Seneca

Green Horizons: Predictive Analytics for Optimal Cannabis Store Locations

BAN210ZCC

PROJECT 2

NAME

- Pareen Suresh Harsosa (112515226)

Company Overview & Key Stakeholders

Richard Boire Inc.:

Industry: Cannabis Retail & Distribution

Founded: 1999

Headquarters: Vancouver, BC

Vision: To be the leading cannabis retailer in Ontario

Key Stakeholders:

Richard Boire: Founder & President, driving force behind the company's strategic decisions

Elena Smith: Chief Financial Officer, oversees company's financial strategies

Michael Chen: Head of Retail Operations, manages the company's retail presence

Sophia Rodriguez: VP of Marketing, responsible for brand presence and customer engagement

Lucas Green: Data & Analytics Lead, harnesses data for strategic insights

Business Challenge:

With a vision to further expand in Ontario, Richard Boire is keen on strategically positioning stores in the province. The challenge lies in pinpointing the **top 5 locations** that not only have high current demand but also showcase potential for future growth. The decision must be data-driven, ensuring optimal investment and high returns.

Introduction to the Problem

Background on the Cannabis Industry:

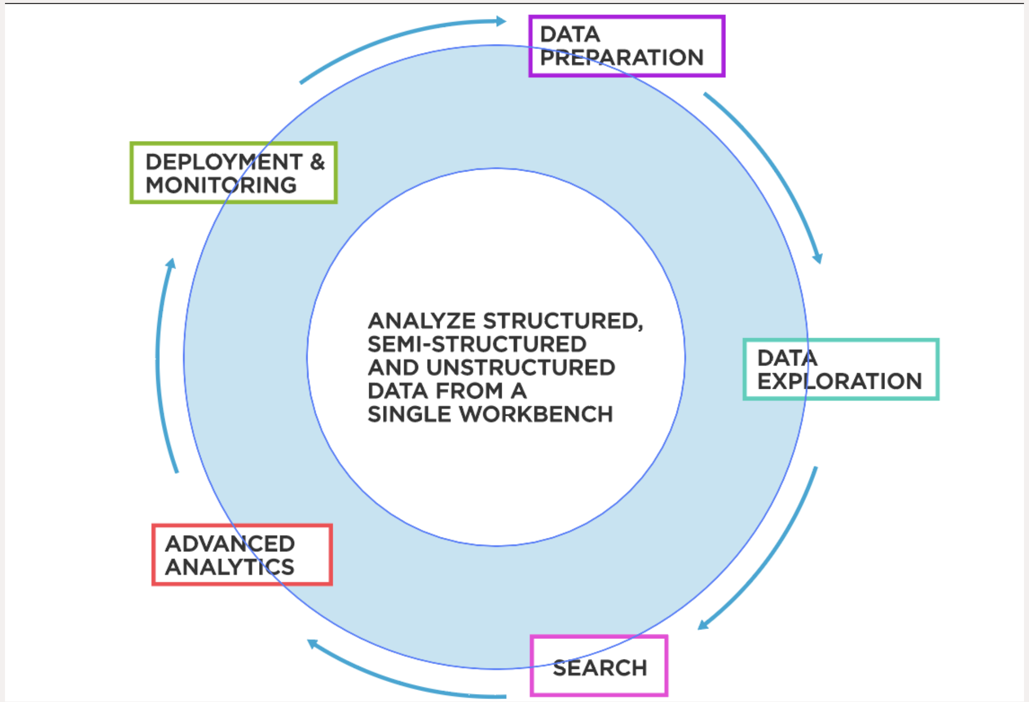
- Rapid industry growth
- Increasing legalization efforts
- Medicinal and recreational markets expanding
- Challenges: regulatory hurdles, market saturation, consistent quality

Business Challenge:

- Critical decision: 5 Best Store locations
- Influence on sales, brand visibility, customer acquisition
- Key goals:
  - Data-driven decision-making
  - Pinpointing optimal store locations for future success

Benefits of Creating an Analytical File in Predictive Analytics

- Data Exploration: Analytical files enable in-depth data exploration and understanding.
- Data Cleaning and Standardization: An analytical file ensures data quality and consistency.
- Efficiency and Scalability: Handling large volumes of data becomes easier and more efficient.
- Reproducibility and Documentation: Analytical files provide a documented record for transparency and collaboration.
- Model Performance and Interpretability: Well-constructed analytical files enhance model performance and deliver interpretable insights.


DATA PREPARATION
DATA EXPLORATION
SEARCH
ADVANCED ANALYTICS
DEPLOYMENT & MONITORING
ANALYZE STRUCTURED, SEMI-STRUCTURED AND UNSTRUCTURED DATA FROM A SINGLE WORKBENCH

Strategic Blueprint: From Data to Decision

Understanding the Challenge: Define the business problem. Identifying top areas for cannabis store openings. Recognize the importance of a data-driven approach.

Data Assimilation: Import necessary datasets. Merge and consolidate data for a holistic view.

Exploratory Data Analysis (EDA): Dive deep into individual variables. Visualize distributions and relationships.

Feature Engineering & Selection: Correlation analysis to gauge variable importance. Employ stepwise forward, backward elimination, and sequential feature selection.

Modeling: Selection of predictive models suitable for the dataset. Training and optimizing models on the data.

Validation & Evaluation: Use advanced techniques like decile charts and Lorenz curves. Assess model accuracy and reliability.

Insight Extraction & Recommendations: Derive actionable insights from model predictions. Recommend the top 5 best store locations in Ontario.

Business Implementation: Strategize store openings based on recommendations. Monitor performance and recalibrate strategy as needed.

Stakeholder Value: The Business Imperative

Strategic Direction:

- Aligns company's mission with actionable insights.
- Ensures all stakeholders work cohesively towards a common goal.

Informed Decision-Making:

- Data-driven insights offer clarity.
- Minimizes risks, maximizes ROI.

Optimal Resource Allocation:

- Ensures investments (time, money, manpower) are directed to the most promising areas.
- Improves efficiency and profitability.

Brand Reputation:

- Making informed, strategic decisions enhances brand image.
- Positions Richard Boire Inc. as an industry leader in cannabis retail.

Future Growth:

- Insights today lay the foundation for tomorrow's expansion.
- Helps stakeholders anticipate and plan for future challenges.

Stakeholder Confidence:

- Transparent, data-driven decisions instill confidence in investors, employees, and partners.
- Bolsters company stability and long-term vision.

Dataset Overview

Development Dataset (ban210dev1.csv):

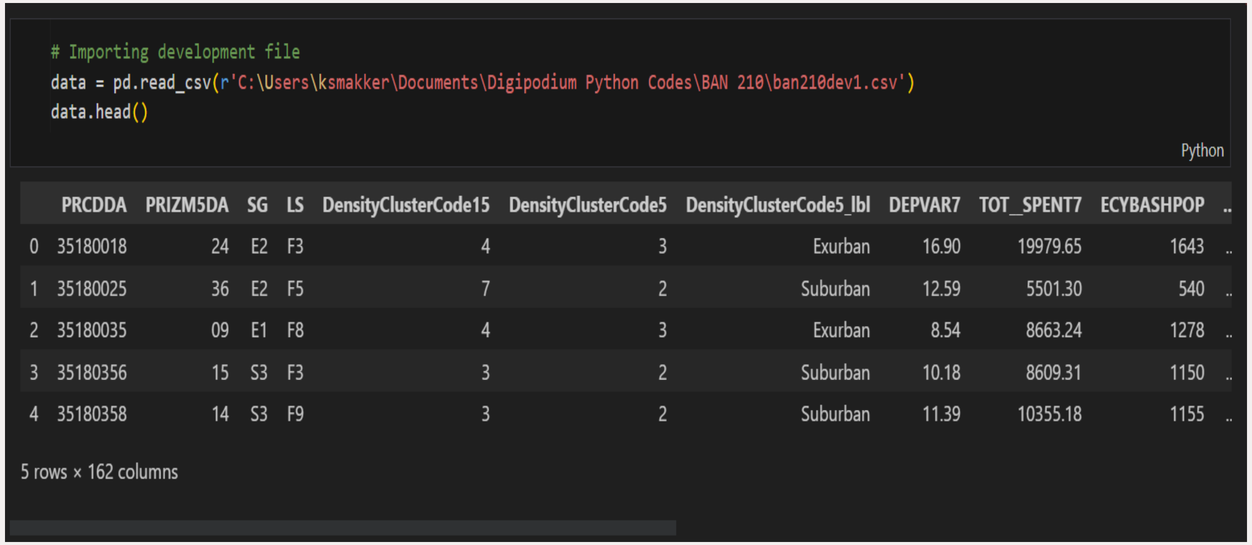
- **Number of Records**: 3,823
- **Number of Variables**: 162

Validation Dataset (ban210val1.csv):

- **Number of Records**: 3,702
- **Number of Variables**: 162

Dataset Description:

The dataset source offers a multifaceted view of consumer demographics and preferences in relation to the cannabis market. With variables ranging from geographic identifiers like the Dissemination Area (SGC Code) to market segmentation indicators such as PRIZM5 codes, it encapsulates various dimensions of the consumer landscape. Detailed descriptors for each variable, such as social group and life stage classifications, provide depth, ensuring a comprehensive understanding of potential cannabis consumers. This rich dataset lays the foundation for data-driven decision-making, enabling Richard Boire Inc. to strategically position its stores for maximum sales potential.

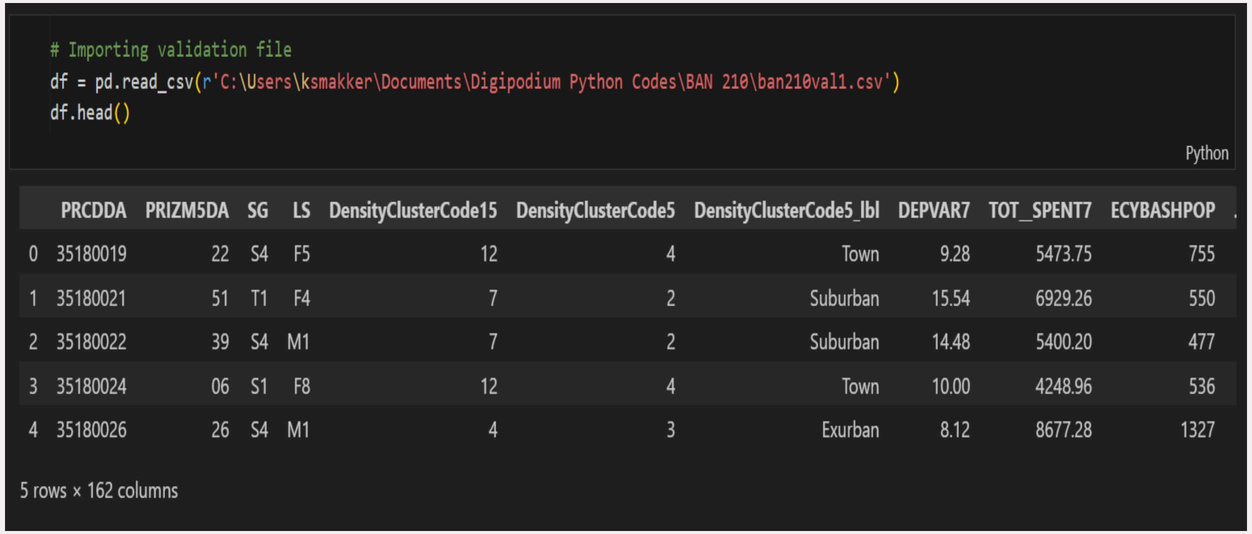


```python
# Importing development file
data = pd.read_csv(r'C:\Users\ksmakker\Documents\Digipodium Python Codes\BAN 210\ban210dev1.csv')
data.head()
```

Python

	PRCDDA	PRIZM5DA	SG	LS	DensityClusterCode15	DensityClusterCode5	DensityClusterCode5_lbl	DEPVAR7	TOT_SPENT7	ECYBASHPOP	..
0	35180018	24	E2	F3	4	3	Exurban	16.90	19979.65	1643	..
1	35180025	36	E2	F5	7	2	Suburban	12.59	5501.30	540	..
2	35180035	09	E1	F8	4	3	Exurban	8.54	8663.24	1278	..
3	35180356	15	S3	F3	3	2	Suburban	10.18	8609.31	1150	..
4	35180358	14	S3	F9	3	2	Suburban	11.39	10355.18	1155	..

5 rows × 162 columns



```python
# Importing validation file
df = pd.read_csv(r'C:\Users\ksmakker\Documents\Digipodium Python Codes\BAN 210\ban210val1.csv')
df.head()
```

Python

	PRCDDA	PRIZM5DA	SG	LS	DensityClusterCode15	DensityClusterCode5	DensityClusterCode5_lbl	DEPVAR7	TOT_SPENT7	ECYBASHPOP
0	35180019	22	S4	F5	12	4	Town	9.28	5473.75	755
1	35180021	51	T1	F4	7	2	Suburban	15.54	6929.26	550
2	35180022	39	S4	M1	7	2	Suburban	14.48	5400.20	477
3	35180024	06	S1	F8	12	4	Town	10.00	4248.96	536
4	35180026	26	S4	M1	4	3	Exurban	8.12	8677.28	1327

5 rows × 162 columns

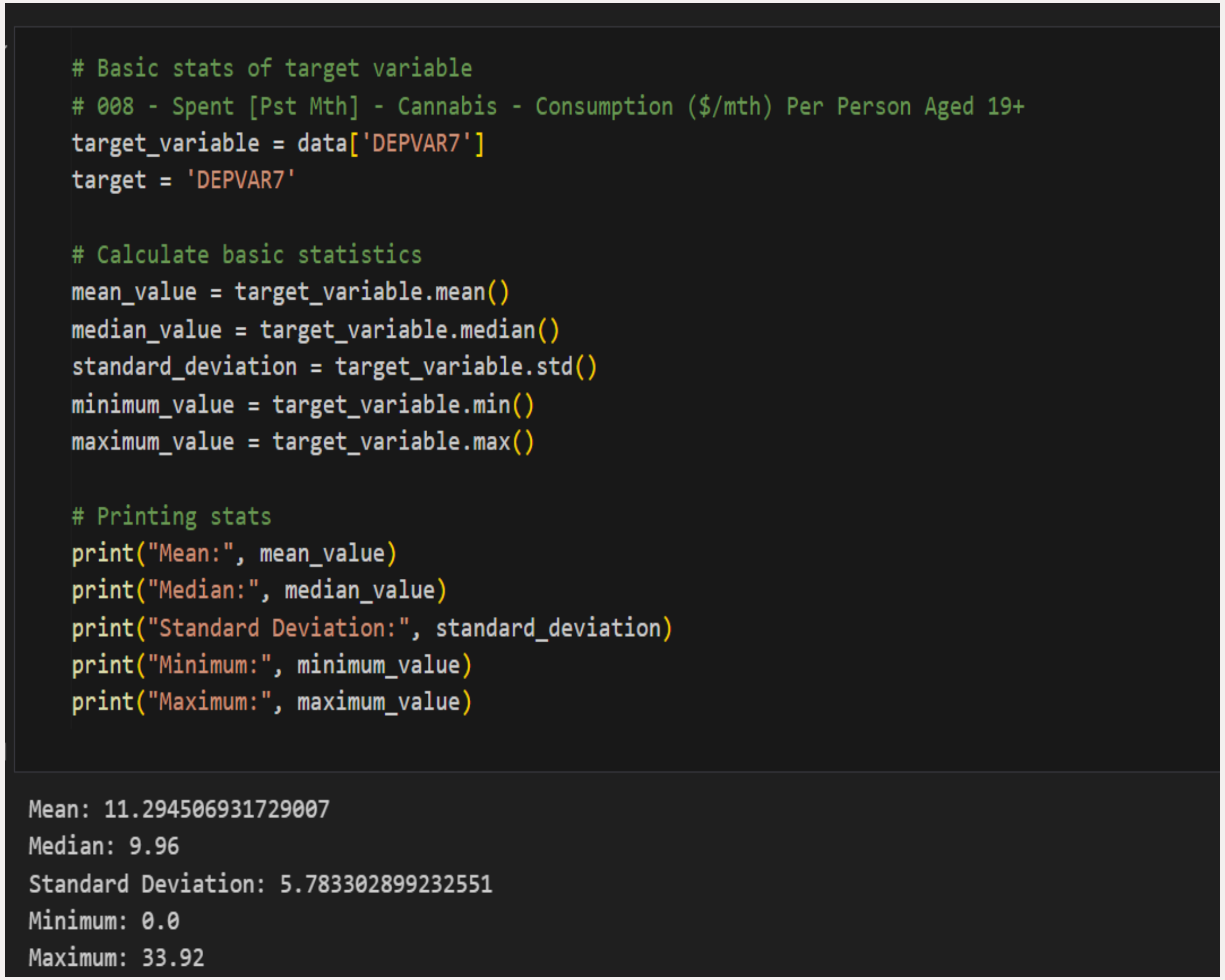
Detailed Statistics of Target Variable- DEPVAR7

Key Statistics:

- **Count:** 3,823 records
- **Mean:** 11.29
- **Median (50th Percentile):** 9.96
- **Mode:** 4.24
- **Standard Deviation:** 5.78
- **Variance:** 33.45
- **Minimum Value:** 0.00
- **1st Quartile (25th Percentile):** 7.44
- **3rd Quartile (75th Percentile):** 14.50
- **Maximum Value:** 33.92

The target variable, DEPVAR7, exhibits a diverse range of values, spanning from 0.00 to 33.92, with an average measure around 11.29. The data's dispersion, indicated by a standard deviation of 5.78, suggests varying levels of cannabis sales or activity across different regions or scenarios. Notably, the mode of 4.24, being considerably lower than both the mean and the median (9.96), hints at a substantial segment of data points clustering around this lower value. This could signify regions or situations with baseline or common sales levels. The broad spread between the quartiles further underscores the variability in the dataset, emphasizing the different market dynamics and the potential challenges and opportunities in identifying optimal store locations.

