VISVESVARAYA TECHNOLOGICAL UNIVERSITY

Jnana Sangama, Belagavi-590018, Karnataka, India

A Mini Project Report on Data Analytics using Python.

"Analyzing Customer Acquisition and Retention: A Comprehensive Cohort Analysis"

Submitted in partial fulfilment of the requirements for the award of the degree of

Master of Computer Applications

Submitted by

NAGENDRA KUMAR K S

[1RR22MC029]

Under the Guidance of

Prof. Shreedhar Kumbhar

Assistant Professor

Department of MCA

RRCE

RAJARAJESWARI COLLEGE OF ENGINEERING

MYSORE ROAD, BANGALORE-560074

(An ISO 9001-2008 Certified Institution)

2023-2024

RAJARAJESWARI COLLEGE OF ENGINEERING

MYSORE ROAD, BANGALORE-560074
(An ISO 9001-2008 Certified Institution)
(Affiliated to Visvesvaraya Technological University)

Department of Master of Computer Applications

CERTIFICATE

This is to certify that the Data Analytics using Python Mini Project work entitled "**Analyzing Customer Acquisition and Retention: A Comprehensive Cohort Analysis**" carried out by **NAGENDRA KUMAR K S [1RR22MC029]** the student of **"RajaRajeswari College of Engineering"** in partial fulfilment of **3rd Semester Master of Computer Applications** of the Visvesvaraya Technological University, Belgaum during the year 2023-2024. It is certified that all corrections/suggestions indicated for internal report has been approved as it is satisfying the academic requirements in respect of Data Analytics using Python Mini Project work prescribed for the said degree.

Signature of Guide **Signature of HOD**

Signature of Internal **Signature of External**

ACKNOWLEDGEMENT

The success and outcome of learning "**Analyzing Customer Acquisition and Retention: A Comprehensive Cohort Analysis"** required a lot of guidance and assistance from many people and I am extremely privileged to have got this all along the completion of my course and few of the projects. All that I have done is only due to such supervision and assistance and I would not forget to thank them.

I cordially thank **Dr. T Subburaj** Head of the Department of MCA and **Assistant Prof. Shreedhar Kumbhar**, intended for giving valuable guidance steadysupport and encouragement to inclusive our internship lucratively.

I am thankful to and fortunate enough to get constant encouragement, support, and guidance from all teaching staffs who helped me in successfully completing my Mini Project on Data Analytics using Python.

NAGENDRA KUMAR K S
[1RR22MC029]

DECLARATION

I, **NAGENDRA KUMAR K S [1RR22MC029]**, student of MCA, **RajaRajeswari College of Engineering**, Bengaluru, hereby declare that Data Analytics using Python Mini Project work entitled "**Analyzing Customer Acquisition and Retention: A Comprehensive Cohort Analysis"** submitted to **Visvesvaraya Technological University** during the academic year 2023-2024, is a record of original work done by us under the guidelines of **Assistant Prof. Shreedhar Kumbhar,** Department MCA, RajaRajeswari College of Engineering, Bengaluru. This Project work is submitted in the partial fulfillment of requirements for the award of the degree of **Master of Computer Applications**. The results embodied in this report have not been submitted to any other university or institute for the award of any degree.

Date: **NAGENDRA KUMAR K S**
Place: Bangalore **[1RR22MC029]**

TABLE OF CONTENTS

SL. NO	CONTENT	PAGE NO
A	Acknowledgement	[I]
B	Declaration	[II]
C	Abstract	01
1.	Introduction	02
2.	Requirement Analysis	3-4
3.	System Requirements	05
4.	Analysis and Design	06
5.	Implementation	7-9
6.	Screenshots	10-16
7.	Testing	17
8.	Conclusion	18
	Bibliography	19

ABSTRACT

This research delves into a detailed exploration of customer acquisition and retention dynamics, employing a robust cohort analysis methodology. Going beyond conventional approaches, the study integrates visual insights in the form of bar and pie charts to enhance the comprehension of customer behavior patterns over time.

Through cohort analysis, the study categorizes customers based on shared characteristics, allowing for a nuanced examination of their journey with the business. Bar charts vividly illustrate the performance of each cohort across various metrics, facilitating a comparative analysis to identify key contributors and potential areas for improvement.

In parallel, pie charts offer a proportional representation of customer distribution within cohorts, enabling stakeholders to quickly grasp the relative significance of each segment. This dual visual approach provides a comprehensive understanding of the customer landscape, aiding decision- makers in formulating targeted strategies for acquisition and retention.

By combining traditional cohort analysis with visually engaging representations, this research aims to deliver actionable insights that empowerbusinesses.

1. INTRODUCTION

Cohort analysis involves segmenting customers into distinct groups based on shared characteristics or experiences, typically focusing on the timing of their interaction with the business. These groups, or cohorts, provide a lens through which businesses can track and analyze customer behavior patterns, trends, and preferences.

Customer acquisition and retention are pivotal aspects of any business strategy, influencing its growth, profitability, and sustainability. Cohort analysis emerges as a powerful technique to delve deep into these areas, providing valuable insights into customer behavior over time.

Understanding Cohorts:

A cohort can represent various customer segments, such as those who signed up for a service during a specific month, made their first purchase in a particular quarter, or belong to a certain demographic category. By grouping customers in this manner, businesses can uncover nuanced insights into how different cohorts engage with their products or services

Studying Customer Behavior Over Time:

Cohort analysis facilitates the tracking of key metrics over successive time periods, such as retention rates, conversion rates, average order value, and customer lifetime value. By observing how these metrics evolve for different cohorts, businesses can identify patterns and trends, enabling informed decision-making.

Applications in Business Strategy:

Cohort analysis finds wide-ranging applications across industries, including e-commerce, subscription services, software platforms, and beyond. It helps businesses optimize marketing strategies, tailor product offerings, enhance customer experience, and improve overall retention and loyalty.

Types of Analysis:

1. **Cohort Creation**: Segmenting customers into cohorts based on the month of their first purchase.

   In the realm of customer analysis, cohort creation serves as a fundamental technique for understanding and segmenting customers based on their behavior over time.

   One common approach involves segmenting customers into cohorts based on the month of their first purchase. This method allows businesses to group customers who initiated their relationship with the company during the same period, providing valuable insights into their subsequent behavior and preferences.

Purpose and Significance:

The primary purpose of cohort creation is to identify and analyze groups of customers who share a common starting point in their interaction with the business. By segmenting customers into cohorts based on the timing of their first purchase, businesses can track the behavior of these cohorts over time, uncovering trends, patterns, and insights that inform strategic decision-making

Understanding Customer Lifecycle:

Cohort creation facilitates the exploration of the customer lifecycle, from acquisition to retention and beyond. By observing how different cohorts evolve over time, businesses can gain insights into factors influencing customer retention, engagement, and loyalty. Additionally, cohort analysis allows businesses to tailor marketing strategies, product offerings, and customer experiences to meet the specific needs of each cohort.

2.Retention Analysis: Analyzing customer retention rates across different cohorts. Retention analysis (or survival analysis) is the process of analyzing user metrics to understand how and why customers churn. Retention analysis is key to gain insights on how to maintain a profitable customer base by improvingretention and new user acquisition rates.

By running consistent retention analysis, you'll learn:

- Why customers are churning.
- When customers are more likely to leave.
- How churn affects your bottom line.
- How to improve your retention strategies.

Overall, this analysis allows you to see how well your customer retention efforts are working
Without it, you may end up spending your marketing budget inefficiently

3.Conversion Analysis:

Conversion analysis is a fundamental aspect of evaluating the effectiveness of marketing and sales efforts in converting prospects into paying customers. It involves examining the rate at which potential customers take desired actions, such as making a purchase, signing up for a service, or completing a specific goal.

