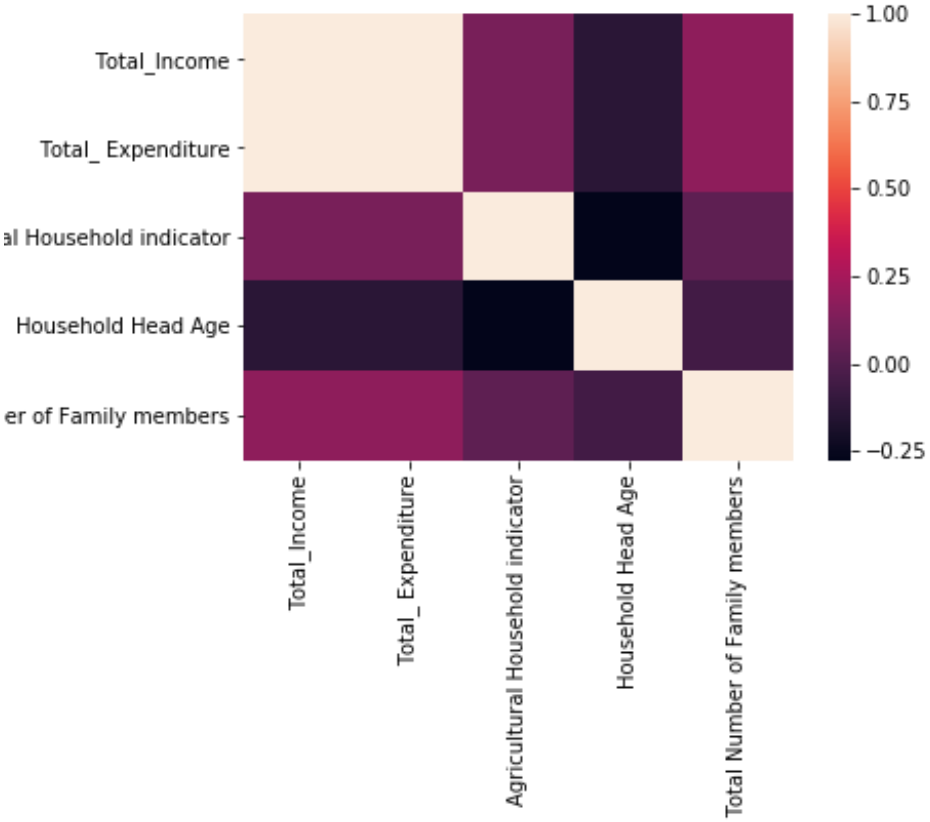
Regression Analysis Report
on
Household Income vs Expenditure

Submitted to Prof. Murat Guner, PhD
Data Analytics & Machine Learning
UMBC

Tamrat Workineh



Analysis Report

Linear regression Analysis on Income and Expenditure

Overview

This paper explores the relationship between Total Income , the independent variable, and Total Expenditure ,Dependent Variables , and predicts the future expenditure using Linear regression model from Scikit Learn package. The models' prediction scores and predict an outcome from the basis of a known variable. Results show that the model predict best with relationship between two variables.

Introduction

The dataset, household income and expenditure, for this problem is extracted from Kaggle. The dataset has 300 rows and 9 columns. Before, conducting the regression analysis model, data preprocessing has been made for concise and better findings.

Goal

Goal for this project is to be able to predict the total expenditure of different household given the Given the household total income with a linear regression model on household income and expenditure and predicts the future expenditure using Linear regression model.

Limitations: Lack of familiarity with Sklean tools to apply linear regression model.

Methods:

Based on the project objectives, Applying the Python Panda software on Jupiter notebook, incomplete and missing data were checked and only 300 rows relevant data were gathered to conduct the analysis. The following command was used to verify the missing values in the dataframe:

- *Data Preprocessing:*
    - *Missing Value:*

data.isnull(). sum() # to check missing value - it has no missing value. If there exist null values:

- *Data Exploration:*
    - *Correlation:*

It was also assessed the entire dataset to find relationship among the variables using :

```
dataCorr = data.corr
dataCorr()
```

	Total_Income	Total_ Expenditure	Agricultural Household indicator	Household Head Age	Total Number of Family members
Total_Income	1.000000	1.000000	0.117138	-0.136051	0.178088
Total_ Expenditure	1.000000	1.000000	0.117148	-0.136058	0.178102
Agricultural Household indicator	0.117138	0.117148	1.000000	-0.276090	0.034451
Household Head Age	-0.136051	-0.136058	-0.276090	1.000000	-0.055790

	Total_Income	Total_ Expenditure	Agricultural Household indicator	Household Head Age	Total Number of Family members
Total Number of Family members	0.178088	0.178102	0.034451	-0.055790	1.000000

Statistics of the dataset to see the relationship. 1.o refers to the strong positive relationship .