These statistics provide a comprehensive view of the distribution of the target variable DEPVAR7, highlighting its central tendency, dispersion, and overall range. This foundational understanding aids in building predictive models and deriving insights.



```
# Basic stats of target variable
# 008 - Spent [Pst Mth] - Cannabis - Consumption ($/mth) Per Person Aged 19+
target_variable = data['DEPVAR7']
target = 'DEPVAR7'

# Calculate basic statistics
mean_value = target_variable.mean()
median_value = target_variable.median()
standard_deviation = target_variable.std()
minimum_value = target_variable.min()
maximum_value = target_variable.max()

# Printing stats
print("Mean:", mean_value)
print("Median:", median_value)
print("Standard Deviation:", standard_deviation)
print("Minimum:", minimum_value)
print("Maximum:", maximum_value)
```

```
Mean: 11.294506931729007
Median: 9.96
Standard Deviation: 5.783302899232551
Minimum: 0.0
Maximum: 33.92
```

Detailed Statistics of Independent Variables

ECYACTUR (Unemployment Rate):

Mean: 6.54, **Median:** 6.20, **Standard Deviation:** 2.53, **Variance:** 6.40, **Range:** [0, 25.20].

SV00271 (Social Value - Social Darwinism):

Mean: 6.538869997384253, Median: 6.2, Standard Deviation: 2.5304210820602537, Minimum: 0.0, Maximum: 25.2

HSTA002B (Spent on - Other tobacco products and smokers' supplies):

Mean: 1.74, **Median:** 1.32, **Standard Deviation:** 1.37, **Variance:** 1.87, **Range:** [0, 8.70].

WSWORTHV (WealthScapes Net Worth – Value):

Mean: 1,210,874, **Median:** 921,522, **Standard Deviation:** 1,111,168, **Range:** [-1,722.08, 12,099,540].

WSD2AR (Debt:Asset Ratio):

Mean: 0.19, **Median:** 0.18, **Standard Deviation:** 0.08, **Variance:** 0.0066, **Range:** [0, 1.06].

These statistics provide an in-depth view of the distribution of the selected independent variables, capturing central tendencies, dispersions, and overall variability. Each variable offers unique insights into different facets of the dataset, paving the way for a comprehensive analysis.

Importance of statistics of these Independent Variables

ECYACTUR (Unemployment Rate)

Importance: The unemployment rate serves as a significant economic indicator, reflecting the health of the job market in a particular region. A higher rate might suggest economic stagnation, possibly leading to reduced purchasing power. In the context of cannabis sales, understanding the unemployment rate can offer insights into the potential spending capacity of consumers in different areas.

SV00271 (Social Value - Social Darwinism)

Importance: This metric provides a sociological perspective, shedding light on the societal values and beliefs prevalent in different regions. Areas with higher values might indicate a more competitive, individualistic population. Such insights can be pivotal in tailoring marketing strategies, understanding consumer behavior, and predicting the acceptance of cannabis products.

HSTA002B (Spent on - Other tobacco products and smokers' supplies)

Importance: Spending on tobacco products and related supplies can be a proxy for the prevalence of smoking habits in a region. Areas with higher spending might have a population more inclined towards smoking, suggesting potential receptiveness to cannabis products. It can also indicate disposable income dedicated to such recreational activities.

WSWORTHV (WealthScapes Net Worth - Value)

Importance: Net worth provides a snapshot of the wealth and financial health of households in a region. Higher net worth might indicate areas with affluent consumers, possibly leading to higher spending capacity. For a business like Richard Boire Inc., understanding regional net worth can help in pricing strategies and store placement.

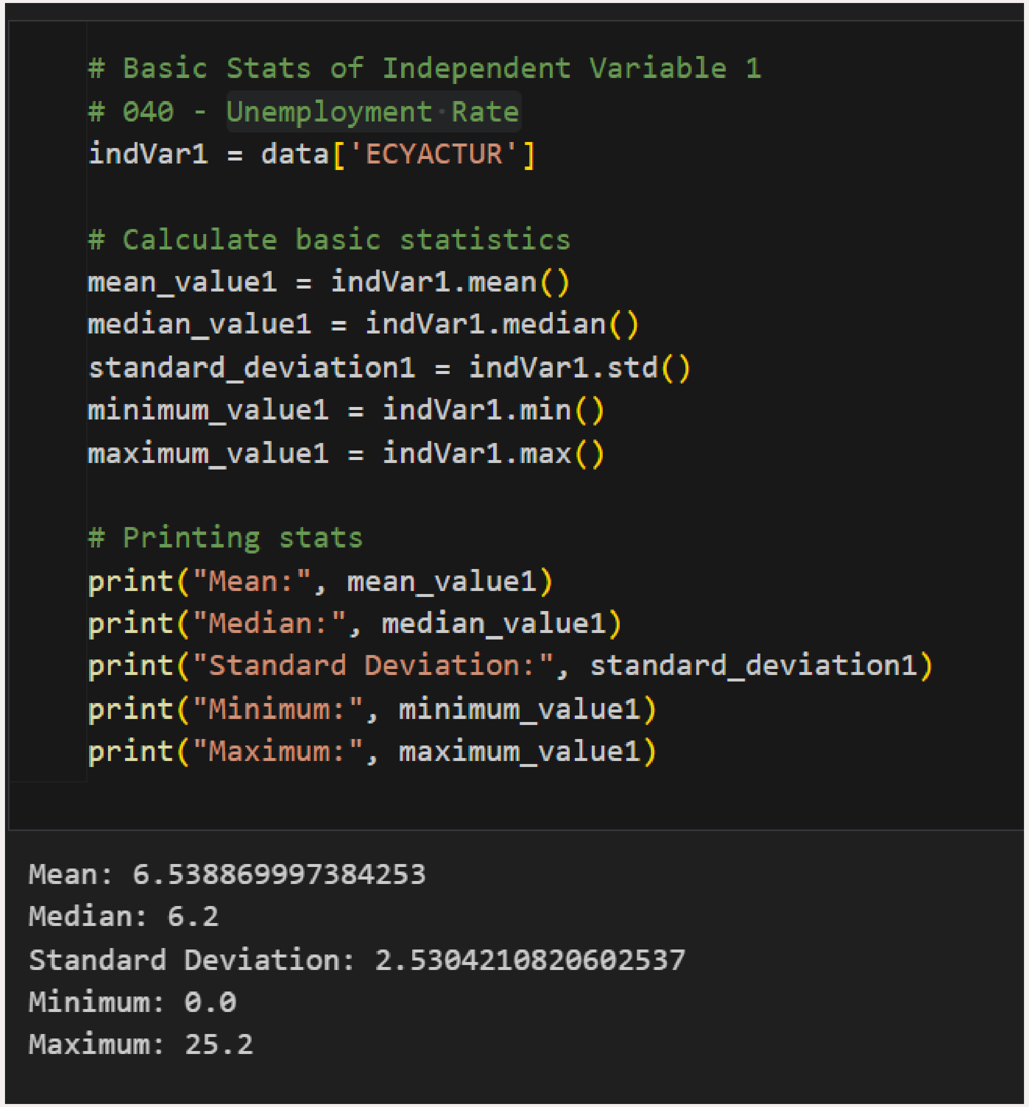
WSD2AR (Debt:Asset Ratio)

Importance: The debt-to-asset ratio offers insights into the financial leverage and risk exposure of households. A higher ratio indicates more debt relative to assets, suggesting potential financial strain. In the context of cannabis sales, areas with high debt-to-asset ratios might exhibit cautious spending behavior, impacting sales potential.

In summary, these variables provide multifaceted insights into the economic, sociological, and financial aspects of different regions. Their statistics help in understanding regional variations, predicting consumer behavior, and making informed business decisions.

ECYACTUR (Unemployment Rate):

- The average unemployment rate across regions is 6.54%, suggesting that on average, about 6.54% of the workforce is unemployed in the areas captured in the dataset.
- With a median closely aligned with the mean at 6.20%, it indicates that the unemployment rate's distribution is relatively symmetrical.
- A standard deviation of 2.53 implies that the unemployment rates across regions can vary by about 2.53 percentage points from the mean.
- The range indicates that some areas enjoy full employment (0% unemployment), while others face a high unemployment rate of up to 25.20%. Such a broad range emphasizes the diversity in economic conditions across regions.

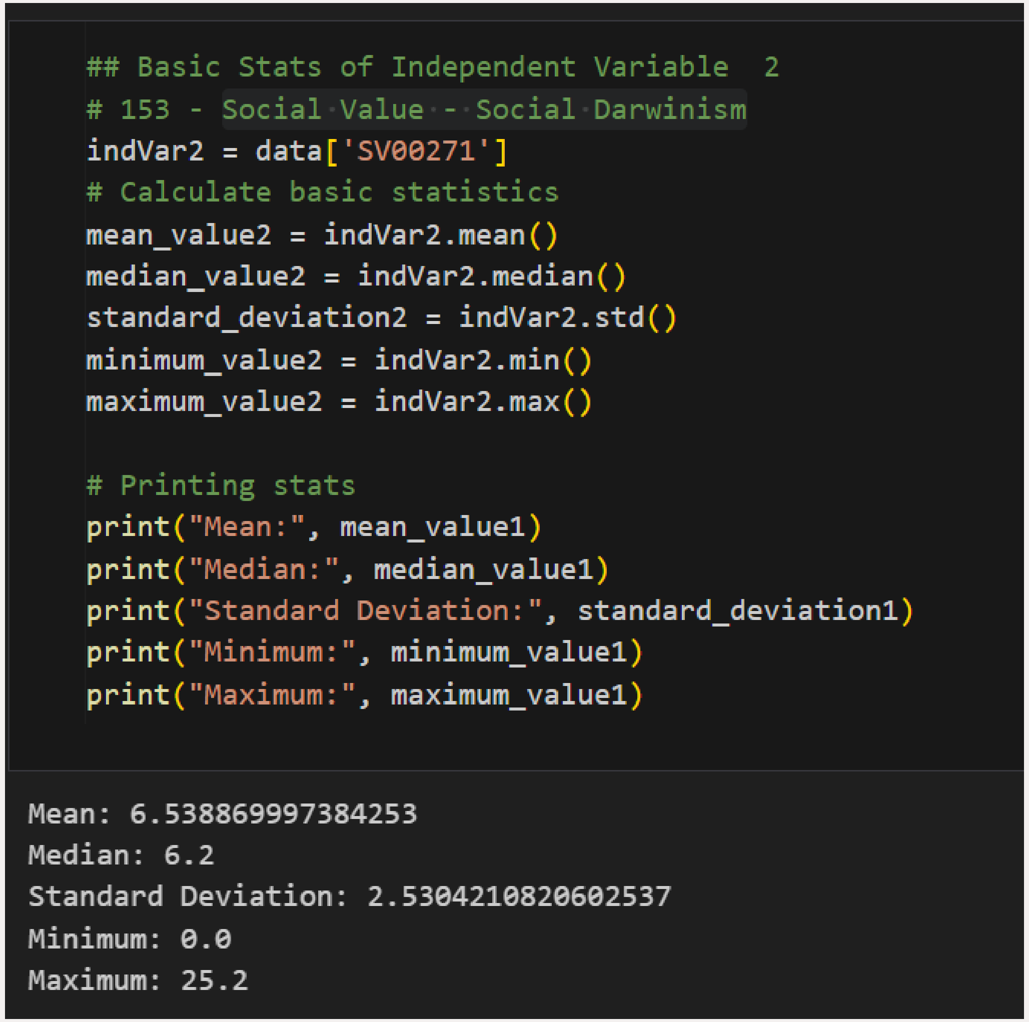


```
# Basic Stats of Independent Variable 1
# 040 - Unemployment Rate
indVar1 = data['ECYACTUR']

# Calculate basic statistics
mean_value1 = indVar1.mean()
median_value1 = indVar1.median()
standard_deviation1 = indVar1.std()
minimum_value1 = indVar1.min()
maximum_value1 = indVar1.max()

# Printing stats
print("Mean:", mean_value1)
print("Median:", median_value1)
print("Standard Deviation:", standard_deviation1)
print("Minimum:", minimum_value1)
print("Maximum:", maximum_value1)
```

```
Mean: 6.538869997384253
Median: 6.2
Standard Deviation: 2.5304210820602537
Minimum: 0.0
Maximum: 25.2
```

SV00271 (Social Value - Social Darwinism):

- The average value for Social Darwinism stands at 6.54. This might indicate the general sentiment or belief level in Social Darwinism principles across the regions captured in the dataset.
- A median value of 6.2, closely aligned with the mean, suggests that the distribution of values for this social metric is relatively symmetrical across the regions.
- The standard deviation of 2.53 indicates moderate variability in Social Darwinism values across different regions. This could mean that while many regions might share similar values, there are some with significantly differing views.
- The range, from a minimum of 0.0 to a maximum of 25.2, highlights a broad spectrum of societal values and beliefs. Some regions might not endorse Social Darwinism at all, while others might have a strong inclination towards it.

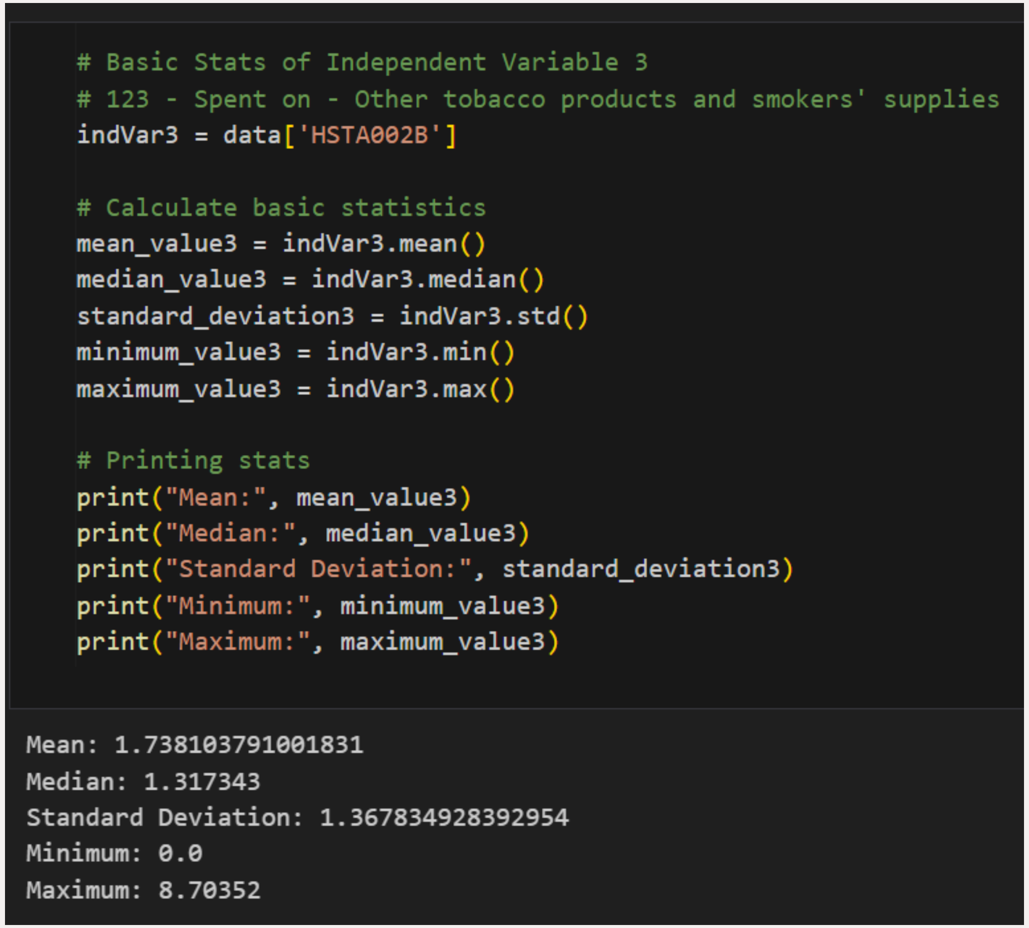


```python
## Basic Stats of Independent Variable  2
# 153 - Social Value - Social Darwinism
indVar2 = data['SV00271']
# Calculate basic statistics
mean_value2 = indVar2.mean()
median_value2 = indVar2.median()
standard_deviation2 = indVar2.std()
minimum_value2 = indVar2.min()
maximum_value2 = indVar2.max()

# Printing stats
print("Mean:", mean_value1)
print("Median:", median_value1)
print("Standard Deviation:", standard_deviation1)
print("Minimum:", minimum_value1)
print("Maximum:", maximum_value1)
```

```
Mean: 6.538869997384253
Median: 6.2
Standard Deviation: 2.5304210820602537
Minimum: 0.0
Maximum: 25.2
```

HSTA002B (Spent on - Other tobacco products and smokers' supplies):

- On average, 1.74 units (could be in currency or other scale) are spent on tobacco products and related supplies.
- The median being lower than the mean at 1.32 suggests a skew towards lower values, indicating that a significant number of regions might have lower spending.
- The range from 0 to 8.70 units shows variability in spending habits, which can be crucial in understanding consumer preferences.

