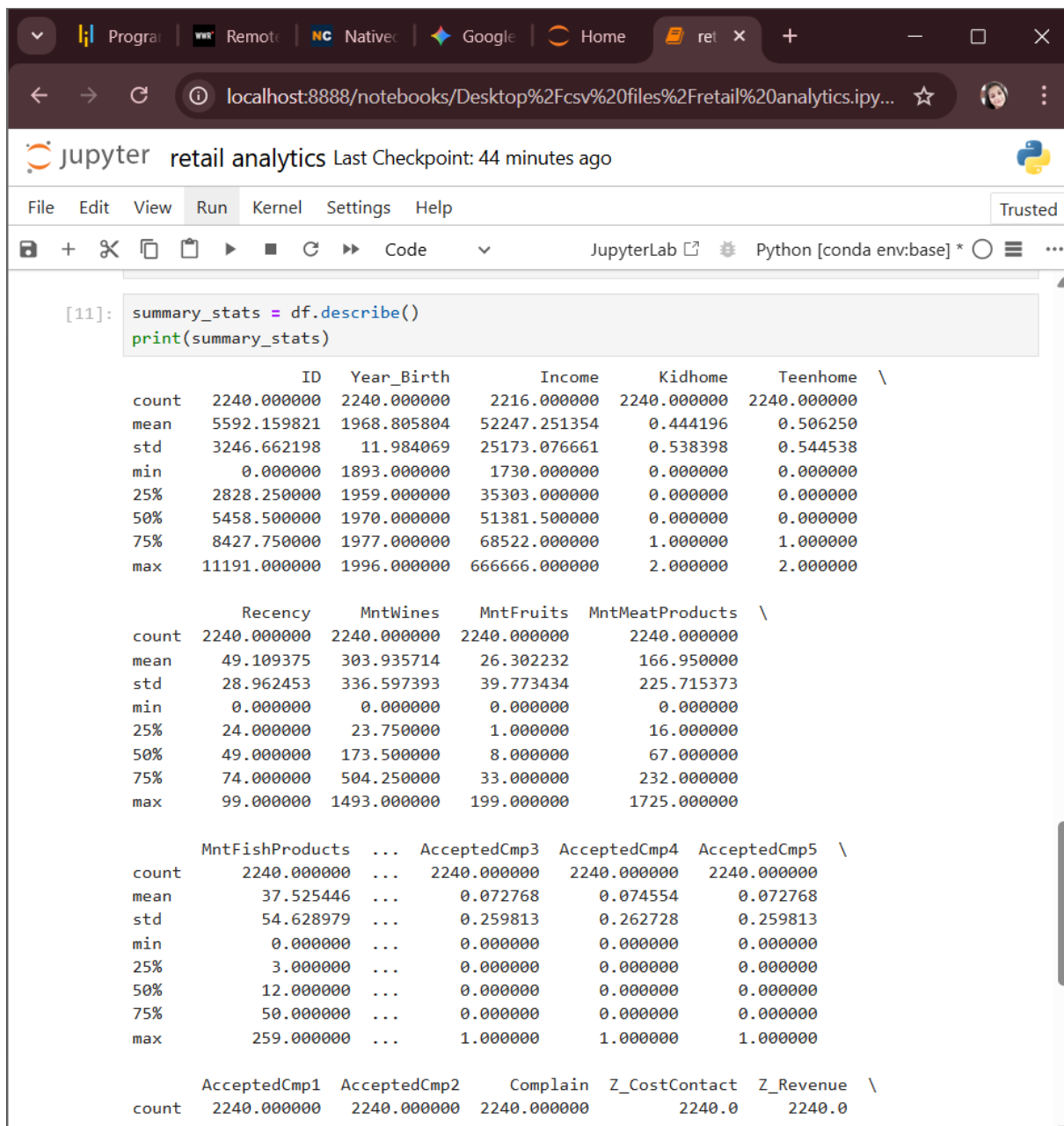
Check the data types of each column

Check for any missing values in the dataset

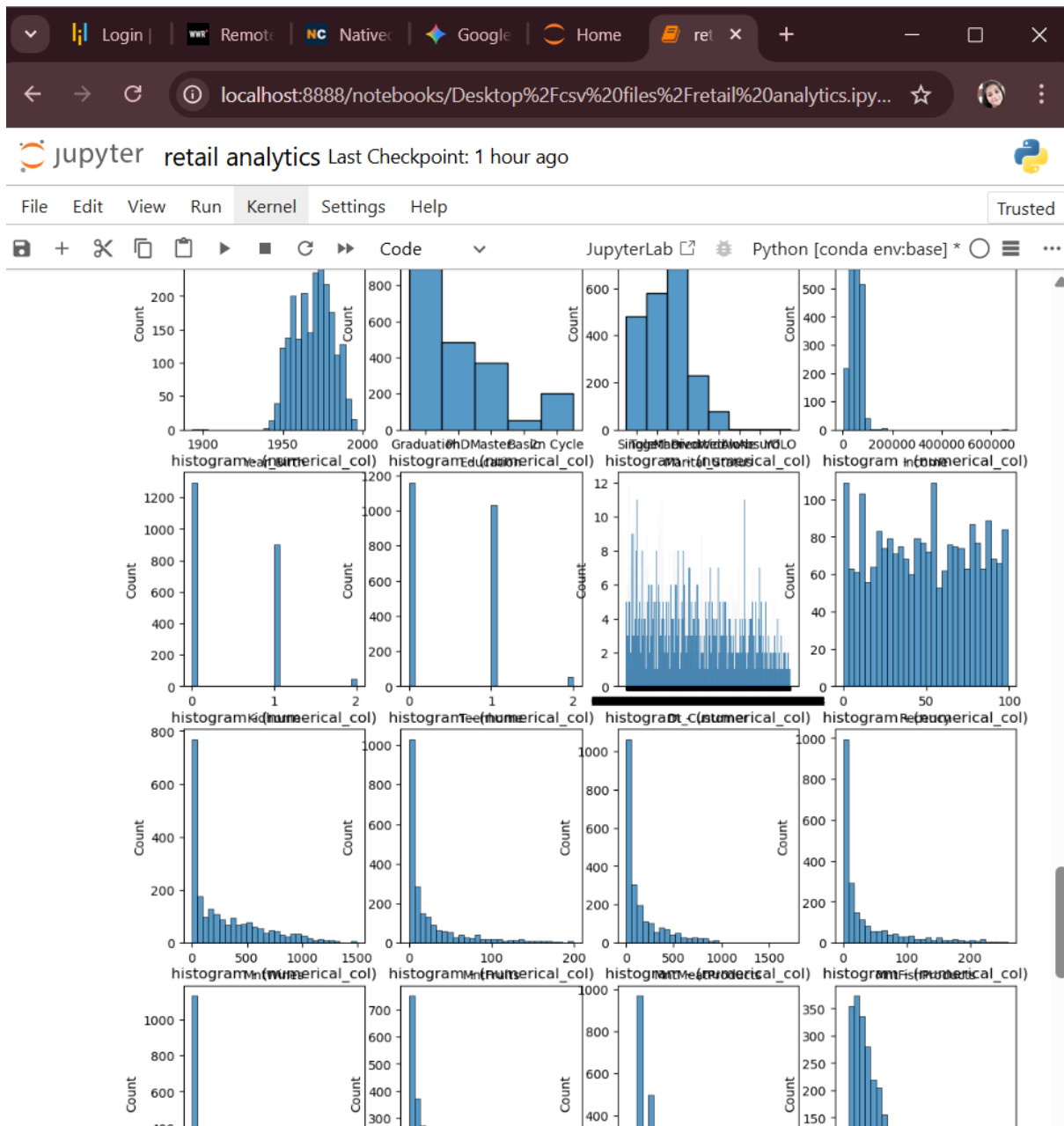

```
[9]: missing_values = df.isnull().sum()  
     print(missing_values)
```

ID	0
Year_Birth	0
Education	0
Marital_Status	0
Income	24
Kidhome	0
Teenhome	0
Dt_Customer	0
Recency	0
MntWines	0
MntFruits	0
MntMeatProducts	0
MntFishProducts	0
MntSweetProducts	0
MntGoldProds	0
