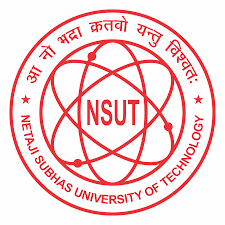
**Netaji Subhas University of Technology**

A STATE UNIVERSITY UNDER DELHI ACT 06 OF 2018, GOVT. OF NCT OF DELHI

Azad Hind Fauj Marg, Sector-3, Dwarka, New Delhi-110078



# REPORT FILE

COMPUTER HARDWARE AND SOFTWARE

**SUBMITTED TO:-**

**Mr. Rishabh Vats**

**GROUP MEMBERS:-**

Komal (2021UCS1623)

Shriya Satija (2021UCS1603)

Aman Deep (2021UCS1594)

## **Project Title: Customer Segmentation**

## **INTRODUCTION**

## Introduction Customer Segmentation is one the most important applications of unsupervised learning. Using clustering techniques, companies can identify the several segments of customers allowing them to target the potential user base. In this machine learning project, we will make use of k-mean Clustering which is the essential algorithm for clustering unlabelled dataset.

## **OBJECTIVE**

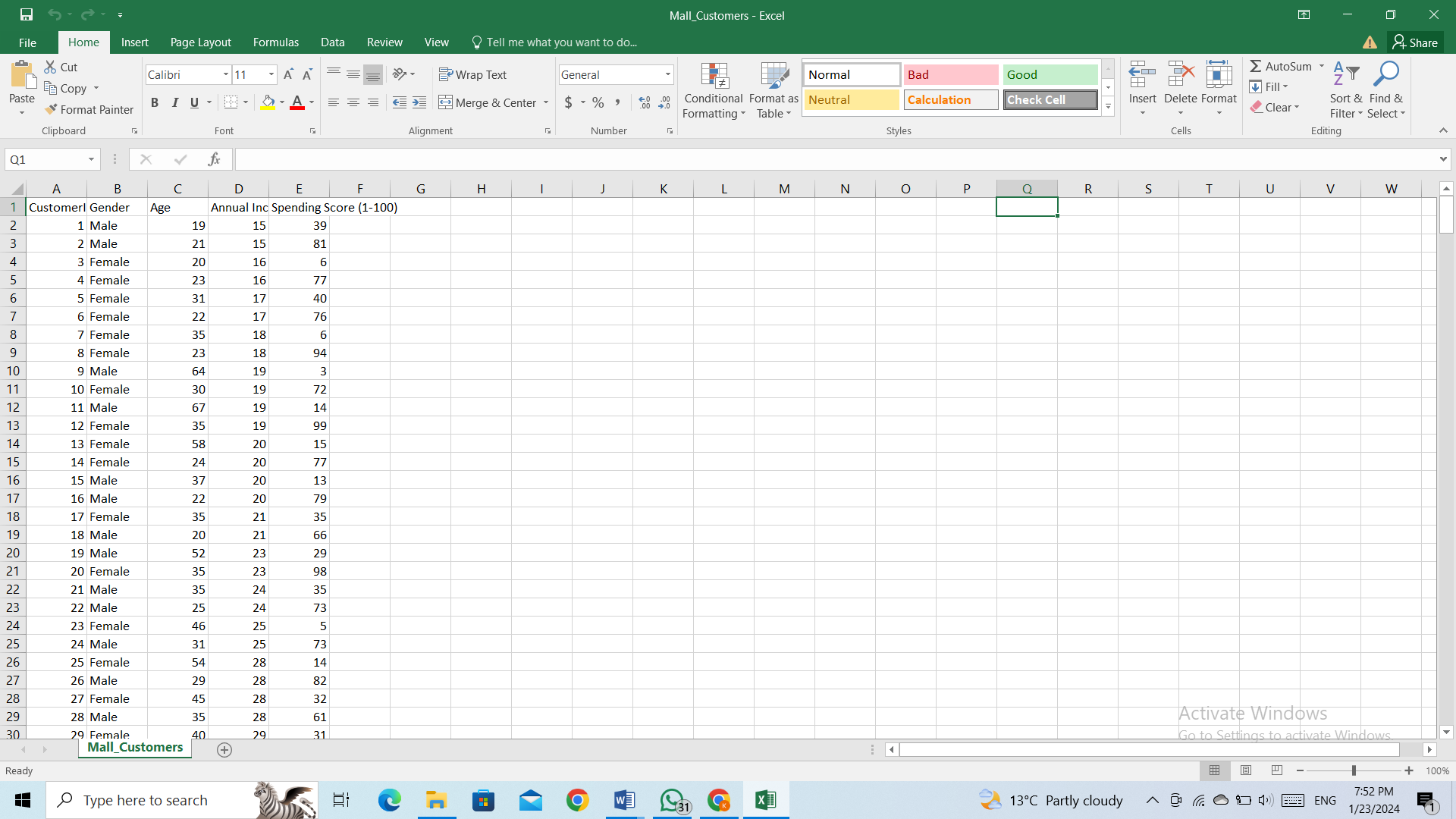
Whenever you need to find your best customer, customer segmentation is the ideal methodology. We will perform one of the most essential applications of machine learning – Customer Segmentation. In this project, we will implement customer segmentation in R.

## **R AND RSTUDIO**

R is a free software programming language and a software environment for statistical computing and graphics. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. R Studio is an integrated development environment (IDE) for R. IDE is a GUI, where you can write your quotes, see the results and also see the variables that are generated during the course of programming.

## **DATASET**

Mall\_Customers.csv



## **PACKAGES REQUIRED**

1. plotrix : it is an R package that provides many plotting, labeling, and axis & color scaling functions.
2. purrr : a complete and consistent functional programming toolkit for R. It provides a consistent and powerful set of tools for working with functions and vectors. It provides a set of functions that are designed to work with functional programming concepts, such as mapping, filtering, and reducing. These functions are designed to work with lists, data frames, and other objects, making it easier to work with complex data structures.
3. Cluster : it provides the methods for Cluster analysis.
4. gridExtra : it provides a number of user-level functions to work with "grid" graphics, notably to arrange multiple grid-based plots on a page, and draw tables.
5. nbClust : it provides 30 indexes for determining the optimal number of clusters in a data set and offers the best clustering scheme from different results to the user.
6. factoextra : extract and visualize the results of multivariate data analyses
7. ggplot2 : create elegant data visualisations using the grammar of graphics
8. dplyr : a grammar of Data Manipulation.

