Exploratory Data Analysis and Multiple Regression Modeling of Bond Fund Characteristics

D-Bonds Fund Study

Submitted to:

Professor Serhat Simsek

INFO 589

Applied Statistics for Business Analytic

Report prepared by:

Team 14

Jose Peralta

Krupa Gor

Vanessa Hartkopf

Raquel Rivera

04/28/2024

★TITLE PAGE★

Overview:

The dataset comprises characteristics of 180 bond funds, including their five-year return, assets, expense ratio, one-year return, three-year return, category, and risk.

Nature of the Data:

The dataset contains both numerical and categorical variables. Numerical variables include five-year return, assets, expense ratio, one-year return, and three-year return. Categorical variables include category and risk.

Explanation:

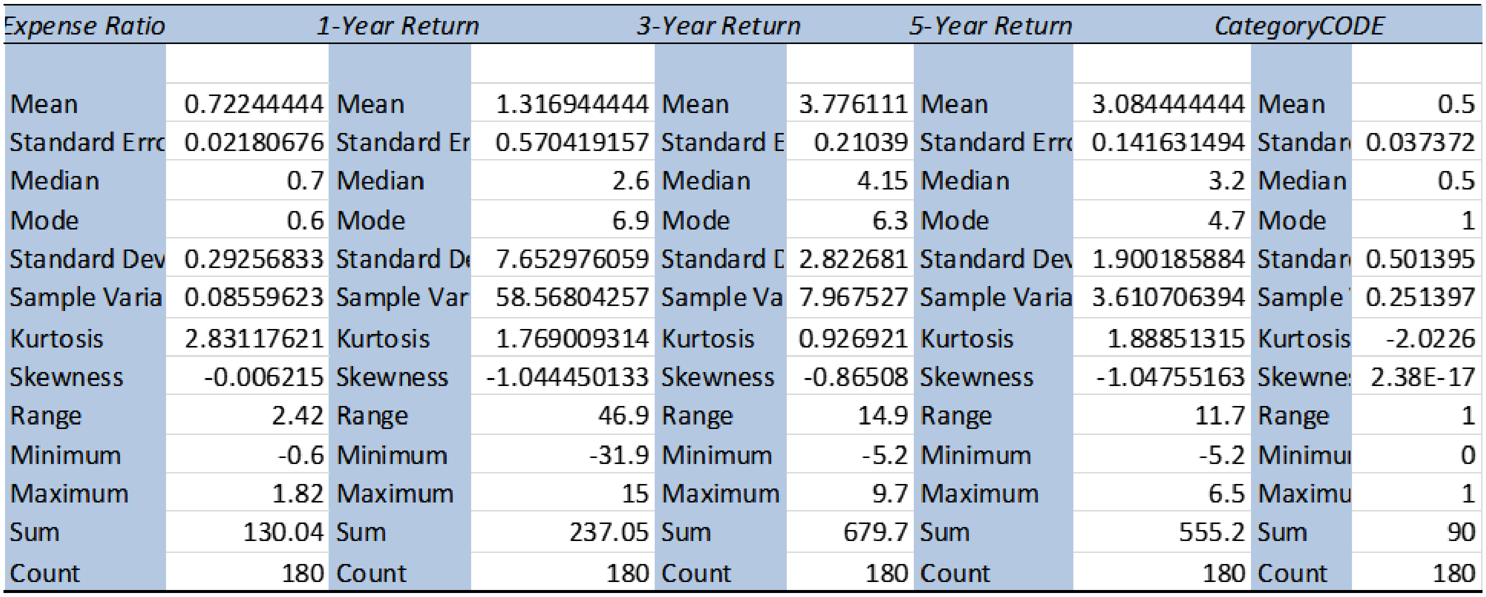
In this comprehensive analysis, the dataset consists of 180 bond funds with various characteristics including assets, expense ratio, one-year return, three-year return, five-year return, category code, and category label. The nature of the data is predominantly numerical, with variables such as assets, expense ratio, and return percentages. Additionally, there are categorical variables like the category code and label, which indicate the type of bond fund. These attributes provide a rich landscape for exploring relationships and patterns within the dataset.

The exploratory data analysis (EDA) begins with a detailed overview of the dataset, summarizing its dimensions and variable types. Summary statistics are calculated and presented for each numerical variable, offering insights into central tendency, variability, and distribution characteristics. Visualizations such as histograms, box plots, and density plots are employed to further explore the distribution of numerical variables and identify potential outliers. Hypothesis testing is conducted on specific variables of interest, including t-tests for comparing means and chi-square tests for independence, to uncover significant associations within the data.

Moving forward, multiple regression modeling is employed to develop predictive models for the response variable, which in this case could be the five-year return. The modeling process involves fitting preliminary regression models, assessing overall model significance, conducting variable selection through backward elimination, and evaluating model fit using R-squared and adjusted R-squared values. Residual analysis is performed to validate model assumptions, including normality, independence, and constant variance. Interpretations are made regarding the regression coefficients, coefficient of determination (R-squared), and standard error of the estimate. Additionally, prediction intervals are determined to gauge the uncertainty around predicted values. Through this thorough analysis, valuable insights are gained into the relationships between bond fund characteristics and their returns, facilitating informed decision-making in investment strategies.

★ EXPLORATORY DATA ANALYSIS ★

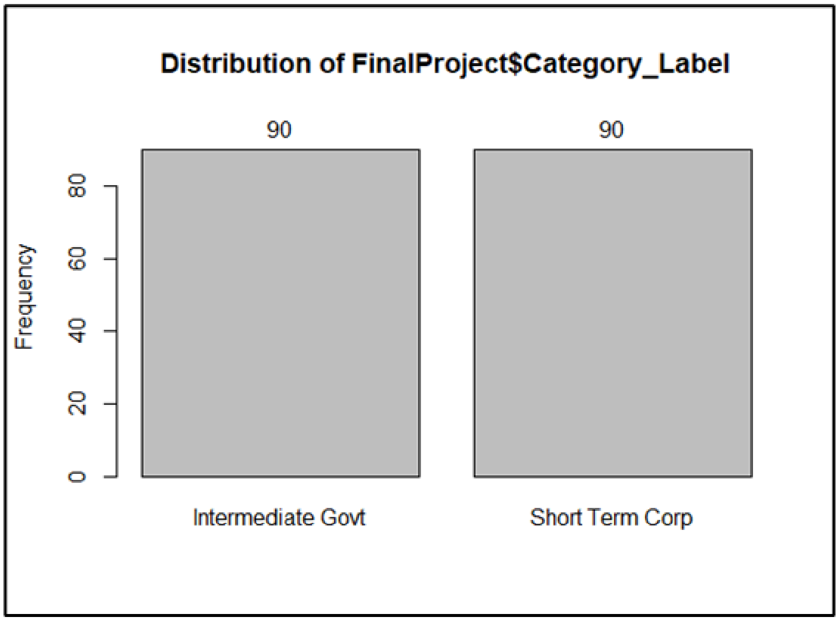
Descriptive Statistics for Continuous Variable:



Expense Ratio		1-Year Return		3-Year Return		5-Year Return		CategoryCODE	
Mean	0.72244444	Mean	1.316944444	Mean	3.776111	Mean	3.084444444	Mean	0.5
Standard Errc	0.02180676	Standard Er	0.570419157	Standard E	0.21039	Standard Errc	0.141631494	Standar	0.037372
Median	0.7	Median	2.6	Median	4.15	Median	3.2	Median	0.5
Mode	0.6	Mode	6.9	Mode	6.3	Mode	4.7	Mode	1
Standard Dev	0.29256833	Standard De	7.652976059	Standard D	2.822681	Standard Dev	1.900185884	Standar	0.501395
Sample Varia	0.08559623	Sample Var	58.56804257	Sample Va	7.967527	Sample Varia	3.610706394	Sample	0.251397
Kurtosis	2.83117621	Kurtosis	1.769009314	Kurtosis	0.926921	Kurtosis	1.88851315	Kurtosis	-2.0226
Skewness	-0.006215	Skewness	-1.044450133	Skewness	-0.86508	Skewness	-1.04755163	Skewne	2.38E-17
Range	2.42	Range	46.9	Range	14.9	Range	11.7	Range	1
Minimum	-0.6	Minimum	-31.9	Minimum	-5.2	Minimum	-5.2	Minimu	0
Maximum	1.82	Maximum	15	Maximum	9.7	Maximum	6.5	Maximu	1
Sum	130.04	Sum	237.05	Sum	679.7	Sum	555.2	Sum	90
Count	180	Count	180	Count	180	Count	180	Count	180

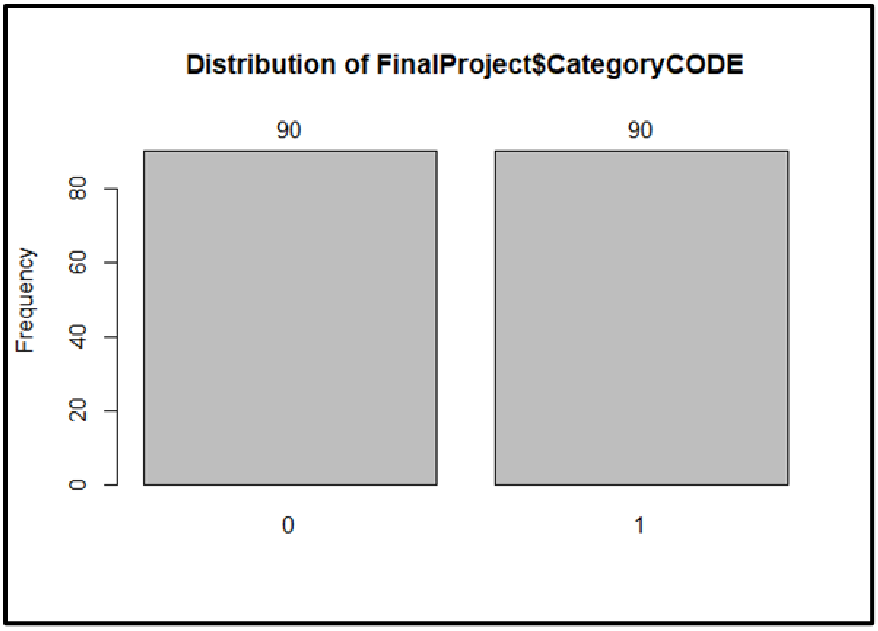
Frequency Table for Categorical Variables:

