Industry Review

The “Food Delivery Company” is a delivery company that delivers Italian, Thai, Indian and continental cuisine to customers in various cities in different region of country. The food delivery dataset was taken from Kaggle. The data consists of 145 weeks. Solving this case study will give us an overlook of how real business problems are solved using EDA, Statistics and Machine learning.

In this case study we will also develop a basic understanding of how food delivery services can be provided in effective way while handling demand and supply as well as man force on work.

We took sample of 145 weeks and need to predict for next 10 weeks. Through this case study we will learn how to use EDA and Machine learning techniques.

Demand forecasting is a key component to every growing online business Without proper demand forecasting processes in place, it can be nearly impossible to have the right amount of stock on hand at any given time. A food delivery service has to deal with a lot of perishable raw materials which makes it all the more important for such a company to accurately forecast daily and weekly demand.

Too much inventory in the warehouse means more risk of wastage, and not enough could lead to out-of-stocks — and push customers to seek solutions from your competitors. In this challenge, get a taste of demand forecasting challenge using a real dataset. For companies, mainly large ones with economies of scale and geographic capillarity, an error in the forecast of demand can cause several consequences, such as:

1. Stock break.
2. Perishable waste
3. Drop in production
4. Idle stock (slow moving)
5. Pricing errors

Literature Survey

The Cornell Hospitality Report (2011) published an article titled ‘Online, Mobile, and Text Food Ordering in the U.S. Restaurant Industry’ (2011) by S. Kimes and P. Laque in which a survey of the top 326 U.S. restaurant chains was carried out. It showed that the industry was gradually adopting electronic ordering, in the form of online, mobile, and text orders. The advantages of electronic ordering include increased sales, particularly through automatic upwelling and by storing order information so that customers are encouraged to repeat their previous orders with a single click. Other than the cost of installation and operation, the chief disadvantage of electronic ordering is the potential for amplifying rush time volume, with the potential of overwhelming the kitchen.

The Cornell University School of Hotel Administration published a paper titled ‘Consumer Perceptions of Electronic Food Ordering’ (2011). The researcher Sheryl E. Kimes surveyed 470 internet users and found that almost half of them had ordered food online through the usage of online food delivery Apps or through text messages. The study showed that the chief reason for electronic ordering given by those have ordered (users) is that they gain convenience and control. The major factor that inhibits those who have not ordered via an electronic channel (non-users) is a desire for interaction (although technology anxiety is also a factor).

In the research paper ‘Consumer experiences, attitude and behavioral intention toward online food delivery (OFD) services’ (2017), released by Vincent Cheow SernYeo, See-Kwong Goh, Sajad Rezaei, the researchers studied the structural relationship between convenience motivation, post-usage usefulness, hedonic motivation, price saving orientation, time saving orientation, prior online purchase experience, consumer attitude and behavioral intention towards Online Food Delivery (OFD) services. They concluded that customers were attracted to technology that could provide them convenience through saving time and effort. Thus, the website must be user friendly and be able to process the customer's request as quickly as possible. This in return would enable customers to complete a transaction quickly, which would be both beneficial to the customers and marketers. Having certain discounts or promotions attracted price- sensitive consumers, as they were likely to choose the channel with provided them with the best value for money.

According to Varsha Chavan, et al, (2015), the use of smart device – based interface for customers to view order and navigate has helped the restaurants in managing orders from customers immediately. The capabilities of wireless communication and smart phone technology in fulfilling and improving business management and service delivery. Their analysis states that this system is convenient, effective and easy to use, which is expected to improve the overall restaurant business in coming times.

In the study by Lee E, Lee S, and Jeon Yon the title ‘Factors Influencing the Behavioral Intention to Use Food Delivery Apps’ (2017), the authors have discussed the relationship between the determinants that affect customers' use of food delivery Apps. Using an extended technology acceptance model, they explored consumers' experiences in purchasing delivery food through mobile Apps. Through the usage of a self-administered questionnaire online and used structural equation modeling the hypotheses was tested. In conclusion it was found that user-generated information, firm- generated information, and system quality had a significant effect on perceived usefulness. In addition, system quality and design quality strongly

influenced the perceived ease of use, which improved perceived usefulness, and in turn, perceived usefulness and perceived ease of use affected attitude toward the use of mobile Apps.

Business problem statement (GOALS)

Business problem:

The food delivery company which operates in multiple cities. They have various fulfillment centers in these cities for dispatching meal orders to their customers. The client wants you to help these centers with demand forecasting for upcoming weeks so that these centers will plan the stock of raw materials accordingly.

