

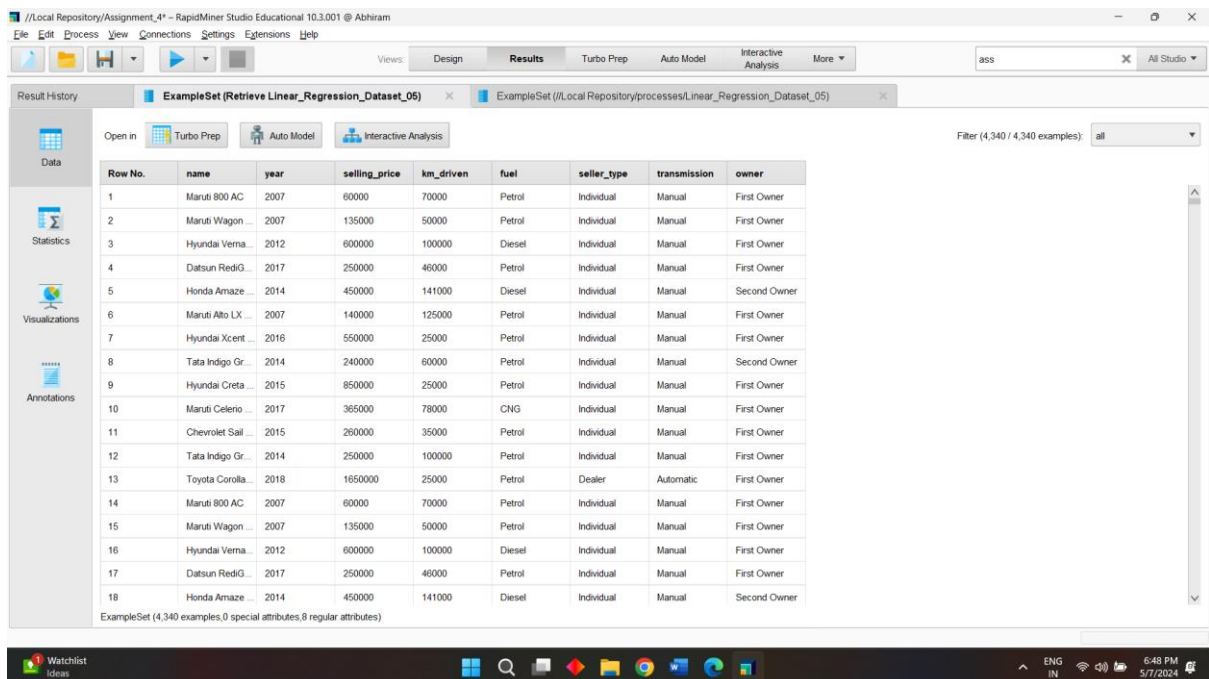
Task 1-4 needs to performed on the dataset assigned in the Question 1 (Linear Regression). (40 Marks)

1. Describing and understanding the data, data preprocessing and data cleaning - 10 marks

- Give a brief note on the dataset - what do you understand about the data.

The dataset displays information about the vehicle, including attributes such as name, year, selling price, Km driven, fuel type, seller type, transmission, and owner. The attribute 'name' indicates the identity of the vehicle, while 'year' indicates the year of manufacture. 'Sales price' gives us an idea of the economic value of a car, while 'Km driven' gives us an idea of usage. Attributes in categories such as 'fuel type', 'seller type', 'transmission', 'owner' may describe other attributes or attributes of a vehicle, such as fuel equipment, sales channel, transmission system, ownership history and overall, the data set Automotive prices of the internal automobile industry, which involve market trends and consumer behaviour are likely structured to facilitate the analysis and modelling task

- Provide the number of attributes and data records present in the dataset. List the variables and their corresponding data types.



The screenshot shows the RapidMiner Studio interface with a dataset table displayed. The table has 18 rows and 8 columns. The columns are: Row No., name, year, selling_price, km_driven, fuel, seller_type, transmission, and owner. The data represents various car models and their attributes.

Row No.	name	year	selling_price	km_driven	fuel	seller_type	transmission	owner
1	Maruti 800 AC	2007	60000	70000	Petrol	Individual	Manual	First Owner
2	Maruti Wagon	2007	135000	50000	Petrol	Individual	Manual	First Owner
3	Hyundai Verna	2012	600000	100000	Diesel	Individual	Manual	First Owner
4	Datsun RediGO	2017	250000	48000	Petrol	Individual	Manual	First Owner
5	Honda Amaze	2014	450000	141000	Diesel	Individual	Manual	Second Owner
6	Maruti Alto LX	2007	140000	125000	Petrol	Individual	Manual	First Owner
7	Hyundai Xcent	2016	550000	25000	Petrol	Individual	Manual	First Owner
8	Tata Indigo Gr	2014	240000	60000	Petrol	Individual	Manual	Second Owner
9	Hyundai Creta	2015	850000	25000	Petrol	Individual	Manual	First Owner
10	Maruti Celerio	2017	365000	78000	CNG	Individual	Manual	First Owner
11	Chevrolet Sail	2015	260000	35000	Petrol	Individual	Manual	First Owner
12	Tata Indigo Gr	2014	250000	100000	Petrol	Individual	Manual	First Owner
13	Toyota Corolla	2018	1650000	25000	Petrol	Dealer	Automatic	First Owner
14	Maruti 800 AC	2007	60000	70000	Petrol	Individual	Manual	First Owner
15	Maruti Wagon	2007	135000	50000	Petrol	Individual	Manual	First Owner
16	Hyundai Verna	2012	600000	100000	Diesel	Individual	Manual	First Owner
17	Datsun RediGO	2017	250000	48000	Petrol	Individual	Manual	First Owner
18	Honda Amaze	2014	450000	141000	Diesel	Individual	Manual	Second Owner

In RapidMiner, the shape of the dataset is (number of instances, number of attributes) So, this csv file has 4340 Examples, and 8 Regular Attributes. So, shape is (4340,8).

Result History

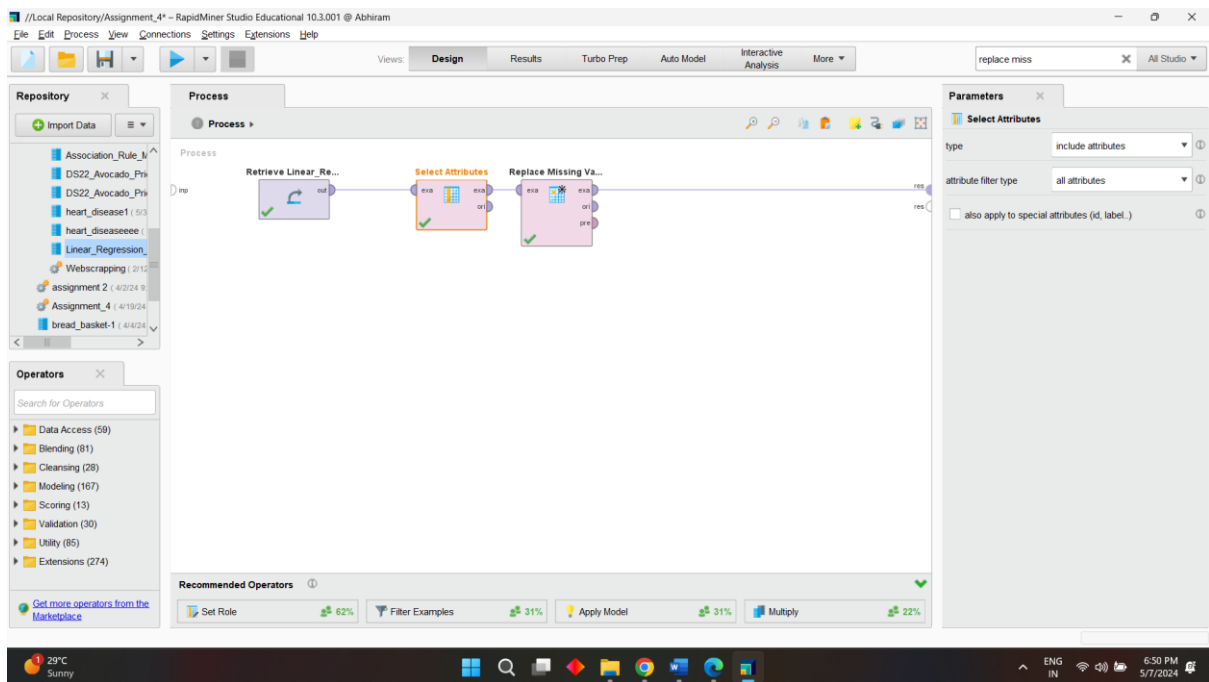
ExampleSet (Retrieve Linear_Regression_Dataset_05)

Name	Type	Missing	Statistics	Filter (8 / 8 attributes)
name	Nominal	0	Volvo XC [...] ption (1)	Most: Maruti S [...] VDI (69) Values: Maruti Swift Dzire VDI (69), Maruti Alto 800 LXI (59), ...[1489 more]
year	Integer	0	Min: 1992 Max: 2020 Average: 2013.091	
selling_price	Integer	0	Min: 20000 Max: 89000000 Average: 504127.312	
km_driven	Integer	0	Min: 1 Max: 806599 Average: 66215.777	
fuel	Nominal	0	Least: Electric (1) Most: Diesel (2153)	Values: Diesel (2153), Petrol (2123), ...[3 more]
seller_type	Nominal	0	Least: Trustmark Dealer (102) Most: Individual (3244)	Values: Individual (3244), Dealer (994), ...[1 more]
transmission	Nominal	0	Least: Automatic (448) Most: Manual (3892)	Values: Manual (3892), Automatic (448)
owner	Nominal	0	Least: Test Drive Car (17) Most: First Owner (2632)	Values: First Owner (2632), Second Owner (1106), ...[3 more]

Showing attributes 1 - 8

Examples: 4,340 Special Attributes: 0 Regular Attributes: 8

- Check for missing values and replace them accordingly as per your understanding.



ResultHistory ExampleSet (Replace Missing Values)

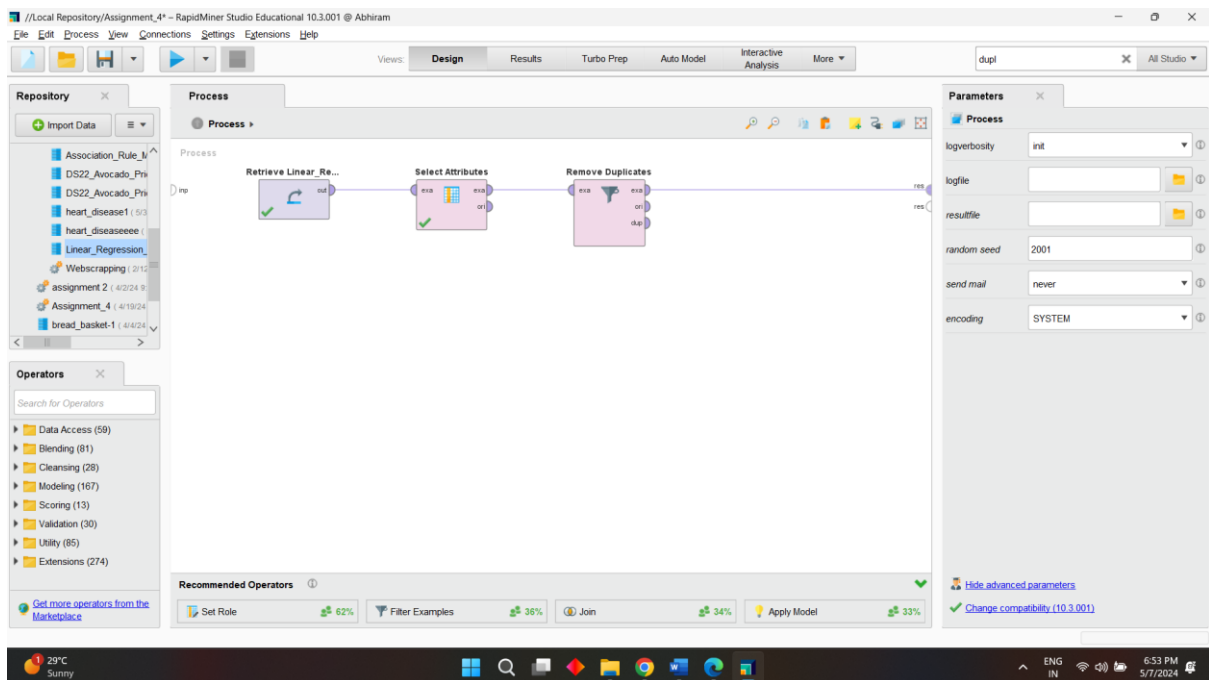
Name	Type	Missing	Statistics	Filter (8 / 8 attributes)
name	Polynomial	0	Least: Volvo XC [...] ption (1) Most: Maruti S [...] VDI (69)	Values: Maruti Swift Dzire VDI (69), Maruti Alto 800 LXI (59), ...[1489 more]
year	Integer	0	Min: 1992 Max: 2020 Average: 2013.091	
selling_price	Integer	0	Min: 20000 Max: 8900000 Average: 504127.312	
km_driven	Integer	0	Min: 1 Max: 806599 Average: 66215.777	
fuel	Polynomial	0	Least: Electric (1) Most: Diesel (2153)	Values: Diesel (2153), Petrol (2123), ...[3 more]
seller_type	Polynomial	0	Least: Trustmark Dealer (102) Most: Individual (3244)	Values: Individual (3244), Dealer (994), ...[1 more]
transmission	Polynomial	0	Least: Automatic (448) Most: Manual (3892)	Values: Manual (3892), Automatic (448)
owner	Polynomial	0	Least: Test Drive Car (17) Most: First Owner (2832)	Values: First Owner (2832), Second Owner (1106), ...[3 more]