- **Category_Label Frequency Table and Plot:**



Category_Lable Frequency Table			
	Frequency	Percentages	Cum. Percentages
Intermediate Govt.	90	50	50
Short Term Corp	90	50	100
Total	180	100	100

- **Category code Frequency Table & Plot:**



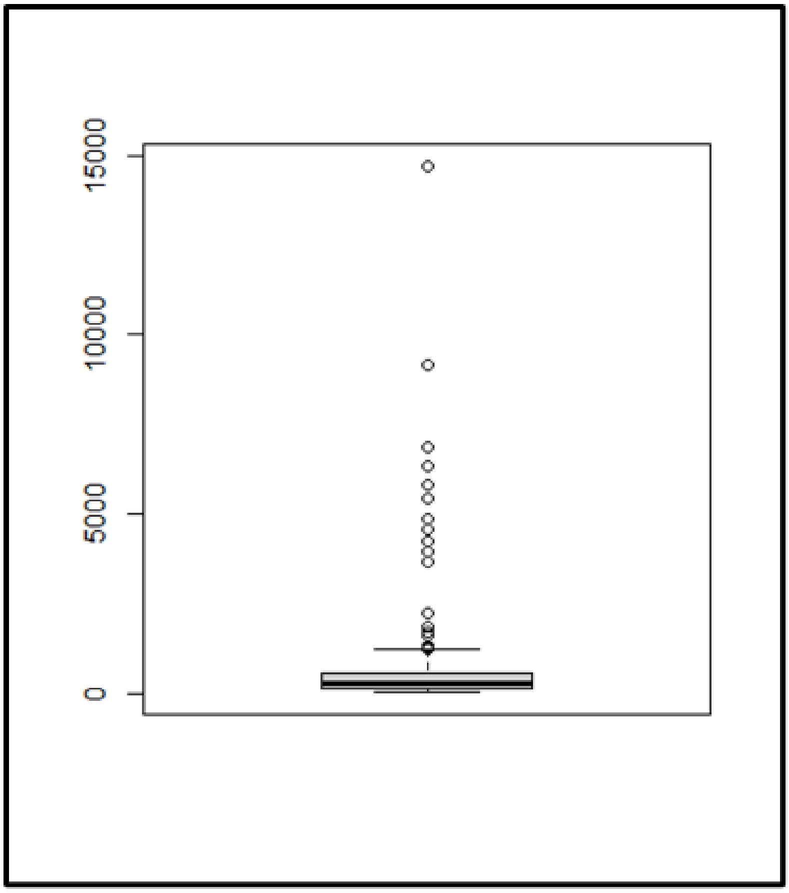
Category Frequency Code Table

	Frequency	Percentages	Cum. Percentages
0	90	50	50
1	90	50	100
Total	180	100	100

Visualization for Numeric Variables:

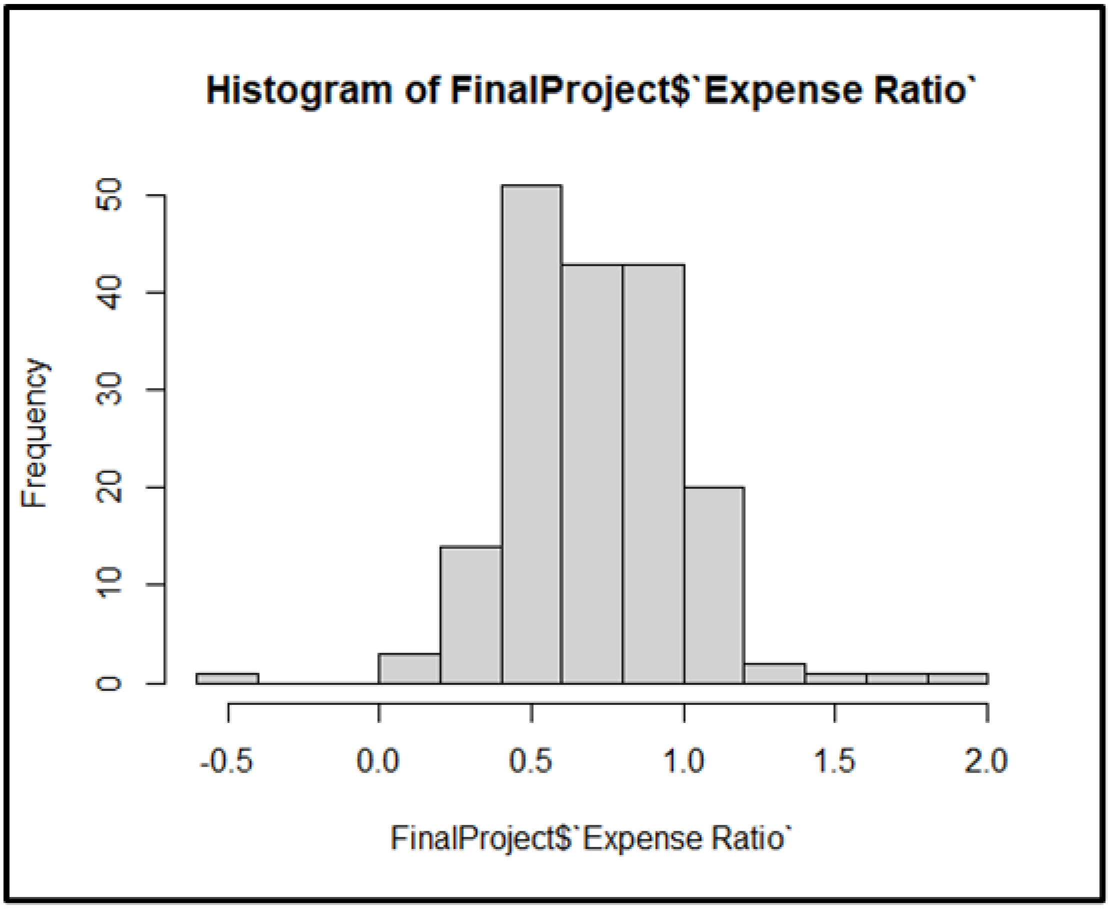
➔ Box Plots:

◆ Box Plot of Assets:



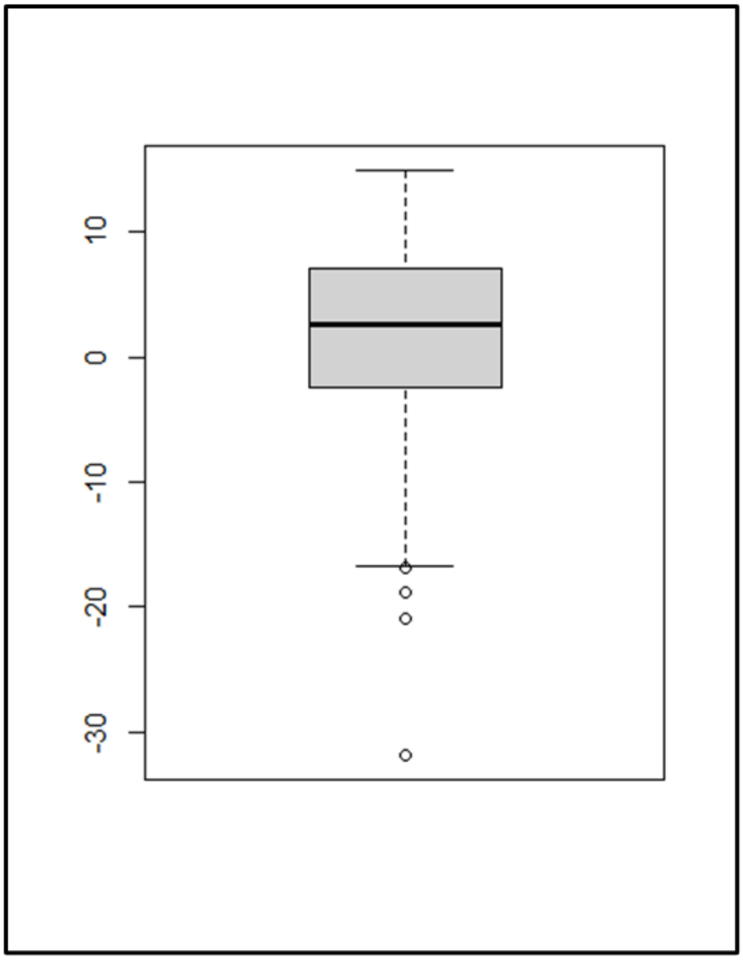
★ The box plot explains the distribution of the expense and the returns. This will be the central tendency of the median is 242.7 and the range is 14712.5. The skewness towards the left is 5.216, which is negative. The central tendency mean is 728.60. The range will be the difference between the Minimum is 12 and the maximum shows to be 1474.5.