The primary objective of conversion analysis is to identify bottlenecks or barriers in the conversion process and to implement strategies to improve conversion rates. This may involve analyzing website traffic sources, optimizing landing pages and calls-to-action, A/B testing different marketing messages or offers, and leveraging customer data to personalize the conversion experience.

4. **Average Order Value Analysis**: Examining changes in average order value across cohorts.

Average order value (AOV) is a key metric used by businesses to measure the average amount spent by customers during each transaction. Analyzing how AOV changes across different cohorts provides valuable insights into customer spending behavior, purchasing patterns, and the overall health of the business.

The primary objective of analyzing AOV across cohorts is to identify factors that influence purchasing decisions and to develop strategies to increase AOV. This may involve analyzing the impact of pricing strategies, promotions, cross-selling and upselling techniques, and customer segmentation on AOV.

Basic Steps to Conduct Cohort Analysis

The basic steps to conduct cohort analysis are pretty straightforward. You're basically following the scientific method—develop a hypothesis, gather data, and then test. Here are the steps you should follow, regardless of where and how you choose to conduct your analysis.

1. **Determine your hypothesis or research question:** What are you hoping to learn from this analysis and how will you use that information?

2. **Define the metrics that will help answer your question:** If you're looking to find out why some users churn, you may want to consider churn and 7-day retention, for example.

3. **Define specific cohorts that are relevant:** Who are the users you're interested in looking at? When did they sign up or complete the target behavior you're interested in?

4. **Analyze the report:** With the information from steps 2 and 3, you can run your report—either by hand or using software that offers cohort analysis features (like Indicative).

5. **Test and review results:** From the report, you should be able to answer your original question and confirm or refute your hypothesis. The next step is to implement any changes and monitor.

2.REQUIREMENTS

Software Requirements:

- **Development Environment:**

  Visual Studio Code.

  Python 3.x (Specify the version if necessary).

- **Dependencies and Libraries:**

  Matplotlib: A 2D plotting library for creating static, animated, and interactive visualizations in Python.

  Seaborn: A statistical data visualization library based on Matplotlib, providing informative statistical graphics.

  Pandas: A data manipulation and analysis library, providing data structures for storing and manipulating large datasets.

- **Version Control:**

  Git: A distributed version control system for tracking changes in source code during software development.

- **Package Management:**

  Pip: The package installer for Python, used for installing and managing Python packages.

- **Operating System Compatibility:**

  Specify if the project is platform-specific (e.g., Windows, macOS, Linux)

Hardware Requirements:

- **Processor (CPU):**

  A multi-core processor (e.g., dual-core or quad-core) is recommended for better parallel processing capabilities.

- **Memory (RAM):**

  At least 8 GB of RAM is recommended, especially when dealing with medium to large datasets.

- **Storage:**

  SSD (Solid State Drive) is preferred over HDD (Hard Disk Drive) for faster data access and improved overall performance.

- **Display:**

  A standard monitor with a resolution suitable for coding and data visualization tasks.

- **Graphics (GPU):**

  This is more relevant if you plan to engage in machine learning tasks using GPU acceleration.

- **Network:**

  A stable internet connection may be required for package installations, updates, and accessing online resources.

3.SOFTWARE REQUIRMENTS & SPEICIFICATION

ANACONDA :

Anaconda is a distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.), that aims to simplify package management and deployment. The distribution includes data-science packages suitable for Windows, Linux, and macos. It is developed and maintained by Anaconda, Inc., which was founded by Peter Wang and Travis Oliphant in 2012. As an Anaconda, Inc. Product, it is also known as **Anaconda Distribution** or **Anaconda Individual Edition**, while other products from the company are Anaconda Team Edition and Anaconda Enterprise Edition, both of which are not free.

Package versions in anaconda are managed by the package management system conda. This package manager was spun out as a separate open-source package as it ended up being useful on its own and for things other than python. There is also a small, bootstrap version of anaconda called miniconda, which includes only conda, python, the packages they depend on, and a small number of other packages.

Conda is an open-source package and environment management system that runs on Windows, macos, and Linux. Conda quickly installs, runs, and updates packages and their dependencies. It also easily creates, saves, loads, and switches between environments on your local computer. It was created for Python programs, but it can package and distribute software for any language.

Open Source

Access the open-source software you need for projects in any field, from data visualization to robotics.

User-friendly

With our intuitive platform, you can easily search and install packages and create, load, and switch between environments.

Anaconda Repository

Our repository features over 8,000 open-source data science and machine learning packages, Anaconda-built and compiled for all major operating systems and architectures.

Fig 3 Anaconda Repository

Jupyter Notebook

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations, and narrative text. Uses include data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more. Jupyter has support for over 40 different programming languages and Python is one of them. Python is a requirement (Python 3.3 or greater, or Python 2.7) for installing the Jupyter Notebook itself.

Installation

Install Python and Jupyter using the Anaconda Distribution, which includes Python, the Jupyter Notebook, and other commonly used packages for scientific computing and data science. You can download Anaconda's latest Python3 version from here. Now, install the downloaded version of Anaconda. Installing Jupyter Notebook using pip:

```
python3 -m pip install --upgrade pip
python3 -m pip install jupyter
```

Fig 7 Jupyter Notebook Installation

Hello World in Jupyter Notebook

After successfully installing and creating a notebook in Jupyter Notebook, let's see how to write code in it. Jupyter notebook provides a cell for writing code in it. The type of code depends on the type of notebook you created. For example, if you created a Python3 notebook then you can write Python3 code in the cell. Now, let's add the following code

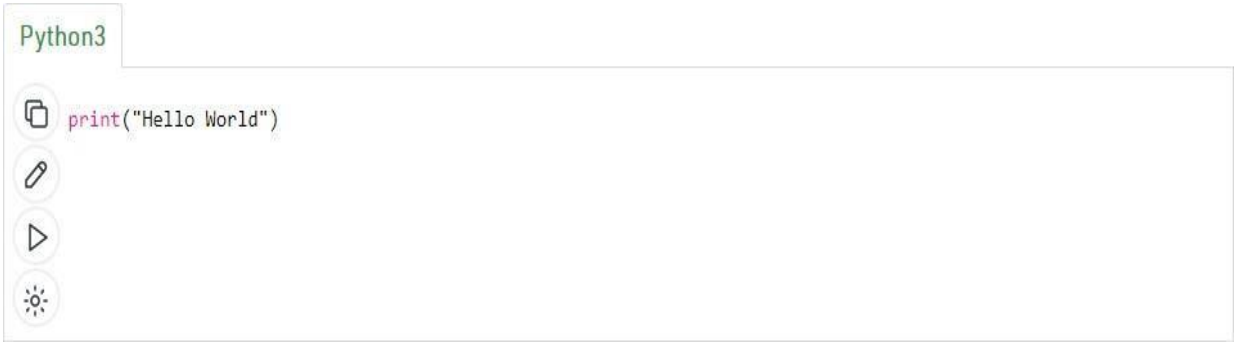


Fig 8

To run a cell either click the run button or press shift ⇧ + enter Q after selecting the cell you want to execute. After writing the above code in the jupyter notebook, the output was:

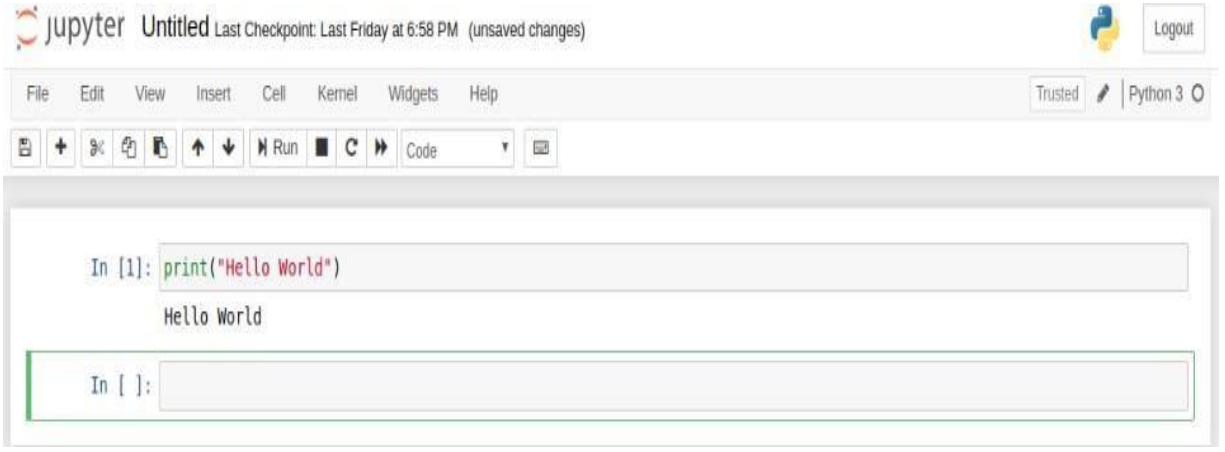


Fig 9 Jupyter Notebook

PANDAS:

Pandas is a popular open-source data analysis and manipulation library for the Python programming language. It provides data structures for efficiently storing and processing large datasets, as well as tools for data manipulation, cleaning, filtering, and aggregation.

The two main data structures in Pandas are Series and Data Frame. A Series is a one-dimensional array-like object that can hold any data type, while a Data Frame is a two-dimensional table-like datastructure consisting of rows and columns, with each column being a Series.

Pandas also provides a wide range of functions for data manipulation and analysis, including data cleaning and preprocessing, filtering and selecting data, grouping and aggregating data, and statistical analysis. It can also handle data from a variety of sources, including CSV, Excel, SQL databases, and more.

Overall, Pandas is a powerful tool for working with data in Python, and it has become a standard tool for many data analysts and data scientists.

MATPLOTLIB:

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK. There is also a procedural "pylab" interface based on a state machine (like OpenGL), designed to closely resemble that of MATLAB, though its use is discouraged. SciPy makes use of Matplotlib.

Matplotlib was originally written by John D. Hunter. Since then it has had an active development community and is distributed under a BSD-style license. Michael Droettboom was nominated as matplotlib's lead developer shortly before
John Hunter's death in August 2012 and was further joined by Thomas Caswell. Matplotlib is a Num FOCUS fiscally sponsored project.

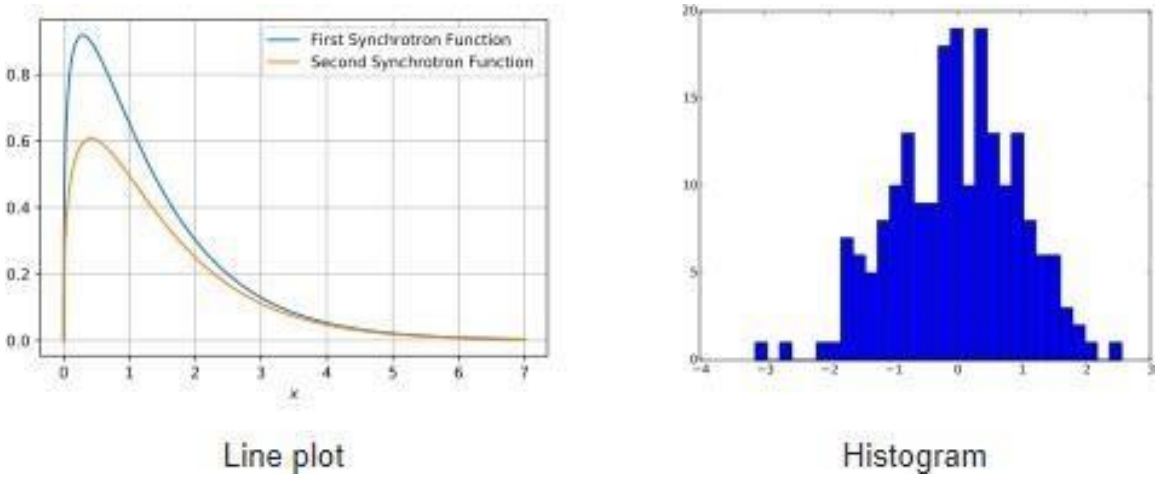


Fig 10 Matplotlib Graphs

NUMPY :

NumPy (Numerical Python) is **an open-source library for the Python programming language**. It is used for scientific computing and working with arrays. Apart from its multidimensional array object, it also provides high-level functioning tools for working with arrays.

