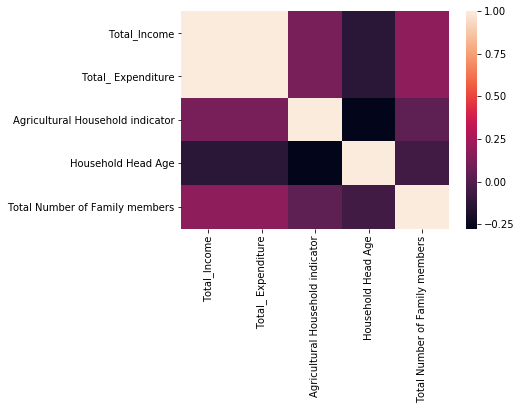
**Analysis Report on**

**Household Income vs Expenditure**

**Submitted to Dr.**

**Prof. Data Analytics & Machine Learning**

**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](https://www.izito.com/ws?q=scikit%20learn%20machine%20learning&asid=iz_us_ba_6_cg1_05&de=c&ac=4318&msclkid=7de9a194d4f91912e0aafee877d7c2e8) 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**

The purpose of this project was to conduct a linear regression model on household income and expenditure and predicts the future expenditure using Linear regression model from [Scikit Learn](https://www.izito.com/ws?q=scikit%20learn%20machine%20learning&asid=iz_us_ba_6_cg1_05&de=c&ac=4318&msclkid=7de9a194d4f91912e0aafee877d7c2e8) package.

**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 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***:**

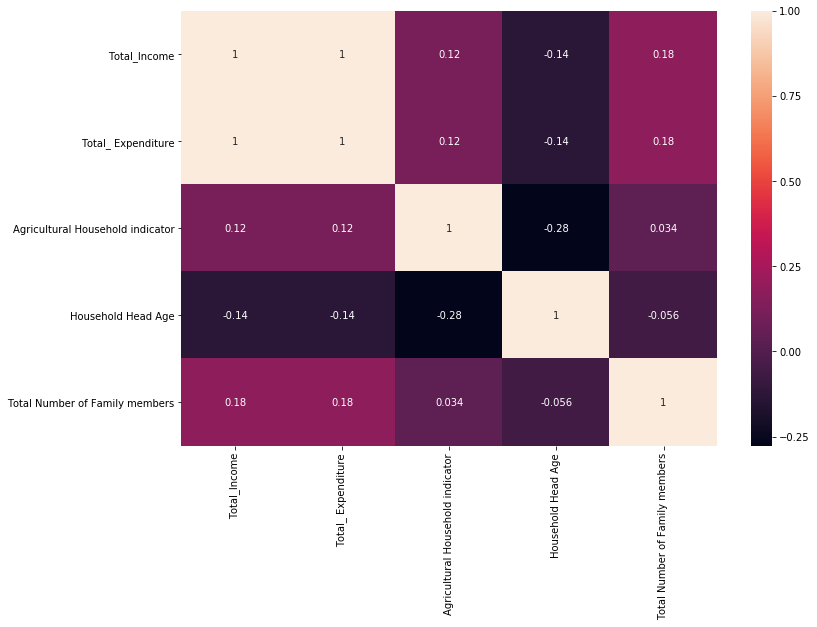
Data visualization a graphical tool to map and visualize a relationship among the Income and expenditure dataset. Based on the map observation and the correlation table above, there is a positive relationship between Total\_Income Total\_ Expenditure.

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.



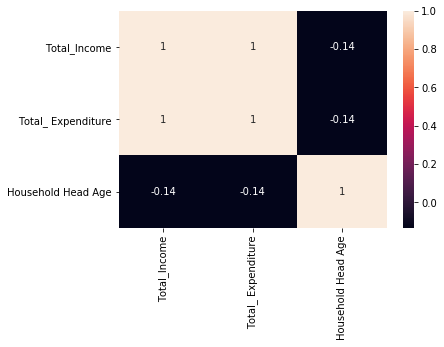
#### 

#### Feature Selection[¶](http://localhost:8889/notebooks/Desktop/Data%20602/Untitled-Copy1.ipynb#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**