- Box Plot of Expense Ratio:



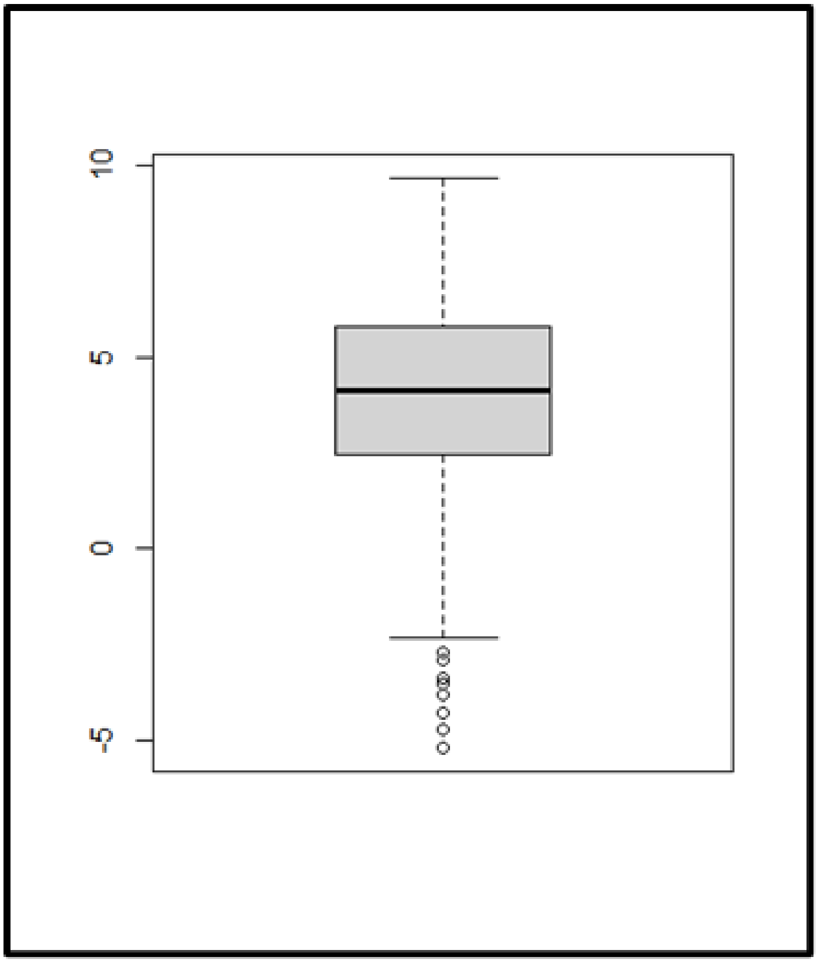
★ For the expense Ratio we are providing a Box Plot for a better overview of the data. The central tendency is the median value of 0.07. The data Range is 2.42. The central tendency mean is 0.722444. The range will be the difference between the Minimum is -0.6 and the maximum is 1.82 This shows that we have a skewness of -0.00621.

➢ Box Plot of 1-Year return:



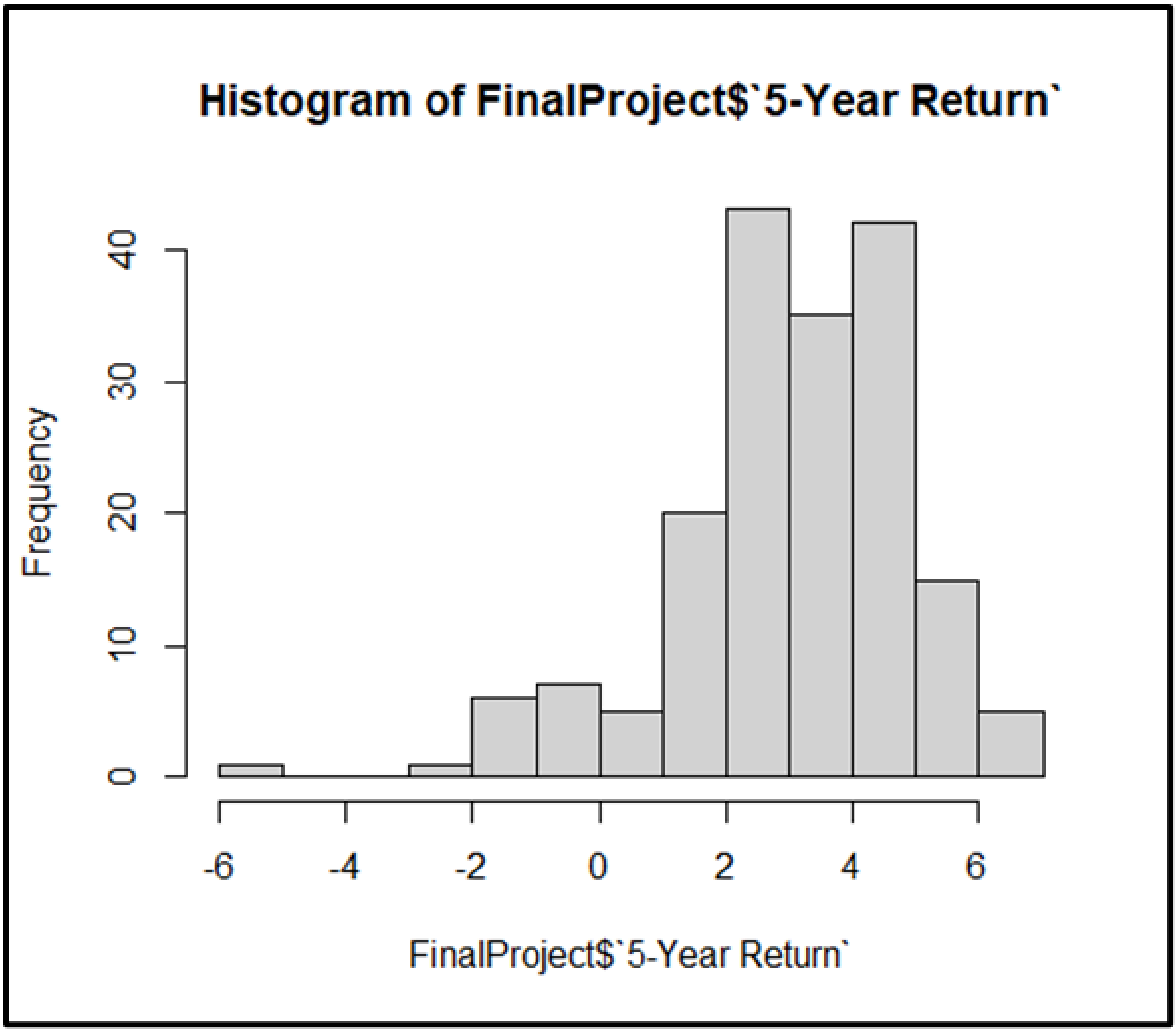
★ For the 1 year return we find the median is 2.6. The interquartile range is 46.9. T The minimum is -31.9 and maximum is 15. The central tendency mean is 1.316944. This shows the skewness would be -0.1.04445.

- Box Plot of 3-Year return:



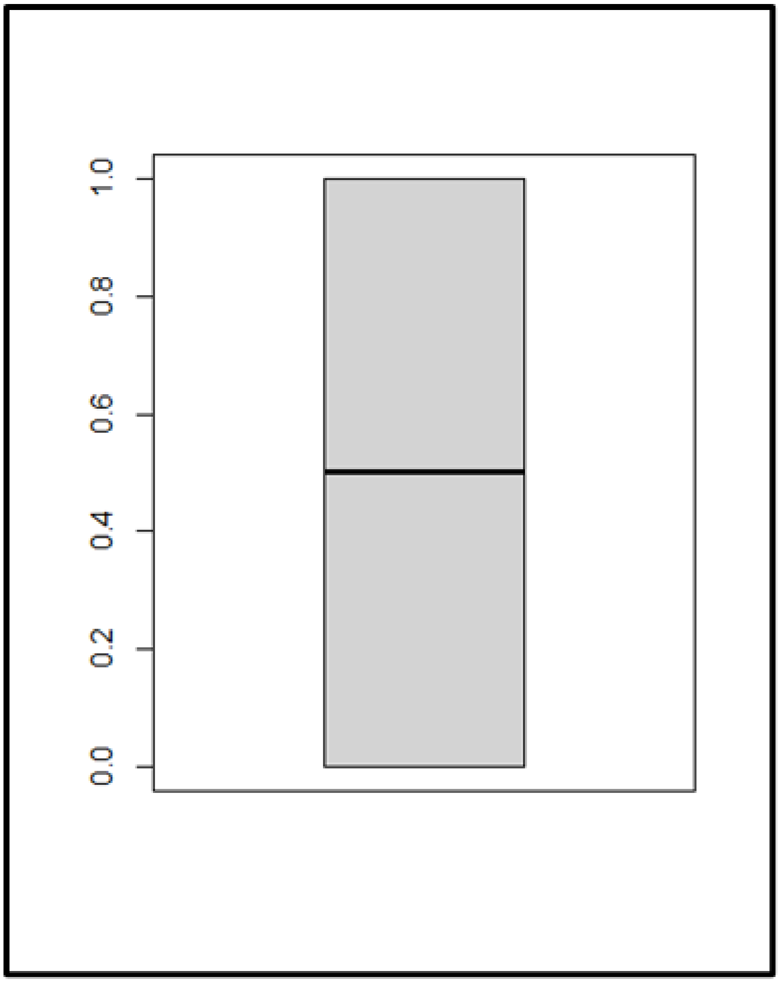
★ The median for a 3 year return is 4.15. The interquartile range is 14.9. In this plot the minimum is -5.2 and the maximum is 9.7. The central tendency mean comes out to 3.776. This shows a skewness of -0.86508.

➢ Box Plot of 5-Year Return:



★ The median for a 5 year return is 3.2. The interquartile range is 11.7. In this plot the minimum is -5.2 and the maximum is 6.5. The central tendency mean is 3.084444. This graph shows a skewness of -1.04755163.

- ★ Box Plot of Category_Code



- ★ The median to category code is 0.5. We found the range to be 1. Looking at the minimum it comes out to 0. We see the box plot above has a maximum of 1. The central tendency mean is 0.5. This graph shows we have a skewness of 2.38E-17.

★ Visualization for Numeric Variables:

- ❖ Visualizations including histograms and box plots were created to explore the distribution of numerical variables. The distributions appear to be skewed in some cases, indicating potential outliers.