## **CUSTOMER SEGMENTATION**

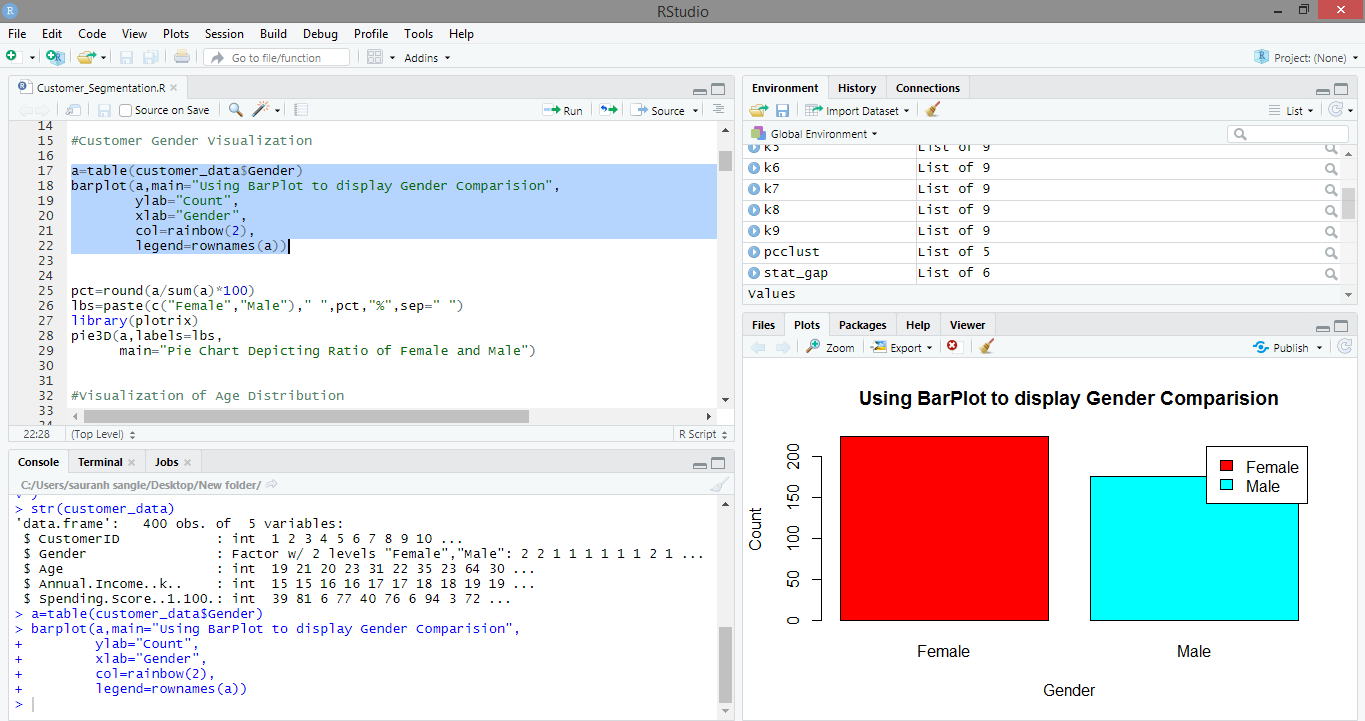
Customer Segmentation is the process of division of customer base into several groups of individuals that share a similarity in different ways that are relevant to marketing such as gender, age, interests, and miscellaneous spending habits.

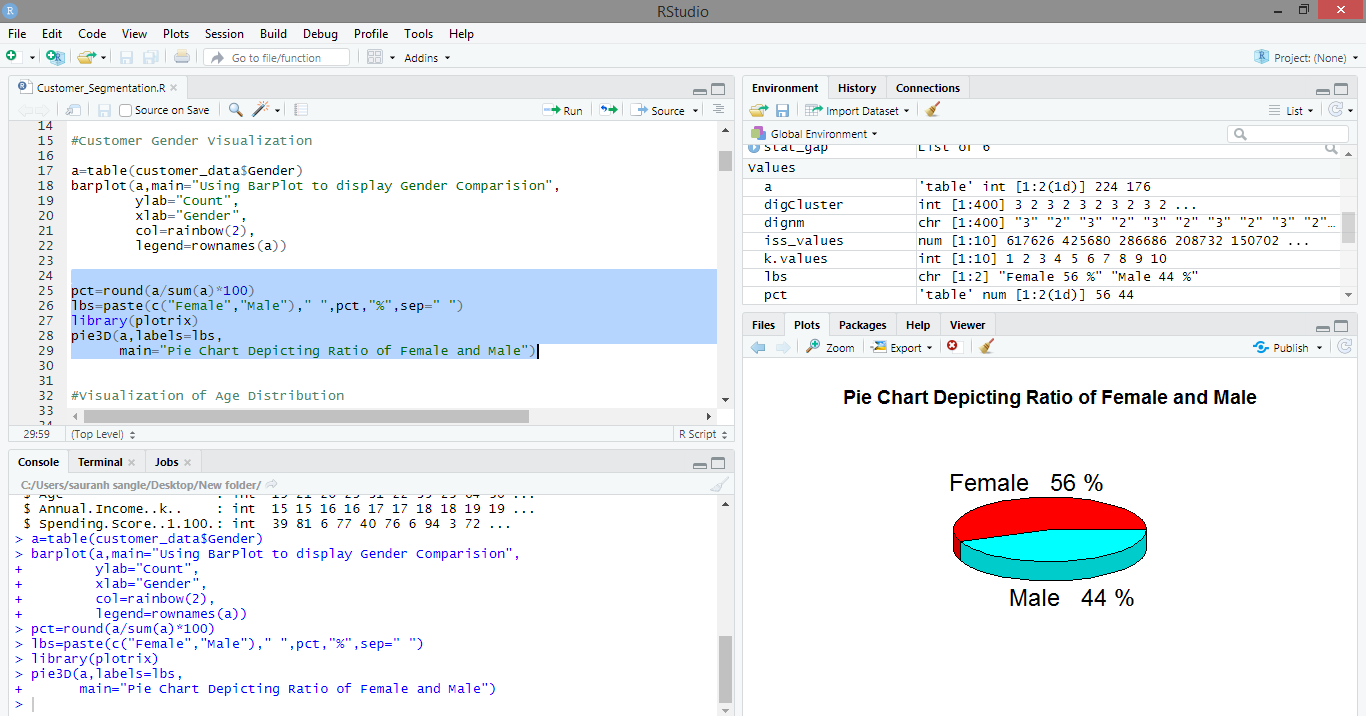
Companies that deploy customer segmentation are under the notion that every customer has different requirements and require a specific marketing effort to address them appropriately. Companies aim to gain a deeper approach of the customer they are targeting. Therefore, their aim has to be specific and should be tailored to address the requirements of each and every individual customer. Furthermore, through the data collected, companies can gain a deeper understanding of customer preferences as well as the requirements for discovering valuable segments that would reap them maximum profit. This way, they can strategize their marketing techniques more efficiently and minimize the possibility of risk to their investment.

