Statistical Analysis: Used Car Price

Desy Ferizqa

Statistics for business project

Introduction

This report dives into used car prices, exploring how rapidly changing technology and consumer preferences have shaped this market. Our main goal is to understand the factors that contribute to pricing in the used car market.

Problem Statement

The problem at hand is the lack of a understanding of the factors contributing to used car pricing. Without a clear understanding of the significant variables that influence pre-owned vehicle prices, stakeholders in the market face difficulties in making informed decisions. Such as:

1. The buyer's lack of knowledge about fair prices in the market increases the possibility of purchasing a used car at an unfair price and not having a basis for negotiating the price.
2. Sellers struggle to set competitive prices for used cars in line with market conditions in order to attract buyers.
3. Lowering the effectiveness of financial institutions in assessing risk and collateral valuation for the purchase of used cars.

Using statistical analysis, we'll pinpoint the important variables that influence pre-owned vehicle prices, like mileage, age, brand, fuel type, and more. By examining patterns, correlations, and trends in the dataset, we'll draw meaningful conclusions and make predictions to help us navigate this dynamic market.

Dataset Description

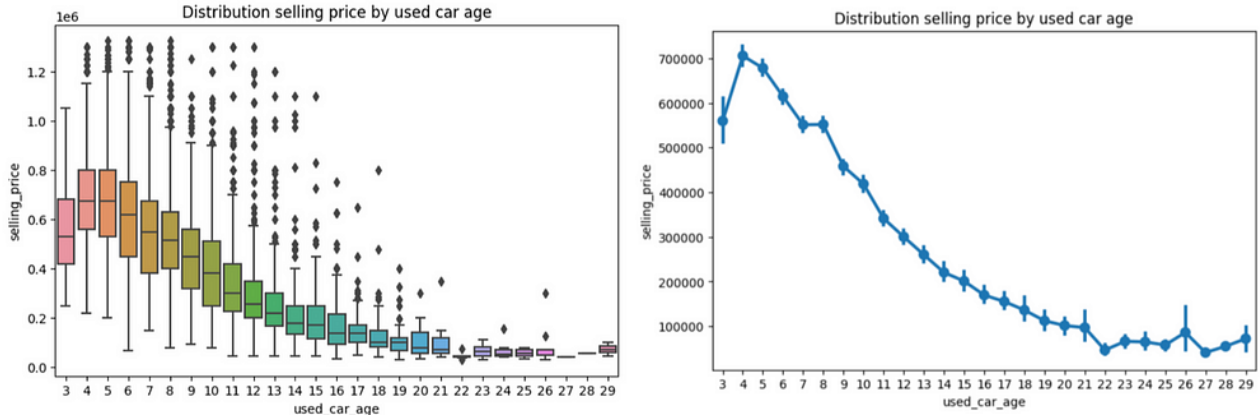
I'm using the dataset from Kaggle. This dataset is obtained from Car DekHo, an e-commerce platform for used cars and was last updated around 6 months ago (Jan 2023). This dataset contains information about :

- : Name of the cars
- : Year of the car when it was bought
- : Price at which car is being sold in Rupee
- : Number of Kilometres the car is driven
- : Fuel type of car (petrol / diesel / CNG / LPG / electric)
- : Tells if a Seller is Individual or a Dealer (Individual/Dealer/Trusted Dealer)
- : Gear transmission of the car (Automatic/Manual)
- : Number of previous owners of the car (First Owner/Second Owner/Third Owner/Fourth & Above Owner/Test Drive Car)
- : Amount of fuel used per km or per kg (for gas)
- : Engine capacity of the car
- : Brake horsepower of vehicle
- : Torque of vehicle
- : number of seats

Before we analyze further, I do some data cleaning and preparation. The process contains of data casting, creating new columns as new parameter (used_car_age), identify and handling outliers, dropping unused columns and so on. You can check the process more on my GitHub.