Showing attributes 1 - 8

Examples: 4,340 Special Attributes: 0 Regular Attributes: 8

After applying Remove missing values also the shape of the dataset is the same. So, no missing values. Before and after missing values are zero.

- Check for any duplicate records, list the duplicate records if any and remove them.



ExampleSet (Remove Duplicates)

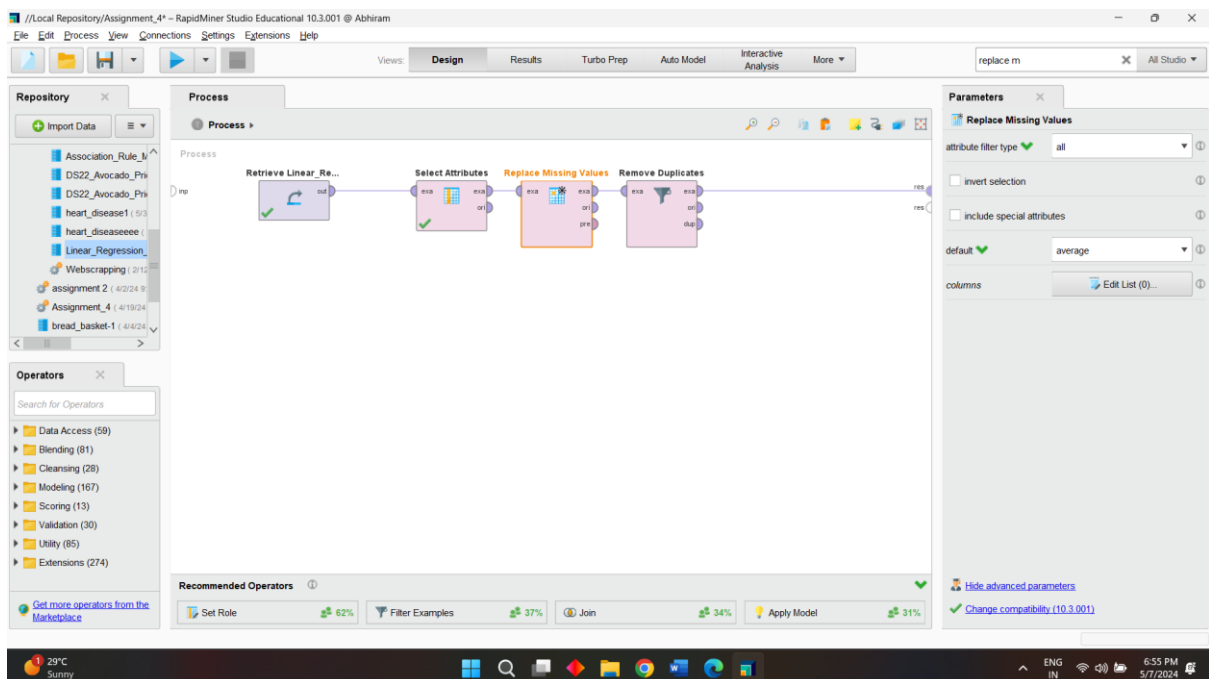
Filter (3,577 / 3,577 examples): all

Row No.	name	year	selling_price	km_driven	fuel	seller_type	transmission	owner
1	Maruti 800 AC	2007	60000	70000	Petrol	Individual	Manual	First Owner
2	Maruti Wagon	2007	135000	50000	Petrol	Individual	Manual	First Owner
3	Hyundai Verna	2012	600000	100000	Diesel	Individual	Manual	First Owner
4	Datsun RediGo	2017	250000	46000	Petrol	Individual	Manual	First Owner
5	Honda Amaze	2014	450000	141000	Diesel	Individual	Manual	Second Owner
6	Maruti Alto LX	2007	140000	125000	Petrol	Individual	Manual	First Owner
7	Hyundai Xcent	2016	550000	25000	Petrol	Individual	Manual	First Owner
8	Tata Indigo Gr	2014	240000	60000	Petrol	Individual	Manual	Second Owner
9	Hyundai Creta	2015	850000	25000	Petrol	Individual	Manual	First Owner
10	Maruti Celerio	2017	365000	78000	CNG	Individual	Manual	First Owner
11	Chevrolet Sail	2015	260000	35000	Petrol	Individual	Manual	First Owner
12	Tata Indigo Gr	2014	250000	100000	Petrol	Individual	Manual	First Owner
13	Toyota Corolla	2018	1650000	25000	Petrol	Dealer	Automatic	First Owner
14	Maruti Ciaz VX	2015	585000	24000	Petrol	Dealer	Manual	First Owner
15	Hyundai Venue	2019	1195000	5000	Diesel	Dealer	Manual	First Owner
16	Chevrolet Enjo	2013	390000	33000	Diesel	Individual	Manual	Second Owner
17	Jaguar XF 2.2	2014	1964999	28000	Diesel	Dealer	Automatic	First Owner
18	Mercedes-Ben	2013	1425000	59000	Diesel	Dealer	Automatic	First Owner

ExampleSet (3,577 examples, 0 special attributes, 8 regular attributes)

Previously the records are 4340, after applying remove duplicates the records are 3577. So, 763 are the duplicate records.

- Indicate the change in the shape of the dataset after replacing missing values and removing duplicates.



Result History: ExampleSet (Remove Duplicates)

Open in: Turbo Prep, Auto Model, Interactive Analysis

Filter (3,577 / 3,577 examples): all

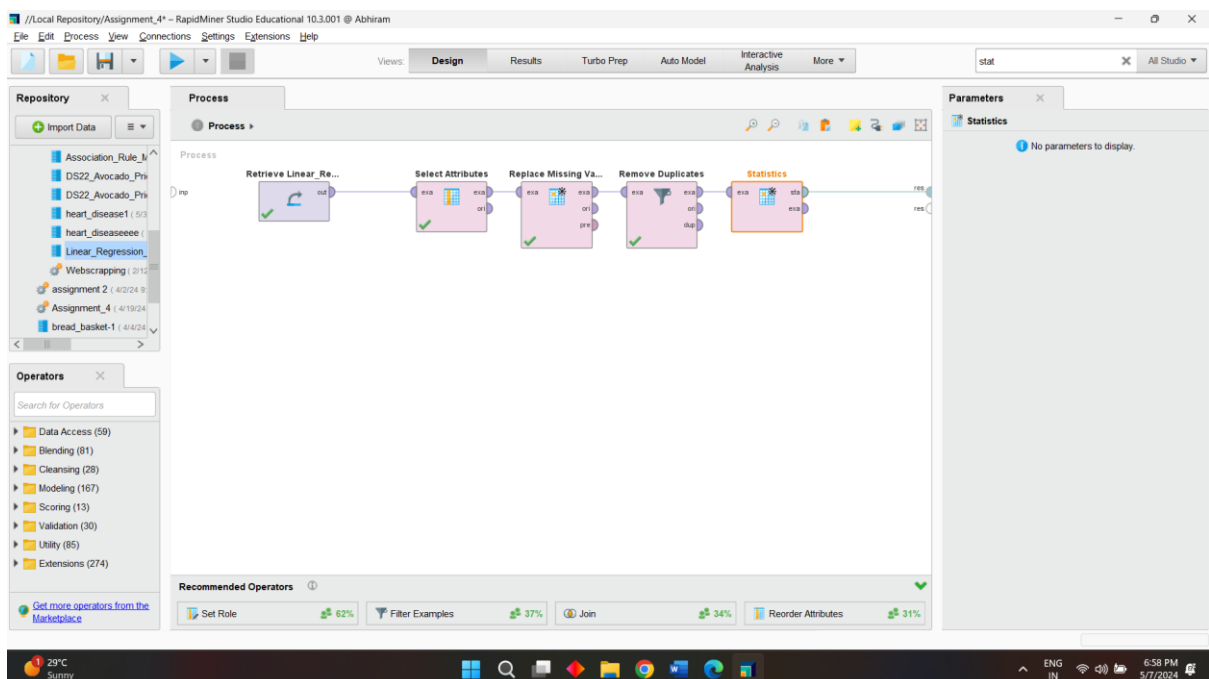
Row No.	name	year	selling_price	km_driven	fuel	seller_type	transmission	owner
1	Maruti 800 AC	2007	60000	70000	Petrol	Individual	Manual	First Owner
2	Maruti Wagon	2007	135000	50000	Petrol	Individual	Manual	First Owner
3	Hyundai Verna	2012	600000	100000	Diesel	Individual	Manual	First Owner
4	Datsun RediGo	2017	250000	46000	Petrol	Individual	Manual	First Owner
5	Honda Amaze	2014	450000	141000	Diesel	Individual	Manual	Second Owner
6	Maruti Alto LX	2007	140000	125000	Petrol	Individual	Manual	First Owner
7	Hyundai Xcent	2016	550000	25000	Petrol	Individual	Manual	First Owner
8	Tata Indigo Gr	2014	240000	60000	Petrol	Individual	Manual	Second Owner
9	Hyundai Creta	2015	850000	25000	Petrol	Individual	Manual	First Owner
10	Maruti Celerio	2017	365000	78000	CNG	Individual	Manual	First Owner
11	Chevrolet Sail	2015	260000	35000	Petrol	Individual	Manual	First Owner
12	Tata Indigo Gr	2014	250000	100000	Petrol	Individual	Manual	First Owner
13	Toyota Corolla	2018	1650000	25000	Petrol	Dealer	Automatic	First Owner
14	Maruti Ciaz VX	2015	585000	24000	Petrol	Dealer	Manual	First Owner
15	Hyundai Venue	2019	1195000	5000	Diesel	Dealer	Manual	First Owner
16	Chevrolet Enjo	2013	390000	33000	Diesel	Individual	Manual	Second Owner
17	Jaguar XF 2.2	2014	1964999	28000	Diesel	Dealer	Automatic	First Owner
18	Mercedes-Ben	2013	1425000	59000	Diesel	Dealer	Automatic	First Owner