- *Data Visualization*:

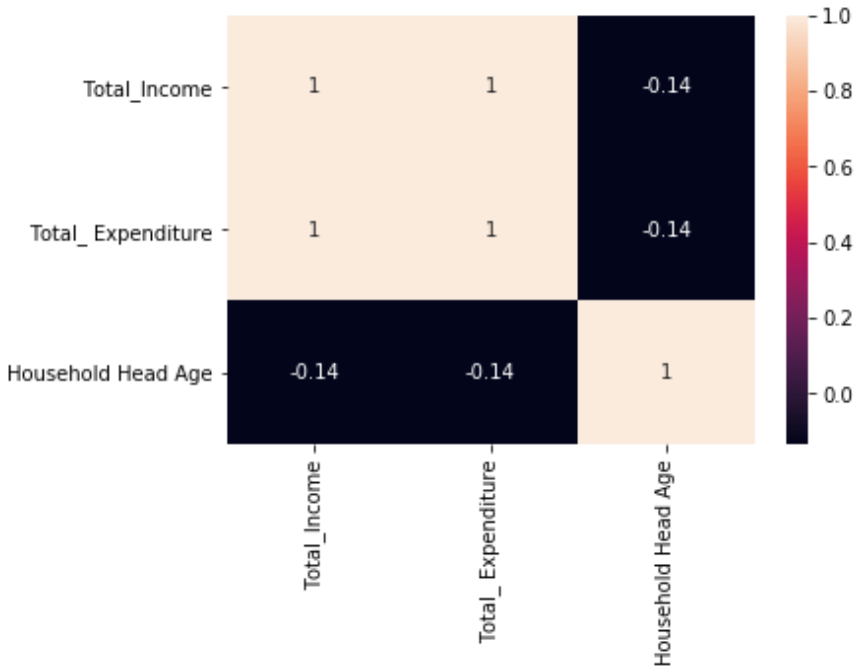
Data visualization a graphical tool to map and visualize a relationship among the Income and expenditure dataset. Based on the map observation and the correlation table above, there is a positive relationship between Total_Income Total_ Expenditure.

```
plt.figure(figsize=(12,8))
sb.heatmap(data.corr(),annot=True )
plt.show()          # From the heatmap, strong correlation the light color show strong
                    # correlation while the dark correlation show us the strong negative correlation.
```

- *Feature Selection¶*

AS it is mentioned above, the dataset contains 300 rows and 9 columns, to effectively predict the model a feature selection technique was used as a sample in which case Total_Income Total_ Expenditure.

myData = myData.corr()
display=sb.heatmap(myData,annot=True)



 # to visualize the feature and Target: the lighter color represents strong correlation

- Train-Test Split dataset:

This technique was used to take a sample data to evaluating the performance of the linear regression model. To this effect, the feature and target variables were a dataset were split into two subsets 80% to 20 % training and test respectively. Accordingly, the dataset was divided to contains X and Y trains 240 data rows each whereas X and Y test to have 60 rows each respectively .

```
X = myData['Total_Income']
y = myData['Total_ Expenditure']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=1)
```

The feature and Target visualization after Train-Test Split. The lighter color resent strong relationship

Linear regression model: The purpose of this model is to performs a regression task applying the aforementioned train data sets and finding out the relationship between variables and forecasting the future expenditure. Based on the existing data , 240 training data ,

```
from sklearn.linear_model import LinearRegression
linReg=LinearRegression()
linReg.fit(X_train.values.reshape(-1,1),y_train.values )
```

In [37]:
```
#reg=LinearRegression()
#reg.fit(X_train.values.reshape(-1,1),y_train.values )
```

In [38]:
```
reg=LinearRegression() # To create  linear Model Class Object
reg.fit(X , y)     #The fit method is used to train the model using the training set i.e. , X the feature and y the target
```

Out[38]:
```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None,
         normalize=False)
```

In [39]:
```
reg.intercept_      # this show the Beta -0
```

Out[39]:
```
array([0.02892759])
```

In [40]:
```
reg.coef_    #  •—*The coefficient represents the B1
```

Out[40]:
```
array([[0.05714274]])
```

In [41]:
```
reg.score(X,y)   # Model Evaluation
```

Out[41]:
```
0.9999999595647799
```

In [42]:
```
   #  Y=β0+β1X
NewPre = 0.02892759 * 66666 + 0.05714274
NewPre                                  # now we have a model to predict future expenditure
```

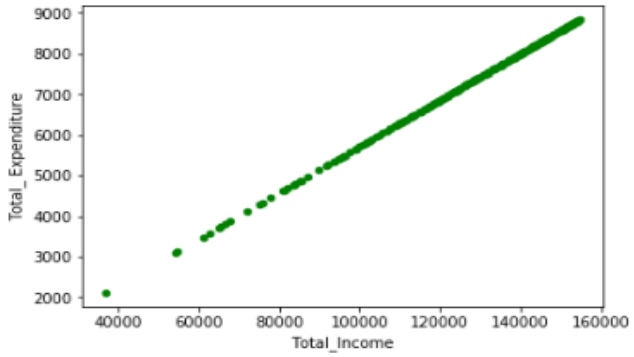
Out[42]:
```
1928.54385768
```

In [43]:
```
# p= reg.predict(X_test)

# X_test['Total_ Expenditure']=p
```

```
In [59]: X=pd.DataFrame(Traindatta['Total_Income'])
         y=pd.DataFrame(Traindatta['Total_ Expenditure'])

In [62]: Traindatta.plot(kind='scatter', x='Total_Income', y='Total_ Expenditure',color='green')
         plt.show()   # Visulizing the graph with the new prediction.
```



Result:

After splitting the data to better compute and predict the future data, a regression model was applied using the formula $Y=\beta_0+\beta_1 X$ where "Y" is the target variable, β_0 – the y- intercept β_1 – the slope and with 'X' the dependent variable , the linear regression model 0.9999999 precision as shown below. This means that the model shows that there is strong positive correlation and given a new data, it precisely predicts the future expenditure of a family.

```
reg.score(X,y)   # Model Evaluation

0.9999999595647799
```

Conclusion:

This paper explores the relationship between Total Income , the independent variable, and Total Expenditure ,Dependent Variables , and predicts the future expenditure using Linear regression model from Scikit Learn package. The models' prediction scores and predict an outcome from the basis of a known variable. Results show that the model predict best with relationship between two variables.