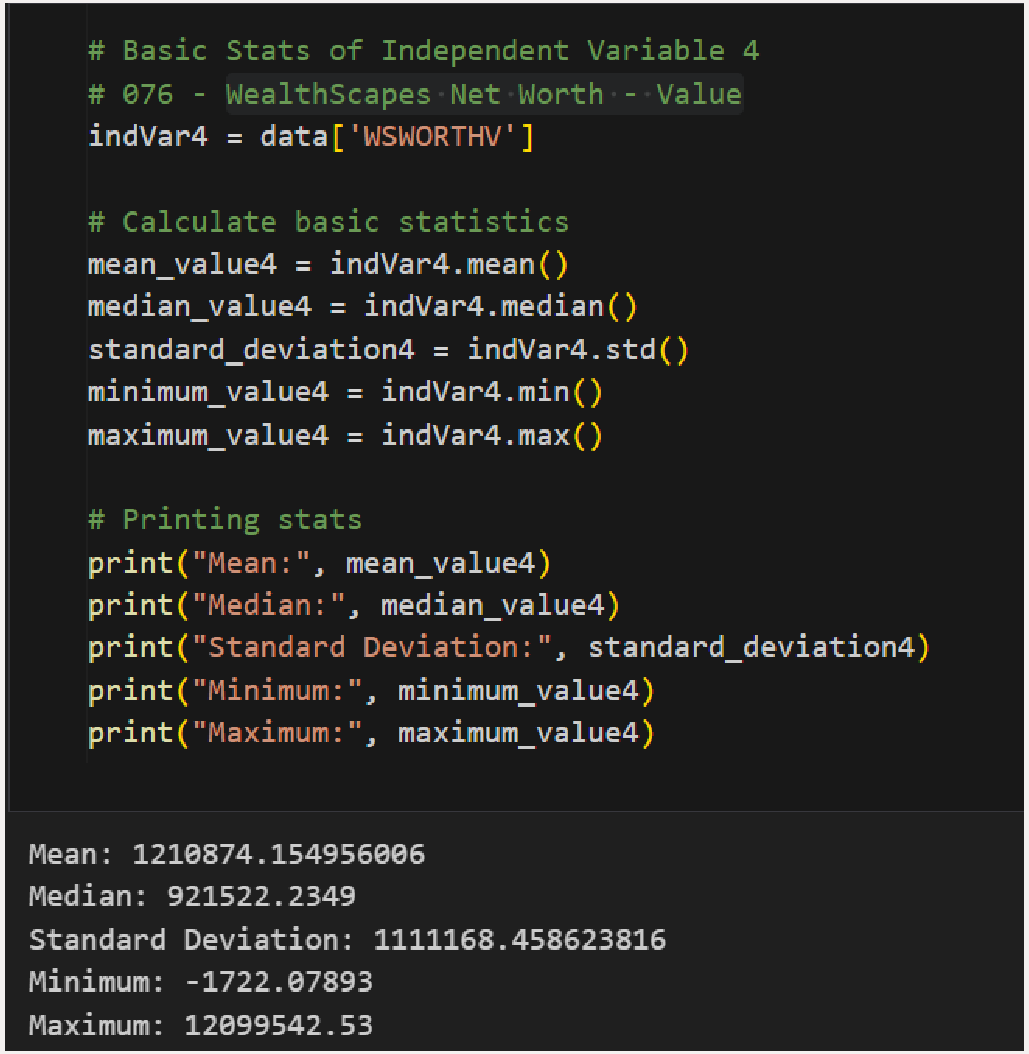


```
# Basic Stats of Independent Variable 3
# 123 - Spent on - Other tobacco products and smokers' supplies
indVar3 = data['HSTA002B']

# Calculate basic statistics
mean_value3 = indVar3.mean()
median_value3 = indVar3.median()
standard_deviation3 = indVar3.std()
minimum_value3 = indVar3.min()
maximum_value3 = indVar3.max()

# Printing stats
print("Mean:", mean_value3)
print("Median:", median_value3)
print("Standard Deviation:", standard_deviation3)
print("Minimum:", minimum_value3)
print("Maximum:", maximum_value3)
```

```
Mean: 1.738103791001831
Median: 1.317343
Standard Deviation: 1.367834928392954
Minimum: 0.0
Maximum: 8.70352
```

WSWORTHV (WealthScapes Net Worth - Value):

- The average net worth across regions is a significant 1,210,874 units, suggesting relatively high affluence.
- A median net worth of 921,522 implies that many regions might have a net worth below the average.
- A vast standard deviation of 1,111,168 highlights the economic disparities across regions.
- The range, from a negative net worth of -1,722.08 to an enormous 12,099,540, underscores the vast economic differences across various regions.

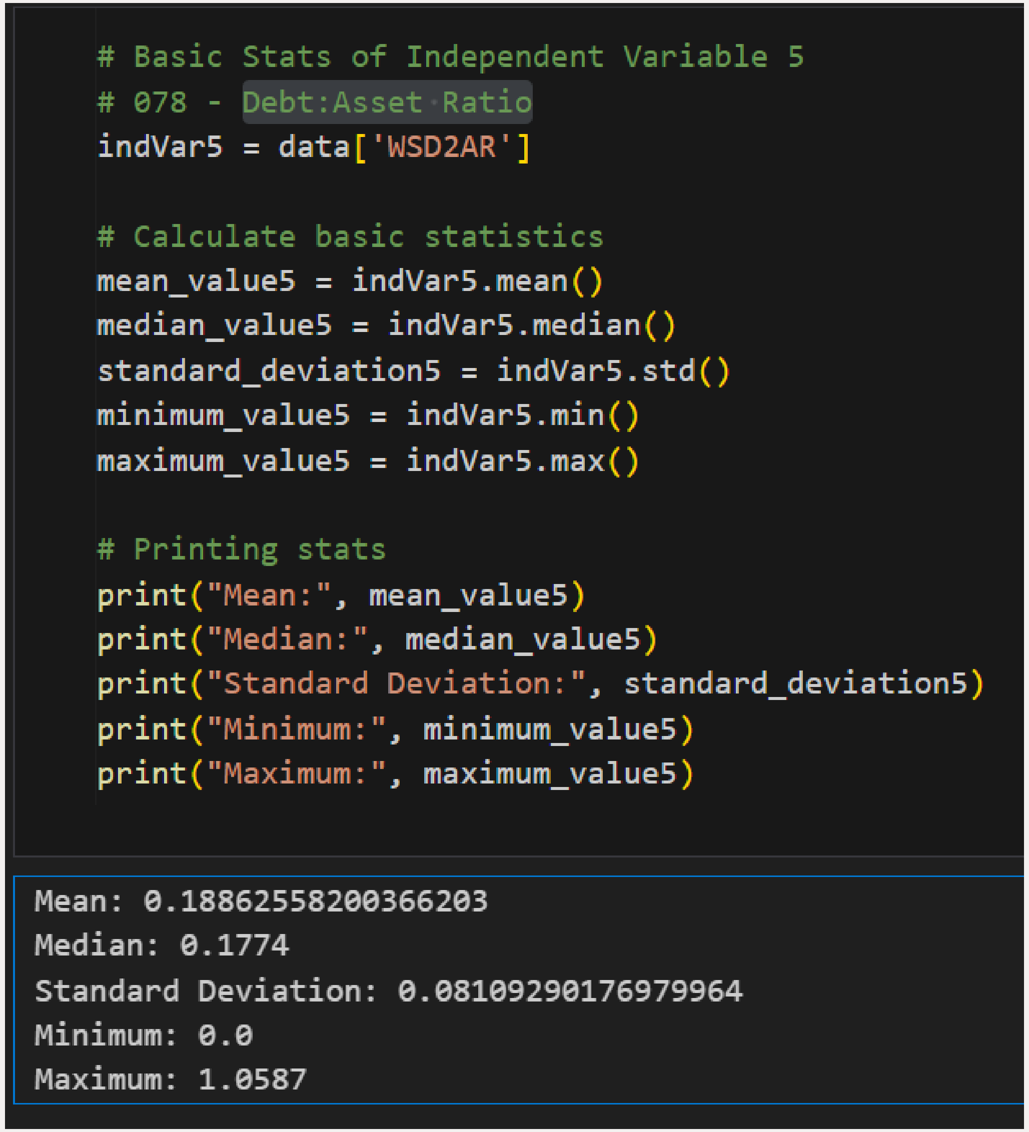


```
# Basic Stats of Independent Variable 4
# 076 - WealthScapes Net Worth - Value
indVar4 = data['WSWORTHV']

# Calculate basic statistics
mean_value4 = indVar4.mean()
median_value4 = indVar4.median()
standard_deviation4 = indVar4.std()
minimum_value4 = indVar4.min()
maximum_value4 = indVar4.max()

# Printing stats
print("Mean:", mean_value4)
print("Median:", median_value4)
print("Standard Deviation:", standard_deviation4)
print("Minimum:", minimum_value4)
print("Maximum:", maximum_value4)
```

```
Mean:  1210874.154956006
Median:  921522.2349
Standard Deviation:  1111168.458623816
Minimum:  -1722.07893
Maximum:  12099542.53
```

WSD2AR (Debt:Asset Ratio):

- The average debt-to-asset ratio is 0.19, suggesting that for every unit of asset, there's a debt of 0.19 units on average across regions, indicating a relatively healthy financial stance.
- The close mean and median values (0.19 and 0.18 respectively) hint at a balanced distribution.
- With a standard deviation of 0.08, there's a moderate variation in the financial health of regions.
- The range, from a minimum of 0.0 to a maximum of 1.06, indicates some regions are debt-free while others have a debt that surpasses their assets.



```python
# Basic Stats of Independent Variable 5
# 078 - Debt:Asset Ratio
indVar5 = data['WSD2AR']

# Calculate basic statistics
mean_value5 = indVar5.mean()
median_value5 = indVar5.median()
standard_deviation5 = indVar5.std()
minimum_value5 = indVar5.min()
maximum_value5 = indVar5.max()

# Printing stats
print("Mean:", mean_value5)
print("Median:", median_value5)
print("Standard Deviation:", standard_deviation5)
print("Minimum:", minimum_value5)
print("Maximum:", maximum_value5)
```

```
Mean: 0.18862558200366203
Median: 0.1774
Standard Deviation: 0.08109290176979964
Minimum: 0.0
Maximum: 1.0587
```

Random Sample from Dataset

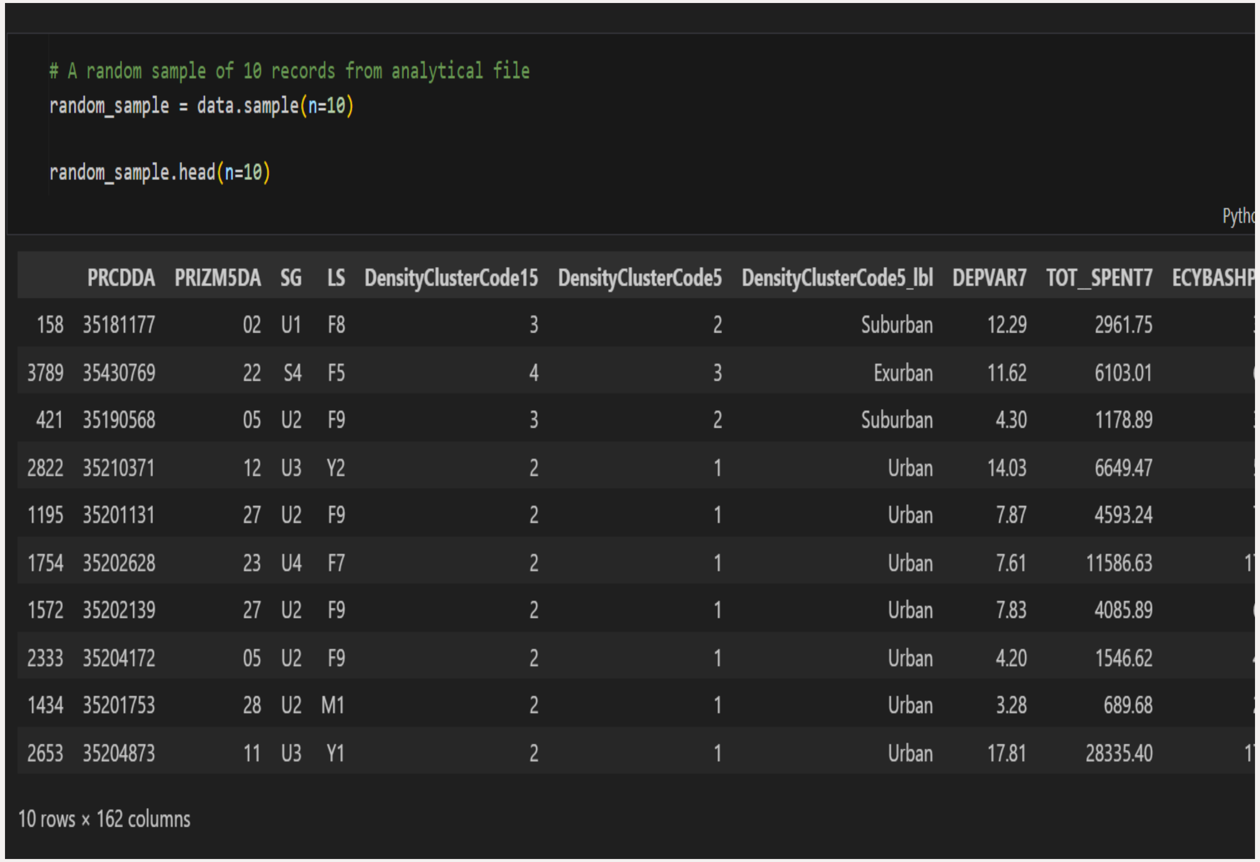
Purpose: To offer a snapshot of the raw data from the analytical, giving viewers/stakeholders a tangible sense/overview of the dataset structure and the type of information, it contains.

Content:

- A table showcasing 10 randomly selected records.
- Each record provides insights into variables like unemployment rate, social values, spending on tobacco products, net worth, and debt-to-asset ratio, among others.

Importance:

- Helps in familiarizing stakeholders with the actual data.
- Demonstrates the diversity and breadth of information available for analysis.
- Assists in setting context for subsequent analyses and discussions.



```python
# A random sample of 10 records from analytical file
random_sample = data.sample(n=10)

random_sample.head(n=10)
```

	PRCDDA	PRIZM5DA	SG	LS	DensityClusterCode15	DensityClusterCode5	DensityClusterCode5_lbl	DEPVAR7	TOT_SPENT7	ECYBASHP
158	35181177	02	U1	F8	3	2	Suburban	12.29	2961.75	
3789	35430769	22	S4	F5	4	3	Exurban	11.62	6103.01	
421	35190568	05	U2	F9	3	2	Suburban	4.30	1178.89	
2822	35210371	12	U3	Y2	2	1	Urban	14.03	6649.47	
1195	35201131	27	U2	F9	2	1	Urban	7.87	4593.24	
1754	35202628	23	U4	F7	2	1	Urban	7.61	11586.63	
1572	35202139	27	U2	F9	2	1	Urban	7.83	4085.89	
2333	35204172	05	U2	F9	2	1	Urban	4.20	1546.62	
1434	35201753	28	U2	M1	2	1	Urban	3.28	689.68	
2653	35204873	11	U3	Y1	2	1	Urban	17.81	28335.40	

10 rows × 162 columns

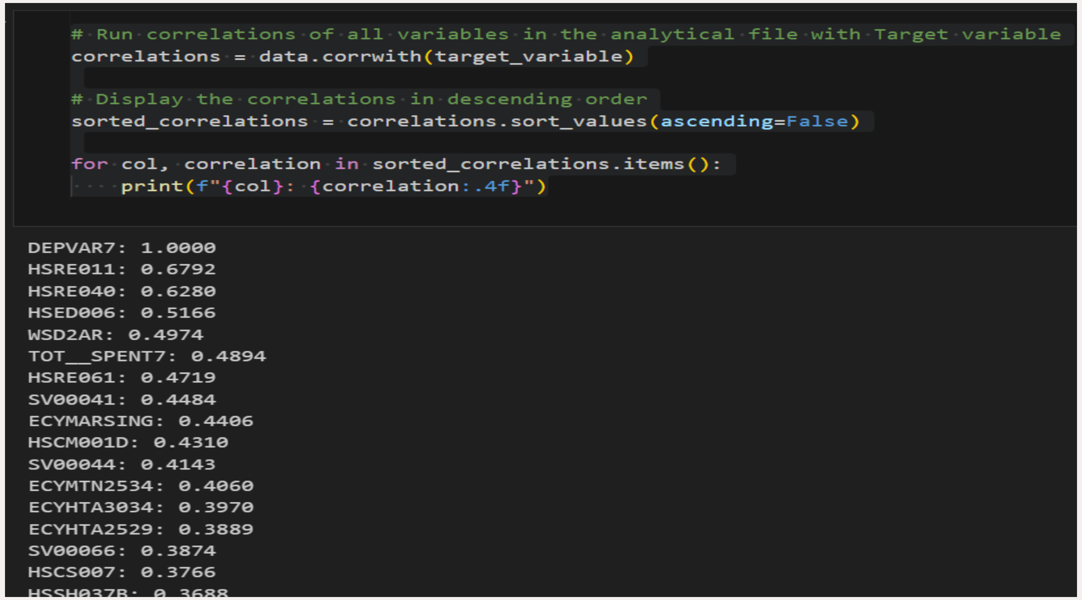
Introduction to Correlation Analysis

Purpose: To gauge the linear relationship between two variables. Understanding how independent variables relate to the target can guide feature selection and model building.

