Business Decision Using Data Analytics

A PROJECT REPORT

Submitted by

Yash Lukhi (18BECE30561)
Mohit Dodiya (18BECE30106)
Saumya Joshi (18BECE30555)

In fulfillment for the award of the degree

Of

BACHELOR OF ENGINEERING IN COMPUTER ENGINEERING

LDRP INSTITUTE OF TECHNOLOGY AND RESEARCH,

GANDHINAGAR

Kadi Sarva Vishwavidyalaya, Gandhinagar

April 2021 – 2022

LDRP Institute of Technology and Research

Computer Engineering Department

CERTIFICATE

This is to certify that the Project Work entitled **"Business Decision using Data Analytics"** has been carried out by **Yash Lukhi (18BECE30561)** under my guidance in fulfilment of the degree of Bachelor of Engineering in Computer Engineering of Kadi Sarva Vishwavidyalaya University, Gandhinagar during the academic year 2021-22.

Mr. Nimesh Patel
Internal Guide
LDRP ITR

Dr. Shivangi Surati
HOD – CE
LDRP ITR

LDRP Institute of Technology and Research

Computer Engineering Department

CERTIFICATE

This is to certify that the Project Work entitled **"Business Decision using Data Analytics"** has been carried out by **Saumya Joshi (18BECE30555)** under my guidance in fulfilment of the degree of Bachelor of Engineering in Computer Engineering of Kadi Sarva Vishwavidyalaya University, Gandhinagar during the academic year 2021-22.

Mr. Nimesh Patel
Internal Guide
LDRP ITR

Dr. Shivangi Surati
HOD – CE
LDRP ITR

LDRP Institute of Technology and Research

Computer Engineering Department

CERTIFICATE

This is to certify that the Project Work entitled **"Business Decision using Data Analytics"** has been carried out by **Dodiya Mohit (18BECE30106)** under my guidance in fulfilment of the degree of Bachelor of Engineering in Computer Engineering of Kadi Sarva Vishwavidyalaya University, Gandhinagar during the academic year 2021-22.

Mr. Nimesh Patel
Internal Guide
LDRP ITR

Dr. Shivangi Surati
HOD – CE
LDRP ITR

ACKNOWLEDGEMENT

We take this opportunity to express our gratitude and regards to our professor Nimesh Patel for his valuable guidance throughout the course of this project.

We would also like to thank our H.O.D Dr. Shivangi Surati and all the Professors of LDRP INSTITUTE OF TECHNOLOGY AND RESEARCH.

We would also like to express our gratitude to our colleagues for their precious and valuable contribution.

ABSTRACT

Data - analytics is becoming a very influential tool for decision-making today both in Business, industry and academia. Data analysis is a vital part of work operations especially nowadays where technological systems and digital touch points are made available for companies and establishments. Data analysis will only require a least amount of effort among all the other data analysis. It describes the main features of a collection of data, describing such data set. This is typically the first kind of data analysis performed on a set of data. Analyzing business data will significantly benefit an organization to come up with the best marketing strategy for the betterment of your business. It will help you identify factors and collect data that are affecting your business operations.

TABLE OF CONTENTS

1. Introduction

1.1 Introduction

LEVERAGING DATA ANALYTICS TO GROW SALES AND REVENUES

Chief marketing officers (CMOs) across the country are increasingly incorporating big data into their decision-making process. For instance, a recent study has revealed that 42 percent of CMOs make marketing decisions based on customer-acquisition numbers, 40.5 percent based on customer insight, 39.1 percent prioritize digital marketing when making such decisions, 35 percent place greater emphasis on customer retention, and 34.5 percent make marketing decisions based on branding. It is worth noting that 46 percent of the polled marketers said that they would use various analytics strategies to gain consumer insight in 2017. Examples of such strategies include location-based targeting, personalization, and an increase in mobile and real-time reporting.

SOURCES OF BUSINESS DATA

Internal and external sources generate 54 percent and 25 percent of business data respectively. The remaining 21 percent of data comes from a combination of the first two sources. The top four ways business leaders source business data are sales and financial transactions (56 percent), leads and sales contacts from customer databases (21 percent), email correspondence (39 percent), and productivity applications (39 percent). Overall, big data boosts a business's performance, improves customer segmentation and enhances the decision-making process. More specifically, 29 percent of marketers in the US say that marketing analytics has helped them grow their organization's sales revenues by as much as 26 percent. Additionally, 54 percent of companies using customer analytics have seen their profits grow considerably.

THE THREE LEVELS OF ANALYTICS

The three levels of analytics, according to tech authority Gartner, are descriptive analytics, predictive analytics and prescriptive analytics. Descriptive analysis entails examining data and content manually with the aim of understanding what happened. Some of the techniques that a business can employ to do this include business intelligence and visualizations. Predictive analysis, on the other hand, attempts to predict the outcome by employing techniques such as regression analysis, forecasting and predictive modelling. Finally, prescriptive analysis is an advanced form of analytics that aims to find suitable solutions to the problems identified in the first and second levels of analytics. Some of the techniques employed in predictive analytics include complex event processing, simulation and recommendation engines.

THE PROS AND CONS OF UTILIZING DATA ANALYTICS IN MARKETING SECTOR

CONS

One of the main challenges of using market analytics revolves around integrating complex interfaces for accessing data. In fact, only 26 percent of the polled marketers believe that their systems are properly set up to work seamlessly together. The second key challenge revolves around a user's ability to employ analytics data effectively. On this front, only 28 percent of the polled marketers said they were able to do this. The third key challenge has to do with data verification and validation. In particularly, outdated, inconsistent and irrelevant data poses a big problem to 59 percent of the businesses interviewed.

PROS

According to polled US executives, American companies that invest in big-data initiatives enjoy enhanced decision-making, improved collaboration and sharing of information, as well as greater customer satisfaction and retention. This is particularly important because 72 percent of the polled executives reported increased competition for customers. Market analytics gives businesses an edge over their competitors that have failed to invest in big-data initiatives.

1.2 Brief Data Set Review

Dataset

• Source: Kaggle open-source datasets
• License: CC0: Public Domain
• Usability: 9.4
• Size: 817kb
• Dimensions: 2206 x 39
• File type: CSV

1.3 Problem definition

DEFINITION: We're a marketing analyst and we've been told by the Senior Marketing Manager that recent marketing campaigns have not been as effective as they were expected to be. We need to analyse the data set in order to understand this problem and propose data -driven solutions. We are required to solve the following questions to generate a report for your management.

Across various public and private sectors, companies capture and maintain enormous amounts of data on their customers, products, and services they provide. To leverage this technical data stored and maintained in various digital platforms such as databases and data warehouses, and to translate it into actionable insights a new field called Business Analytics (BA) also known as Business Intelligence (BI) or Big Data (BD) has emerged in recent years.

1.4 Strategy

1. **Problem Statement**. What problem(s) does adblocking software present to a firm like GYF in its ability to attract and maintain advertisement-buying customers?
2. Our **Strategy** for addressing the problem.
3. The anticipated **Effects** of strategy.
4. How you will **Measure** the success of the strategy?
5. Are there any null **values or outliers**? How will you **wrangle/handle** them?
6. Are there any variables that **warrant transformations**?
7. Are there any useful variables that you can **engineer** with the given data?
8. Do you notice any **patterns or anomalies** in the data? Can you plot them?
9. What **factors** are significantly related to the number of store purchases?
10. Does US **fare** significantly better than the Rest of the World in terms of total purchases?

11. Your supervisor insists that people who buy gold are more conservative. Therefore, people who spent an above average amount on gold in the last 2 years would have more in store purchases. Justify or refute this statement using an appropriate **statistical test.**

11. Fish has Omega 3 fatty acids which are good for the brain. Accordingly, do "Married PhD candidates" have a significant relation with amount spent on fish? What other factors are significantly related to amount spent on fish? (Hint: use your **knowledge of interaction variables/effects**)

Is there a significant relationship between **geographical regional and success of a campaign**?

Process of Data Analysis

2. Technology and Literature Review

2.1 Business Intelligence

There is another issue with a great number of definitions; they tend to change after some time, in light of the fact that the way of what they consider changes. This is the situation with BI for instance. Initially, software business engaged with BI, BI used to be comprehended as private insight, rather than state or open knowledge. Even after many years, BI is still used by engineers and programmers (Solberg Solien, 2015).

BI is characterized as frameworks that gather, change, and present organized information from various sources lessening the required time to acquire significant business data and enable their efficiency use in management decision making process (Den Hamer, 2004), permitting dynamic enterprise information look, recovery, examination, and clarification of the necessities of administrative choices (Nofal and Yusof, 2013).

As indicated by Tyson (1986), BI concentrates on gathering, process and present information concerning customers, contenders, the business sectors, technology, and products. Pirttimäki (2007) depicts BI as a procedure that incorporates a series of activities, being driven by the particular data needs of decision makers and the objective of achieving competitive advantage.

BI is a framework that transforms information into data and afterward into learning, consequently enhancing company's basic decision-making process (Singh and Samalia, 2014). BI is characterized as a framework which gathers, changes and shows organized information from various sources. BI is a system and an answer that helps decision makers to comprehend the economic circumstance of the firm (Nofal et al., 2013).

BI is termed to as a set of numerical and methodological models for examination utilized for extracting data and valuable information from raw information for utilizing confused basic leadership prepare (Vercelli's, 2013). Similarly, Wixom and Watson (2010, p.14) mention that ―Business intelligence (BI) is a broad category of technologies, applications, and processes for gathering, storing, accessing, and analysing data to help its users make better decisions.

We can upgrade the bits of knowledge gave by BI applications—particularly by utilizing information mining procedures, through simulation and modelling of real world under a "systems thinking" approach, enhancing forecasts, and adding to a superior comprehension of the business progression of any organization (Raisinghani, 2004).

BI helps administrators by breaking down information from various resources in better basic leadership at both tactical and strategic level, for customary utilization, conventional data frameworks farewell, yet for hierarchical and functional planning; new tools are required for business analysis (Rasoul and Mohammad, 2016).

2.2 Data, Information, and Knowledge

In BI context, we always see the word data, information, and knowledge which could lead us getting confused on its use and implication. Carlo (2009) distinguishes their definition.

Data: It refers to a structured codification of single primary entities and as well as of transactions involving two or more primary entities Carlo (2009). BI is popular among companies mainly because of analysis of data that is of any form and formulate a strategy accordingly. Generally, data is classified into three types—structured data, semi-structured data, and unstructured data.

Structured data are information that is fixed form, the data may be a collection of forms of websites, and detailed address that can be easily read by the computers since the data is already standardized.

Unstructured data are information that cannot be easily read by computers, which may be text, documents, video tapes, websites, and pictures (Jermol et al. 2003), or any other type of information that cannot be clearly sorted or organized into rows and columns.

Information is used many times to Company data are found across different locations and places in the form of Customer Relation Management (CRM) programs, marketing automation systems and social media platforms.

Information: It refers to the result of extraction and processing activities carried out on data, and it appears meaningful for those who receive it in a specific domain.

Knowledge: It is formed from information which is used to make decisions and develop the corresponding actions. Hence, we could say that knowledge consists of information that puts to work into a specific domain, and it is enhanced by the experience and competence of decision makers in tackling and solving complex problems.

2.3 Business Intelligence Architectures

Carlo (2009) uses the following pyramid to describe how business intelligence system is constructed.

Data sources: The sources mostly consist of data belonging to operationalize systems, but may also include unstructured data, such as emails, and data received from external providers.