The technique of customer segmentation is dependent on several key differentiators that divide customers into groups to be targeted. Data related to demographics, geography, economic status as well as behavioral patterns play a crucial role in determining the company direction towards addressing the various segments.

## **CUSTOMER GENDER VISUALISATION**

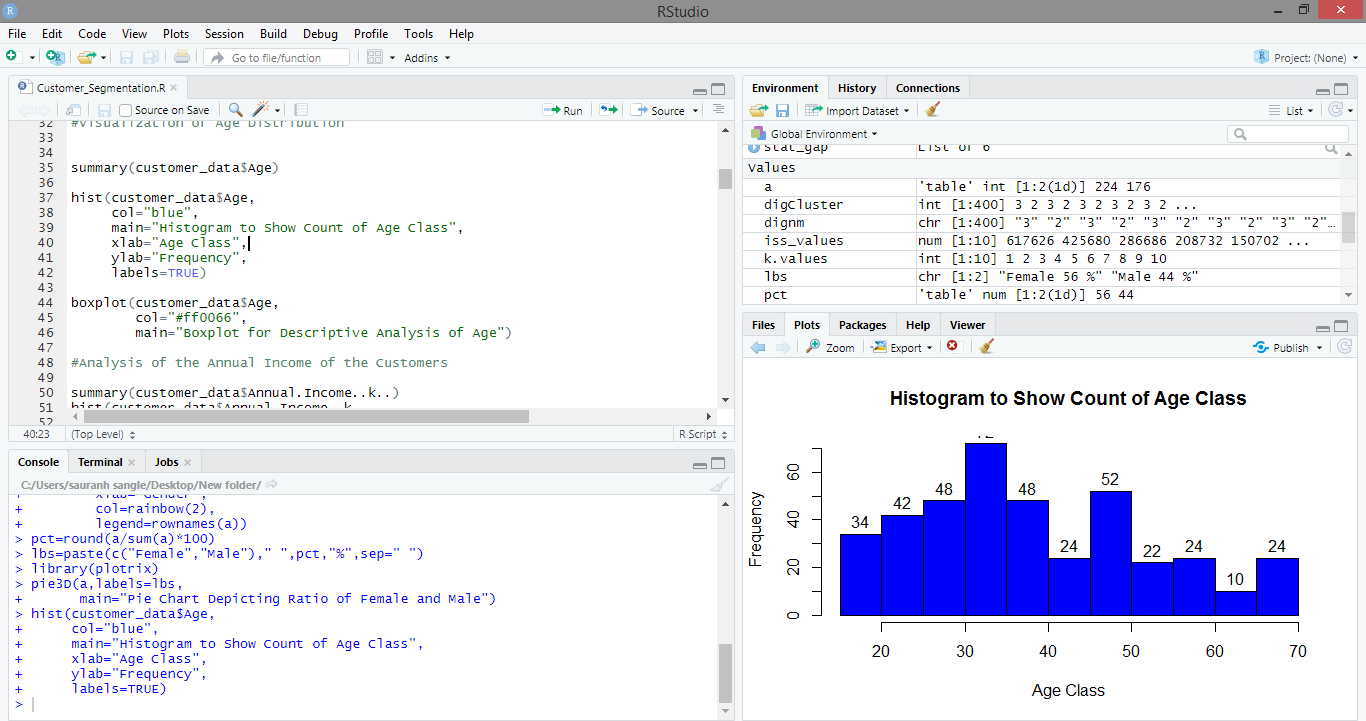
In this, we will create a barplot and a piechart to show the gender distribution across our customer\_data dataset. A bar chart represents data in rectangular bars with length of the bar proportional to the value of the variable. R uses the function barplot() to create bar charts. R can draw both vertical and Horizontal bars in the barchart. In bar chart each of the bars can be given different colors