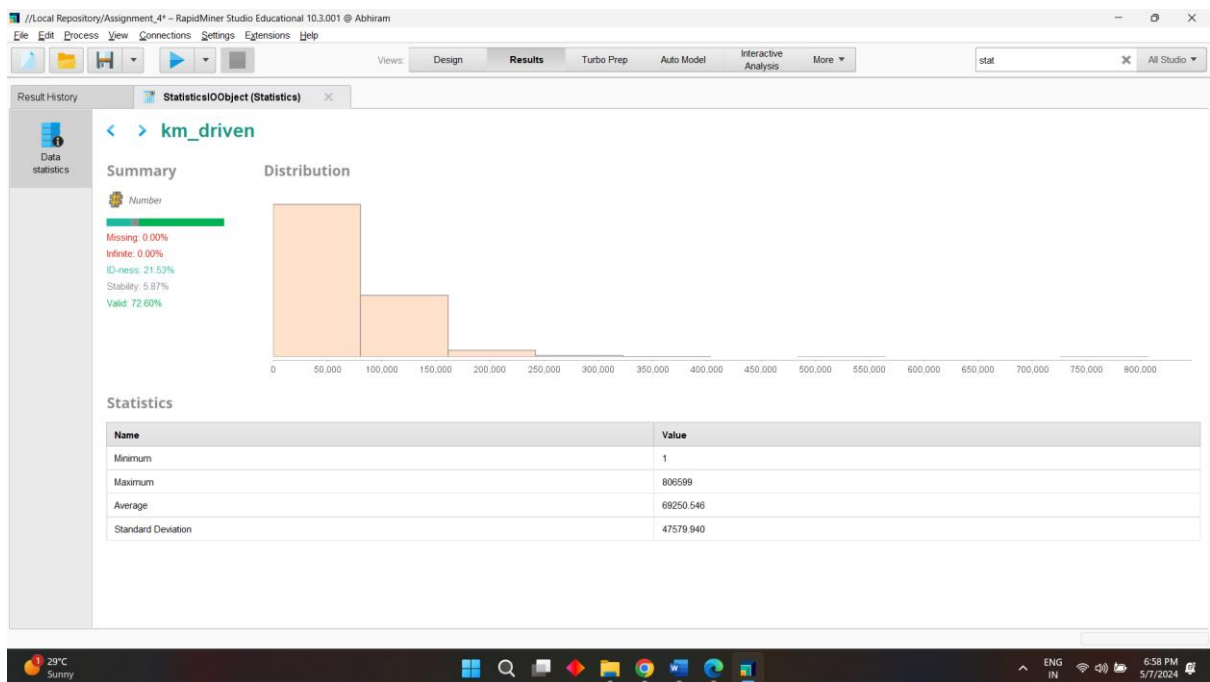
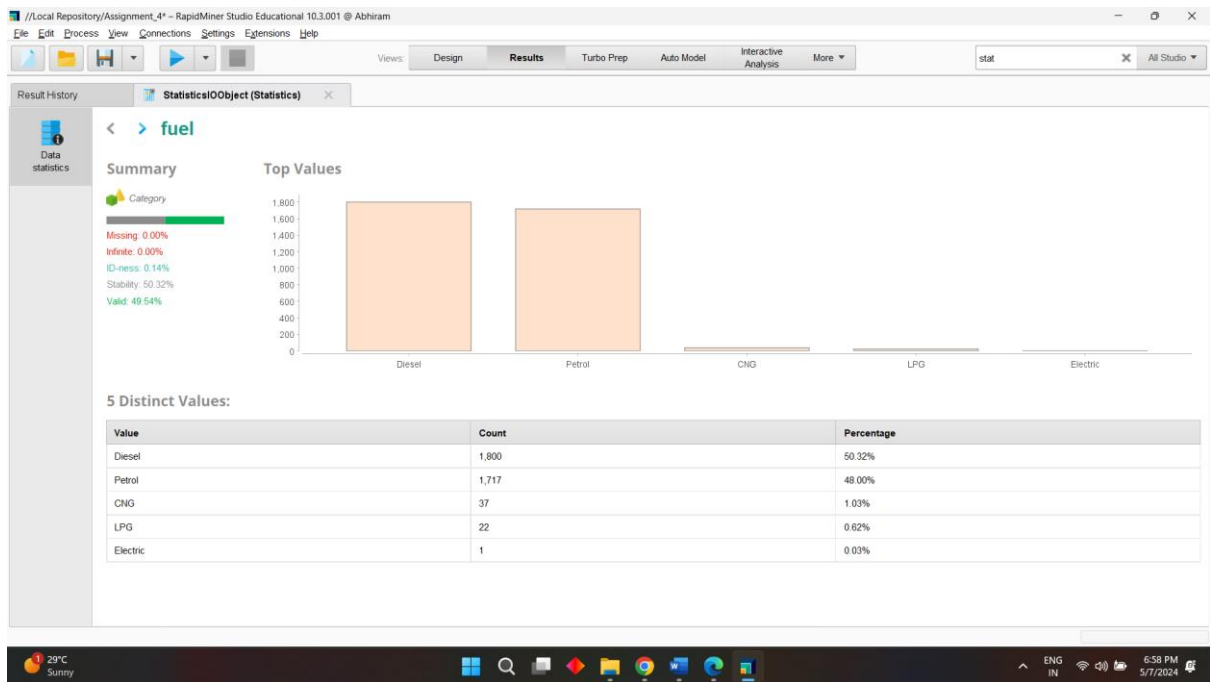
ExampleSet (3,577 examples, 0 special attributes, 8 regular attributes)

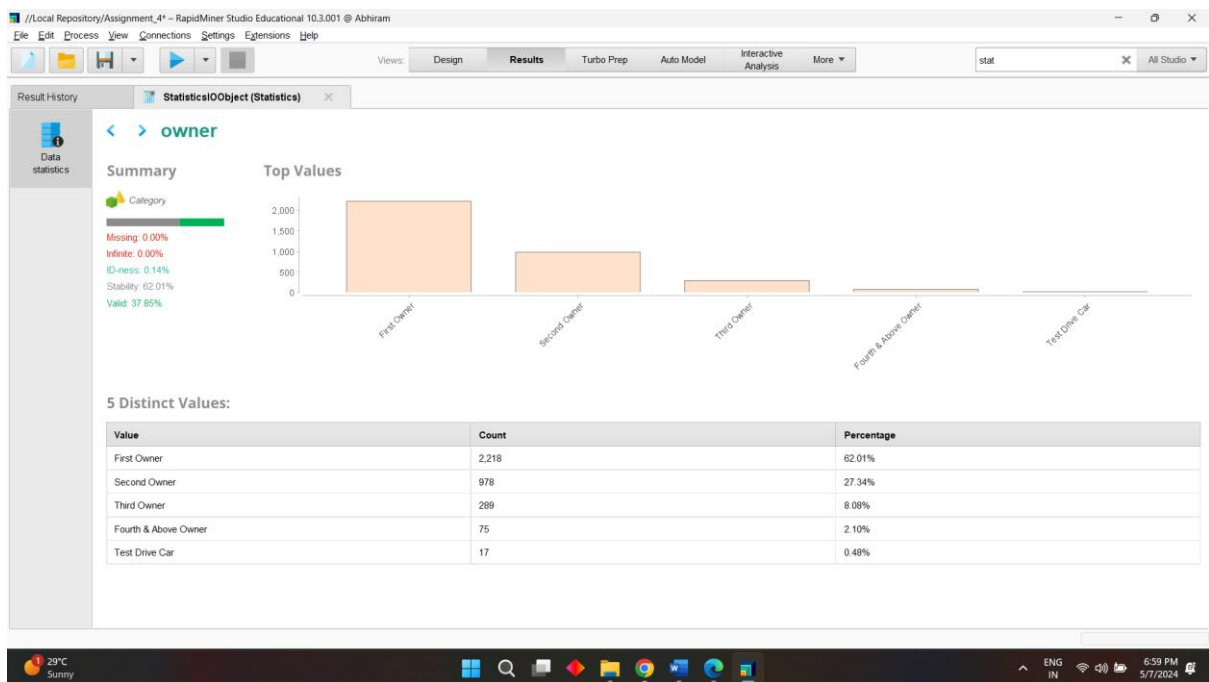
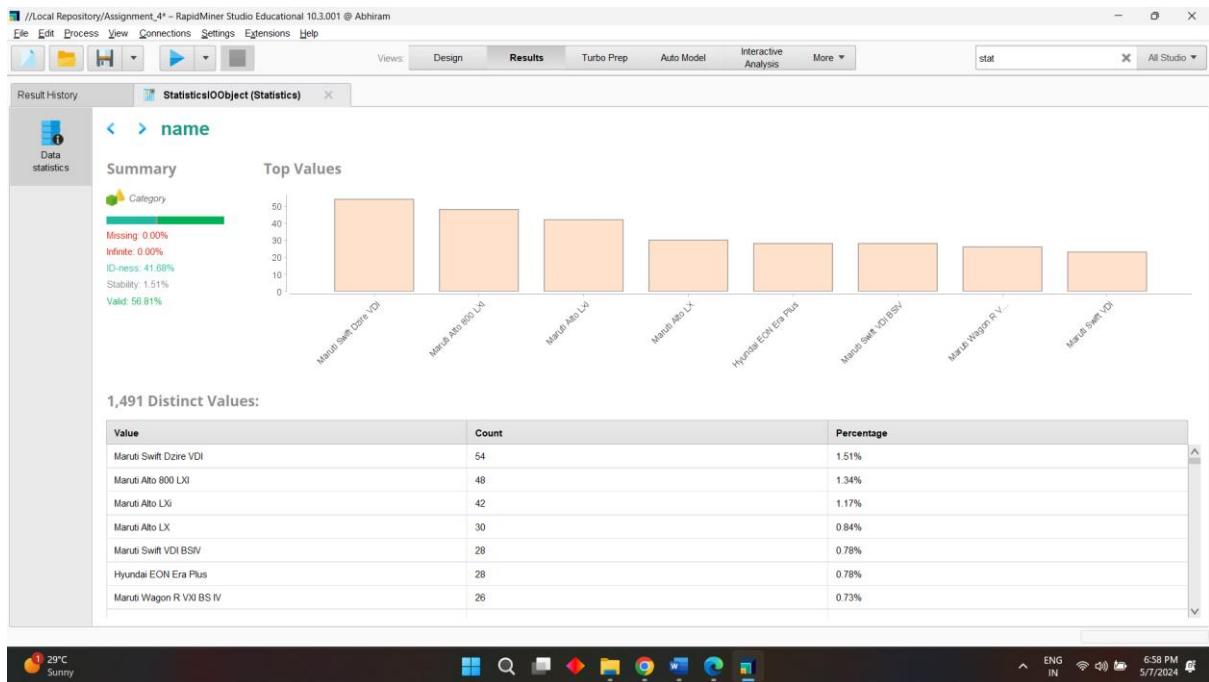
After removing duplicates and missing values the data set shape is 3577 records, 8 regular attributes and 0 special attributes.