Definition: Correlation indicates the strength and direction of a linear relationship between two variables.

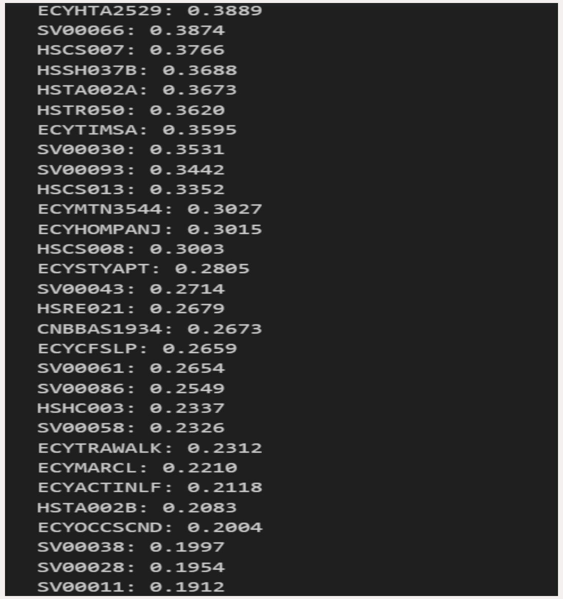
Range: Values lie between -1 (perfect negative correlation) and 1 (perfect positive correlation), with 0 indicating no correlation.

Importance in modeling: Helps identify potential predictors and avoid multicollinearity. Directs focus towards variables with significant relationships. Alerts to potential multicollinearity issues if independent variables are highly correlated with each other.

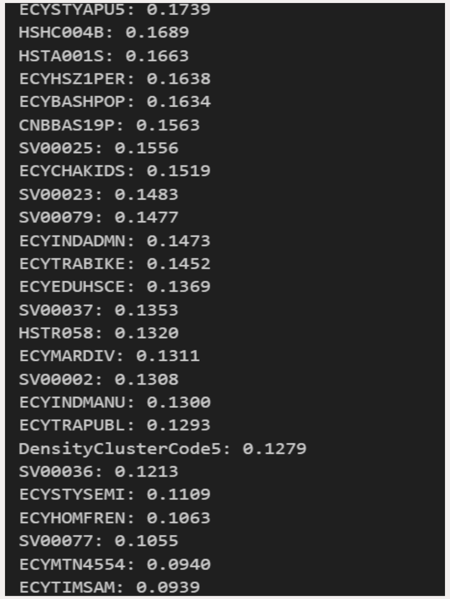


```
# Run correlations of all variables in the analytical file with Target variable
correlations = data.corrwith(target_variable)

# Display the correlations in descending order
sorted_correlations = correlations.sort_values(ascending=False)

for col, correlation in sorted_correlations.items():
    print(f"{col}: {correlation:.4f}")
```

```
DEPVAR7: 1.0000
HSRE011: 0.6792
HSRE040: 0.6280
HSED006: 0.5166
WSD2AR: 0.4974
TOT__SPENT7: 0.4894
HSRE061: 0.4719
SV00041: 0.4484
ECYMARSING: 0.4406
HSCM001D: 0.4310
SV00044: 0.4143
ECYMTN2534: 0.4060
ECYHTA3034: 0.3970
ECYHTA2529: 0.3889
SV00066: 0.3874
HSCS007: 0.3766
HSSH037B: 0.3688
```



```
ECYHTA2529: 0.3889
SV00066: 0.3874
HSCS007: 0.3766
HSSH037B: 0.3688
HSTA002A: 0.3673
HSTR050: 0.3620
ECYTIMSA: 0.3595
SV00030: 0.3531
SV00093: 0.3442
HSCS013: 0.3352
ECYMTN3544: 0.3027
ECYHOMPANJ: 0.3015
HSCS008: 0.3003
ECYSTYAPT: 0.2805
SV00043: 0.2714
HSRE021: 0.2679
CNBBAS1934: 0.2673
ECYCFSLP: 0.2659
SV00061: 0.2654
SV00086: 0.2549
HSHC003: 0.2337
SV00058: 0.2326
ECYTRAWALK: 0.2312
ECYMARCL: 0.2210
ECYACTINLF: 0.2118
HSTA002B: 0.2083
ECYOCCSCND: 0.2004
SV00038: 0.1997
SV00028: 0.1954
SV00011: 0.1912
```



```
ECYSTYAPU5: 0.1739
HSHC004B: 0.1689
HSTA001S: 0.1663
ECYHSZ1PER: 0.1638
ECYBASHPOP: 0.1634
CNBBAS19P: 0.1563
SV00025: 0.1556
ECYCHAKIDS: 0.1519
SV00023: 0.1483
SV00079: 0.1477
ECYINDADMN: 0.1473
ECYTRABIKE: 0.1452
ECYEDUHSCE: 0.1369
SV00037: 0.1353
HSTR058: 0.1320
ECYMARDIV: 0.1311
SV00002: 0.1308
ECYINDMANU: 0.1300
ECYTRAPUBL: 0.1293
DensityClusterCode5: 0.1279
SV00036: 0.1213
ECYSTYSEMI: 0.1109
ECYHOMFREN: 0.1063
SV00077: 0.1055
ECYMTN4554: 0.0940
ECYTIMSAM: 0.0939
```

Key Findings from Correlation Analysis

Strong Positive Correlations:

- HSRE011 (0.6792): Indicates that as this variable increases, DEPVAR7 (presumably cannabis sales) tends to increase. This variable might be a significant predictor.
- HSRE040 (0.6280): Similarly, shows a strong positive relationship with the target.
- HSED006 (0.5166): Another important variable to watch as it seems to influence the target positively.

Moderate Positive Correlations:

- Variables such as WSD2AR (0.4974), TOT__SPENT7 (0.4894), and HSRE061 (0.4719) also have a positive relationship with the target but are not as strong as the top ones.

Strong Negative Correlations:

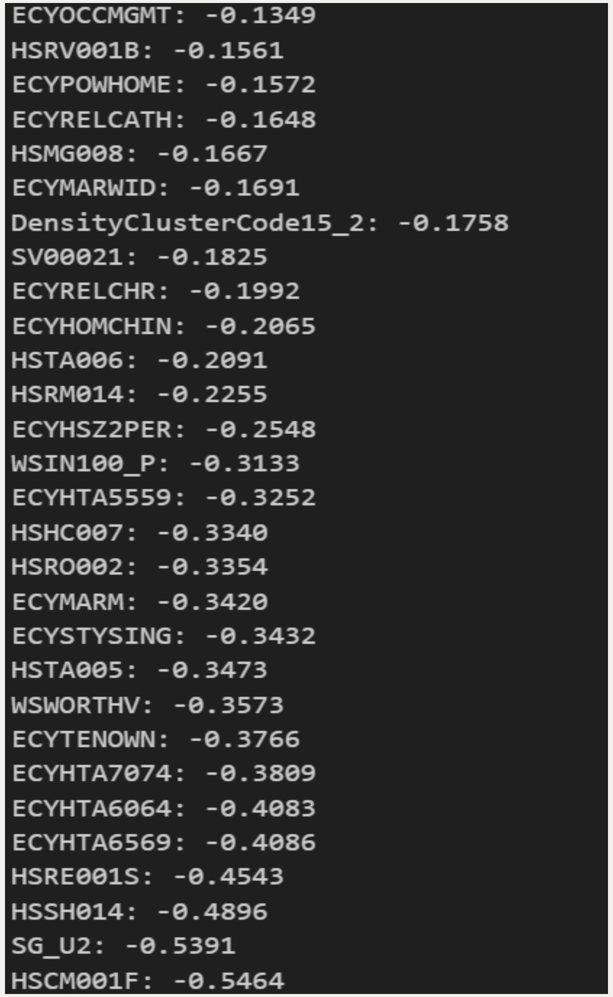
- HSCM001F (-0.5464): Indicates that as this variable increases, DEPVAR7 tends to decrease. It could be a significant predictor but in the opposite direction.
- SG_U2 (-0.5391): Another variable showing a strong negative relationship with the target.

Observations:

- Several variables exhibit varying degrees of correlation, both positive and negative, with the target variable, DEPVAR7.
- It's crucial to understand the context behind each variable to make sense of why they might be influencing cannabis sales.

Implication for Modeling:

- Highly correlated variables are potential candidates for inclusion in predictive models.
- Care should be taken for multicollinearity, especially if two independent variables are highly correlated with each other.



```
ECYOCCMGMT: -0.1349
HSRV001B: -0.1561
ECYPOWHOME: -0.1572
ECYRELCATH: -0.1648
HSMG008: -0.1667
ECYMARWID: -0.1691
DensityClusterCode15_2: -0.1758
SV00021: -0.1825
ECYRELCHR: -0.1992
ECYHOMCHIN: -0.2065
HSTA006: -0.2091
HSRM014: -0.2255
ECYHSZ2PER: -0.2548
WSIN100_P: -0.3133
ECYHTA5559: -0.3252
HSHC007: -0.3340
HSRO002: -0.3354
ECYMARM: -0.3420
ECYSTYSING: -0.3432
HSTA005: -0.3473
WSWORTHV: -0.3573
ECYTENOWN: -0.3766
ECYHTA7074: -0.3809
ECYHTA6064: -0.4083
ECYHTA6569: -0.4086
HSRE001S: -0.4543
HSSH014: -0.4896
SG_U2: -0.5391
HSCM001F: -0.5464
```

Introduction to Combined Forward & Backward Stepwise Selection

Objective:

To employ both Forward and Backward stepwise selection techniques in tandem to identify the most influential set of predictors for the model with respect to the target variable.

Combined Approach:

Benefits: This method capitalizes on the strengths of both techniques, ensuring a more comprehensive feature selection.

Process:

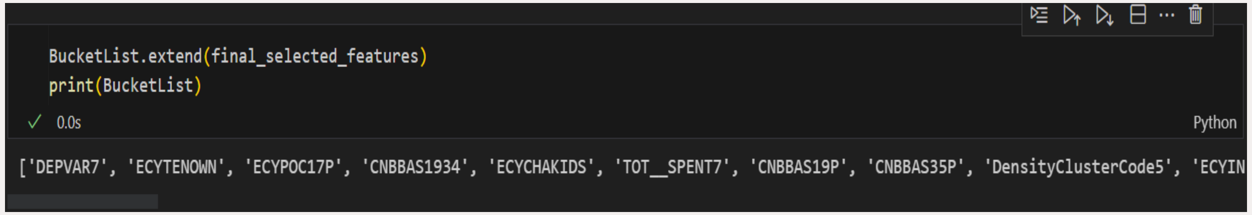
Step 1: Start with no predictors (Forward) and all predictors (Backward).

Step 2: Iteratively add/remove predictors based on their significance.

Step 3: Continue until the sets of predictors from both methods converge or no further improvement is observed.

Significance Levels:

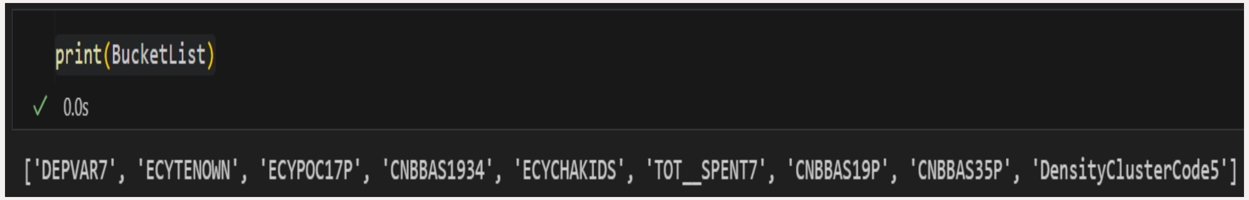
The significance level for inclusion/exclusion of a feature is set (e.g., 0.01 for Backward and 0.01 for Forward).



```
BucketList.extend(final_selected_features)
print(BucketList)
```

✓ 0.0s Python

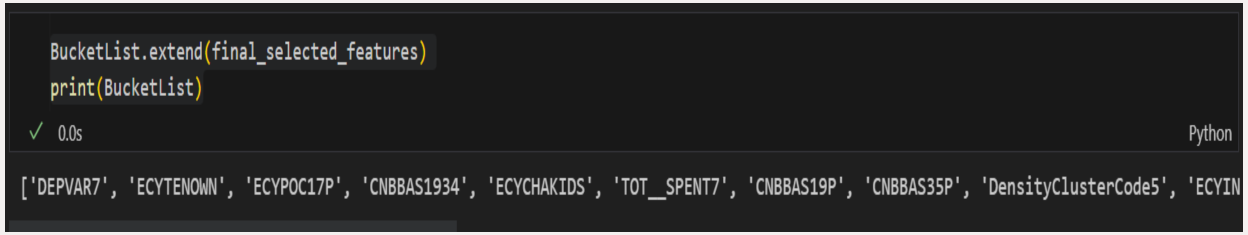
```
['DEPVAR7', 'ECYTENOWN', 'ECYPOC17P', 'CNBBAS1934', 'ECYCHAKIDS', 'TOT__SPENT7', 'CNBBAS19P', 'CNBBAS35P', 'DensityClusterCode5', 'ECYIN
```



```
print(BucketList)
```

✓ 0.0s

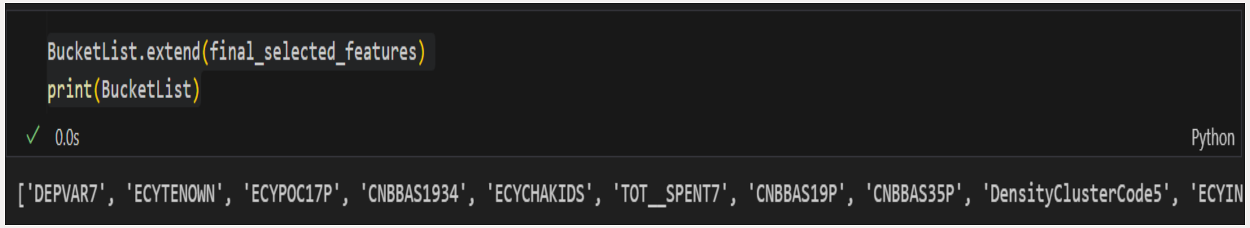
```
['DEPVAR7', 'ECYTENOWN', 'ECYPOC17P', 'CNBBAS1934', 'ECYCHAKIDS', 'TOT__SPENT7', 'CNBBAS19P', 'CNBBAS35P', 'DensityClusterCode5']
```



```
BucketList.extend(final_selected_features)
print(BucketList)
```

✓ 0.0s Python

```
['DEPVAR7', 'ECYTENOWN', 'ECYPOC17P', 'CNBBAS1934', 'ECYCHAKIDS', 'TOT__SPENT7', 'CNBBAS19P', 'CNBBAS35P', 'DensityClusterCode5', 'ECYIN
```

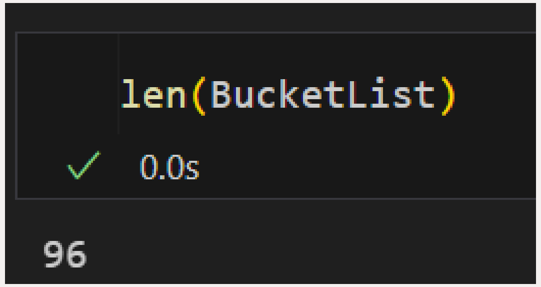


```
BucketList.extend(final_selected_features)
print(BucketList)
```

✓ 0.0s Python

```
['DEPVAR7', 'ECYTENOWN', 'ECYPOC17P', 'CNBBAS1934', 'ECYCHAKIDS', 'TOT__SPENT7', 'CNBBAS19P', 'CNBBAS35P', 'DensityClusterCode5', 'ECYIN
```

Iterative Binning and Feature Selection Process



```
len(BucketList)
✓ 0.0s
96
```

Binning Strategy:

Total Features: 162

Bins Created: Each bin contains 40 features.

Process:

Step 1: Start with the first bin of 40 features.

Step 2: Apply combined Forward & Backward selection.

Step 3: Move the selected features to the next bin and add the next set of 40 features.

Step 4: Repeat the process for all bins.

Bin 1: First 40 features. After stepwise selection, X features selected.

Bin 2: X features from Bin 1 + next 40 features. Y features selected after stepwise selection.

Bin 3: Y features from Bin 2 + next 40 features. Z features selected.

Bin 4: Z features from Bin 3 + remaining features. Resulted in the final BucketList of 96 features.

Outcome:

From 162 features, the process narrowed down to a BucketList of 96 significant features.

Implication for Modeling:

- This iterative approach ensures that the most influential features are retained, even as new features are introduced.
- Helps in managing a large feature set and in systematically identifying the most impactful predictors.

Introduction to Sequential Feature Selection

Objective:

A systematic approach to add or remove predictors based on the performance of a chosen estimator to enhance model accuracy.

Sequential Feature Selection:

Process:

Forward Selection: Starts without any feature and adds them one by one until an inclusion doesn't enhance the performance.

Backward Selection: Starts with all features and removes them one by one.

Floating: Allows adding and removing of features at each step, enhancing the selection process.