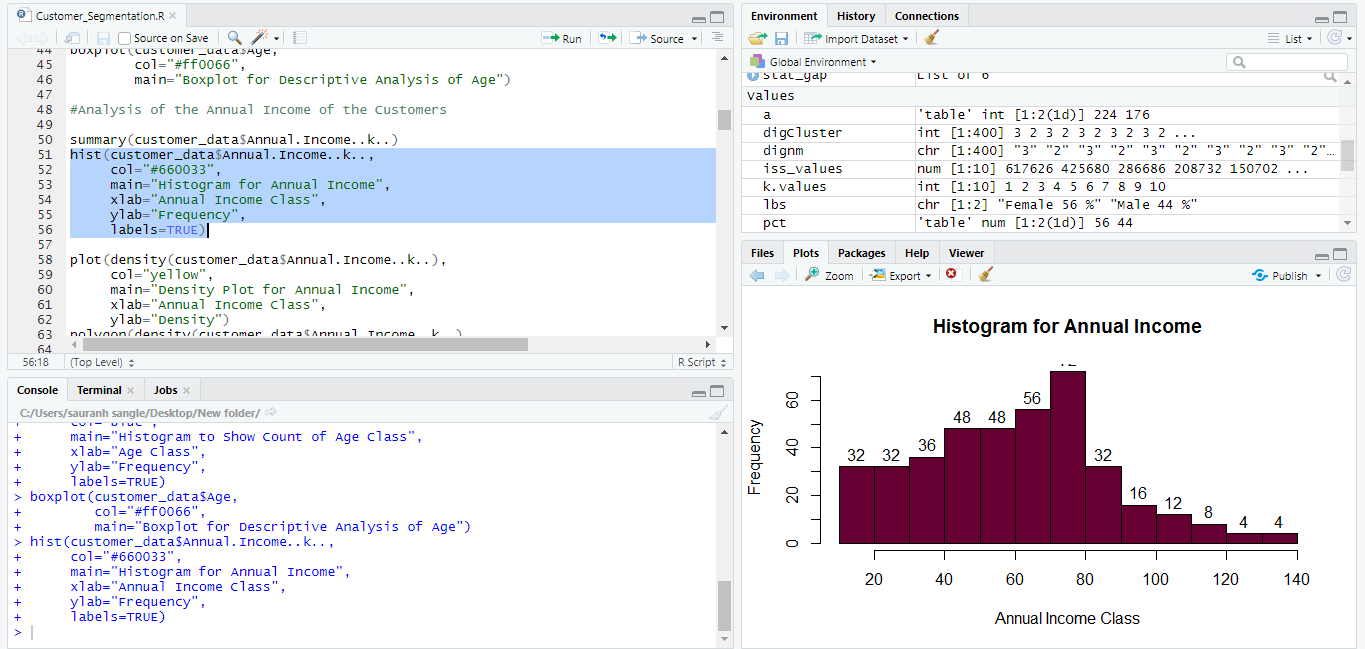


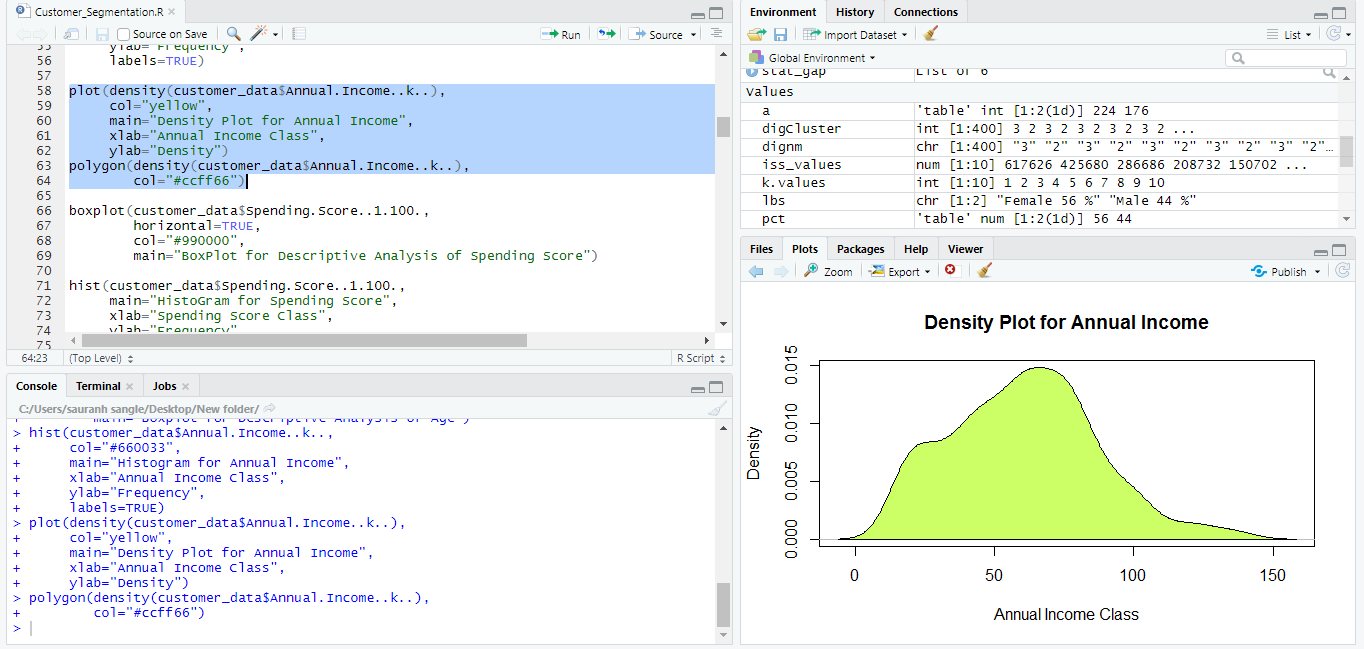
## **VISUALISATION OF AGE DISTRIBUTION**

