Statistics and Trends

February 29, 2024

1 Statistics and Trends Assignment

1.1 Applied Data Science 1

1.1.1 Module Leader: Dr. William Cooper

1.2 Report written by:

- **Name:** Nazmul Hossain
- **Student ID:** 23015862
- **GitHub Repository:** nazmul-nil/statistics-and-trends

1.3 Project overview

This is a project report on Consumer Behavior and Shopping Habits (Zeesolver, 2022). With this dataset we can find answers of what people buy and why. We can get a relationship between people's age and choices of things, purchasing amount, their color choices, subscription status, frequency of purchases and a bunch of other things.

By analising this data, we can learn why people choose certain things. Whether it's clothes, shoes, or other stuff, this data helps us understand why we make the choices we do when we shop.

1.4 First let's import the necessary libraries

[21]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

1.5 Let's read dataframe using pandas

[22]:
```
# Set the Customer ID as the index column
df = pd.read_csv("archive/shopping_trends.csv", index_col = "Customer ID") #␣
 ↪shopping_trends.csv (Zeesolver, 2022)

# print the first 5 rows and try to get some insight
# df.head()
```

[23]:
```
# Find some basic statistics of the numerical data series from our dataframe
df.describe()
```

[23]:

	Age	Purchase Amount (USD)	Review Rating	Previous Purchases
count	3900.000000	3900.000000	3900.000000	3900.000000
mean	44.068462	59.764359	3.749949	25.351538
std	15.207589	23.685392	0.716223	14.447125
min	18.000000	20.000000	2.500000	1.000000
25%	31.000000	39.000000	3.100000	13.000000
50%	44.000000	60.000000	3.700000	25.000000
75%	57.000000	81.000000	4.400000	38.000000
max	70.000000	100.000000	5.000000	50.000000

[24]:
```
# Make a sub dataframe with only two columns: Age and Purchase Amount(USD)

sub_dataframe = df[['Age', 'Purchase Amount (USD)']]
sub_dataframe.corr(method='pearson')
```

[24]:

	Age	Purchase Amount (USD)
Age	1.000000	-0.010424
Purchase Amount (USD)	-0.010424	1.000000

Correlation Coefficient (Pearson): - The correlation coefficient measures the strength and direction of the linear relationship between two variables. (Mukaka, 2012; Anwar et al, 2022)

From the correlation matrix found from our two variables Age and Purchase Amount(USD):

- For the 'Age' and 'Purchase Amount (USD)' variables, the correlation coefficient is approximately -0.0104.
- A correlation coefficient close to 0 means a weak linear relationship between the variables. (Székely et al., 2007; Khong, 2009; Baron, 2016)
- Also a correlation coefficient zero means that there is no linear relation at all between the variables. (Mukaka, 2012)
- The negative correlation coefficient suggests a slight negative linear relationship between age and purchase amount. (Mukaka, 2012)
- However, in our case the correlation is very close to zero. Which indicats that the relationship between age and purchase amount is extremely weak.
- We can say that, there is very little predictable relationship between a person's age and the amount they spend on purchases.