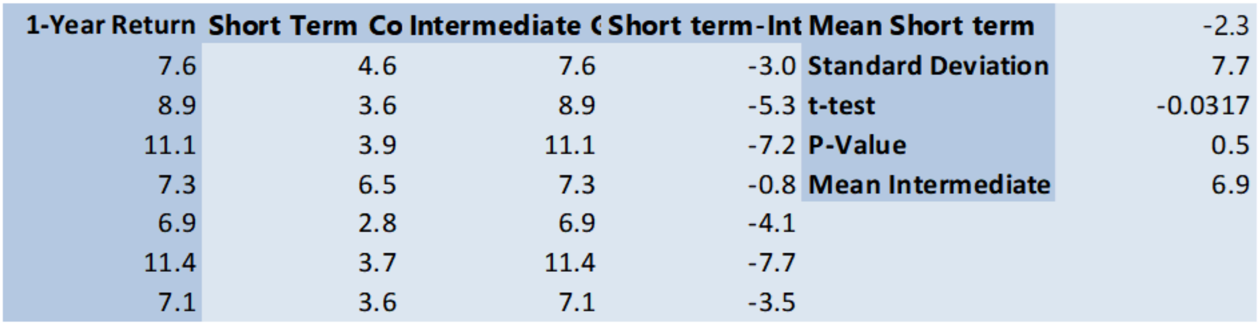
- ❖ Box plots were utilized to identify potential outliers in the dataset. Outliers were observed in several numerical variables, including assets and one-year return.

★ Hypothesis Testing:

- ★ When Conducting a hypothesis test, we define a Null Hypothesis denoted as H_0 and Alternative Hypothesis denoted as H_A. We conduct a hypothesis Test to determine whether or not sample evidence contradicts H_0.
- ★ Our goal is to determine if the Null Hypothesis can be Rejected in favor of the Alternative Hypothesis.
- ★ If the sample evidence is inconsistent with the Null hypothesis, we reject the Null Hypothesis. If sample evidence is not inconsistent with the Null Hypothesis, then we do not reject the Null Hypothesis. Most importantly, we will never conclude that “We accept the Null Hypothesis” because while the sample information may not be inconsistent with the Null Hypothesis, it does not necessarily prove that the Null Hypothesis is true.

★ Interest of Variable for Hypothesis Testing:

❖ Two-Tailed Hypothesis Test.



1-Year Return	Short Term Co	Intermediate (	Short term-Int	Mean Short term	-2.3
7.6	4.6	7.6	-3.0	Standard Deviation	7.7
8.9	3.6	8.9	-5.3	t-test	-0.0317
11.1	3.9	11.1	-7.2	P-Value	0.5
7.3	6.5	7.3	-0.8	Mean Intermediate	6.9
6.9	2.8	6.9	-4.1		
11.4	3.7	11.4	-7.7		
7.1	3.6	7.1	-3.5		

Ho: $\mu_{IG} = \mu_{ST}$

Ha: $\mu_{IG} \neq \mu_{ST}$

The Null hypothesis is one year Intermediate Government is equal to one year Short term return. While the alternative hypothesis is one year return is not equal to a one year short term.

The P- value is greater than the level of significance for 0.05 .We fail to reject the null hypothesis for one year return intermediate government is equal to one year short term return. We do not have enough evidence to conclude.

6a. Null hypothesis and Alternative Hypothesis:

Ho: Short Term Corp and Intermediate Govt independent variable of each other

Ha: Short Term Corp and Intermediate Govt dependent variable of each other

6b. Descriptive statistics for Intermediate Govt and Short Term

We will break down Intermediate Govt. and Short Term Govt. ‘s 1-Year Return in three levels with conditions being applied of **5 or more** sample sizes.

- **Low Performing 1-Year Return:** Values include from -31 to -1
- **Average Performing 1-Year Return:** Values include from -1 to 3
- **High Performing 1-Year Return:** Values include from 3 to 15

	Low Performance	**Average Performance**	**High Performance**	**Row Total**
Intermediate Govt.	12	5	73	**90**
Short-Term Govt.	51	26	13	**90**
Column Total	**63**	**31**	**86**	**180**

- **Observed Observation:**

Observed Observation

	Low Performance	Average Performance	High Performance	Row Total
Intermediate Govt.	0.175	0.086	0.239	90.00
Short-Term Govt.	1.175	0.086	0.239	90.00
Column Total	63.00	31.00	86.00	180.00

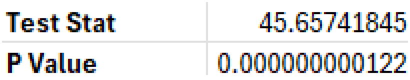
- **Expected Observation:**

Expected Observation

	Low Performance	Average Performance	High Performance	Row Total
Intermediate Govt.	31.50	15.50	43.00	90.00
Short-Term Govt.	31.50	15.50	43.00	90.00
Column Total	63.00	31.00	86.00	180.00

❖ Chi-Square test statistics and p-value

❖



Test Stat	45.65741845
P Value	0.000000000122

6c. Since P-value is < less than 0.05 significance value, we REJECT the null. Since we reject the null, we conclude that two categories are dependent on one another.

1. ANOVA Test

Null Hypothesis (H0): There is no significant difference in the mean of 1-year returns across different expense ratio categories.

Alternative Hypothesis (HA): There is a significant difference in the mean of 1-year returns across different expense ratio categories.

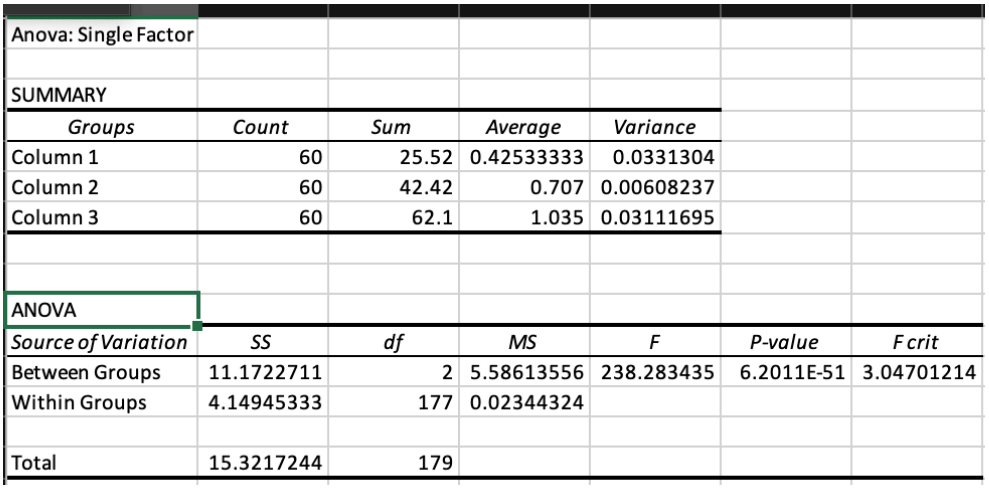
The way we segregated the three levels of expense ratio by using

Low- The low fee is -.060 to .59

Median - The median is .59 to .85

High- The high is **.**85 to 1.82

The ANOVA test will help determine whether the variations in 1-year returns can be attributed to differences in expense ratio categories or if the variations are simply due to random chance.



Anova: Single Factor						
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
Column 1	60	25.52	0.42533333	0.0331304		
Column 2	60	42.42	0.707	0.00608237		
Column 3	60	62.1	1.035	0.03111695		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	11.1722711	2	5.58613556	238.283435	6.2011E-51	3.04701214
Within Groups	4.14945333	177	0.02344324			
Total	15.3217244	179				

F-Statistic: 238.283

P-value: 6.2011E-51

Given the p-value of 6.2011E-51, which is much smaller than the typical significance level of 0.05, we reject the null hypothesis. This means that we do have enough evidence to conclude that there is a non- significant difference in the means of 1-year returns across different levels of the categorical variable (expense ratio categories).

2. Summary

The exploratory data analysis (EDA) conducted on the bond fund dataset offers valuable insights into its characteristics and relationships. Here are the key findings and observations:

1. **Descriptive Statistics:**

- Assets: The mean assets under management for bond funds is $728.6 million, with a wide range from $12 million to $14,724.5 million.
- Expense Ratio: The average expense ratio is 0.72%, with a standard deviation of 0.2925, indicating some variability among funds.
- Returns: The average one-year, three-year, and five-year returns are 1.3%, 3.8%, and 3.1% respectively. Returns vary widely, with some funds showing negative returns.
- Category: The dataset consists of two categories - Intermediate Govt. and Short Term Corp, each with 50% frequency.

2. **Visualization:**

- Histograms, box plots, and density plots were used to visualize the distribution of numerical variables. Skewed distributions and potential outliers were observed, especially in assets and one-year return.

3. **Hypothesis Testing:**

- A two-tailed hypothesis test was conducted to compare the one-year returns between Intermediate Govt. and Short Term Corp funds. The test did not provide enough evidence to reject the null hypothesis, suggesting no significant difference in one-year returns between the two categories.

- Chi-square test indicated a significant dependency between category labels (Intermediate Govt. and Short Term Corp), implying that the choice of category may affect fund performance.
- ANOVA test showed no significant difference in mean one-year returns across different expense ratio categories, suggesting that variations in returns may not be attributed to expense ratios.

4. **Insights and Further Analysis:**

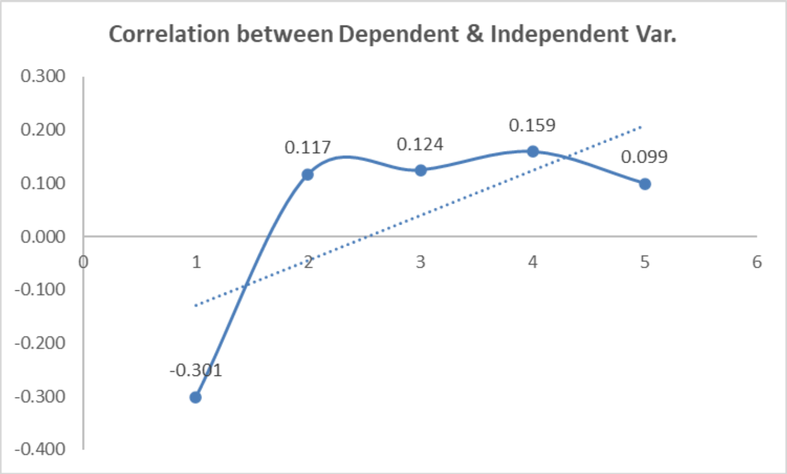
- The dataset contains a mix of bond fund types, with potential implications for investment strategies based on category.
- Variability in returns suggests the importance of diversification and careful selection of bond funds.
- Further analysis could explore the relationship between fund characteristics (such as assets, expense ratio) and returns using regression modeling.
- Investigating outliers and their impact on overall analysis could provide deeper insights into the dataset's characteristics.

