**Abstract**

Statistical Analysis has been used to tackle simple to complex problems in diverse areas of interest and provide an evaluative quantitative solution using statistical techniques. We propose a narrative way of segmenting customers using retail customer data. We first identify factors of demographic customer behavior and validate the finding using statistical techniques. We enlist methodology for mapping the attribute of customer data to present a holistic customer segment. Demographic segments are generated from the demographic data fetched from the system in a transaction. Which includes transactional attributes such as age, gender, location & salary of these customers to generate personas. In this work, we present a statistical overview of some theoretical concepts, to provide a brief understanding of the techniques that can precisely identify demographic customer segments using real online transactional data from which we can achieve the personas of a real group of people.

Keywords: Personas, Customer Segmentation, Marketing.

**Introduction**

**1.1 The Business Problem**

Customer segmentation is a tactic for dividing the whole customer group based on distinctions across segments that are defined by a certain set of characteristics. In order to target particular customer groups with content and items that a consumers within a segment are likely to find relevant, e-commerce companies and other businesses rely on customer segmentation.

Organizations can interact with their clients or users more successfully based on this information. In order to maximize some performance parameters, such as task speed, buying preferences, or ease of use, ongoing efforts are made to discover and evaluate customers, an audience, or market segments in software design, marketing strategy, and advertising development.

For reliable results, statistical analysis must be carefully planned from the beginning of the research process. As well as defining your hypothesis, you must determine your study design, sample size, and sampling technique.

Once you have collected the data from your sample, you can organize and summarize it using descriptive statistics. Using inferential statistics, you can then test hypotheses and create population estimates. Lastly, you can generalize and analyze your results.

**1.2 Gathering of Data**

I have selected publicly available datasets for hypothesis testing using statistical techniques.

**Dataset:**

[**https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python**](https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python)

It is very important to collect data in a proper manner but also a framework that satisfies a wide range of monetary, moral, and computational requirements.

**1.3 Scope of Analysis**

In general, it would be simple to adapt the techniques employed to collect the data for this assessment to other relevant contexts or studies. Both investigating buying habits and developing an item-based collaborative filtering recommender system could benefit from using the same data, although neither of these is the subject of this work.

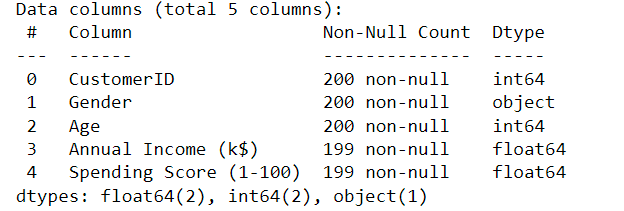
* Group customers for upcoming operations/marketing projects based on common purchasing habits.
* Find the correlation between the attributes of the given dataset for eg Gender & Income or Age & Customer Score.

**Methodology**

General Information on various methods that are important for this assessment, I’ve described the methods used in the data analysis. The following analytical models using statistical techniques were performed on the basis of quantitative data.

* T Test ( Student T-test)
* Chi-Square Test.
* Anova Test & Levene Test ( TEST OF VARIANCE )

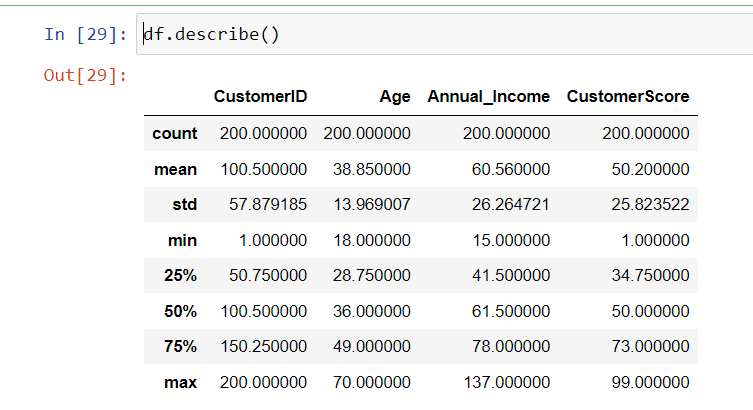
**Shape of Dataset : 200 Rows & 5 Columns**



**2.1 Dealing with the Null/Nan values**

Replacing the Nan Values present into the dataset with the population mean, to deal with the Null Values.