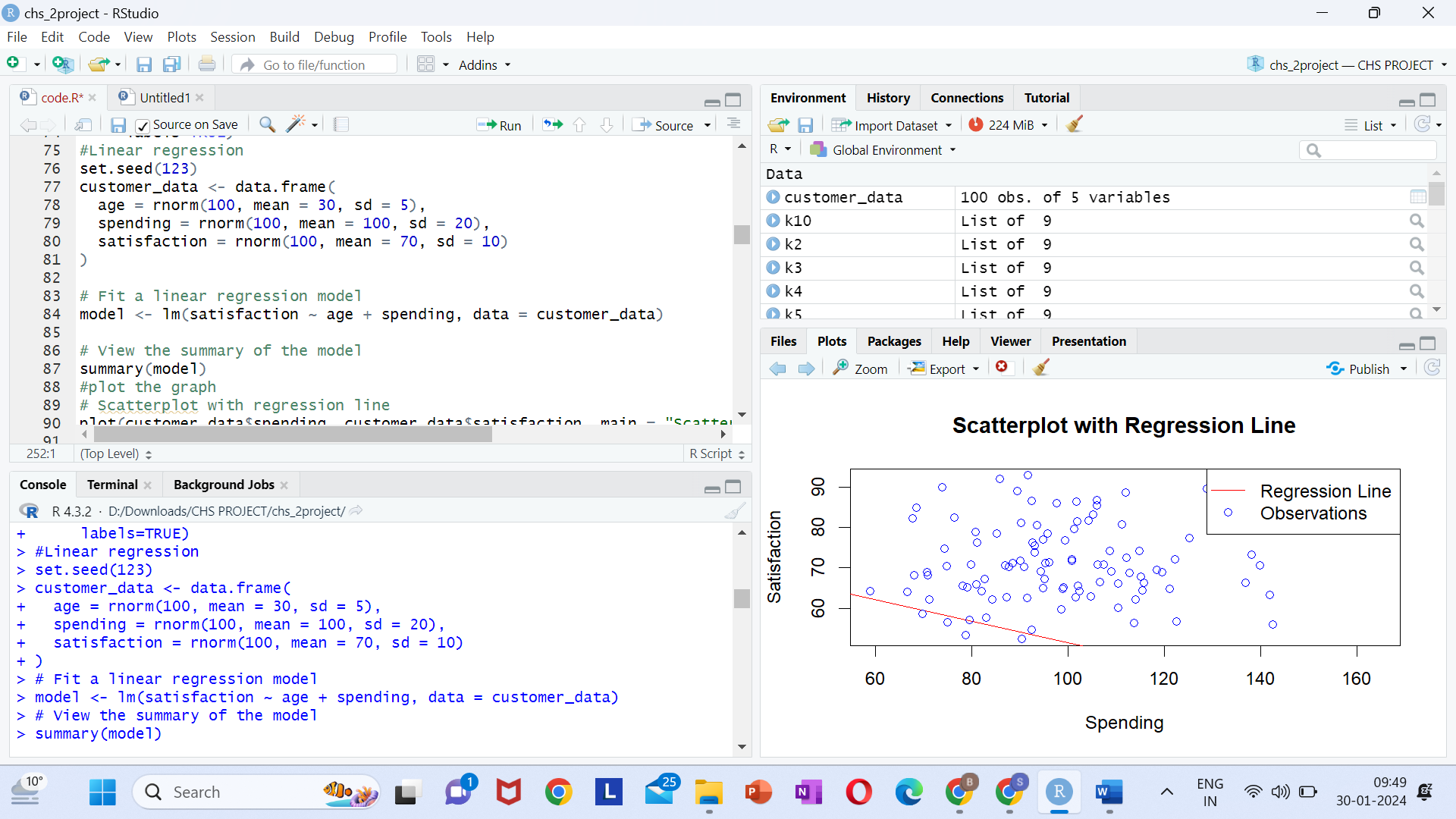
Let us plot a histogram to view the distribution to plot the frequency of customer ages. We will first proceed by taking summary of the Age variable.