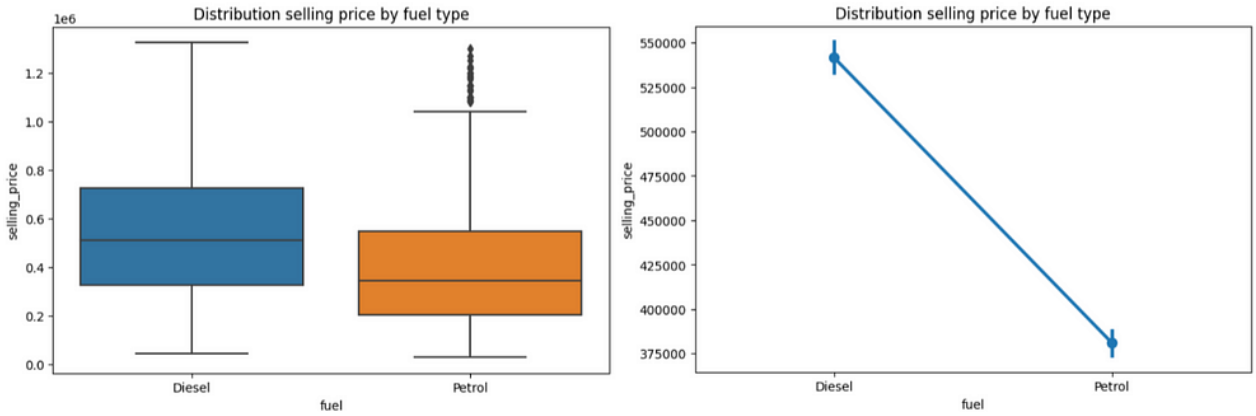
Data Exploration

We're going to see visualization of relationship between selling prices and others variable. Whether to see if a certain variable has higher or lesser selling price.



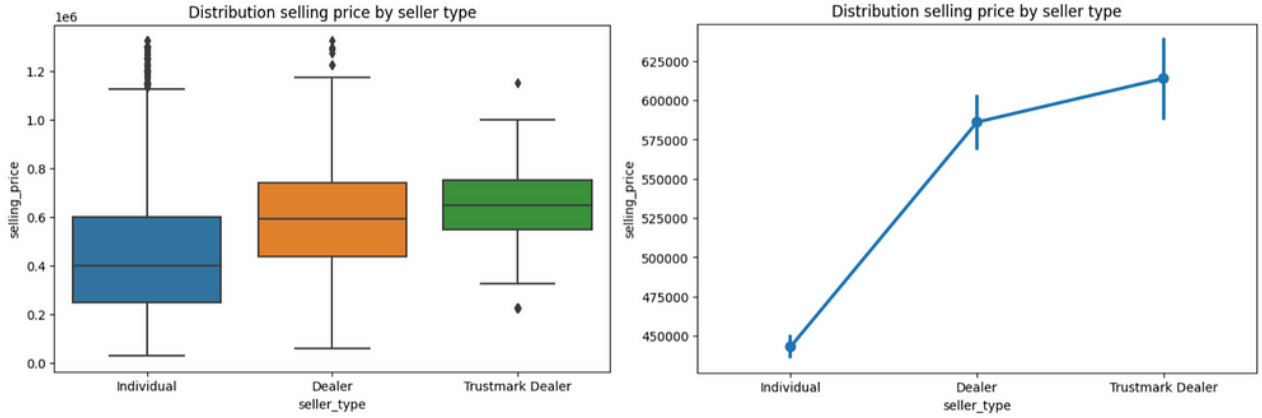
Picture 1. Selling price based on used car age

Based on the boxplot, we can see as the age of a vehicle increases, the selling price tends to decrease. This observation suggests that prospective buyers place significant value on newer models, possibly due to factors such as improved technology, better fuel efficiency, or overall desirability. Consequently, sellers should consider the age of their vehicles as a critical determinant when setting competitive prices in the used car market.