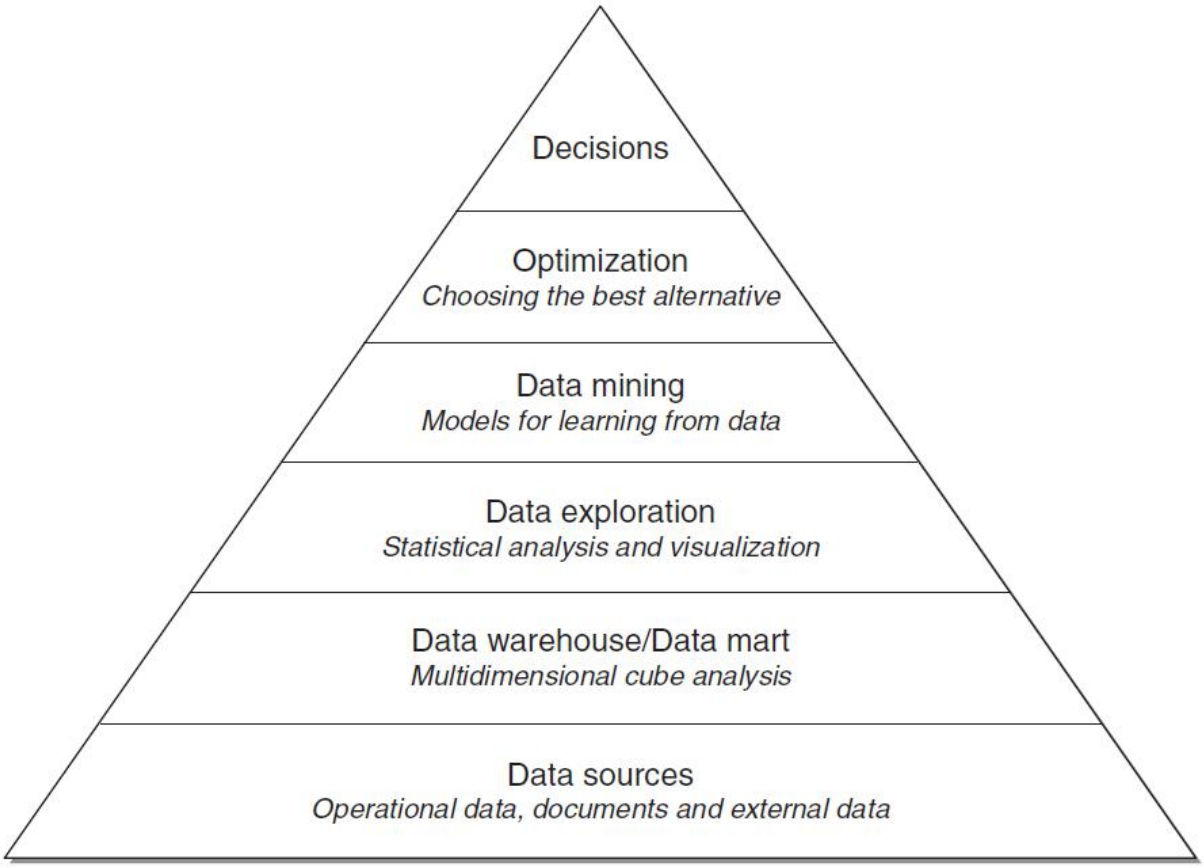
Data warehouse/Data mart: Data warehouses are used to consolidate different kinds of data into a central location using a process known as extract, transform and load (ETL) and standardize these results across systems that are allowed to be queried. Data marts are generally small warehouses that focus on information on a single department, instead of collecting data across a company. They limit the complexity of databases and are cheaper to implement than full warehouses.

Data exploration: Data exploration is a passive BI analysis consisting of query and reporting systems, as well as statistical method.

Data mining: Data mining is active BI methodologies with the purpose of information and knowledge extraction from data.

Optimization: Optimization model allows us to determine the best solution out of a set of alternative actions, which is usually fairly extensive and sometimes even infinite.

Decisions: When business intelligence methodologies are available and successfully adopted, the choice of a decision pertains to the decision makers, who may also take advantage of informal and unstructured information available to adapt and modify the recommendations and the conclusions achieved through the use of mathematical models.
Figure: The main components of a Business Intelligence System (Carlo, 2009:10)



2.4 Business Intelligence Capabilities

One underlying theme that is evident through the research is that BI used in an organization should be suited for decision making, which in turn contributes to BI success (Clark, Jones & Armstrong, 2007). However, many scholars gained that this success is yet to be realized by many organizations (Hostmann, Herschel, & Rayner, 2007). BI capacities are basic capacities that help organizations enhance both its adjustment to change and its execution (Watson & Wixom, 2007).

Many researchers state that failure in adopting BI in an organization because of an absence of fit between organization's BI and its characteristics and objectives. An organization that has made progress with their BI usage have attempted to guarantee that their BI is steady with their corporate business targets and much research on BI achievement concentrates on the alignment amongst BI and business targets (McMurchy, 2008).

However, little is known about the part BI abilities play in accomplishing this objective. In-spite, the fact that there is a collection of research tending to BI abilities, it has remained to a great extent quiet on the part of BI capacities in accomplishing the important match amongst BI and the decision environment in which it is implemented.

According to Oyku et al. (2012), BI can be examined from both organizational and technological views.

Technological BI capabilities are referring to the data quality (data standard), technical platforms that could be integrated with other systems in the organization and user access. Organizational BI is the assets supporting the BI application that runs in the organization such as flexibility and shared risks and responsibilities (Ross, Beath and Goodhue, 1996).

Data Quality

BI has largely relied on numerical and/or structured data, which can be measured on a numerical scale and analysed with statistical methods and computing equipment (Isik et al. 2013, p.14). Ponniah (2001) stated that data quality is the most important element leading to BI success.

Similarly, Kimball et al. (2008) also stated that the data quality is the most important factor, and they added that the massive data from many different sources of a large enterprise can be integrated into a coherent body to provide a clear view of its business, therefore, meaningful information can be delivered at the right time, in the right location, and in the right form to assist individuals, departments, divisions or even larger units to facilitate improved decision making.

Data quality refers to the data which is consistent and comprehensive. Poor data reliability is because of poor data handling processes, poor data maintenance procedures, and errors in the migration process from one system to another. If the information that we collect is not accurately or consistently analysed, organizations cannot satisfy their customers' expectations nor keep up with new information-centric regulations.

According to Oyku et al. (2012), in order to improve the business agility, the organization should develop the technological ability that could deliver accurate, consistent and timely information to its users.

Moreover, clean and relevant data are one of the most essential factors of BI success. As companies incorporate data from a wider variety of sources, they will continue to face new and ever-increasing issues surrounding the quality of the data on which they rely.

Integration with other systems

Since BI system is a new system for organization, the integration between BI system and other systems in the organization is another crucial activity behind the BI success. The integrating activity is involving with the connection between various systems and their application or data together, either physically or functionally, thus each individual system can create and provide value to the organization (White, 2005).

Furthermore, the organization using data from multiple sources and feeding the data into multiple information systems, the performance of integration will be affected directly by the quality of the communication between these systems (Oyku, 2012).

User access

BI tools according to Oyku et al. (2012) have different capabilities and serve different purposes so that one size does not fit with all BI. Whether the organization prefers to use a single BI suite or best-of-breed applications, it is essential to match tool capabilities with user types. While some organizations limit user access through practicing authorization/authentication and access control, others prefer to allow full access to all types of users through a web-centric approach.

It is critical that organizations achieve the necessary balance to allow the way BI users access information to fit the types of decisions they make using BI.

Flexibility

In order to achieve the competitive advantages provided by BI, organizations must consider carefully on selecting the underlying technology to support BI and also be flexible with the strictness of the business process rules and regulations since flexibility is one of the key factors to run BI successfully in the organization (Oyku et al. 2012).

Risk Management Support

Risk management is one of the major supports in BI, as it helps in decision making, where the conditions tend to be uncertain, for example, when all the factors are known (Harding, 2003). Risk management is crucial for organizations that operate in high-risk environments, as well as, it is important for organizational success (Davenport, 2006).

Despite, hazard, and instability exist in each business decisions, and organizations may utilize BI to limit vulnerability and settle on better choices. The impact of BI in decision-making capabilities affects its success.

According to Alaskar and Efthimios (2015), not all of BI solutions succeed in all organizations, and, there are signs, before a project begins, that could indicate whether the project will succeed, struggle, or fail and it is essential that organizations are aware of the key indicators of success in adopting BI, so as to overcome the challenges or risks that are associated with the BI project during its implementation.

2.5 Enabling factors in business intelligence projects

Some factors such as technologies, analytics and human resources that Carlo (2009) mentions are more critical than others to the success of a BI project.

Technologies

The crucial enabling factors that have facilitated the development of BI systems in the complex organization and enterprise are hardware and software technologies. This pattern has empowered the utilization of advanced processes which are required to utilize inductive learning strategies and enhancement models, keeping the processing times inside a sensible range.

Additionally, it allows the appropriation of best-in-class graphical perception strategies, featuring real-time animations. A further important factor gets from the exponential increment in the limit of mass storage's, again at low costs, enabling any organization to store terabytes of information for business insight analysis.

What's more, system network, as Extranets or Intranets, has played an essential part in the diffusion inside organizations of data and learning separated from BI. Finally, the simple integration of hardware and software obtained by various providers, or grew inside by an organization, is another factor influencing the diffusion of data analysis of tools.

Analytics

Mathematical model and analytical methodologies play an important role in information advancement and knowledge taking out from the accessible data inside most organizations. The mere visualization of the data according to timely and flexible logical views, plays a relevant role in facilitating the decision-making process, but still, represents a passive form of support. Hence, it is essential to apply more advanced models of inductive learning and optimization in order to achieve active forms of support for the decision-making process.

Human resources

The human resources of an organization are built up by the competencies of those who operate within its boundaries, whether as individuals or collectively. When employees possess the ability of knowledge that could acquire information and then translate it into the practical way, they will have a major influence on the quality of decision-making process. The organization must emphasize the personal skills of its knowledge workers to work out creative solutions and to devise effective action plan if it implements an advanced BI system. Every company could access to available analytical tools equally, but if a company wants to have the competitive advantage over its competitors, it should employ human resources endowed with a greater mental agility and willing to accept changes in decision-making style.

3. System Requirements Study

3.1 Hardware and Software Requirements

3.1.1 Hardware Requirements

- **For Running Application**
  - o Intel 3rd gen or higher
  - o 1gb ram
  - o 50mb of free HDD space
  - o Intel HD 4000
- **For Building Application**
  - o Intel 3th gen or higher
  - o 4gb ram
  - o 200mb of free HDD space

3.1.2 Software Requirements

- **For Running Application**
  - o Operating system: windows 7 or higher (32bit or 64bit)
  - o Microsoft Visual C++ Redistributable 2012 (x86)

- **For Building Application**

    - Operating system: windows 7 and higher (64bit)
    - Microsoft Visual C++ Redistributable 2012 (x86)

3.2 Assumption and Dependencies

3.2.1 Assumption

- User's CPU or GPU will support at least OpenGL 3.0

3.2.2 Dependencies

- Windows SDK
- MS Mouse Pointers

4. System Analysis

4.1 Study of Current System:

Some benefits of data analysis are:

• **Delivering Relevant Products**: Products are the life-blood of any organization and often the largest investment companies make. Effective data collation combined with analytics will help companies stay competitive

• **Personalization and Services**: Being able to react in real time and make the customer feel personally valued is only possible through advanced analytics. Big data offers the opportunity for interactions to be based on the personality of the customer, by understanding their attitudes and considering factors such as real-time location to help deliver personalization in a multi-channel service environment.

• **Optimizing and Improving Customer Experience**: Advanced analytical techniques can be deployed to improve field operations productivity and efficiency as well as optimize an organizational workforce according to business needs and customer demand.

4.2 Limitations of Current Techniques:

- Data collection for small companies can be expensive for small areas, particularly for one time analysis.
- **The data could be incomplete**. Missing values, even the lack of a section or a substantial part of the data, could limit its usability.
- **Data collected from different sources can vary in quality and format**. Data collected from such diverse sources as surveys, e-mails, data-entry forms, and the company website will have different attributes and structures.

- **Lack of commitment and patience**. Analytics solutions are not difficult to implement, however, they are costly, and the ROI is not immediate. Especially, if existing data is not available, it may take time to put processes and procedures in place to start collecting the data.
- **Privacy concerns**. Sometimes, data collection might breach the privacy of the customers as their information such as purchases, online transactions, and subscriptions are available to companies whose services they are using. Some companies might exchange those datasets with other companies for mutual benefit.
- **Complexity & Bias**. Some of the analytics tools developed by companies are more like a black box model. What is inside the black box is not clear or the logic the system uses to learn from data and create a model is not readily evident. For example, a neural network model that learns from various scenarios to decide who should be given a loan and who should be rejected.

5. System Design

Section 01: Exploratory Data Analysis

1. Are there zero values or outliers? How are you going to fight / deal with them?
2. Are there variables that justify the transformations?
3. Is there a useful variable that I can build with the given data?
4. Do you see patterns or anomalies in the data? Can you trace them?

Section 02: Statistical Analysis

Perform statistical tests in the form of regressions to answer these questions and suggest data-driven recommendations for action to your CMO. Be sure to interpret your results using non-statistical jargon so your CMO can understand your results.

1. What factors are largely related to the amount of purchases in the store?
2. Is the United States doing much better than the rest of the world in terms of total purchases?
3. His boss insists that people who buy gold are more conservative. Therefore, people who have spent above average on gold in the last 2 years would buy more from the store. Justify or refute this claim with an adequate statistical test
4. Fish has omega-3 fatty acids that are good for the brain. Does this mean that "married PhD students" have a significant connection to spending on fish? What other factors are significantly related to the amount spent on fish?
5. Is there a significant correlation between geographic region and the success of a campaign?

