Personal Project

David Carbajal

Goal / Motivation

Impact society in a positive way, by improving a local and small clothing business. Enhancing their knowledge of clothing trends, so they can make better educated store decisions.

How will this goal be accomplished?

Through the use of Statistical analysis on consumer shopping data, I would have the ability to make informed business and store suggestions.

Data Discovery and Discussion:

- Discovered the dataset on Kaggle Customer Shopping (Latest Trends)
- The dataset is a customer shopping dataset designed for analytical purposes. It includes details about the customer demographics, shopping behavior, and transactional information
- No missing information or values in this data set

The Data:

Through the use of a Customer Shopping Dataset, I was able to analyze Consumer Shopping Trends, which provided insights into Purchase Behavior and Patterns.

The Data set dove deep into consumer shopping patterns by analyzing 3,900 purchases, and 19 specific details of each purchase. These details consist of customer age, gender, item purchased, color, season, frequency of purchase, location, payment type, and numerous more details.

The specific store / stores the data originated from are not disclosed, therefore this analysis is not meant to be used for any specific clothing businesses. This project demonstrates the ability to make business suggestions through the use of statistical analysis.

The Data:

Customer I	Age	Gender	Item Purch	Category	Purchase A	Location	Size	Color	Season	Review Rat	Subscriptic	Payment M	Shipping Ty	Discount A	Promo Cod	Previous Pu	Preferred P	Frequency of Purchases
1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.1	Yes	Credit Carc	Express	Yes	Yes	14	Venmo	Fortnightly
2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	3.1	Yes	Bank Trans	Express	Yes	Yes	2	Cash	Fortnightly
3	50	Male	Jeans	Clothing	73	Massachus	S	Maroon	Spring	3.1	Yes	Cash	Free Shippi	Yes	Yes	23	Credit Carc	Weekly
4	21	Male	Sandals	Footwear	90	Rhode Islar	M	Maroon	Spring	3.5	Yes	PayPal	Next Day A	Yes	Yes	49	PayPal	Weekly
5	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring	2.7	Yes	Cash	Free Shippi	Yes	Yes	31	PayPal	Annually
6	46	Male	Sneakers	Footwear	20	Wyoming	M	White	Summer	2.9	Yes	Venmo	Standard	Yes	Yes	14	Venmo	Weekly
7	63	Male	Shirt	Clothing	85	Montana	M	Gray	Fall	3.2	Yes	Debit Card	Free Shippi	Yes	Yes	49	Cash	Quarterly
8	27	Male	Shorts	Clothing	34	Louisiana	L	Charcoal	Winter	3.2	Yes	Debit Card	Free Shippi	Yes	Yes	19	Credit Carc	Weekly
9	26	Male	Coat	Outerwear	97	West Virgin	L	Silver	Summer	2.6	Yes	Venmo	Express	Yes	Yes	8	Venmo	Annually
10	57	Male	Handbag	Accessorie	31	Missouri	M	Pink	Spring	4.8	Yes	PayPal	2-Day Ship	Yes	Yes	4	Cash	Quarterly
11	53	Male	Shoes	Footwear	34	Arkansas	L	Purple	Fall	4.1	Yes	Credit Carc	Store Picku	Yes	Yes	26	Bank Trans	Bi-Weekly
12	30	Male	Shorts	Clothing	68	Hawaii	S	Olive	Winter	4.9	Yes	PayPal	Store Picku	Yes	Yes	10	Bank Trans	Fortnightly
13	61	Male	Coat	Outerwear	72	Delaware	M	Gold	Winter	4.5	Yes	PayPal	Express	Yes	Yes	37	Venmo	Fortnightly
14	65	Male	Dress	Clothing	51	New Hamp	M	Violet	Spring	4.7	Yes	Debit Card	Express	Yes	Yes	31	PayPal	Weekly
15	64	Male	Coat	Outerwear	53	New York	L	Teal	Winter	4.7	Yes	PayPal	Free Shippi	Yes	Yes	34	Debit Card	Weekly
16	64	Male	Skirt	Clothing	81	Rhode Islar	M	Teal	Winter	2.8	Yes	Credit Carc	Store Picku	Yes	Yes	8	PayPal	Monthly
17	25	Male	Sunglasses	Accessorie	36	Alabama	S	Gray	Spring	4.1	Yes	Venmo	Next Day A	Yes	Yes	44	Debit Card	Bi-Weekly
18	53	Male	Dress	Clothing	38	Mississippi	XL	Lavender	Winter	4.7	Yes	Debit Card	2-Day Ship	Yes	Yes	36	Venmo	Quarterly
19	52	Male	Sweater	Clothing	48	Montana	S	Black	Summer	4.6	Yes	Bank Trans	Free Shippi	Yes	Yes	17	Cash	Weekly
20	66	Male	Pants	Clothing	90	Rhode Islar	M	Green	Summer	3.3	Yes	Venmo	Standard	Yes	Yes	46	Debit Card	Bi-Weekly
21	21	Male	Pants	Clothing	51	Louisiana	M	Black	Winter	2.8	Yes	Credit Carc	Express	Yes	Yes	50	Cash	Every 3 Months
22	31	Male	Pants	Clothing	62	North Caro	M	Charcoal	Winter	4.1	Yes	Credit Carc	Store Picku	Yes	Yes	22	Debit Card	Quarterly
23	56	Male	Pants	Clothing	37	California	M	Peach	Summer	3.2	Yes	Cash	Store Picku	Yes	Yes	32	Debit Card	Annually
24	31	Male	Pants	Clothing	88	Oklahoma	XL	White	Winter	4.4	Yes	Credit Carc	Express	Yes	Yes	40	Credit Carc	Weekly