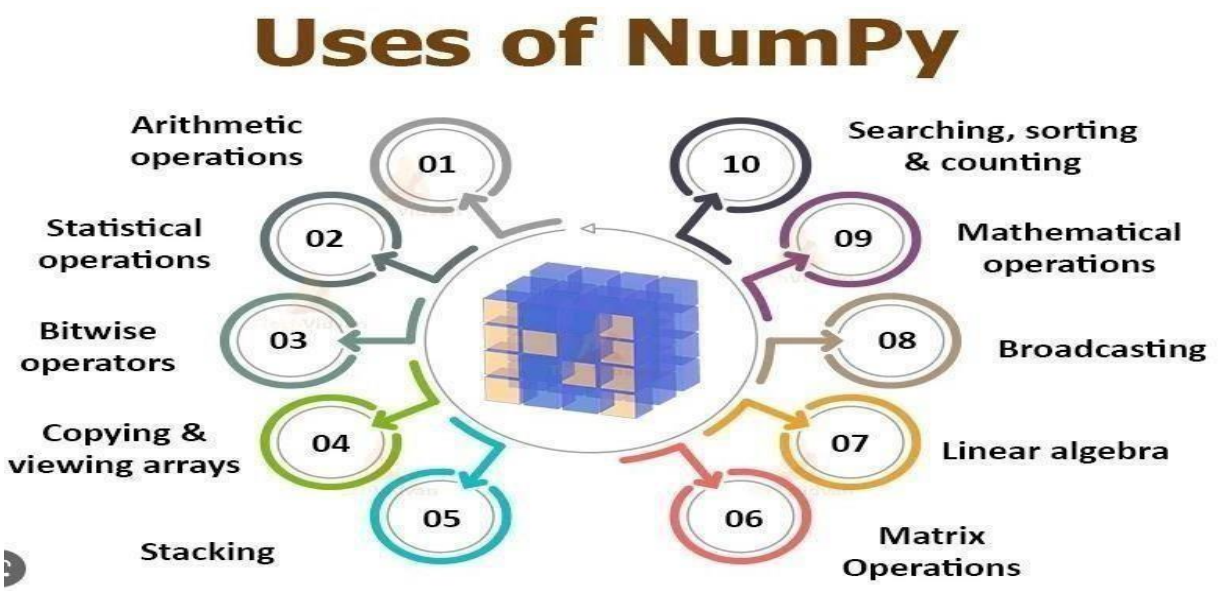


Fig 11 NumPy Uses

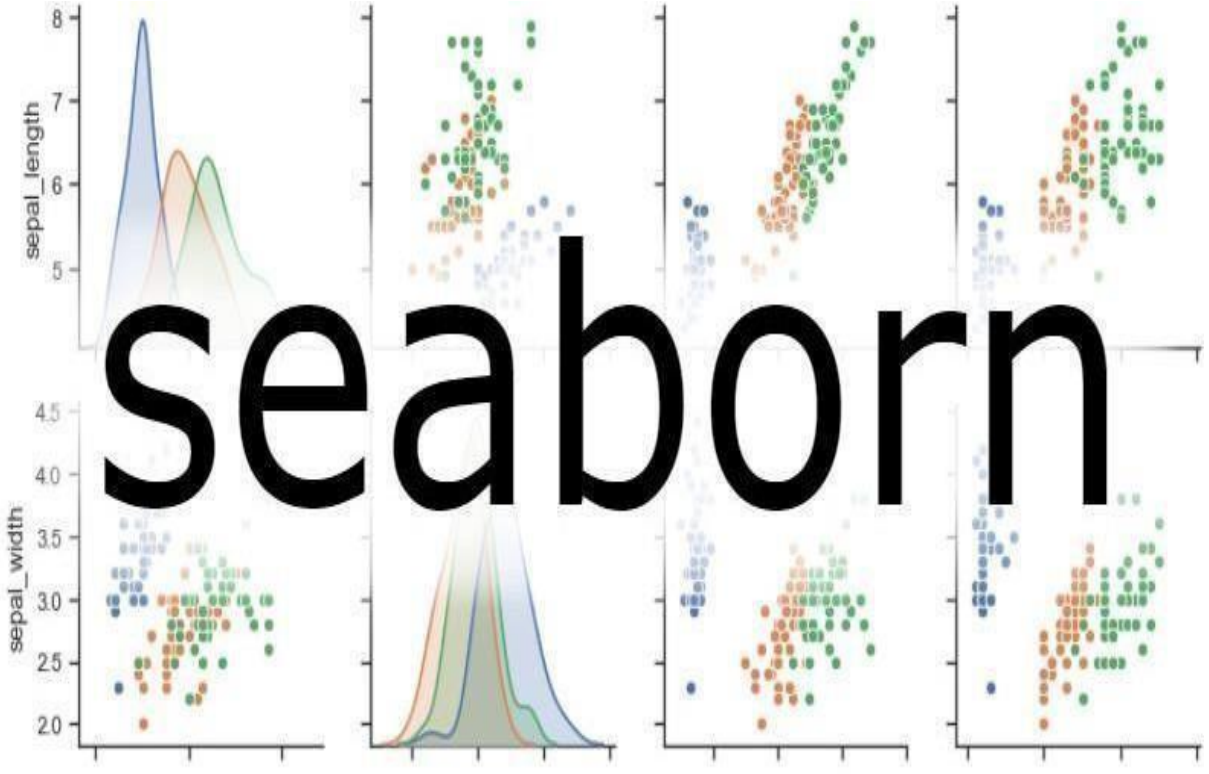
SEABORN:

Seaborn is a popular Python data visualization library built on top of Matplotlib. It provides a high-level interface for creating informative and attractive statistical graphics. Seaborn is widely used in data analysis and machine learning projects for exploring and visualizing data, as well as for building predictive models.

Seaborn provides a range of built-in functions for creating various types of visualizations, including:

1. Scatter plots
2. Line plots
3. Histograms
4. Box plots
5. Violin plots
6. Heat maps
7. Pair plots
8. Regression plots
9. Bar plots
10. Count plots

Seaborn's strength lies in its ability to create complex and aesthetically pleasing visualizations with minimal coding. It provides a range of color palettes, styles, and themes that can be easily applied to any plot. Additionally, it supports statistical estimation and testing by automatically calculating and displaying confidence intervals, p-values, and other statistical measures.



Seaborn Plots

OPERATING SYSTEM

Microsoft Windows is a group of several proprietary graphical operating system families developed and marketed by Microsoft. Each family caters to a certain sector of the computing industry. For instance, Windows NT for consumers, Windows Server for servers, and Windows IoT for embedded systems. Defunct Windows families include Windows 9x, Windows Mobile, and Windows Phone. The first version of Windows was released on November 20, 1985, as a graphical operating system shell for MS-DOS in response to the growing interest in graphical user interfaces (GUIs). Windows is the most popular desktop operating system in the world, with a 70% market share as of March 2023, according to Stat Counter. However, Windows is not the most used operating system when including both mobile and desktop OSes, due to Android's massive growth. As of September 2022, the most recent version of Windows is Windows 11 for consumer PCs and tablets, Windows 11 Enterprise for corporations, and Windows Server 2022 for servers

Windows is a family of operating systems developed by Microsoft Corporation. It is one of the most widely used operating systems for personal computers and has been a dominant player in the desktop and laptop operating system market. Here are key aspects of Windows OS:

1. **Graphical User Interface (GUI):** Windows is known for its graphical user interface, which uses icons, buttons, and windows to allow users to interact with the computer. The desktop metaphor is a fundamental aspect of the Windows GUI.

2. **Versions and Editions:** Over the years, Microsoft has released multiple versions and editions of the Windows operating system. Some notable versions include Windows 3.1, Windows 95, Windows XP, Windows 7, Windows 8, and Windows 10. Each version brought new features, improvements, and changes.

3. **Windows 10:** The latest major version as of my last knowledge update in January 2022 is Windows 10. It was released in 2015 and has received regular updates. Windows 10 is designed to be a universal operating system, compatible with various devices such as desktops, laptops, tablets, and even some smartphones.

4. **Start Menu:** The Start Menu is a central element of the Windows user interface, providing access to installed applications, system settings, and search functionality. It underwent changes in different Windows versions but returned in a revamped form in Windows 10.

5. **File Explorer:** File Explorer is the file management application in Windows, allowing users to navigate, organize, and manipulate files and folders. It provides a graphical interface for accessing the file system.

6. **Taskbar:** The Taskbar is a horizontal bar located at the bottom of the screen, providing quick access to frequently used applications and system notifications.

7. **Control Panel** and Settings: Windows provides both the traditional Control Panel and the modern Settings app for configuring system settings. The Settings app is more streamlined and user-friendly.

8. **Security Features:** Windows includes security features such as Windows Defender (antivirus and anti-malware), BitLocker (disk encryption), Windows Firewall, and regular security updates to protect users from threats.

9. **Compatibility:** Windows is compatible with a vast array of third-party software and hardware devices, making it a versatile platform for various applications.

10. **Updates and Support:** Microsoft regularly releases updates and patches to improve performance, fix bugs, and address security vulnerabilities. Windows typically receives long- term support, and users are encouraged to keep their systems up to date.

11. **Microsoft Store:** The Microsoft Store provides a platform for users to download and install applications, including both traditional desktop applications and Universal Windows Platform (UWP) apps.

4.ANALYSIS AND DESIGN

Purpose and Goal: The purpose of this code is to perform a comprehensive cohort analysis to understand customer acquisition and retention dynamics over time. The goal is to identify patterns, trends, and insights related to customer behavior, retention rates, and purchasing patterns.

Data Acquisition and Preparation: The code begins by importing necessary libraries such as Pandas, Matplotlib, and Seaborn. Data is read from an Excel file containing customer transaction records. Initial data exploration is conducted to understand the structure of the dataset and identify anymissing values.

Data Cleaning and Preprocessing: Missing values in the 'Description' and 'Customer ID' columns are handled by filling them with appropriate values. Independent and dependent attributes are identified for further analysis.

Cohort Creation: Customers are segmented into cohorts based on the month of their first purchase, creating distinct groups for analysis. The cohort index is calculated to determine the duration since the customer's first purchase.

Cohort Analysis Techniques:

Retention Analysis: Cohort-based retention rates are calculated to understand how well the business retains customers over time. Visualization: Heatmaps are generated to visualize cohort data, allowing for easy identification of trends and patterns.

Percentage Analysis: Cohort data is converted into percentages to compare the retention across different cohorts.

Bar Graphs and Pie Charts: Visual representations such as bar plots and pie charts are utilized to present retention rates and customer retention rate percentages in a more intuitive manner.