5.1 Exploratory Data Analysis

```
# linear algebra
import numpy as np

# data processing
import pandas as pd

# data visualization (for EDA)
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
plt.style.use('ggplot')
sns.set(color_codes=True)

#ignore warnings
import warnings
warnings.filterwarnings('ignore')
import datetime


# Importing sklearn methods
#from sklearn import linear_model
#from sklearn.ensemble import RandomForestRegressor
#from sklearn.metrics import mean_squared_error
#from sklearn import svm
#from sklearn.ensemble import GradientBoostingRegressor
#from sklearn.model_selection import cross_val_score
#from sklearn import model_selection
#from sklearn.model_selection import GridSearchCV
#from sklearn.model_selection import train_test_split

# import labelencoder
from sklearn.preprocessing import LabelEncoder

# the spearman's correlation between two variables
from scipy.stats import spearmanr
```

```
from google.colab import files
uploaded = files.upload()

df=pd.read_csv("marketing_data.csv")
Df.shape
```

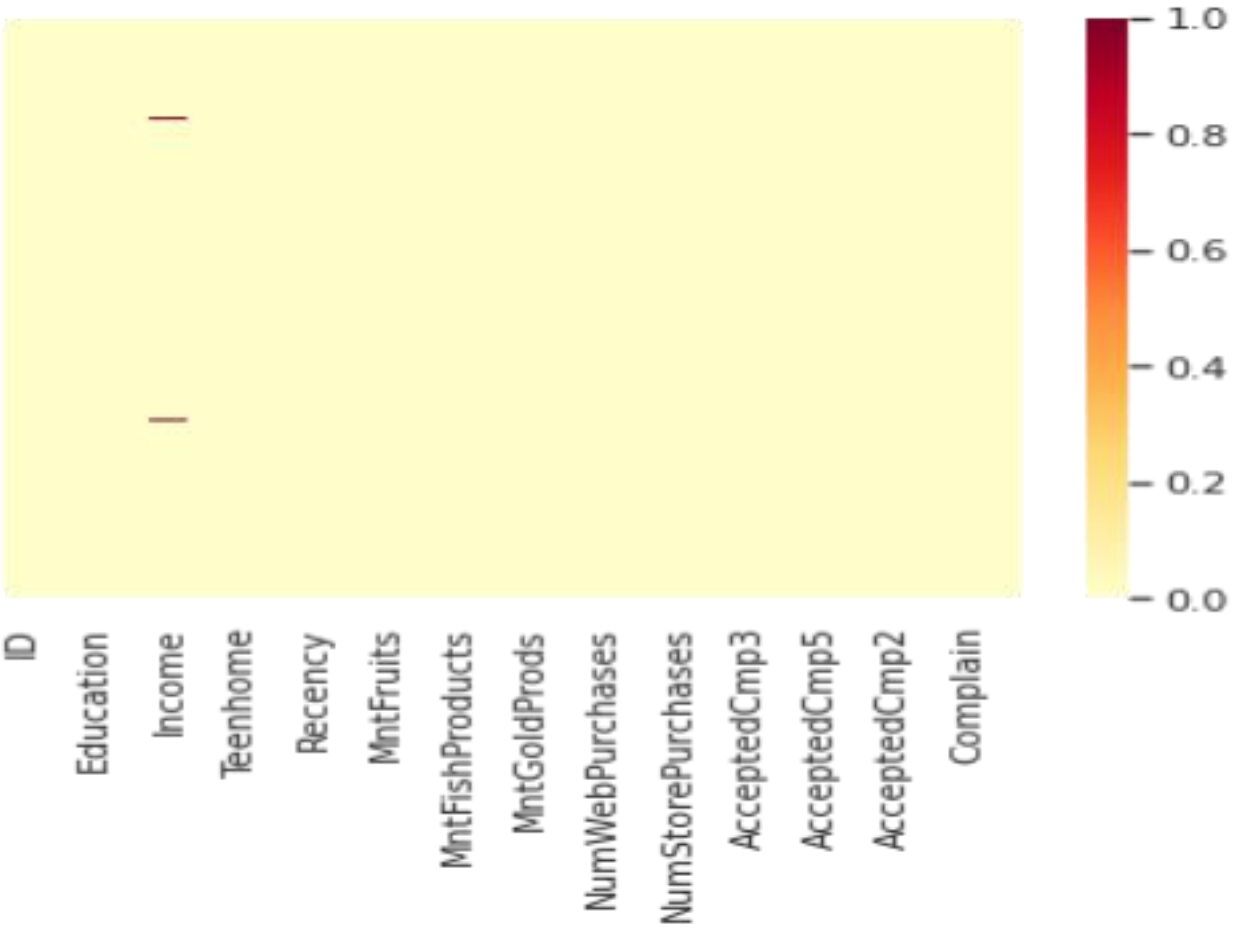
Cleaning Data

1. Income column change to numeric
2. Dt_Customer has string data type; we have to change it type to date

```
df.rename({' Income ':'Income'}, axis=1, inplace=True)
df['Income'] = df['Income'].str.replace('$','').str.replace(',','')
.astype(float)
```

5.1.1 Are there any null values or outliers? How will you wrangle/handle them?

```
df.head(3)
#null values
sns.heatmap(df.isnull(),yticklabels=False,cmap='YlOrRd');
```



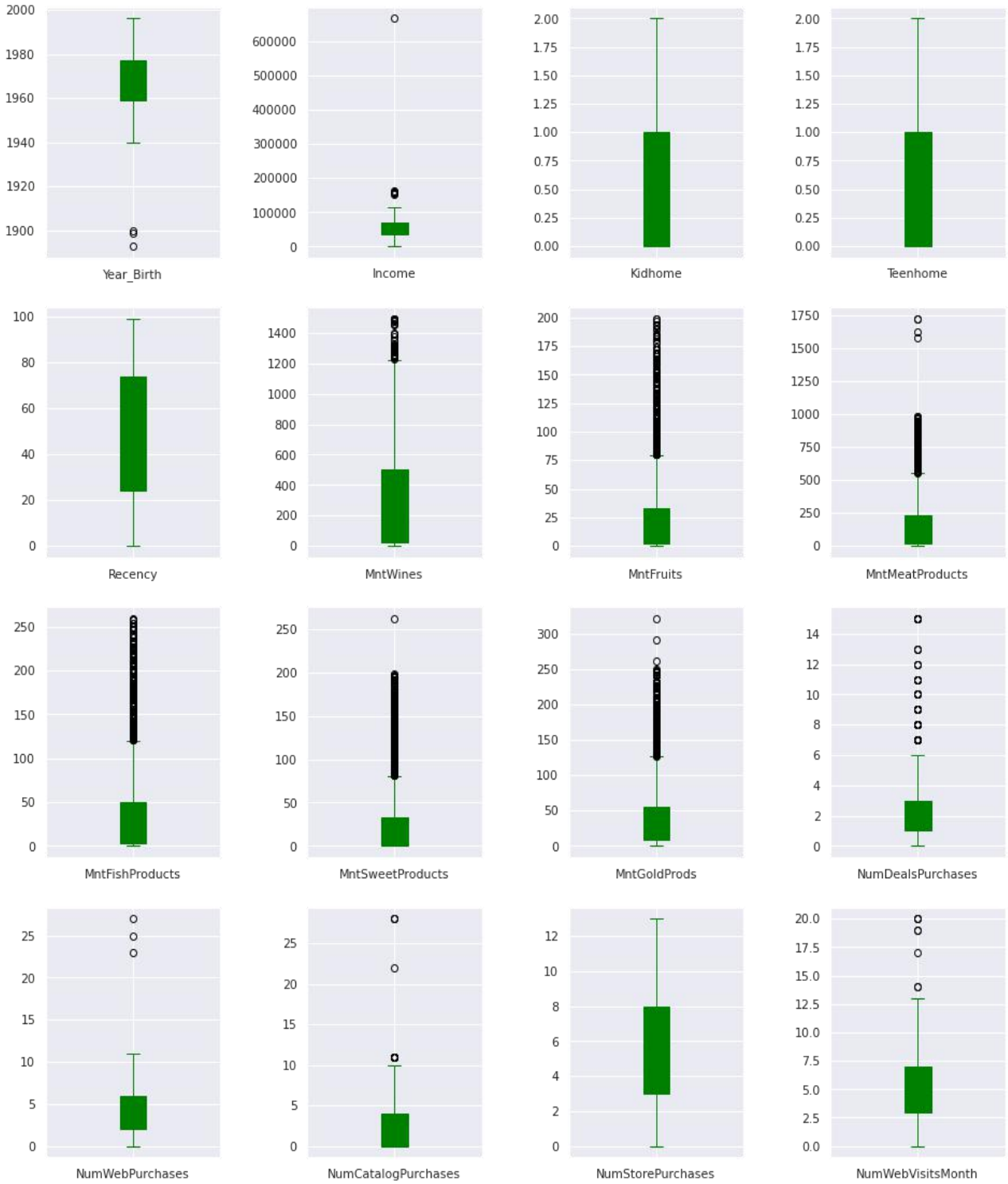
As we can see that from the above plot, we have null values in Income column

```
#We can see that Income has 24 null values so we drop them
df = df[df['Income'].notna()]
df.columns[df.isnull(). any()].tolist()
```

Outliers & Anomalies

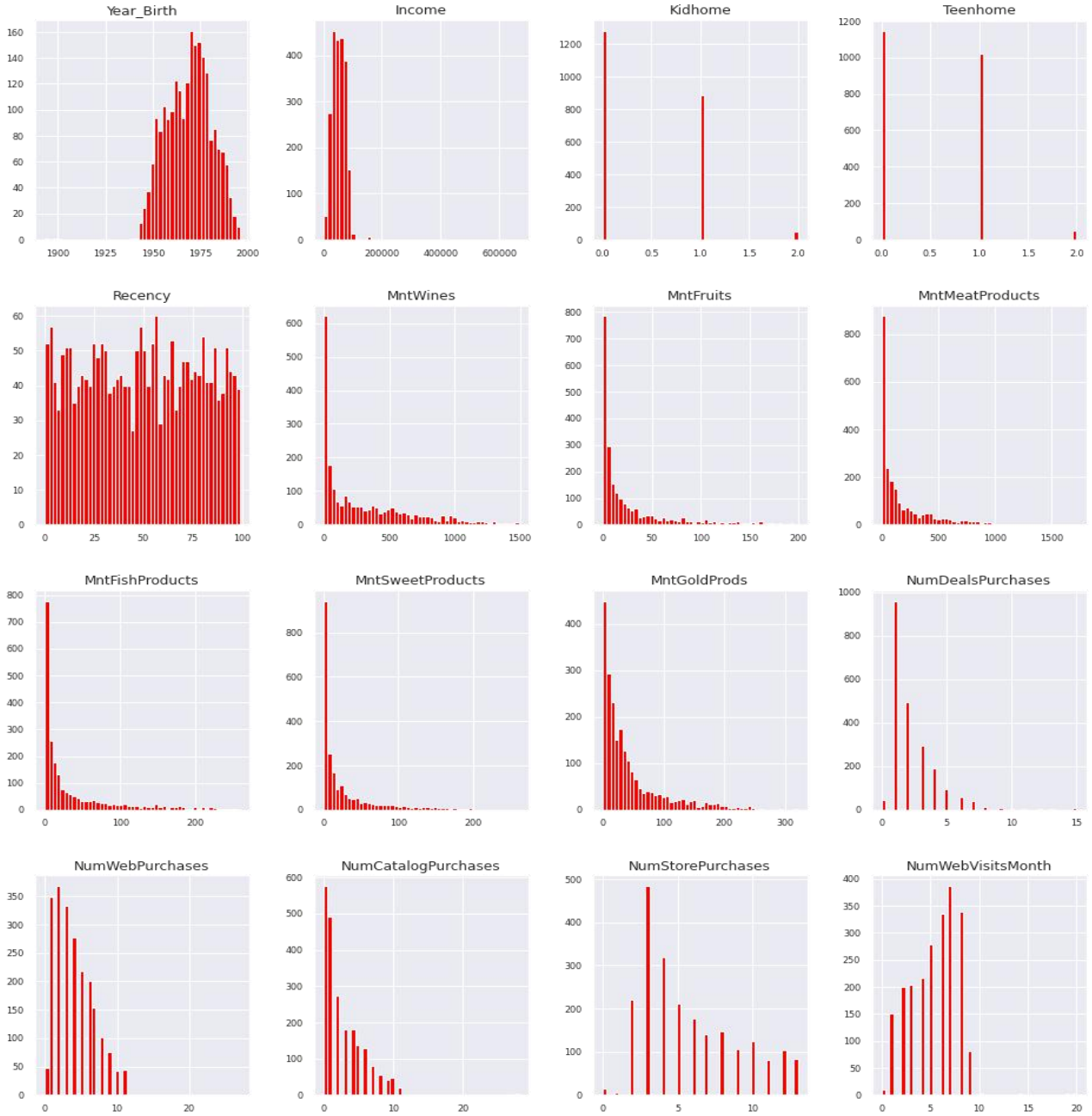
From the graphs, it is clear that multiple features contain outliers but income and births may indicate data entry error

```
import matplotlib.pyplot as plt
list(set(df.dtypes.tolist()))
df_num = df.drop(columns=
['ID', 'AcceptedCmp1', 'AcceptedCmp2', 'AcceptedCmp3', 'AcceptedCm
p4', 'AcceptedCmp5', 'Response', 'Complain']).select_dtypes(includ
e = ['float64', 'int64'])

df_num.plot(subplots=True, layout=
(4,4), kind='box', figsize=(16,18), patch_artist=True,color="Green
")
plt.subplots_adjust(wspace=0.5);
```