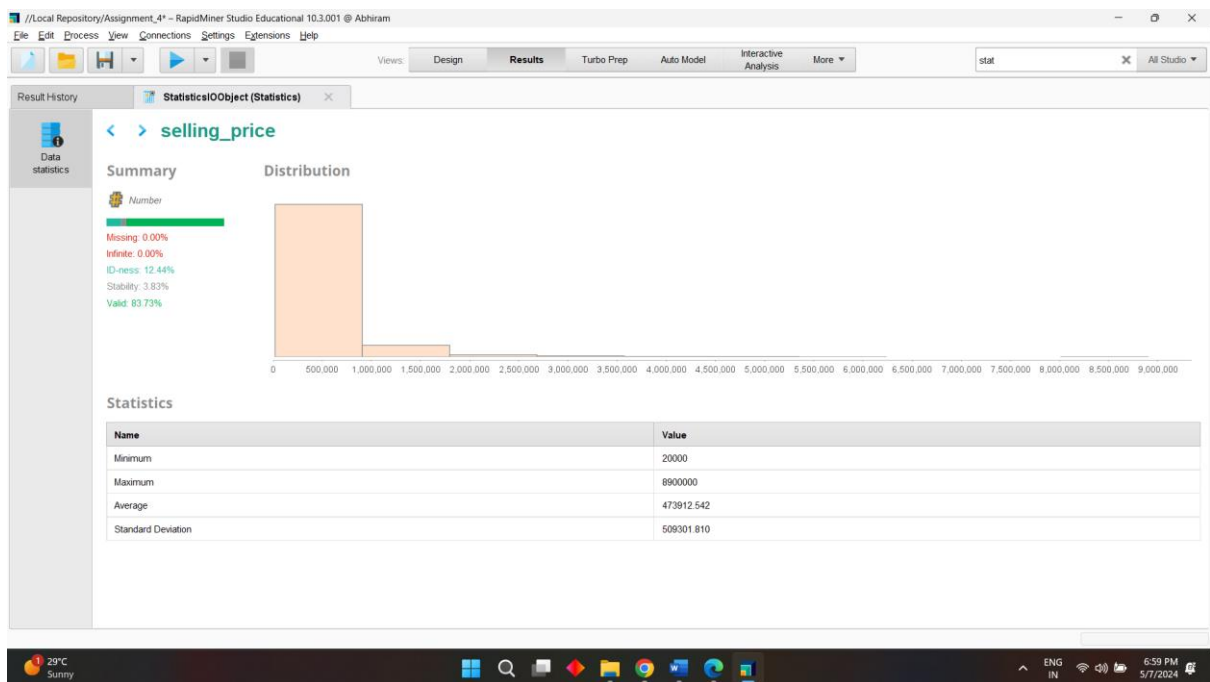
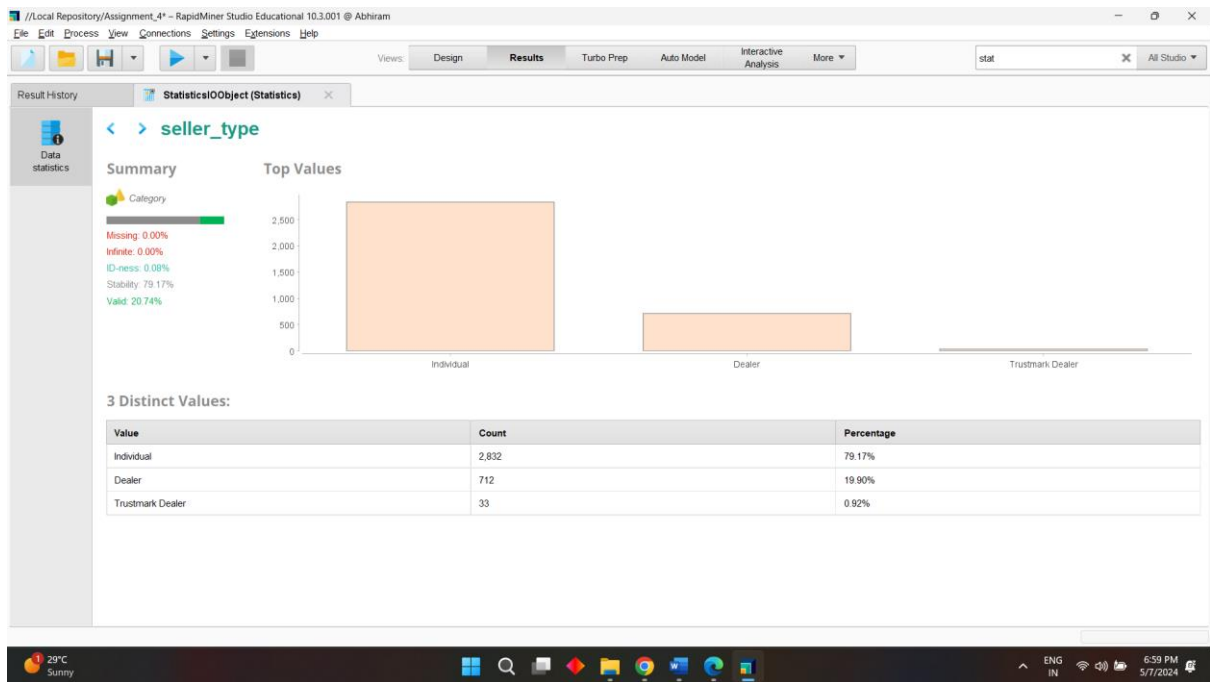
2. Data quality checking and evaluation - 10 marks

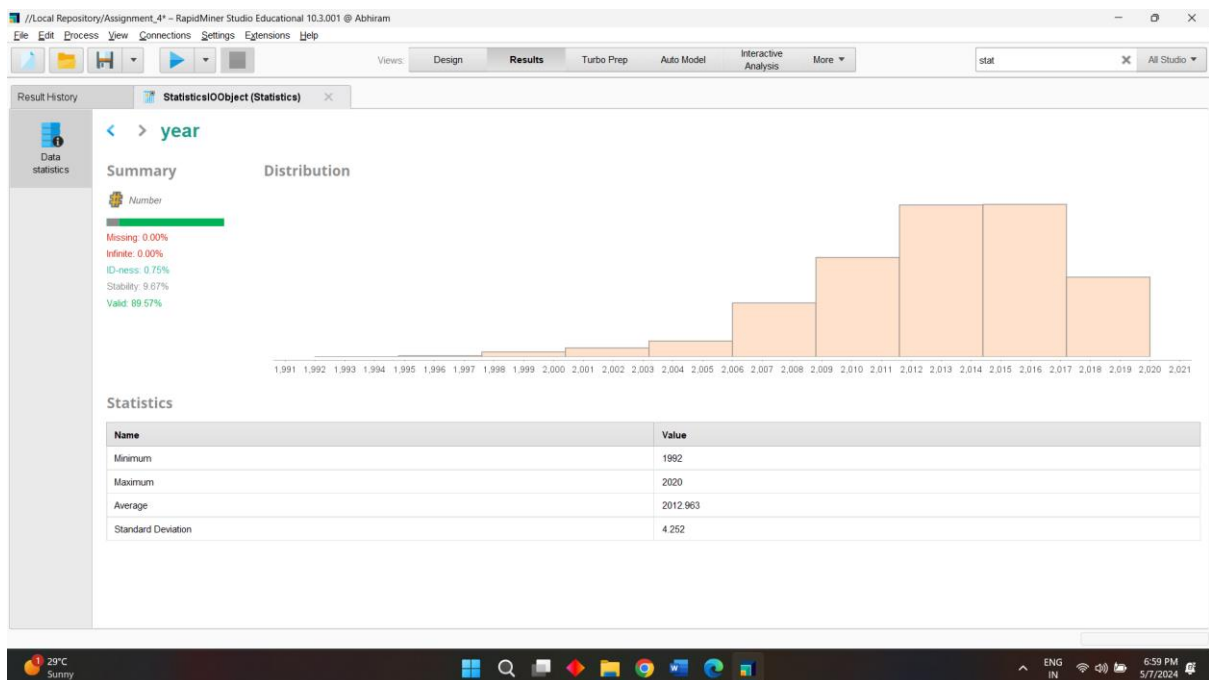
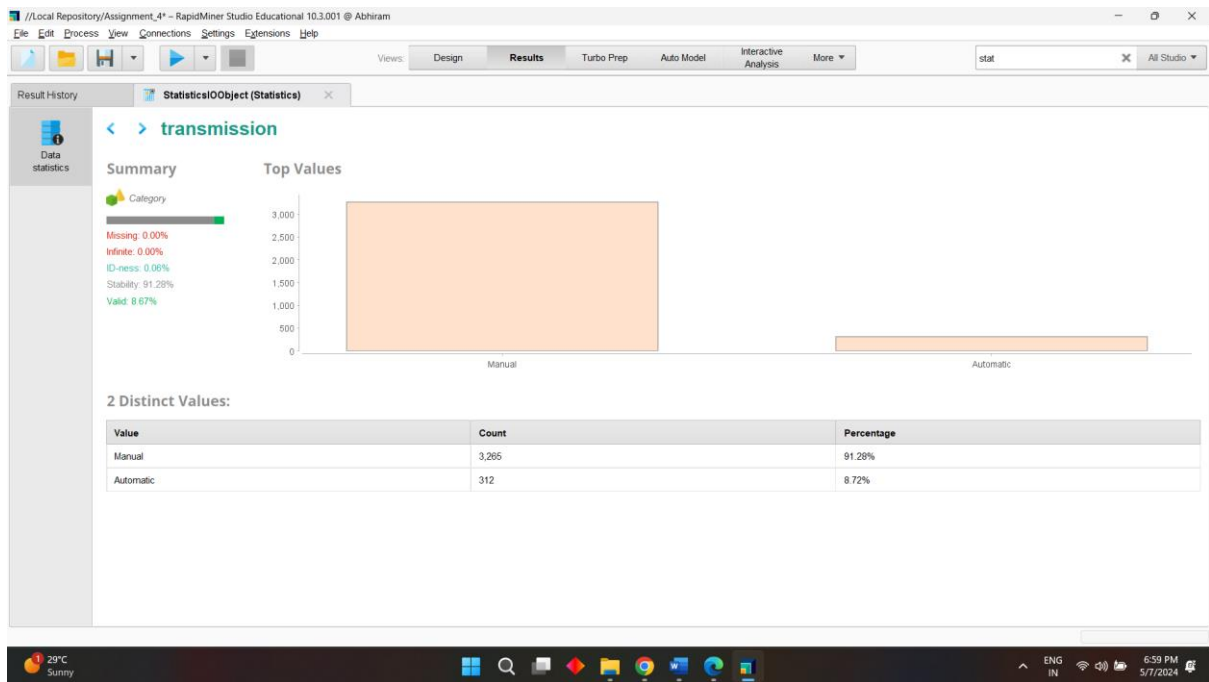
- Check for biases if any, look for outliers in data and comment how to deal with it (should it be kept or be removed from the dataset). Is the dataset good enough to go ahead?
- You can use the operator in RapidMiner or use statistics to comment on this. *No need of staying long if consuming time.*











Biases: Based on the data, there is no obvious bias in the dataset. Reasoning and conclusions appear to vary, and there is no indication that certain groups are disproportionately represented. However, additional research or domain knowledge may be required to fully confirm these findings.

Outliers: Based on the statistical staff results, there is no clear indication of outliers in the data set. There are some outliers that stand out significantly from the rest of the data set and can skew research results. The fact that no outliers were mentioned or observed suggests that the dataset is relatively clean and free of extreme values.

Data set quality: Based on the results of the audit staff, the quality of the data set appears to be good. Support, trust, lift, profitability, and other metrics show that there is a meaningful

relationship between location and findings. Additionally, the stability of the dataset indicates that the associations are stable and reliable under different circumstances.

Considering these factors, we can conclude that the data set is good with respect to bias, outliers, and overall quality of the data set. However, it is always necessary to take such precautions or careful examinations as are necessary to ensure the reliability and accuracy of the findings.

3. Correlation Matrix and insights from it about the attributes - 10 marks

- **How can you use a correlation matrix to inform feature selection in a machine learning context for the data?**

In machine learning, the correlation matrix is a powerful selection tool by revealing the relationships among features in a data set. Examination of the correlation matrix can identify highly correlated elements, indicating a wealth of information. Removing redundant features not only simplifies the model but also reduces the risk of overfitting and improves computational efficiency.

In addition, it is important to retain those with strong correlations with the target variables while removing irrelevant factors, as they contribute significantly to the accuracy of the forecast. Factors that exhibit high correlations with the target variables are often valuable predictors, increasing the ability of the model to predict accurately. However, it is important to focus on multicollinearity, where three or more variables are highly correlated, as it may lead to incorrect model estimates. Addressing multicollinearity by removing or replacing correlated factors ensures model stability and facilitates interpretation.

The selection process guided by the correlation matrix is iterative, applying different sub-factors and investigating their effect on model performance. Mapping correlations through heat maps helps facilitate capture patterns and make informed product selection decisions.

In summary, the advantages of correlation matrix over feature selection optimize feature space, increase pattern interpretation capabilities, and improve prediction performance in machine learning tasks.

- Mention the highly correlated attributes.

The screenshot displays the RapidMiner Studio interface. The top panel shows a workflow with the following operators: Retrieve Linear Regression, Select Attributes, Replace Missing Values, Remove Duplicates, and Correlation Matrix. The right panel shows the parameters for the Correlation Matrix operator, including 'type' (include attributes), 'attribute filter type' (all attributes), 'normalize weights' (checked), and 'squared correlation' (unchecked).