Visualization: Heatmaps are generated to visualize cohort data, allowing for easy identification of trends and patterns

1. **Interpretation and Insights:**

The analysis aims to provide actionable insights into customer acquisition and retention strategies. Trends and patterns identified through cohort analysis can guide decision-making processes, including marketing strategies, product development, and customer engagement initiatives.

2. **Future Iterations:**

The code can be further refined and extended by incorporating additional analysistechniques such as customer segmentation, lifetime value analysis, and cohort-based predictive modeling. Integration with real-time data sources and automation of data processing pipelines can enhance the scalability and efficiency of the analysis.

3. Deployment and Reporting:

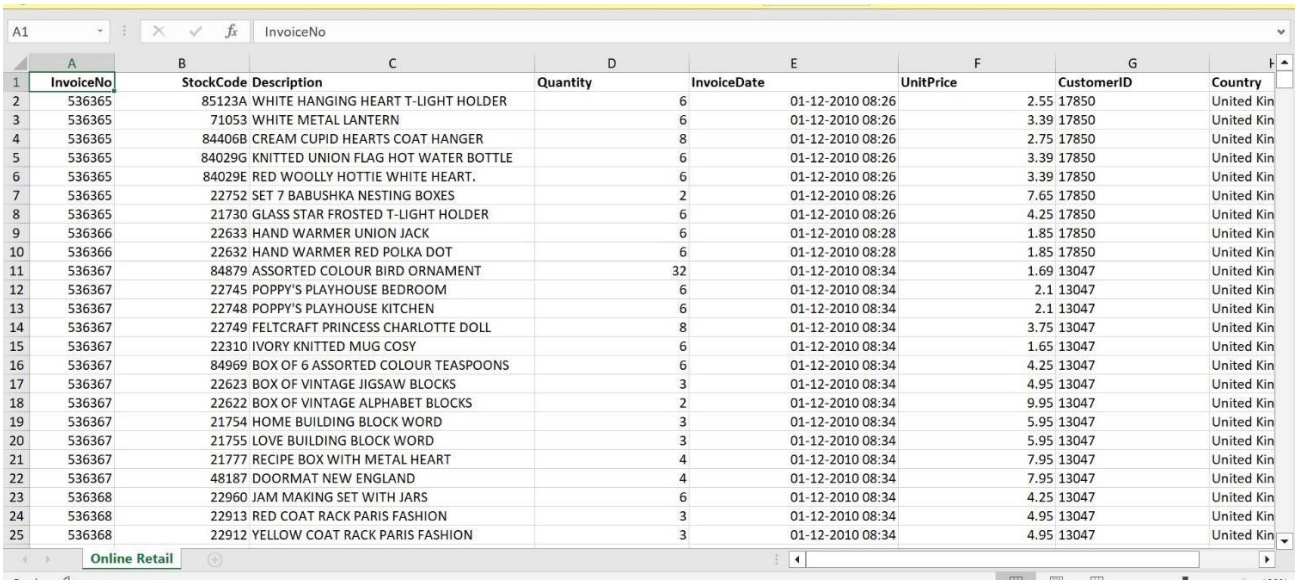
The results of the cohort analysis can be compiled into a comprehensive report or dashboard for stakeholders, providing a holistic view of customer acquisition and retention dynamics.

Insights derived from the analysis can inform strategic planning and drive business growth initiatives.

Overall, the code serves as a foundation for conducting cohort analysis to gain valuable insights into customer behavior and retention dynamics, enabling businesses to make informed decisions and optimize their strategies for sustainable growth and success.

DATA SETS:

The dataset contains customer transaction records with attributes such as Invoice No, Stock Code, Description, Quantity, Invoice Date, and Unit Price. Data may include information on customer demographics, purchase history, and other relevant factors.



A1 | InvoiceNo

	A	B	C	D	E	F	G	H
1	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
2	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	01-12-2010 08:26	2.55	17850	United Kin
3	536365	71053	WHITE METAL LANTERN	6	01-12-2010 08:26	3.39	17850	United Kin
4	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	01-12-2010 08:26	2.75	17850	United Kin
5	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	01-12-2010 08:26	3.39	17850	United Kin
6	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	01-12-2010 08:26	3.39	17850	United Kin
7	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	01-12-2010 08:26	7.65	17850	United Kin
8	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	01-12-2010 08:26	4.25	17850	United Kin
9	536366	22633	HAND WARMER UNION JACK	6	01-12-2010 08:28	1.85	17850	United Kin
10	536366	22632	HAND WARMER RED POLKA DOT	6	01-12-2010 08:28	1.85	17850	United Kin
11	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	01-12-2010 08:34	1.69	13047	United Kin
12	536367	22745	POPPY'S PLAYHOUSE BEDROOM	6	01-12-2010 08:34	2.1	13047	United Kin
13	536367	22748	POPPY'S PLAYHOUSE KITCHEN	6	01-12-2010 08:34	2.1	13047	United Kin
14	536367	22749	FELTCRAFT PRINCESS CHARLOTTE DOLL	8	01-12-2010 08:34	3.75	13047	United Kin
15	536367	22310	IVORY KNITTED MUG COSY	6	01-12-2010 08:34	1.65	13047	United Kin
16	536367	84969	BOX OF 6 ASSORTED COLOUR TEASPOONS	6	01-12-2010 08:34	4.25	13047	United Kin
17	536367	22623	BOX OF VINTAGE JIGSAW BLOCKS	3	01-12-2010 08:34	4.95	13047	United Kin
18	536367	22622	BOX OF VINTAGE ALPHABET BLOCKS	2	01-12-2010 08:34	9.95	13047	United Kin
19	536367	21754	HOME BUILDING BLOCK WORD	3	01-12-2010 08:34	5.95	13047	United Kin
20	536367	21755	LOVE BUILDING BLOCK WORD	3	01-12-2010 08:34	5.95	13047	United Kin
21	536367	21777	RECIPE BOX WITH METAL HEART	4	01-12-2010 08:34	7.95	13047	United Kin
22	536367	48187	DOORMAT NEW ENGLAND	4	01-12-2010 08:34	7.95	13047	United Kin
23	536368	22960	JAM MAKING SET WITH JARS	6	01-12-2010 08:34	4.25	13047	United Kin
24	536368	22913	RED COAT RACK PARIS FASHION	3	01-12-2010 08:34	4.95	13047	United Kin
25	536368	22912	YELLOW COAT RACK PARIS FASHION	3	01-12-2010 08:34	4.95	13047	United Kin

Online Retail

TOOLS USED

PYTHON PROGRAMMING:

Python is a High-Level, Interpreted, Interactive and Object-Oriented Scripting Language. Python is designed to be highly readable. It uses English keywords frequently whereas other languages use punctuation, and it has fewer syntactical constructions than other languages.

VISUAL STUDIO CODE:

Visual Studio is a Powerful Developer Tool that you can use to complete the entire development cycle in one place. It is a comprehensive Integrated Development Environment (IDE) that you can use to Write, Edit, Debug, and Build Code, and then deploy your app. Beyond code editing and debugging, Visual Studio includes Compilers, Code Completion Tools, Source Control, Extensions, and many more features to enhance every stage of the software development process.

5. IMPLEMENTATION

MODULES :

- **Data Collection:**

  Kaggle Data Collection Module: Describing the process of obtaining data from Kaggle datasets, including any preprocessing steps.

- **Data Exploration and Cleaning:**

  Data Exploration Module: Using Pandas for exploring the dataset, understanding its structure, and gaining initial insights.

  Data Cleaning Module: Addressing missing values, handling outliers, and ensuring data quality.

- **Data Visualization:**

  Matplotlib Module: Creating static plots and charts for data visualization.

  Seaborn Module: Utilizing Seaborn for enhancing the aesthetics of visualizations and exploring statistical relationships in the data.

