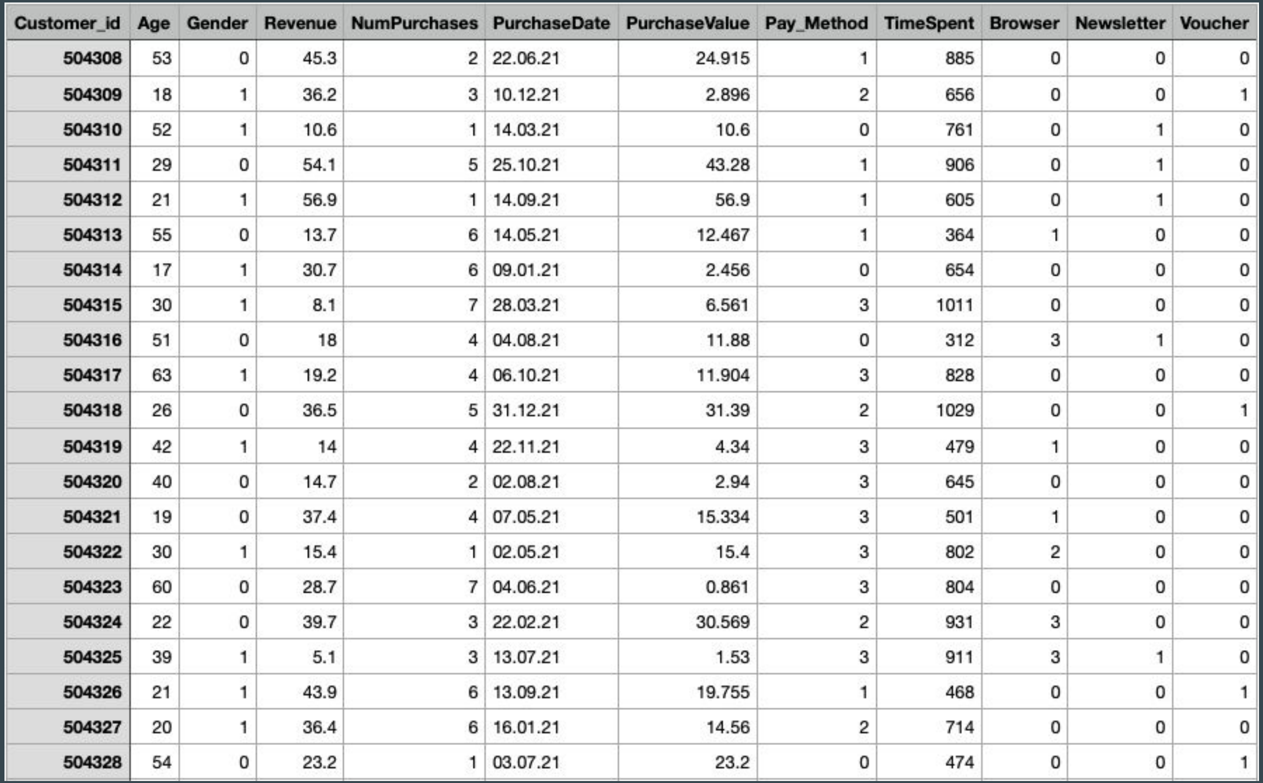
Online Sales Data 2021

•••

Conner Jamison

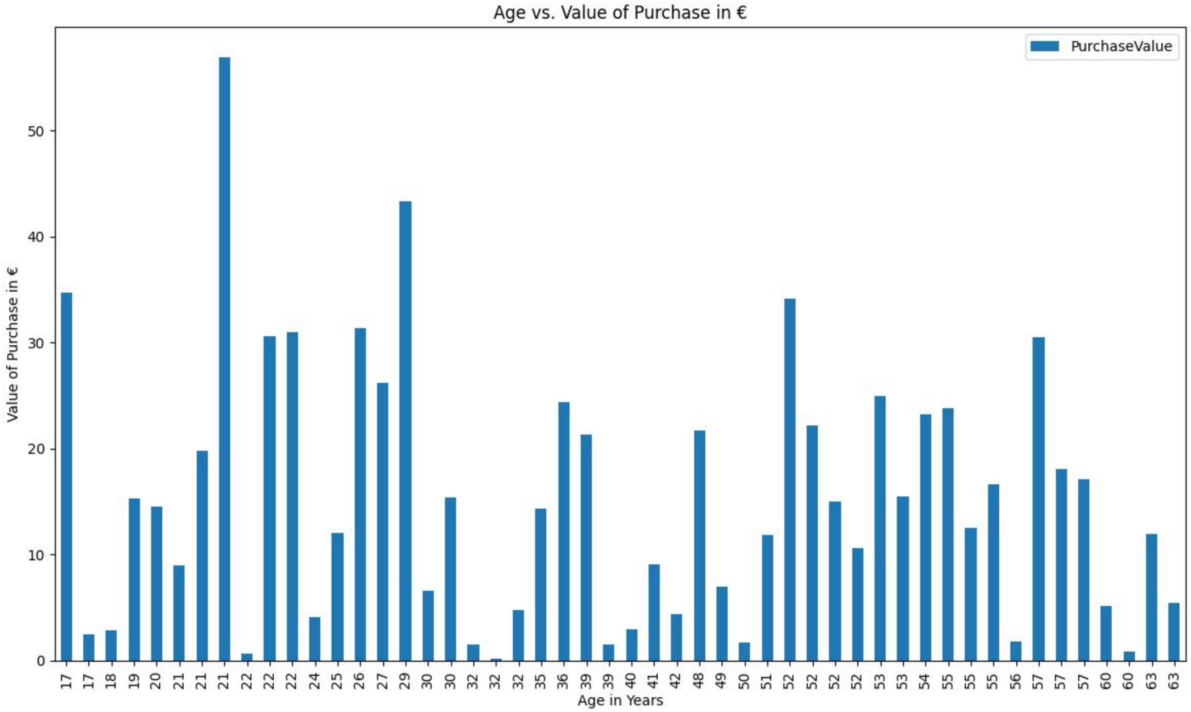
Intro/Glimpse of the Data



Customer_id	Age	Gender	Revenue	NumPurchases	PurchaseDate	PurchaseValue	Pay_Method	TimeSpent	Browser	Newsletter	Voucher
504308	53	0	45.3	2	22.06.21	24.915	1	885	0	0	0
504309	18	1	36.2	3	10.12.21	2.896	2	656	0	0	1
504310	52	1	10.6	1	14.03.21	10.6	0	761	0	1	0
504311	29	0	54.1	5	25.10.21	43.28	1	906	0	1	0
504312	21	1	56.9	1	14.09.21	56.9	1	605	0	1	0
504313	55	0	13.7	6	14.05.21	12.467	1	364	1	0	0
504314	17	1	30.7	6	09.01.21	2.456	0	654	0	0	0
504315	30	1	8.1	7	28.03.21	6.561	3	1011	0	0	0
504316	51	0	18	4	04.08.21	11.88	0	312	3	1	0
504317	63	1	19.2	4	06.10.21	11.904	3	828	0	0	0
504318	26	0	36.5	5	31.12.21	31.39	2	1029	0	0	1
504319	42	1	14	4	22.11.21	4.34	3	479	1	0	0
504320	40	0	14.7	2	02.08.21	2.94	3	645	0	0	0
504321	19	0	37.4	4	07.05.21	15.334	3	501	1	0	0
504322	30	1	15.4	1	02.05.21	15.4	3	802	2	0	0
504323	60	0	28.7	7	04.06.21	0.861	3	804	0	0	0
504324	22	0	39.7	3	22.02.21	30.569	2	931	3	0	0
504325	39	1	5.1	3	13.07.21	1.53	3	911	3	1	0
504326	21	1	43.9	6	13.09.21	19.755	1	468	0	0	1
504327	20	1	36.4	6	16.01.21	14.56	2	714	0	0	0
504328	54	0	23.2	1	03.07.21	23.2	0	474	0	0	1