NumDealsPurchases	0
NumWebPurchases	0
NumCatalogPurchases	0
NumStorePurchases	0
NumWebVisitsMonth	0
AcceptedCmp3	0
AcceptedCmp4	0
AcceptedCmp5	0
AcceptedCmp1	0
AcceptedCmp2	0

Compute summary statistics for numerical columns (mean, median, min, max, and standard deviation)



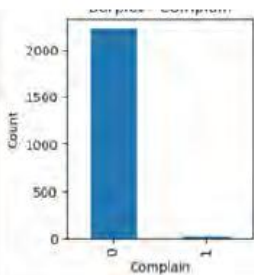
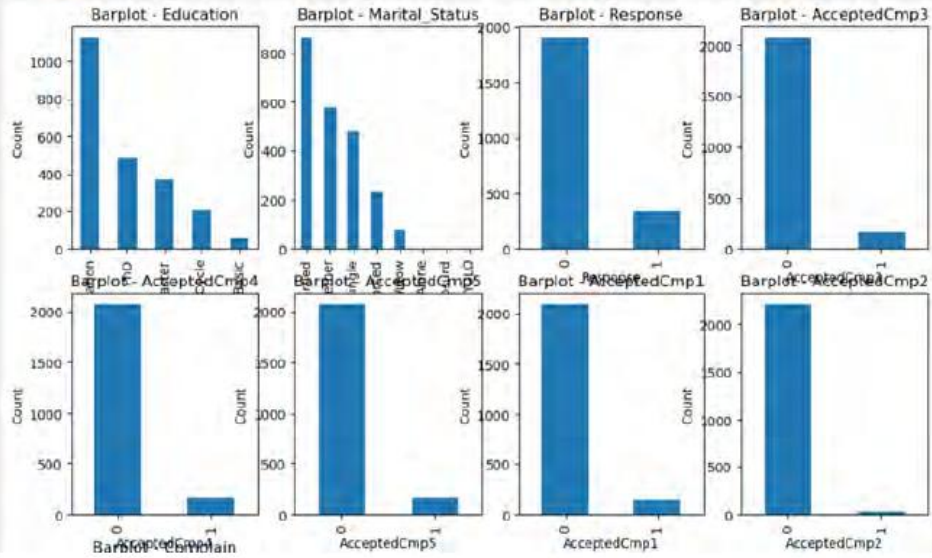
Explore the distribution of numerical variables using histograms or box plots