Year_Birth
Income
Kidhome
Teenhome
Recency
MntWines
MntFruits
MntMeatProducts
MntFishProducts
MntSweetProducts
MntGoldProds
NumDealsPurchases
NumWebPurchases
NumCatalogPurchases
NumStorePurchases
NumWebVisitsMonth

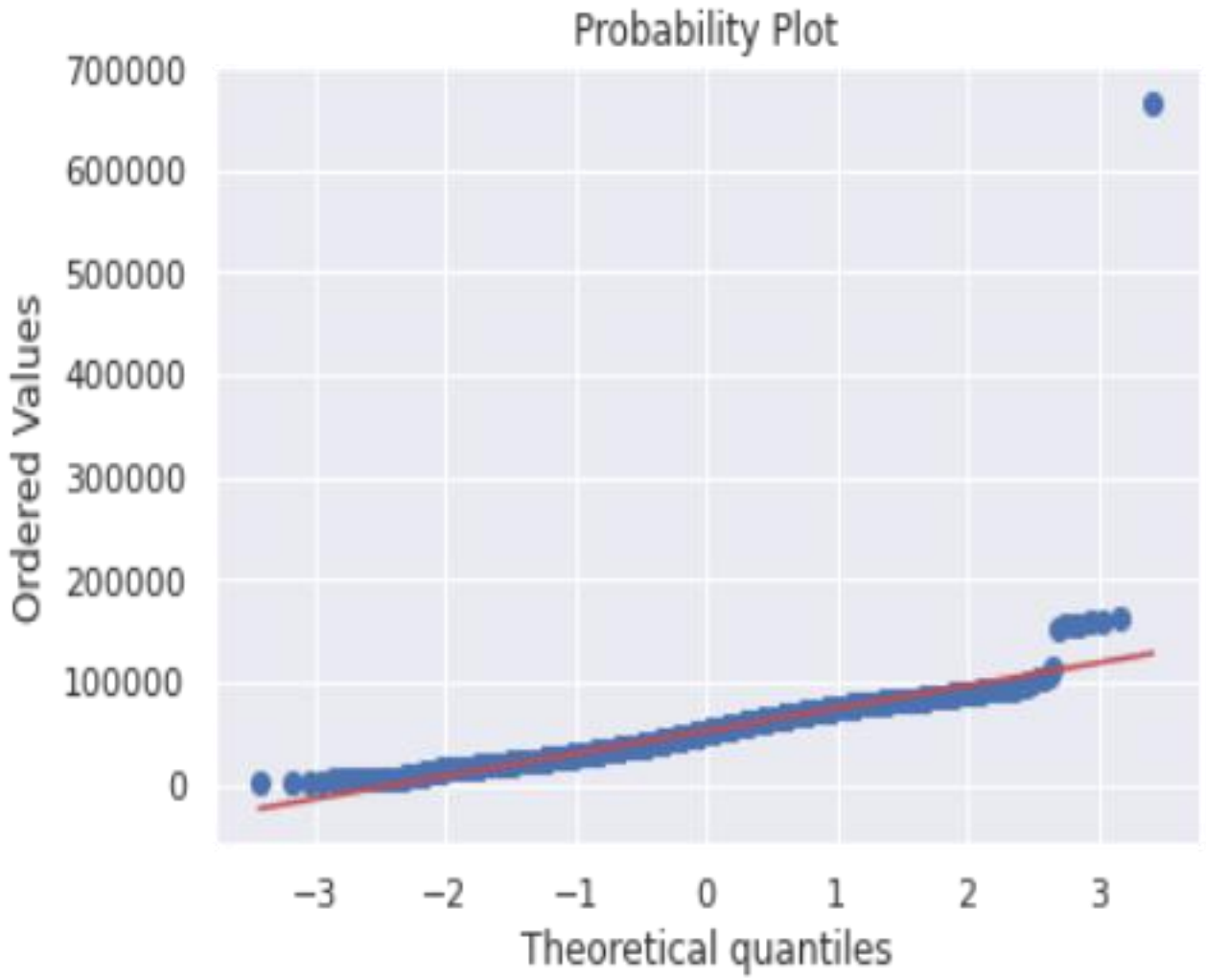
Numeric Data Distribution

```
df_num.hist(figsize=(16, 20), bins=50, xlabelsize=8, ylabelsize=8,
color="Red");
```



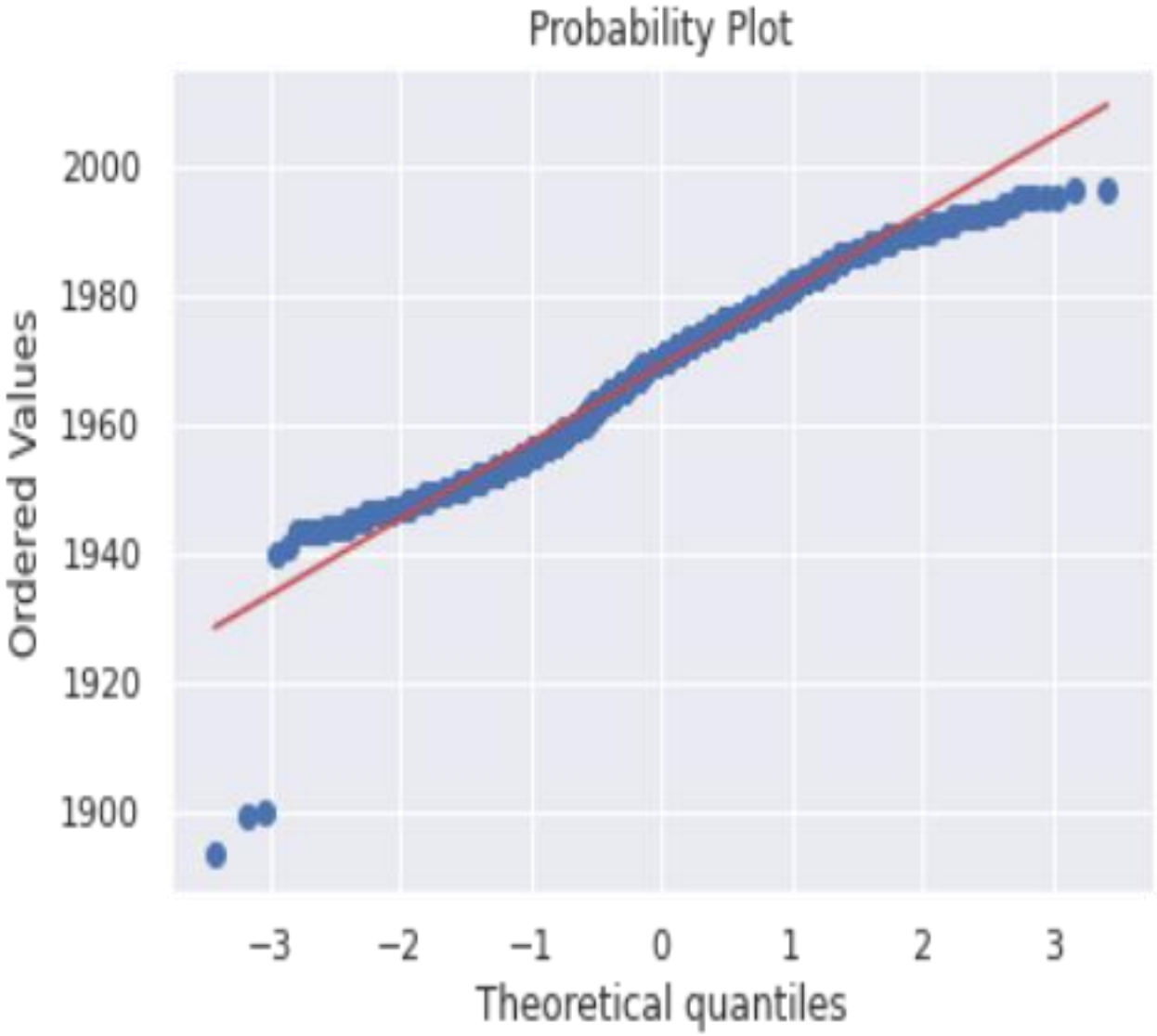
Handling Outliers

```
from scipy import stats
import seaborn as sns
stats.probplot(df['Income'], plot=sns.mpl.pyplot);
```



From the above plot we see that one person has an income of over 600,000 which is an anomaly. Since it's a single record, I'll simply delete it.

```
stats.probplot(df['Year_Birth'], plot=sns.mpl.pyplot);
```



Finding: The Year_Birth column contains three anomalies, so we'll simply drop these records

```
df.Marital_Status.value_counts()
Married     857
Together    573
Single      471
Divorced    232
Widow        76
Alone         3
Absurd        2
YOLO          2
Name: Marital_Status, dtype: int64
```

We can see that the martial_status has outliers (alone, absurd, Yolo) as there are only seven records. Therefore, we will simply exclude these outliers from our data.

```
df = df[~df['Marital_Status'].isin(['Absurd', 'Alone', 'YOLO'])]
```

```
df = df[df['Year_Birth'] > 1910].reset_index(drop=True)
df = df[df['Income'] < 600000].reset_index(drop=True)
```

Are there any useful variables that you can engineer with the given data?

Feature Engineering

1. With the help of given features, we can drive some important variables like:
2. The total number of children in the household can be calculated from the sum of “Kid home” and “Teen home”.
3. The total amount of the expense can be calculated from the sum of all the characteristics that contain the keyword Mnt.
4. The total number of purchases is calculated from the sum of all the characteristics that contain the keyword "purchases".
5. The total number of accepted campaigns is calculated from the sum of all the characteristics that contain the keyword "Cmp".
6. From Dt_Customer we can find the year in which we became a customer
7. From Year_Birth we can derive the age

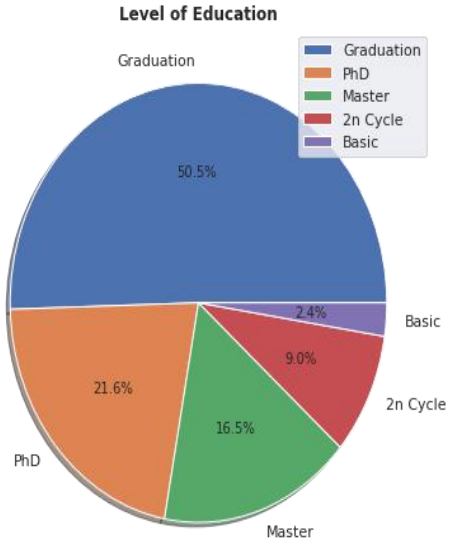
```
#Total kids
df['Totalkids'] = df['Kidhome'] + df['Teenhome']

#
df['YearCustomer'] = pd.DatetimeIndex(df['Dt_Customer']). year


# total amount spent
mnt_cols = [col for col in df.columns if 'Mnt' in col]
df['TotalMnt'] = df[mnt_cols].sum(axis=1)

# Total Purchases
purchases_cols = [col for col in df.columns if 'Purchases' in col]
df['TotalPurchases'] = df[purchases_cols].sum(axis=1)

# Total Campaigns Accepted
campaigns_cols = [col for col in df.columns if 'Cmp' in col] + ['R
esponse']
df['TotalCampaignsAcc'] = df[campaigns_cols].sum(axis=1)

#Age
year=datetime.datetime.today(). year
df['Age'] =year-df['Year_Birth']

#Age_groupe
bins= [18,39,59,90]
labels = ['Adult','Middle Age Adult','Senior Adult']
df['AgeGroup'] = pd.cut(df['Age'], bins=bins, labels=labels, right
=False)
df['AgeGroup'] = df['AgeGroup'].astype('object')
```

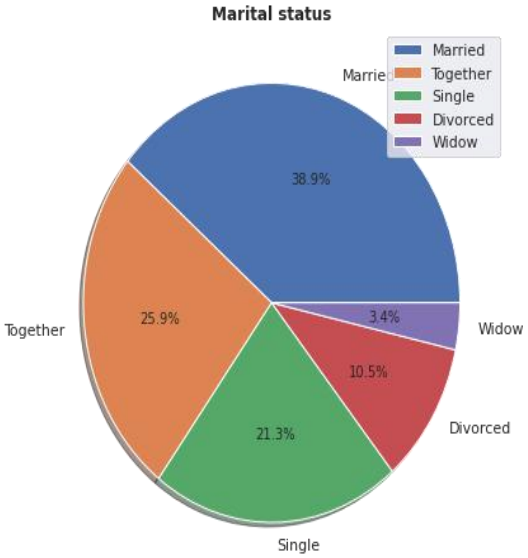
5.1.2 Do you see any patterns or anomalies in the data? Can you trace them?