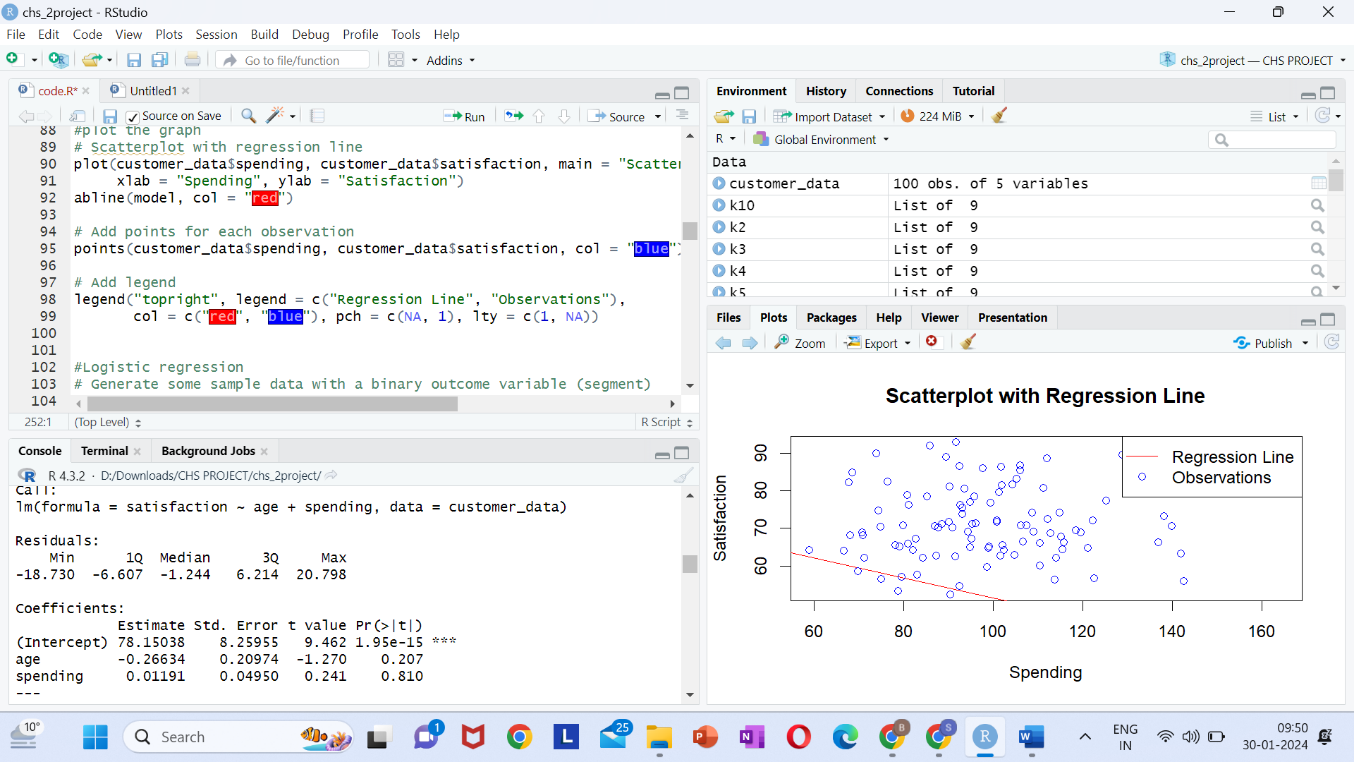
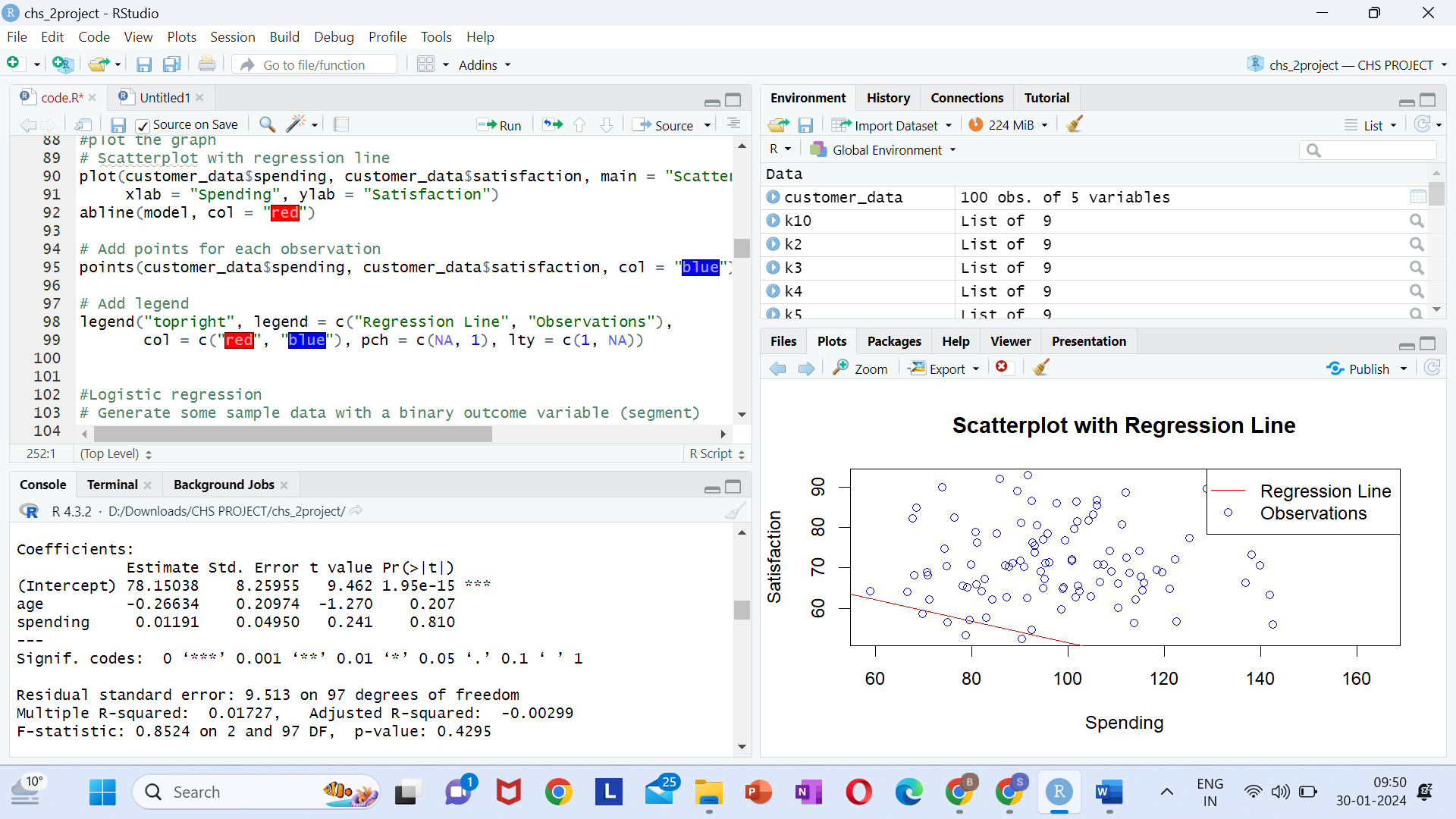


## **ANALYSIS OF ANNUAL INCOME**

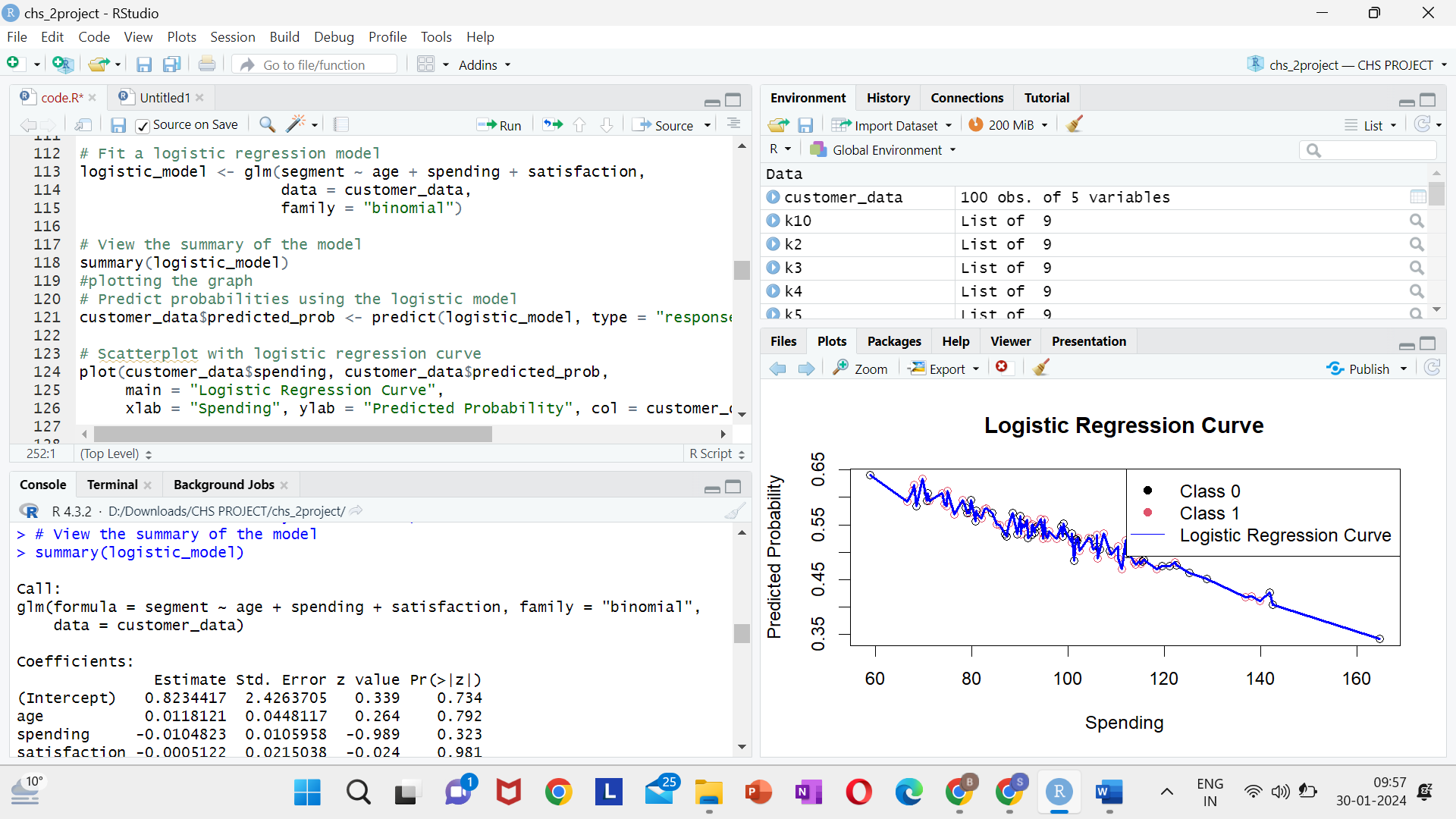
In this section of the R project, we will create visualizations to analyze the annual income of the customers. We will plot a histogram and then we will proceed to examine this data using a density plot.