  Visual Studio Code Integration: Explaining how Visual Studio Code was used for writingand executing code related to data visualization.

- **Statistical Analysis:**

  Descriptive Statistics Module: Calculating and interpreting key statisticalmeasures using Pandas.

  Correlation Analysis: Exploring relationships between variables using statistical methods and visualizations. ***Explanation****:* Visualizations aid in presenting complex data insights in an accessible manner, enabling stakeholders to grasp trends, correlations, and outliers effectively.

- **Model Evaluation:**

  Model Evaluation Module: Assessing model performance through metrics like accuracy, precision, recall, etc.

  Visualizing Model Results: Utilizing Matplotlib and Seaborn to visualize model outputs.

- **Conclusion and Recommendations:**

  Summary Module: Summarizing key findings and insights from the analysis.

  Recommendations: Providing actionable recommendations based on the analysis.

- **Documentation:**

  Code Documentation: Describing the code structure, functions, and usage.

  Report Generation: Creating a comprehensive report using Markdown or otherdocumentation tools.

- **Future Work:**

  Future Work Module: Suggesting potential avenues for future exploration, improvements,or additional analyses.

6.IMPLEMENTATION(CODE)

```
# import libraries
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```
# Read the file
df = pd.read_excel('C:/Users/nagen/OneDrive/Desktop/MCA_Projects/DataAnalytics_Python/Datasets/data.xlsx')
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    541909 non-null  object
 1   StockCode    541909 non-null  object
 2   Description  540455 non-null  object
 3   Quantity     541909 non-null  int64
 4   InvoiceDate  541909 non-null  datetime64[ns]
 5   UnitPrice    541909 non-null  float64
 6   CustomerID   406829 non-null  float64
 7   Country      541909 non-null  object
dtypes: datetime64[ns](1), float64(2), int64(1), object(4)
```

```
# returns the top 5 data present in the record
df.head()
```

```
df.isnull().sum()
```

```
InvoiceNo          0
StockCode          0
Description     1454
Quantity           0
InvoiceDate        0
UnitPrice          0
CustomerID     135080
Country            0
dtype: int64
```

```
# # drop rows with no customer ID or null values
# df = df.dropna(subset = ['CustomerID'])
# df.info()

# Fill null values in the 'Description' column with a placeholder string
df['Description'].fillna('No Description', inplace=True)

# Fill null values in the 'CustomerID' column with a specific value (-1 in this case)
df['CustomerID'].fillna(-1, inplace=True)
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    541909 non-null  object
 1   StockCode    541909 non-null  object
 2   Description  541909 non-null  object
 3   Quantity     541909 non-null  int64
 4   InvoiceDate  541909 non-null  datetime64[ns]
 5   UnitPrice    541909 non-null  float64
 6   CustomerID   541909 non-null  float64
 7   Country      541909 non-null  object
```

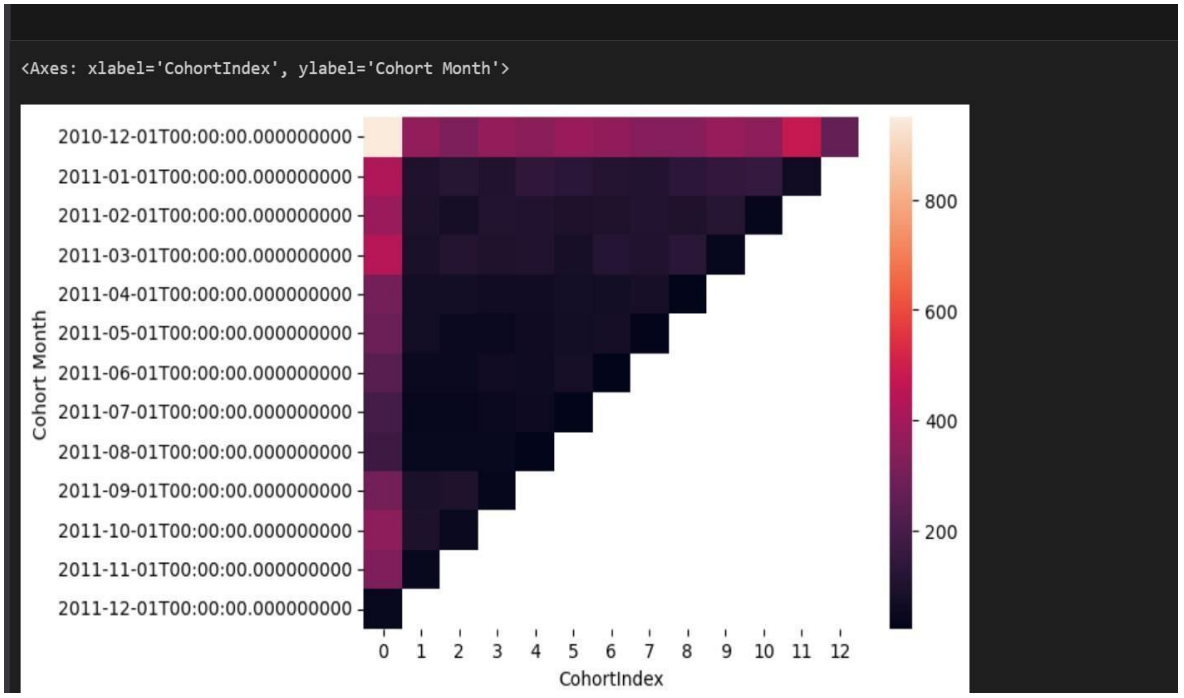
```
#check if it works
Invoice_year[:10]  #cohort_year[:10]
```

```
0    2010
1    2010
2    2010
3    2010
4    2010
5    2010
6    2010
7    2010
8    2010
9    2010
Name: InvoiceMonth, dtype: int32
```

```
#create a cohort index ,this index shows when the user acquired
year_diff = Invoice_year - cohort_year
month_diff = Invoice_month - cohort_month
```

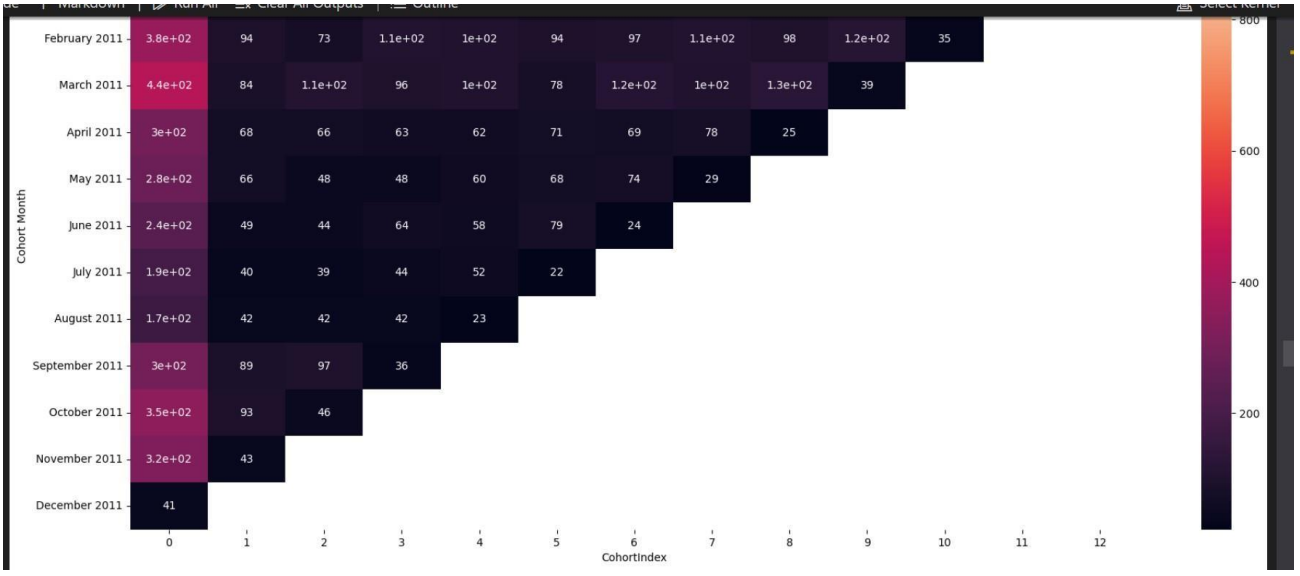
```
#create a cohort index ,this index shows when the user acquired
year_diff = Invoice_year - cohort_year
month_diff = Invoice_month - cohort_month

# because 1 year consists of 12 months and plus the months
# we add 1 because if the customer is new memeber and the diff would be o so
#df['CohortIndex'] = year_diff * 12 + month_diff + 1
df['CohortIndex'] = year_diff * 12 + month_diff
df.head()
```

```
#count the "CustomerID" by grouping "Cohort Month" and "CohortIndex" represents how long the customers been active
# df.groupby(['Cohort Month', 'CohortIndex'])['CustomerID'].nunique()
cohort_data = df.groupby(['Cohort Month', 'CohortIndex'])['CustomerID'].apply(pd.Series.nunique).reset_index()
cohort_data
```



