SELF ANALYSIS ON DAILY EXPENSE(Jan 2022-Dec 2022)

BY DHRUVI PATEL

OBJECTIVE

The primary goal of this analysis is to comprehensively examine and understand the patterns, trends, and relationships in expense data over the past year. The key goals of this analysis include:

- Understanding category based spending
- Statistical analysis
- Identifying spending trends
- Validate assumptions through hypothesis testing.
- Forecast future expenses

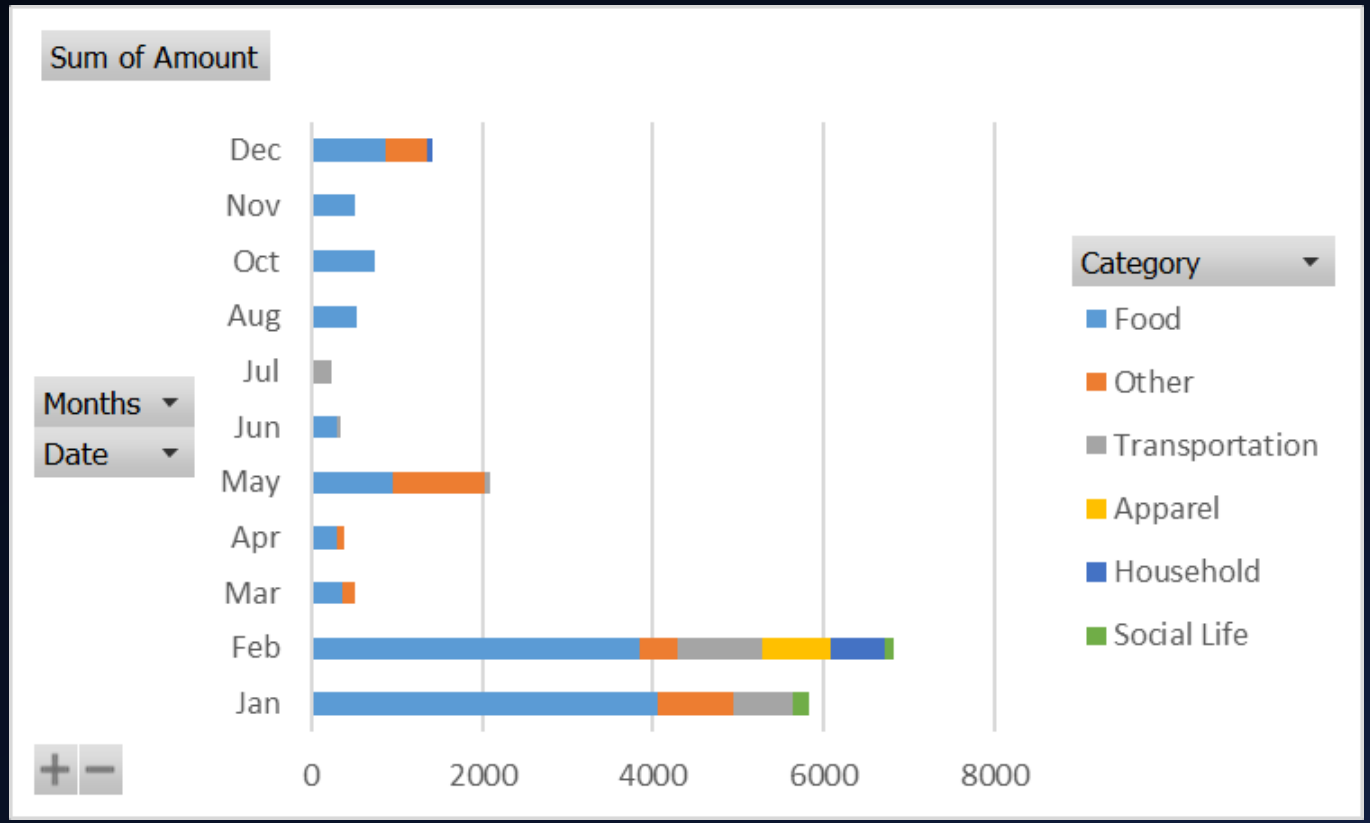
DATA OVERVIEW

1. Number of entries: 110
2. Categories of expense :
    - Food
    - Transportation
    - Apparel
    - Social life
    - Household
    - Others
3. Time period of data: January 2022-December 2022
4. Outliers has been removed by IQR method where lower bound(-314.0375) and upper bound(708.0625)