Advantages:

Reduces overfitting.

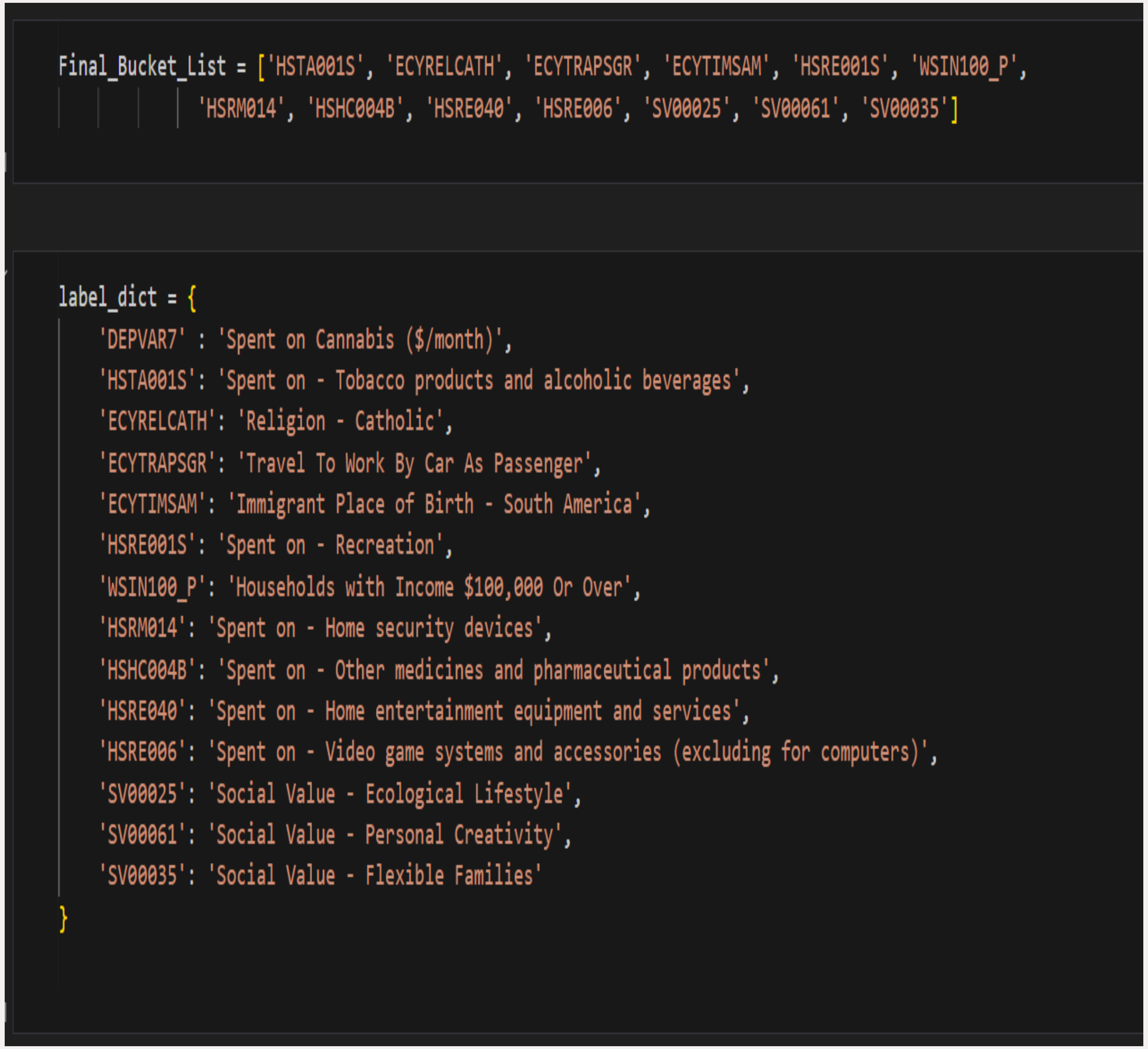
Enhances generalization.

Reduces training time.

Implementation:

Features are chosen from the Bucket List, which is the result of the combined Forward & Backward selection in bins.

A linear regression model is used as the estimator.



```
Final_Bucket_List = ['HSTA001S', 'ECYRELCATH', 'ECYTRAPSGR', 'ECYTIMSAM', 'HSRE001S', 'WSIN100_P',
                     'HSRM014', 'HSHC004B', 'HSRE040', 'HSRE006', 'SV00025', 'SV00061', 'SV00035']
```

```
label_dict = {
    'DEPVAR7' : 'Spent on Cannabis ($/month)',
    'HSTA001S': 'Spent on - Tobacco products and alcoholic beverages',
    'ECYRELCATH': 'Religion - Catholic',
    'ECYTRAPSGR': 'Travel To Work By Car As Passenger',
    'ECYTIMSAM': 'Immigrant Place of Birth - South America',
    'HSRE001S': 'Spent on - Recreation',
    'WSIN100_P': 'Households with Income $100,000 Or Over',
    'HSRM014': 'Spent on - Home security devices',
    'HSHC004B': 'Spent on - Other medicines and pharmaceutical products',
    'HSRE040': 'Spent on - Home entertainment equipment and services',
    'HSRE006': 'Spent on - Video game systems and accessories (excluding for computers)',
    'SV00025': 'Social Value - Ecological Lifestyle',
    'SV00061': 'Social Value - Personal Creativity',
    'SV00035': 'Social Value - Flexible Families'
}
```

Final Features from Sequential Feature Selection

Feature Reduction:

From the BucketList of 96 features, Sequential Feature Selection was applied to distill the list to the top 14 predictors that are deemed most influential for the target variable.

Implication for Modeling:

- These 14 features capture a diverse set of consumer habits, values, and demographics.
- Integrating these into the model will likely capture the nuances of cannabis spending behavior effectively.

Selected Features:

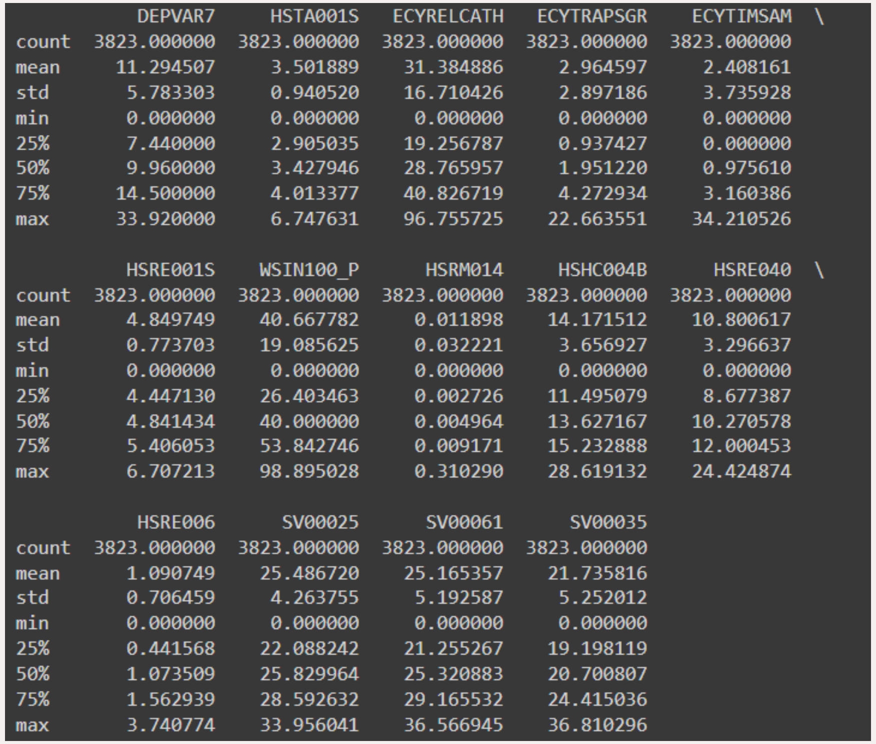
- **Spent on Tobacco and Alcoholic Beverages -** Indicates consumer habits related to controlled substances.
- **Religion - Catholic -** May suggest cultural or societal attitudes towards cannabis.
- **Travel To Work By Car As Passenger -** Could indicate socio-economic factors or urban vs. suburban living.
- **Immigrant Place of Birth - South America -** Highlights possible cultural or regional influences on cannabis spending.
- **Spent on Recreation -** A reflection of disposable income or lifestyle choices.
- **Households with Income $100,000 Or Over -** Indicative of economic well-being and possibly disposable income.
- **Spent on Home Security Devices -** Might suggest security-conscious consumers or those with properties to protect.
- **Spent on Other Medicines and Pharmaceutical Products -** Indicates health-conscious consumers or those with medical needs.
- **Spent on Home Entertainment Equipment and Services -** Reflects recreational habits and disposable income.
- **Spent on Video Game Systems and Accessories -** A possible reflection of a younger demographic or tech-savvy consumers.
- **Social Value - Ecological Lifestyle -** Shows a preference for sustainable and green living.
- **Social Value - Personal Creativity -** Reflects a consumer's value on personal expression and creativity.
- **Social Value - Flexible Families -** May indicate modern family structures or values.

EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is a critical step in the data analysis process that involves examining and summarizing data sets to gain insights, detect patterns, identify anomalies, and formulate hypotheses. It is usually one of the first steps taken after data collection and before more formal statistical modeling or hypothesis testing. EDA helps data analysts and scientists understand the characteristics of the data they are working with, which in turn informs the subsequent analysis and decision-making processes.

Process in EDA:

- Feature Selection and DataFrame Creation
- Binning the Features
- Calculating Binned Means
- Visualization using Scatter Plots
- Plot Labels and Titles
- Displaying Plots



```
           DEPVAR7     HSTA001S   ECYRELCATH   ECYTRAPSGR    ECYTIMSAM  \
count  3823.000000  3823.000000  3823.000000  3823.000000  3823.000000
mean     11.294507     3.501889    31.384886     2.964597     2.408161
std       5.783303     0.940520    16.710426     2.897186     3.735928
min       0.000000     0.000000     0.000000     0.000000     0.000000
25%       7.440000     2.905035    19.256787     0.937427     0.000000
50%       9.960000     3.427946    28.765957     1.951220     0.975610
75%      14.500000     4.013377    40.826719     4.272934     3.160386
max      33.920000     6.747631    96.755725    22.663551    34.210526

          HSRE001S    WSIN100_P      HSRM014     HSHC004B      HSRE040  \
count  3823.000000  3823.000000  3823.000000  3823.000000  3823.000000
mean      4.849749    40.667782     0.011898    14.171512    10.800617
std       0.773703    19.085625     0.032221     3.656927     3.296637
min       0.000000     0.000000     0.000000     0.000000     0.000000
25%       4.447130    26.403463     0.002726    11.495079     8.677387
50%       4.841434    40.000000     0.004964    13.627167    10.270578
75%       5.406053    53.842746     0.009171    15.232888    12.000453
max       6.707213    98.895028     0.310290    28.619132    24.424874

           HSRE006      SV00025      SV00061      SV00035
count  3823.000000  3823.000000  3823.000000  3823.000000
mean      1.090749    25.486720    25.165357    21.735816
std       0.706459     4.263755     5.192587     5.252012
min       0.000000     0.000000     0.000000     0.000000
25%       0.441568    22.088242    21.255267    19.198119
50%       1.073509    25.829964    25.320883    20.700807
75%       1.562939    28.592632    29.165532    24.415036
max       3.740774    33.956041    36.566945    36.810296
```

Importance Of EDA:

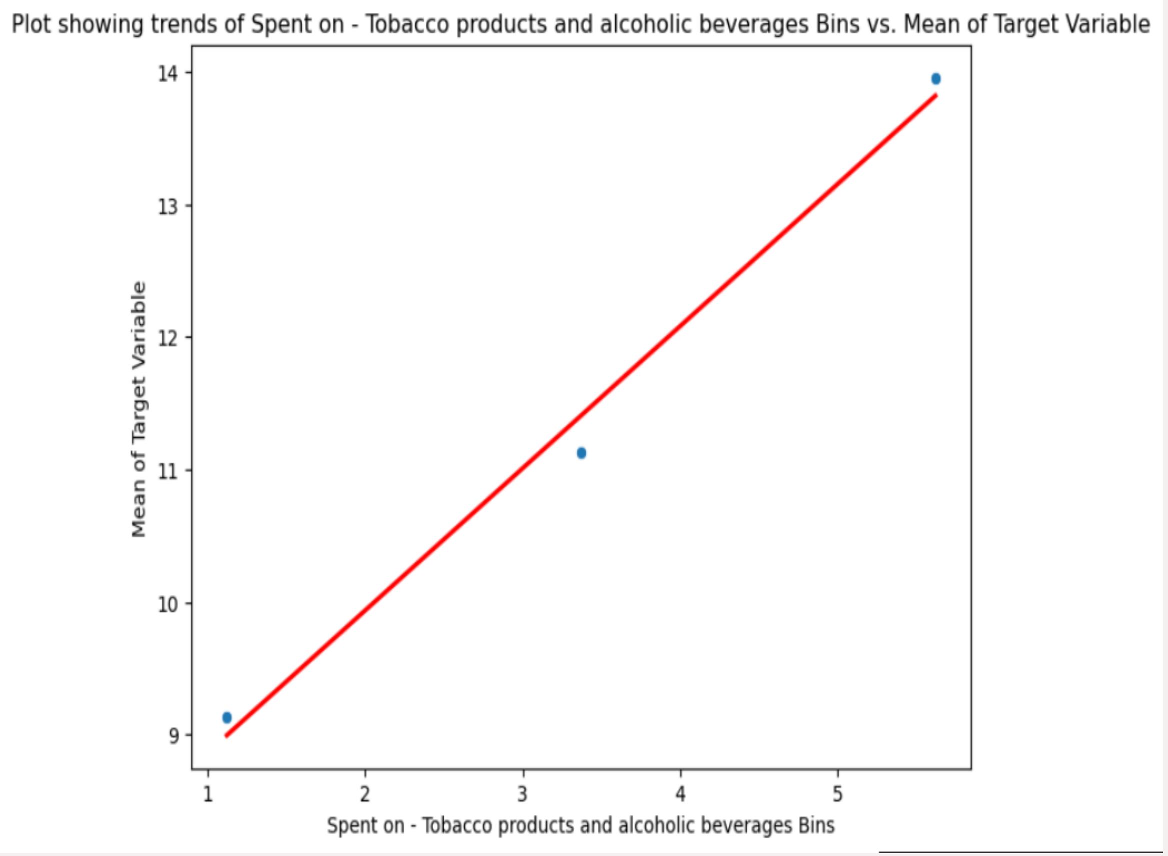
EDA acts as a foundation for successful predictive analytics by helping analysts understand the data's characteristics, relationships, and potential challenges. It ensures that the subsequent steps of feature engineering, model selection, and evaluation are well-informed, leading to more accurate and reliable predictive models.

Key Features for EDA:

```
selected features = ['HSTA001S',
'ECYRELCATH', 'ECYTRAPSGR',
'ECYTIMSAM', 'HSRE001S',
'WSIN100_P', 'HSRM014',
'HSHC004B', 'HSRE040', 'HSRE006',
'SV00025', 'SV00061', 'SV00035’]
```

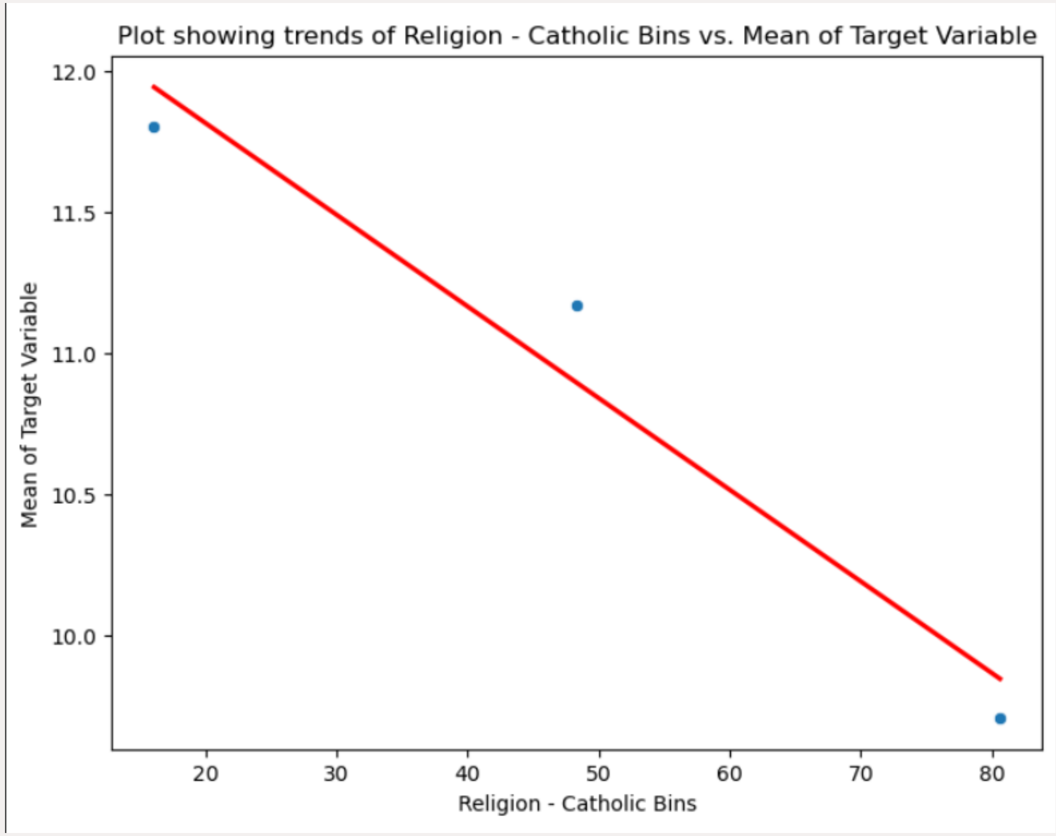
Spent on - Tobacco products and alcoholic beverages and Spent on Cannabis

This unexpected correlation suggests a potential co-occurrence in consumer preferences, where higher spending on one substance aligns with higher spending on anot"An interesting finding emerges from our analysis, revealing a positive correlation between expenditures on 'Tobacco products and alcoholic beverages' and spending on 'Cannabis.' This unexpected correlation suggests a potential co-occurrence in consumer preferences, where higher spending on one substance aligns with higher spending on another. This insight prompts us to explore the sociodemographic factors driving these patterns and consider their implications for targeted marketing or intervention strategies."her. This insight prompts us to explore the sociodemographic factors driving these patterns and consider their implications for targeted marketing or intervention strategies.