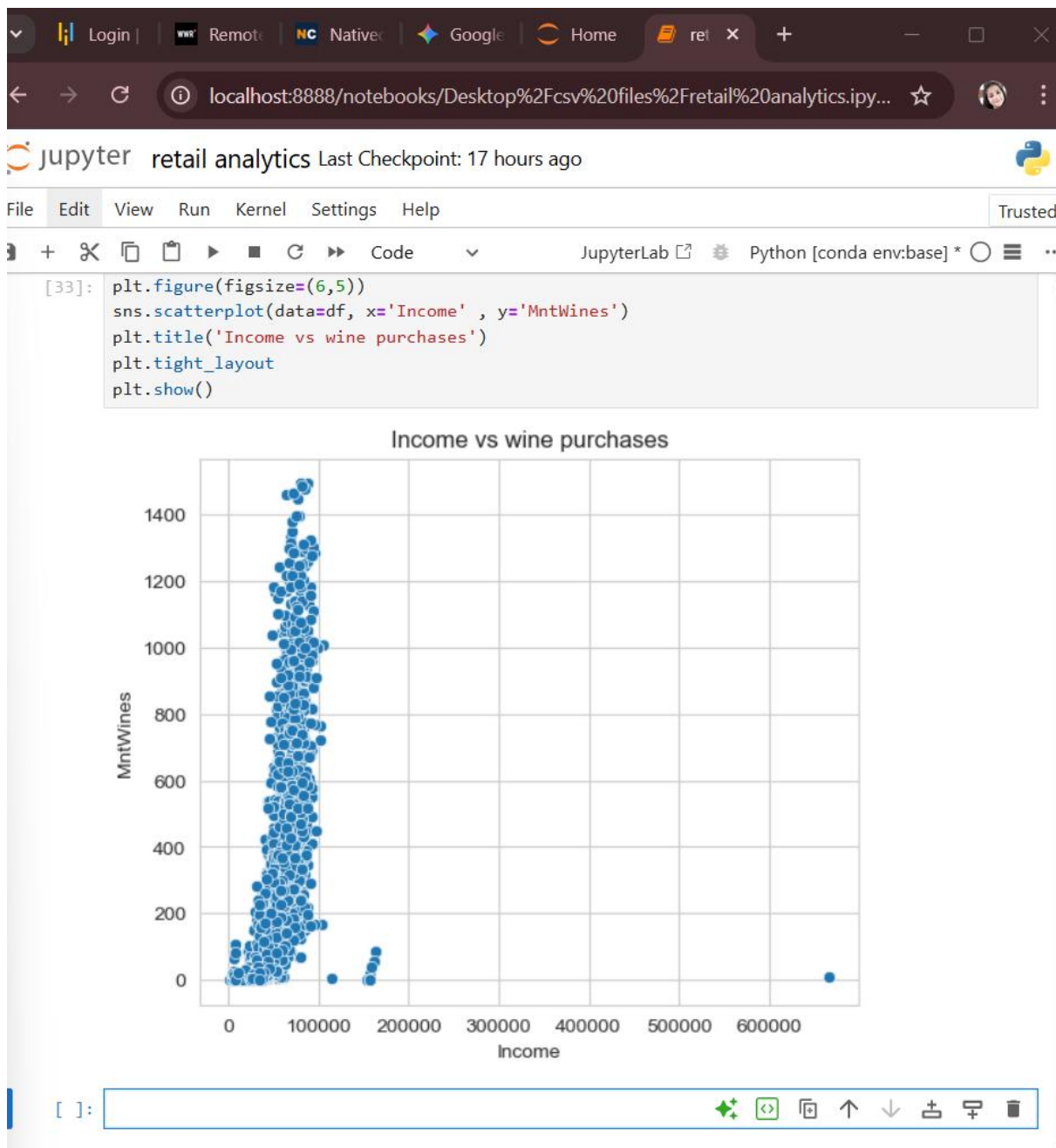
Explore the distribution of each categorical variable using bar plots or pie charts