CATEGORY ANALYSIS

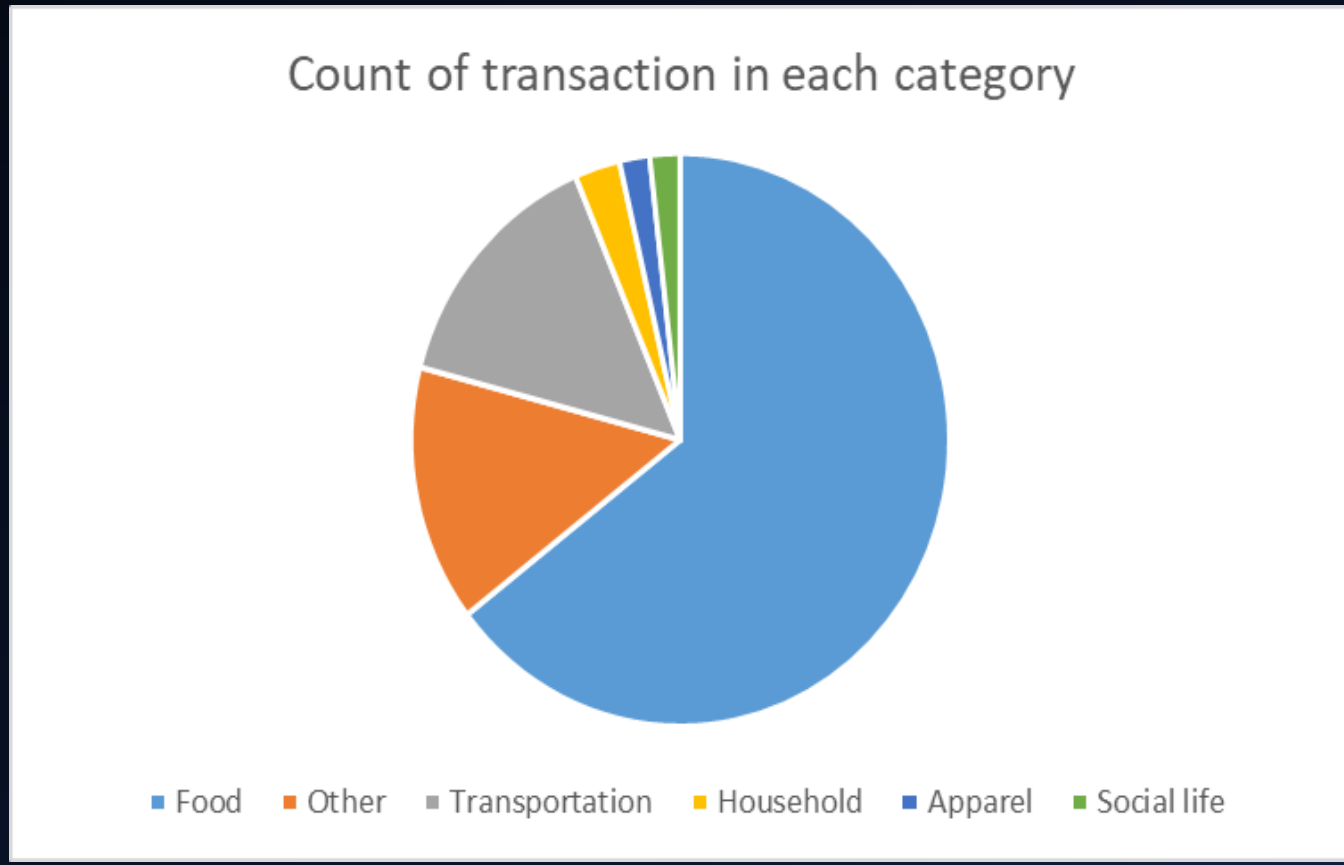
Monthly expense distribution by category

Sum of Amount	Column Labels						
Row Labels	Food	Other	Transportation	Apparel	Household	Social Life	Grand Total
Jan	4050.15	900	688			200	5838.15
Feb	3839.85	458	989.8	798	639	100	6824.65
Mar	360	150					510
Apr	293	80					373
May	953	1070	60				2083
Jun	305		42				347
Jul	15		214				229
Aug	535.3						535.3
Oct	747						747
Nov	508.5						508.5
Dec	872.75	479			70		1421.75
Grand Total	12479.55	3137	1993.8	798	709	300	19417.35



Count of monthly transaction per category

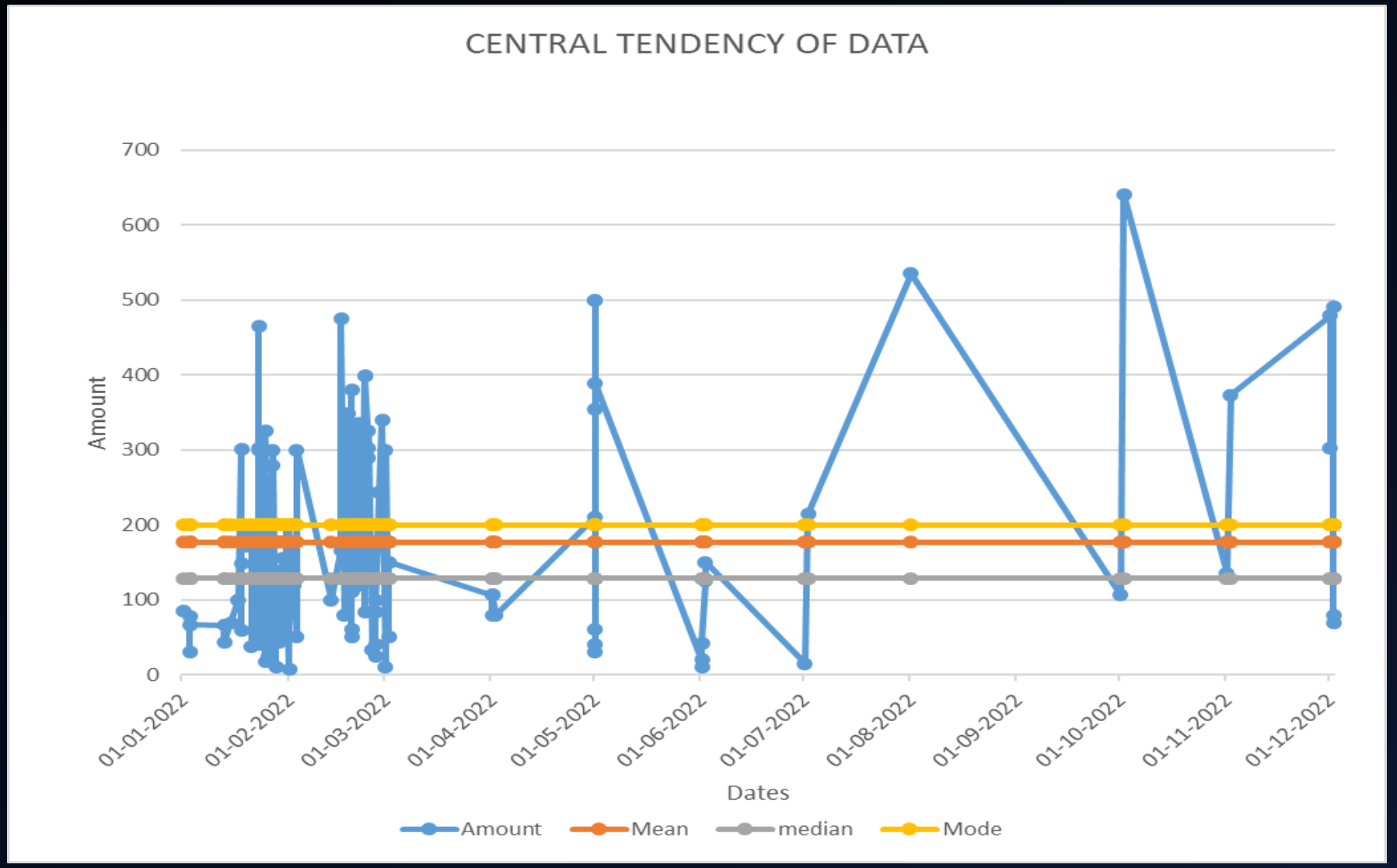
Count of Amount	Column Labels						
Row Labels	Food	Other	Transportation	Household	Apparel	Social Life	Grand Total
Jan	29	5	7			1	42
Feb	20	4	6	2	2	1	35
Mar	3	1					4
Apr	3	1					4
May	3	4	1				8
Jun	4		1				5
Jul	1		1				2
Aug	1						1
Oct	2						2
Nov	2						2
Dec	3	1		1			5
Grand Total	71	16	16	3	2	2	110



