

**Applied Data Science with Python**

Course-End Project 01 – Marketing Campaign Analysis





**Phase-End Project: Marketing Campaigns**

### Problem Scenario: ‘Marketing mix’ is a popular concept used in implementing marketing strategies. A marketing mix includes multiple areas of focus as part of a comprehensive marketing plan. This all revolves around the four Ps of marketing - product, price, place, and promotion.

**Problem Objective:** As a data scientist, you should perform exploratory data analysis and hypothesis testing. The goal is to gain a better understanding of the various factors that contribute to customer acquisition.

**Data Description:**

The variables birth-year, education, income, and so on are related to the first 'P' or 'People' in the tabular data provided to the user. The amount spent on wine, fruits, gold, etc., is related to ‘Product’. The information pertinent to sales channels, like websites, stores, etc., is related to ‘Place’, and the fields which talk about promotions and results of different campaigns are related to ‘Promotion’.

**Steps to Perform:**

* Once data is imported, investigate variables like Dt\_Customer and Income, etc., and check if they are imported correctly.
* Income values for a few customers are missing. Perform missing value imputation. Assume that the customers with similar education and marital status make the same yearly income, on average. You may have to clean the data before performing this. For data cleaning, look into the categories of education and marital status.
* Create variables to populate the total number of children, age, and total spending.

Hint: From the number of purchases through the three channels, people can derive the total purchases.

* Create box plots and histograms to understand the distributions and outliers. Perform outlier treatment.
* Use ordinal encoding and one hot encoding according to different types of categorical variables.
* Create a heatmap to showcase the correlation between different pairs of variables.
* Test the following hypotheses:
* Older people are not as tech-savvy and probably prefer shopping in-store.
* Customers with kids probably have less time to visit a store and would prefer to shop online.
* Other distribution channels may cannibalize sales at the store.
* Does the US fare significantly better than the rest of the world in terms of total purchases?
* Use appropriate visualization to help analyze the following:
* Which products are performing the best, and which are performing the least in terms of revenue?
* Is there any pattern between the age of customers and the last campaign acceptance rate?
* Which Country has the greatest number of customers who accepted the last campaign?
* Do you see any pattern in the no. of children at home and total spend?
* Education background of the customers who complained in the last 2 years.

**Solution**

Marketing Campaign Analysis - Solution Code

## 1. Load and Inspect Data

import pandas as pd  
df = pd.read\_csv('marketing\_data.csv')  
df.columns = df.columns.str.strip()  
df.info()  
df.head()

## 2. Clean and Format Columns

df['Income'] = df['Income'].replace('[\$,]', '', regex=True).astype(float)  
df['Dt\_Customer'] = pd.to\_datetime(df['Dt\_Customer'], format='%m/%d/%y')

## 3. Feature Engineering

df['Age'] = 2025 - df['Year\_Birth']  
df['Total\_Children'] = df['Kidhome'] + df['Teenhome']  
df['Total\_Spending'] = df[['MntWines', 'MntFruits', 'MntMeatProducts',  
 'MntFishProducts', 'MntSweetProducts', 'MntGoldProds']].sum(axis=1)  
df['Total\_Purchases'] = df[['NumWebPurchases', 'NumCatalogPurchases', 'NumStorePurchases']].sum(axis=1)

## 4. Impute Missing Income

income\_means = df.groupby(['Education', 'Marital\_Status'])['Income'].mean()  
df['Income'] = df.apply(lambda row: income\_means.loc[(row['Education'], row['Marital\_Status'])] if pd.isnull(row['Income']) else row['Income'], axis=1)

## 5. Outlier Treatment (IQR Capping)

def cap\_outliers(series):  
 Q1 = series.quantile(0.25)  
 Q3 = series.quantile(0.75)  
 IQR = Q3 - Q1  
 lower = Q1 - 1.5 \* IQR  
 upper = Q3 + 1.5 \* IQR  
 return series.clip(lower, upper)  
  
for col in ['Income', 'Age', 'Total\_Spending', 'Total\_Purchases']:  
 df[col] = cap\_outliers(df[col])