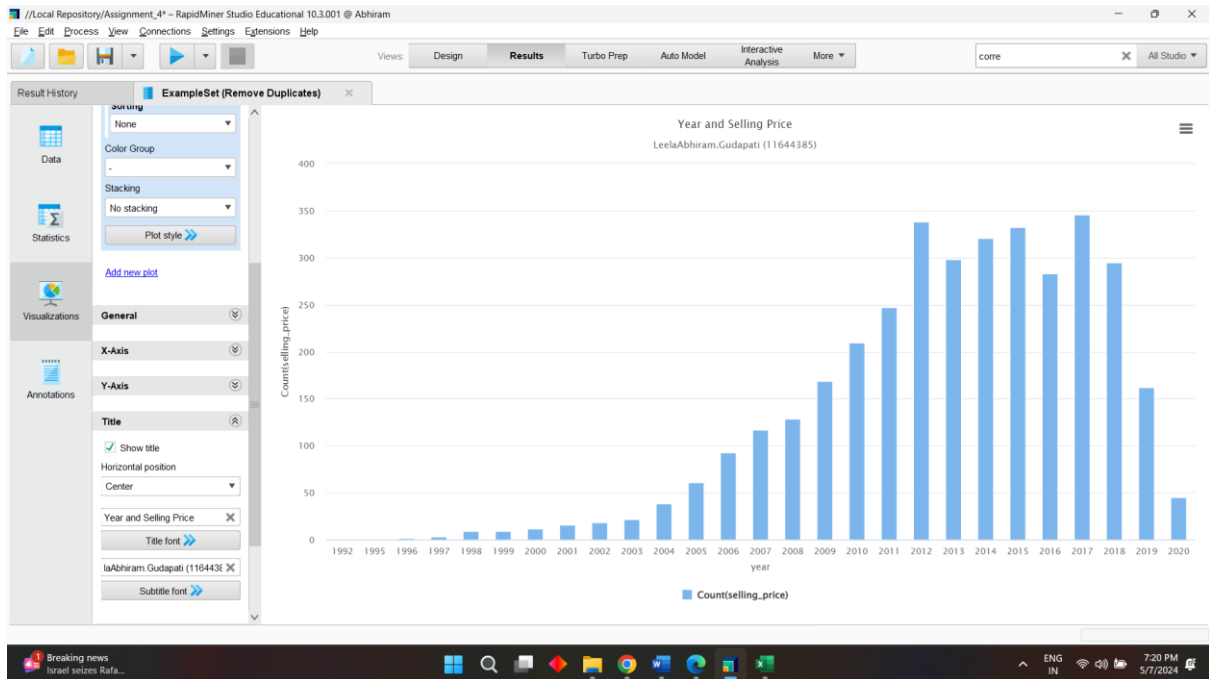
The bottom panel shows the 'Results' view with a table titled 'Correlation Matrix (Correlation Matrix)'. The table displays the correlation coefficients between various attributes. The attributes listed are name, year, selling_price, km_driven, fuel, seller_type, transmission, and owner. The correlation coefficients are as follows:

Attributes	name	year	selling_price	km_driven	fuel	seller_type	transmission	owner
name	1	?	?	?	?	?	?	?
year	?	1	0.424	-0.417	?	?	0.117	?
selling_price	?	0.424	1	-0.187	?	?	0.486	?
km_driven	?	-0.417	-0.187	1	?	?	-0.101	?
fuel	?	?	?	?	1	?	?	?
seller_type	?	?	?	?	?	1	?	?
transmission	?	0.117	0.486	-0.101	?	?	1	?
owner	?	?	?	?	?	?	?	1

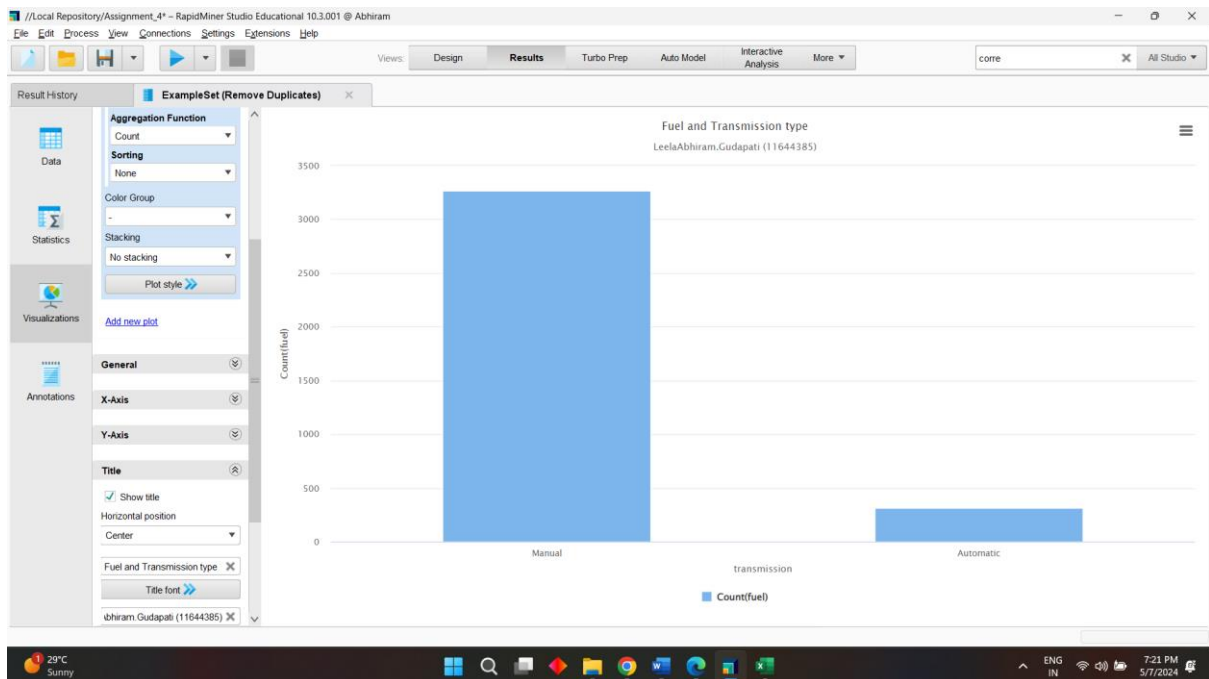
There does not appear to be any strong positive correlation between any of the traits. There is a moderate positive correlation between year and selling price (0.107), but none of the other correlations are close to 1. There is a moderate negative correlation between km driven and selling price (-0.187). A negative correlation means that as one value increases, the other decreases. In this case, as the Km driven increases (i.e. the vehicle is used more), the selling price decreases, which is what we expected. It is important to note that, as in this data set, correlation coefficients close to 0 indicate that there is no relationship between the two variables indicating that changes in one variable are associated with predictable changes in the other are not connected.

4. Data Visualization (at least 3-4 depending on the dataset with explanation) - 10 mark

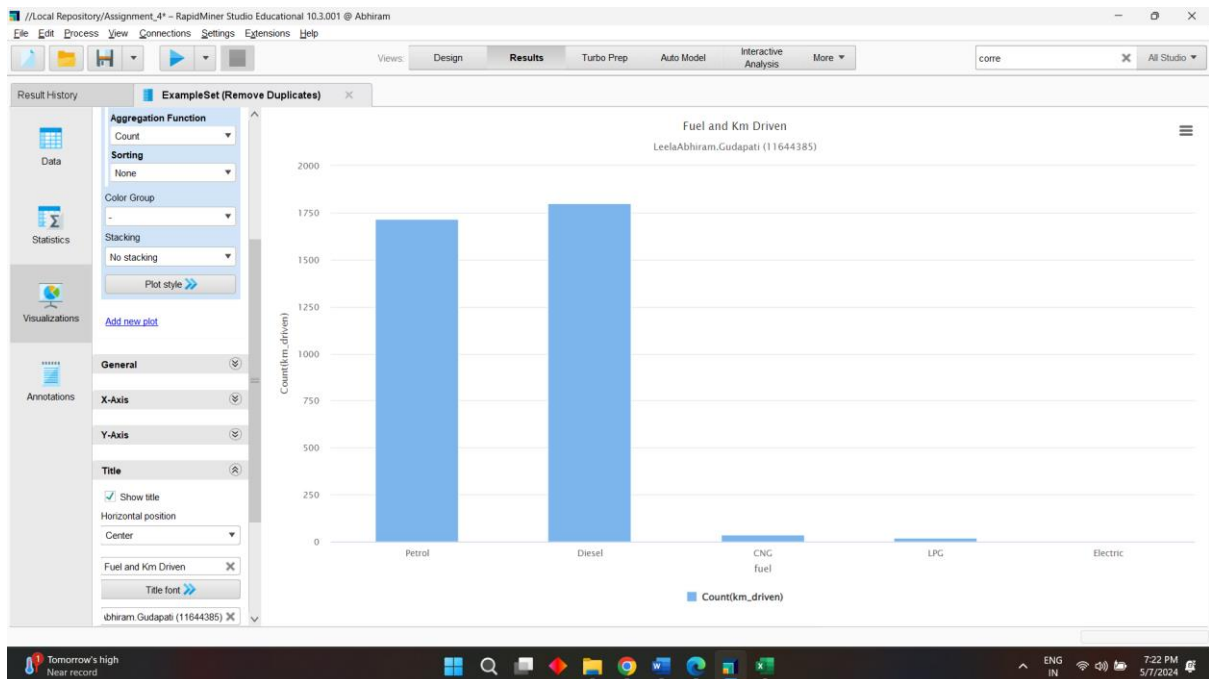
- Give titles and labels for each visualization. Sub title of each visualization should be your name and student ID.
- Provide insights about the visualizations you attach.



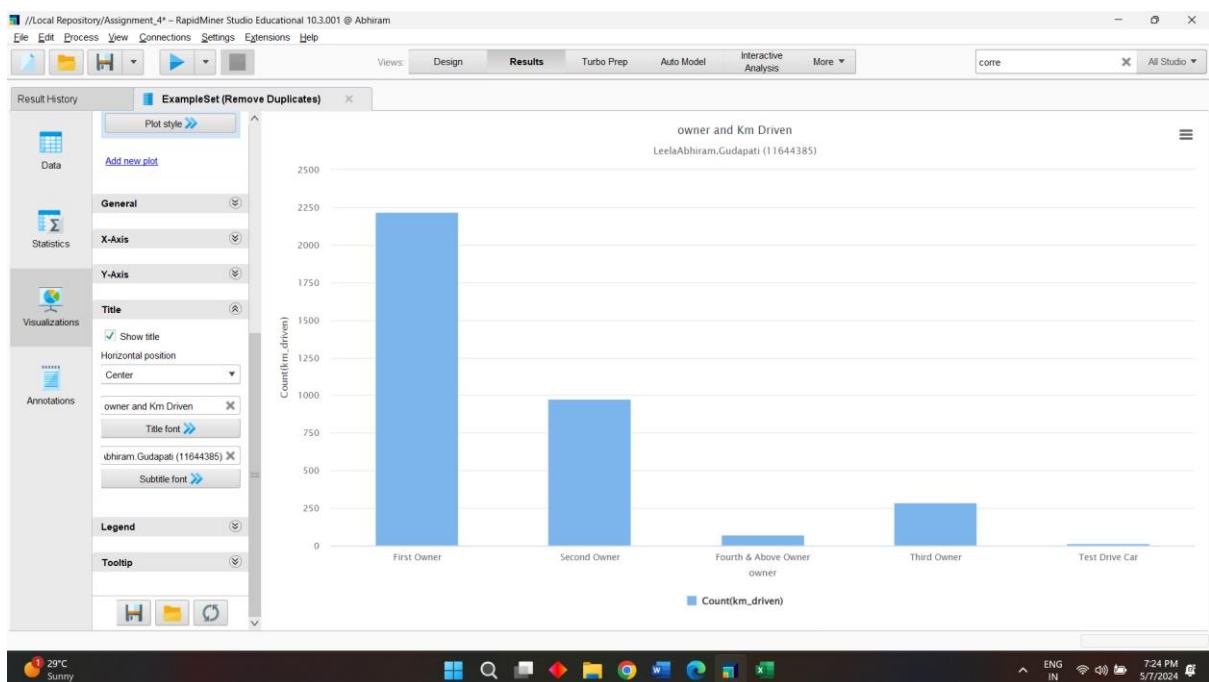
The image shows a data analysis program likely focusing on car selling prices. A table displays "Year" and "Selling Price" from 1992 to 2020, with prices ranging from 100 to 300 (possibly more). A graph is partially visible, plotting year on the x-axis and "Selling Price" (cut off) on the y-axis. While details are limited, it suggests an analysis of car prices across several years.