Picture 2. Selling price based on fuel type

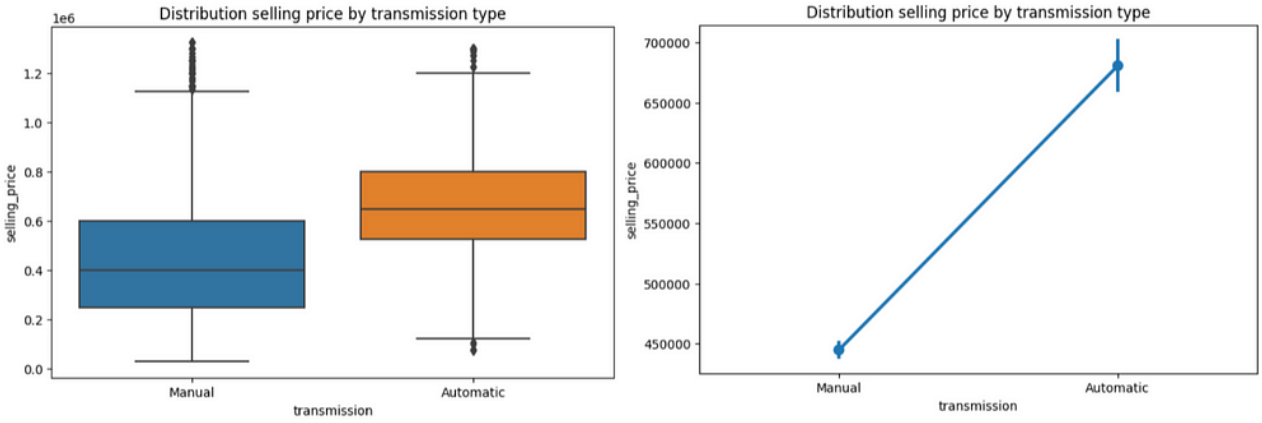
Diesel fuel cars has higher prices compared to petrol. This differences suggests that certain factors associated with diesel-fueled vehicles, such as their higher torque, fuel efficiency, or longer lifespan, contribute to their increased market value.



Picture 3. Selling price based on seller type

Vehicles sold by trustmark dealers has higher prices compared to those listed by individual sellers. This differences can be attributed to several factors, including the trust and credibility associated with dealerships, additional services provided, and potential warranties offered.

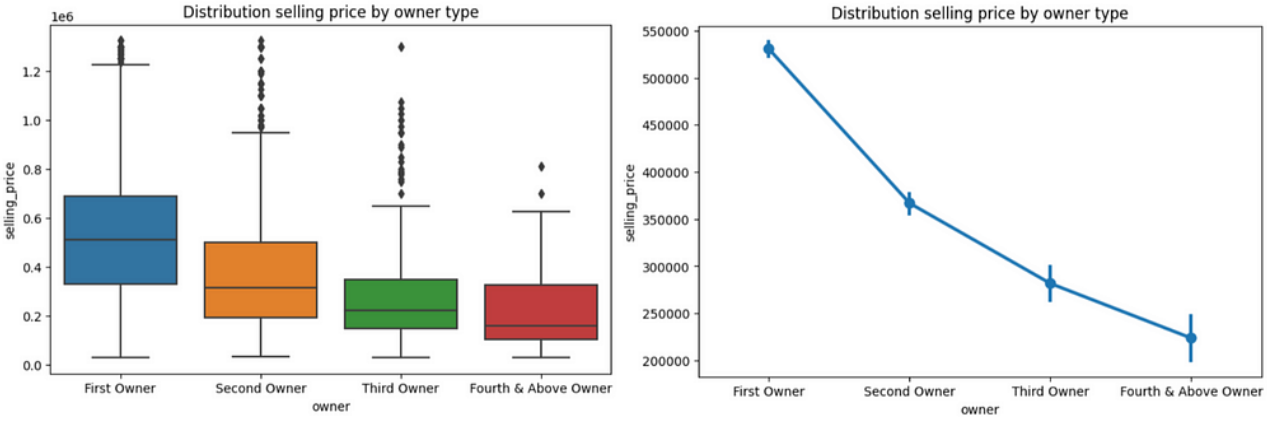
On the other hand, individual sellers tend to set relatively lower prices for their vehicles. However, we can see the presence of a higher number of outliers among individual sellers, indicating a wider range of pricing strategies. Despite the outliers, trustmark dealers price still higher.