Overall, the EDA provides a comprehensive understanding of the bond fund dataset, highlighting key variables and relationships that can inform investment decisions and guide further analysis.

★ MULTIPLE REGRESSION MODELING ★

- A Multiple linear regression model allows us to examine how the response variable is influenced by two or more explanatory variables. The multiple linear regression model is defined as **y=β0+β1x1+β2x2+...+βkxk**
- Visualization of the Correlation between Dependent and independent variables.:

★ **Positive correlation**



- **Preliminary Multiple Regression model** using potential variables :

➢ The Variables we will be using for our Preliminary Multiple Regression Model will be as follows:

❖ Response Variable (y)/Dependent Variable : Assets

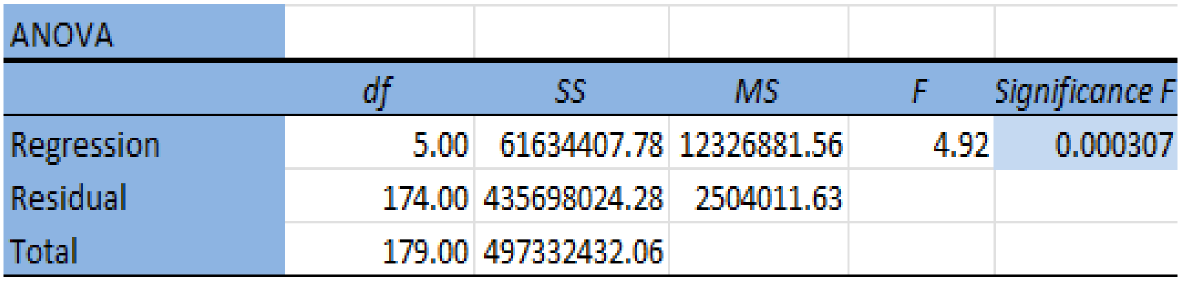
❖ Independent Variable (x) : Expense Ratio, 1-Year Return, 3-Year Return, 5-Year Return & Category Code.

Preliminary Multiple Regression model Equation

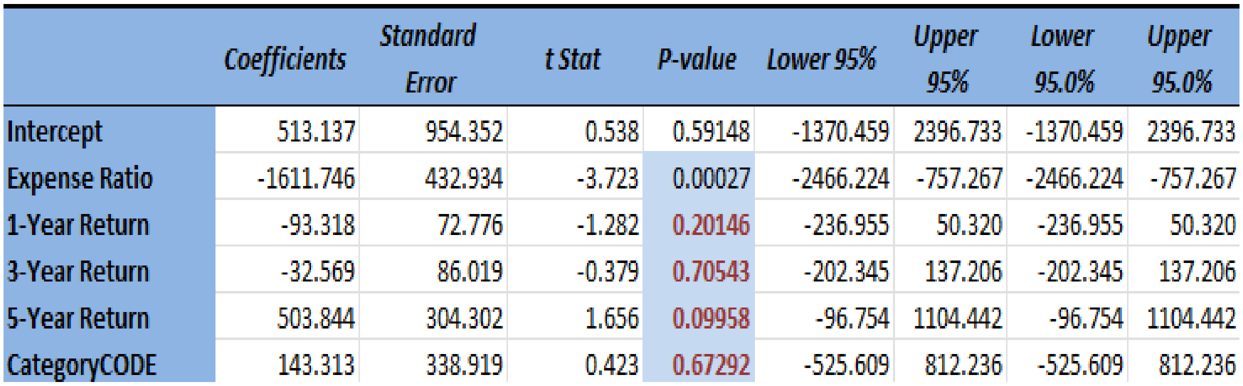
- 513.137 + (-1611.746) x1 + (-93.318) x2 + (-32.569) x3 + (503.844) x4 + (143.313) x5

Regression Statistics	
Multiple R	0.3520
R Square	0.1239
Adjusted R Square	0.0988
Standard Error	1582.4069
Observations	180.00

- Here, we can see the Adjusted R Square is 0.0988 and the Standard Error is 1582.4069. Also we can see that the R Square value is 0.1239.



ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	5.00	61634407.78	12326881.56	4.92	0.000307
Residual	174.00	435698024.28	2504011.63		
Total	179.00	497332432.06			



	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	513.137	954.352	0.538	0.59148	-1370.459	2396.733	-1370.459	2396.733
Expense Ratio	-1611.746	432.934	-3.723	0.00027	-2466.224	-757.267	-2466.224	-757.267
1-Year Return	-93.318	72.776	-1.282	0.20146	-236.955	50.320	-236.955	50.320
3-Year Return	-32.569	86.019	-0.379	0.70543	-202.345	137.206	-202.345	137.206
5-Year Return	503.844	304.302	1.656	0.09958	-96.754	1104.442	-96.754	1104.442
CategoryCODE	143.313	338.919	0.423	0.67292	-525.609	812.236	-525.609	812.236

- Here, we can see that the **Significance F** is 0.000307 which is less than 5% significance level. It means that the model is significant overall.
- But, if we look at the individual significance of the model, the p-value for all variables except Expense Ratio are not individually significant since they are greater than our significance level 0.05.

- **Performing Backward Variable Selection :**

#1

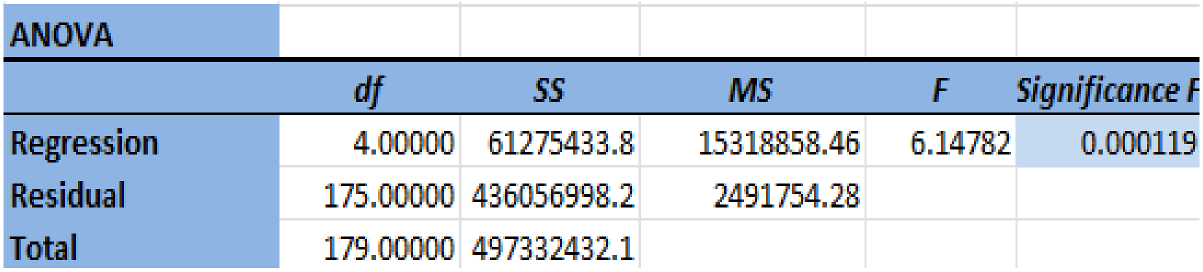
- Here we will perform a backward variable selection method by removing the highest p-value from the model and we will fit a new Regression Model without that variable and check the p-values to check their individual significance.
- We will be removing the highest p-value (0.70543) variable from the previous model and fit a new model. The variable we will be **removing** is: **3-Year return**
- Response Variable (y)/Dependent Variable : Assets
- Independent Variable (x) : Expense Ratio, 1-Year Return, 5-Year Return & Category Code.

New Equation without 3-Year Return Variable

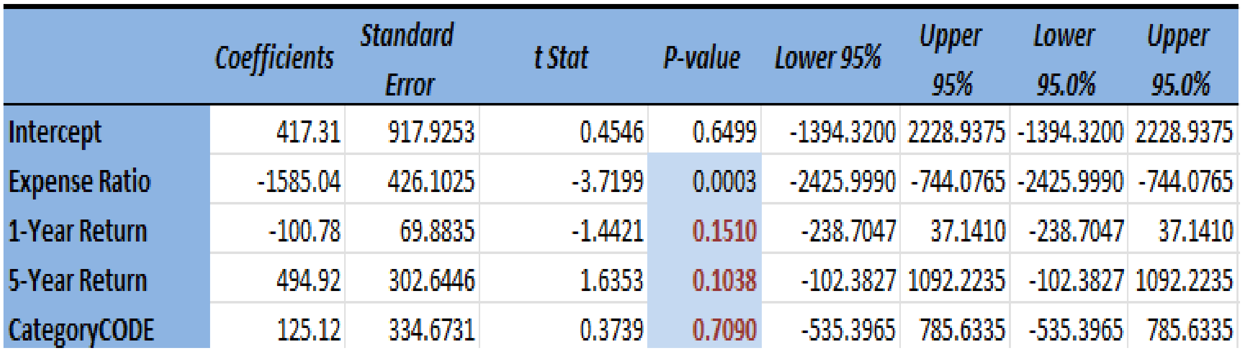
- ❖ 417.31 + (-1585.04) x1 + (-100.78) x2 + (494.92) x3 + (125.12) x4

Regression Statistics	
Multiple R	0.3510
R Square	0.1232
Adjusted R Square	0.1032
Standard Error	1578.5291
Observations	180.00

- Here, we can see the Adjusted R Square is 0.1032 and the Standard Error is 1578.5291. Also we can see that the R Square value is 0.1232.
- We can observe here that the Adjusted R Square has +0.0044 difference and as we know **"Higher the Adjusted R Square, Better the Model".**



ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	4.00000	61275433.8	15318858.46	6.14782	0.000119
Residual	175.00000	436056998.2	2491754.28		
Total	179.00000	497332432.1			



	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	417.31	917.9253	0.4546	0.6499	-1394.3200	2228.9375	-1394.3200	2228.9375
Expense Ratio	-1585.04	426.1025	-3.7199	0.0003	-2425.9990	-744.0765	-2425.9990	-744.0765
1-Year Return	-100.78	69.8835	-1.4421	0.1510	-238.7047	37.1410	-238.7047	37.1410
5-Year Return	494.92	302.6446	1.6353	0.1038	-102.3827	1092.2235	-102.3827	1092.2235
CategoryCODE	125.12	334.6731	0.3739	0.7090	-535.3965	785.6335	-535.3965	785.6335

- Here, we can see that the **Significance F** is 0.000119 which is less than 5% significance level. It means that we are 95% confident that the model is significant overall.
- But, if we look at the individual significance of the model, the p-value for all variables except Expense Ratio are not individually significant since they are greater than our significance level 0.05.