STATISTICAL ANALYSIS

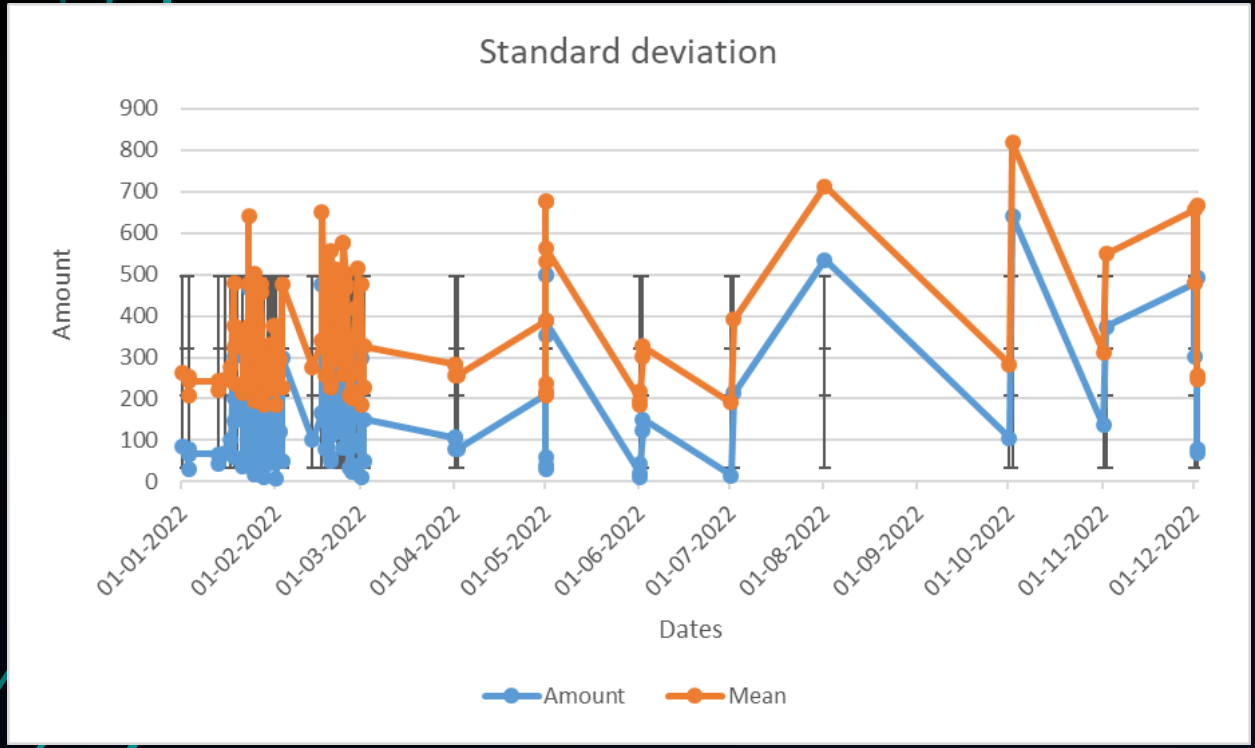
DATA	VALUE
Mean	176.5213636
Standard Error	13.67487873
Median	129
Mode	200
Standard Deviation	143.4233381
Sample Variance	20570.25392
Kurtosis	0.312679786
Skewness	1.004894197
Range	633
Minimum	8
Maximum	641
Sum	19417.35
Count	110