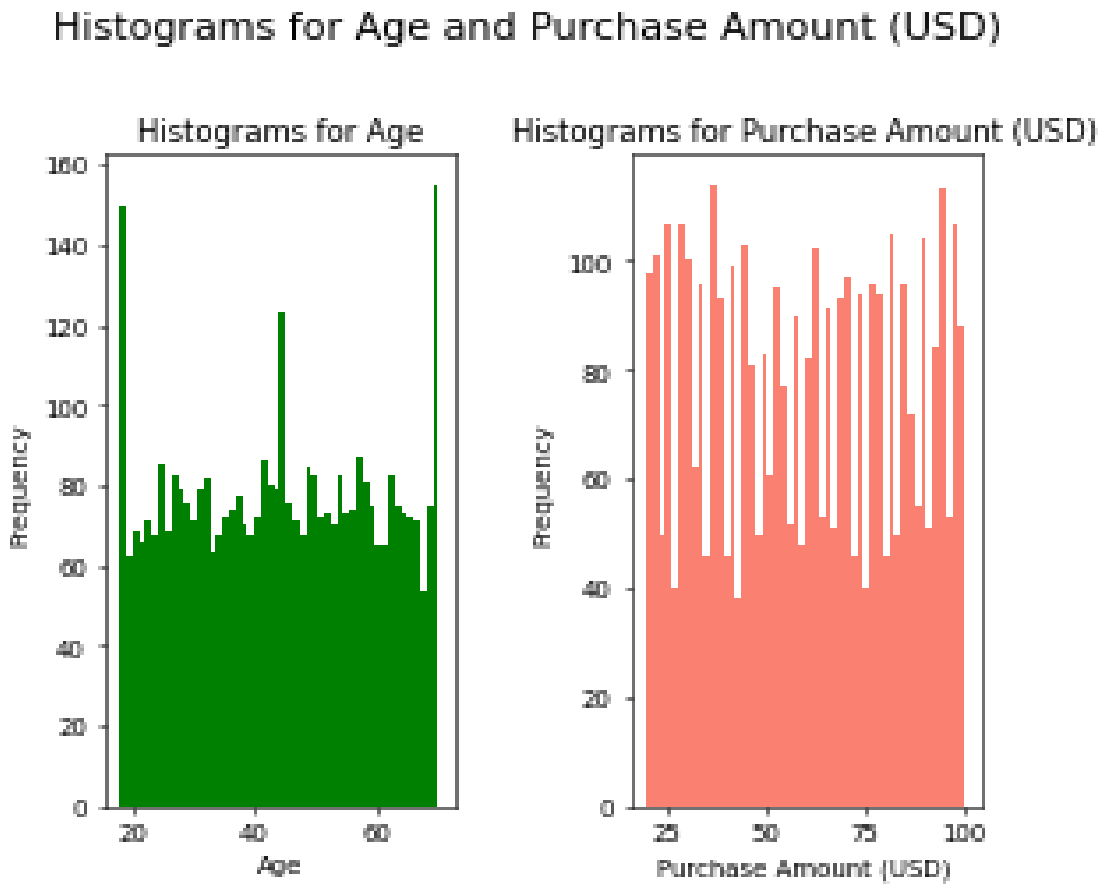
Conclusion: -We can draw a conclusion that age does not have a significant impact on the purchase amount based on this correlation analysis.

1.6 Let's build some functions and do some visualisations

[25]:
```
def age_purchase_amount_hist(df_age, df_purchase_amount):
    """This function will plot a histogram of purchasing amount of consumers vs␣
 ↪their age"""
```

```
    # Initialize figures with 2 subplots in a single row
    fig, axs = plt.subplots(1, 2, dpi=50)

    axs = axs.flatten()

    subdata_series = [df_age, df_purchase_amount]

    colors = ['green', 'salmon']

    # Labels for the axes
    xlabels = ['Age', 'Purchase Amount (USD)']

    for i, ax in enumerate(axs):
            #Plots the histogram
            ax.hist(subdata_series[i], color=colors[i], bins=50)

            # Set label for the axes and title for each of the plot
            ax.set_xlabel(xlabels[i])
            ax.set_ylabel('Frequency')
            ax.set_title('Histograms for ' + xlabels[i])

    # Adjust spacing between the subplots and set title in the middle
    plt.subplots_adjust(wspace=0.5)
    fig.suptitle('Histograms for Age and Purchase Amount (USD)', fontsize=16,␣
 ↪y=1.05)

    return
```

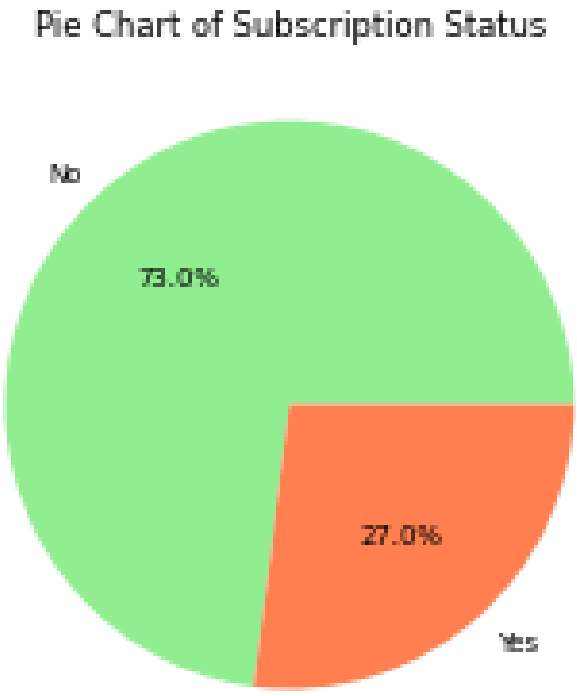
[26]:
```
# Make sub data series from the Age and Purchase Amount (USD) columns of our␣
 ↪dataframe
df_age = df['Age']
df_purchase_amount = df['Purchase Amount (USD)']

age_purchase_amount_hist(df_age, df_purchase_amount)
```