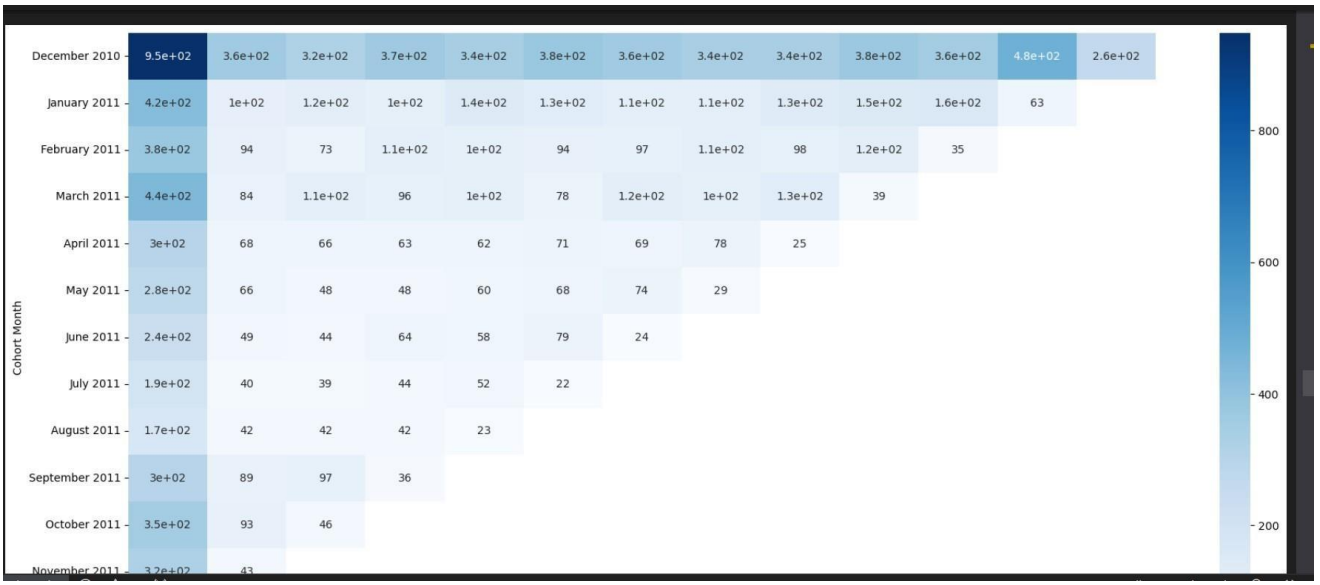

```
#increase the fig size to make fig more clear
plt.figure(figsize=(21,10))
sns.heatmap(cohort_table,annot=True)
```

```
<Axes: xlabel='CohortIndex', ylabel='Cohort Month'>
```



```
#increase the fig size to make fig more clear
plt.figure(figsize=(21,10))
sns.heatmap(cohort_table,annot=True, cmap='Blues')
```

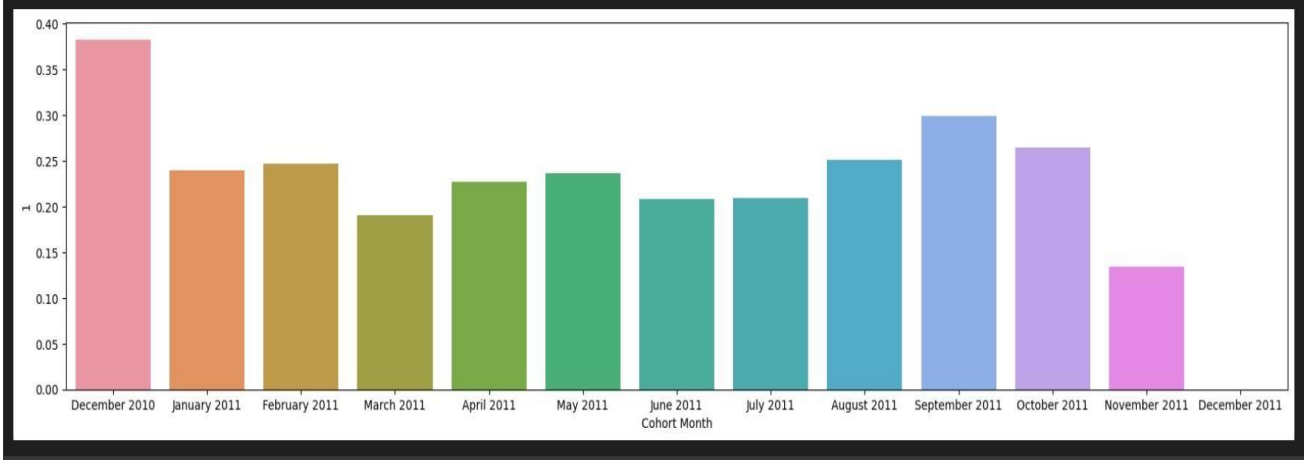
```
<Axes: xlabel='CohortIndex', ylabel='Cohort Month'>
```




```
# from above cohort analysis converting it to bar graph

plt.figure(figsize=(21,5))
sns.barplot(x=final_cohort_table.index, y=final_cohort_table[1])
```