- **Performing Backward Variable Selection :**

#2

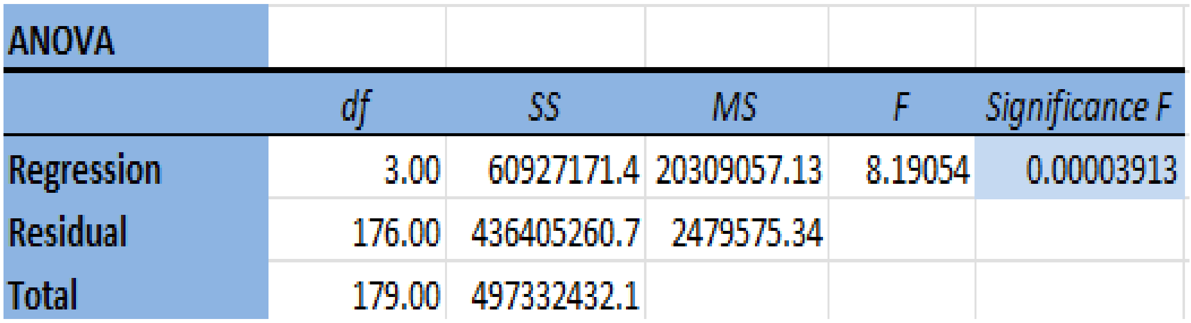
- ❖ Here we will perform a second backward variable selection method by removing the highest p-value from the model and we will fit a new Regression Model without that variable and check the p-values to check their individual significance.
- ❖ We will be removing the highest p-value (0.7090) variable from the previous model and fit a new model. The variable we will be **removing** is: **Category Code**
- ❖ Response Variable (y)/Dependent Variable : Assets
- ❖ Independent Variable (x) : Expense Ratio, 1-Year Return, 5-Year Return.

New Equation without 3-Year Return & Category Code Variable

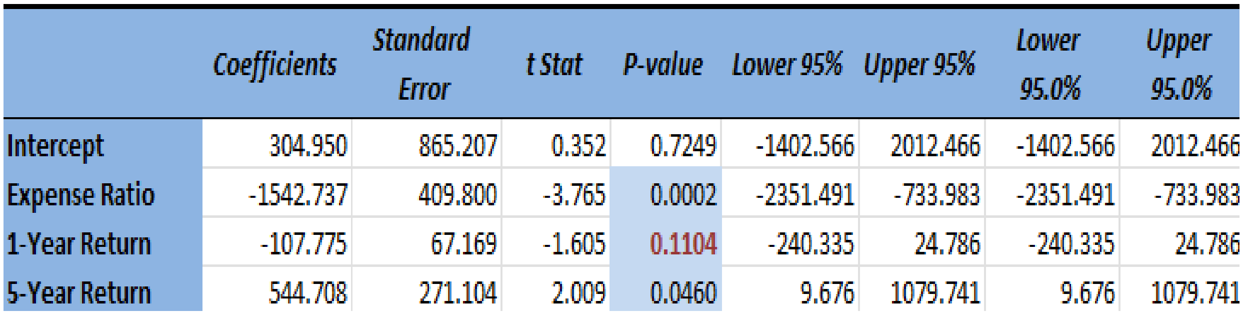
- **304.950 + (-1542.737) x1 + (-107.775) x2 + (544.708) x3**

Regression Statistics	
Multiple R	0.3500
R Square	0.1225
Adjusted R Square	0.1076
Standard Error	1574.6667
Observations	180.00

- Here, we can see the Adjusted R Square is 0.1076 and the Standard Error is 1574.6667. Also we can see that the R Square value is 0.1225.
- We can observe here that the Adjusted R Square has again +0.0044 difference and as we know **"Higher the Adjusted R Square, Better the Model".**



ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	3.00	60927171.4	20309057.13	8.19054	0.00003913
Residual	176.00	436405260.7	2479575.34		
Total	179.00	497332432.1			



	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	304.950	865.207	0.352	0.7249	-1402.566	2012.466	-1402.566	2012.466
Expense Ratio	-1542.737	409.800	-3.765	0.0002	-2351.491	-733.983	-2351.491	-733.983
1-Year Return	-107.775	67.169	-1.605	0.1104	-240.335	24.786	-240.335	24.786
5-Year Return	544.708	271.104	2.009	0.0460	9.676	1079.741	9.676	1079.741

- Here, we can see that the **Significance F** is 0.00003913 which is less than 5% significance level. It means that we are 95% confident that the model is significant overall.
- But, if we look at the individual significance of the model, the p-value for 1-Year Return variable is not individually significant since it is greater than our significance level 0.05. Although, p-value for **5-Year Return is Individually significant** since it is **less than alpha 0.05.** That means the 5-Year return Variable is individually significant to this model.

- **Performing Backward Variable Selection :**

#3

- ❖ Here we will perform a third backward variable selection method by removing the highest p-value from the model and we will fit a new Regression Model without that variable and check the p-values to check their individual significance.
- ❖ We will be removing the highest p-value (0.1104) variable from the previous model and fit a new model. The variable we will be **removing** is: **1-Year Return**

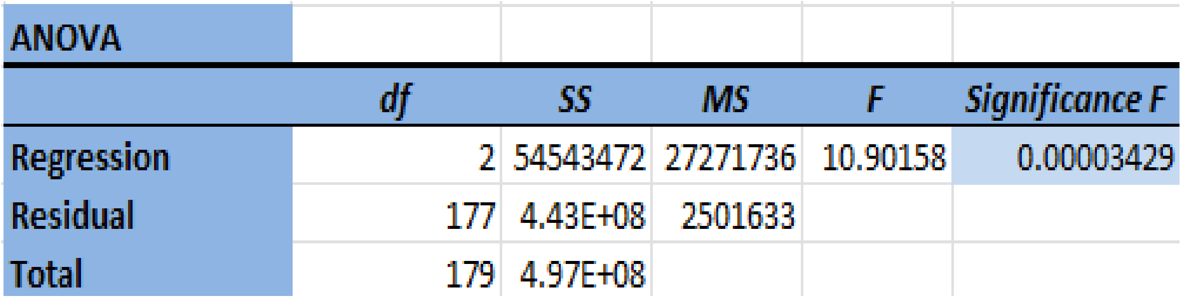
- Response Variable (y)/Dependent Variable : Assets
- Independent Variable (x) : Expense Ratio, 5-Year Return.

New Equation without 3-Year Return & Category Code Variable

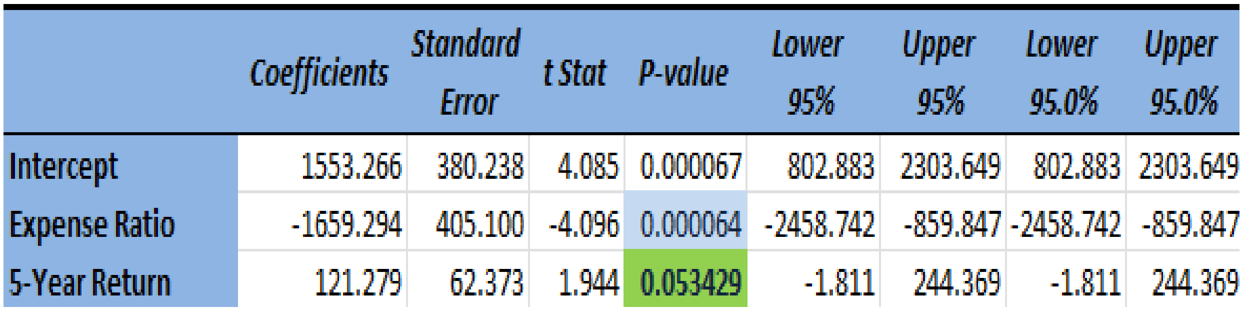
- **1553.266 + (-1659.294) Expense Ratio + (121) 5-Year Ratio**

Regression Statistics	
Multiple R	0.3312
R Square	0.1091
Adjusted R Square	0.0996
Standard Error	1581.65507
Observations	180.00

- Here, we can see the Adjusted R Square is 0.0996 and the Standard Error is 1581.6550. Also we can see that the R Square value is 0.1091.
- We can observe here that the Adjusted R Square has again +0.0079 difference and as we know **"Higher the Adjusted R Square, Better the Model".**



ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	54543472	27271736	10.90158	0.00003429
Residual	177	4.43E+08	2501633		
Total	179	4.97E+08			



	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	1553.266	380.238	4.085	0.000067	802.883	2303.649	802.883	2303.649
Expense Ratio	-1659.294	405.100	-4.096	0.000064	-2458.742	-859.847	-2458.742	-859.847
5-Year Return	121.279	62.373	1.944	0.053429	-1.811	244.369	-1.811	244.369

- Here, we can see that the **Significance F** is 0.00003429 which is less than 5% significance level. It means that we are 95% confident that the model is significant overall.
- Model#3 saw a significant increase in intercept compared to Model #2. The difference between two model's intercept is +1248.317.
- But, if we look at the individual significance of the model, the p-value for Expense Ratio & **5-Year Return is Individually significant** since it is **less than alpha 0.05.** That means the **5-Year return Variable is individually significant to this model.**

Therefore, the best Model we will be choosing is Model#3.