The replenishment of majority of raw materials is done on weekly basis and since the raw material is perishable, the procurement planning is of utmost importance. Secondly, staffing of the centers is also one area wherein accurate demand forecasts are helpful.

Business Objective:

Given the following information, the task is to predict the demand for the next 10 weeks (Weeks: 146-155) for the center-meal combinations in the test set:

- Historical data of demand for a product-center combination (Weeks: 1 to 145)
- Product (Meal) features such as category, sub-category, current price and discount
- Information for fulfillment center like center area, city information etc.

Dataset and Domain

Our Dataset has 4,56,548 records and 15 attributes.

Column name	Description
ID	Serial number of the customer
WEEK	Week serial number
CENTER_ID	ID of the restaurant
MEAL_ID	Code of the meal
CHECKOUT PRICE	Prices include with tax and discount
BASE PRICE	Price of the meal
EMAILER_FOR_PROMOTION	Email address of the customer
HOMEPAGE_FEATURED	Food displayed on the front page of website

NUM_ORDERS	Number of the order in the week
CITY_CODE	Code of the city
REGION_CODE	Code of the region
CENTER_RATING	Rating of the center on the basis of food
OP AREA	Distance of customer from restaurant in KM
CATEGORY	Type of food available
CUISINE	Country name of food

Variable categorization (count of numeric and categorical)

In This dataset many variables are encoded from categorical values into Numerical Values

Pre-Processing Data Analysis (count of missing/ null values, redundant columns, etc.)

Alternate sources of data that can supplement the core dataset (at least 2-3 columns)

Quantitative/Numerical: 11 features:

- ‘Week'
- 'ID’
- ‘Center id'
- ‘Meal id'
- ‘Checkout price’
- ‘Num orders'
- ‘City code'
- ‘Region code’
- ‘Op area’
- ‘Email For Promotion’
- ‘Homepage Featured’

Qualitative/ Categorical: 3 features:

- ‘Center type’
- ‘Category’
- ‘Cuisine'

Target:

- ‘Num orders’

Project Justification

- In this Restaurant Dataset, we have 15 columns which will help us to explore the impact of sale.
- This is a Regression Problem.
- Using our Analytical skills to explore data, find correlation between features, and predict total number of orders to know the impact of business.
- We can use Linear model algorithms like Linear Regression, Decision Tree and Random Forest
- We use Shapiro test, Chi2 test, and Kruskal Wallis test for Statistical test.
- We can use boosting techniques for increasing the accuracy and performance of the model.

CRITICAL ASSESSMENT OF TOPIC SURVEY:

The food delivery company attracts customers by proving them various types of cuisine in different locations with exciting offers. Well providing such service comes with challenges. Challenges are not one but many like keeping enough number of stocks if demand increases suddenly, many a times demands even reduces at such cases all the raw material may go in waste, if that's not enough managing right number of staffs at all the locations in various city in different location, delivering food on time the list goes on.

If company can predict that what will be the number of orders they are going to receive, solving above mentioned problem will be quite easy. so Here comes the role of machine learning to make business grow more efficiently.

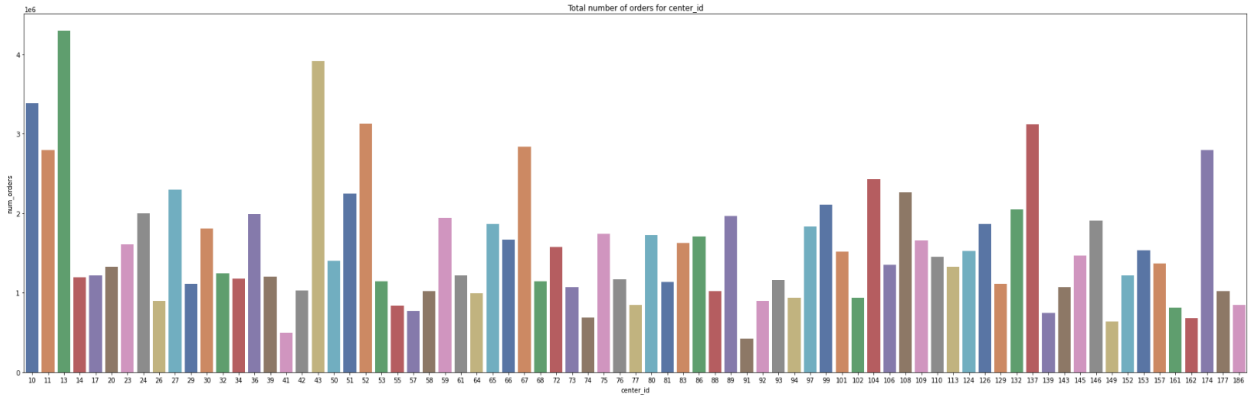
Exploration Data Analysis:

Number of Orders per week:

Total number of orders per week is important to analysis and predict the future orders.