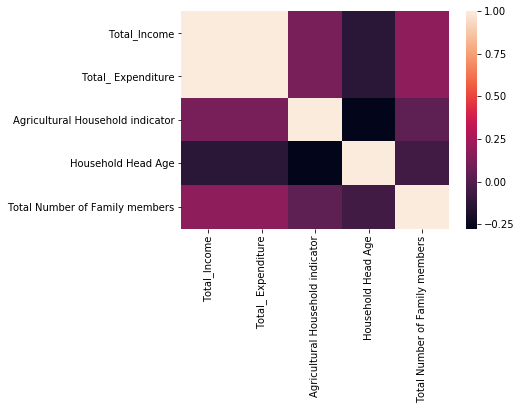
* + Train-Test Split dataset:

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)



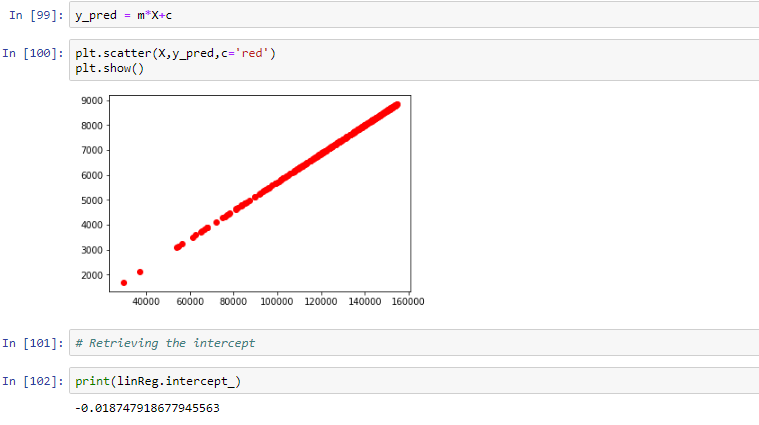
The feature and Target visualization after Train-Test Split. The lighter color resent strong relationship

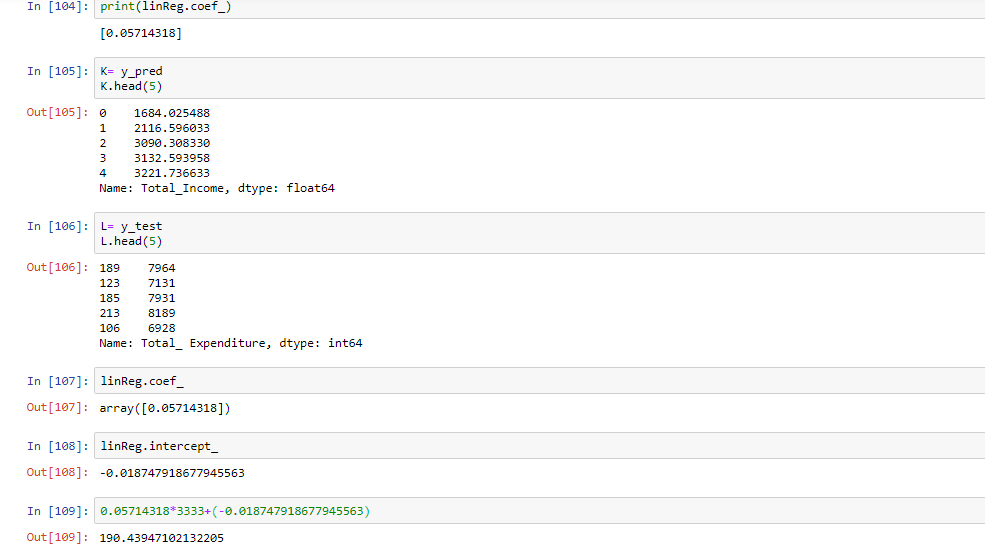
Linear regression model : The purpose of this model is to performs a regression task applying the aforementioned train data sets and finding out the relationship between variables and  forecasting the future expenditure.

from sklearn.linear\_model import LinearRegression

linReg=LinearRegression()

linReg.fit(X\_train.values.reshape(-1,1),y\_train.values )





**Result**:

The models’ has perfectly predicted future expenditure from the basis of a known variable. Results show that the model predict best with relationship between two variables.

**Conclution**:

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](https://www.izito.com/ws?q=scikit%20learn%20machine%20learning&asid=iz_us_ba_6_cg1_05&de=c&ac=4318&msclkid=7de9a194d4f91912e0aafee877d7c2e8) 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.