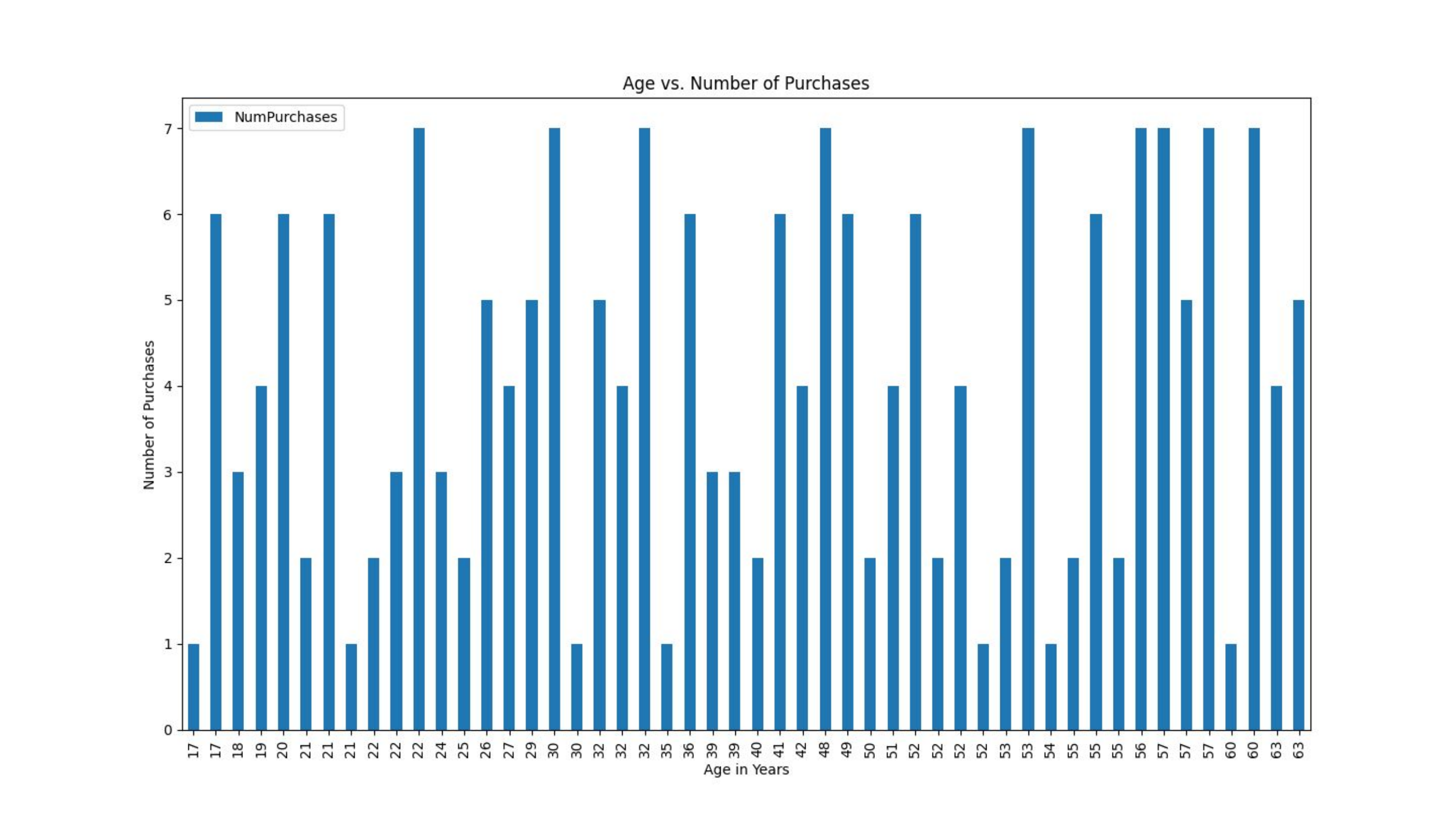
- All from 2021
- 12 columns or categories
- 65000 rows
- Avg items per transaction = 4
  - 260000 items sold
- Used only 50 rows for graphs