Data Discovery and Discussion

I specifically focused on the following categories: Age, Gender, Purchase Amount, Categories of Items purchased, and Discounts / Promo Codes.

I began the project by specifically focusing on the ages of the customers in the dataset and dividing this research by gender. I found the following:

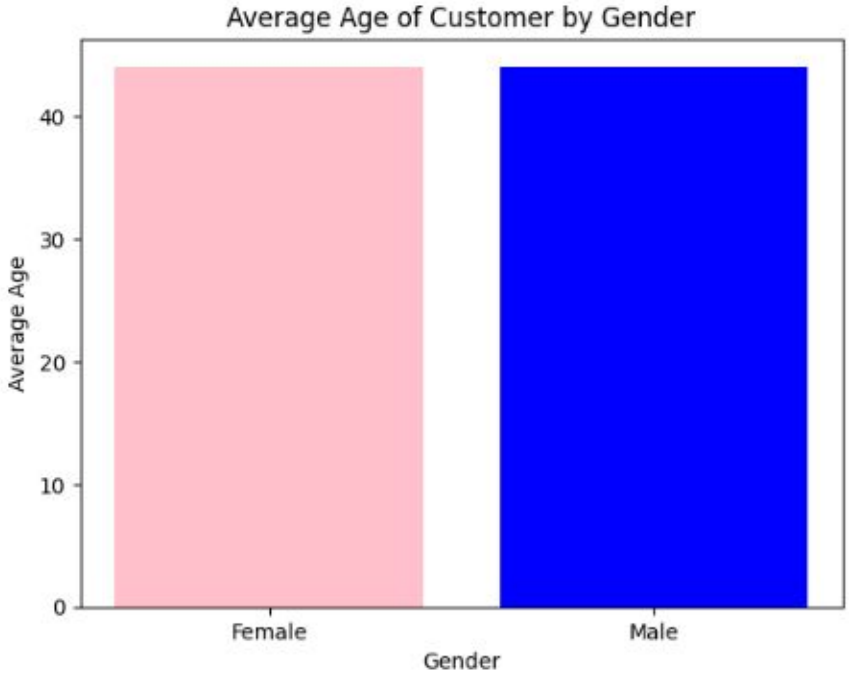
The ages of male customers range from: 18 to 70
The ages of female customers range from: 18 to 70

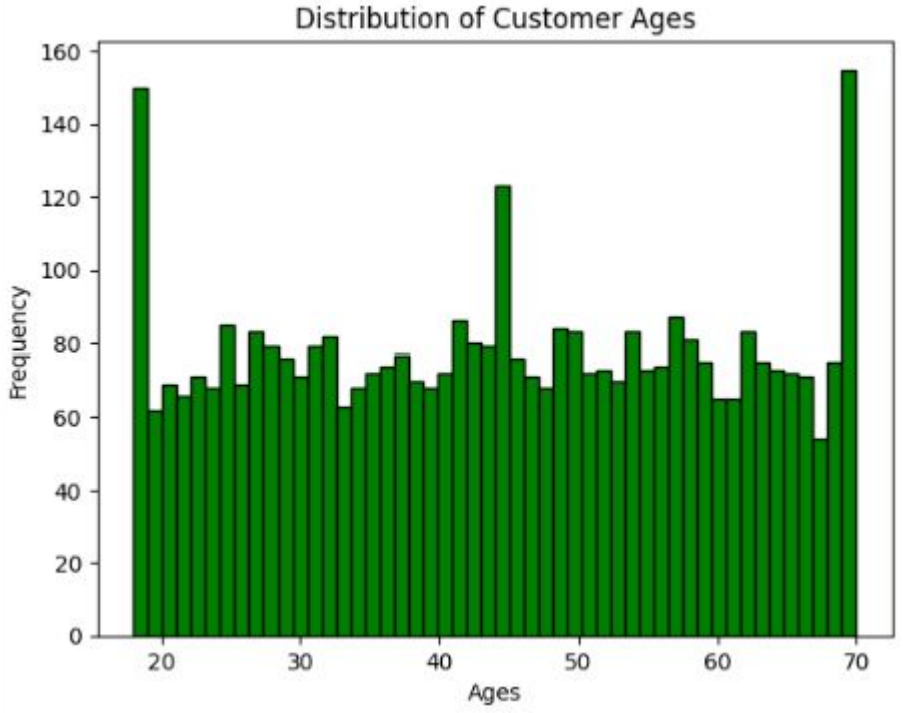
The average age of male customers is:
44.0972850678733
The average age of female customers is:
44.00721153846154