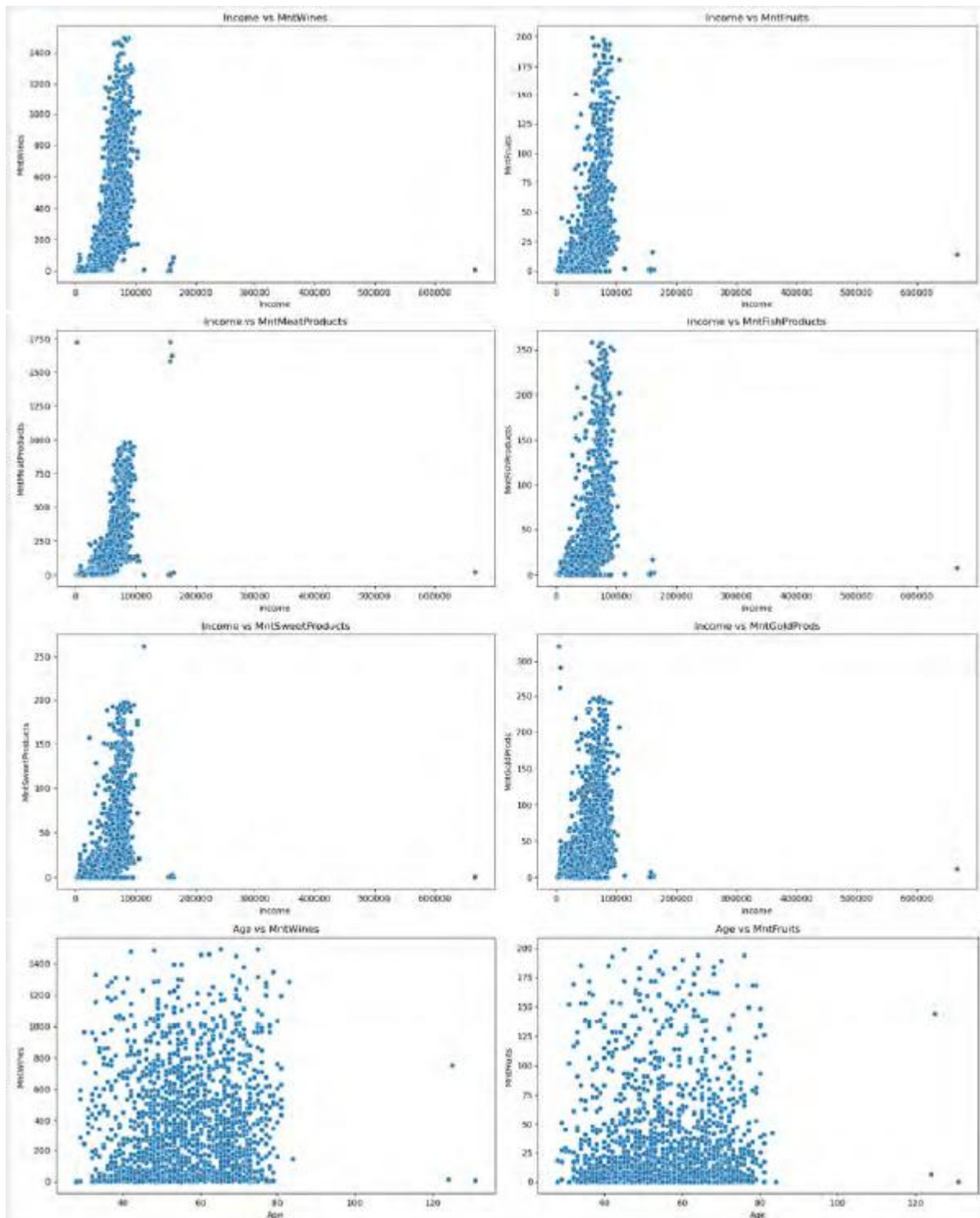
```
(11) # Plot a Histogram
```

```
plt.figure(figsize=(12,18))
for i, col in enumerate(categorical_col):
    plt.subplot(5,4,i+1)
    data[col].value_counts().plot(kind='bar')
    plt.title(f'Barplot - {col}')
    plt.xlabel(col)
    plt.ylabel('Count')
plt.tight_layout()
plt.show()
```



Identify outliers in numerical variables using box plots or scatter plots



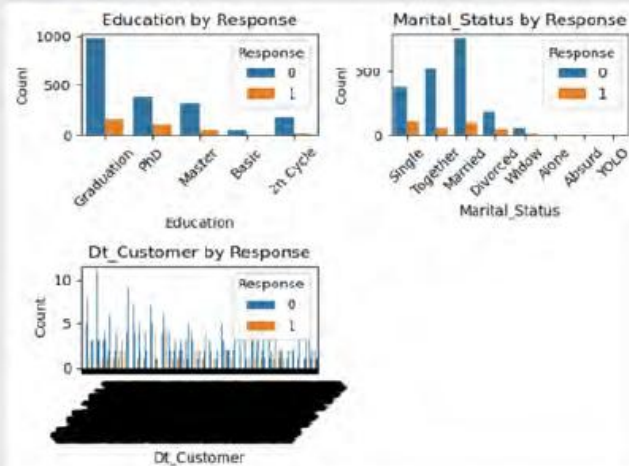


Explore the relationship between categorical variables and the target variable using bar plots or chi-square tests