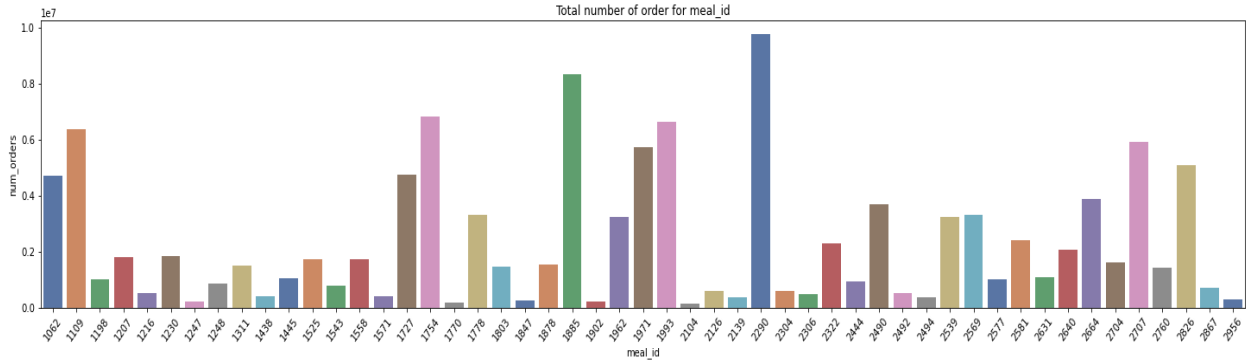
This will be one of the import aspects to prevent food wastage and making staff available in advance. From the above graph we can see that highest number of orders are in 45th to 50th week and lowest was in 61th week.

Number of orders per center id:



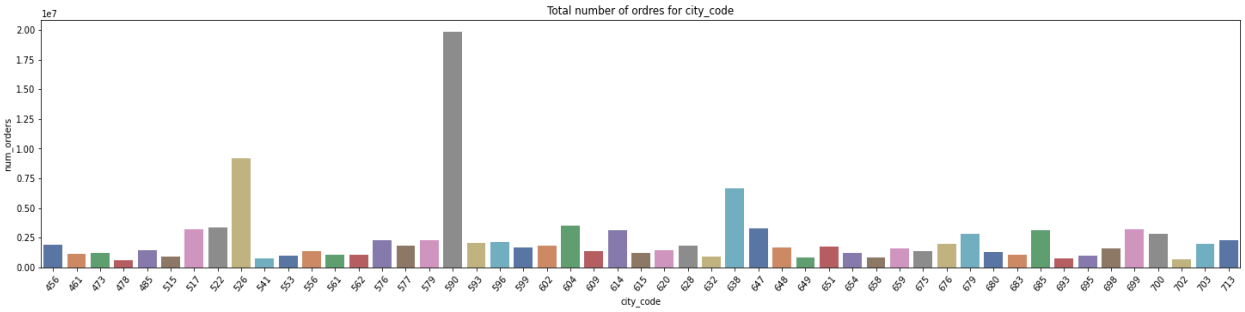
As there are many centres in different region and city it's very important to know from which centres are getting what orders. This will help us to manage the working staffs at various locations. From the graph we found that centre id 13 received highest order and lowest were received from 41 and 91.