Assets = 1553.266 + (-1659.294) Expense Ratio + (121) 5-Year Ratio

- **Performing Statistical test to confirm it is better than the first model:**
  - Test of Individual Significance:
  - We will fit Two Regression models:
    - Full Model : **y=β0 + β1 Expense Ratio + β2 5-Year Return**
    - Restricted Model : **y=β0 + β1 Expense Ratio**

- The competing **Hypothesis** is as follows:

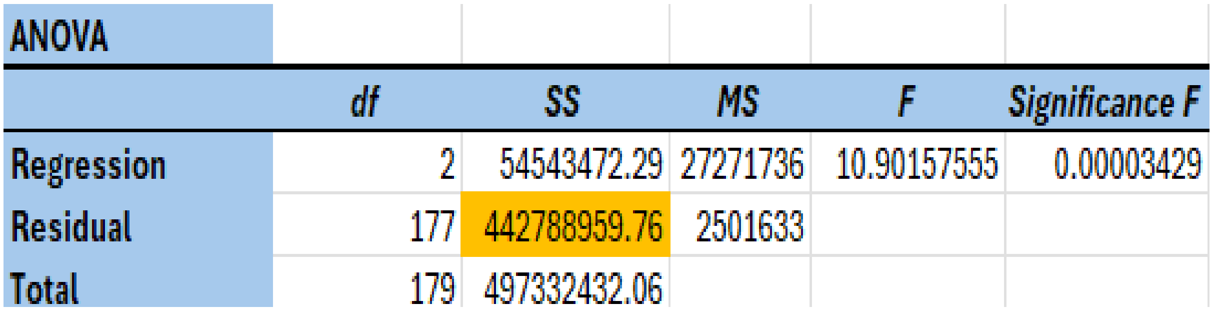
  Ho: β2 = β3 = 0

  HA : At least one βi ≠ 0

- n= 180

- **Full Model : y=β_0 + β_1 Expense Ratio + β_2 5-Year Return**

Regression Statistics	
Multiple R	0.331
R Square	0.1097
Adjusted R Square	0.0996
Standard Error	1581.655
Observations	180.00



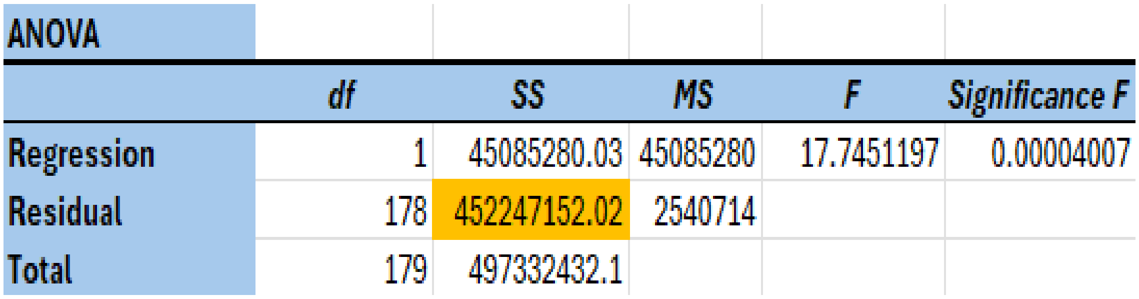
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	54543472.29	27271736	10.90157555	0.00003429
Residual	177	442788959.76	2501633		
Total	179	497332432.06			

.

- Here, we can see that the value of Significance F is 0.00003429 which is less than alpha 0.05. Therefore, this model is overall significant.

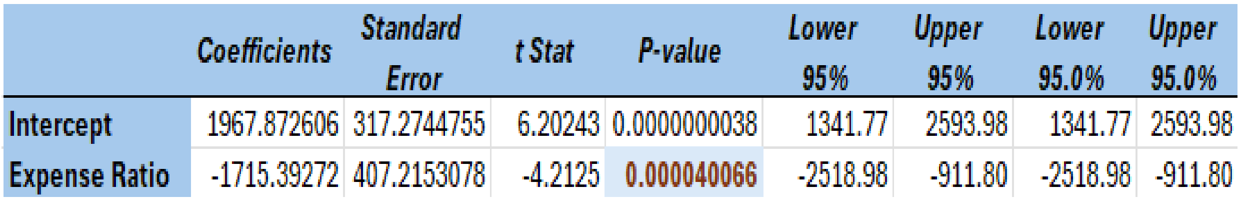
- **Residual for the Full / Unrestricted Model (SSE_U): 442788959.76**
- **Restricted Model : $y=\beta_0 + \beta_1$ Expense Ratio**

Regression Statistics	
Multiple R	0.301
R Square	0.091
Adjusted R Square	0.086
Standard Error	1593.962
Observations	180.00



ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	45085280.03	45085280	17.7451197	0.00004007
Residual	178	452247152.02	2540714		
Total	179	497332432.1			

- Here, we can see that the value of Significance F is 0.00004007 which is less than alpha 0.05. Therefore, this model is overall significant.
- **Residual for the Full / Unrestricted Model (SSE_R): 452247152.02**



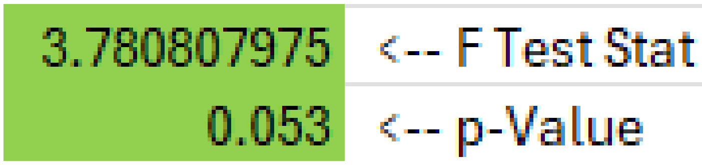
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	1967.872606	317.2744755	6.20243	0.0000000038	1341.77	2593.98	1341.77	2593.98
Expense Ratio	-1715.39272	407.2153078	-4.2125	**0.000040066**	-2518.98	-911.80	-2518.98	-911.80

- **Now we will calculate F Test-stat**

F(df1,df2) = [(SSER-SSEU) / df1] / [SSEU / df2]

$$F_{(df_1,df_2)} = \frac{(SSE_R - SSE_U)/df_1}{SSE_U/df_2}$$

- **SSE$_R$ = 452247152.02**
- **SSE$_U$ = 442788959.76**
- **n= 180**
- **df1= 1**
- **df2= 177**

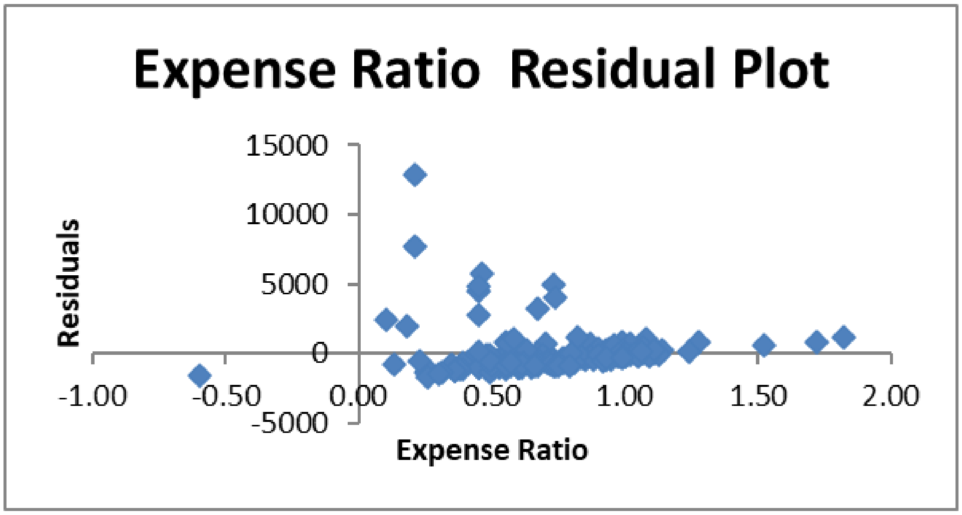


3.780807975	<-- F Test Stat
0.053	<-- p-Value

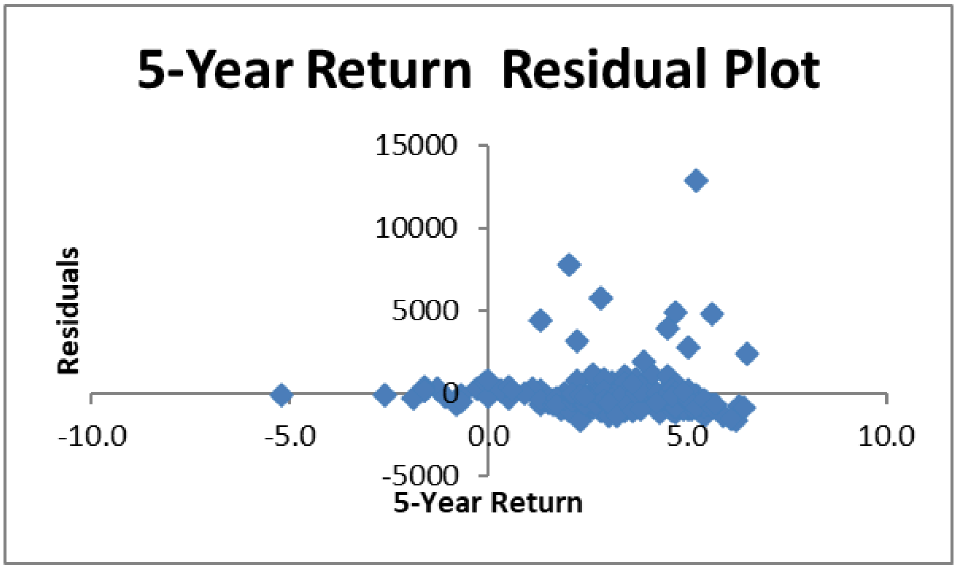
- **Since p-value ≥ alpha, we DO NOT REJECT the null and conclude that we do not have enough evidence to claim that the influence of 5-Year Return on Assets are not different. So Expense Ratio and 5-Year Return are creating the same impact on Assets. Since we cannot confirm it is better than the first model statistically, we will proceed with the full model –> y=β_0 + β_1 Expense Ratio + β_2 5-Year Return.**

★ RESIDUAL PLOTS:

➢ Expense Ratio:



➢ 5-Year Return:



Regression Statistics	
Multiple R	0.3312
R Square	0.1091
Adjusted R Square	0.0996
Standard Error	1581.65507
Observations	180.00

Sample Multiple Regression equation for the “Final Best” model we have developed.

