MustCart

MustCart is a Trend analysis project where it does analysis of panic buying trend of people during pandemic/lockdown situations and guides them to buy essential products.

We extracted useful information that answered the questions such as: what gender shops more during lockdown/pandemic situations? Which age group is the highest spender?

In the end, we created a simple machine learning algorithm that predicted the amount of money that a person is likely to spend on lockdown depending on features such as gender, age, and occupation.

We developed it using Python, SQL server and Machine Learning Algorithm.

Dataset is from a private source; it contains one day sales data of a supermarket. It contains user info, product info and sales info. We did a pre data analysis of this dataset, and we came to know about the different products from various department sold on a single day.

There are 18,240 entries, 1000 customers bought various products from 23 different categories. There are total 4000 products divided in different aisles (department wise).

Main aim of this project is to find some trend in this dataset.

The graphs showed that there are almost 3 times more male customers than female customers.

we can conclude that the highest number of customers belong to the age group between 26 and 35, for both genders. Based on these results, the supermarket should sell most of the products that target people in their late twenties to early thirties. To increase profits, the number of products targeting people around their thirties can be increased while the number of products that target the older or younger population can be reduced.

1. The Data Analysis Process

This process contains 4 major steps

1. Questions
2. Wrangle
3. Explore
4. Draw Conclusions

we will look into each steps one by one going through the analysis.

1. Questions

Asking right questions is the key part of the analysis process this will define what you are going to present to the audience. Answers to the right questions will provide key inputs to the company or the audience to improve their business.

if we consider our Indian supermarket dataset the appropriate questions and their results will lead to better functioning of the retail store on their sales.

As we can see the dataset contains 11 features. so, from these features let us form few questions that would help the store to better understand their customers.

1. which age group of customers are more likely to purchase with More amount per person?
2. which age group and gender have high visiting rate to the retail store?
3. Top 10 products which made highest sales in the store?
4. Based on marital status and gender who has high purchase rate?
5. Which product is popular for each age group?
6. What is the purchase percent for each age group and for Gender Group in total purchase amount?

2. Wrangle

Wrangling is the part where we make sure that the data, we collected for analysis is of good quality. Here we assess the data quality and clean the data. Wrangling is the part where we take care of missing data, duplicate data, incorrect datatypes etc.,

let us check the basic summary which our dataset tells us.

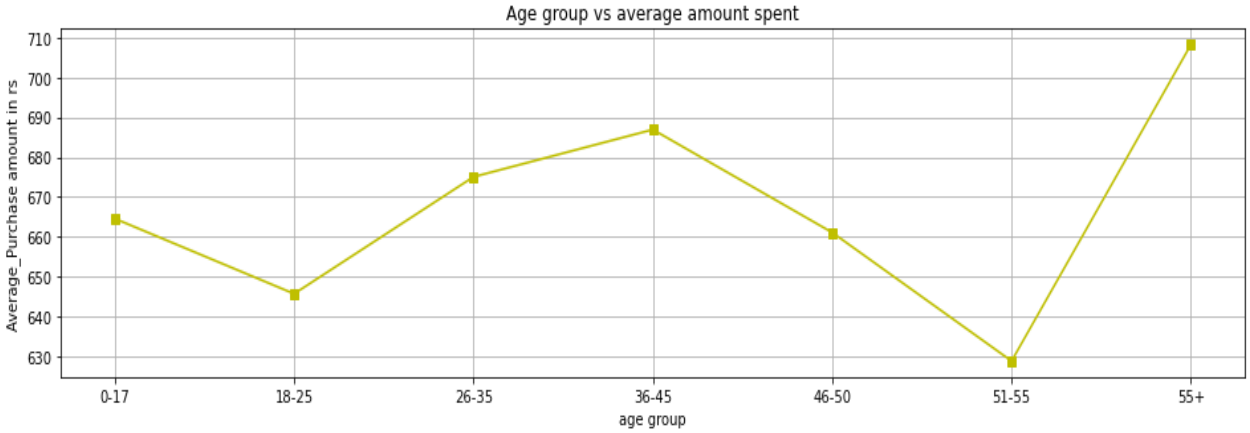
what we can understand from the summary?