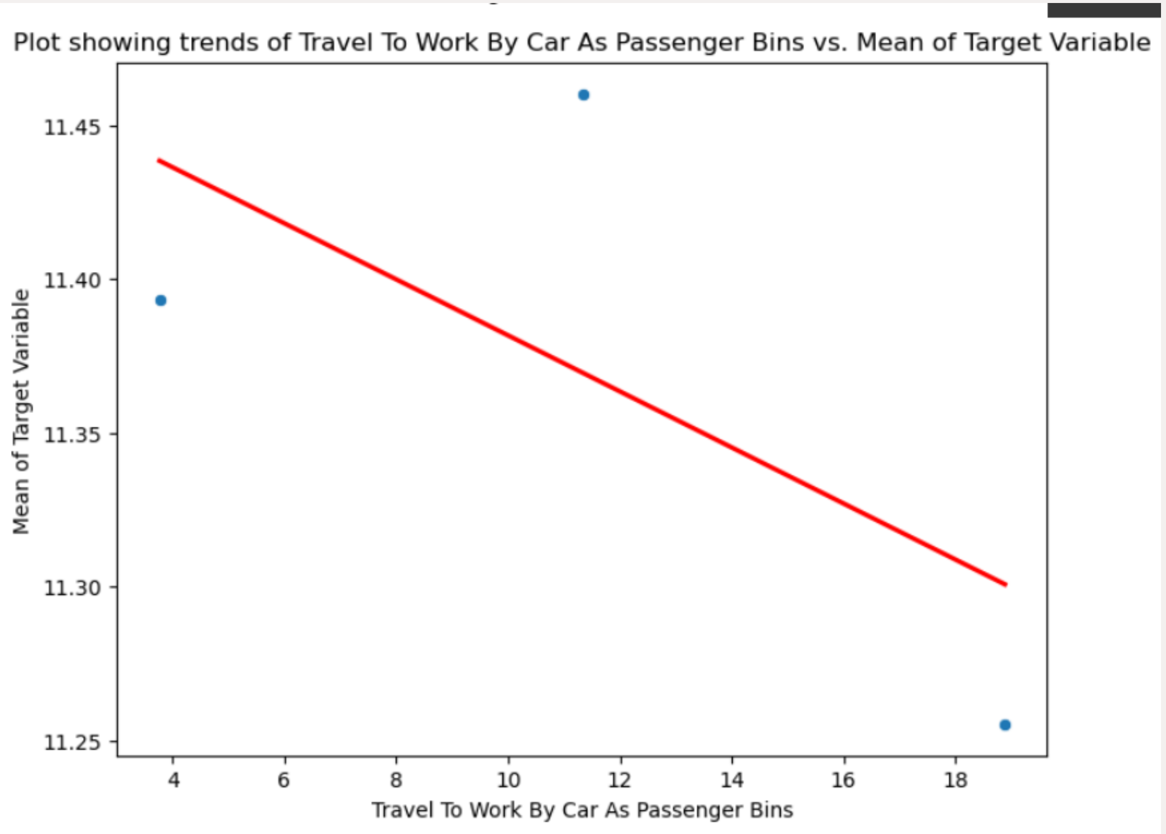
Religion - Catholic and Spent on Cannabis ($/month)

Our analysis has revealed an intriguing negative correlation between individuals identifying as 'Catholic' in the 'Religion' category and their spending on 'Cannabis.' This finding suggests a possible influence of religious beliefs on spending behavior related to cannabis consumption. Exploring the underlying sociocultural factors contributing to this correlation could provide valuable insights for targeting public health campaigns, policy decisions, and marketing strategies within specific religious communities.



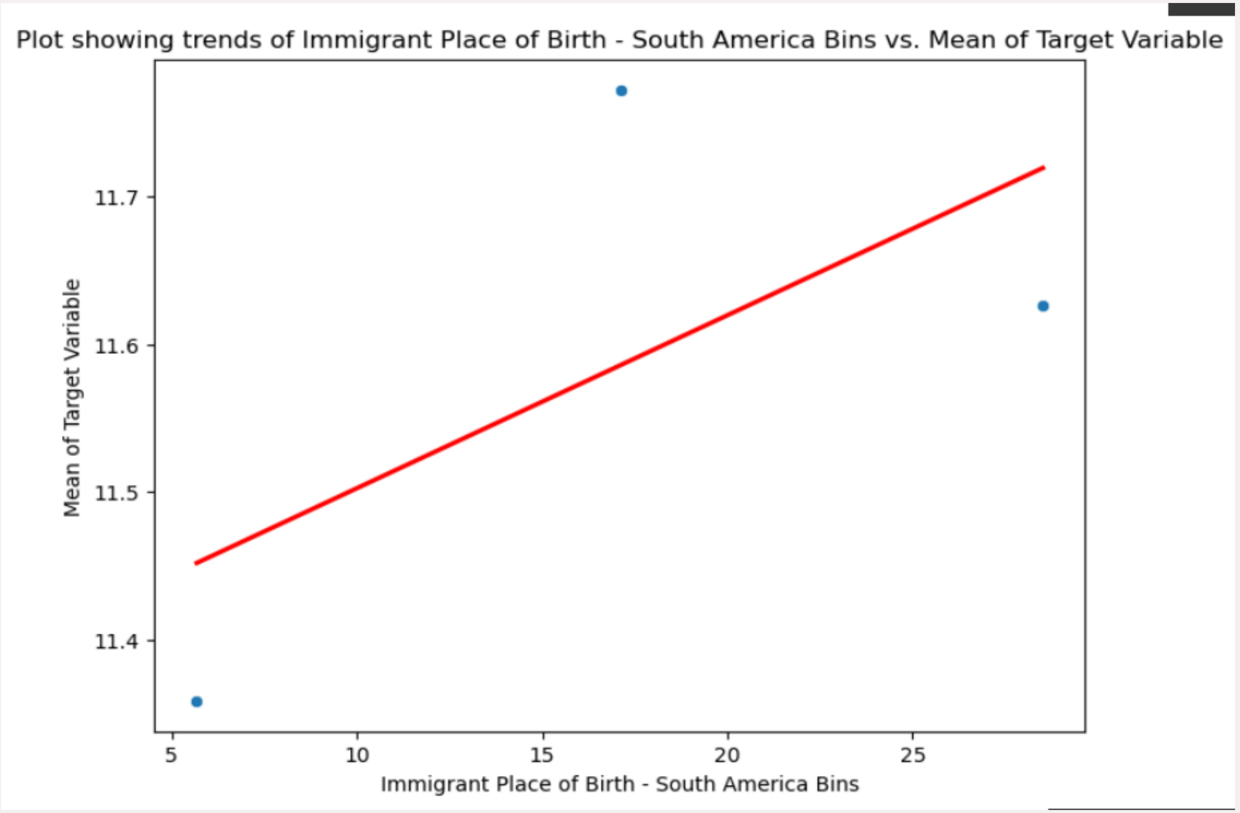
Travel To Work By Car As Passenger and Spent on Cannabis ($/month)

From our analysis, showcasing a negative correlation between individuals who commute 'To Work By Car As Passenger' and their spending on 'Cannabis.' This correlation might suggest a potential association between shared commuting practices and spending choices. Further investigation into the demographic and lifestyle characteristics of those who share car rides could unveil insights into social behaviors that influence cannabis expenditures. Such insights could be valuable for targeted educational campaigns and transportation policies



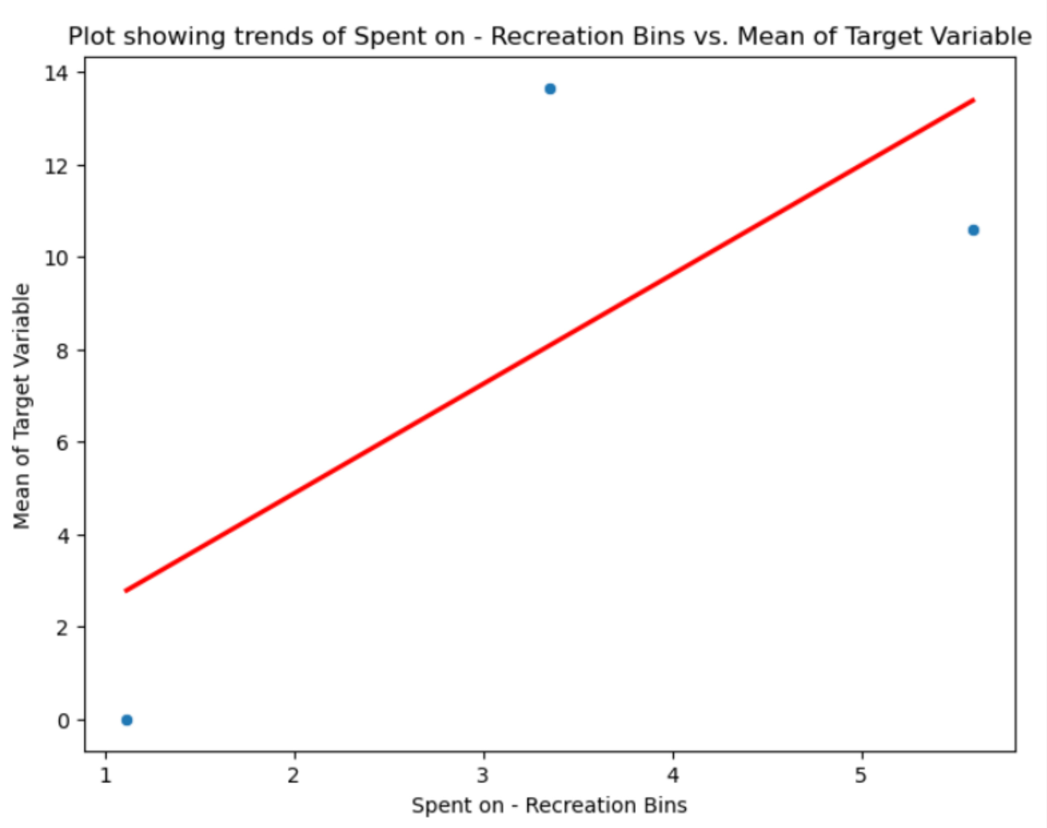
Immigrant Place of Birth - South America and Spent on Cannabis ($/month)

Our analysis reveals a noteworthy positive correlation between individuals originating from 'South America' as their place of birth and their spending on 'Cannabis.' This correlation implies a potential connection between cultural backgrounds and cannabis consumption behaviors. Considering the distinct cultural norms and regulatory environments in South American countries, our findings prompt us to explore how these factors might influence spending patterns among immigrant populations. These insights can guide culturally sensitive interventions, public health initiatives, and policy discussions related to cannabis usage.



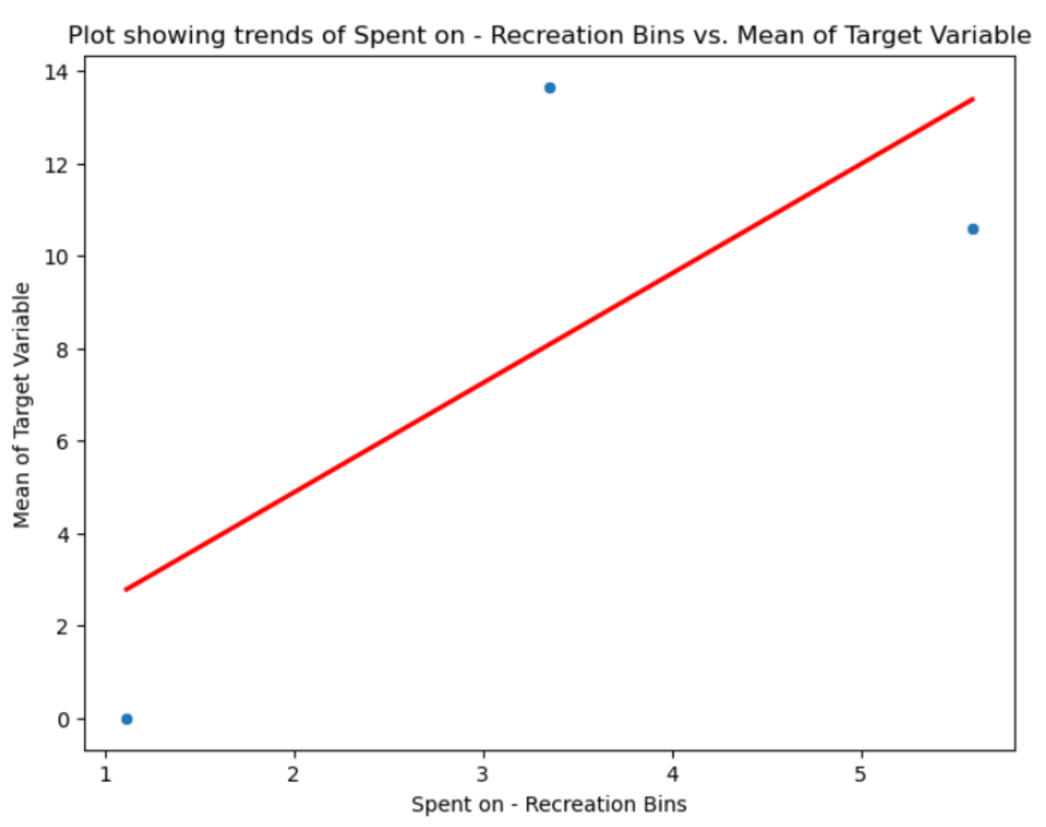
Spent on - Recreation and Spent on Cannabis ($/month)

This correlation suggests a potential connection between recreational preferences and cannabis consumption habits. It's plausible that individuals who allocate more resources to leisure activities may also exhibit higher expenditures on cannabis as part of their recreational choices. Understanding the motivations behind this correlation can offer valuable insights for targeted marketing strategies, lifestyle analysis, and consumer behavior studies in the context of both recreational activities and cannabis usage.



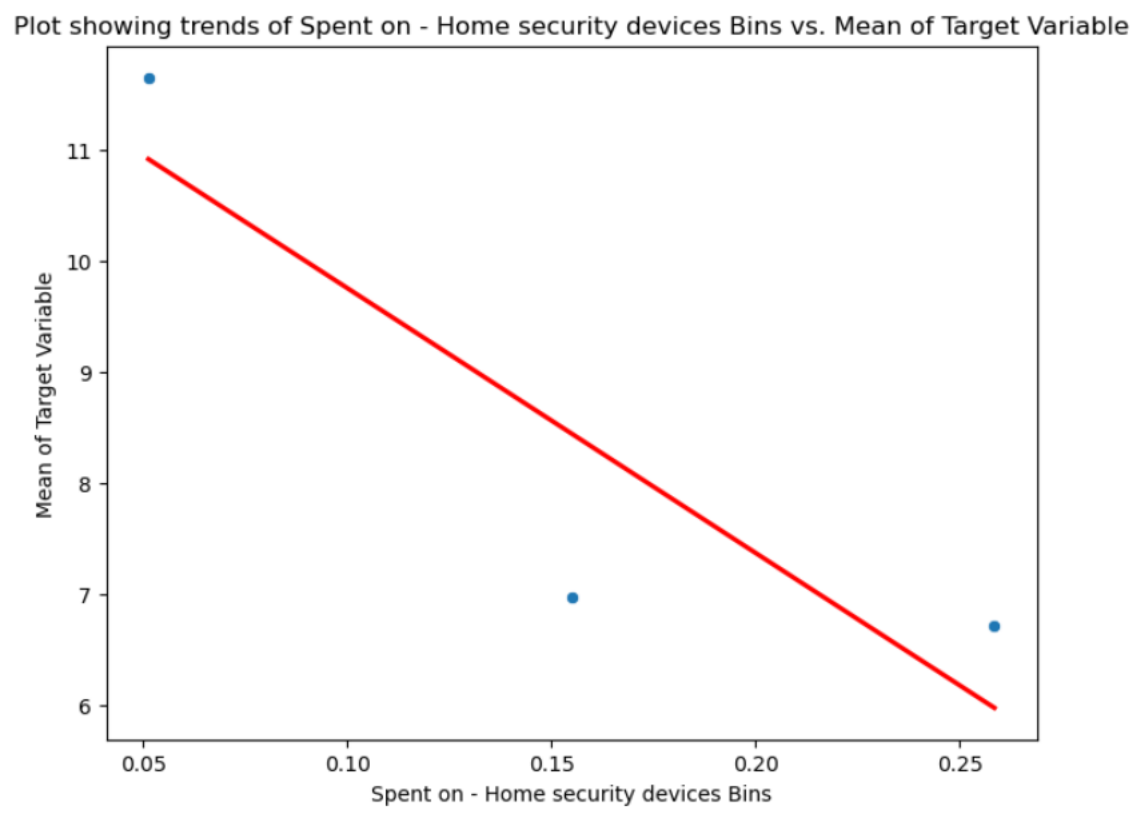
Households with Income $100,000 Or Over and Spent on Cannabis ($/month)

An intriguing observation emerges as we identify a negative correlation between 'Households with Income $100,000 Or Over' and 'Spent on Cannabis.' This finding suggests that higher-income households tend to allocate fewer financial resources toward cannabis consumption. Possible explanations include differing priorities, perceptions, or regulations surrounding cannabis expenditure within affluent households. This insight prompts us to explore the socio-economic dynamics shaping this relationship and their potential implications for consumer behavior, market strategies, and policy discussions.