The median age of male customers is: 44.0
The median age of female customers is: 44.0

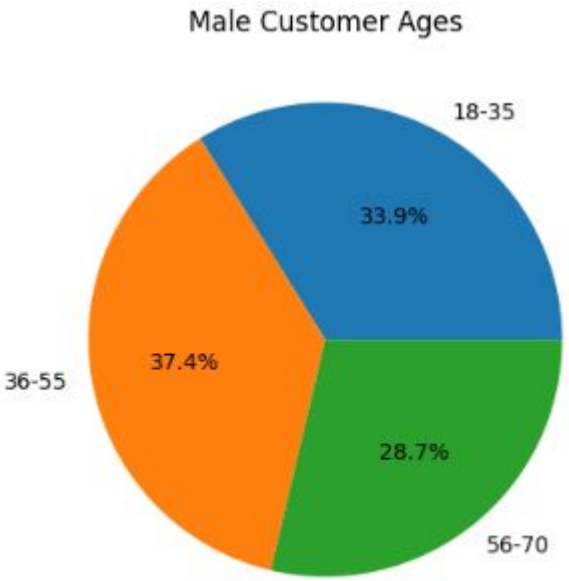
The mode age of male customers is: 54
The mode age of female customers is: 45

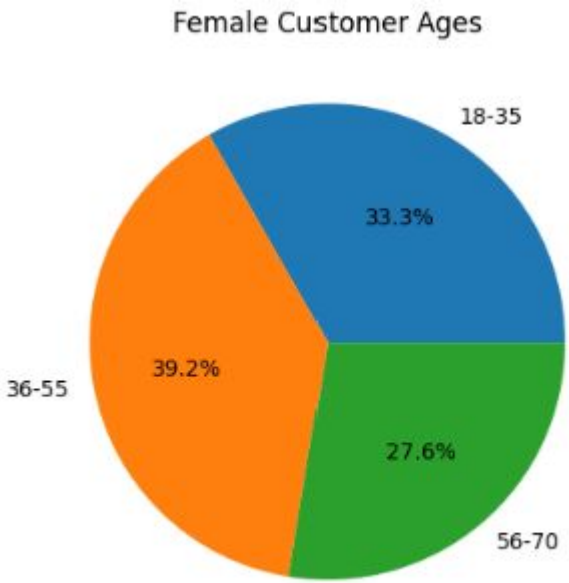
Exploratory Analysis:



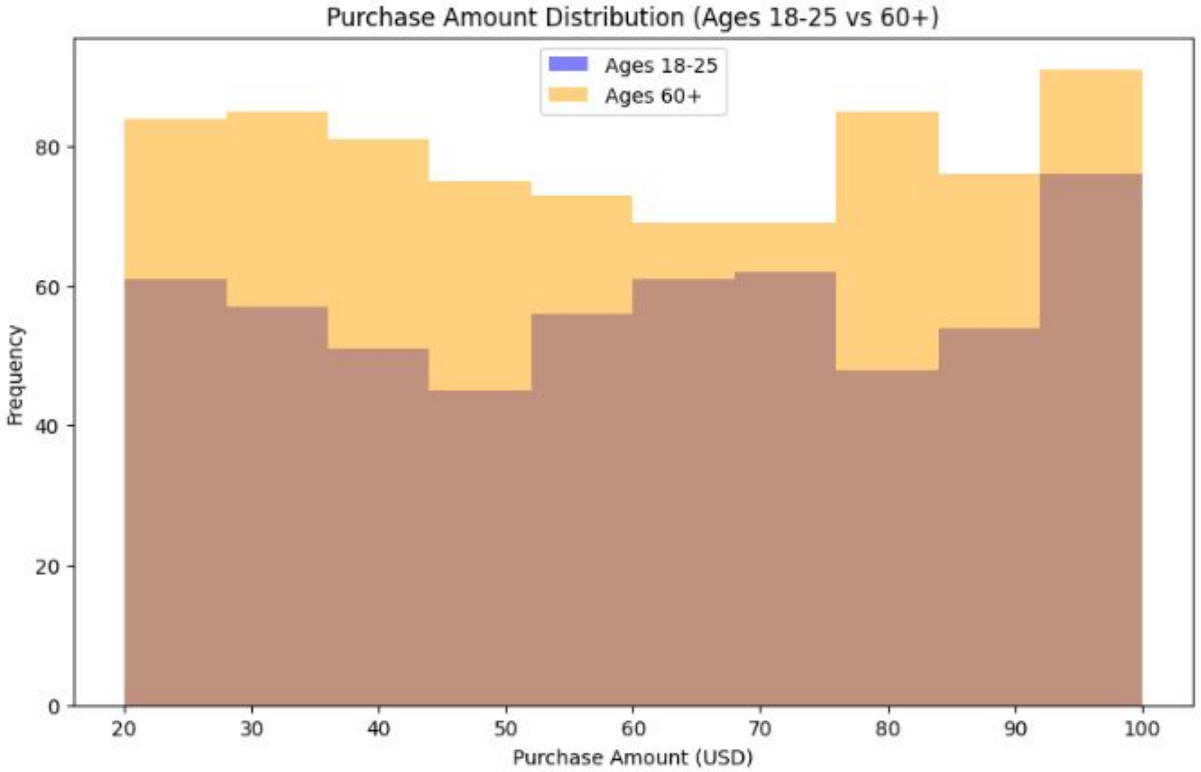


Exploratory Analysis:

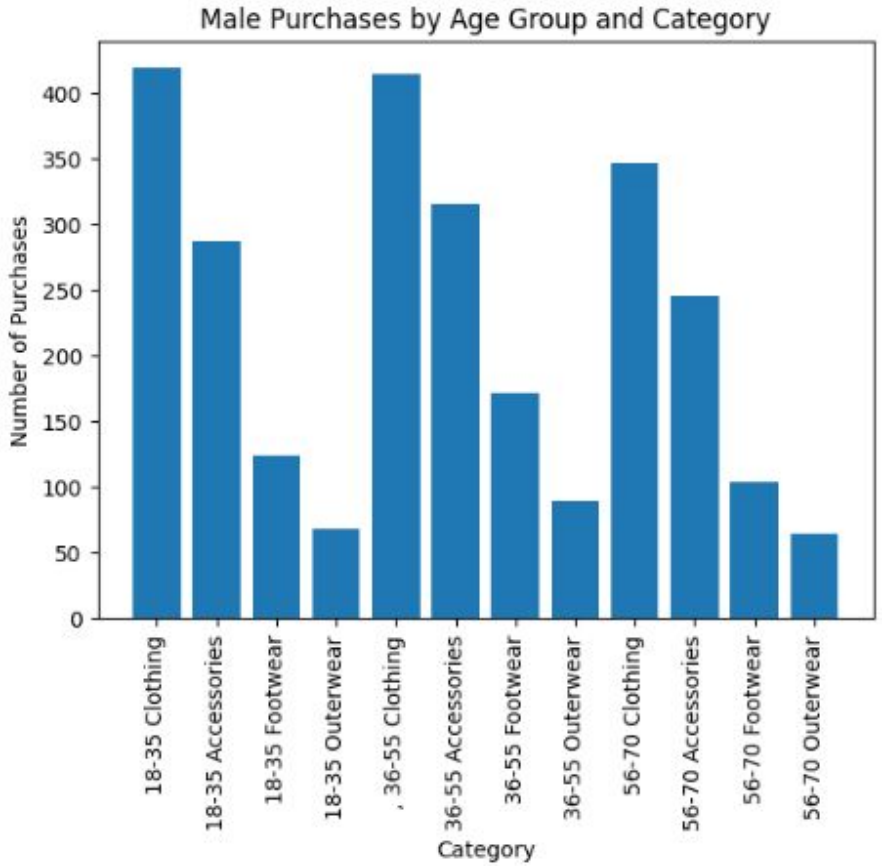


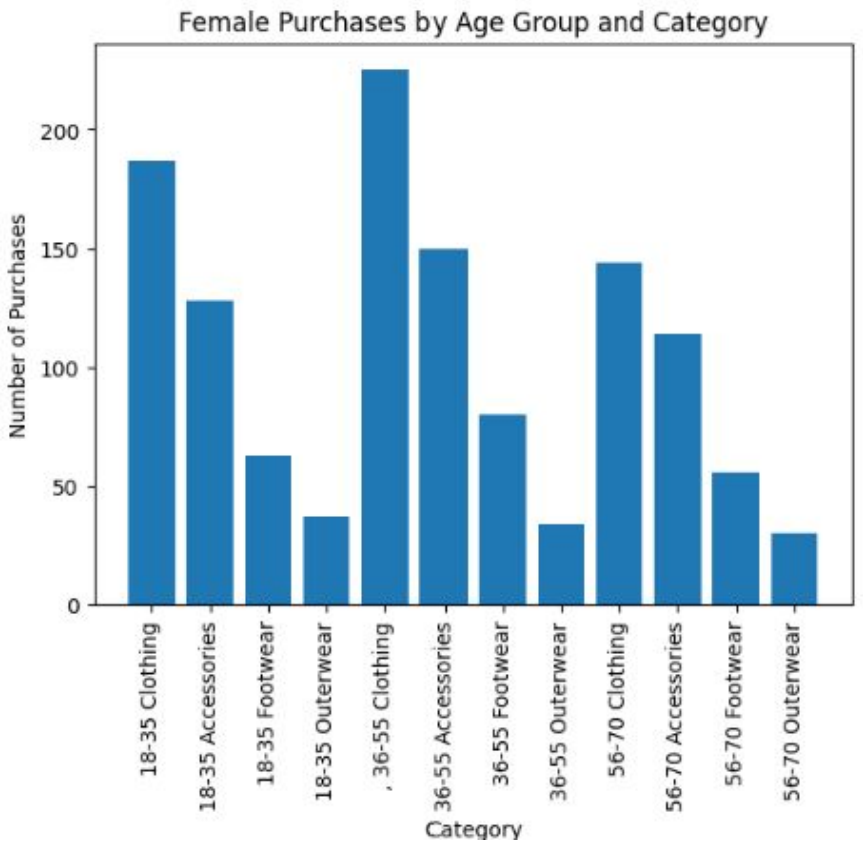


Exploratory Analysis



Exploratory Analysis



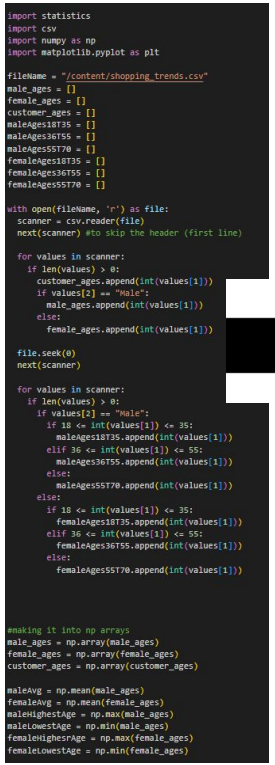


Model Design and Implementation

First implementation: Focusing on the age and gender of the customers.

Imported necessary python libraries such as statistics, csv, numpy, and matplotlib.pyplot that allowed me to read the data set file, import that information to the program, and use my own lists and arrays to analyze and graph the data

I first read the data, calculated statistics of the ages and gender, and input this information I found onto graphs and charts



```
import statistics
import csv
import numpy as np
import matplotlib.pyplot as plt

fileName = "/content/shopping_trends.csv"
male_ages = []
female_ages = []
customer_ages = []
maleAges18T35 = []
maleAges36T55 = []
maleAges55T70 = []
femaleAges18T35 = []
femaleAges36T55 = []
femaleAges55T70 = []

with open(fileName, 'r') as file:
  scanner = csv.reader(file)
  next(scanner) #to skip the header (first line)

  for values in scanner:
    if len(values) > 0:
      customer_ages.append(int(values[1]))
      if values[2] == "Male":
        male_ages.append(int(values[1]))
      else:
        female_ages.append(int(values[1]))

  file.seek(0)
  next(scanner)

  for values in scanner:
    if len(values) > 0:
      if values[2] == "Male":
        if 18 <= int(values[1]) <= 35:
          maleAges18T35.append(int(values[1]))
        elif 36 <= int(values[1]) <= 55:
          maleAges36T55.append(int(values[1]))
        else:
          maleAges55T70.append(int(values[1]))
      else:
        if 18 <= int(values[1]) <= 35:
          femaleAges18T35.append(int(values[1]))
        elif 36 <= int(values[1]) <= 55:
          femaleAges36T55.append(int(values[1]))
        else:
          femaleAges55T70.append(int(values[1]))



#making it into np arrays
male_ages = np.array(male_ages)
female_ages = np.array(female_ages)
customer_ages = np.array(customer_ages)

maleAvg = np.mean(male_ages)
femaleAvg = np.mean(female_ages)
maleHighestAge = np.max(male_ages)
maleLowestAge = np.min(male_ages)
femaleHighesrAge = np.max(female_ages)
femaleLowestAge = np.min(female_ages)
```

```
#printing the results (means)
print("The ages of male customers range from: ", maleHighestAge, " to ", maleLowestAge)
print("The ages of female customers range from: ", femaleHighesrAge, " to ", femaleLowestAge)
print()
print("The average age of male customers is: ", maleAvg)
print("The average age of female customers is: ", femaleAvg)
print()
print("The median age of male customers is: ", np.median(male_ages))
print("The median age of female customers is: ", np.median(female_ages))
print()
print("The mode age of male customers is: ", statistics.mode(male_ages))
print("The mode age of female customers is: ", statistics.mode(female_ages))
print()

#creating the bar chart to demonstrate
plt.bar(["Female", "Male"], [femaleAvg, maleAvg], color=["pink","blue"])
plt.xlabel("Gender")
plt.ylabel("Average Age")
plt.title("Average Age of Customer by Gender")
plt.show()
print()

#creating histogram to demonstrate variety of ages
plt.hist(customer_ages, bins=50, color = 'green', edgecolor='black')
plt.xlabel("Ages")
plt.ylabel("Frequency")
plt.title("Distribution of Customer Ages")
plt.show()
print()

#creating pie charts to demonstrate variety of ages based on gender
plt.pie([len(maleAges18T35), len(maleAges36T55), len(maleAges55T70)], labels=["18-35","36-55", "56-70"], autopc
plt.title("Male Customer Ages")
plt.labels = ["18-35","36-55", "56-70"]
plt.show()
print()
plt.pie([len(femaleAges18T35), len(femaleAges36T55), len(femaleAges55T70)], labels=["18-35","36-55", "56-70"], a
plt.title("Female Customer Ages")
plt.labels = ["18-35","36-55", "56-70"]
plt.show()
print()
```