Graph 1

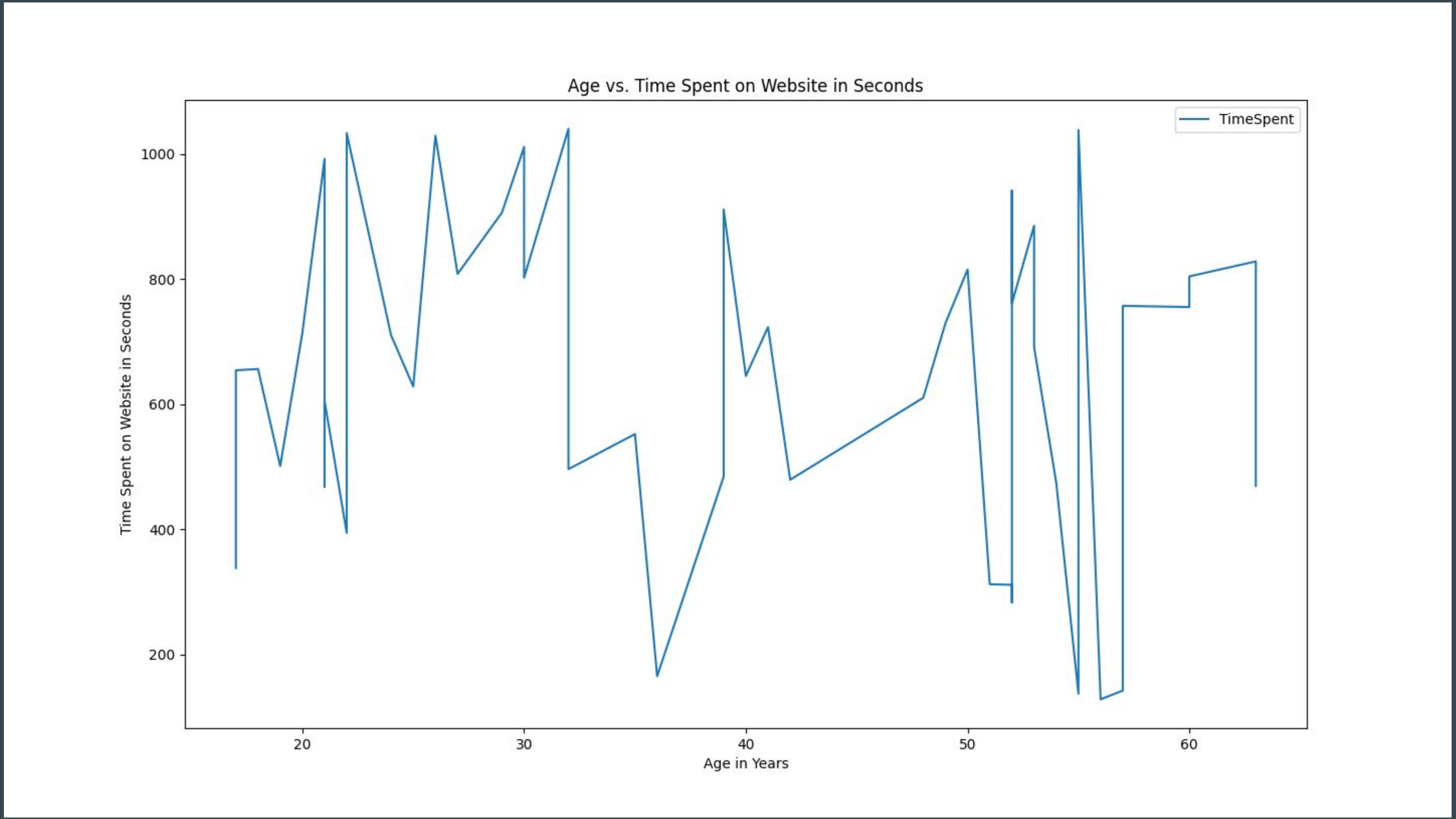


Age vs. Value of Purchase in €
PurchaseValue
Value of Purchase in €
0
10
20
30
40
50
Age in Years
17
17
18
19
20
21
21
21
22
22
22
24
25
26
27
29
30
30
32
32
32
35
36
39
39
40
41
42
48
49
50
51
52
52
52
52
53
53
54
55
55
55
56
57
57
57
60
60
63
63