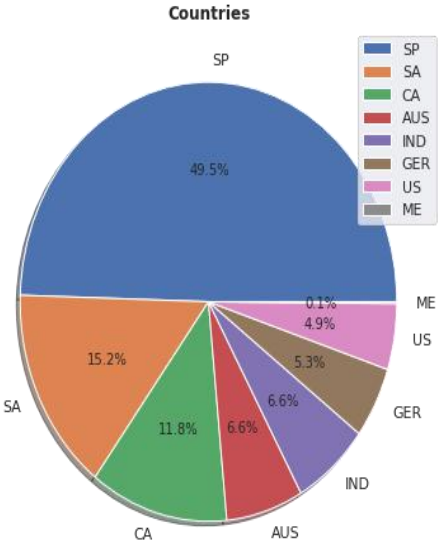
Findings

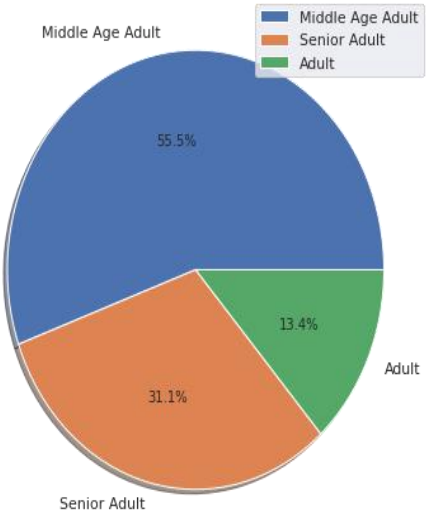
1. Almost 50% of clients' education levels have a college degree and few clients have a primary education.
2. The number of married clients is higher than that of widowers and divorcees.
3. There is a remarkably high percentage of clients in Spain, while the percentage of clients in the United States and Montenegro is very low.
4. There is a very high percentage of customers between 39 and 59 compared to other age groups.

```
f,ax=plt.subplots(2,2, figsize=(20,15))


df['Education'].value_counts().plot.pie(autopct='%1.1f%%',ax=ax[0]
[0], shadow=True,legend=True)
ax[0][0].set_title('Level of Education',fontweight ="bold")
ax[0][0].set_ylabel('')
df['Marital_Status'].value_counts().plot.pie(autopct='%1.1f%%',ax=
ax[0][1], shadow=True,legend=True)
ax[0][1].set_title('Marital status',fontweight ="bold")
ax[0][1].set_ylabel('')
df['Country'].value_counts().plot.pie(autopct='%1.1f%%',ax=ax[1][0
, shadow=True,legend=True)
ax[1][0].set_title('Countries',fontweight ="bold")
ax[1][0].set_ylabel('')
df['AgeGroup'].value_counts().plot.pie(autopct='%1.1f%%',ax=ax[1][
1], shadow=True,legend=True)
ax[1][1].set_title('Age Group',fontweight ="bold")
ax[1][1].set_ylabel('');
```


Level of Education
Graduation
PhD
Master
2n Cycle
Basic
50.5%
21.6%
16.5%
9.0%
2.4%
Graduation
PhD
Master
2n Cycle
Basic


Marital status
Married
Together
Single
Divorced
Widow
38.9%
25.9%
21.3%
10.5%
3.4%
Together
Single
Divorced
Widow


Countries
SP
SA
CA
AUS
IND
GER
US
ME
49.5%
15.2%
11.8%
6.6%
6.6%
5.3%
4.9%
0.1%
SP
SA
CA
AUS
IND
GER
US
ME

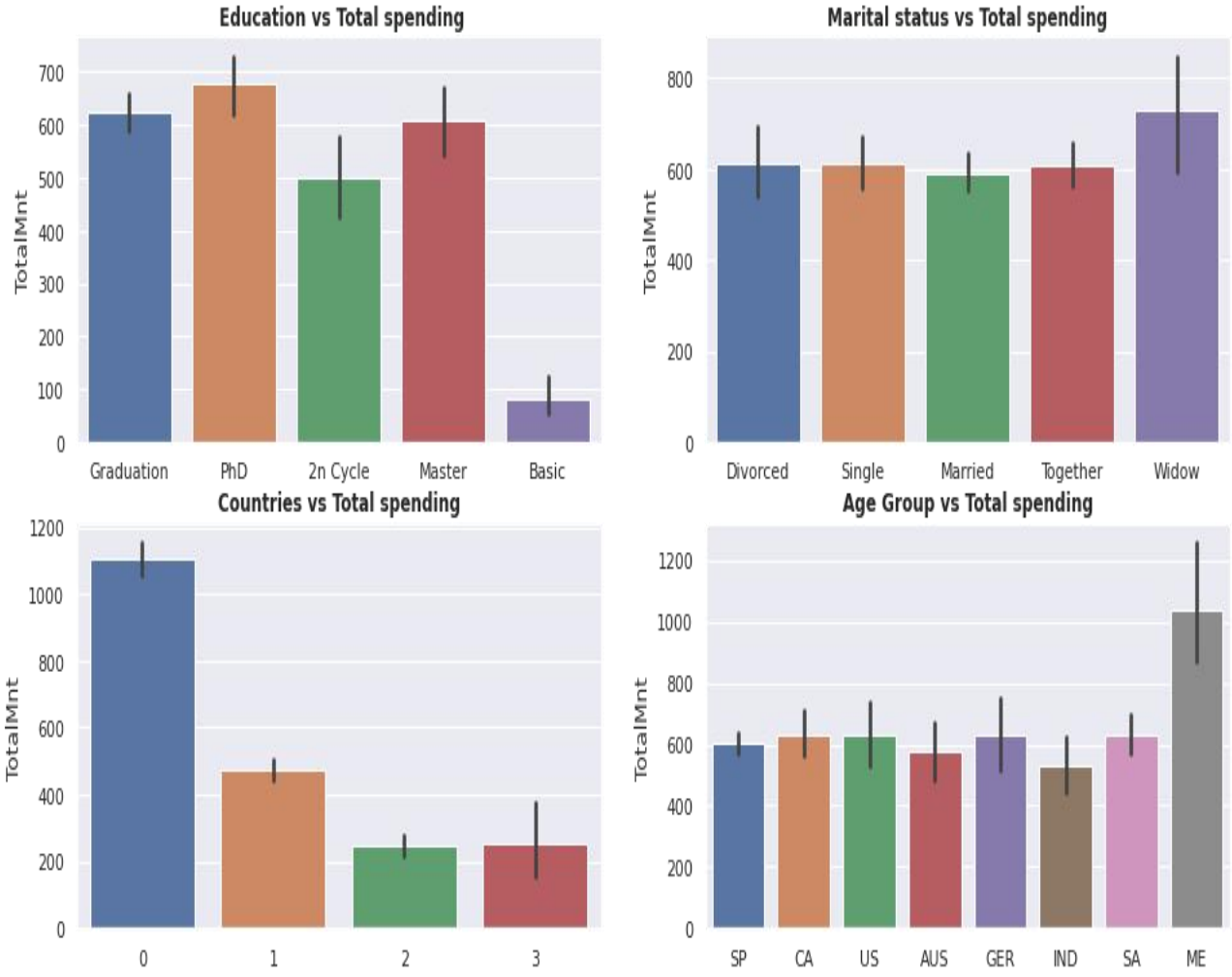

Age Group
Middle Age Adult
Senior Adult
Adult
55.5%
31.1%
13.4%
Middle Age Adult
Senior Adult
Adult

Total spending

Findings

1. Doctors previously spent more than other groups of people
2. Total expenses for divorced, single and married group members are roughly the same, while the expenses for widows are slightly higher than for these people.
3. People without children spend more money than people with children
4. Montenegro spends significantly more than other countries.

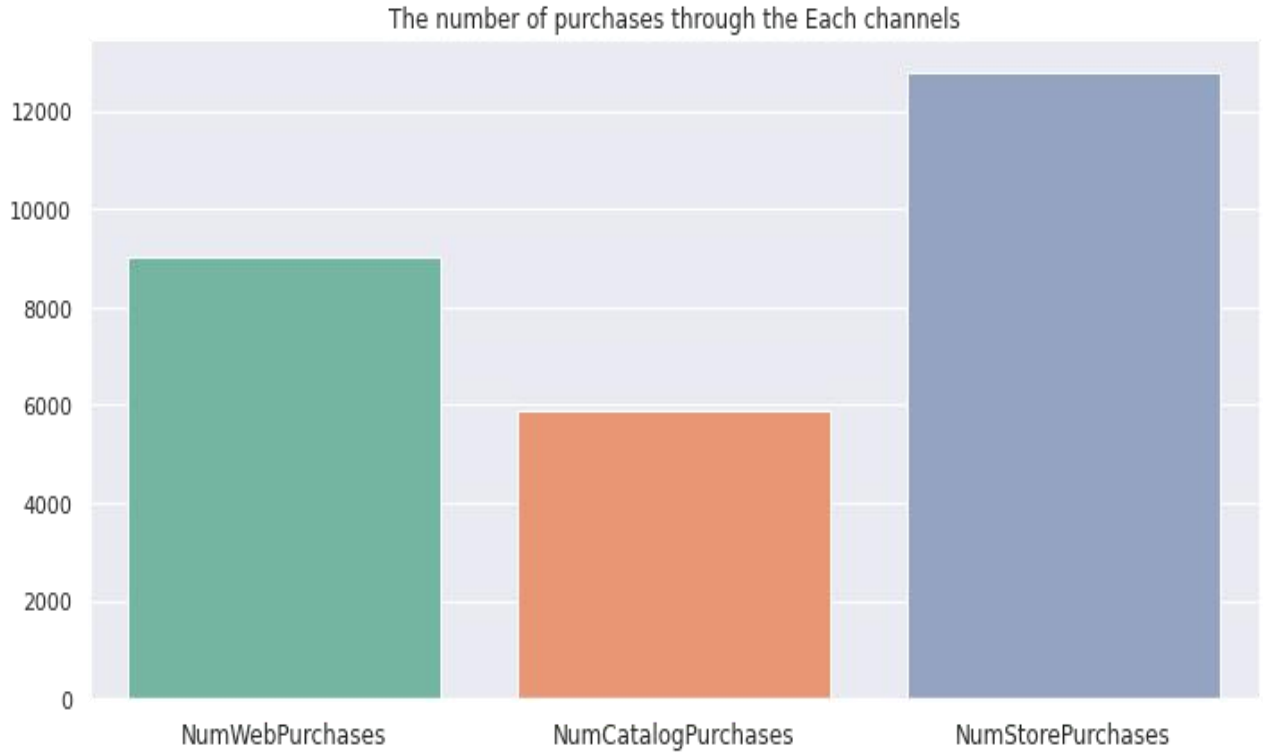
```
f,ax=plt.subplots(2,2, figsize=(16,8))
sns.barplot(x='Education', y='TotalMnt', data=df,ax=ax[0][0]);
ax[0][0].set_title(' Education vs Total spending',fontweight ="bol
d")
ax[0][0].set_xlabel('')
sns.barplot(x='Marital_Status', y='TotalMnt', data=df,ax=ax[0][1]);
ax[0][1].set_title('Marital status vs Total spending',fontweight =
"bold")
ax[0][1].set_xlabel('')
sns.barplot(x='Totalkids', y='TotalMnt', data=df,ax=ax[1][0]);
ax[1][0].set_title('Countries vs Total spending',fontweight ="bold
")
ax[1][0].set_xlabel('')
sns.barplot(x='Country', y='TotalMnt', data=df,ax=ax[1][1]);
ax[1][1].set_title('Age Group vs Total spending',fontweight ="bold
");
ax[1][1].set_xlabel('');
```



The number of purchases through the Each channel

Plot represent that most customer buy product from store.