[27]:
```
# Create a pie chart of Subscription Status

def subscription_counts_pie(subscription_counts):
    """This function will show the number of subscriptions in a pie chart"""

    # Initialize a figure
    plt.subplots(dpi=50)

    # Plot the pie chart
    plt.pie(subscription_counts, labels=subscription_counts.index, autopct='%1.
↪1f%%', colors=['lightgreen', 'coral'])

    # Set the title for the plot
    plt.title('Pie Chart of Subscription Status')
    plt.show()
    return
```

[28]:
```
subscription_counts = df['Subscription Status'].value_counts()
```

```
subscription_counts_pie(subscription_counts)
```

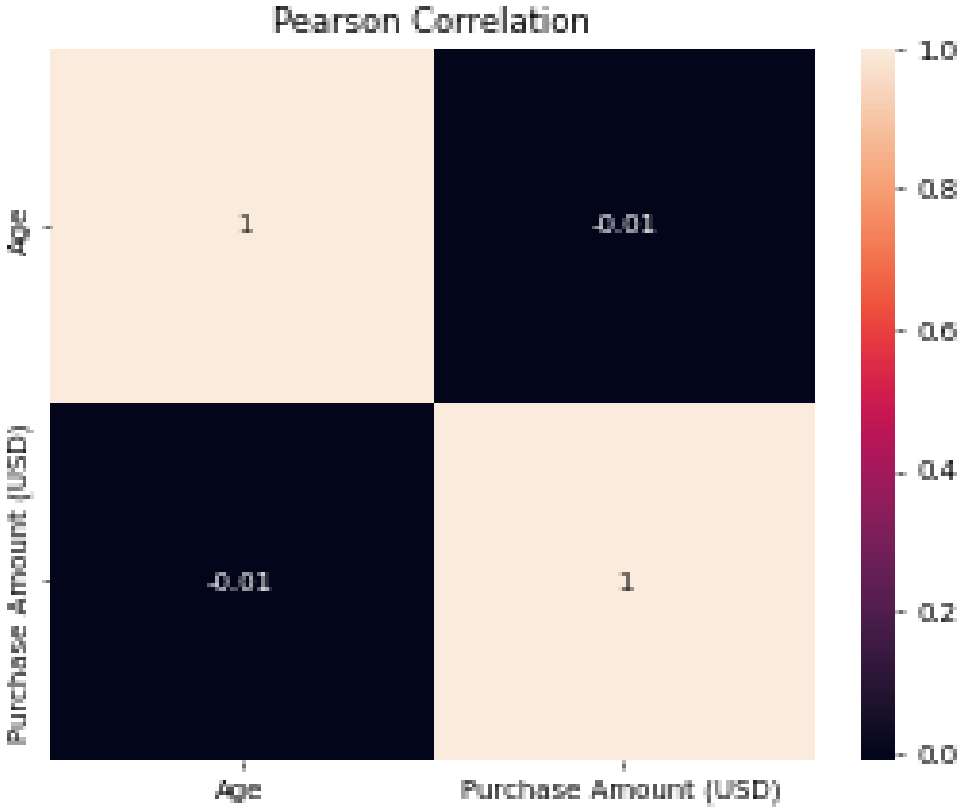


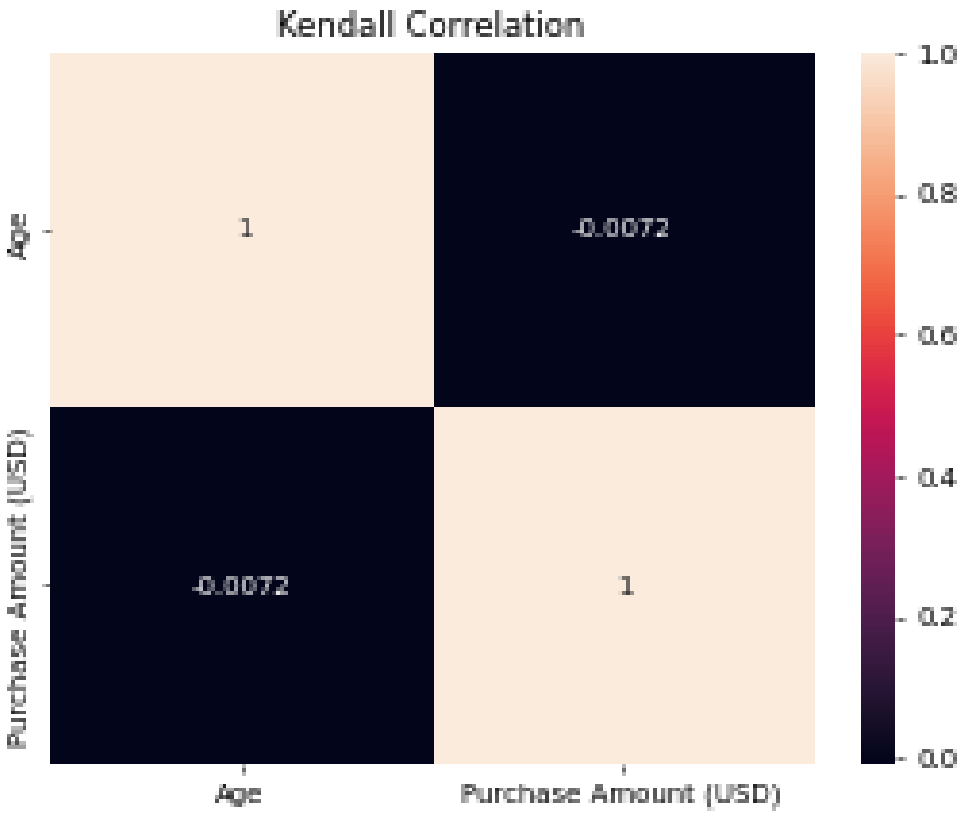
[29]:
```
def age_vs_purchase_amount_scatter(df_age, df_purchase_amount):
    """This function will plot a scatter plot of purchasing amount of consumers␣
 ↪vs their age"""

    # Initialize a figure
    fig, ax = plt.subplots(dpi=50)

    ax.scatter(df_age, df_purchase_amount)

    # Set the label for x-axis and y-axis
    ax.set_xlabel('Age')
    ax.set_ylabel('Purchase Amount (USD)')

    # Set the title for the plot
    ax.set_title('Scatter Plot of Purchase Amount vs Age')
    plt.show()
    return
```

[30]:
```
age_vs_purchase_amount_scatter(df_age, df_purchase_amount)
```

[31]:
```
sub_dataframe = df[['Age', 'Purchase Amount (USD)',]]
sub_dataframe.corr(method='pearson')

def age_purchase_corr_heatmap(sub_dataframe, method):
    """This function will draw us a heatmap using seaborn package
    about correlation between different people of age and their purchasing␣
 ↪amount
    """

    # Initialize a figure

    fig, ax = plt.subplots(dpi=50)

    # Plots the heatmap of the correlation between the variables Age and␣