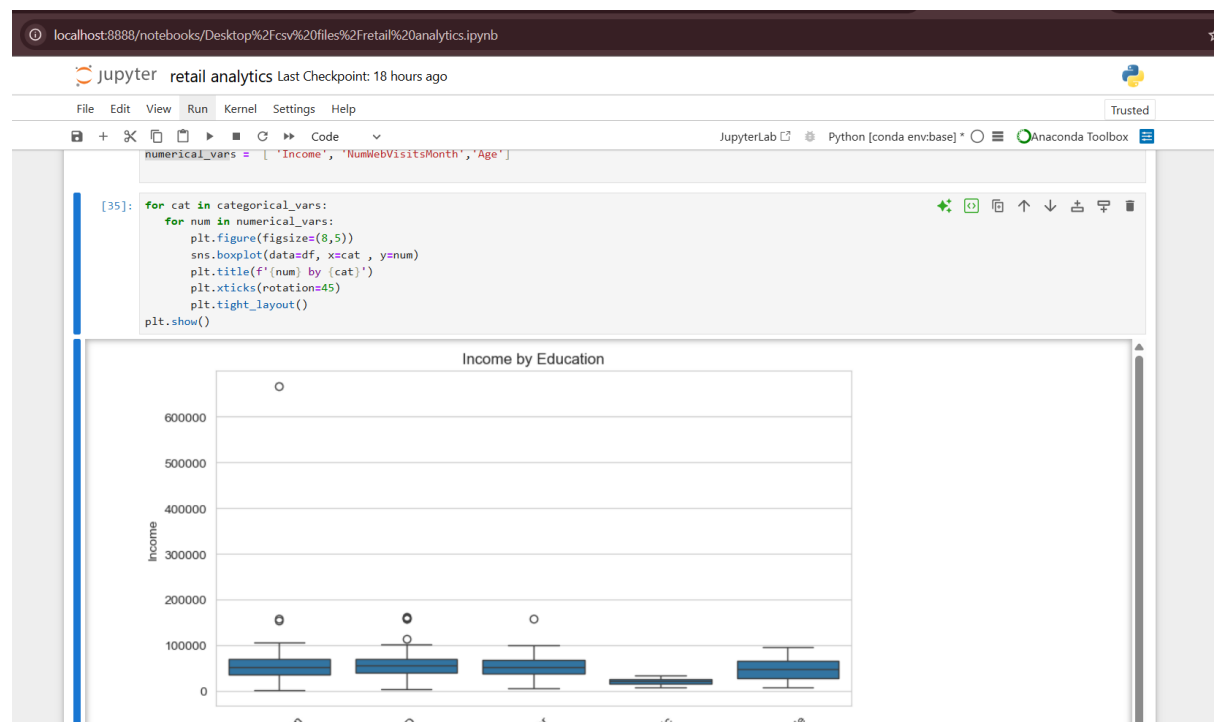

```
categorical_cols = data.select_dtypes(include=['object']).columns
```

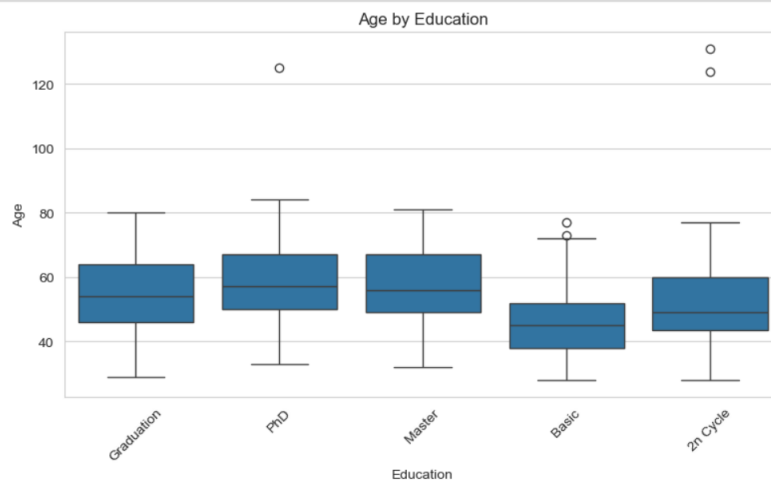
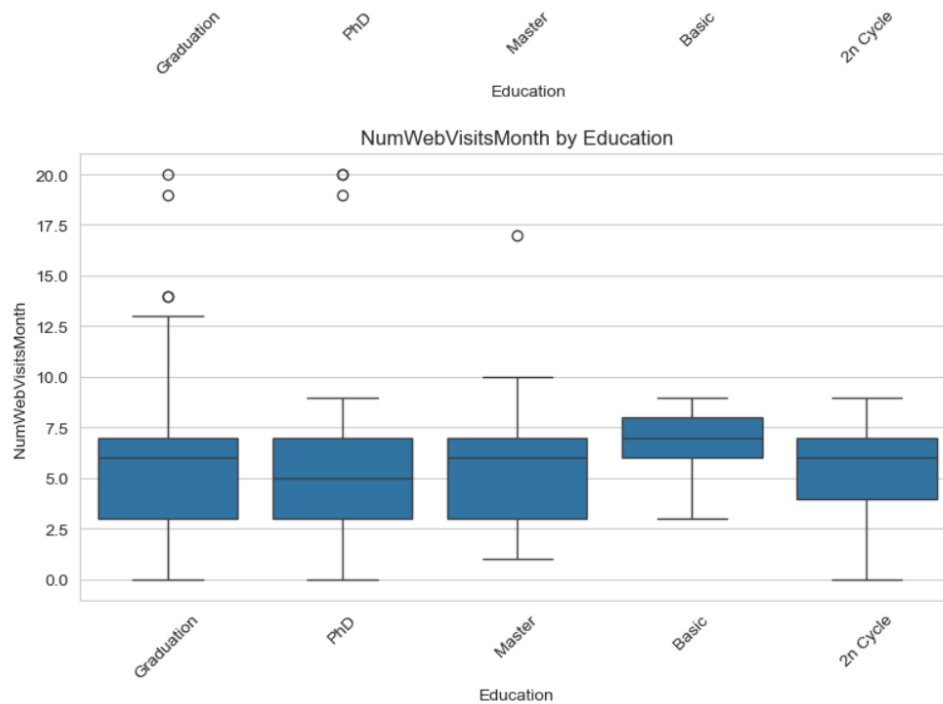
```
# Plot Response distribution by each categorical variable
```

```
for idx, col in enumerate(categorical_cols,1):
    plt.subplot(2,2,idx)
    sns.countplot(x=col, hue='Response', data=data)
    plt.title(f'{col} by Response')
    plt.xticks(rotation=45)
    plt.ylabel('Count')
    plt.tight_layout()
plt.show()
```