Central tendency



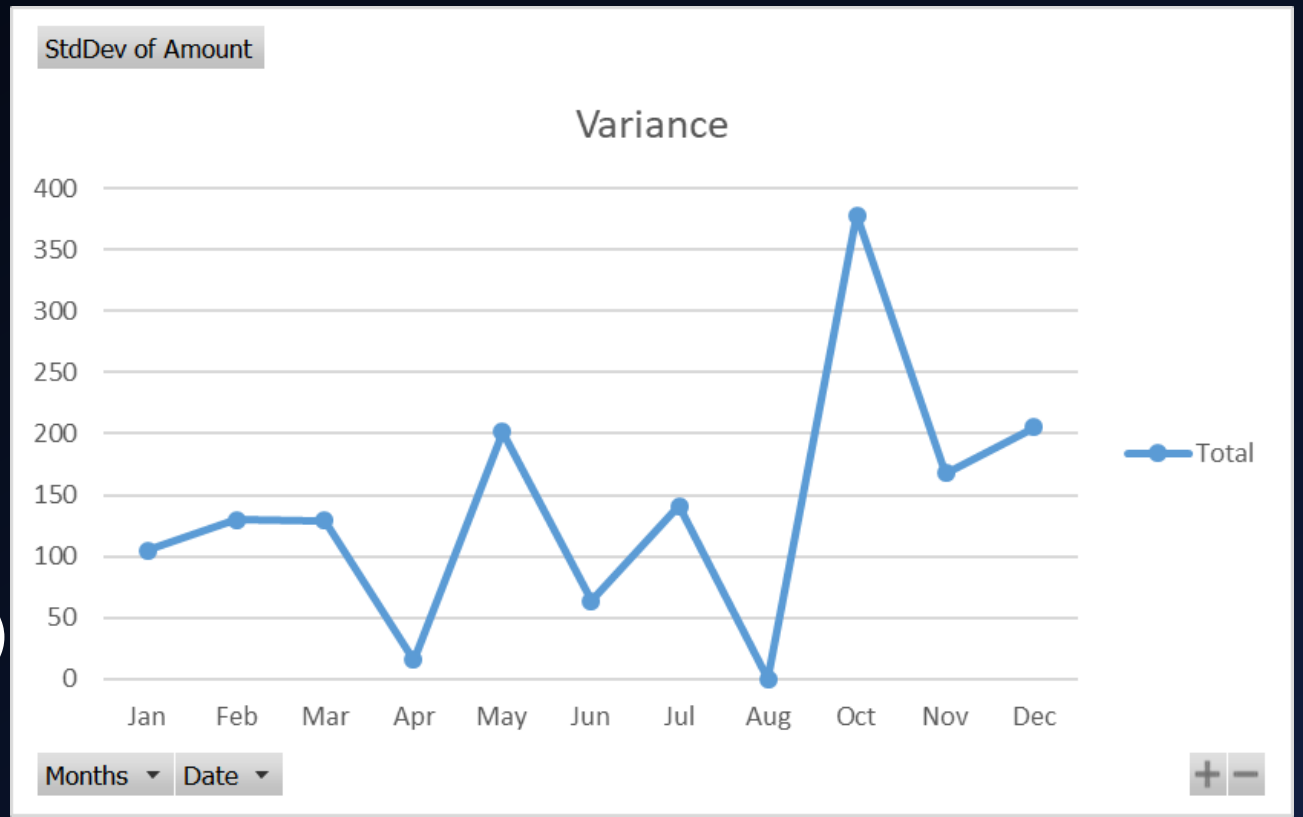
Mean: 176.52
Median: 129
Mode: 200

Standard deviation and variance

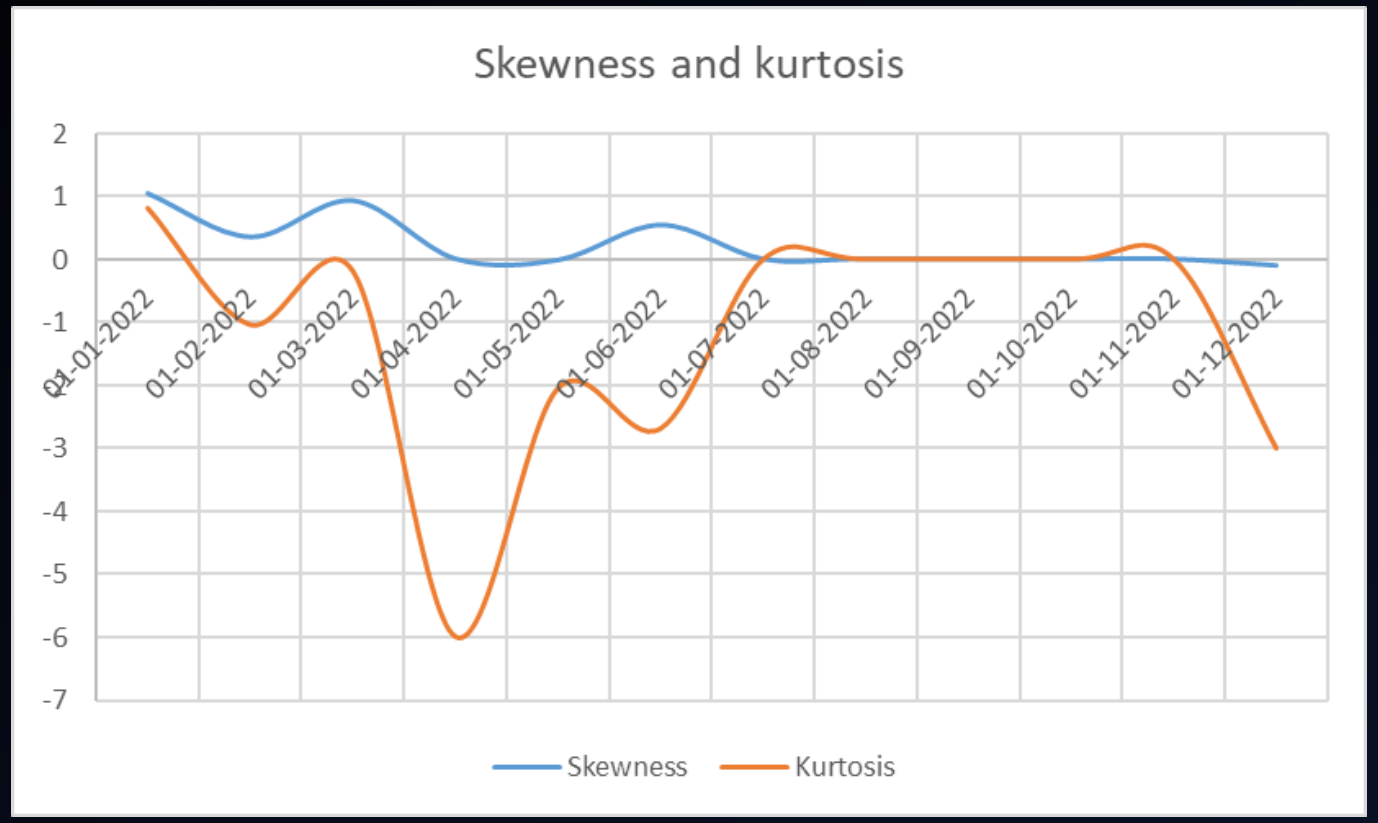


- Standard deviation measures how much the data points deviate from the mean.
- A higher standard deviation indicates greater variability, while a lower standard deviation indicates data points are clustered closer to the mean.

- Variance is a statistical measure that quantifies the amount of variation or spread in a dataset.
- A variance of 20570.25392 is relatively high, indicating a significant amount of variation in dataset. This means that the data points are widely spread out, and the average value (mean) might not be a very representative measure of the central tendency.



Skewness and kurtosis



Skewness:
The skewness values fluctuate over time, indicating varying degrees of asymmetry in the data.
There are periods of positive skewness (e.g., around January and April), suggesting a longer tail on the right side.
There are also periods of negative skewness (e.g., around March and December), suggesting a longer tail on the left side.

Kurtosis:
The kurtosis values are predominantly negative, indicating a flatter distribution compared to a normal distribution (platykurtic).
The kurtosis values fluctuate, suggesting varying degrees of tailedness in the data.

The combination of negative kurtosis and fluctuating skewness suggests that the data is likely skewed in different directions at various times, but the overall shape tends to be flatter than a normal distribution.