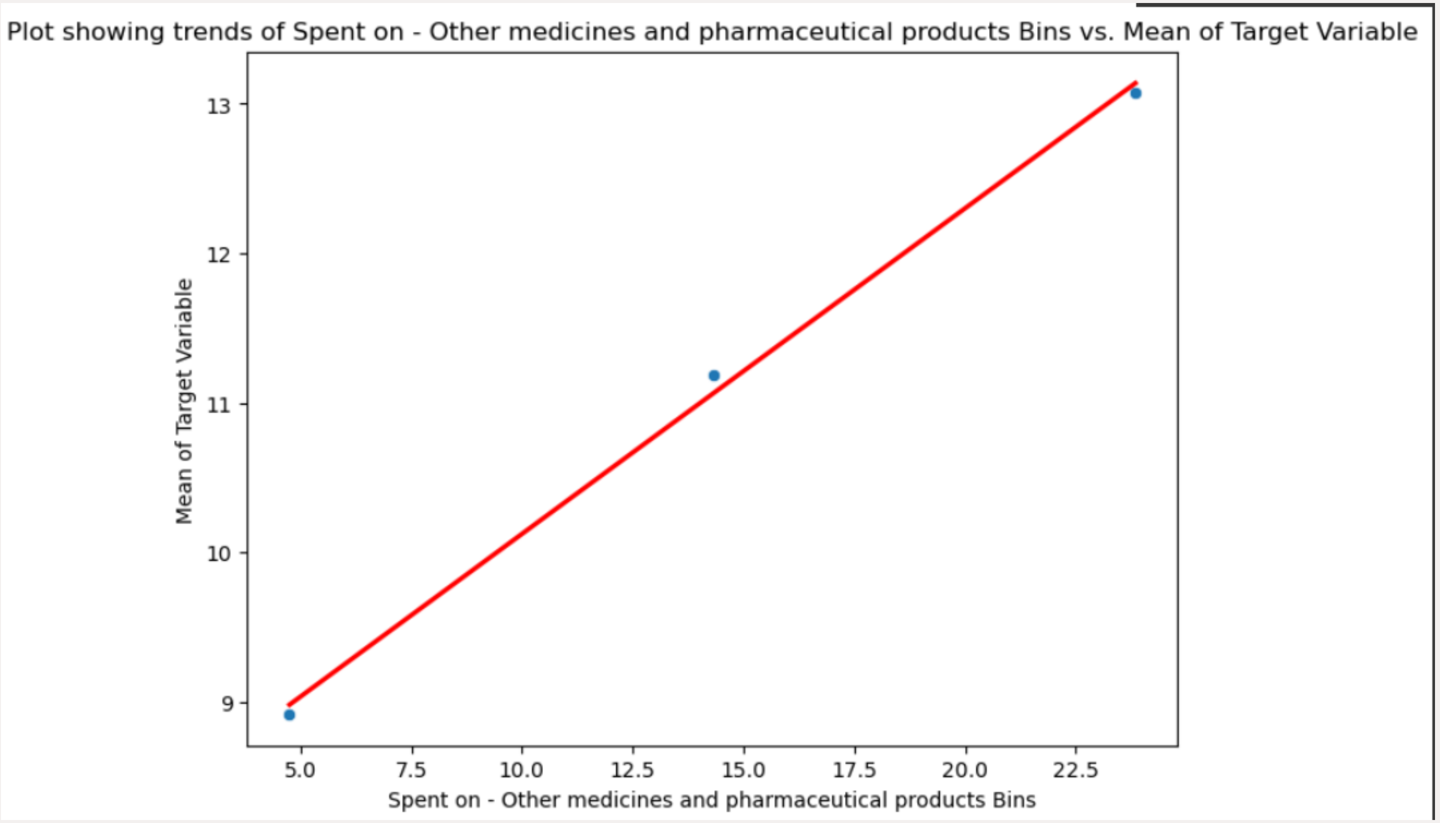
Spent on - Home security devices and Spent on Cannabis ($/month)

Our analysis reveals an intriguing negative correlation between 'Spent on - Home security devices' and 'Spent on Cannabis.' This finding suggests a potential trade-off between investments in home security and spending on cannabis-related products. Individuals who prioritize home security may allocate resources differently, potentially influencing their choices concerning cannabis expenditure. This insight prompts us to explore the underlying motivations and demographics driving this correlation, offering insights into lifestyle choices, security concerns, and potential intersections between these distinct consumer behaviors.



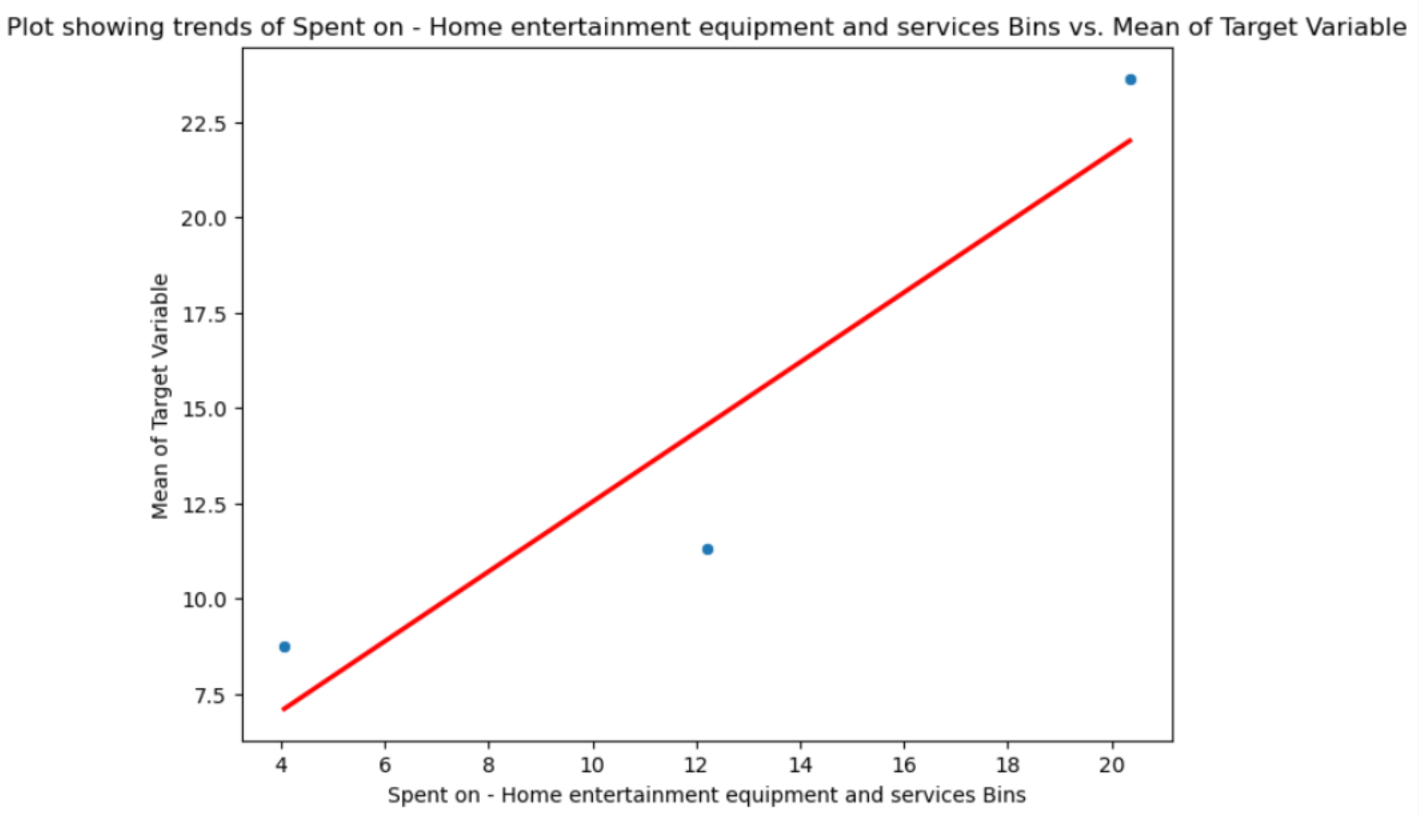
Spent on - Other medicines and pharmaceutical products and Spent on Cannabis ($/month)

This correlation suggests a potential link between health-related expenditures and cannabis consumption patterns. Individuals who allocate resources toward other medicines might also engage in spending on cannabis for various wellness-related reasons. Understanding the motivations behind this connection can provide insights into consumer behavior, potential complementary usage, and intersections between traditional and alternative health approaches.



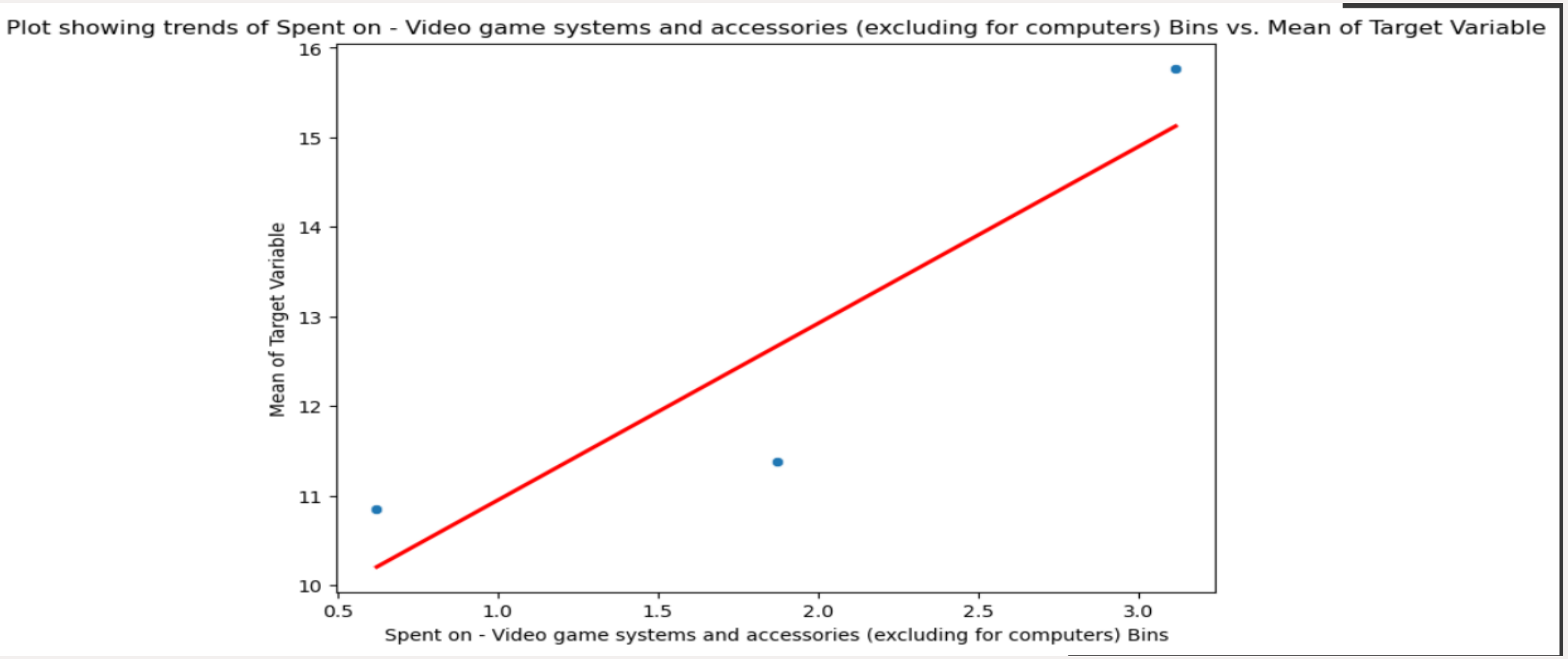
Spent on - Home entertainment equipment and services and Spent on Cannabis ($/month)

The positive correlation between spending on Home entertainment equipment and services and spending on Cannabis highlights a contemporary trend of blending leisure choices. As individuals invest in immersive home entertainment, they also exhibit an affinity for recreational experiences like cannabis, showcasing a modern fusion of relaxation and entertainment. This correlation suggests a growing preference for tailored leisure experiences. Consumers who allocate resources for advanced home entertainment are also inclined to embrace cannabis consumption, reflecting a desire to curate their leisure activities according to personal preferences.



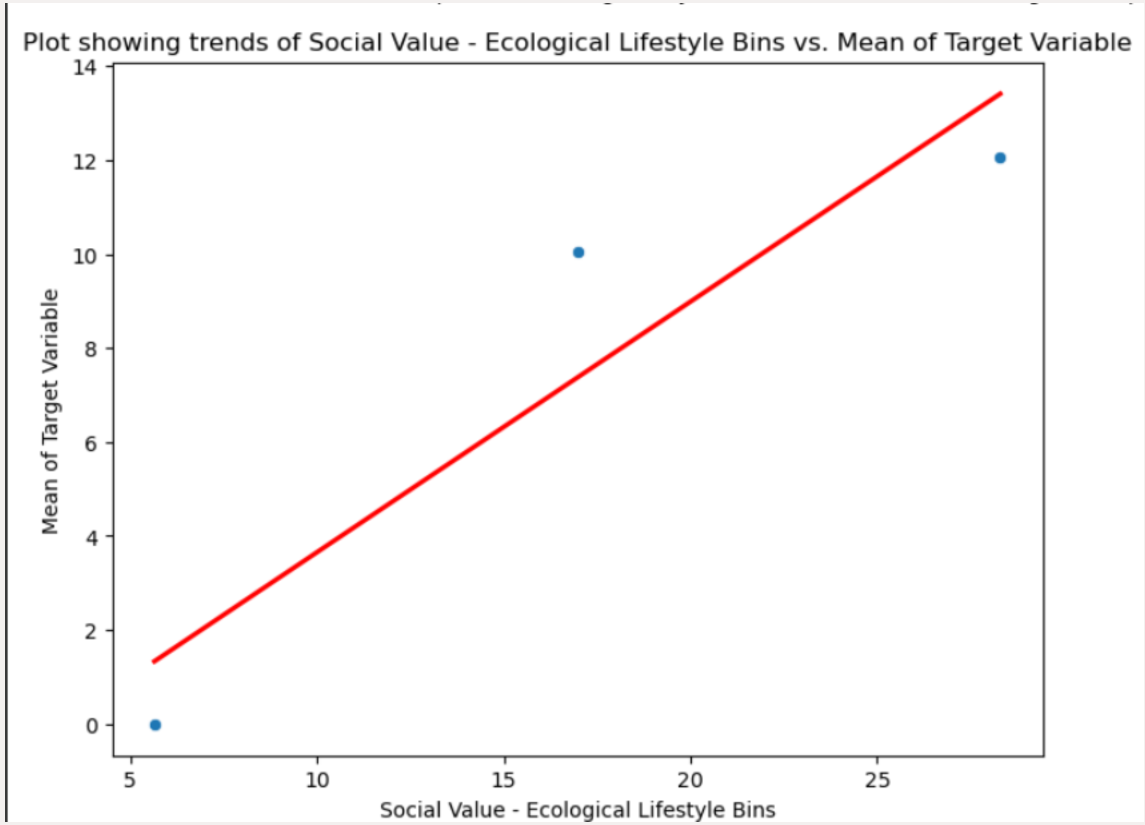
Spent on - Video game systems and accessories (excluding for computers) and Spent on Cannabis ($/month)

The positive correlation between spending on Video game systems and accessories (excluding for computers) and spending on Cannabis unveils a unique connection between contemporary recreational choices. As gamers invest in immersive gaming experiences, there's a simultaneous inclination towards exploring recreational activities like cannabis, showcasing an intriguing synergy of entertainment preferences. The correlation reflects evolving cultural norms and attitudes. With diminishing societal stigma around both gaming and cannabis, individuals are embracing these activities more openly, mirroring a broader shift towards accepting a range of leisure interests.