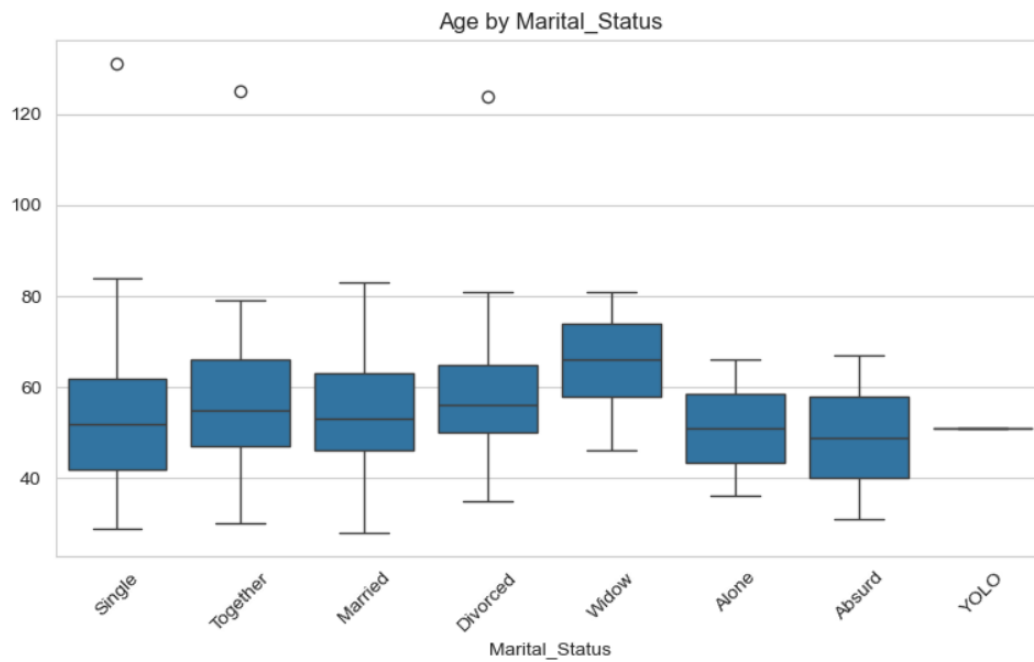
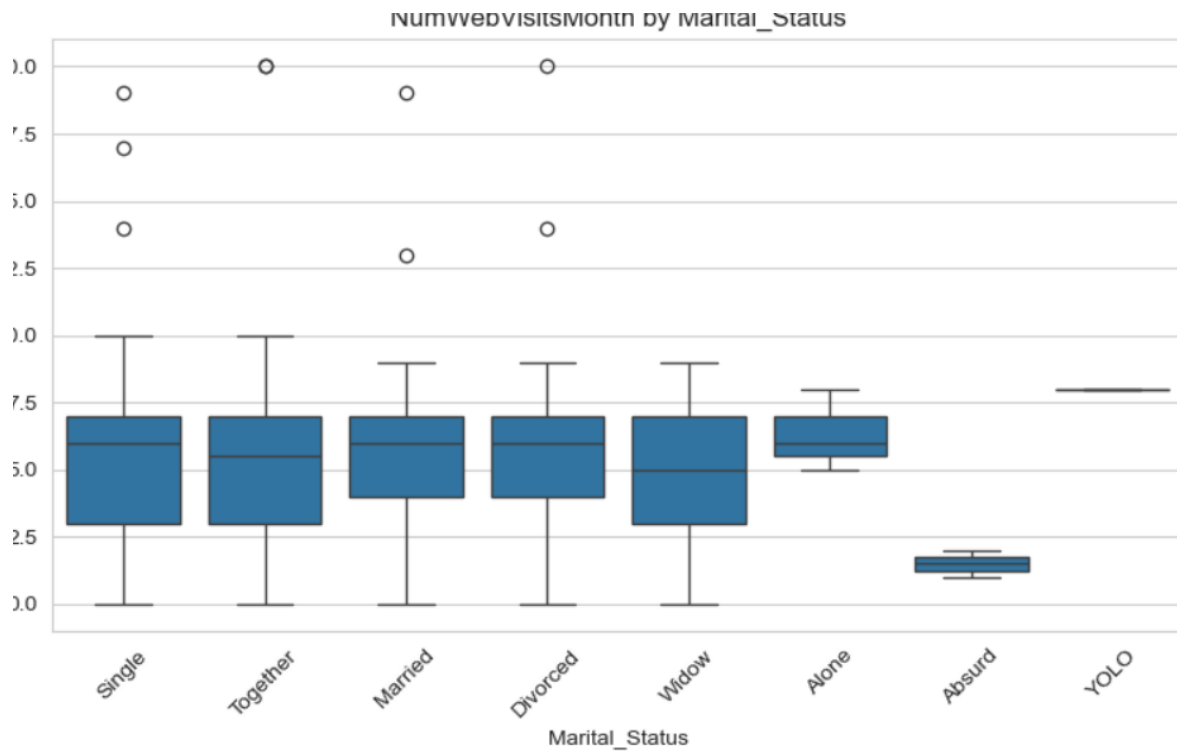
Explore the relationship between numerical and categorical variables using box plots or violin plots