Python

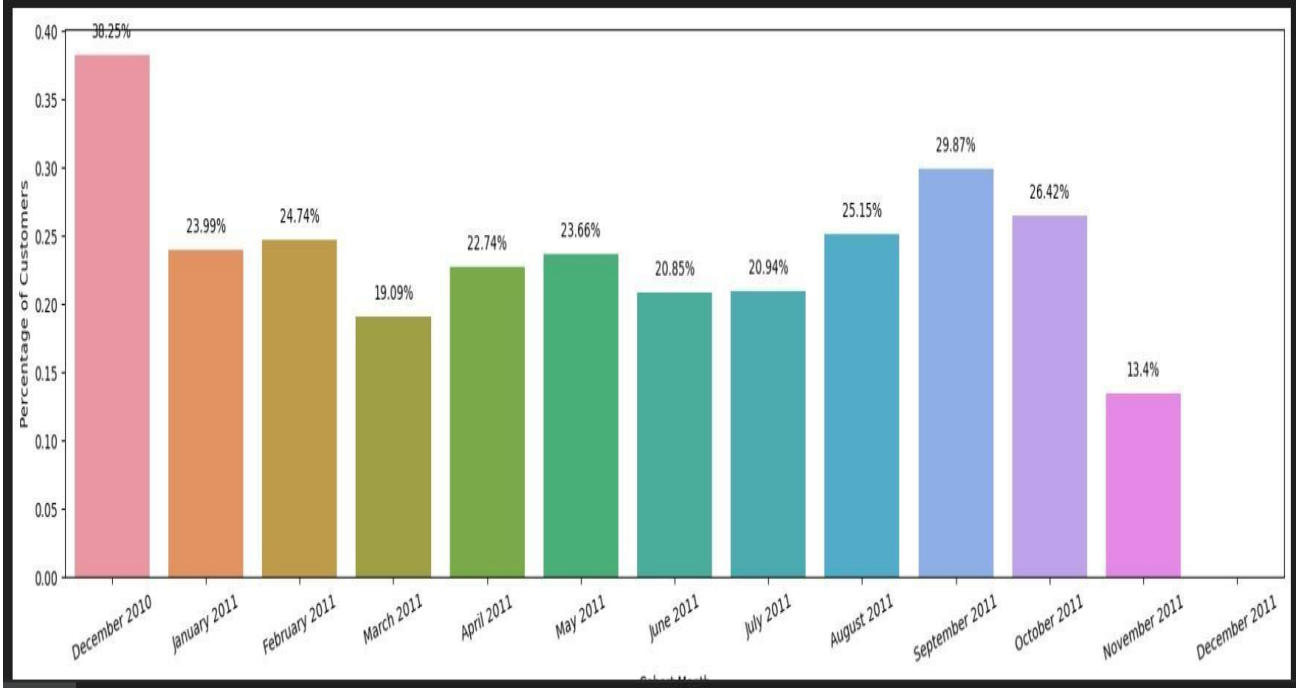
```
<Axes: xlabel='Cohort Month', ylabel='1'>
```



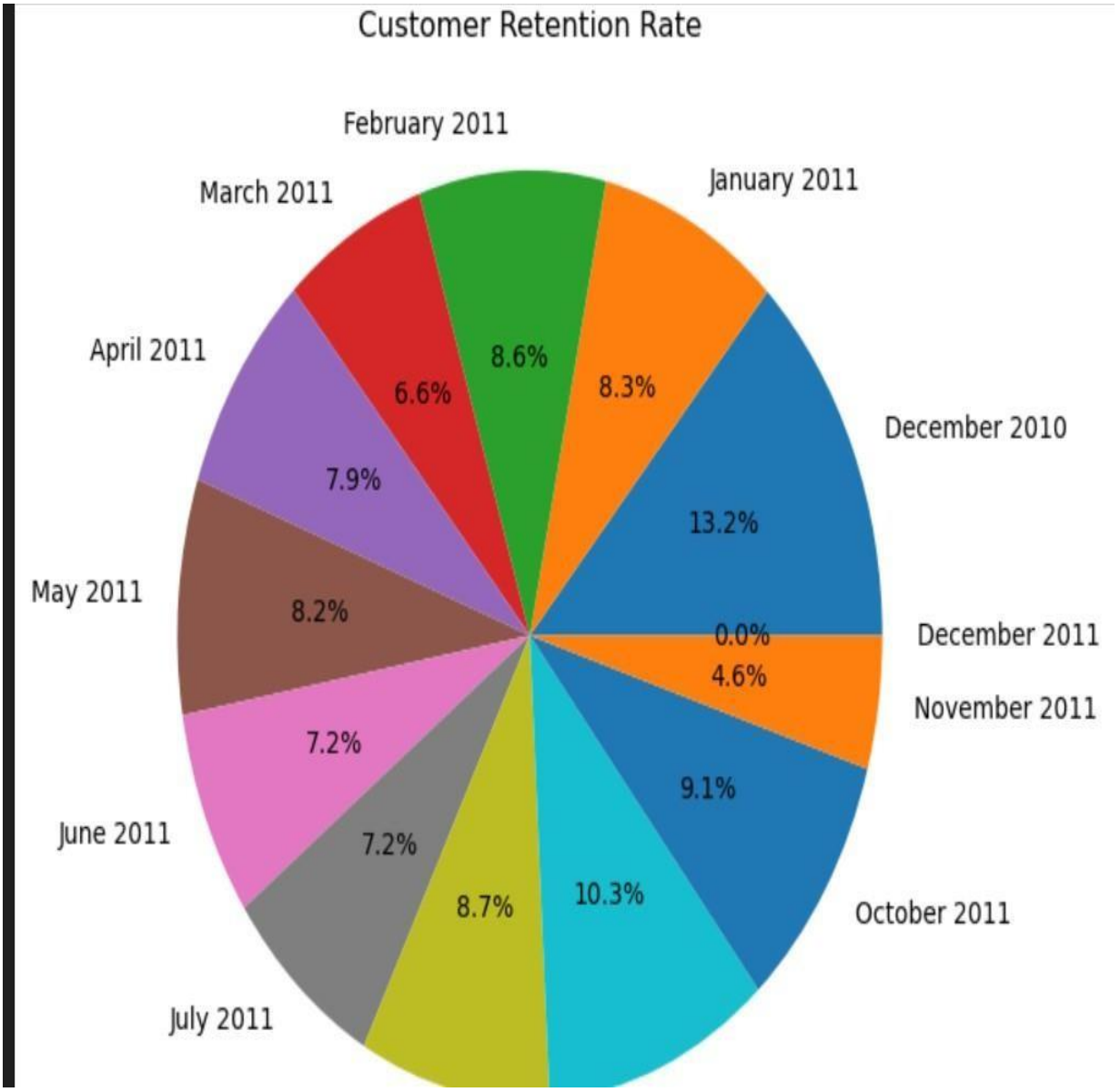
```
plt.xticks(rotation=20)
for i, v in enumerate(final_cohort_table[1]):
    plt.text(i, v + 0.01, str(round(v * 100, 2)) + '%', color='black', ha='center', va='bottom')
```

Python

```
posx and posy should be finite values
posx and posy should be finite values
posx and posy should be finite values
```



```
#creating a pie chart for above cohort analysis

plt.figure(figsize=(25,7))
plt.pie(final_cohort_table[1].fillna(0), labels=final_cohort_table.index, autopct='%1.1f%%')
plt.title('Customer Retention Rate')
plt.show()
```



7.CONCLUSION

Analyzing Customer Acquisition and Retention: A Comprehensive Cohort Analysis" provides valuable insights into understanding customer behavior over time, particularly focusing on acquisition and retention patterns. Through this analysis, businesses can gain a deeper understanding of their customer base, identify trends, and make data-driven decisions to improve marketing strategies, enhance customer satisfaction, and optimize product development.

In conclusion, "Analyzing Customer Acquisition and Retention: A Comprehensive Cohort Analysis". By segmenting customers into cohorts and analyzing their actions, businesses can gain valuable insights into acquisition patterns, retention rates, and customer lifetime value.

Through cohort analysis, businesses can identify key trends, opportunities, and challenges in their customer journey, allowing for targeted marketing strategies, personalized experiences, and enhanced customer engagement. This data-driven approach enables businesses to make informed decisions, allocate resources effectively, and ultimately improve customer satisfaction and loyalty.

REFERENCES

https://amplitude.com/blog/cohorts-to-improve-your-retention

https://www.researchgate.net/publication/304537951_Customer_Acquisition_and_Customer_Retention_in_a_Competitive_Industry

https://fastercapital.com/content/Cohort-Analysis--How-Cohort-Analysis-Can-Improve-Your-Retention-Modeling-Strategy.html

https://www.moengage.com/blog/growth-tactic-1-how-to-use-cohort-analysis-to-measure-customer-retention/

https://experienceleague.adobe.com/docs/analytics/analyze/analysis-workspace/visualizations/cohort-table/cohort-analysis.html?lang=en