1. Total samples in our dataset are 18240 (no of rows)
2. This dataset has 11 features (no of columns)
3. There are 2 features of float type, 6 features of int type, 3 features of object type (String)
4. Age in dataset is range so it is Object (String)

3. Explore

Now we will analyse and visualize the questions we posed and find out what data tells us.

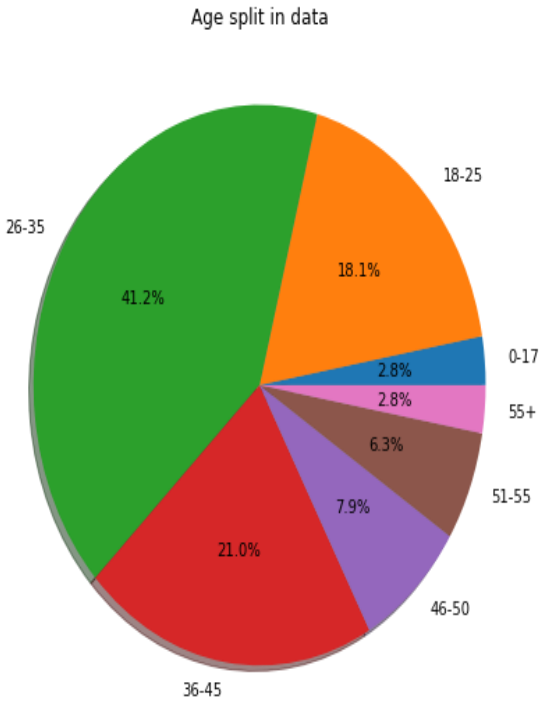
1. which age group of customers are more likely to purchase with More amount per person?

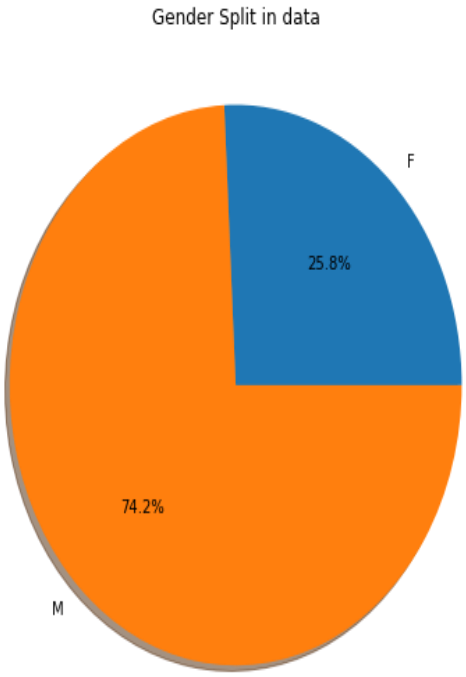


People of age group 55+ have spent more on purchase, approximately 700 rs on average spent by this age group People.

The graph values tend to increase from 18-25 to 36-45 and decreases from 46-50 to 51-55 and rises further in 55+ age group. There is an increase purchase variation between 18-25 and 36-45 age people.

2. which age group and gender have high visiting rate to the retail store?





The first Pie gave interesting understanding when we compare it with first question solution.

This shows 41% of customers are 26-35 age group and 21% are from 36-45 => 62% of customers from 26-45 age group. only 3% of customers are of 55+ Age group.

The second pie tells that the store gets most of the male customers i.e. 74.2% male customers & 25.8% Female customers.