 ↪Purchase Amount (USD)
    sns.heatmap(sub_dataframe.corr(method = method), ax = ax, annot=True)

    plt.title(method.capitalize() + " Correlation")

    plt.show()
```

```
    return

age_purchase_corr_heatmap(sub_dataframe, 'pearson')

age_purchase_corr_heatmap(sub_dataframe, 'kendall')

# (Dr. William Cooper, 2024)
```





1.6.1 Reference list

Zeesolver. (2022). Consumer Behavior and Shopping Habits Dataset. Kaggle. https://www.kaggle.com/datasets/zeesolver/consumer-behavior-and-shopping-habits-dataset

Mukaka, M.M., 2012. A guide to appropriate use of correlation coefficient in medical research. Malawi medical journal, 24(3), pp.69-71. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3576830/

Anwar, I., Hamid, K., Ahmad, N., & Riaz, A. (2022). Sustainable Business and Society in Emerging Economies. Vol. 4, No 1, March 2022. Institute of Business Management Sciences, University of Agriculture, Faisalabad-Pakistan; Lyallpur Business School, Government College University, Faisalabad-Pakistan.

Székely, G. J. Rizzo; Bakirov, N. K. (2007). "Measuring and testing independence by correlation of distances". Annals of Statistics. 35 (6): 2769–2794. arXiv:0803.4101. doi:10.1214/009053607000000505. S2CID 5661488.

Khong, H. H. (2009). A computer simulation model for microwave link path loss prediction (p. 29) [Master's thesis, Department of Electrical and Computer Engineering, FAMU FSU College of Engineering, Florida State University]

BaronScorpion2240. (2018, December 28). MANAGEMENT C207. Retrieved from Western Governors University website.

Dr W. Cooper, 2024. “Basic Statistics and Statistical Graphs”. https://herts.instructure.com/courses/112173/pages/week-3-lecture-slash-tutorial-notes?module_item_id=3634554