## 6. Encoding Categorical Variables

from sklearn.preprocessing import OrdinalEncoder  
edu\_order = [['Basic', '2n Cycle', 'Graduation', 'Master', 'PhD']]  
df['Education\_Encoded'] = OrdinalEncoder(categories=edu\_order).fit\_transform(df[['Education']])  
df = pd.get\_dummies(df, columns=['Marital\_Status', 'Country'], drop\_first=True)  
df.drop(columns=['Education'], inplace=True)

## 7. Correlation Heatmap

import seaborn as sns  
import matplotlib.pyplot as plt  
  
plt.figure(figsize=(16,12))  
sns.heatmap(df.corr(numeric\_only=True), cmap='coolwarm', linewidths=0.5)  
plt.title('Correlation Heatmap')  
plt.show()

## 8. Hypothesis Testing

from scipy.stats import spearmanr, ttest\_ind  
  
# Hypothesis 1  
spearmanr(df['Age'], df['NumStorePurchases'])  
  
# Hypothesis 2  
spearmanr(df['Total\_Children'], df['NumWebPurchases'])  
  
# Hypothesis 3  
spearmanr(df['NumWebPurchases'], df['NumStorePurchases'])  
spearmanr(df['NumCatalogPurchases'], df['NumStorePurchases'])  
  
# Hypothesis 4  
us = df[df['Country\_US'] == 1]['Total\_Purchases']  
non\_us = df[df['Country\_US'] == 0]['Total\_Purchases']  
ttest\_ind(us, non\_us, equal\_var=False)

## 9. Product Performance

product\_cols = ['MntWines', 'MntFruits', 'MntMeatProducts', 'MntFishProducts', 'MntSweetProducts', 'MntGoldProds']  
df[product\_cols].sum().sort\_values(ascending=False)

## 10. Campaign and Demographic Insights

# Age vs Response  
df.groupby('Response')['Age'].mean()  
  
# Country with most accepted campaigns  
df['Country'] = df[[col for col in df.columns if col.startswith('Country\_')]].idxmax(axis=1).str.replace('Country\_', '')  
df[df['Response'] == 1]['Country'].value\_counts()  
  
# Children vs Spending  
df.groupby('Total\_Children')['Total\_Spending'].mean()  
  
# Education of Complaints  
edu\_order = ['Basic', '2n Cycle', 'Graduation', 'Master', 'PhD']  
df['Education'] = df['Education\_Encoded'].apply(lambda x: edu\_order[int(x)])  
df[df['Complain'] == 1]['Education'].value\_counts()

Jupyter notebook is uploaded as a separate file.