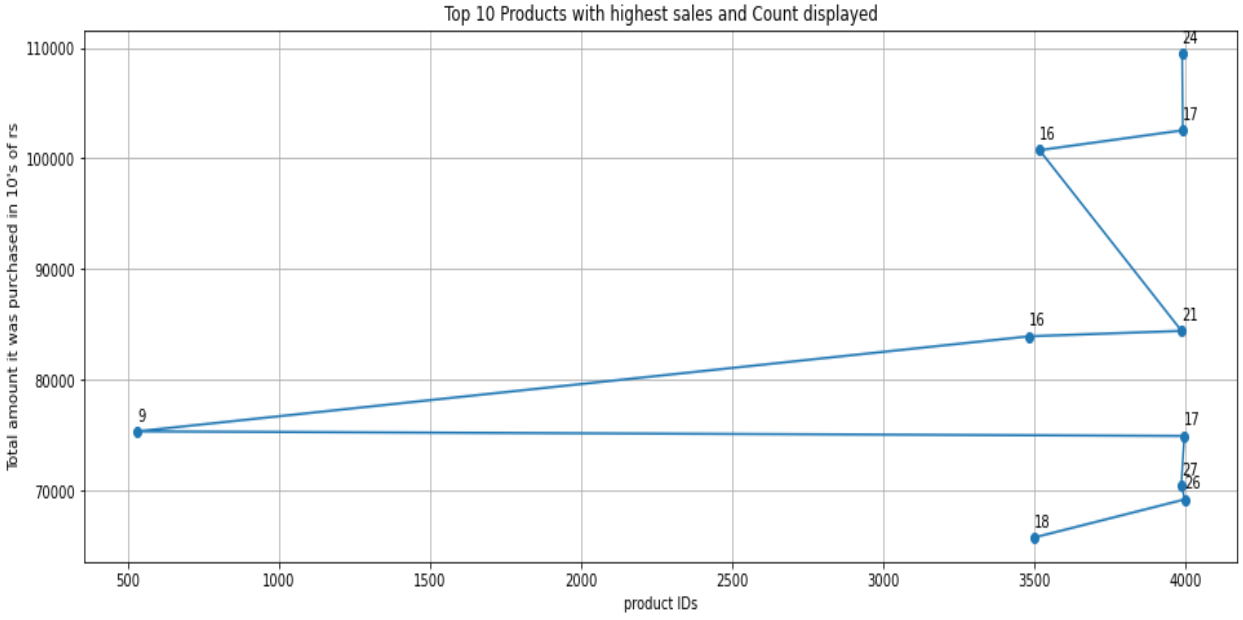
From 1st and 2nd questions we can tell 62% of customers from 26-45 who have a medium purchase rate. 3% of customers are from 55+ age group who have high purchase rate. This Gives an interesting insight on sales to store owners.