Picture 4. Selling price based on transmission type

We can see vehicles with automatic transmission price is higher than manual

Vehicles equipped with automatic transmission tend to have higher prices compared to manual transmission. This price differences can be attributed to various factors, such as the convenience and ease of driving which may appeal to a broader range of buyers.



Picture 5. Selling price based on owner type

It becomes apparent that as the number of previous owners increases, the selling price of a vehicle tends to decrease. This price decrease can be attributed to factors such as increased wear and tear, potential maintenance issues, and a perceived lower level of trust associated with multiple ownership transfers. First-owner vehicles are likely to be in better overall condition, having experienced fewer usage years and potentially receiving regular maintenance and care.

Statistical Test

To solidify our earlier discoveries and provide robust statistical evidence, we will conduct hypothesis tests on selected variables.

T-test on selling price by fuel type

```
Hypothesis test result
sample size for diesel 3571
sample size for petrol 3240
----------------------------------------
Hypothesis:
H0 = Used car price with fuel type diesel <= Used car price with fuel type petrol
H1 = Used car price with fuel type diesel > Used car price with fuel type petrol
----------------------------------------
t-statistic:  25.62565462232831
p-value:  2.752725671450283e-138
alpha : 0.05
----------------------------------------
Based on t-test, it can be concluded:
Null hypothesis rejected
```

Picture 6. T-test on selling price by fuel type

Based on the test results above, the p-value obtained is <0.05, which means there is an average difference in the selling price of vehicle with fuel type diesel more higher than petrol.

T-test on selling price by transmission type

```
Hypothesis test result
sample size for manual 6258
sample size for automatic 553
----------------------------------------
Hypothesis:
H0 = Used car price with transmission type manual <= Used car price with type automatic
H1 = Used car price with transmission type automatic > Used car price with transmission type manual
----------------------------------------
t-statistic:  -21.018078899494657
p-value:  2.4232236210803974e-75
alpha : 0.05
----------------------------------------
Based on t-test, it can be concluded:
Reject H0
```

Picture 7. T-test on selling price by transmission type

Based on the test results above, the p-value obtained is <0.05, which means there is an average difference in the selling price of vehicle with transmision type automatic more higher than manual.

ANOVA test on selling price by seller type

```
Hypothesis:
H0 = There is no significant difference in the average selling_price between seller_type
H1 = There is a significant difference in the average selling_price between at least two seller_type
----------------------------------------------------------------------------------------------------
p-value:  4.8641946035899874e-57
alpha : 0.05
----------------------------------------------------------------------------------------------------
Based on ANOVA test, it can be concluded:
Reject H0
```

Picture 8. ANOVA test on selling price by seller type

Based on the test results above, the p-value obtained is <0.05, which means there is a significant difference in the selling price of vehicle between seller type.