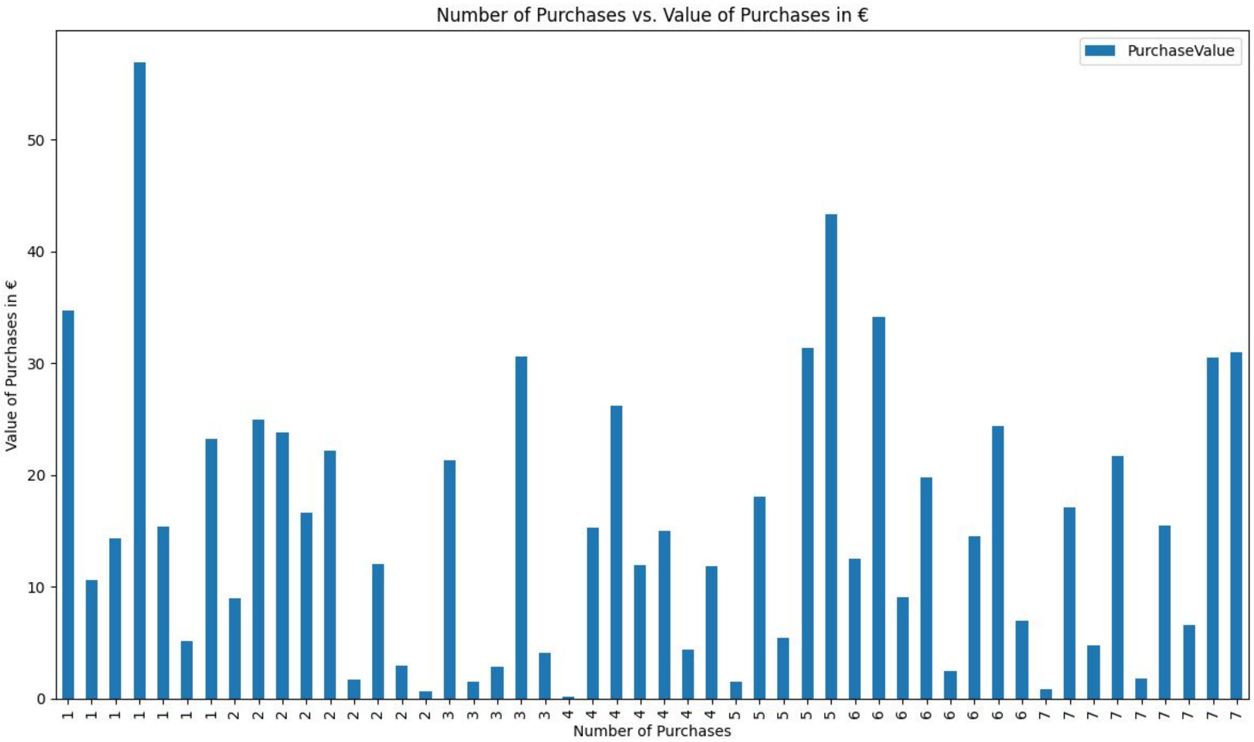
Graph 2


Age vs. Number of Purchases
NumPurchases
Number of Purchases
0
1
2
3
4
5
6
7
17 17 18 19 20 21 21 21 22 22 22 24 25 26 27 29 30 30 32 32 32 35 36 39 39 40 41 42 48 49 50 51 52 52 52 52 53 53 54 55 55 55 56 57 57 57 60 60 63 63
Age in Years