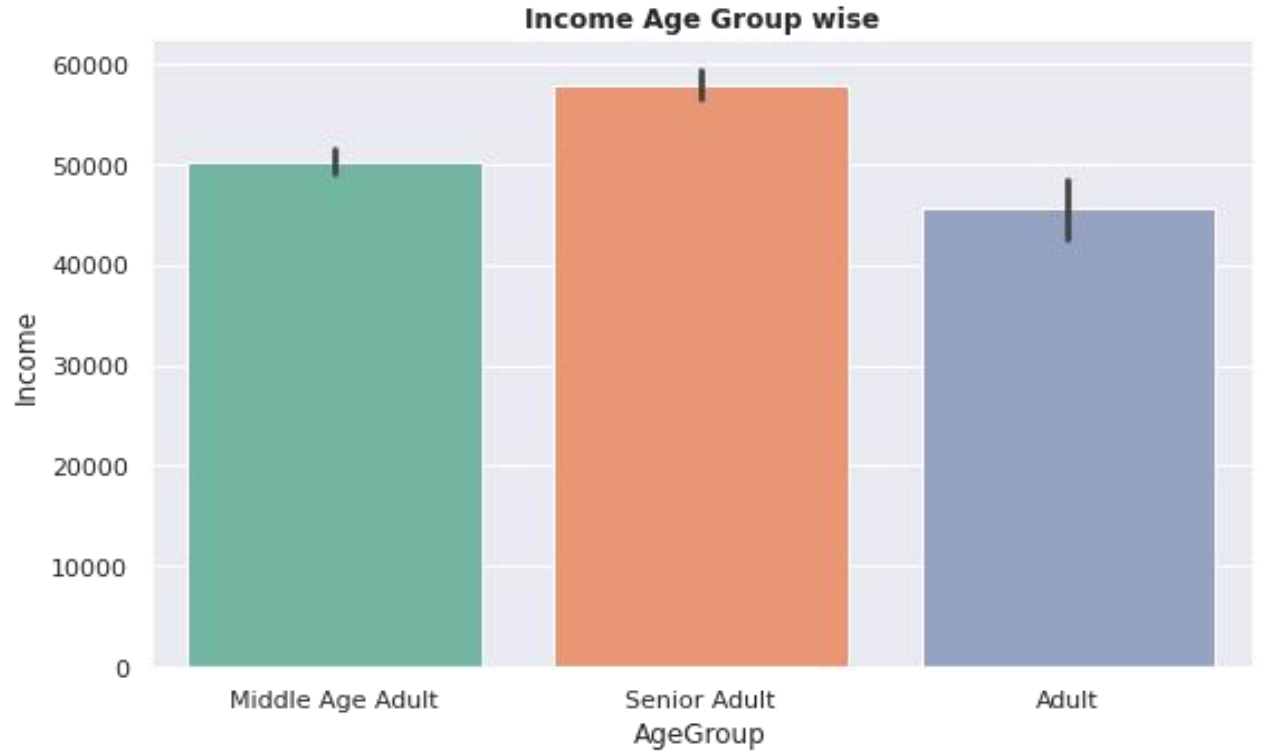
```
channels = ['NumWebPurchases', 'NumCatalogPurchases',  'NumStorePu
rchases']
data = df[channels]. sum ()
plt.figure(figsize=(10,5))
plt.title('The number of purchases through the Each channels')
x=sns.barplot(x=channels,y=data.values,palette='Set2')
x.set_xticklabels(channels, size=12)
plt.tight_layout();
```



The Total Amount of each product spent

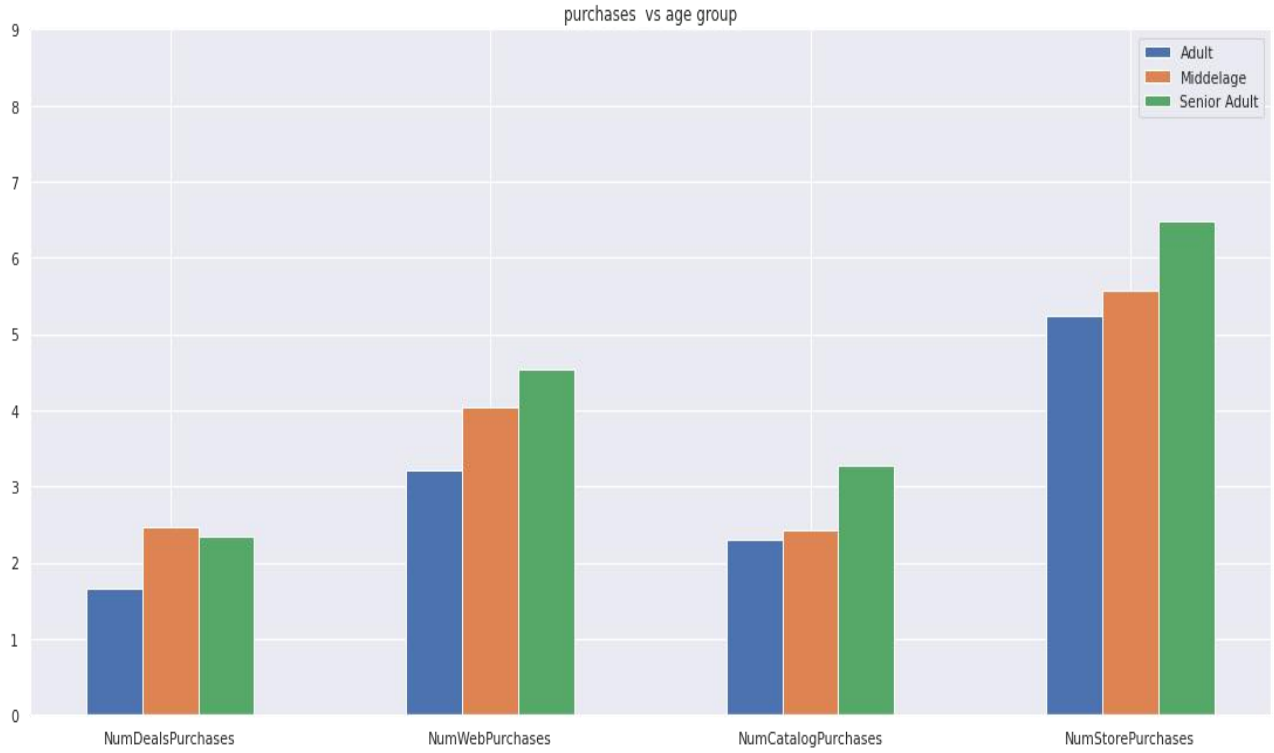
```
col_products = ['MntWines', 'MntFruits', 'MntMeatProducts', 'MntFi
shProducts', 'MntSweetProducts', 'MntGoldProds']

data = df[col_products].sum()
plt.figure(figsize=(15,5))
plt.title('The Total Amount of each product spent',fontweight ="bo
ld")
x=sns.barplot(x=col_products,y=data.values,palette='Set2')
x.set_xticklabels(col_products, size=15)
plt.tight_layout()
```


```
plt.figure(figsize=(8,5))
plt.title('Income Age Group wise',fontweight ="bold")
x=sns.barplot(data=df,x='AgeGroup',y='Income',palette='Set2')
plt.tight_layout()
```



purchases vs age group

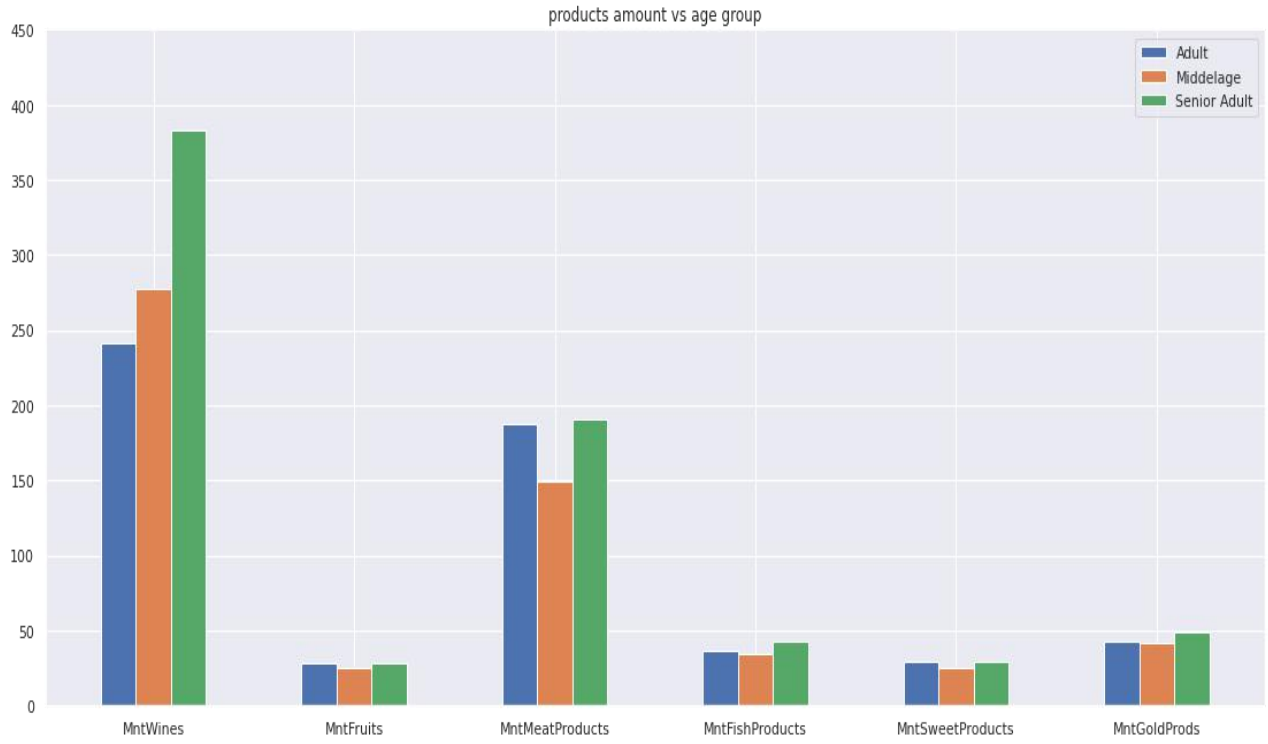
```
Purchases = ['NumDealsPurchases','NumWebPurchases','NumCatalogPurc
hases','NumStorePurchases']
dataset = df.groupby('AgeGroup')[Purchases]. mean ()

score_label = np.arange(0, 10, 1)
Adult_mean  = list(dataset.T['Adult'])
Middleage_mean  = list(dataset.T['Middle Age Adult'])
SeniorAdult_mean  = list(dataset.T['Senior Adult'])

# set width of bar
barWidth = 0.35

fig, ax = plt.subplots(figsize=(19,8))

# Set position of bar on X axis
r1 = np.arange(0, len(Purchases)*2,2)
r2 = [x + barWidth for x in r1]
r3 = [x + barWidth for x in r2]


# Make the plot

Adult = ax.bar(r1, Adult_mean, width=barWidth, label='Adult')
Middleage = ax.bar(r2,
Middleage_mean, width=barWidth, label='Middelage')
SeniorAdult= ax.bar(r3, SeniorAdult_mean,width=barWidth, label='Se
nior Adult')


# inserting x axis label
plt.xticks([r + barWidth for r in range (0,
len(Purchases)*2,2)], dataset)
ax.set_xticklabels(Purchases)

# inserting y axis label
ax.set_yticks(score_label)
ax.set_yticklabels(score_label)
# inserting legend
ax.legend()
plt.title('purchases vs age group')
plt.show()
```



products amount vs age group

```
Products = ['MntWines','MntFruits','MntMeatProducts','MntFishProducts','MntSweetProducts','MntGoldProds']
dataset = df.groupby('AgeGroup')[Products]. mean ()

score_label = np.arange(0, 500, 50)
Adult_mean  = list(dataset.T['Adult'])
Middleage_mean  = list(dataset.T['Middle Age Adult'])
SeniorAdult_mean  = list(dataset.T['Senior Adult'])

# set width of bar
barWidth = 0.35
fig, ax = plt.subplots(figsize=(19,8))

# Set position of bar on X axis
r1 = np.arange(0, len(Products)*2,2)
r2 = [x + barWidth for x in r1]
r3 = [x + barWidth for x in r2]
```

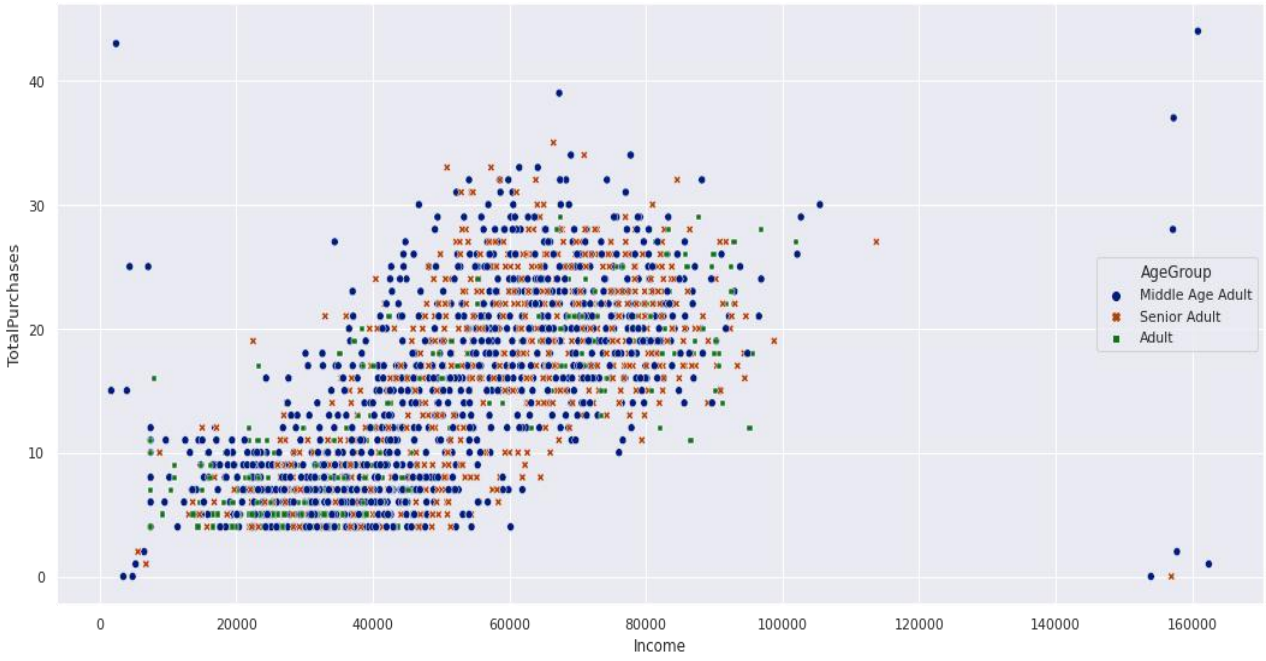
```
# Make the plot

Adult = ax.bar(r1, Adult_mean, width=barWidth, label='Adult')
Middleage = ax.bar(r2,
Middleage_mean, width=barWidth, label='Middelage')
SeniorAdult= ax.bar(r3, SeniorAdult_mean,width=barWidth, label='Se
nior Adult')


# inserting x axis label
plt.xticks([r + barWidth for r in range (0,
len(Products)*2,2)], dataset)
ax.set_xticklabels(Products)

# inserting y axis label
ax.set_yticks(score_label)
ax.set_yticklabels(score_label)

# inserting legend
ax.legend()

plt.title('products amount vs age group')
plt.show()
```