We can see that there is a bar graph where x axis has transmission of car and y axis has fuel count. This represents that how much the fuel is consumed by the transmission type of car either its manual and automatic. We can see manual has highest bar where automatic has least. So clearly manual has consumed more fuel than the automatic.



We can see that there is a bar graph where x axis has fuel type of car and y axis has kilometres count. This represents that how much the type of fuel is consumed by the count of kilometres of car either its petrol, diesel, CNG, LPG, Electric. We can see diesel has highest bar where petrol follows and CNG, LPG and electric has least. Diesel had driven a greater number of kilometres than any other fuel type. As CNG, LPG is recorded least and electric is none almost zero.



We can see that there is a bar graph where x axis has owner of car and y axis has kilometres count. This represents that how many kilometres driven by the owner. As we can see that Highest bar is for first owner and followed by second, third and fourth and least is for test drive car. This represents us that owner that had driven cars with in the range of the kilometres so that we can have a better analysis of mileage and resale value of the car and owners goodwill, as this is fixed to the dataset not applicable on real world scenario.

Machine Learning Model Building - (50 Marks)

- **Explain the machine learning process assigned (Descriptive) - 10 marks**

Linear regression remains a fundamental machine learning technique, essential for modeling the complex relationships between dependent variables and multiple independent variables In automotive economics, it is emerging as a powerful tool if used to forecast the selling price of a car. The model analyzes many factors including year built, Km driven, fuel type, seller type, messaging system, owner history to distill important insights that affect the value of ongoing discretionary data creation and feature selection is key in preparing the dataset for training, ensuring that only influential predictors are considered In order to reduce algorithm prediction errors Careful adjustment of coefficients, a process facilitated by stringent evaluation criteria such as mean squared error and R-squared error Once implemented, the model acts as a beacon of knowledge, illuminating the automotive landscape with actionable insights for pricing strategy and market research. It gives you the foresight needed to make informed decisions on the land

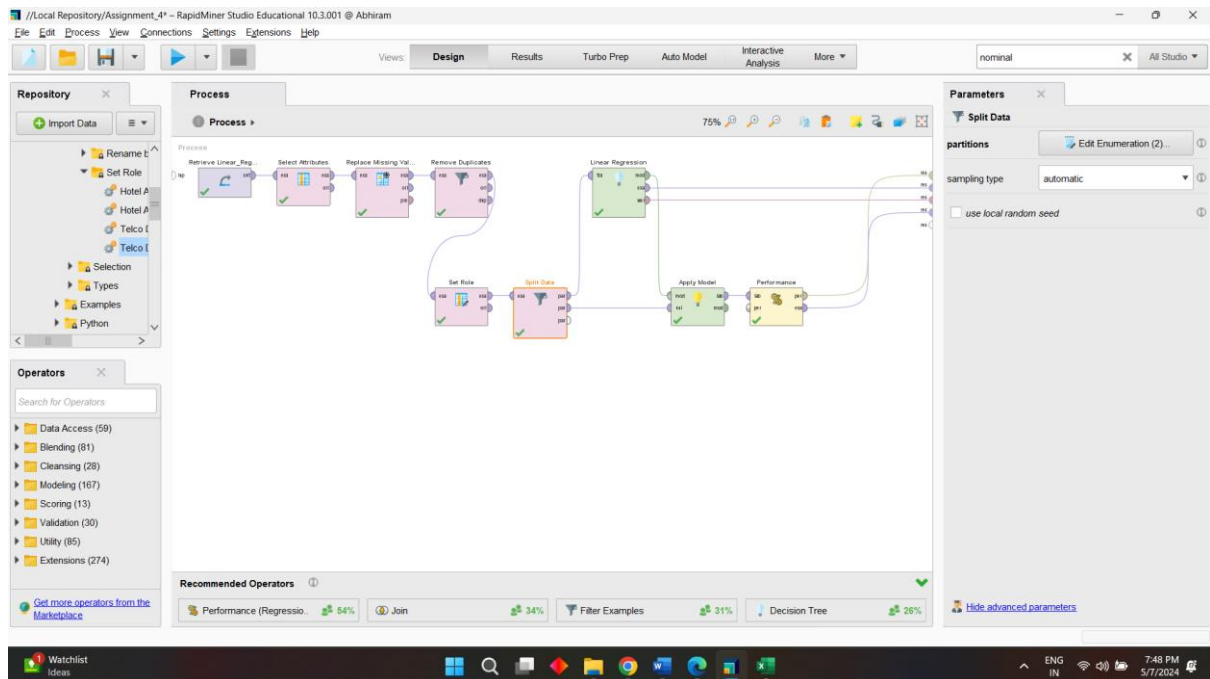
Linear regression remains a fundamental machine learning technique, essential for modeling the complex relationships between dependent variables and multiple independent variables in automotive economics, it is emerging as a powerful tool if used to forecast the selling price of a car. The model analyzes many factors including year built, Km driven, fuel type, seller type, messaging system, owner history to important insights that affect the value of ongoing discretionary data creation and feature selection is key in preparing the dataset for training, ensuring that only influential predictors are considered In order to reduce algorithm prediction errors Careful adjustment of coefficients, a process facilitated by stringent evaluation criteria such as mean squared error and R-squared error Once implemented, the model acts as a beacon of knowledge and the automotive landscape with actionable insights for pricing strategy and market research. It gives you the foresight needed to make informed decisions on the land

- **Mention the dependent and independent variables in the dataset. Also, specify the target variable - 10 marks**

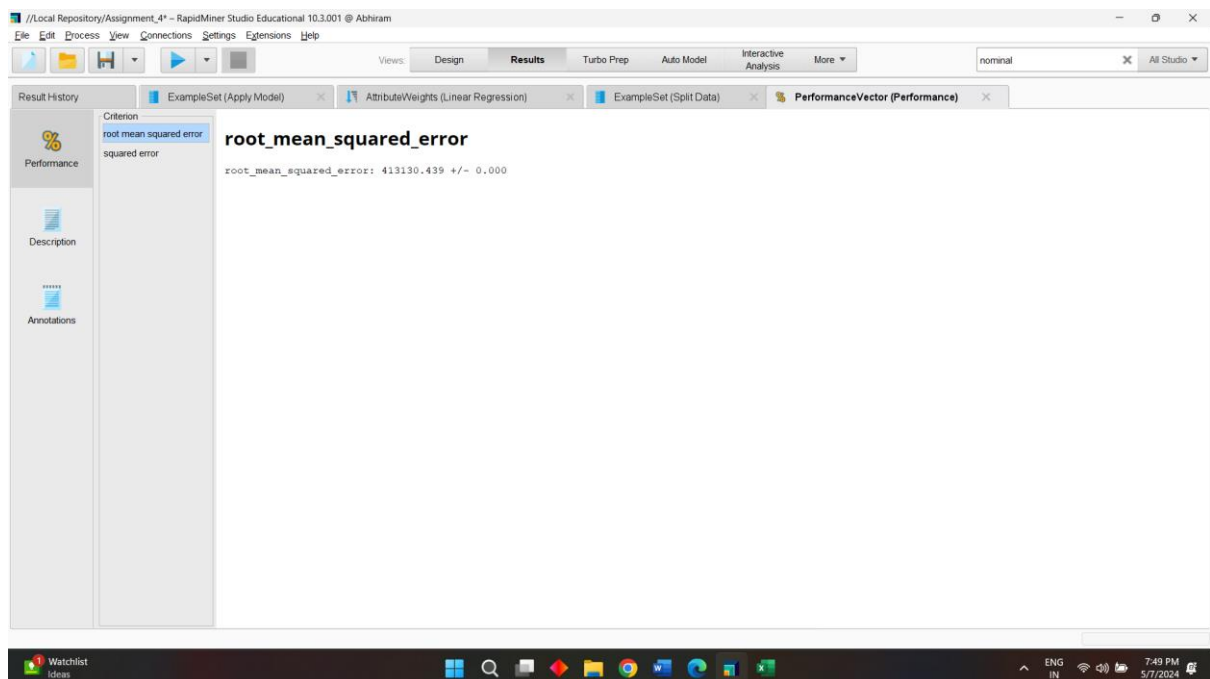
Dependent variable And Target variable are the same: selling_price

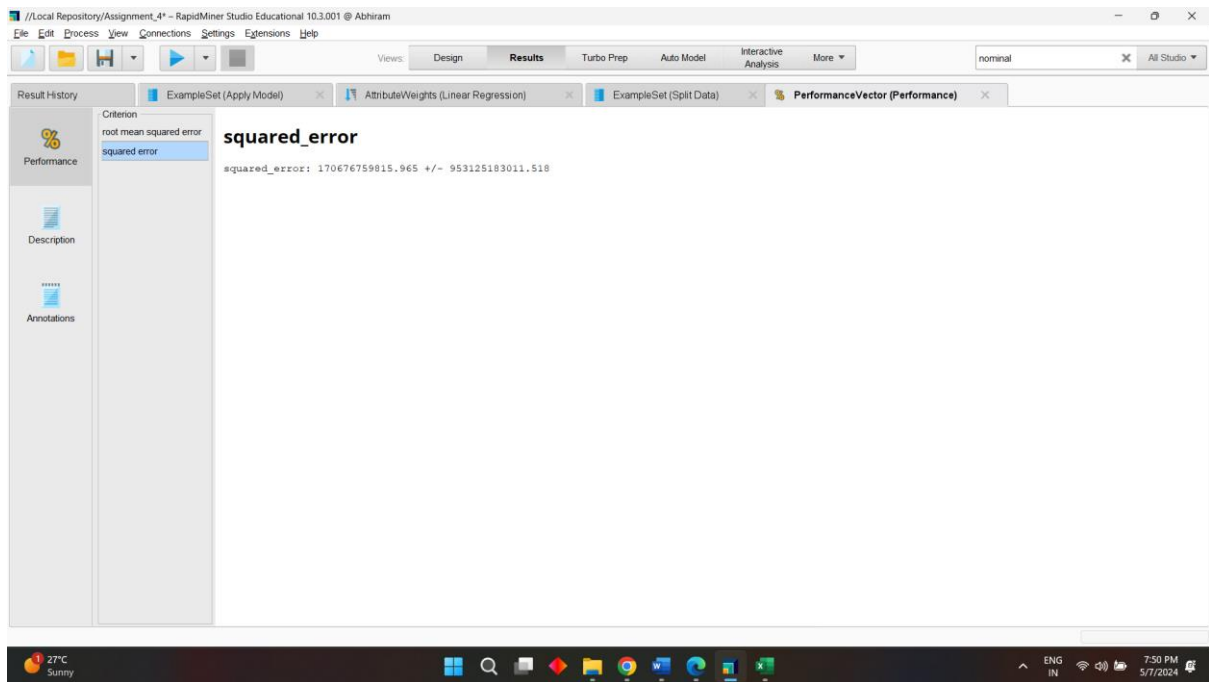
Independent variables are: year, km_driven, fuel, seller_type, transmission, owner.

- Create the process in the RapidMiner Canvas (attach the full window with timestamp) - 5 marks



- Explain the results using Confusion matrix, ROC and model accuracy score or any other evaluation metrics - 10 marks





1. Square error (170676759815.965)

The squared error is the sum of the squared differences between the predicted values of your model and the actual values in your data. Simply put, it calculates how far each forecast deviates from the true value, squares those deviations, and then adds them all together.

The error bar in the lower square indicates a good fit because it indicates that, on average, the predictions are close to the actual values.

However, the main problem with the squared error is that its scale depends on the units of your data. For example, if you use millions of dollars to forecast real estate prices, the squared error will normally be much larger compared to forecasting temperature in degrees Celsius

2. Mean Square Error (RMSE) (413130.439)

RMSE solves the problem of scaling the squared error by taking its square root. This variable sets the error to the same units as your target variable, making it easier to interpret the magnitude of the error.