CORRELATION AND REGRESSION ANALYSIS

Detailed Regression Analysis of Daily Expenses

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.273086262					
R Square	0.074576106					
Adjusted R Square	0.066007367					
Standard Error	138.6090388					
Observations	110					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	1	167211.3895	167211.389	8.70328	0.00389537	
Residual	108	2074946.288	19212.4656			
Total	109	2242157.677				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-20736.8497	7088.974641	-2.9252256	0.0042	-34788.4273	-6685.27219
X Variable 1	0.468516212	0.158811975	2.95013151	0.0039	0.15372333	0.7833091

1. Correlation Insights

Weak Positive Relationship: The correlation coefficient is **0.273**, indicating a weak positive relationship between the date (time) and daily expenses. While there is a relationship, it is not strong, implying that other factors are also influencing daily expenses.

2. Key Regression Statistics

Multiple R (0.273): Reflects a weak linear relationship between the date and the amount spent daily. The correlation is positive but not strong.

R-Squared (7.46%): Only **7.46%** of the variability in daily expenses is explained by the progression of dates. This suggests that the date alone is not a strong predictor of spending patterns, and other factors are likely more influential.

Standard Error (₹138.61): On average, the actual daily expenses deviate from the predicted expenses by **₹138.61**. This indicates that predictions based on this model can be off by this amount, reflecting significant variability in daily expenses.

