**Data Analysis and Visualization of Shop Customer Information**

**Question 1**

**(a) Explain how to read the dataset and identify the dimensions of the dataset with Python program. State the answer and discuss the rationale.**

**Code**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

file\_path = 'ANL252\_Customers.csv'

customer = pd.read\_csv(file\_path)

customer.head()

print("Number of dimensions:", customer.ndim)

print("Shape of the DataFrame:", customer.shape)

**Discussion**

Number of dimensions: 2

Shape of the DataFrame: (2000, 8)

Pandas to read and inspect data is a foundational skill in Python data analysis. The rationale behind using pd.read\_csv to load data is its efficiency and simplicity in handling CSV files .The use of .ndim and .shape attributes to inspect DataFrame's dimensions provides immediate insights into data's structure and guiding further data manipulation and analysis tasks. (Taherdoost, 2022)

**(b) Discuss why it is necessary to handle missing values. Use Python program to identify the variables with missing values in the given dataset.**

**Code**

missing\_values = customer.isnull().sum()

print("Missing values in each column:\n", missing\_values)

**Discussion**

Executing the code, we can see that the variables in the dataset that have missing values are ‘Gender’, ‘Annual Income’, ‘Profession’, with 1, 220, and 35 missing values respectively.

**Accuracy:** Missing values can lead to inaccurate analyses or predictions if not handled properly. Some statistical techniques and machine learning algorithms cannot handle missing data directly.

**Bias:** Ignoring missing values can introduce bias into your analysis and especially if the missingness is not random.

**Quality of Insights:** Quality decision-making and insights require complete information. Missing data can obscure patterns and relationships present in the data.

**Model Performance:** Most machine learning algorithms require complete datasets to train effectively. Missing values can reduce model accuracy and generalizability. Handling missing data through techniques such as imputation can improve model performance.

**(c) Propose ways to treat the missing data with Python and explain rationale(s) of the treatment(s).**

**Code**

# Removing rows where 'Gender' is missing

customer.dropna(subset=['Gender'], inplace=True)

# Replacing missing 'Profession' values with 'Unknown Profession'

customer['Profession'].fillna('Unknown Profession', inplace=True)

# Filling missing 'Annual Income ($)' values based on median of 'Work Experience' and 'Profession'

customer['filled annual income'] = customer.groupby(['Work Experience', 'Profession'])['Annual Income ($)'].transform(lambda x: x.fillna(x.median()))

customer.drop(columns=['Annual Income ($)'], inplace=True)

customer.rename(columns={'filled annual income': 'Annual Income ($)'}, inplace=True)

# Checking if there are still missing values for 'Annual Income ($)'

if customer['Annual Income ($)'].isnull().sum() > 0:

# Filling remaining missing 'Annual Income ($)' values based on median of 'Profession' only

customer['Annual Income ($)'] = customer.groupby('Profession')['Annual Income ($)'].transform(lambda x: x.fillna(x.median()))

# Final check for missing values

print("Missing values after treatment:\n", customer.isnull().sum())

**Discussion**

Now the missing value is 0 for every attribute

Given that there's likely a very small number of missing values in the 'Gender' column (as indicated by the preliminary analysis), removing these rows should have minimal impact on the overall dataset size and analysis outcomes.

Profession is a categorical variable, and missing values in such variables are often handled by adding a separate category, like 'Unknown Profession', for missing data. This approach allows for retention of rows with valuable information in other fields avoiding loss of data due to a missing 'Profession' value.

Annual Income ($) is a continuous numerical variable where missing values can significantly impact analysis outcomes especially in studies focused on financial aspects. Using median income of groups based on 'Work Experience' and 'Profession' to impute missing values is a methodologically sound choice because.