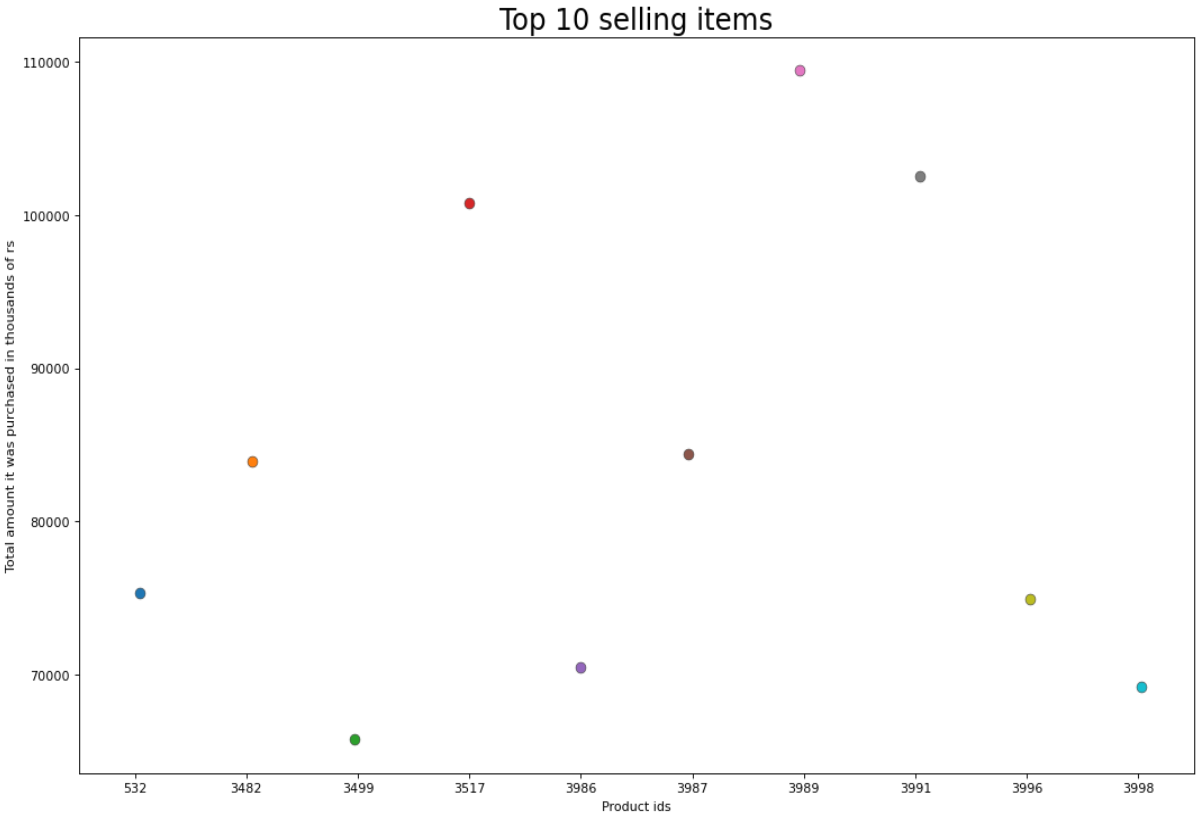
3. Top 10 products which made highest sales in the store?

We can see 10 products their purchase amount and count of products sold.

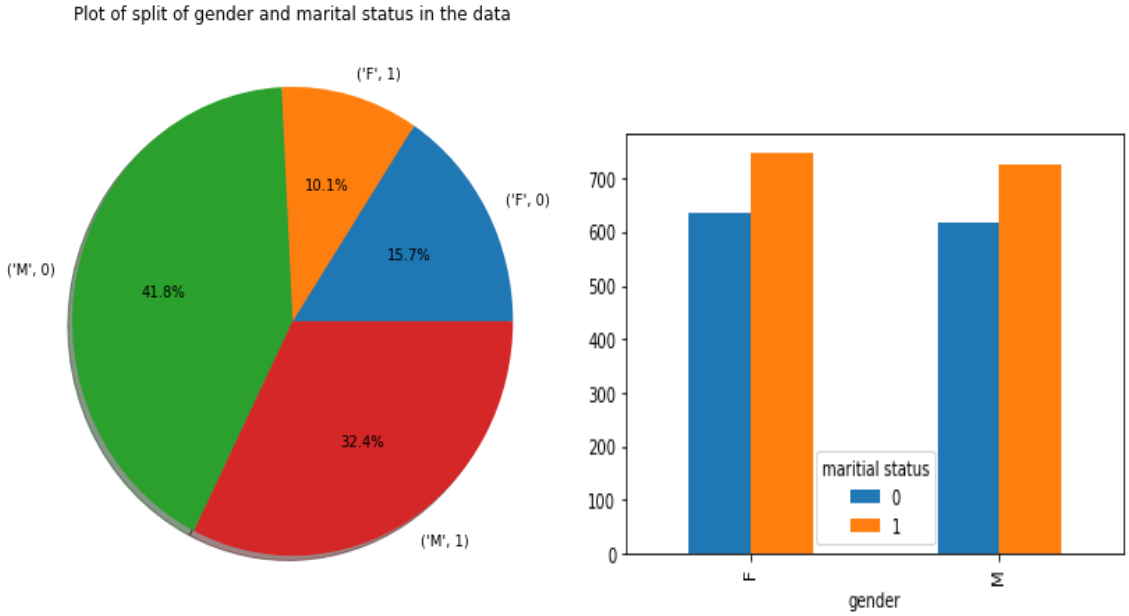
1. 1st product had 24 pieces sold with total sale of 1,09,427 rs and 2nd product with 17 pieces sold but with 1,02,546 rs. which means 1st product might have higher product cost.
2. 3rd product had 16 units sold and 4th product had 21 units sold but 4th product had low price than 3rd product so even it had higher products sold it had lesser sale amount than 3rd.
3. similarly, we observed for all the products.