ANOVA test on selling price by owner type

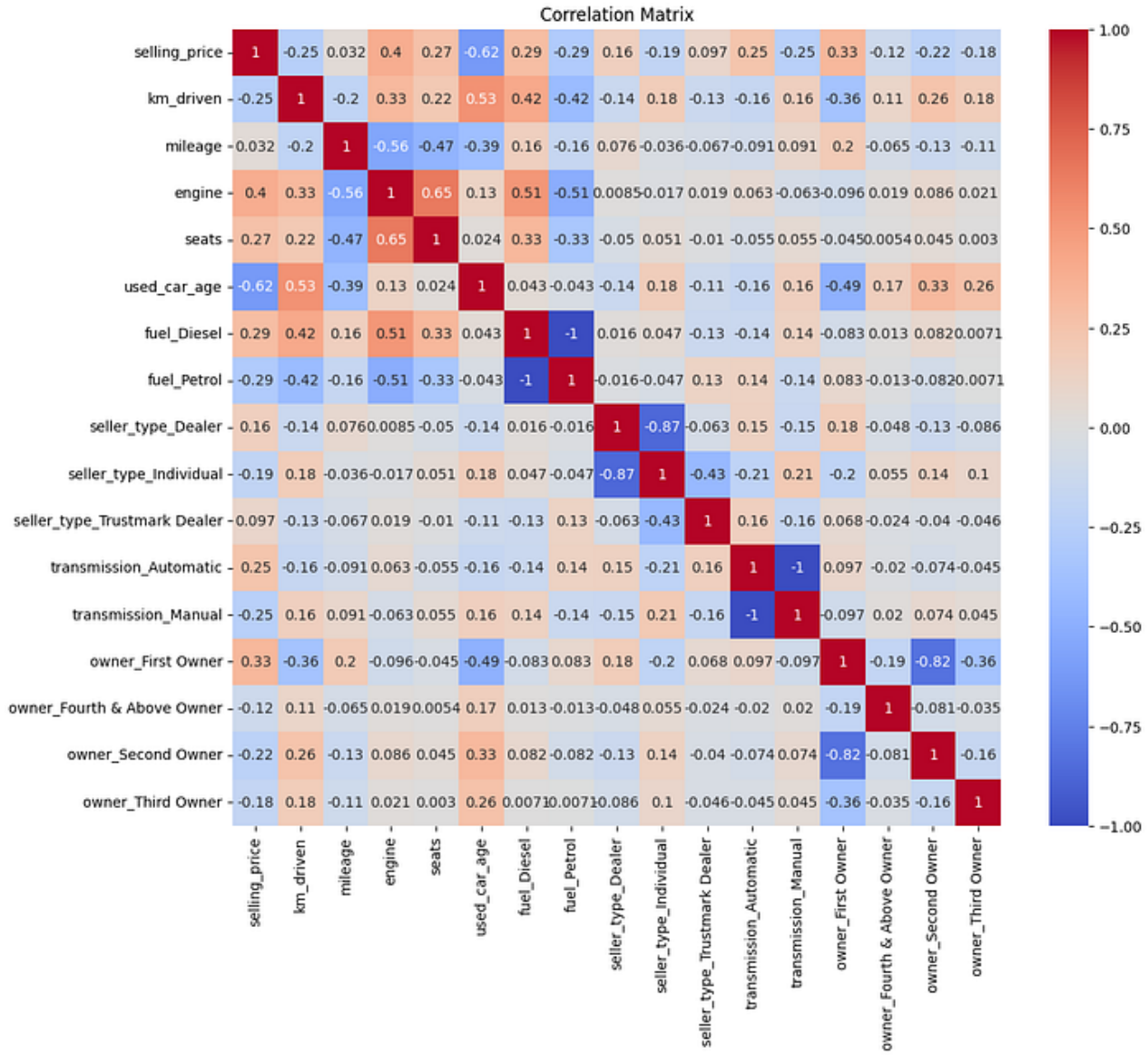
```
Hypothesis:
H0 = There is no significant difference in the average selling_price between owner type
H1 = There is a significant difference in the average selling_price between at least two owner type
-----------------------------------------------------------------------------------------------------
p-value:  4.0097115367599766e-185
alpha : 0.05
-----------------------------------------------------------------------------------------------------
Based on ANOVA test, it can be concluded:
Reject H0
```

Picture 9. ANOVA test on selling price by owner type

Based on the test results above, the p-value obtained is <0.05, which means there is a significant difference in the selling price of vehicle between owner type.

Correlation Matrix

We'll see correlation between numeric variable in the dataset.



Picture 11. Correlation matrix

We'll drop columns that has a relatively low correlation; "seats", "km_driven", "mileage", "seller_type_Dealer", "seller_type_Individual", "seller_type_Trustmark Dealer", "owner_Fourth & Above Owner", "owner_Second Owner", "owner_Third Owner".

Regression Model

We'll use linear regression analysis, using the predictor variables that were identified in the previous stage, with the ultimate goal of predicting house prices.