Marketing Campaign Analysis - Visual Report

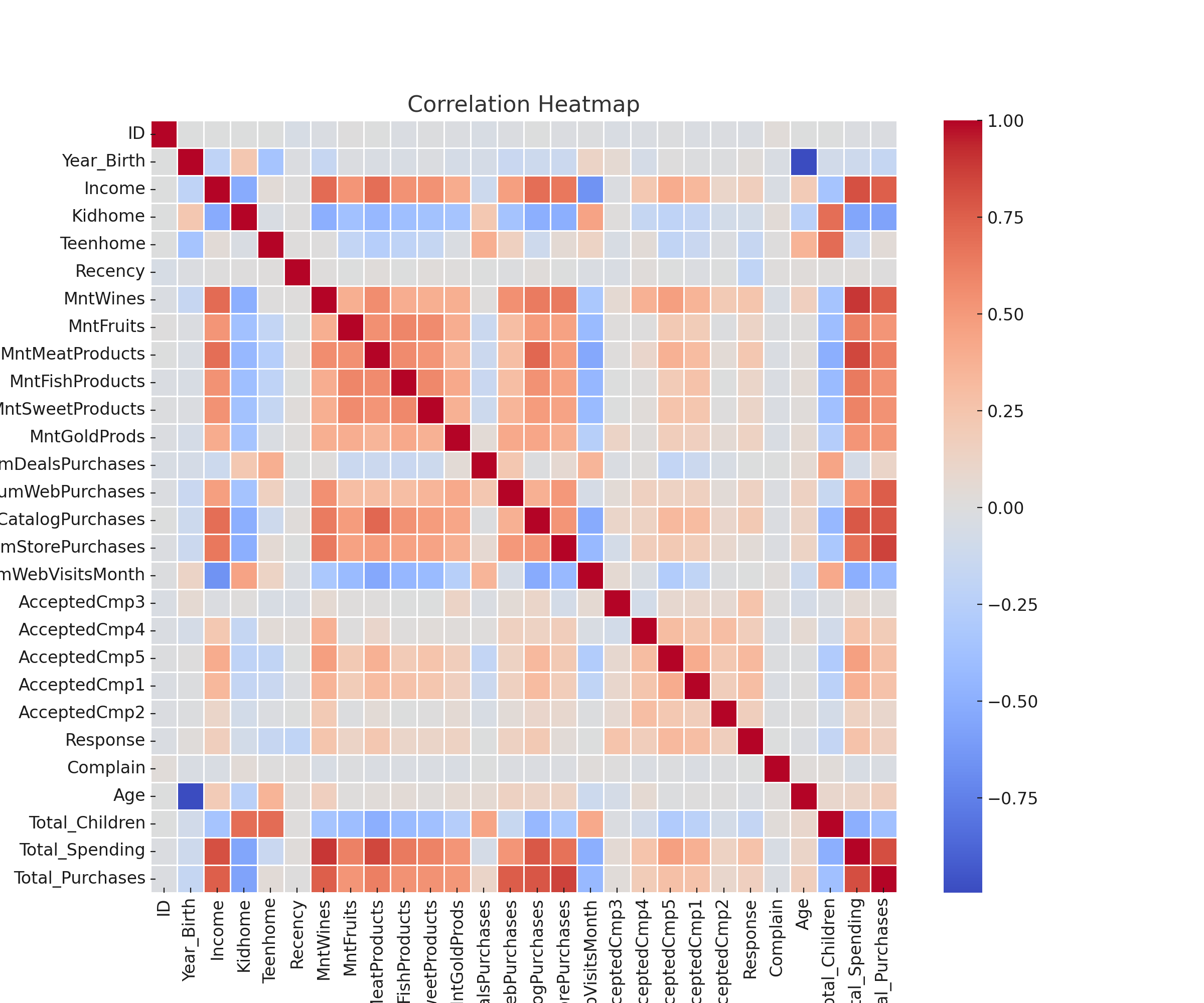
# Feature Distributions

Histograms and KDE plots showing the distribution of key variables:



# Correlation Heatmap

This heatmap illustrates the correlation between key numerical variables:



Marketing Campaign Analysis Summary

# Key Insights Summary

## Exploratory Data Analysis

- Income is right-skewed with outliers above $115,000 (capped via IQR).  
- Age ranges from 29 to ~90 (after outlier treatment), with an average of ~56.  
- Total Spending is highly skewed, peaking under $1,000 but reaching $2,500.  
- Missing Income values were imputed using mean by Education and Marital\_Status.

## Hypothesis Testing

|  |  |  |
| --- | --- | --- |
| **Hypothesis** | **Result** | **Insight** |
| Older people prefer shopping in-store | Supported | Positive correlation between age and in-store purchases. |
| Customers with kids prefer to shop online | Rejected | More children → fewer web purchases. Contradicts the assumption. |
| Online/catalog channels cannibalize store sales | Rejected | Web and catalog purchases are positively correlated with store ones. |
| US performs significantly better in total purchases | Rejected | US average is slightly higher, but not statistically significant. |

## Product Performance

* Top Performers:  
   1. Wine  
   2. Meat Products  
   3. Gold Products
* Low Performers: Fruits and Sweet Products

## Customer Behavior Insights

1. Age vs Campaign Acceptance: No strong pattern; age not a decisive factor.
2. Country with Most Acceptances: Spain (SP) had the highest number of campaign acceptances by far.
3. Children at Home vs Spending: More children → lower total spending.
4. Complaints by Education Level: Most complaints came from Graduates; none from customers with Basic education.