Total purchases vs Income

```
fig, ax = plt.subplots(figsize=(18,8))
sns.scatterplot(data=df,x='Income', y='TotalPurchases',ax=ax,hue='
AgeGroup',style="AgeGroup",palette='dark')
```

```
<matplotlib. axes._subplots.AxesSubplot at 0x7fc32cedaed0>
```



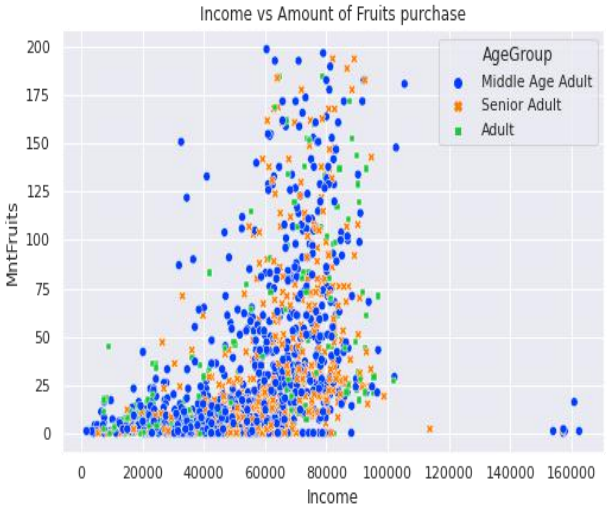
Income versus the quantity of products purchased

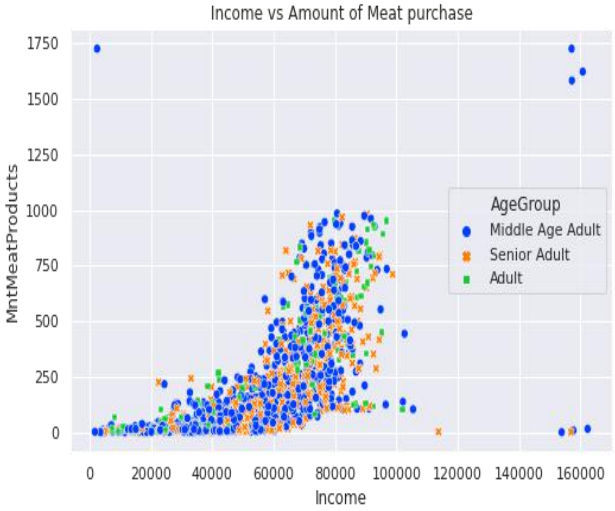
```
f,ax=plt.subplots(3,2, figsize=(18,17))

sns.scatterplot(data=df, x='Income', y='MntWines', hue='AgeGroup',
markers= ["o", "s", "D"],ax=ax[0][0])
sns.scatterplot(data=df, x='Income', y='MntWines', hue='AgeGroup',
style="AgeGroup",ax=ax[0][0], palette="dark")
ax[0][0].set_title('Income vs Amount of wines purchase')
sns.scatterplot(data=df, x='Income', y='MntFruits', hue='AgeGroup',
style="AgeGroup",ax=ax[0][1], palette="bright")
ax[0][1].set_title('Income vs Amount of Fruits purchase')
sns.scatterplot(data=df, x='Income', y='MntMeatProducts', hue='Age
Group',style="AgeGroup",ax=ax[1][0], palette="bright")
ax[1][0].set_title('Income vs Amount of Meat purchase')
```

```
sns.scatterplot(data=df, x='Income', y='MntSweetProducts', hue='Ag
eGroup',style="AgeGroup",ax=ax[1][1], palette="bright")
ax[1][1].set_title('Income vs Amount of Sweet purchase')
sns.scatterplot(data=df, x='Income', y='MntGoldProds', hue='AgeGro
up',style="AgeGroup",ax=ax[2][0], palette="bright")
ax[2][0].set_title('Income vs Amount of Gold purchase')
sns.scatterplot(data=df, x='Income', y='MntFishProducts', hue='Age
Group',style="AgeGroup",ax=ax[2][1], palette="bright")
ax[2][1].set_title('Income vs Amount of Fish purchase')
```

```
Text (0.5, 1.0, 'Income vs Amount of Fish purchase')
```


Income vs Amount of wines purchase
MntWines
Income
AgeGroup
Middle Age Adult
Senior Adult
Adult


Income vs Amount of Fruits purchase
MntFruits
Income
AgeGroup
Middle Age Adult
Senior Adult
Adult


Income vs Amount of Meat purchase
MntMeatProducts
Income
AgeGroup
Middle Age Adult
Senior Adult
Adult


Income vs Amount of Sweet purchase
MntSweetProducts
Income
AgeGroup
Middle Age Adult
Senior Adult
Adult


Income vs Amount of Gold purchase
MntGoldProds
Income
AgeGroup
Middle Age Adult
Senior Adult
Adult


Income vs Amount of Fish purchase
MntFishProducts
Income
AgeGroup
Middle Age Adult
Senior Adult
Adult

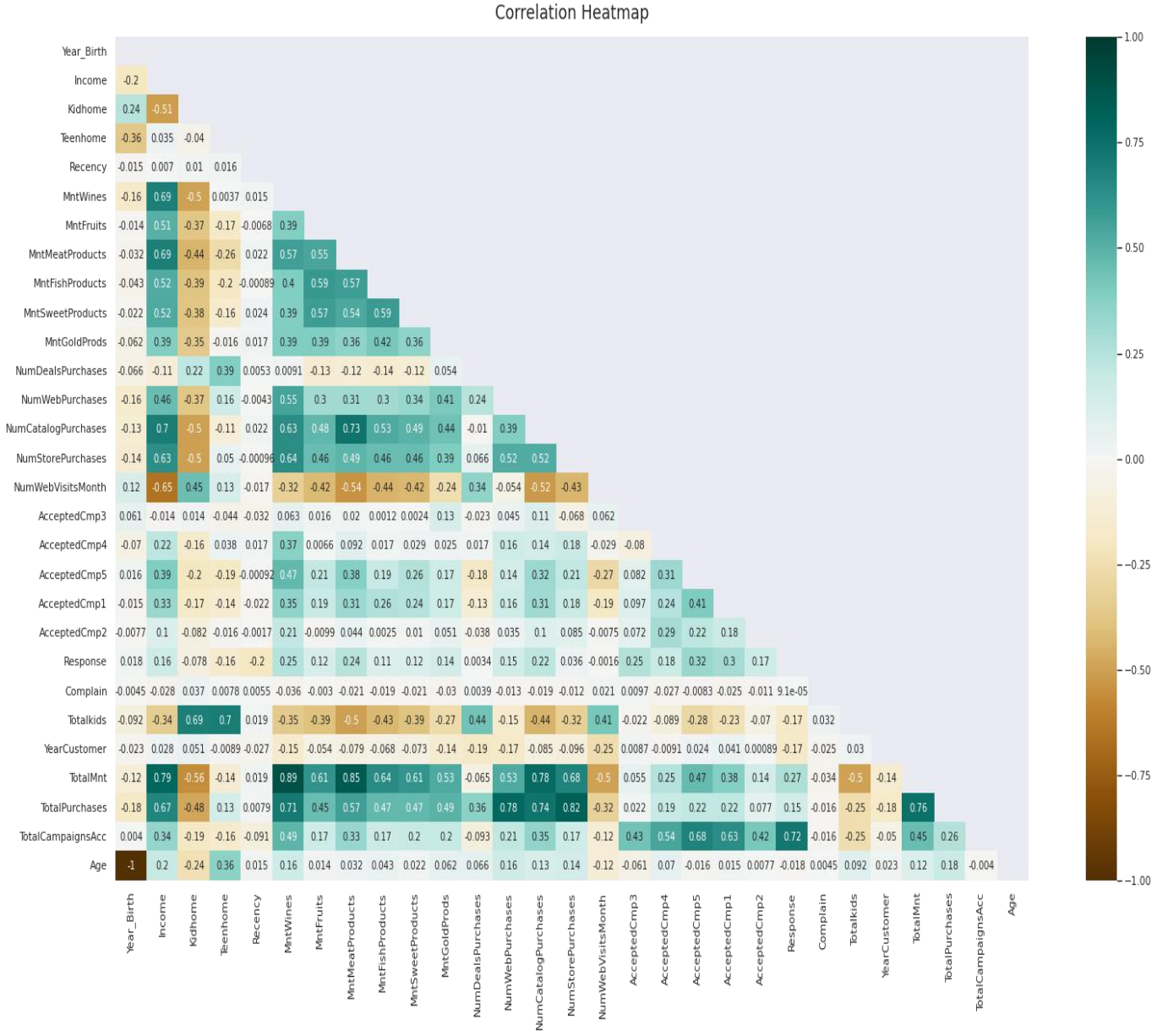
5.2 Statistical Analysis

5.2.1 What factors are largely related to the amount of purchases in the store?

Let's draw heat diagram graphs to see the correlation of numerical variables in store purchases.

```
df_num = df.drop(columns=['ID']).select_dtypes(include = ['float64
', 'int64'])


plt.figure(figsize=(25,14))


mask = np.triu(np.ones_like(df_num.corr(), dtype=np.bool))
heatmap = sns.heatmap(df_num.corr(), mask=mask, vmin=-
1, vmax=1, annot=True, cmap='BrBG')
heatmap.set_title('Correlation Heatmap', fontdict={'fontsize':18},
 pad=16);
```

5.2.2 Is there a significant correlation between geographic region and the success of a campaign?

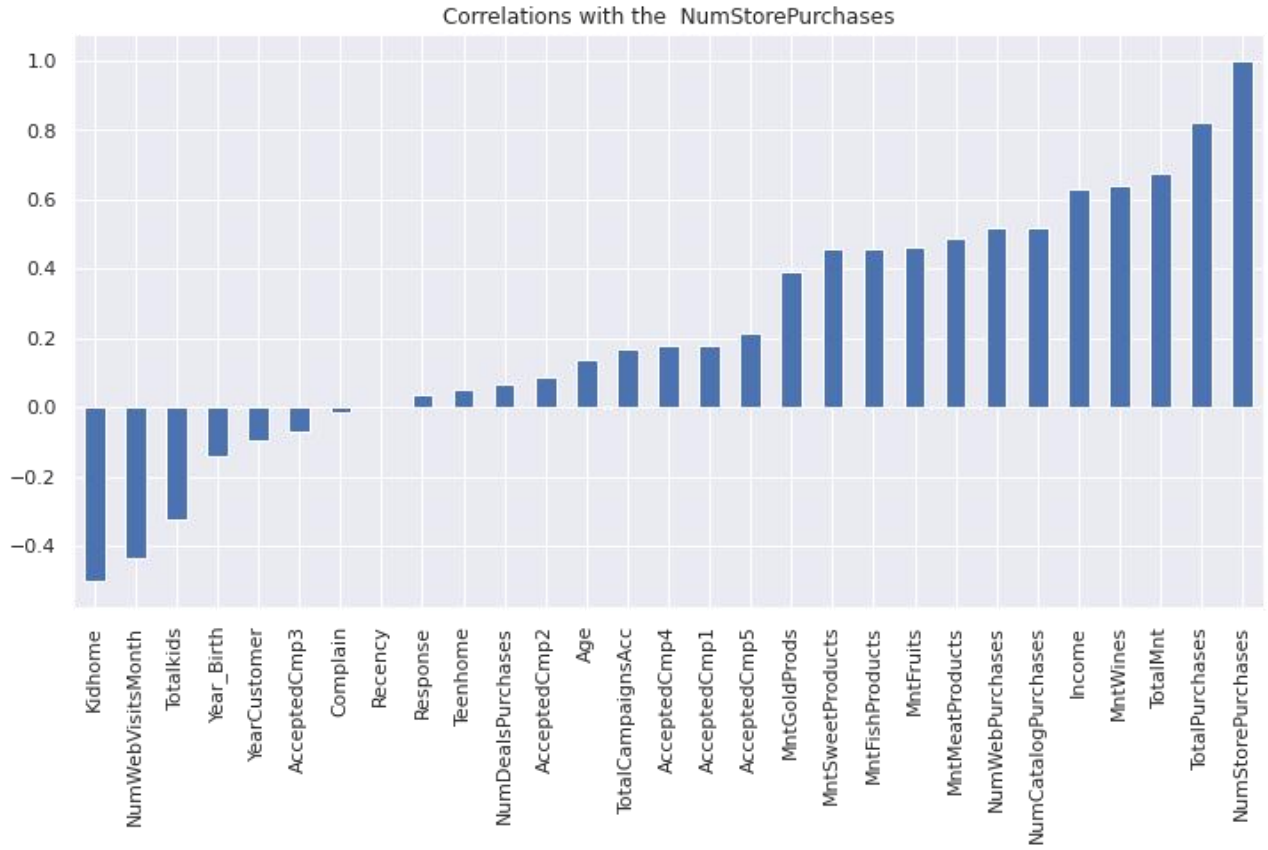


Correlation with NumStorePurchases

Let us check the correlation of numerical variables with NumStorePurchases.

```
corr_with_SalePrice = df_num.corr()
```

```
plot_data = corr_with_SalePrice["NumStorePurchases"].sort_values(a
scending=True)
plt.figure(figsize=(12,6))
plot_data.plot.bar()
plt.title("Correlations with the  NumStorePurchases")
plt.show()
```