OLS Regression Results

Dep. Variable:	selling_price	R-squared:	0.644
Model:	OLS	Adj. R-squared:	0.644
Method:	Least Squares	F-statistic:	2465.
Date:	Wed, 12 Jul 2023	Prob (F-statistic):	0.00
Time:	01:13:18	Log-Likelihood:	-90960.
No. Observations:	6811	AIC:	1.819e+05
Df Residuals:	6805	BIC:	1.820e+05
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.951e+05	5259.083	56.122	0.000	2.85e+05	3.05e+05
engine	240.0664	5.154	46.582	0.000	229.964	250.169
used_car_age	-4.725e+04	634.006	-74.519	0.000	-4.85e+04	-4.6e+04
fuel_Diesel	1.831e+05	3975.711	46.049	0.000	1.75e+05	1.91e+05
fuel_Petrol	1.121e+05	2789.117	40.181	0.000	1.07e+05	1.18e+05
transmission_Automatic	2.105e+05	5026.883	41.870	0.000	2.01e+05	2.2e+05
transmission_Manual	8.468e+04	3648.290	23.210	0.000	7.75e+04	9.18e+04
owner_First Owner	3.477e+04	4465.364	7.786	0.000	2.6e+04	4.35e+04

Omnibus:	462.371	Durbin-Watson:	1.881
Prob(Omnibus):	0.000	Jarque-Bera (JB):	758.867
Skew:	0.531	Prob(JB):	1.64e-165
Kurtosis:	4.244	Cond. No.	4.31e+19

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 7.64e-30. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

Picture 12. OLS

From the results above, the Prob (F-Statistics) value is less than 0.05, so it can be concluded that our model is fit enough to use.

The P-value (P>|t|) shows that all coefficients have very small P-values (0.000), indicating that all predictor variables have a statistically significant impact on “selling_price”.

The R-squared value is 0.630, which means that about 63.0% of the variability in the dependent variable “selling_price” can be explained by the predictor variables used in this model.

Interpretation of the coefficient value of each predictor :

1. Coefficient “engine” has a value 240.0664. This means that every 1 unit increase in the “engine” variable (engine CC) is expected to lead to a 240.0664 unit increase in “selling_price”, while keeping the other predictor variables constant. In other words, the bigger the car’s engine type, the higher the expected used car “selling_price”.
2. Coefficient “used_car_age” has a value of -47,250. This means that each 1 year increase in the “used_car_age” variable is expected to lead to a 47,250 unit decrease in “selling_price”, while holding the other predictor variables constant. In other words, the older the used car is, the lower the expected “selling_price” of the used car will be.

3. Coefficients “fuel_Diesel” has a value of 183,100 and the “fuel_Petrol” has a value of 112,100 It shows the difference in used car prices based on fuel type. If used cars run on diesel fuel (fuel_Diesel), it is expected that “selling_price” will increase by 183,100 units compared to used cars running on petrol, while keeping other predictor variables constant.
4. Coefficients “transmission_Automatic” has a value of 210,500 and the “transmission_Manual” has a value of 84,680. It shows the difference in used car prices based on the type of transmission. If a used car uses an automatic transmission, it is expected that “selling_price” will increase by 210,500 units compared to a used car that uses a manual transmission, while keeping other predictor variables constant.
5. Coefficient “owner_First Owner” has a value of 34,770. This shows that used cars with first ownership (owner_First Owner) are expected to have a higher “selling_price” of 34,770 units compared to used cars with other ownership, while keeping other predictor variables constant.

Conclusion and Recommendation

Based on our analysis, we have identified several key factors that influence the selling price of used cars. The engine capacity, age of the vehicle, fuel type, transmission type, and ownership history all play significant roles in determining the selling price.

Based on these findings, we can make the following recommendations:

1. Sellers should consider the factors that positively impact selling price, such as larger engine sizes, diesel fuel, automatic transmission, and first ownership. By highlighting these features in their listings, sellers can potentially attract buyers who are willing to pay a premium.
2. Buyers should take into account the age of the vehicle when evaluating its price. Older cars tend to have lower prices, but it’s important to consider other factors like condition, maintenance history, and specific requirements.
3. Financial institutions should consider these factors when assessing risk and collateral valuation for used car purchases. Understanding the variables that impact pricing can help in making more accurate assessments and informed lending decisions

It should be noted that the OLS results also mention the possibility of a strong multicollinearity problem or a singular design matrix and shows the Prob(Omnibus) and Prob(JB) values are very small (<0.05), it indicates that there is an abnormality in the residual model.

This indicates that further evaluation and validation is needed on this model.

You can check my code .

Feel free to share your suggestions and critic so that I can improve in the future. Thank you!

References

investopedia.com