**2.2** **Checking the descriptive statistics about the data**



**2.2 Univariate Analysis**

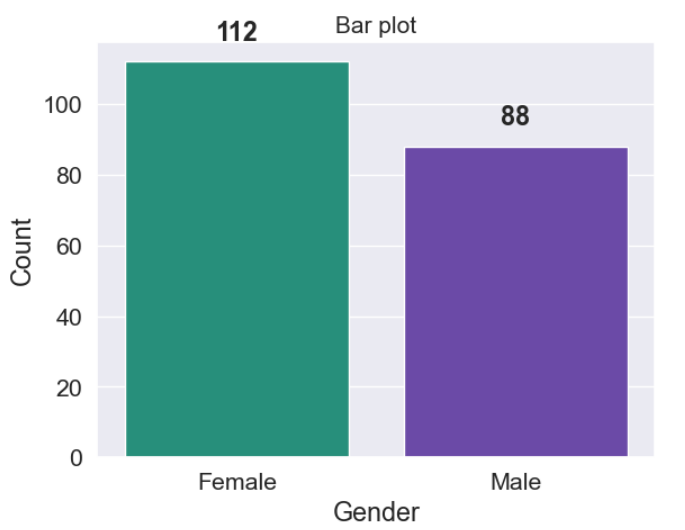
****

**Inference from 5-number theory:**

* Customer Age is Positively Skewed (Right skewed).
* Customer Annual Income & Customer Score is Roughly Normally Distributed.

**Distribution of Gender**

The Barplot illustrate graphically the distribution of Gender i.e Absolute Count

****

**STATISTICAL TECHNIQUES IMPLEMENTATION**

**T-Test : Observed Sample contains**

* One Categorical & One Continous(Numerical)
* Means from a continous variable ( for eg Age here ) taken from two groups.

**Step 1: State the null and alternative hypothesis:**

H0: μ1 = μ2 (Two population means of two Group with respect to Age is significant).

H1: μ1 ≠ μ2 (Two population means of two Group with respect to Age is not significant).

**Step 2: Calculate the t-value. & p-value:**

Variance of Male Age is 238.0

Variance of Female Age is 158.4

The difference between the bigger and smaller sample variances is expressed as a ratio = 238.0/158.4 ,1.5 which is less than 4. This means we can assume that population variance is equal

|  |  |
| --- | --- |
| **t-test statistic** | **0.86** |
| **A p-value** | **0.39** |

***Inference of T-test:***

* *Due to the p value > 0.05, the H0 hypothesis cannot be rejected*
* *Therefore, we do not have enough proof to claim that female and male ages are different in the population*

**Chi Square Test : Observed Sample contains**

* Both Variable are Categorial Data
* Both Variables are qualitative variables

**Checking association between gender and customer score.**

**Step 1: State Null hypothesis & Alternative hypothesis.**

H0: Gender and customer score are independent.

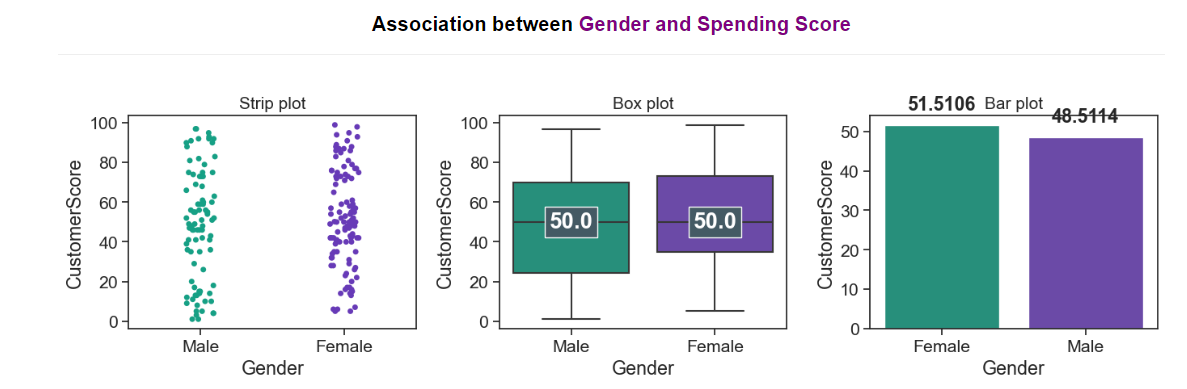
H1: Customer score depends on Gender.

**Step 2: Calculate the 𝜒2 test-statiscs & p-value:**

|  |  |
| --- | --- |
| **𝜒2 value** | **87.62** |
| **A p-value** | **0.39** |
| **Degree of Freedom** | **83** |

***Inference of Chi-Square-test:***

* *Since the pvalue is greater than 0.05 we fail to reject the Null hypothesis*
* *Null Hypothesis sustains i.e Gender and Customer Score are independent.*

**

**Pearson's Corelation Test : Observed Sample contains**

* Both Variables are Continuous (Numerical)

**Checking for Association between Annual Income and Customer Score**

**Step 1: State the null hypothesis & alternative hypothesis.**

H0: Annual Income and Customer Score are independent.