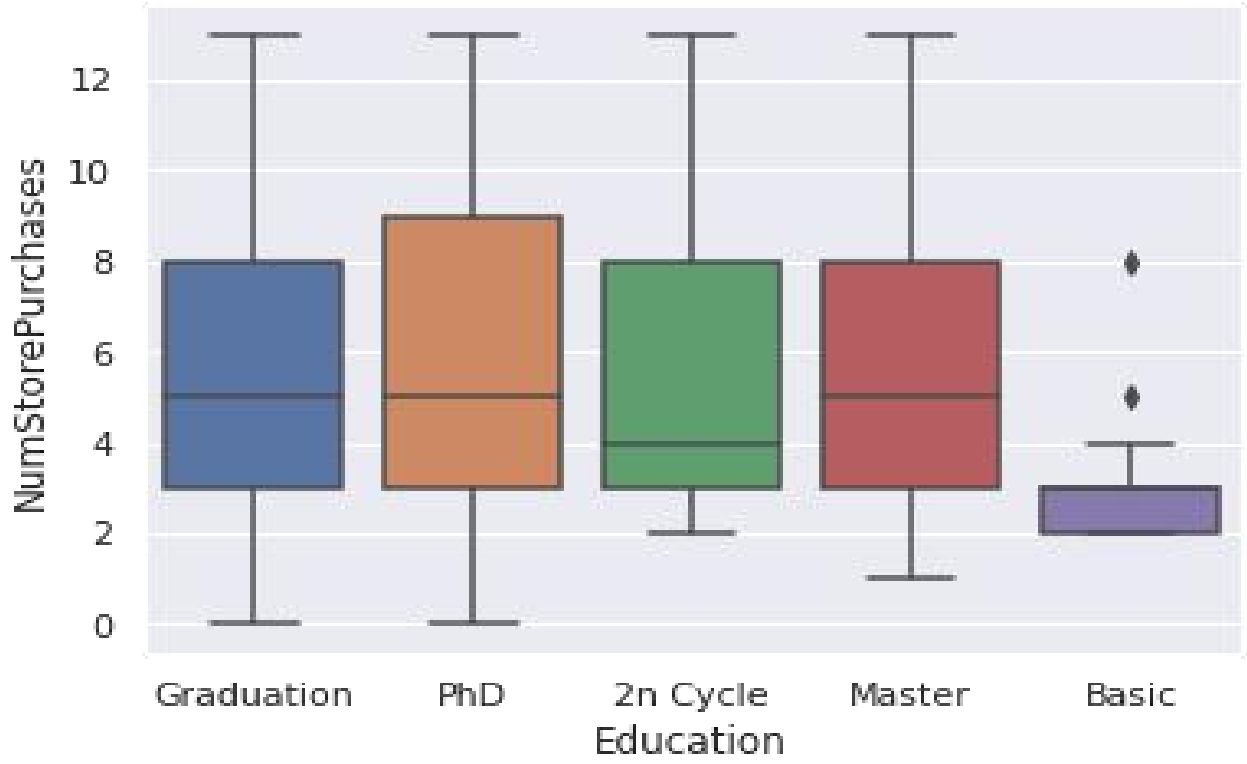


We can see the mapping of numeric columns / decorations in NumStorePurchases. The columns with a clear correlation (strongly positive or strongly negative) are important for the predictive model, but only some of those with a small correlation (approximately zero) do not have a great impact on the Sale Price, so we can still have some of them drop you.

```
Data=df.drop(columns=
['Response', 'Complain','Recency','Teenhome'])
#Droping uninformative features
Data=df.drop(columns=['ID','Dt_Customer'])
```

Let us now look at the 'NumStorePurchases' variation on different categories of categorical variables/columns.

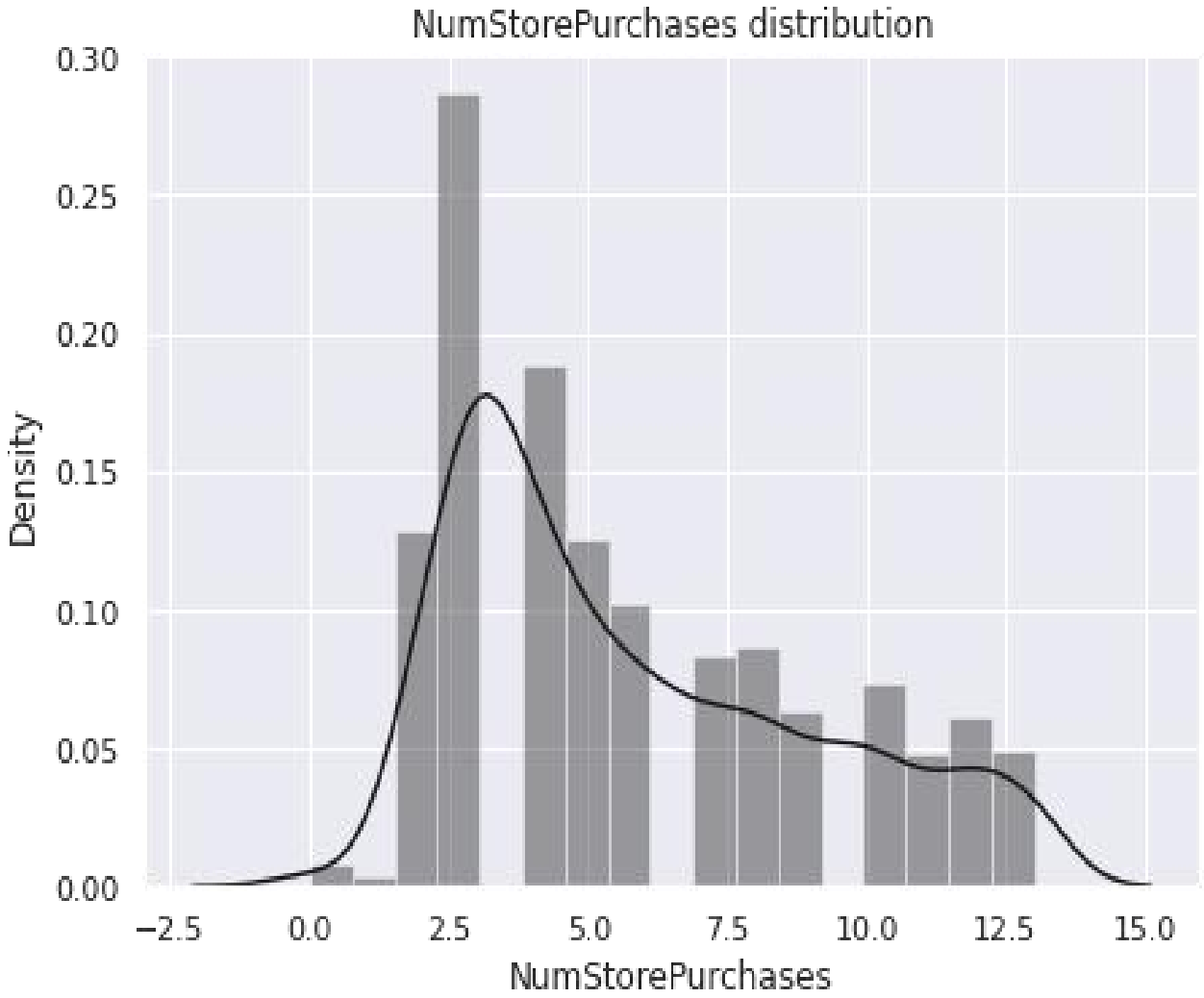
```
few_cat_variables = ['Education','Marital_Status','Country' ,'AgeG
roup']

for i in range(len(few_cat_variables)):
    sns.boxplot(x=few_cat_variables[i], y='NumStorePurchases',
    data=df)
 plt.show()
```



Now we will change the categorical variables to numerical ones by using LabelEncoder for the regression models

```
# Categorical boolean mask
categorical_feature_mask = Data.dtypes==object

# filter categorical columns using mask and turn it into a list
categorical_cols =Data.columns[categorical_feature_mask].tolist()

# instantiate labelencoder object
```

```
le = LabelEncoder()

# apply le on categorical feature columns
Data[categorical_cols] =Data[categorical_cols].apply(lambda col: l
e.fit_transform(col))
plt.figure(figsize = (7, 5))
sns.distplot(df['NumStorePurchases'], color = 'k')
plt.title('NumStorePurchases distribution');
```



6. Conclusion

While business analytics is being increasingly used to gain data-driven insights to support decision-making, little research exists regarding the mechanism through which business analytics can be used to improve decision-making effectiveness at the organizational level. Drawing on the information processing view and contingency theory, this paper develops a research model linking business analytics to organizational decision-making effectiveness. The research model is tested using structural equation modelling based on 740 responses collected from UK businesses. The key findings demonstrate that business analytics, through the mediation of a data-driven environment, positively influences information processing capabilities, which in turn have a positive effect on decision-making effectiveness. The findings also demonstrate that the paths from business analytics to decision-making effectiveness has no statistical differences between large and medium companies but some differences between manufacturing and professional service industries. Our findings contribute to the business analytics literature by providing useful insights into business analytics applications and the facilitation of data-driven decision-making. They also contribute to managers' knowledge and understanding by demonstrating how business analytics should be implemented to improve decision-making effectiveness.

7. Bibliography

[1] T. H. Davenport and J. G. Harris, Competing on Analytics: The New Science of Winning. Boston, MA: Harvard Business School Review Press, 2007.

[2] T. H. Davenport, "Analytics 3.0," Harvard Business Review, vol. 91, pp. 64-72, 2013.

[3] H. Chen, R. H. L. Chiang, and V. C. Storey, "Business intelligence and analytics: from big data to big impact," MIS Quarterly, vol. 36, pp. 1165-1188, 2012.

[4] H. J. Watson, "Tutorial: Big Data Analytics: Concepts, Technologies, and Applications," Communications of the Association for Information Systems, vol. 34, pp. 1247-1268, 2014.

[5] C. Holsapple, A. Lee-Post, and R. Pakath, "A unified foundation for business analytics," Decision Support Systems, 2014.

[6] T. H. Davenport, "Competing on analytics," Harvard Business Review, vol. 84, pp. 98-107, 2006.

[7] T. H. Davenport, J. G. Harris, D. W. De Long, and A. L. Jacobson, "Data to Knowledge to Results: Building an analytic capability," California Management Review, vol. 43, pp. 117-138, 2001.

[8] D. Kiron, P. K. Prentice, and R. B. Ferguson, "Innovating with Analytics. (Cover story)," MIT Sloan Management Review, vol. 54, pp. 47-52, 2012.

[9] S. Lavalle, E. Lesser, R. Shockley, M. S. Hopkins, and N. Kruschwitz, "Special Report: Analytics and the New Path to Value," MIT Sloan Management Review, vol. 52, pp. 22-32, 2011.

[10] D. Kiron, P. K. Prentice, and R. B. Ferguson, "Raising the Bar with Analytics," MIT Sloan Management Review, vol. 55, pp. 29-33, 2014.

[11] D. Barton and D. Court, "Making Advanced Analytics Work for You," Harvard Business Review, vol. 90, pp. 78-83, 2012.

[12] K. Gillon, S. Aral, L. Ching-Yung, S. Mithas, and M. Zozulia, "Business Analytics: Radical Shift or Incremental Change?" Communications of the Association for Information Systems, vol. 34 pp. 287-296, 2014.
[13] G. George, M. R. Haas, and A. Pentland, "Big data and management," Academy of Management Journal, vol. 57, pp. 321-326, 2014.
[14] D. A. Marchand and J. Peppard, "Why IT Fumbles Analytics," Harvard Business Review, vol. 91, pp. 104-112, 2013.
[15] K. M. Eisenhardt and M. J. Zbaracki, "Strategic Decision Making," Strategic Management Journal, vol. 13, pp. 17-37, 1992.