	product_id	Total_amount	Count	brand	product	department
3783	3989	109427.00	24	Bikaner	Appetizing Festivity	Sweets
3785	3991	102546.00	17	Bikaner	Gift Of Variety	Sweets
3311	3517	100752.00	16	Pampers	Premium Care Xtra Large Monthly Box Pack - 72 ...	Diapers
3781	3987	84435.00	21	Bikaner	Moti Choor Pleasure	Sweets
3276	3482	83944.00	16	Pampers	Active Baby Large (9-14 kg) - 78 Diapers	Diapers
531	532	75363.75	9	BORGES	Olive Oil - Extra Light	Oil&Ghee
3790	3996	74950.00	17	Bikaner	Assorted Sweets	Sweets
3780	3986	70467.00	27	Bikaner	Creamy Milk Cake Celebration	Sweets
3792	3998	69223.00	26	Bikaner	2 Layer Bamboo & Treats	Sweets
3293	3499	65775.15	18	Pampers	New Medium - 76 Diaper Pants	Diapers





4. Based on marital status and gender who has high purchase rate?



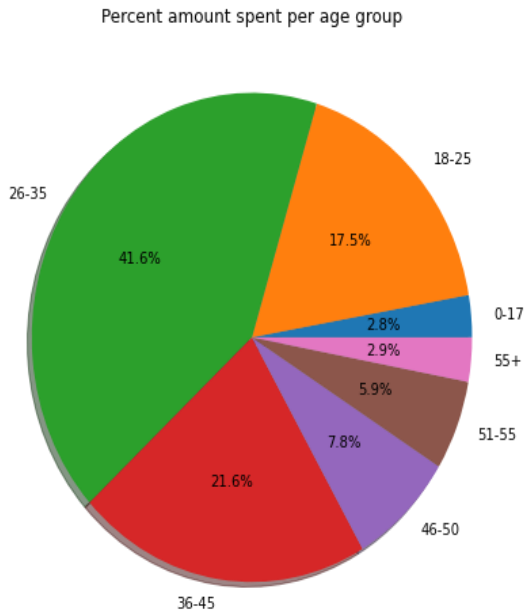
Males tend to purchase more and Unmarried Males are around 42% in the data and they show to purchase 700 rs on average.

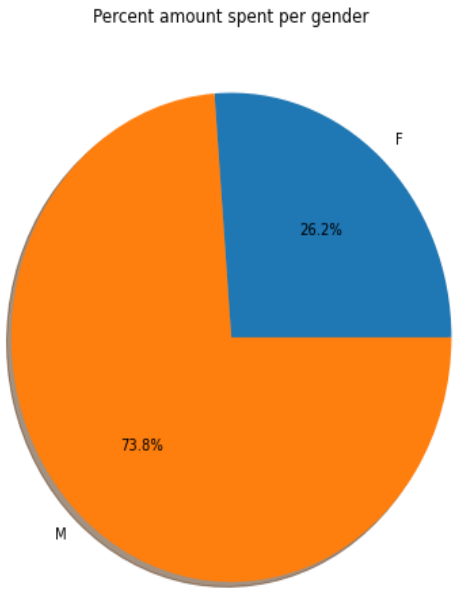
5. Which product is popular for each age group?

	age	product_id	brand	product	price	department	dept id	essential
0	0-17	1	Fresho	Sun Melon	30.0	Fruit	1	Y
1	18-25	3989	Bikaner	Appetizing Festivity	1499.0	Sweets	16	N
2	26-35	3471	Mamypoko	Extra Absorb XXL - 12 Diaper Pants	275.0	Diapers	19	Y
3	36-45	1716	bb Royal	Jaggery - Kolhapuri 450G + Till Regular 500G +...	260.0	Salt&Sugar	8	Y
4	46-50	1714	bb Royal	Organic - Brown Sugar	95.0	Salt&Sugar	8	Y
5	51-55	3998	Bikaner	2 Layer Bamboo & Treats	899.0	Sweets	16	N
6	55+	2227	Red Label	Tea	115.0	Tea & Coffee	13	Y

Here each age group has their own most bought product.