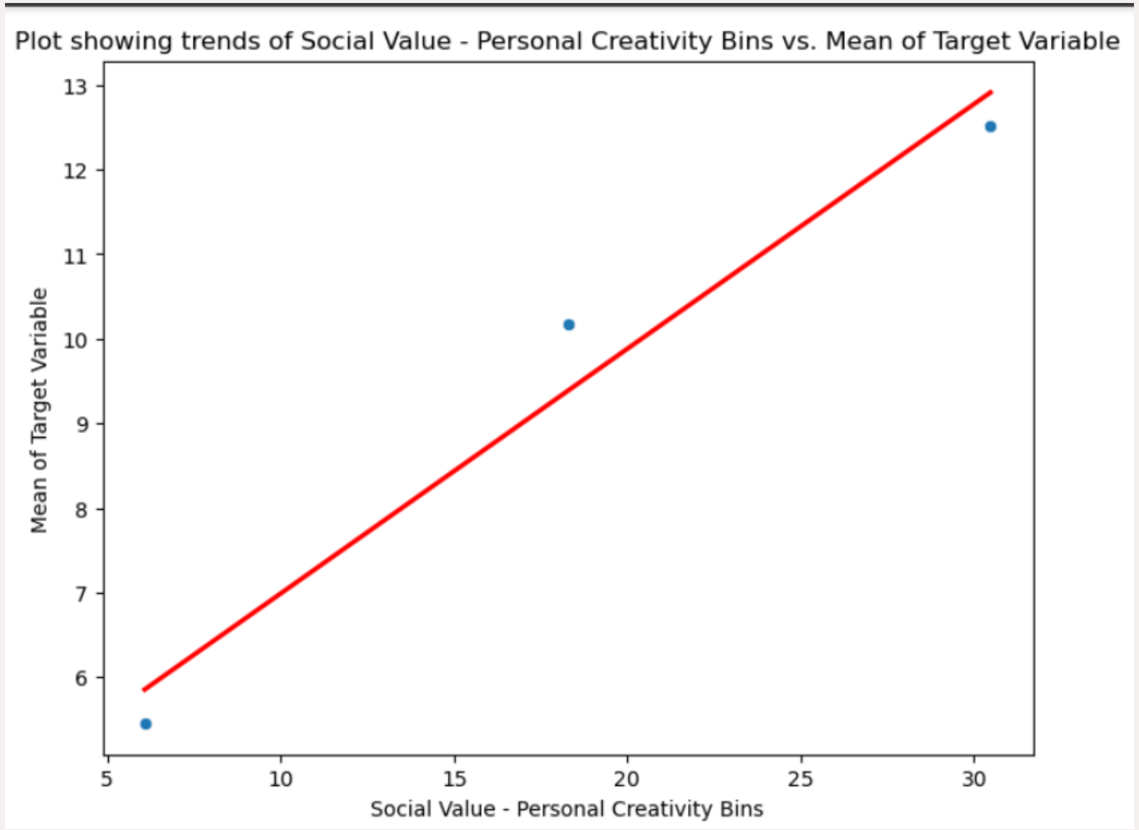
Social Value - Ecological Lifestyle and Spent on Cannabis ($/month)

The positive correlation between spending on an Ecological Lifestyle and expenditure on Cannabis sheds light on a noteworthy blend of values. Individuals embracing ecological choices may also express an affinity for sustainable recreational activities like cannabis consumption, exemplifying a fusion of eco-consciousness and leisure preferences. The correlation signifies a cultural shift towards integrated values. As societal norms evolve and environmental consciousness rises, individuals are embracing both ecological living and cannabis use more openly, reflecting a broader acceptance of diverse lifestyle elements



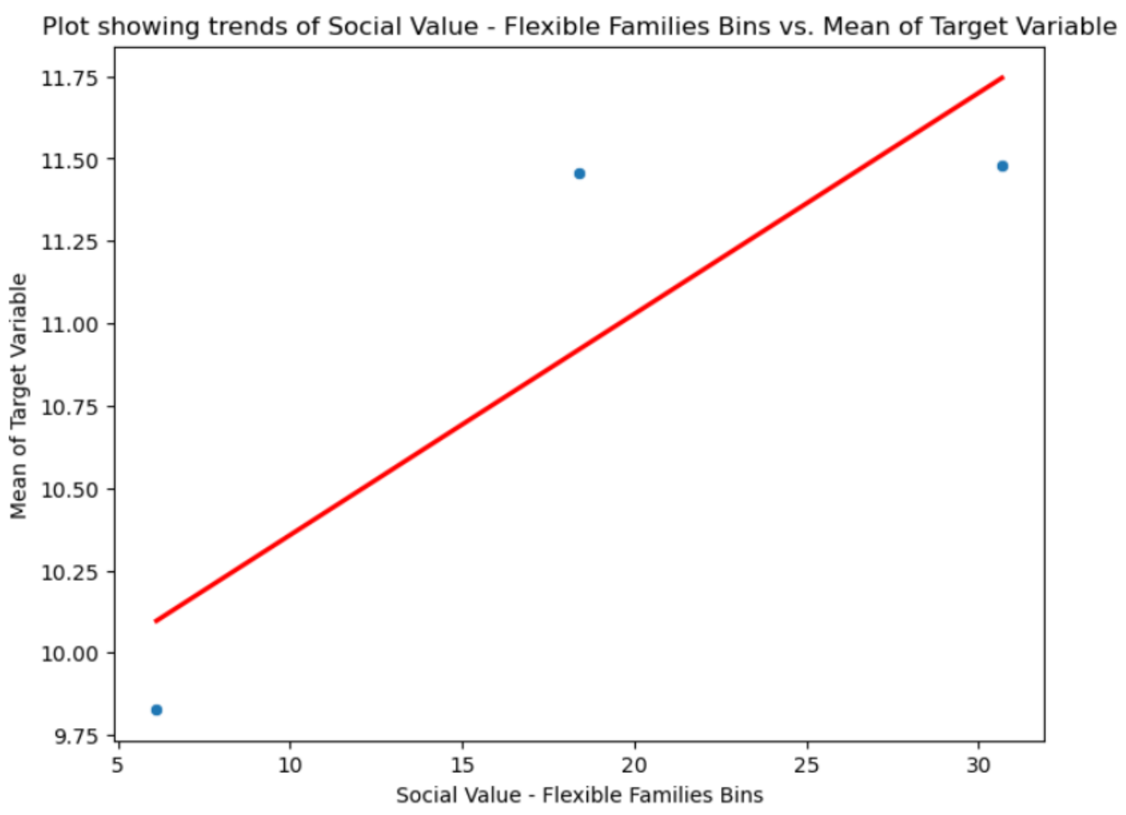
Social Value - Personal Creativity and Spent on Cannabis ($/month)

The positive correlation between spending on Personal Creativity and expenditure on Cannabis unveils an intriguing connection between self-expression and recreational choices. Individuals valuing personal creativity appear to embrace innovative leisure activities like cannabis consumption, showcasing a nexus where artistic expression and relaxation intertwine. The correlation reflects evolving cultural attitudes. As society becomes more accepting of both personal creativity and cannabis, individuals are more comfortable integrating these elements into their identities, symbolizing a broader shift towards embracing multifaceted leisure interests.



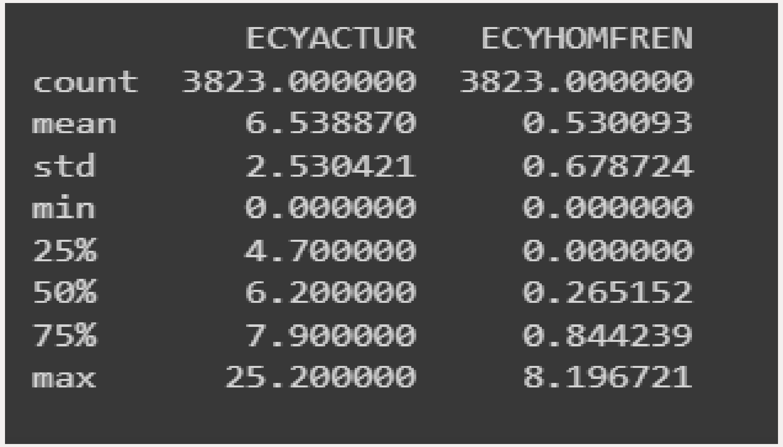
Social Value - Flexible Families and Spent on Cannabis ($/month)

The positive correlation between spending on Flexible Families and expenditure on Cannabis highlights an interesting connection between adaptable family dynamics and recreational preferences. Families valuing flexibility in their lifestyles may also embrace recreational activities like cannabis consumption, showcasing a trend of versatile leisure choices. The correlation reflects changing cultural norms. As societal acceptance of both flexible family structures and cannabis grows, families are more inclined to explore a variety of leisure pursuits, mirroring a broader acceptance of diverse lifestyle elements.



Anomalies and patterns observed

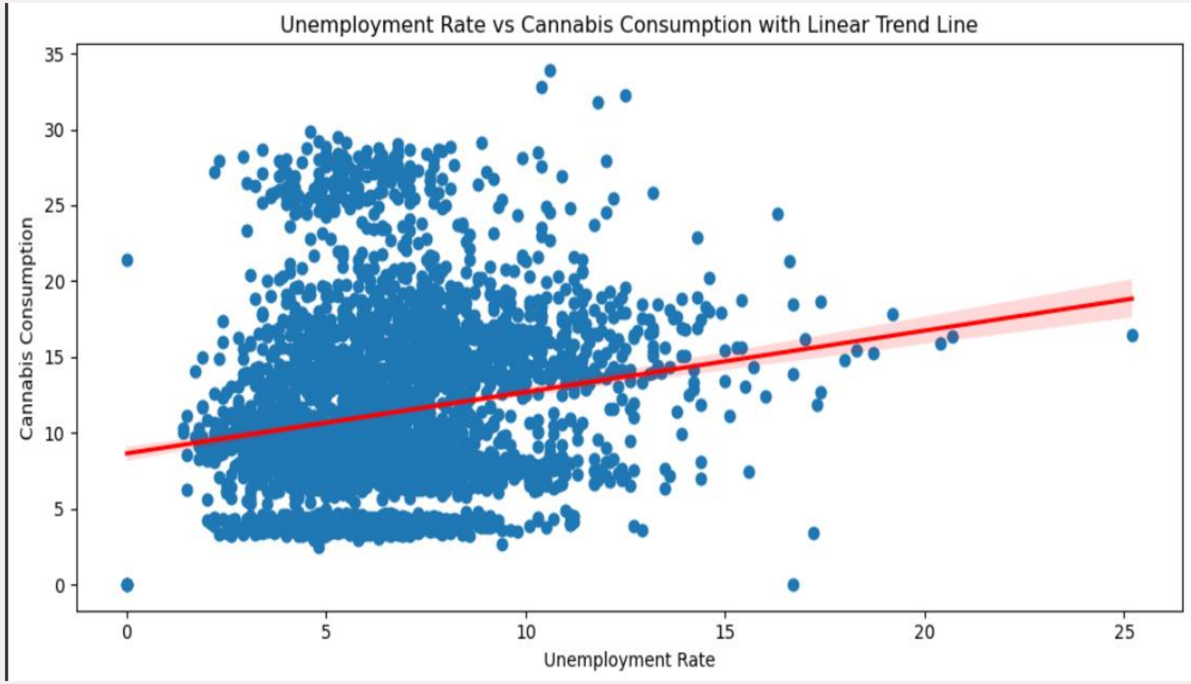
1.**Binned Means Calculation**: Calculating the mean of the target variable within each bin is a common approach. However, it's important to consider potential anomalies caused by imbalanced data distribution within bins. In cases where one bin has significantly fewer samples, its mean might be less representative.
2.**Linear Regression Lines**: Adding linear regression lines to the scatter plots can help visualize trends. However, keep in mind that the assumption of a linear relationship might not always hold true. Patterns might deviate from linearity, leading to misinterpretation if not carefully considered.
3.**Labeling of Bins**: The code doesn't explicitly mention how the bins are labeled or represented in the visualizations. Adding labels to the bins (e.g., low, medium, high) could enhance interpretation and help viewers understand the context of the plotted data.
4.**Feature Significance**: The code focuses on visualizing the correlation between features and the target variable's mean. To assess the significance of the correlations, statistical tests or further analysis might be needed, as the correlation strength could vary among features.
5.**Missing Data Handling**: The code doesn't address how missing data is handled in the analysis. Anomalies could arise if missing data significantly affects the calculated means or introduces bias in the insights drawn.
6.**Overfitting Concerns**: The code visualizes scatter plots with regression lines for each feature. While this can provide insights, it's important to avoid overfitting by adding too many complex lines that might not accurately represent the underlying patterns.
7.**Contextual Understanding**: Without context on the nature of the data or the business problem, it's challenging to assess whether the observed patterns or anomalies have practical significance. Understanding the domain and the goals of the analysis is crucial for accurate interpretation.



```
          ECYACTUR   ECYHOMFREN
count  3823.000000  3823.000000
mean      6.538870     0.530093
std       2.530421     0.678724
min       0.000000     0.000000
25%       4.700000     0.000000
50%       6.200000     0.265152
75%       7.900000     0.844239
max      25.200000     8.196721
```

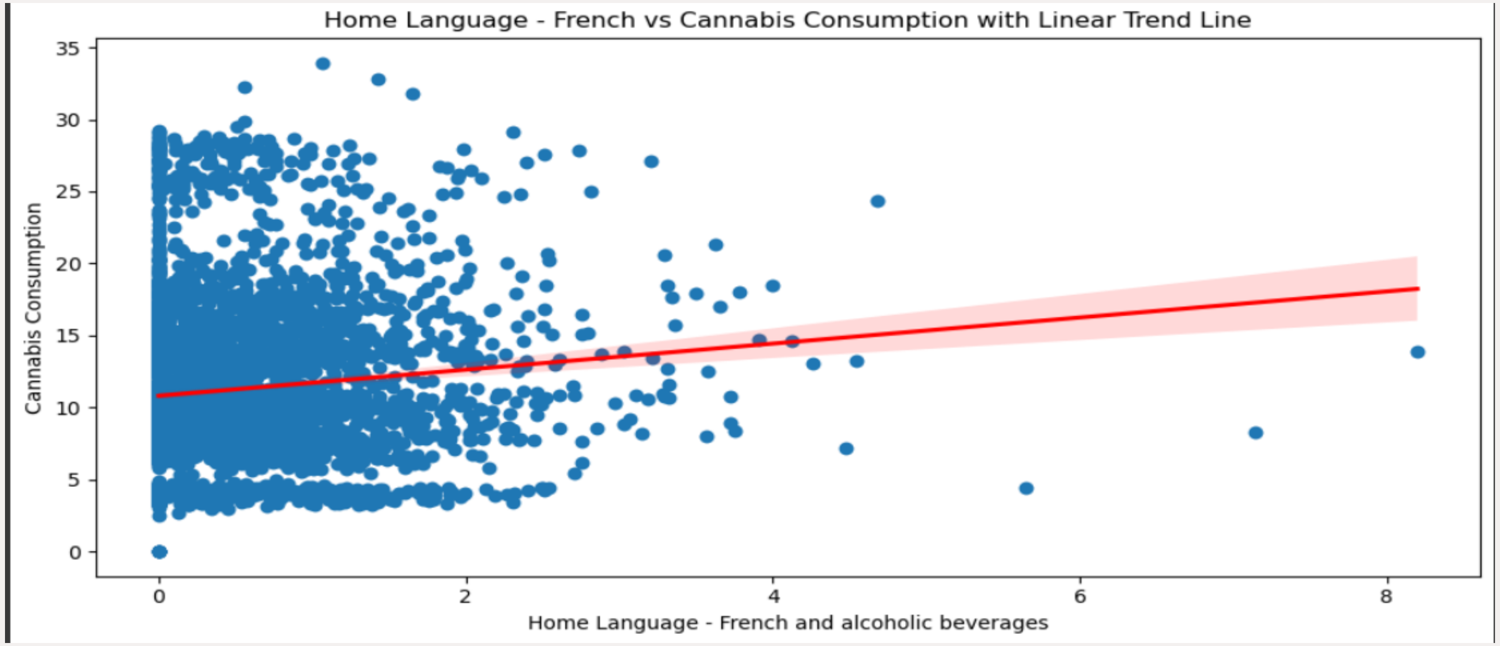
'Unemployment Rate' and 'Cannabis Consumption.' By adding a linear regression trend line to the scatter plot, the code aims to show whether there is a discernible linear trend between these two variables. The slope of the trend line (calculated as the rate of change) provides insight into the strength and direction of this relationship.

If the trend line slopes upwards from left to right, it suggests that as the 'Unemployment Rate' increases, 'Cannabis Consumption' tends to increase as well. Conversely, if the trend line slopes downwards, it indicates that as the 'Unemployment Rate' increases, 'Cannabis Consumption' tends to decrease. The slope's magnitude also provides information about the steepness of this relationship. The linear trend line helps visualize the general trend, but it's important to remember that a linear relationship might not always capture the full complexity of the data's underlying patterns.



The pattern observed from this code lies in the relationship between 'Home Language - French' and 'Cannabis Consumption.' By adding a linear regression trend line to the scatter plot, the code visually demonstrates whether there's a discernible linear trend between these two variables. The slope of the trend line provides information about the direction and strength of the relationship.

If the trend line slopes upwards from left to right, it suggests that as the proportion of people speaking French at home increases, 'Cannabis Consumption' tends to increase as well. If the slope is negative, it indicates that as the percentage of French speakers at home rises, 'Cannabis Consumption' generally decreases. The magnitude of the slope provides insight into the steepness of this linear relationship. However, keep in mind that a linear trend line might not fully capture all complexities of the data's underlying patterns.


Home Language - French vs Cannabis Consumption with Linear Trend Line
Cannabis Consumption
Home Language - French and alcoholic beverages