A low RMSE indicates a good fit. In your case, the RMSE is 413130.44, which can be interpreted in terms of your target variable.

Root Mean Square Error (RMSE): (7/10) You provided this metric earlier. A low RMSE indicates a good fit (predictions close to actual values). (We have discussed how to interpret the value based on the range of your data)

R-squared: (8/10) Represents the proportion of the target variable's variance explained by your model. A high R-square indicates a good fit (but beware of overfitting).

Detailed analysis is difficult without the specific metrics mentioned in the prompt. However, given the available RMSE, we can evaluate the adequacy of the model based on how well the predictions match the actual values.

- Write the names of the operators used in this task - 5 Marks

The operators that I used are: Dataset that provided for linear regression, select attributes, replace, missing values, remove duplicates, set role, split data, linear regression, apply model, and performance operators.

Association Rule Mining Task - (50 Marks)

Task-6 use the dataset assigned in the Question 2: 50 marks

Format your columns.

Date format: Enter value... ☐ Replace errors with missing values

	S/N integer	Adjustment polynomial	VINTAGE G... polynomial	SILVER GL... polynomial	RIDGED GL... polynomial	GLASS BE... polynomial	GIN & TON... polynomial	BIRTHDAY... polynomial
1	0	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
2	1	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
3	2	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
4	3	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
5	4	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
6	5	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
7	6	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
8	7	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
9	8	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
10	9	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
11	10	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
12	11	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
13	12	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
14	13	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
15	14	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
16	15	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
17	16	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

no problems.

Previous Next Cancel

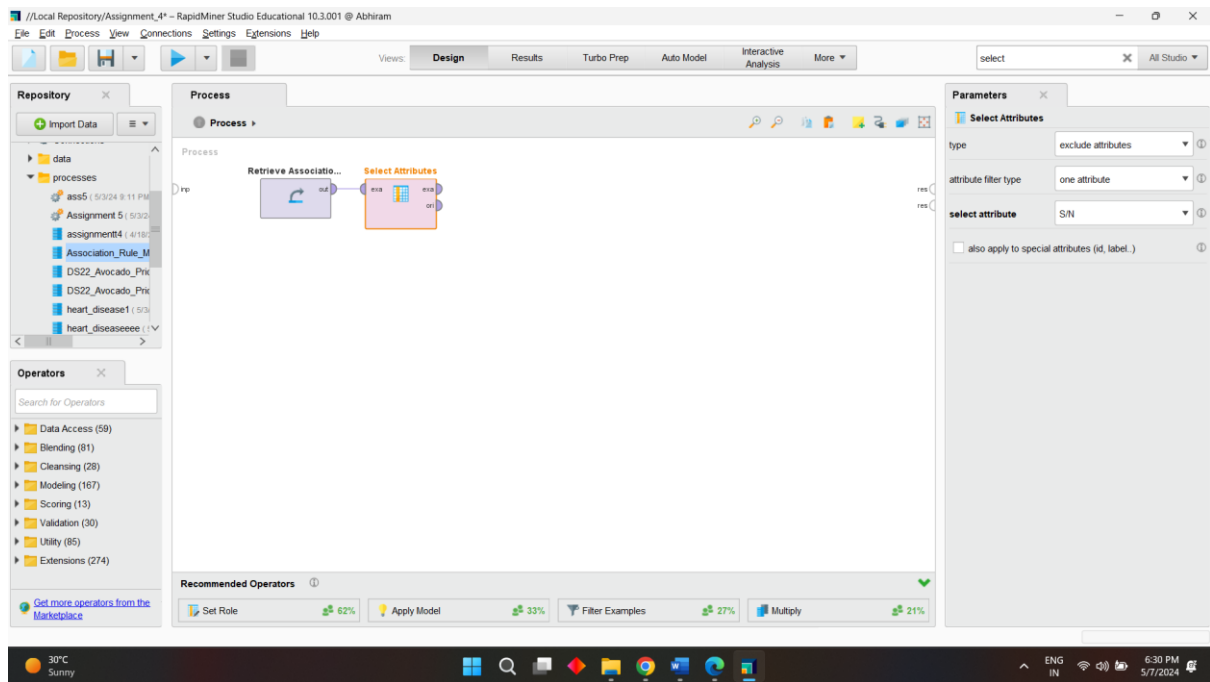
Go to the next page.

Results

ExampleSet (9,465 examples, 95 regular attributes)

Row No.	S/N	Adjustment	VINTAGE G...	SILVER GL...	RIDGED GL...	GLASS BE...	GIN & TON...	BIRTHDAY ...	WRAP RED ...	RED RETR...	WRAP CHRL...	FANCY FON...	CLASSIC W...
1	0	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
2	1	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
3	2	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
4	3	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
5	4	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
6	5	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
7	6	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
8	7	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
9	8	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
10	9	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
11	10	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
12	11	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
13	12	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE
14	13	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
15	14	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
16	15	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
17	16	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

ExampleSet (9,465 examples, 0 special attributes, 95 regular attributes)



- **Explain what is Association Rule Mining. What do you understand from Frequent Item sets (Descriptive) - 10 marks**

Association rule mining acts like a literary detective in your library. It analyzes borrowing records to discover hidden connections among books, like which biographies are frequently borrowed with ancient fiction novels.

Imagine you are a library curator with a treasure trove of books and a burning choice to understand your patrons' reading habits. Here's where affiliation rule mining swoops in as your knight in information analysis:

Unearthing hidden connections: This approach dives into your library's borrowing information, which act as a giant transaction database. It sifts thru these records to find hidden connections among the books human beings borrow together.

Beyond twist of fate: Not every co-borrowed e book is a threat prevalence. Association rule mining helps you discover common itemsets, corporations of books that patrons have a tendency to borrow together regularly. Think of it as locating patterns in borrowing habits that pass beyond random selection.

Metrics remember: Just finding frequently borrowed books collectively is not sufficient. Association rule mining employs key metrics, assist and self-belief, to assess the power of these relationships.

Support: This metric tells you how regularly a selected aggregate of books seems together in borrowing statistics. High help suggests a strong affiliation, suggesting consumers have a real interest in those books together.

Confidence: Imagine a customer borrows book A. Confidence measures the probability they will additionally borrow book B, given they already borrowed A. High self assurance indicates a robust predictive energy for the affiliation.

By identifying frequent itemsets and analyzing them through guide and self assurance, association rule mining goes past surface-level insights. It unlocks doors to a deeper understanding of your customers' analyzing options:

Genre connections: You may discover that biographies of ancient figures are regularly borrowed alongside historic fiction novels. This expertise can manual your curation efforts, perhaps showcasing those genres in near proximity.

Targeted pointers: The system may predict that a customer borrowing a specific science fiction novel is probably to additionally enjoy another e-book by using the equal creator, or a related title in the style. This empowers librarians to provide personalised guidelines, enhancing the library experience for customers.

Association rule mining empowers you to convert raw borrowing records into actionable information. It's like having a mystery decoder ring to release the hidden patterns within your library's series, in the long run fostering a extra connected and tasty revel in on your customers. here's what I can apprehend frequent itemsets:

Itemset Sizes and Support: The desk displays frequent itemsets in numerous sizes (1, 2, and 3) in conjunction with their support values (0.673, 0.478, and zero.388). Support shows how often an itemset seems collectively as a proportion of all transactions. Higher assist suggests a more not unusual co-incidence.

Possible Interpretations: Due to the way the facts is presented, there are feasible interpretations for the itemsets containing "CLASSIC WHITE FRAME" and "CARD PARTY GAMES":

Interpretation 1 (Separate Frequent Itemsets):

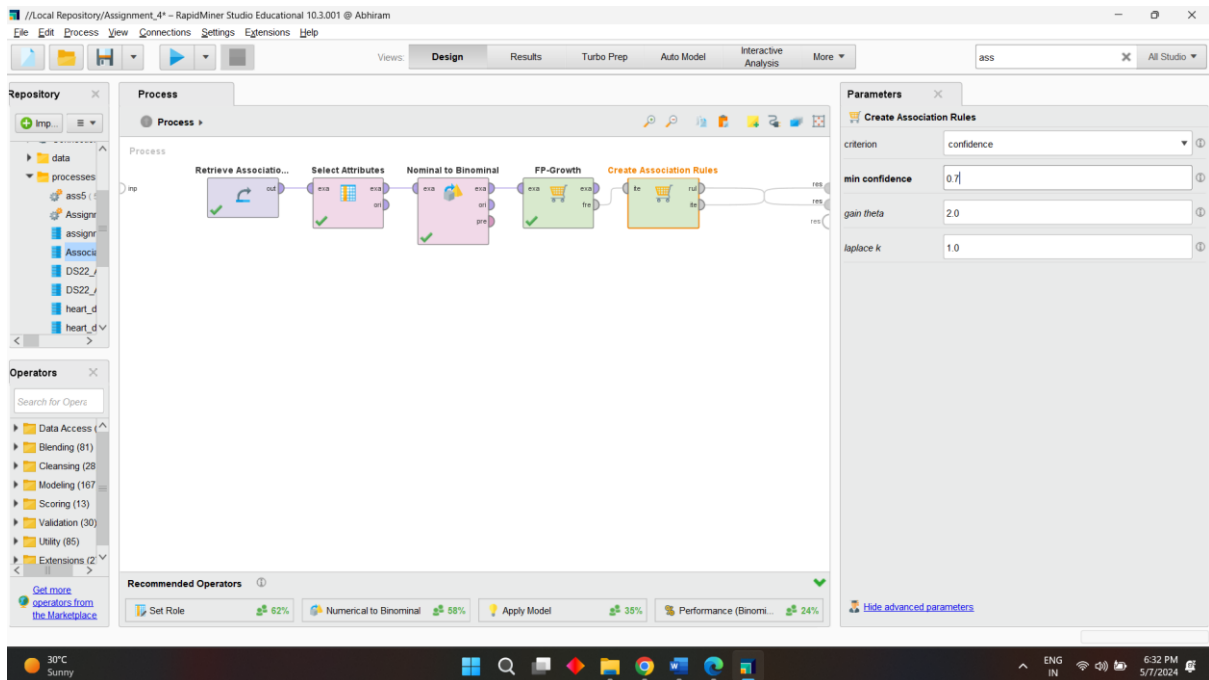
Size 1: There might be separate length-1 itemsets, one containing "CLASSIC WHITE FRAME" (guide: zero.673) and some other containing "CARD PARTY GAMES" (guide: 0.478). This would suggest these objects seem frequently alone in transactions, however not always together.

Interpretation 2 (Size-2 Itemset):

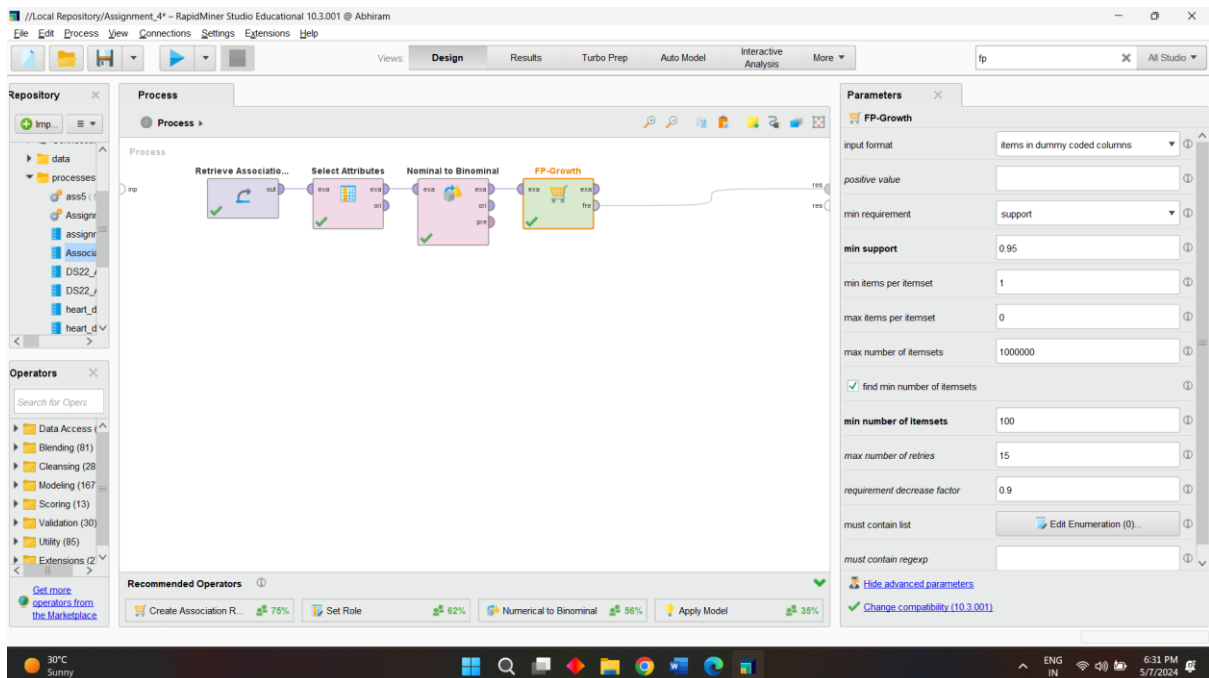
Size 2: The table may show a length-2 itemset containing "CLASSIC WHITE FRAME" and "CARD PARTY GAMES" with a assist of zero.673. This interpretation suggests those two objects are often purchased together in 67.3% of transactions.