3. ANOVA (Analysis of Variance) Results

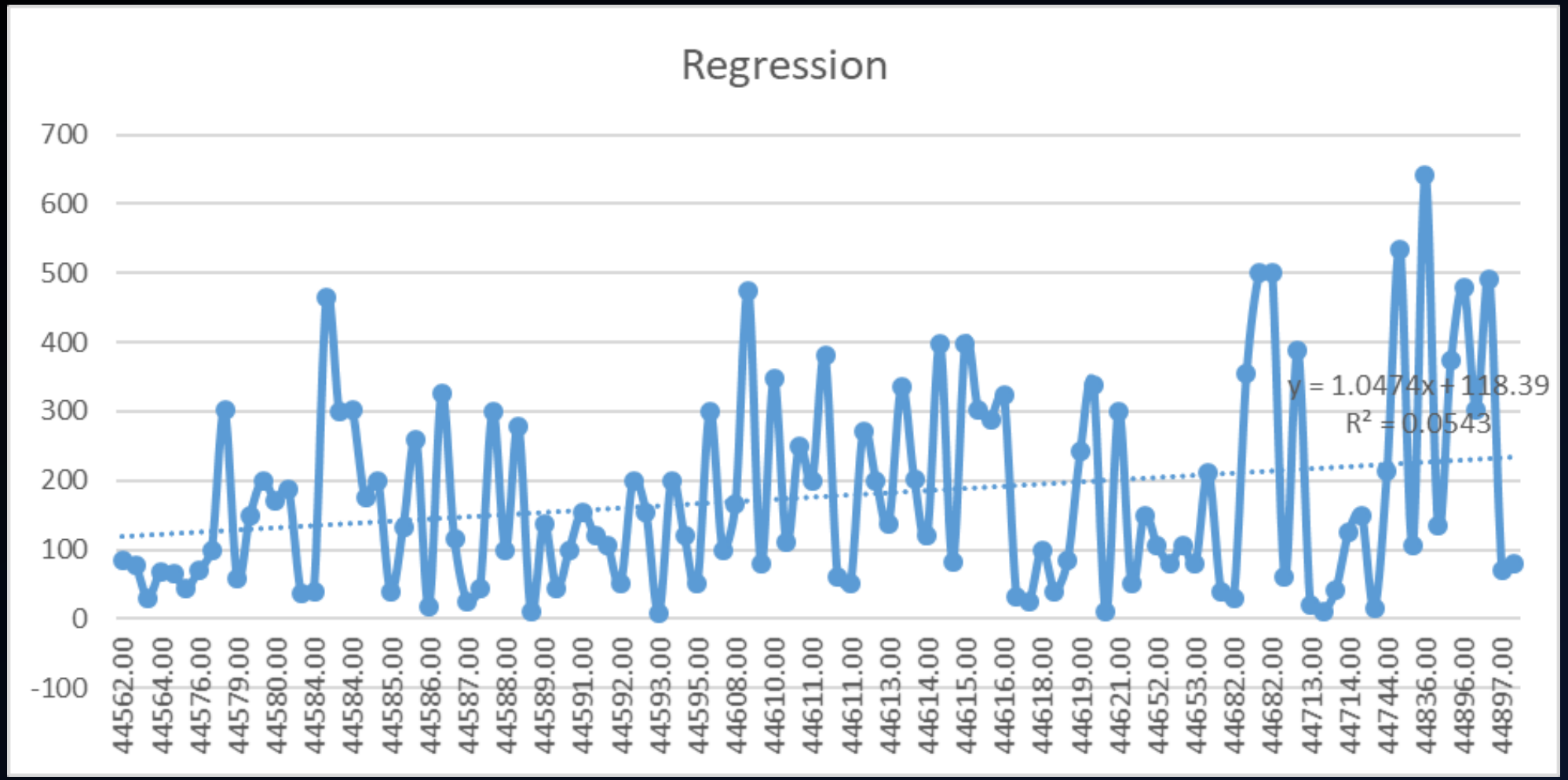
Significant F-Statistic (8.703, p = 0.0039): The model is statistically significant, meaning there is a real relationship between the date and daily expenses, even though it is weak. The low p-value (< 0.05) confirms that this relationship is not due to random chance.

Interpretation: While the date has a statistically significant impact on expenses, it explains only a small fraction of the total variance, indicating that other factors likely play a more important role.

4. Model Coefficients Breakdown

Intercept (₹-20,736.85): This negative value suggests that if we extrapolate the model backward to a date of zero (which isn't meaningful in this context), expenses would theoretically be negative. This highlights that the date variable alone might not be sufficient to explain expense behavior.

Date Coefficient (₹0.47/day): Each passing day is associated with an average increase of **₹0.47** in daily expenses. This small but statistically significant increase suggests a very gradual upward trend in spending over the year.



- The plot would show a slight upward trend in expenses over time, but with considerable scatter around the trendline. The low R^2 value (0.0746) highlights the considerable variation in daily expenses that is not explained by the date.

- While the model is statistically significant, the date alone explains only a small portion of the variability in daily expenses. This suggests that other variables (e.g., specific expense categories, external events) may be more influential.

HYPOTHESIS TESTING

1. Null Hypothesis ($H0$): There is no significant trend in expenses over time (the slope of the trend line is zero).Alternative Hypothesis ($H1$H 1): There is a significant trend in expenses over time (the slope of the trend line is not zero).

2. Independent Variable: Time (Date).
   Dependent Variable: Expense Amount.

3. Expense=β 0 +β 1 ×Time+ϵ
β 0 : Intercept
β 1 : Slope of the trend line
ϵ: Error term

4. Slope ($\beta 1$): The coefficient for the time variable indicates the direction and magnitude of the trend.

- If $\beta 1>0$, there is an upward trend in expenses over time.
- If $\beta 1<0$, there is a downward trend in expenses over time.
- If $\beta 1=0$, there is no trend in expenses over time.

P-Value:

- If the p-value is less than your significance level (commonly 0.05), you reject $H0$ and conclude that there is a significant trend in your expenses over time.
- If the p-value is greater than your significance level, you fail to reject $H0$, indicating that there is no significant trend.

Detailed Regression Analysis of Daily Expenses

SUMMARY OUTPUT						
Correlation	0.273086262					
Regression Statistics						
Multiple R	0.273086262					
R Square	0.074576106					
Adjusted R Square	0.066007367					
Standard Error	138.6090388					
Observations	110					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	1	167211.3895	167211.389	8.70328	0.00389537	
Residual	108	2074946.288	19212.4656			
Total	109	2242157.677				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-20736.8497	7088.974641	-2.9252256	0.0042	-34788.4273	-6685.27219
X Variable 1	0.468516212	0.158811975	2.95013151	0.0039	0.15372333	0.7833091

1. Hypotheses Recap

H0: There is no significant trend in expenses over time (the slope of the trend line is zero).

H1: There is a significant trend in expenses over time (the slope of the trend line is not zero).