**LINEAR REGRESSION**

****

****

## **LOGISTIC REGRESSION**



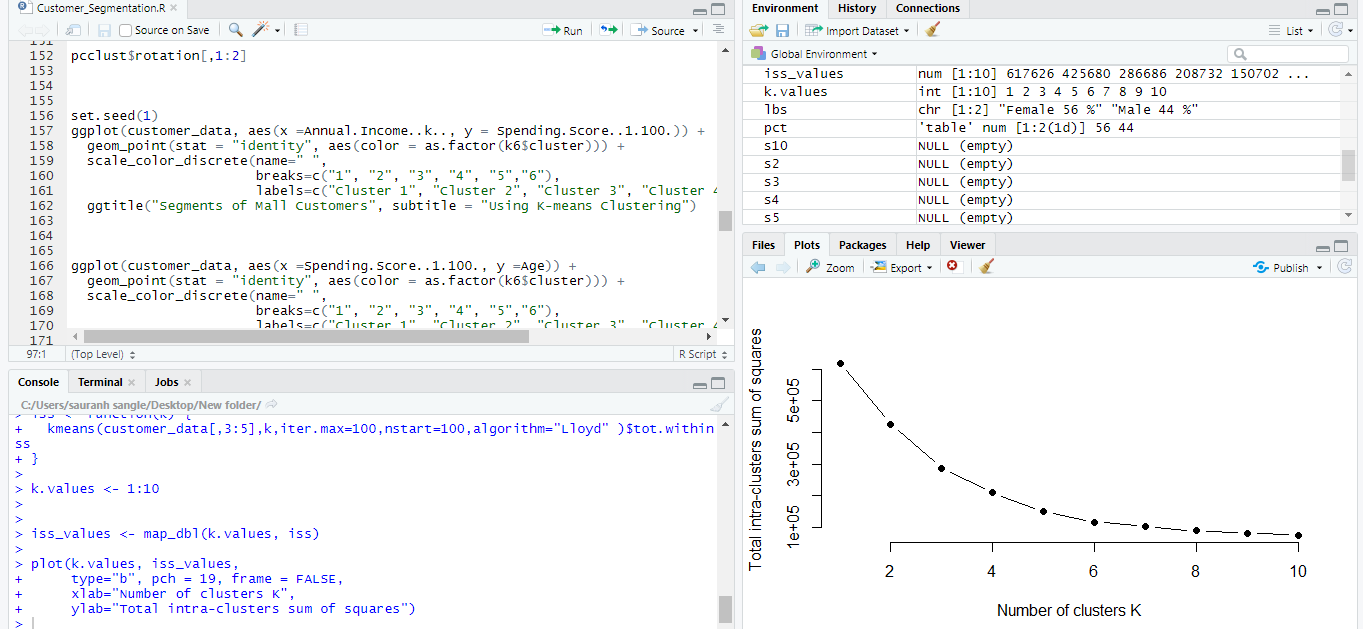
## 

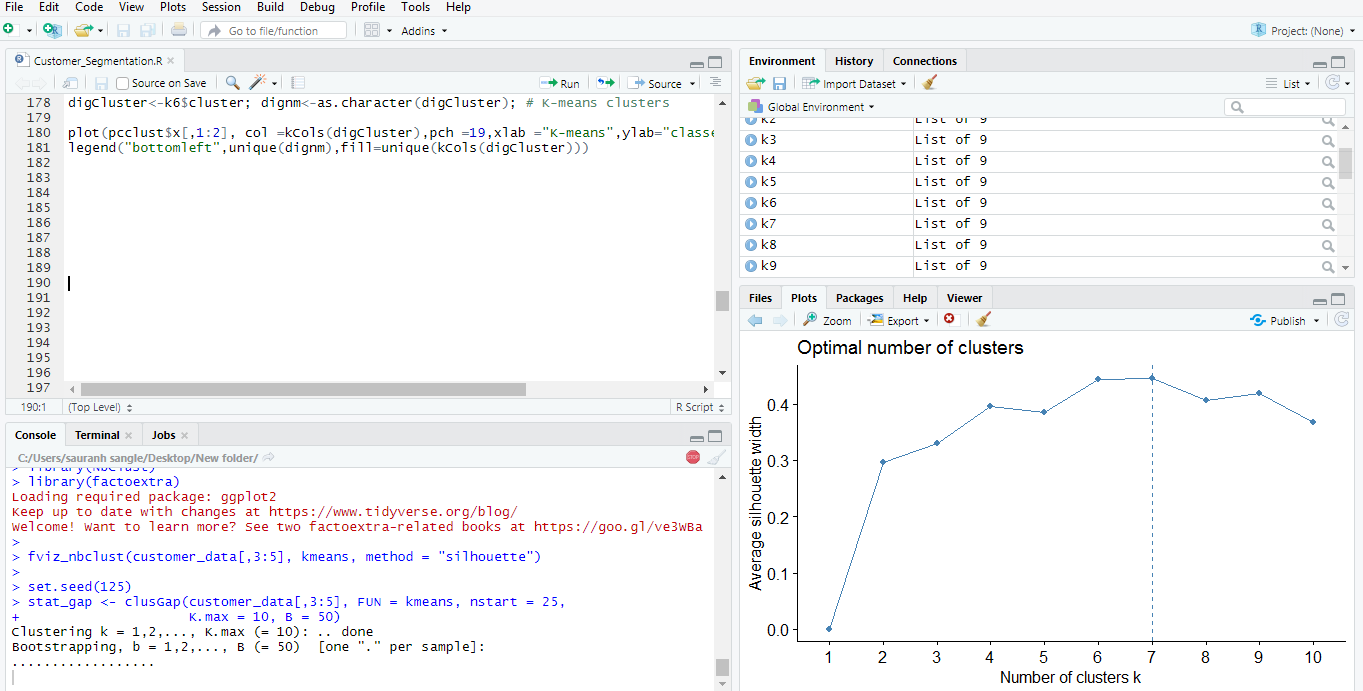
## **K-MEANS ALGORITHM**

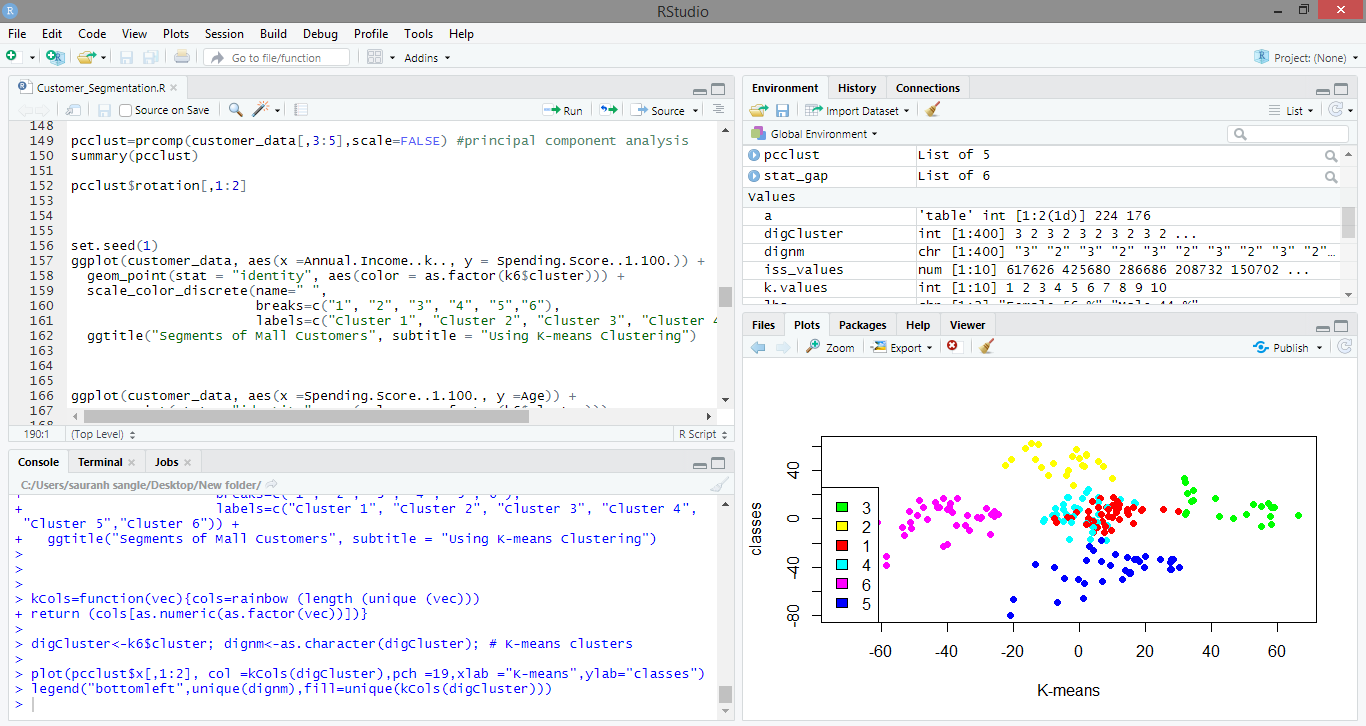
K-Means Clustering is an Unsupervised Machine Learning algorithm, which groups the unlabeled dataset into different clusters. We are given a data set of items, with certain features, and values for these features (like a vector). The task is to categorize those items into groups. To achieve this, we will use the K-means algorithm, an unsupervised learning algorithm. ‘K’ in the name of the algorithm represents the number of groups/clusters we want to classify our items into.(It will help if you think of items as points in an n-dimensional space). The algorithm will categorize the items into k groups or clusters of similarity. To calculate that similarity, we will use the Euclidean distance as a measurement.

The algorithm works as follows:

* First, we randomly initialize k points, called means or cluster centroids.
* We categorize each item to its closest mean, and we update the mean’s coordinates, which are the averages of the items categorized in that cluster so far.
* We repeat the process for a given number of iterations and at the end, we have our clusters.







## **CONCLUSION**

In this data science project, we went through the customer segmentation model. We developed this using a class of machine learning known as unsupervised learning. Specifically, we made use of a clustering algorithm called K-means clustering. We analysed and visualized the data and then proceeded to implement our algorithm. Hope you enjoyed this customer segmentation project of machine learning using R