Graph 3

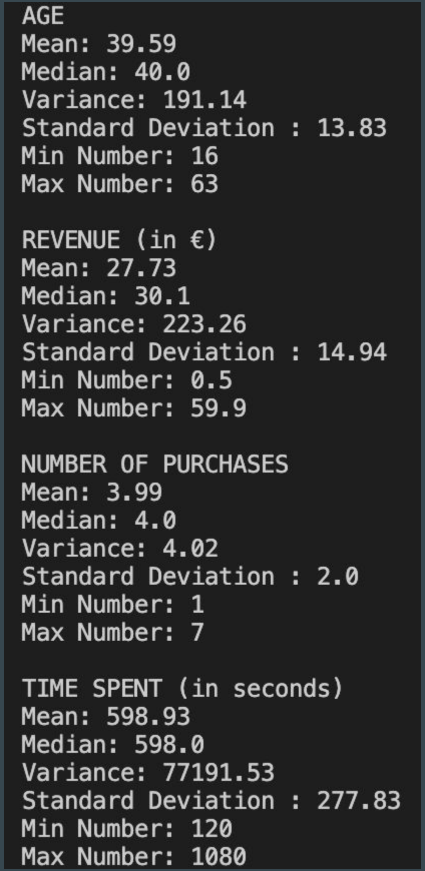

Age vs. Time Spent on Website in Seconds
TimeSpent
Time Spent on Website in Seconds
1000
800
600
400
200
20
30
40
50
60
Age in Years

Graph 4



Number of Purchases vs. Value of Purchases in €
PurchaseValue
Value of Purchases in €
0
10
20
30
40
50
Number of Purchases
1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 3 3 3 3 3 4 4 4 4 4 4 4 5 5 5 5 5 6 6 6 6 6 6 6 6 7 7 7 7 7 7 7 7 7

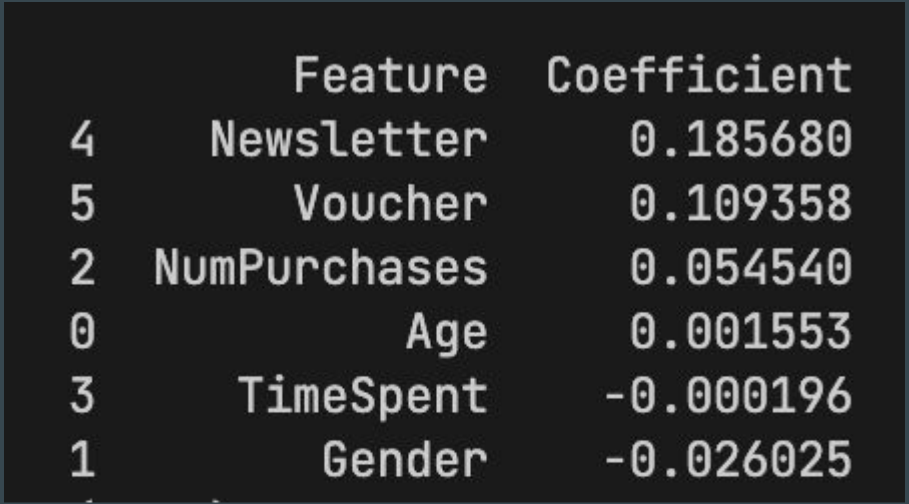
Statistics



```
AGE
Mean: 39.59
Median: 40.0
Variance: 191.14
Standard Deviation : 13.83
Min Number: 16
Max Number: 63

REVENUE (in €)
Mean: 27.73
Median: 30.1
Variance: 223.26
Standard Deviation : 14.94
Min Number: 0.5
Max Number: 59.9

NUMBER OF PURCHASES
Mean: 3.99
Median: 4.0
Variance: 4.02
Standard Deviation : 2.0
Min Number: 1
Max Number: 7

TIME SPENT (in seconds)
Mean: 598.93
Median: 598.0
Variance: 77191.53
Standard Deviation : 277.83
Min Number: 120
Max Number: 1080
```

- Using age, revenue, number of purchases, and time spent columns from data set
- Used all 65000 rows
- Average revenue was € 27.73/transaction
  - Over 65000 transactions - over 1.8 million in revenue