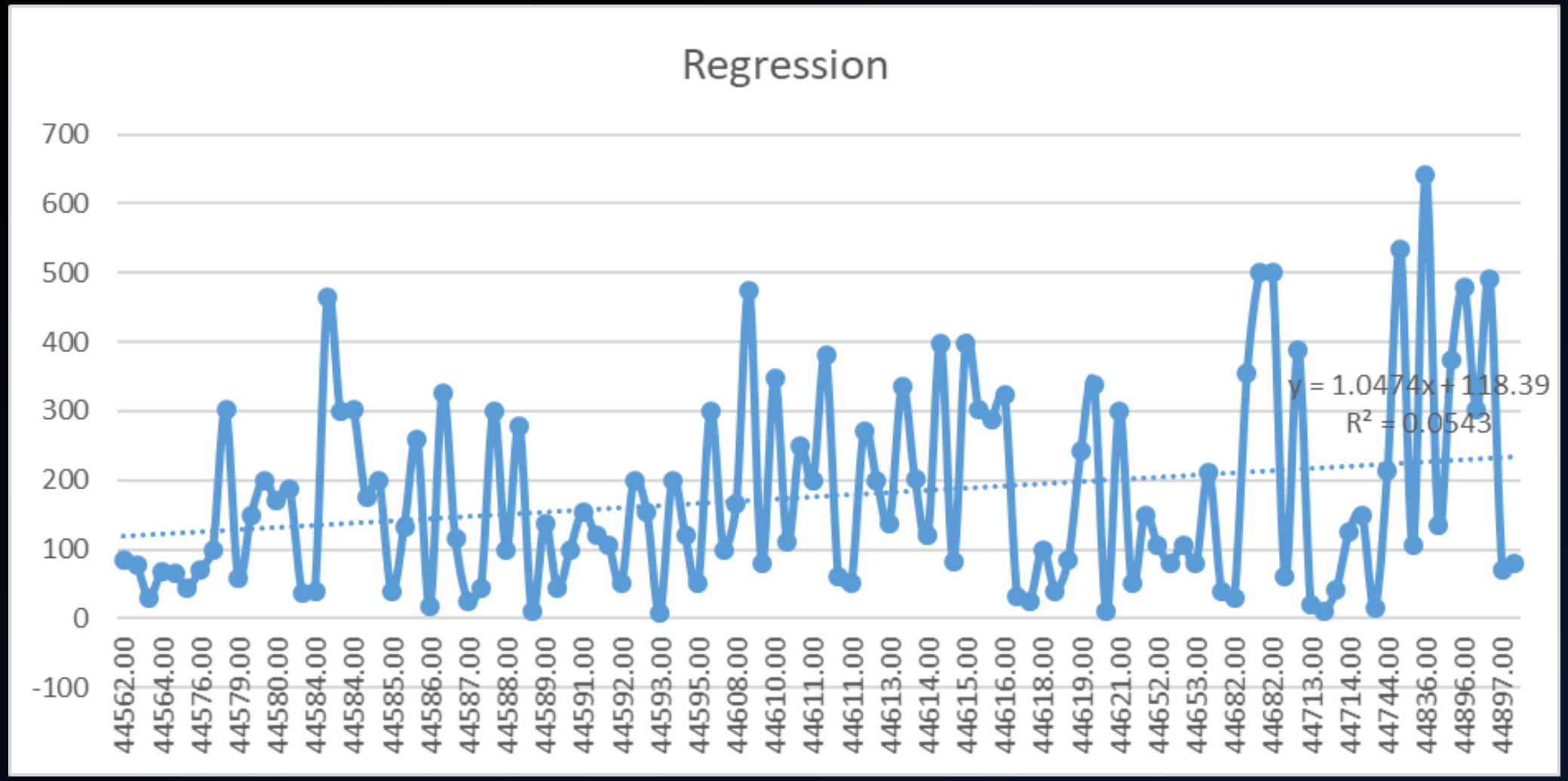
2. Key Metrics

- **Multiple R (Correlation):** 0.273
    - This value indicates a positive but weak correlation between time and expenses.
- **R Square:** 0.0746
    - This means that about 7.46% of the variance in expenses is explained by time. This is a low value, indicating that time is not a strong predictor of expenses, but there is still a measurable effect.
- **Significance F (P-value for ANOVA):** 0.0039
    - The p-value is below the significance level (0.05), so you reject the null hypothesis H0. This means that the regression model is statistically significant, indicating that there is a significant trend in your expenses over time.
- **Coefficients:**
    - **Intercept:** -20,736.85 (with a p-value of 0.0042)
    - **X Variable 1 (Slope):** 0.4685 (with a p-value of 0.0039)

        The slope is positive, indicating an upward trend in expenses over time. The p-value associated with the slope is also below 0.05, confirming that the trend is significant.

3. Interpretation

- **Significant Trend:** The p-value for the slope (0.0039) is less than 0.05, meaning you reject the null hypothesis and conclude that there is a significant upward trend in expenses over time.
- **Strength of Trend:** While the trend is statistically significant, the R Square value is quite low (0.0746), suggesting that time explains only a small portion of the variance in expenses. Other factors might also influence your expenses.