6. What is the purchase percent for each age group and for Gender Group in total purchase amount?





It Looks like count of people in different Age groups in data is in correlation with total percent of amount spent. Similarly, with Gender males were 74% their spending in total is 73.8%, females were 26% their spending in total is 26.2%.

4. Conclusion

From the questions and Solutions lets write a summary of our findings.

Findings

- People of Age group 55+ have purchased with high amount per person (700 rs per person).

- 74% of total people visited were Male and 62% of total people were between Age 26-45.

- People from Age group 26-35 collectively have spent more amount (41% of sale purchase is from this group).

- Unmarried Male who are 42% in the dataset have spent 630 rs per person.

- 1st product had 24 pieces sold with total sale of 1,09,427 rs and 2nd product with 17 pieces sold but with 1,02,546 rs. which means 1st product had higher product cost.

2. Predicting the amount spent by the customer using ML algorithm

We removed the columns that do not help in the prediction.

‘User_id’ is the number assigned automatically to each customer, and it is not useful for prediction purposes.

The ‘product_id’ column contains information about the product purchased. It is not a feature of the customer. Therefore, we removed it too.

Our final selection is based on 9 columns - one variable we want to predict (the ‘total_amount’ column) and 8 variables which we used for training our machine learning model.

we were dealing with 2 categorical columns. However, basic machine learning models are capable of processing numerical values. Therefore, we converted the categorical columns to numeric ones.

We used a get_dummies Python function which converts categorical values to one-hot encoded vectors. All categorical variables were transformed into numerical. So, if a customer is between 0 and 18 years old, only that column value will be equal to 1, other age group columns have a value of 0. Similarly, if it is a male customer, the column named 'M' will be equal to 1 and column 'F' will be 0.

We used one of the simplest machine learning models, i.e. the linear regression model, to predict the amount spent by the customer on during panic buying. Linear regression represents a very simple method for supervised learning and it is an effective tool for predicting quantitative responses.

This model, like most of the supervised machine learning algorithms, makes a prediction based on the input features. The predicted output values are used for comparisons with desired outputs and an error is calculated. The error signal is propagated back through the model and model parameters are

updating in a way to minimize the error. Finally, the model is considered to be fully trained if the error is small enough.

We created input and output vectors for our model. The training set was used to fit our model. Training data is always used for learning, adjusting parameters of a model and minimizing an error on the output. The rest of the data (the Test set) was used to evaluate performances. We used a 60:40 ratio for training & test dataset.

We trained our model, created intercept parameters, and we derived values of all coefficients of our model. each category of our data set was defined with one regression coefficient. The training process was looking for the best values of these coefficients during the learning phase and we got the most optimum values for the coefficients of our machine learning model.

In order to see how well our model performs, we used test dataset as input.

Final predicted Output: Predicted purchases (in rupees) for new costumers: [6 39.6875 715.375 673.5 ... 673.5 673.5 660.0625]

From the output, we can clearly say that customer is going to spend around 600 to 700 rupees during panic buying/post lockdown announcement.

Performance Estimation of ML model

In the end, we found out the estimate our results by finding the mean absolute error (MAE) and mean squared error (MSE) of our predictions.

MAE: 591.4207291666668
MSE: 927104.7207132351

Machine learning can be used for a variety of tasks. In this project, we used a machine learning algorithm to predict the amount that a customer is likely to spend during panic buying. We also performed exploratory data analysis to find interesting trends of customers from the dataset.