H1: Customer Score depends on Annual Income.

**Step 2: Calculate p-value:**

|  |  |
| --- | --- |
| **A p-value** | **0.01** |

We reject Null Hypothesis, Since p-value is smaller than 0.05 level of significance.

***Inference of Pearson's Corelation Test:***

* *we can reject the null hypothesis and conclude that we have sufficient evidence to say that the alternative hypothesis is true : Annual Income and Spending Score are correlated.*

**ANOVA : Observed Sample contains**

* one quantitative dependent variable and one category independent variable
* There ought to be a minimum of three levels or Categories for the independent variable

**Checking the Association between Age(Binned) and Customer Spending Score¶**

**1. Homogeneity of variance can be checked with the help of Levene's test.**

H0 : the variances are equal

H1 : the variances are not equal

|  |  |
| --- | --- |
| **A p-value** | **0.0004** |

**Since the p\_value < 0.05 we reject the Null hypothesis(h0) & Accept the h1 : the variances are not equal.**

**2. Performing ANOVA**

**Step 1: Define the Null hypothesis & Alternative hypothesis:**

H0: The mean score of all Age groups is equal).

H1: At least one to mean Annual Income of Age groups differ.).

**Step 2: Calculate the p-value:**

|  |  |
| --- | --- |
| **A p-value** | **1.425** |

Since the p\_value > 0.05 we reject the Null hypothesis(h0) & Accept the h1 : the variances are not equal.

***Inference of ANOVA:***

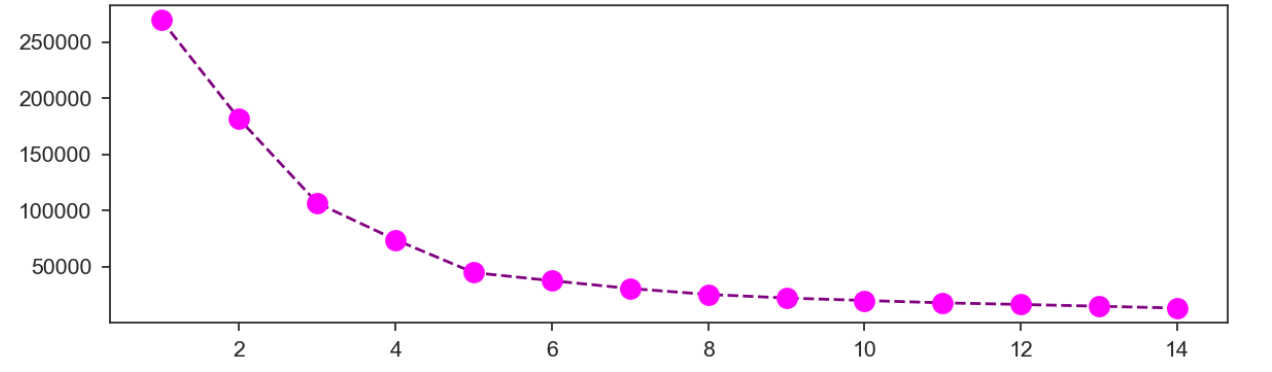
* *Due to the p-value > 0.05, we are unable to reject the H0 hypothesis*
* *Therefore, we do not have enough evidence to claim that male and female ages are different in the population*

**K-MEANS CLUSTERING**

* We can consider any number of features as an input to the Clustering Algorithm but to visualize the results at most I can consider only three features.
  + Features Considered:
    - Annual Income
    - Customer Score
* Gender is the least important feature here as per the statistical tests performed so I'm not considering it

**Determining optimum value of K using the Elbow Method**

1. The KMeans Algorithm's parameter k tells it how many clusters of the given data should be generated.
2. init = 'k-means++' ensures that the initial clusters are selected carefully rather than at random, increasing the likelihood of convergence.



***Inference of Elbow Method*** *:*

* *We notice that the inertia, or more precisely the mean squared difference between each point in a cluster and its cluster centre, diminishes at a slower rate after k=5.*
* *With k=5, which behaves like the elbow, we achieve a good compromise between the number of clusters that can form and inertia.*

**

***Inference of K-Means Cluster:***

* *We find that the annual income vs. spending plot best captures the groups that have formed.*
* *Customers who, although having lesser incomes, tend to spend a higher percentage of them make up the green group.*
* *What we typically anticipate is that customers that fall into the red, violet, and orange categories tend to exhibit a linear relationship between annual income and spend.*