Model Design and Implementation

Second Implementation: Focusing on a correlation between ages and purchase power, I created two lists that were determined by age. I did this by reading the csv file and extracting information, based on the customers age, their purchase information would be input into one of the two lists. With these lists, I converted them to arrays, and used numpy to find the mean purchase amounts. I also dove into any correlations between age and purchase amounts. I decided to visualize this information through a histogram.

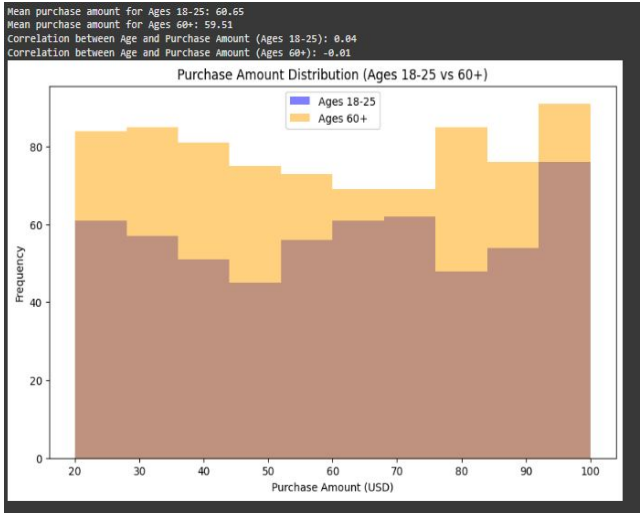
18-25 Correlation Coefficient: 0.04

```python
# Investigating if there is a correlation between ages and purchase power by dividing customer into two
# age groups
# Do older people spend more money than younger people?

import csv
import numpy as np
import matplotlib.pyplot as plt

fileName = "/content/shopping_trends.csv"

# Lists to store age and purchase amounts for the two age groups
age_18_25 = []
purchase_amount_18_25 = []

age_60_plus = []
purchase_amount_60_plus = []

# Reading the data
with open(fileName, "r") as file:
    scanner = csv.reader(file)
    next(scanner)  # Skip the header

    for values in scanner:
        if len(values) > 0:
            age = int(values[1])
            purchase_amount = float(values[5])

            # Add to corresponding age group
            if 18 <= age <= 25:
                age_18_25.append(age)
                purchase_amount_18_25.append(purchase_amount)
            elif age >= 60:
                age_60_plus.append(age)
                purchase_amount_60_plus.append(purchase_amount)
```

```python
# Convert to NumPy arrays for easier analysis
age_18_25 = np.array(age_18_25)
purchase_amount_18_25 = np.array(purchase_amount_18_25)

age_60_plus = np.array(age_60_plus)
purchase_amount_60_plus = np.array(purchase_amount_60_plus)

# Calculate mean purchase amounts
mean_18_25 = np.mean(purchase_amount_18_25)
mean_60_plus = np.mean(purchase_amount_60_plus)

print(f"Mean purchase amount for Ages 18-25: {mean_18_25:.2f}")
print(f"Mean purchase amount for Ages 60+: {mean_60_plus:.2f}")

# Calculate the correlation for each group
corr_18_25 = np.corrcoef(age_18_25, purchase_amount_18_25)[0, 1]
corr_60_plus = np.corrcoef(age_60_plus, purchase_amount_60_plus)[0, 1]

# Output the correlation coefficients
print(f"Correlation between Age and Purchase Amount (Ages 18-25): {corr_18_25:.2f}")
print(f"Correlation between Age and Purchase Amount (Ages 60+): {corr_60_plus:.2f}")

# Create the histogram plot
plt.figure(figsize=(10, 6))

# Histogram for ages 18-25
plt.hist(purchase_amount_18_25, bins=10, color='blue', alpha=0.5, label='Ages 18-25')

# Histogram for ages 60+
plt.hist(purchase_amount_60_plus, bins=10, color='orange', alpha=0.5, label='Ages 60+')

# Titles and labels
plt.title('Purchase Amount Distribution (Ages 18-25 vs 60+)')
plt.xlabel('Purchase Amount (USD)')
plt.ylabel('Frequency')

# Show legend
plt.legend()

# Display the plot
plt.show()
```