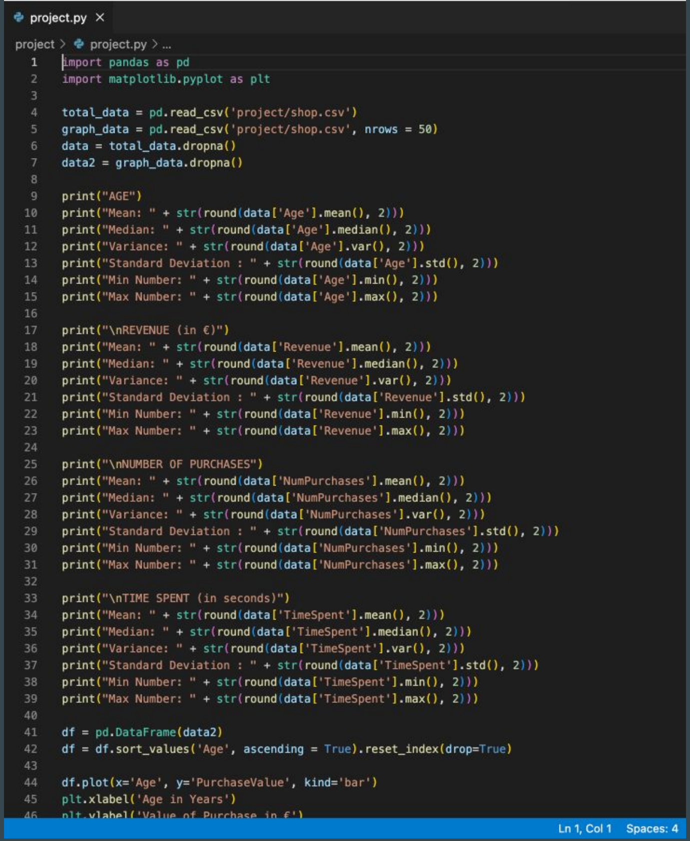
Regression Analysis



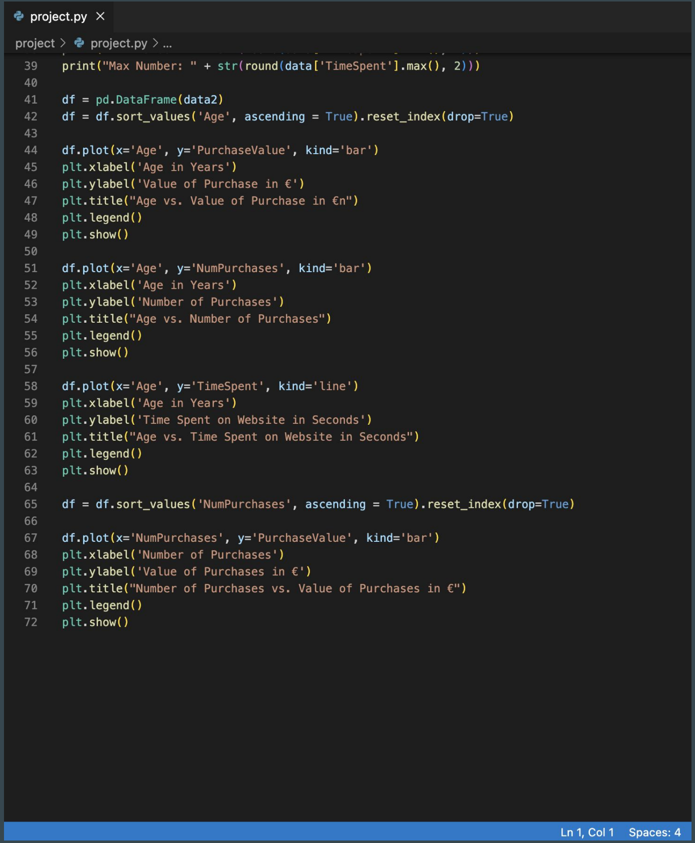
```
       Feature  Coefficient
4   Newsletter     0.185680
5      Voucher     0.109358
2 NumPurchases     0.054540
0          Age     0.001553
3    TimeSpent    -0.000196
1       Gender    -0.026025
```

