Project Proposal

Analysis on Store sales and Customer Data

Spring 2023

Group Members

|  |  |  |
| --- | --- | --- |
| **Student ID** | **Name** | **Group** |
| 11546725 | Pragnesh Kumar Devarakonda | Group2 |
| 11580431 | Viswak sena palaparthi | Group2 |
| 11648583 | Pavan Kalyan Kumar boddu | Group2 |
| 11609230 | Madhuri Sri Yarramareddy | Group2 |

Table of Contents

Summary Of Research 4

Motivation of research 4

**Dataset and Details 5**

**Research questions and results 5**

# Summary Of Research

# Introduction

This document is used to explain the analysis of the dataset which contains the data related to the Customer details who have membership in a shop. The data explains the details of customers with few fields like profession, Annual income.

The dataset which will help the shop owner to understand about the business in the store.

# Motivation and Background

The dataset includes 2000 records and 8 columns which will explain about the customers who have membership with the store. The details include Customer ID, Gender, Age, Annual income, Spending Score-Score assigned by the shop, based on customer behavior and spending nature, Profession, Work experience, Family Size.

With this data the shop owner will have complete details about his business. There comes the analysis of the purchases and profits in the business. The dataset will play a key role in understanding the business, which is very important for the store owner, includes increasing the stocks based on the spending score of the customers, selling different types of items based on the customers need, etc.

# Dataset and details

The Dataset includes the details of customers of a store.



The research includes data of customer as the sample data to analysis the business of the store.

We got this dataset from the URL below.

<https://www.kaggle.com/datasets/datascientistanna/customers-dataset?resource=download>

**Exploratory Data Analysis:**

We performed an insights analysis on the customer data obtained from Kaggle by plotting and graphing the data. Based on the information provided in the data, we drew conclusions for the questions outlined in our proposal. The analysis was conducted using Python code to visualize the data.

* Python libraries such as NumPy, Pandas, Matplotlib, Plotly, and Seaborn were used for plotting and analysis of the customer data.
* Read csv. File using pandas libraries and imported that to a Dataframe df.
* Customer datasets have a total of 2000 records with fields Customer ID, Gender, Age, Annual Income ($), Spending Score (1-100), Profession, Work Experience, Family Size fields.
* Python library Pandas will help us to read the dataset csv file into a data frame.
* Below is the dataframe,

Using pandas libraries read the customer.csv file into a dataframe df, below is the dataframe.

Table

Description automatically generated with medium confidence

Found 35 records are having Null values in professions, below are the null value records from the dataframe.

A screenshot of a computer

Description automatically generated with low confidence

We tried to fill the Null values based on percentage of each profession and percentage of gender in the dataset with the missing records in the professions.

Below is the code,

A picture containing text, screenshot, font, line

Description automatically generated

The First line of code calculates the percentage of each profession in the available data. It filters out any rows where the profession value is missing, then groups the remaining rows by profession and counts the number of rows in each group. The count for each profession is then divided by the total number of rows where profession value is not missing to get percentage.

The Second line of code calculates the percentage of each gender in the entire dataset by grouping the rows by gender and counting the number of rows in each group, then dividing the count for each gender by the total number of rows in the dataset.

The last block of code fills in missing profession values by iterating over each row in the dataset where the profession value is missing. For each row, it randomly selects a profession from the available professions based on the percentage of each profession calculated in the first line of code. It also assigns the gender values from the row to the ‘Gender’ column and updates the ‘profession’ column with the randomly selected profession.

Before the filling, the count of Male and female in each profession are below.

A picture containing screenshot, line, plot, diagram

Description automatically generated

After filling the Null values below are the count of Male, Female with respect to the Professions.

A picture containing screenshot, line, plot, diagram

Description automatically generated

**Prediction Questions:**

**Can we predict which profession customers with an age above 25 more likely to have a membership.**

Tried to insight the data using pandas, Seaborn libraries. The graph shows Age verse Professions.

Mean, Median and interquartile of each profession which explain the ages which are more than 25. Minimum and Maximum of each profession are explained.

**A picture containing screenshot, colorfulness, diagram, design

Description automatically generated**

Using tableau, we tried to visualize the data for Profession and Average(Age) to match the data which we generated using python code and libraries.

A screenshot of a graph

Description automatically generated with medium confidence

Interquartile of ages verse professions.

A screenshot of a graph

Description automatically generated with medium confidence

**Can we predict the likelihood of a customer making a purchase based on their Family Size and Spending Score.**

Using the Bar plot, tried to figure out the Mean, max of the spending score verse family size.

A picture containing text, font, screenshot, line

Description automatically generated

**A picture containing diagram, line, rectangle, square

Description automatically generated**

From the plot the individual family size are having the mean, median and Interquartile.

A screenshot of a graph

Description automatically generated with medium confidence

A picture containing plot, line, diagram

Description automatically generated

The above line plot explains about how the spending scores are varying with respect to Family size.

**What is the likelihood of a customer having a high spending score based on their profession?**

A screen shot of a computer code

Description automatically generated with low confidence

Mean, Median, Mode of Profession with respect to spending score (1-100). Plot explains the mean, median and interquartile.

A picture containing screenshot, colorfulness, diagram, line

Description automatically generated

A screenshot of a graph

Description automatically generated with medium confidence

**How does the annual Income of customers relate to their spending score, particularly for those with the highest spending score? Can we use their relationship to predict the possible range of annual income for a customer with the highest spending score?**

**A screenshot of a computer program

Description automatically generated with low confidence**

The mean, Median and Interquartile range are statistical measures that describe the distribution of spending scores relative to the annual income of customers.

**A picture containing rectangle, square, line, diagram

Description automatically generated**

**A picture containing diagram, rectangle, square, screenshot

Description automatically generated**

**A picture containing screenshot, diagram, line

Description automatically generated**

**What Family Size group has the highest frequency of shopping among customers?**

**A picture containing rectangle, screenshot, square, line

Description automatically generated**

It has been observed that people who frequently shop have a higher likelihood of belonging to households with 2, 3, or 1 family members in that order, compared to households with 4 or 5 family members.

# What is the most common profession that has both a high mean income and a high mean spending score among customers?

A screenshot of a computer

Description automatically generated with medium confidence

A picture containing text, screenshot, colorfulness, parallel

Description automatically generated