60+ Correlation Coefficient: -0.04

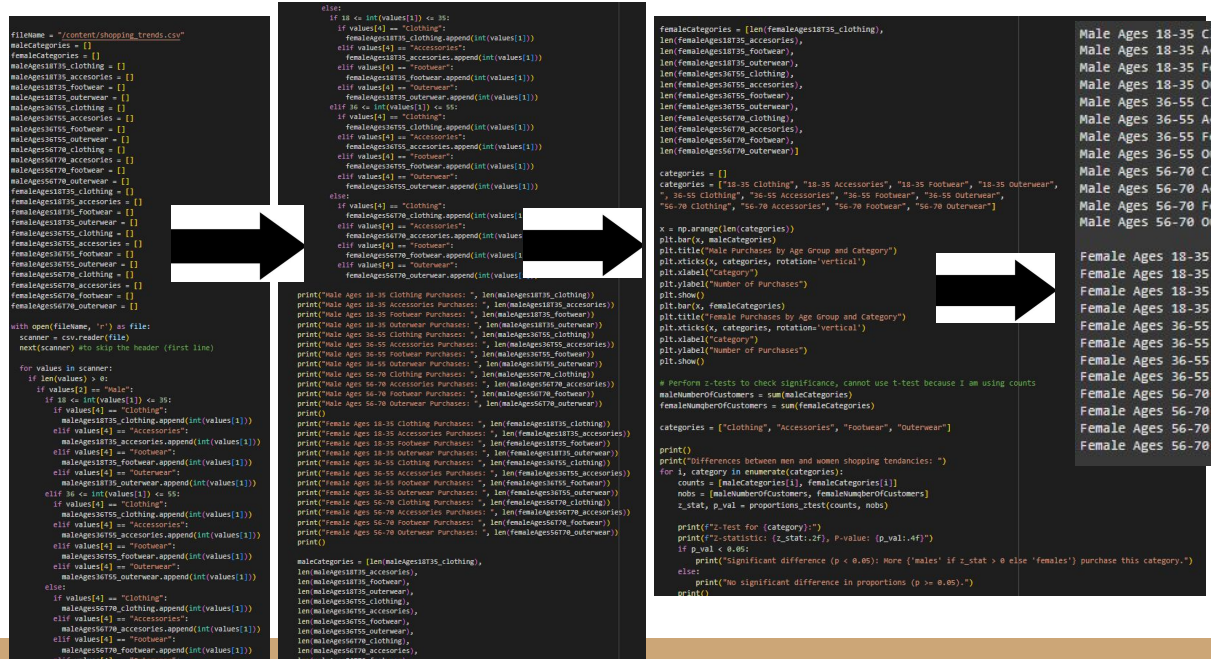
```
Mean purchase amount for Ages 18-25: 60.65
Mean purchase amount for Ages 60+: 59.51
Correlation between Age and Purchase Amount (Ages 18-25): 0.04
Correlation between Age and Purchase Amount (Ages 60+): -0.01
```



Model Design and Implementation

Third Implementation:

Wanted to determine if men or women had different clothing purchase habits. I focused on the categories of items purchased for this (Clothing, Accessories, Footwear, and Outerwear) and made numerous lists that depended on gender, age, and what category of item was purchased. I compared these lists using Z tests to determine whether or not there were different spending habits.



```
Male Ages 18-35 Clothing Purchases:  419
Male Ages 18-35 Accessories Purchases:  287
Male Ages 18-35 Footwear Purchases:  124
Male Ages 18-35 Outerwear Purchases:  68
Male Ages 36-55 Clothing Purchases:  415
Male Ages 36-55 Accessories Purchases:  316
Male Ages 36-55 Footwear Purchases:  172
Male Ages 36-55 Outerwear Purchases:  90
Male Ages 56-70 Clothing Purchases:  347
Male Ages 56-70 Accessories Purchases:  245
Male Ages 56-70 Footwear Purchases:  104
Male Ages 56-70 Outerwear Purchases:  65

Female Ages 18-35 Clothing Purchases:  187
Female Ages 18-35 Accessories Purchases:  128
Female Ages 18-35 Footwear Purchases:  63
Female Ages 18-35 Outerwear Purchases:  37
Female Ages 36-55 Clothing Purchases:  225
Female Ages 36-55 Accessories Purchases:  150
Female Ages 36-55 Footwear Purchases:  80
Female Ages 36-55 Outerwear Purchases:  34
Female Ages 56-70 Clothing Purchases:  144
Female Ages 56-70 Accessories Purchases:  114
Female Ages 56-70 Footwear Purchases:  56
Female Ages 56-70 Outerwear Purchases:  30
```

```
Differences between men and women shopping tendancies:
Z-Test for Clothing:
Z-statistic: 0.66, P-value: 0.5120
No significant difference in proportions (p >= 0.05).

Z-Test for Accessories:
Z-statistic: 0.53, P-value: 0.5931
No significant difference in proportions (p >= 0.05).

Z-Test for Footwear:
Z-statistic: -0.51, P-value: 0.6117
No significant difference in proportions (p >= 0.05).

Z-Test for Outerwear:
Z-statistic: -0.72, P-value: 0.4709
No significant difference in proportions (p >= 0.05).
```

Model Design and Implementation

Last Implementation: Using the T-test, I wanted to find out if promo codes or discounts prompted higher purchase amounts. Therefore I analyzed both purchases with discounts / promo codes and purchases without using the T-test.

```python
#Hypothesis Discounts and promo code applied increases the Purcahse Amount by incentivizing purchases

from scipy.stats import ttest_ind
import statistics
import csv
import numpy as np
import matplotlib.pyplot as plt

fileName = "/content/shopping_trends.csv"

purchase_with_discount = []
purchase_without_discount = []

with open(fileName, "r") as file:
  scanner = csv.reader(file)
  next(scanner) #to skip the header 5

  for values in scanner:
    if len(values) > 0:
      if(values[14] == "Yes" and values[15] == "Yes") :
        purchase_with_discount.append(float(values[5])) #storing the purchase amount if they used a discount code
      else:
        purchase_without_discount.append(float(values[5])) #storing the purchase amount if they DID NOT

purchase_with_discount = np.array(purchase_with_discount)
purchase_without_discount = np.array(purchase_without_discount)

mean_with_discount = np.mean(purchase_with_discount)
mean_without_discount = np.mean(purchase_without_discount)

print(f"Mean price with discount {mean_with_discount: .2f}")
print(f"Mean price without discount {mean_without_discount: .2f}")

# Perform t-tests to check significance
t_stat_discount, p_val_discount = ttest_ind(purchase_with_discount, purchase_without_discount)

print(f"\nT-test for Discounts: t-stat={t_stat_discount:.2f}, p-value={p_val_discount:.4f}")

# Interpretation
if p_val_discount < 0.05:
    print("Discounts significantly impact Purchase Amount (p < 0.05).")
else:
    print("Discounts do not significantly impact Purchase Amount (p >= 0.05).")
```

```
Mean price with discount  59.28
Mean price without discount  60.13

T-test for Discounts: t-stat=-1.11, p-value=0.2665
Discounts do not significantly impact Purchase Amount (p >= 0.05).
```

Conclusion:

Through the use of the project, the main goal was to see if I can contribute and help small clothing businesses in local communities. Through the use of statistical analysis I was able to achieve this goal, by showing I have the ability to share educated and informed knowledge about purchasing trends within clothing.

For Example, I came to multiple conclusions regarding the data analyzation through statistics.

I found that the ages of customers varied immensely, and that all age groups were purchasing items at a number around the mean. It also did not matter whether they were male or female. With this information, I can confidently encourage companies to continue to advertise to all age groups.

I also found that there was no correlation between age and the purchase amount. Therefore, I can recommend that there is no need for the company to focus on a specific age group when stocking items and advertising.

I discovered that for all age groups, and for both men and women, clothing was always the most purchased item category. Therefore, I can recommend that stocking up on more clothing than other items by allocating their money differently can help them in the long run.

I also found no differences in mens and womens shopping tendencies based on item categories, therefore there is no need to suggest focusing on different categories when advertising to the different genders.

Lastly, I found that when using promo codes, customers generally do not increase purchase amounts by incentivizing purchases, so it may not be in the companies best interest to promote numerous discounts or promos if their goal is to have customers spend more.

Conclusion:

With this information, small businesses can more confidently make decisions about what clothing items they put for sale, they bulk order, and how they advertise or put items on sale. All of these decisions truly make up how well a businesses does, and making more informed decisions would definitely help local businesses find more success.

By sharing my findings, my goal would be met, proving I can impact society in a positive way by helping local / small businesses if they gave me the opportunity to work with their sales data.