- **Assets = 1553.266 + (-1659.294) Expense Ratio + (121) 5-Year Ratio**

➔ **Interpret the meaning of the Y intercept and interpret the meaning of all the slopes for our fitted model**

- ◆ 1% growth in Expense ratio, -1659.294 is the expected decrease in the Assets(in billions of $).
- ◆ 1% growth in 5-Year Return, 121 is the expected increase in the Assets(in billions of $).

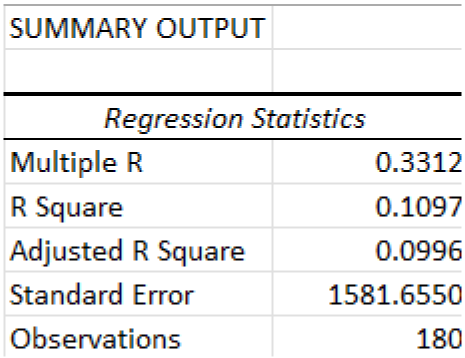
- ➔ **Interpret the meaning of the coefficient of multiple determination R^2 & Interpret the meaning of the standard error of the estimate SYX.**
  - ◆ R^2 : The multiple R2 value is 0.1091. When we look at this value 10.91% it means that the Regression output is closer to zero. We can determine that these variables are weak.
  - ◆ Standard Error of the Estimate: The standard deviation of the residual; used as goodness-of-fit measure for regression analysis. S_e can assume any value between 0 and infinity. $0 \leq S_e < \infty$. Our final model has the smallest S_e hence the preferred model.

- ❖ **Select one value for each of your independent variables in their respective relevant ranges (do no extrapolate):**

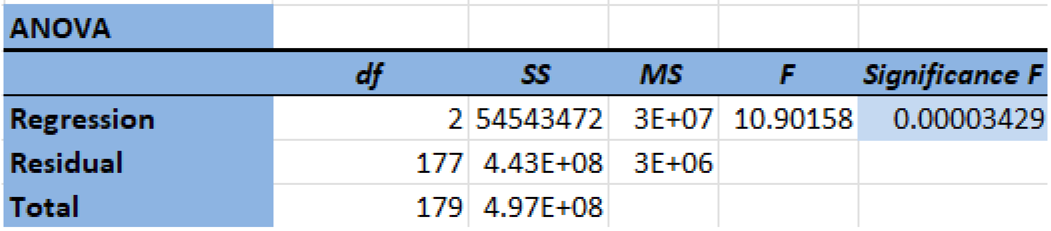
- ➔ Expense Ratio = 0.72
- ➔ 5 - Year return = 3.08

- ❖ **Predicted ŷ:** 2834494.52

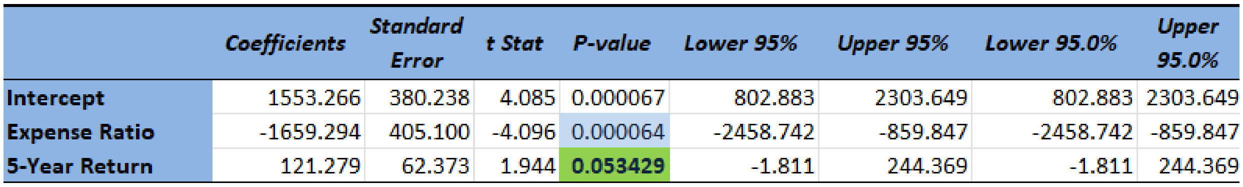
❖ Confidence Interval:



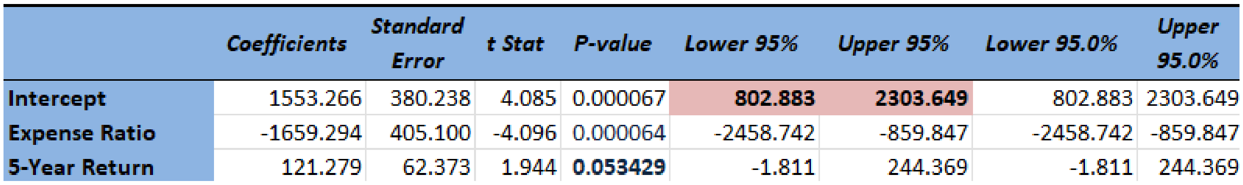
SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.3312
R Square	0.1097
Adjusted R Square	0.0996
Standard Error	1581.6550
Observations	180



ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	54543472	3E+07	10.90158	0.00003429
Residual	177	4.43E+08	3E+06		
Total	179	4.97E+08			



	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	1553.266	380.238	4.085	0.000067	802.883	2303.649	802.883	2303.649
Expense Ratio	-1659.294	405.100	-4.096	0.000064	-2458.742	-859.847	-2458.742	-859.847
5-Year Return	121.279	62.373	1.944	**0.053429**	-1.811	244.369	-1.811	244.369



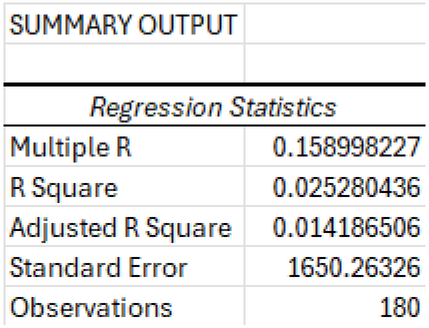
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	1553.266	380.238	4.085	0.000067	**802.883**	**2303.649**	802.883	2303.649
Expense Ratio	-1659.294	405.100	-4.096	0.000064	-2458.742	-859.847	-2458.742	-859.847
5-Year Return	121.279	62.373	1.944	**0.053429**	-1.811	244.369	-1.811	244.369

☆As we see above in the attached image, the confidence interval is

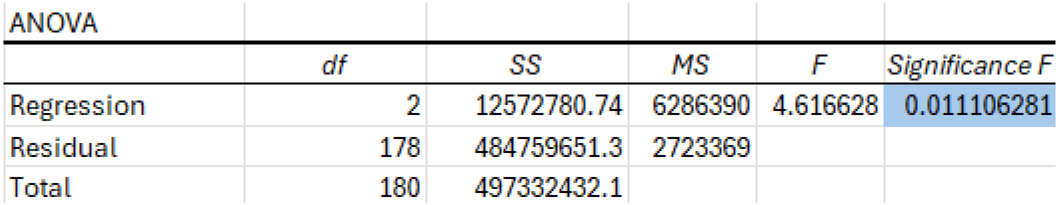
(802.883 , 2303.649)

❖ Prediction Interval:

❖ **We came up with one value for each of your independent variables in their respective relevant : 0.72 & 8.03**

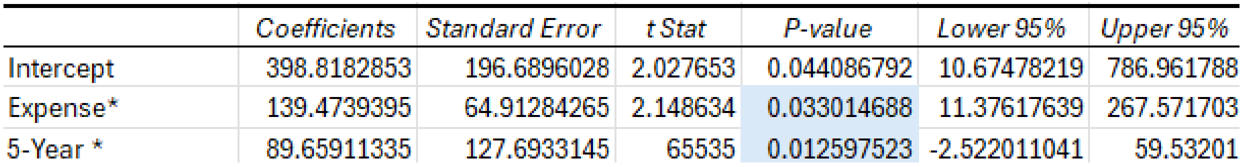


SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.158998227
R Square	0.025280436
Adjusted R Square	0.014186506
Standard Error	1650.26326
Observations	180

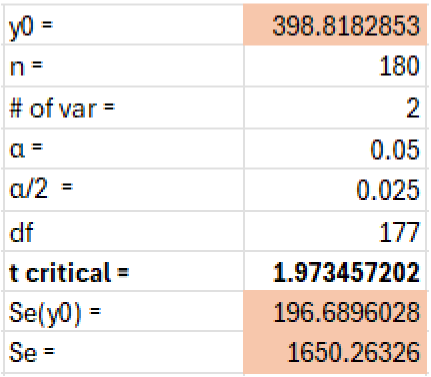


ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	12572780.74	6286390	4.616628	0.011106281
Residual	178	484759651.3	2723369		
Total	180	497332432.1			

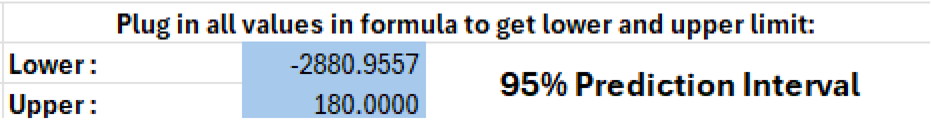
☆As we can see here, the Significance F value is 0.01 which means the model is overall significant.



	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	398.8182853	196.6896028	2.027653	0.044086792	10.67478219	786.961788
Expense*	139.4739395	64.91284265	2.148634	0.033014688	11.37617639	267.571703
5-Year *	89.65911335	127.6933145	65535	0.012597523	-2.522011041	59.53201



y0 =	398.8182853
n =	180
# of var =	2
α =	0.05
α/2 =	0.025
df	177
t critical =	**1.973457202**
Se(y0) =	196.6896028
Se =	1650.26326



Plug in all values in formula to get lower and upper limit:		
Lower :	-2880.9557	95% Prediction Interval
Upper :	180.0000	

Prediction Interval : (-2880.9557 , 180.00)

☆**Prediction Intervals are usually wider than the Confidence interval and as we can see here, Prediction Interval is much wider than the Confidence interval.**

☆ In order to determine the 95% confidence interval estimate for the predicted ŷ we found it to be 802.883 , 2303.649. For the 95% prediction interval estimate for the predicted ŷ,we determine with one value for each of your independent variables in their respective relevant is 0.72 & 8.03. We also find that the Significance F value is 0.01 which would mean the model is overall significant. Finally the prediction interval is -288.9557, 180.00 we find the prediction interval is a lot wider than the Confidence interval.

Bibliography

1. Jaggia, Sanjiv. *Business Statistics: Communicating with Numbers*. fourth ed., McGraw-Hill Education, 2022, 978126421882.