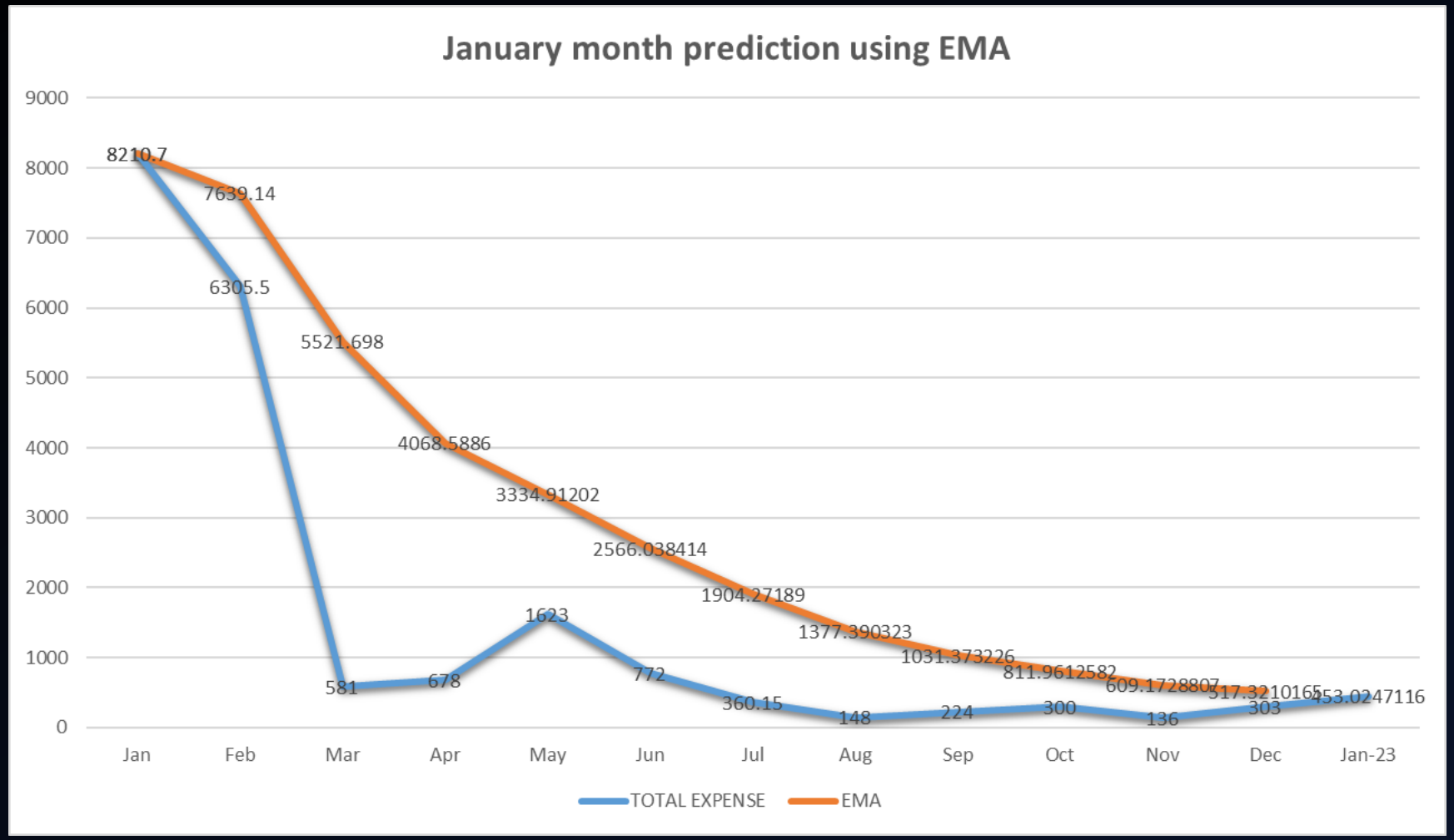


- The plot show that there is a statistically significant upward trend in expenses over the year, but the strength of the trend is weak. This indicates that while expenses tend to increase over time, other variables may also play a significant role in influencing spending.

PREDICTIVE ANALYSIS

Prediction of January 2023 using data imputation and EMA

- Data Imputation: Used average of August and October to fill missing September data.
- EMA Model: Employed Exponential Moving Average for prediction due to non-linearity.
- Evaluation: Essential to assess model performance using appropriate metrics.
- EMA assigns more weight to recent data points, making it more responsive to changes in the trend.
- EMA smooths out the data, reducing the impact of noise and outliers.
- EMA is used to forecast future values based on past data.



- The predicted value using EMA for Jan 2023 is 453.0247
- Trend: The overall trend of the "Total Expense" appears to be decreasing from January to December.
- EMA Smoothing: The EMA line is smoother than the actual "Total Expense" line, indicating that it reduces the impact of short-term fluctuations.
- Prediction: The prediction for January-23 is just above the December-22 value, suggesting a potential increase in "Total Expense" for the upcoming month.

Evaluating performance

MONTH	TOTAL EXPENSE	EMA	error	mse	rmse	mae	r^2
Jan	8210.7	8210.7	0	0	0	0	#DIV/0!
Feb	6305.5	7639.14	-1333.64	1778596	1333.64	1333.64	23.35079
Mar	581	5521.698	-4940.7	24410497	4940.698	4940.698	1.013805
Apr	678	4068.5886	-3390.59	11496091	3390.589	3390.589	1.039927
May	1623	3334.91202	-1711.91	2930643	1711.912	1711.912	1.898269
Jun	772	2566.038414	-1794.04	3218574	1794.038	1794.038	1.18493
Jul	360.15	1904.27189	-1544.12	2384312	1544.122	1544.122	1.05425
Aug	148	1377.390323	-1229.39	1511401	1229.39	1229.39	1.014395
Sep	224	1031.373226	-807.373	651851.5	807.3732	807.3732	1.076631
Oct	300	811.9612582	-511.961	262104.3	511.9613	511.9613	1.34223
Nov	136	609.1728807	-473.173	223892.6	473.1729	473.1729	1.082004
Dec	303	517.3210165	-214.321	45933.5	214.321	214.321	2.992141
Jan-23	453.0247116						

1. Error Analysis:
   - The "error" column shows that the EMA model's predictions deviate from the actual "Total Expense" at different times.
   - The "mse," "rmse," and "mae" metrics provide quantitative measures of the overall error. Lower values of these metrics indicate better model performance.
2. Model Fit:
   - The "r^2" value ranges from 1.0138 to 2.9921. This suggests that the EMA model explains a significant portion of the variance in the "Total Expense" data. However, it's important to note that R-squared can be misleading in certain cases, especially when the data is not linear or has outliers.

Based on the provided data, the EMA model appears to be a reasonable fit for predicting the "Total Expense." However, further analysis and evaluation are necessary to fully assess its performance and identify potential areas for improvement.

THANK YOU