- Utilized sklearn for regression analysis
  - X data = 'Age', 'Gender', 'NumPurchases', 'TimeSpent', 'Newsletter', 'Voucher'
  - Y data = ‘Revenue’
- Found small correlation between newsletter subscription + voucher use and sales
  - Newsletter subscription results in 18.5% increase in sales
  - Voucher use results in 11% increase in sales

Code



```python
import pandas as pd
import matplotlib.pyplot as plt

total_data = pd.read_csv('project/shop.csv')
graph_data = pd.read_csv('project/shop.csv', nrows = 50)
data = total_data.dropna()
data2 = graph_data.dropna()

print("AGE")
print("Mean: " + str(round(data['Age'].mean(), 2)))
print("Median: " + str(round(data['Age'].median(), 2)))
print("Variance: " + str(round(data['Age'].var(), 2)))
print("Standard Deviation : " + str(round(data['Age'].std(), 2)))
print("Min Number: " + str(round(data['Age'].min(), 2)))
print("Max Number: " + str(round(data['Age'].max(), 2)))

print("\nREVENUE (in €)")
print("Mean: " + str(round(data['Revenue'].mean(), 2)))
print("Median: " + str(round(data['Revenue'].median(), 2)))
print("Variance: " + str(round(data['Revenue'].var(), 2)))
print("Standard Deviation : " + str(round(data['Revenue'].std(), 2)))
print("Min Number: " + str(round(data['Revenue'].min(), 2)))
print("Max Number: " + str(round(data['Revenue'].max(), 2)))

print("\nNUMBER OF PURCHASES")
print("Mean: " + str(round(data['NumPurchases'].mean(), 2)))
print("Median: " + str(round(data['NumPurchases'].median(), 2)))
print("Variance: " + str(round(data['NumPurchases'].var(), 2)))
print("Standard Deviation : " + str(round(data['NumPurchases'].std(), 2)))
print("Min Number: " + str(round(data['NumPurchases'].min(), 2)))
print("Max Number: " + str(round(data['NumPurchases'].max(), 2)))

print("\nTIME SPENT (in seconds)")
print("Mean: " + str(round(data['TimeSpent'].mean(), 2)))
print("Median: " + str(round(data['TimeSpent'].median(), 2)))
print("Variance: " + str(round(data['TimeSpent'].var(), 2)))
print("Standard Deviation : " + str(round(data['TimeSpent'].std(), 2)))
print("Min Number: " + str(round(data['TimeSpent'].min(), 2)))
print("Max Number: " + str(round(data['TimeSpent'].max(), 2)))

df = pd.DataFrame(data2)
df = df.sort_values('Age', ascending = True).reset_index(drop=True)

df.plot(x='Age', y='PurchaseValue', kind='bar')
plt.xlabel('Age in Years')
plt.ylabel('Value of Purchase in €')
```



```python
print("Max Number: " + str(round(data['TimeSpent'].max(), 2)))

df = pd.DataFrame(data2)
df = df.sort_values('Age', ascending = True).reset_index(drop=True)

df.plot(x='Age', y='PurchaseValue', kind='bar')
plt.xlabel('Age in Years')
plt.ylabel('Value of Purchase in €')
```

```python
plt.title("Age vs. Value of Purchase in €n")
plt.legend()
plt.show()

df.plot(x='Age', y='NumPurchases', kind='bar')
plt.xlabel('Age in Years')
plt.ylabel('Number of Purchases')
plt.title("Age vs. Number of Purchases")
plt.legend()
plt.show()

df.plot(x='Age', y='TimeSpent', kind='line')
plt.xlabel('Age in Years')
plt.ylabel('Time Spent on Website in Seconds')
plt.title("Age vs. Time Spent on Website in Seconds")
plt.legend()
plt.show()

df = df.sort_values('NumPurchases', ascending = True).reset_index(drop=True)

df.plot(x='NumPurchases', y='PurchaseValue', kind='bar')
plt.xlabel('Number of Purchases')
plt.ylabel('Value of Purchases in €')
plt.title("Number of Purchases vs. Value of Purchases in €")
plt.legend()
plt.show()
```

Source

- https://www.kaggle.com/datasets/onlineretailshop/online-shop-customer-sales-data