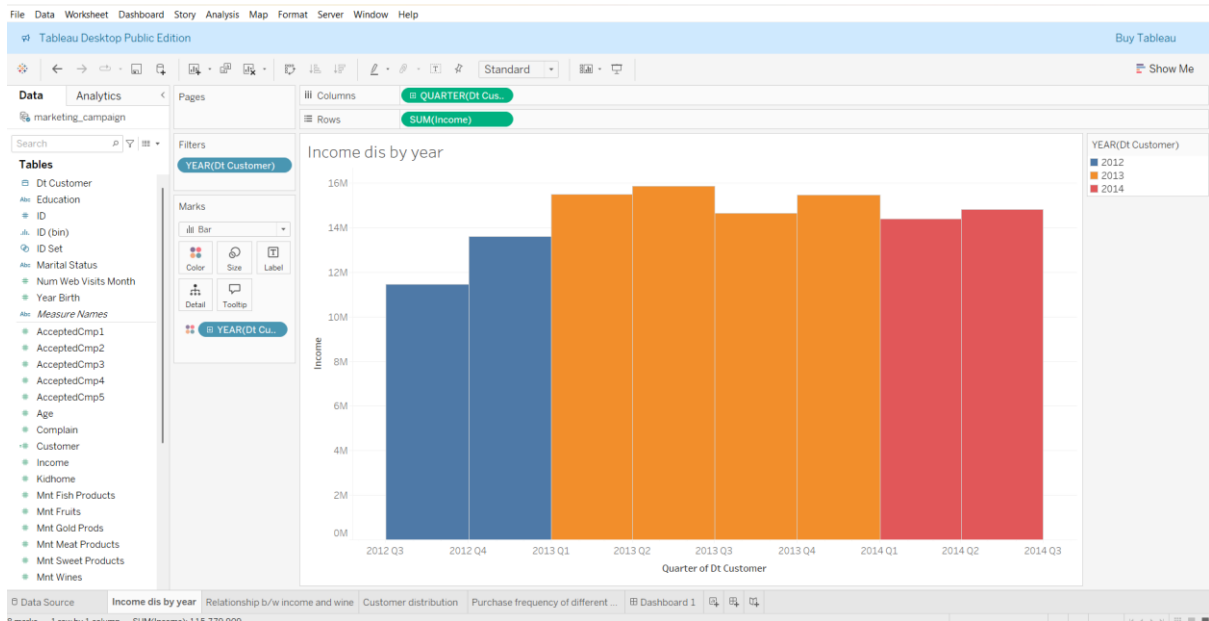


Income by Marital_Status

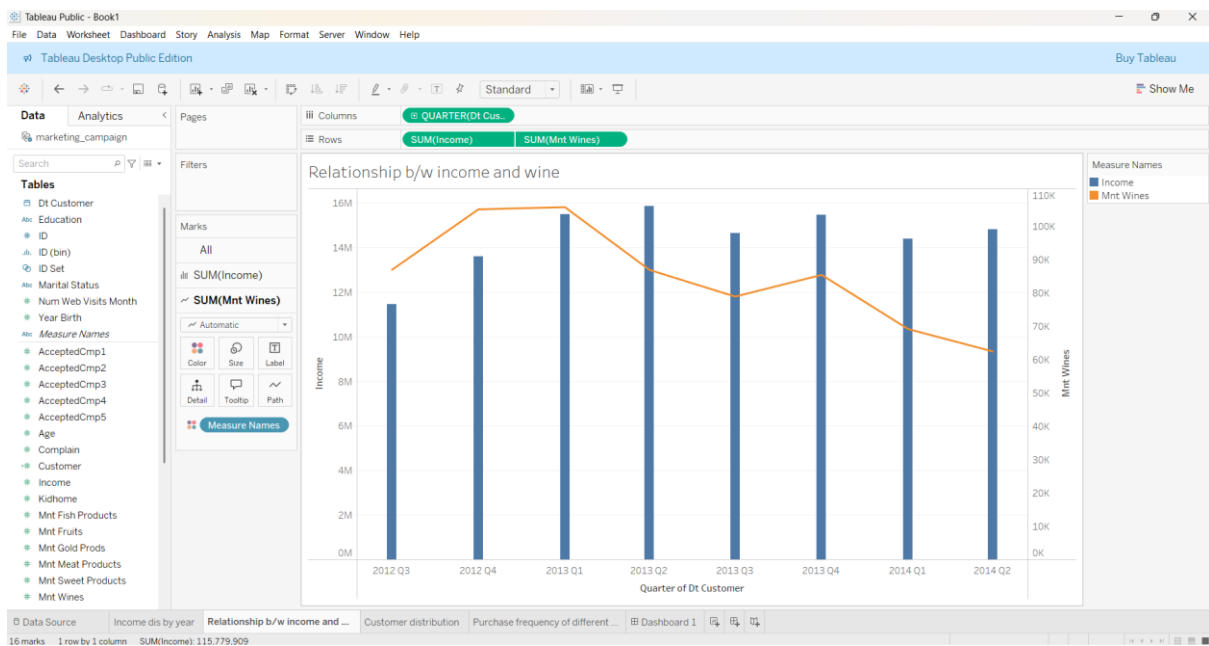




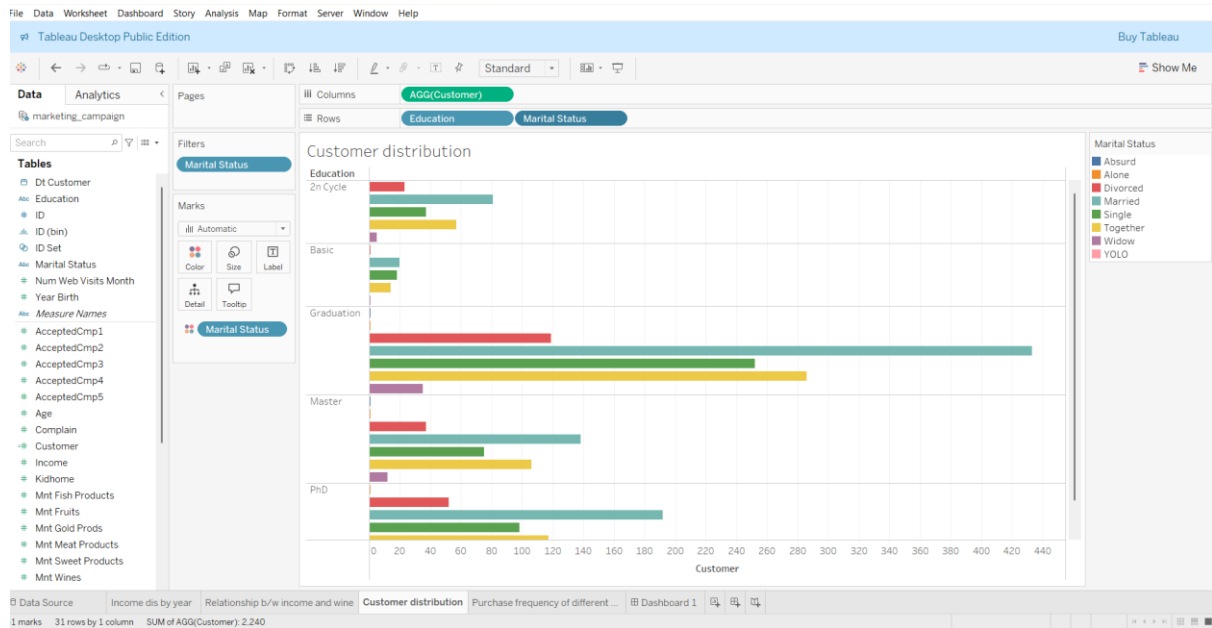
Task 1: Visualize Customer Income Distribution grouped by Year of Registration



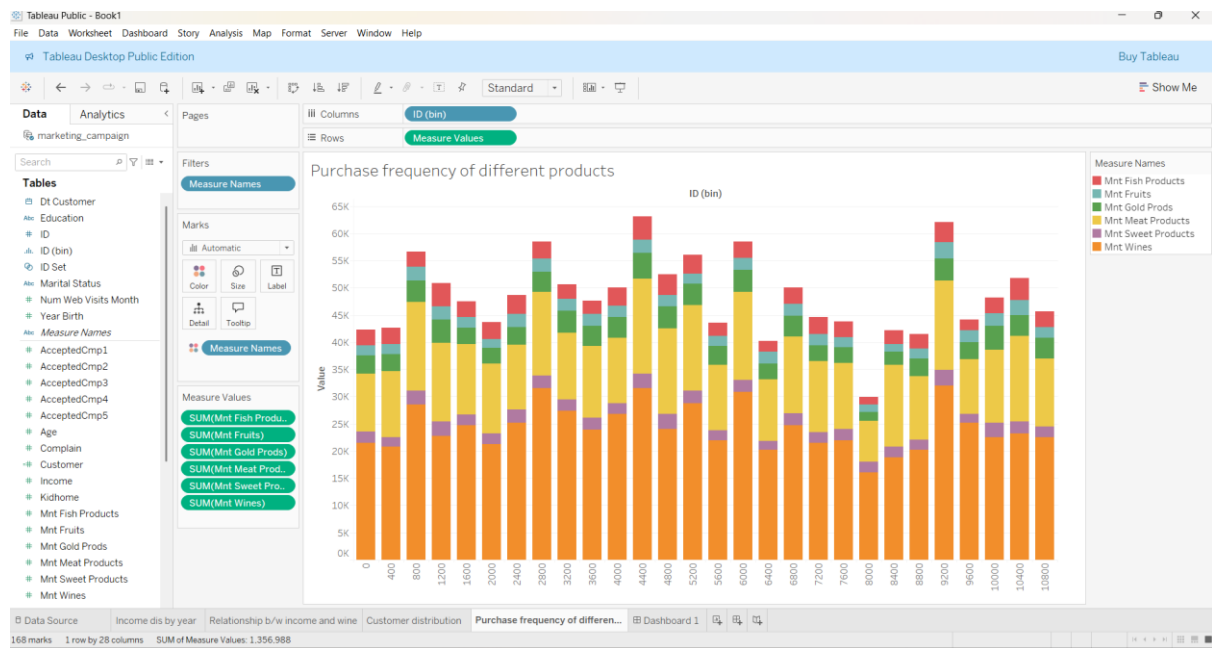
Task 2: Examine the breakdown of education levels and marital status using side-by-side circles to represent count



Task 3: Explore the relationship between Income and Wine Spending



Task 4: Analyze the frequency of purchases across various product categories



Task 5: Create a dashboard using all the visualizations