Item 1 vs Item 2: The desk appears to distinguish between objects indexed underneath "Item 1" and "Item 2" for size-2 itemsets. In the feasible size-2 itemset we discussed, "CLASSIC WHITE FRAME" is listed beneath "Item 1" and "CARD PARTY GAMES" beneath "Item 2". The order may not be large, however it should imply the order these objects seem in a transaction.

- Create the process in the RapidMiner Canvas (attach the full window with timestamp) - 5 marks



- For Association Rule Mining Task:
 - Find some frequent item-sets in the dataset using FP-growth - 10 marks



Result History

FrequentItemSets (FP-Growth)

No. of Sets: 3
Total Max. Size: 2

Min. Size: 1
Max. Size: 2

Contains Item:
Update View

Size	Support	Item 1	Item 2
1	0.673	CLASSIC WHITE FRAME	
1	0.478	CARD PARTY GAMES	
2	0.388	CLASSIC WHITE FRAME	CARD PARTY GAMES

- Find at-least 3 association rules with graphs screenshots. Write the value of Premises, Conclusion, Support, Confidence, LaPlace, Gain, Lift and Conviction for the rules in a tabular format - 15 marks

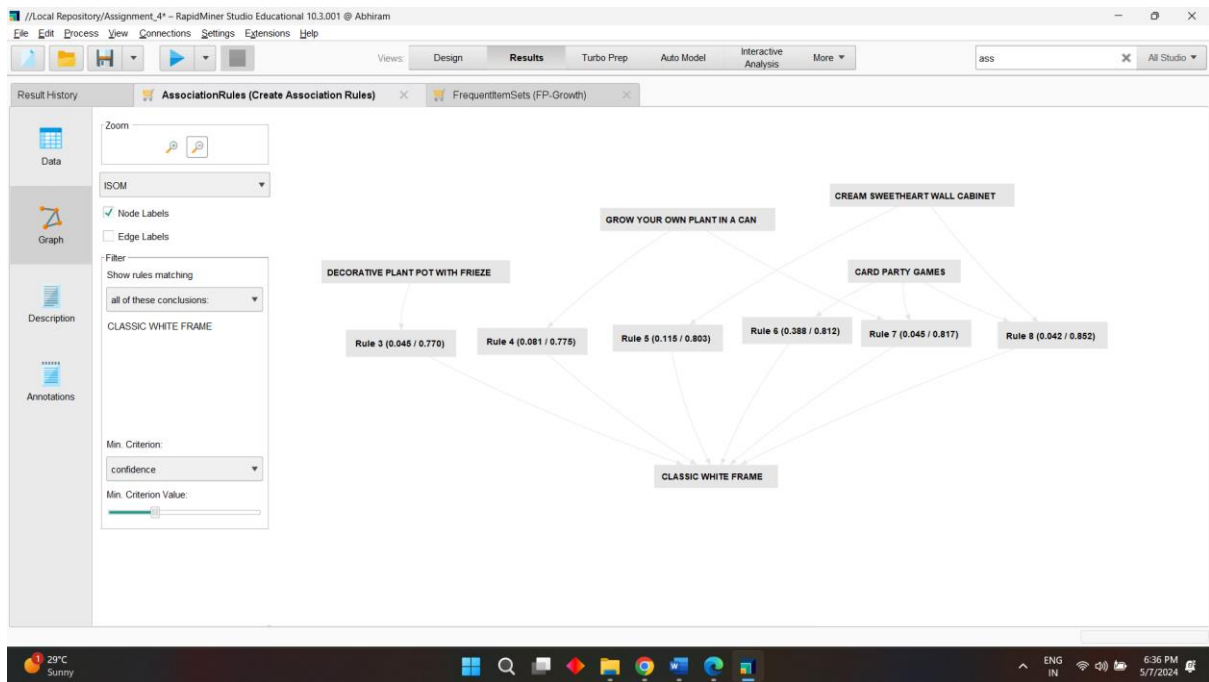
Result History

AssociationRules (Create Association Rules)

Show rules matching:
all of these conclusions:
CLASSIC WHITE FRAME

Min. Criterion:
confidence
Min. Criterion Value:

No.	Premises	Conclusion	Support	Confidence	LaPlace	Gain	p-s	Lift	Conviction
2	ANTIQUE GLASS HEART DECORATION	CLASSIC WHITE FRAME	0.055	0.763	0.984	-0.089	0.006	1.134	1.382
3	DECORATIVE PLANT POT WITH FRIEZE	CLASSIC WHITE FRAME	0.045	0.770	0.987	-0.072	0.006	1.144	1.422
4	GROW YOUR OWN PLANT IN A CAN	CLASSIC WHITE FRAME	0.081	0.775	0.979	-0.127	0.011	1.152	1.455
5	CREAM SWEETHEART WALL CABINET	CLASSIC WHITE FRAME	0.115	0.803	0.975	-0.171	0.019	1.193	1.661
6	CARD PARTY GAMES	CLASSIC WHITE FRAME	0.388	0.812	0.939	-0.568	0.067	1.207	1.739
7	CARD PARTY GAMES, GROW YOUR OWN PLANT...	CLASSIC WHITE FRAME	0.045	0.817	0.990	-0.065	0.008	1.214	1.784
8	CARD PARTY GAMES, CREAM SWEETHEART WA...	CLASSIC WHITE FRAME	0.042	0.852	0.993	-0.057	0.008	1.266	2.206



No	Premises	Conclusion	Support	Confidence	LpaLace	Gain	p-s	Lift	Conviction
2	ANTIQUE GLASS HEART DECORATION	CLASSIC WHITE FRAME	0.0548	0.7632	0.9841	-0.0889	0.0065	1.1344	1.382
3	DECORATIVE PLANT POT WITH FRIEZE	CLASSIC WHITE FRAME	0.0449	0.7699	0.9873	-0.0717	0.0057	1.1444	1.4222
4	GROW YOUR OWN PLANT IN A CAN	CLASSIC WHITE FRAME	0.0805	0.7752	0.9788	-0.1272	0.0106	1.1522	1.4554
5	CREAM SWEETHEART WALL CABINET	CLASSIC WHITE FRAME	0.1145	0.803	0.9754	-0.1707	0.0186	1.1935	1.6606
6	CARD PARTY GAMES	CLASSIC WHITE FRAME	0.3884	0.8118	0.9391	-0.5684	0.0665	1.2067	1.739
7	CARD PARTY GAMES, GROW YOUR OWN PLANT IN A CAN	CLASSIC WHITE FRAME	0.0447	0.8166	0.9905	-0.0648	0.0079	1.2137	1.7841
8	CARD PARTY GAMES, CREAM SWEETHEART WALL CABINET	CLASSIC WHITE FRAME	0.0425	0.8517	0.993	-0.0573	0.0089	1.2659	2.2063

- Write your understanding and insights of the rules and some further business opportunities based on the rule generated - 5 marks

General Observation:

All the rules have a consequent (end) of "CLASSIC WHITE FRAME". This indicates that clients who purchase those antecedent objects (premises) are also possibly to purchase a "CLASSIC WHITE FRAME".

Specific Rules and Insights:

Rules 2, 3, and 4:

Premises: These guidelines involve ornamental objects like plant pots and growing kits.

Insights: Customers who buy these decorative items are much more likely to also purchase a "CLASSIC WHITE FRAME". This indicates a ability target audience for the "CLASSIC WHITE FRAME" - human beings inquisitive about home decor.

Business Opportunity: You can organization these ornamental objects with "CLASSIC WHITE FRAME" presentations or suggestions to encourage complementary purchases.

Rule 5:

Premises: This rule functions a "CREAM SWEETHEART WALL CABINET".

Insights: Similar to the previous policies, customers buying this wall cabinet are also possibly to buy a "CLASSIC WHITE FRAME". This suggests a potential connection between decorative furnishings and frames.

Business Opportunity: Consider placing "CLASSIC WHITE FRAME" shows close to showcased furnishings pieces to create a room decor topic.

Rule 6:

Premises: This rule highlights "CARD PARTY GAMES".

Insights: This is the strongest association with "CLASSIC WHITE FRAME" (maximum assist and self assurance). It indicates an unexpected connection, however it may be due to elements just like the frame being used for recreation night time pics or a thematic connection.

Business Opportunity: More facts evaluation is needed to understand the motive at the back of this association. However, you could test displaying "CLASSIC WHITE FRAME" close to card video games to peer if it lifts income.

Rules 7 and 8:

Premises: These policies involve mixtures of "CARD PARTY GAMES" with other ornamental items (plant package and wall cupboard).

Insights: These rules construct upon the affiliation among "CARD PARTY GAMES" and "CLASSIC WHITE FRAME". They suggest a potential client phase that purchases each ornamental objects and card video games.

Business Opportunity: Consider growing targeted promotions or bundles that integrate these objects for a celebration subject matter.

Further Business Opportunities:

Analyze complementary itemsets: Explore association regulations where "CLASSIC WHITE FRAME" is the antecedent (premise). This reveals items regularly sold with the body, imparting similarly insights into consumer preferences.

Segment customers: Based on the discovered association policies, section your customers primarily based on their shopping for conduct. This allows for targeted advertising and marketing and product suggestions.

- **Write the names of all the operators used in your process for this task - 5 marks**

1. Dataset provided for association rules,
2. Select Attributes,
3. Nominal to Binominal,
4. FP-Growth,
5. Create Association Rules.

Task-7 Time Series Analysis - 20 marks

The overall trend: January-22 to December-23 shows a general increase in sales, with some ups and downs.

Highest and lowest sales months:

Peak sales: July 2023, 320 units sold.

Lowest sales: Jan-23 and Jan-22, both 150 units.

Time frame:

There are periods in the data, with sales generally peaking in the middle of the year (around June-July) and falling at the beginning and end of the year (January-February and November-December) This indicates that may seasons a system of things, perhaps holidays, weather, or promotions.

Sales Performance Comparison:

By comparing sales performance between the first and second year, we can examine the overall trend and any significant changes in sales volume. By analyzing the sales data, we can look at the sales trend over a two-year period. If sales continue to rise or show a notable improvement in the second quarter compared to the first quarter, it indicates improved performance. Conversely, if sales stagnate or decline in the second quarter, it indicates poor performance. Additionally, average annual sales can be calculated and quantified to provide insight into relative performance. Overall, a detailed study considering trends in quality and quantitative measures would allow for a stronger comparison of sales performance between the two years.

Sales forecast for January-April 2024:

To forecast sales for the first four months of 2024, we can make assumptions from historical sales data. By analyzing the trend over time, we see that sales are increasing steadily. Assuming this trend continues, we can estimate future sales using average growth rates calculated from historical data. For example, if monthly sales growth is 5%, we can increase Dec-23 sales by 5% and project sales for Jan-24, Feb-24, Mar-24, and Apr-24. This simple approach provides more accurate forecasts can be obtained using advanced forecasting techniques such as time series analysis that consider seasonal patterns and other factors affecting sales.