**(d) After missing data treatment in Question 1(c), analyse the dataset by providing three (3) charts and their corresponding tables using Python. Describe the insights and highlight any interesting observations.**

1. **Age Distribution Graph**

**Code**

# Age Distribution Chart

plt.figure(figsize=(10, 6))

sns.histplot(customer['Age'], bins=20, kde=True)

plt.title('Age Distribution of Customers')

plt.xlabel('Age')

plt.ylabel('Frequency')

plt.show()

# Age Distribution Table

age\_description = customer['Age'].describe()

print(age\_description)

**Discussion**

The combination of histogram and KDE provides a dual representation of data histogram shows actual data distribution in discrete intervals while KDE provides a continuous probability density curve.

The descriptive statistics table offers a numerical summary that might be used to support visual findings or provide additional details.

By examining height of the bars you can identify most common age ranges among customers. You can also look for patterns such as skewness or other distribution shapes. The KDE helps identify peaks (modes) in data which might not be as evident in histogram due to discretization into bins.(Diamond and Mattia, 2017)

1. **Annual Income vs. Spending Score Graph**

**Code**

# Annual Income vs. Spending Score Chart

plt.figure(figsize=(10, 6))

sns.scatterplot(x='Annual Income ($)', y='Spending Score (1-100)', data=customer, hue='Gender', style='Gender', alpha=0.6)

plt.title('Annual Income vs. Spending Score by Gender')

plt.xlabel('Annual Income ($)')

plt.ylabel('Spending Score (1-100)')

plt.legend(title='Gender')

plt.show()

# Correlation Table

income\_spending\_correlation = customer[['Annual Income ($)', 'Spending Score (1-100)']].corr()

print(income\_spending\_correlation)

**Discussion**

**Distribution of Data Points:** Each point represents a customer plotted according to their annual income and spending score.

**Clusters:** If any graph could show clusters of data points, which could indicate groups of customers with similar income and spending behaviors.

**Gender Differences:** By coloring the points by gender, it's possible to discern if there's a notable difference in spending score or annual income between genders.

A correlation coefficient near 1 or -1 indicates a strong positive or negative linear relationship respectively. A coefficient near 0 suggests little to no linear relationship. The more important value here is the correlation between Annual Income and Spending Score which appears to be approximately 0.023 suggesting a very weak positive linear relationship.

1. **Profession Distribution Graph**

**Code**

# Profession Distribution Chart

plt.figure(figsize=(12, 8))

sns.countplot(y='Profession', data=customer, order = customer['Profession'].value\_counts().index)

plt.title('Profession Distribution Among Customers')

plt.xlabel('Count')

plt.ylabel('Profession')

plt.show()

# Profession Distribution Table

profession\_counts = customer['Profession'].value\_counts()

print(profession\_counts)

**Discussion**

The graph presented is a horizontal bar chart that displays distribution of various professions among customers. Each bar's length represents count of customers within a particular profession. The chart is likely sorted by count in descending order showing most common professions at top and least common at bottom.

**Most to Least Common Professions:** You can quickly see which professions are most common among customers and which are less common.

**Distribution Spread:** The graph shows how customers are spread across different professional fields.

**Unknown Profession:** The category "Unknown Profession" is included resulting from missing data treatment where missing profession data was labeled as "Unknown." It's at bottom indicating that there are relatively few customers with an unspecified profession.

**References**

Taherdoost, Hamed, Different Types of Data Analysis; Data Analysis Methods and Techniques in Research Projects (August 1, 2022). International Journal of Academic Research in Management, 9(1):1-9, 2022 http://elvedit.com/journals/IJARM/wp-content/uploads/Different-Types-of-Data-Analysis-Data-Analysis-Methods-and-Tec, Available at SSRN: <https://ssrn.com/abstract=4178680>

Diamond, M. and Mattia, A. (2017) ‘Data visualization: An exploratory study into the software tools used by businesses’, *Journal of Instructional Pedagogies*, 18(1), pp. 1–7. Available at: http://www.aabri.com/copyright.html%0Ahttp://www.aabri.com/copyright.html%0Ahttps://eric.ed.gov/?id=EJ1151731.