Total number of orders based on meal id:



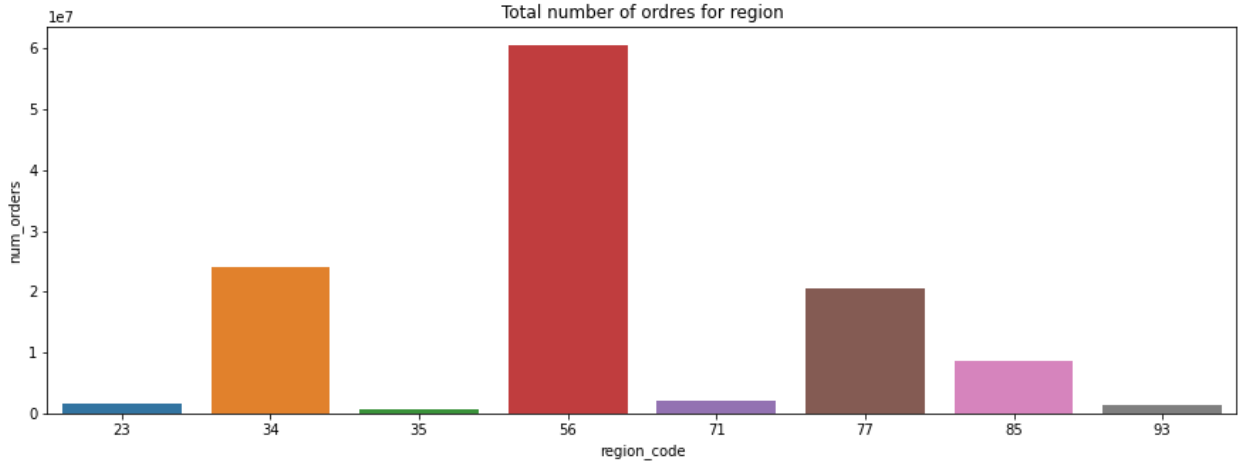
Which meal will be in demand if previous known will play an important role to reduce service time. We can see that we got highest number of orders from meal id 2290, 1885, 1993, 1754, 1109, 1971, 1727, 1062, 2707, 2826 and lowest from 1247, 1770, 1438, 1847, 1902, 2104, 2956.

Total number of orders based on city code:



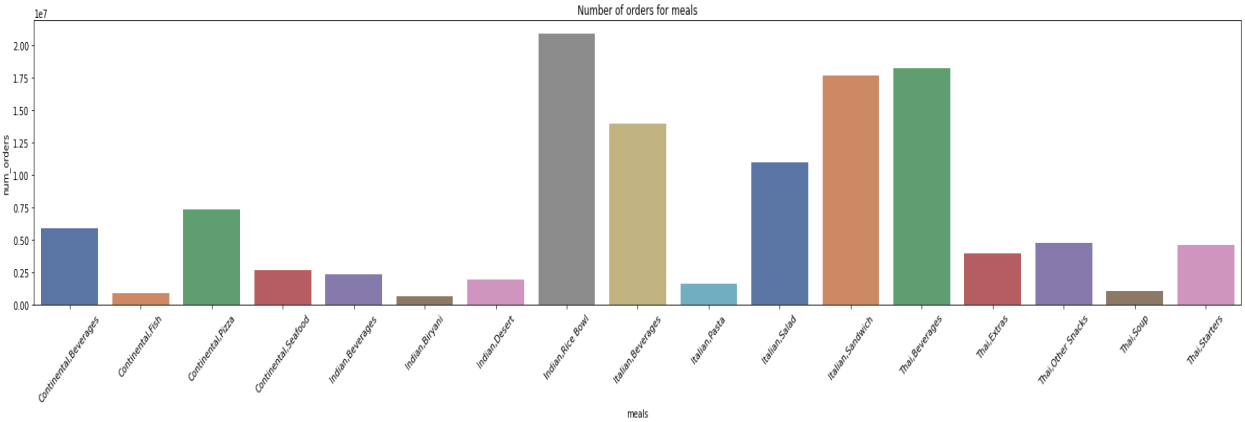
For any delivery company specially food its import to deliver their product on time. As time plays an important role in delivery service. Highest number of orders were received from city code 590, 526, 638 and lowest orders where from 478, 541, 702, 693.

Total number of orders based on region code:



For any company it's very important to know that from which geographical region are they earning more. Which directly helps them to provide facilities accordingly. Region code 56 was getting highest number of orders and lowest orders were reported from 35, 23, 93, 71.

Number of orders based on meals (category and cuisine):



Knowing which dishes does customers like or prefers to order is must for food business. If company is able to provide what customers like that to on time makes customer feel like satisfactory.

Indian (Rice Bowl), Italian (Beverages), Italian (Sandwich), Thai (Beverages), Italian (salad) were among most preferred meals and Continental (Fish), Indian (Biryani), Thai (Soup) was least preferred meals.

1. Beverages from Italian and Thai are most preferred unlike from Indian and continental.
2. Rice Bowl from India are in high demand.
3. Italian Sandwich is also most preferred.

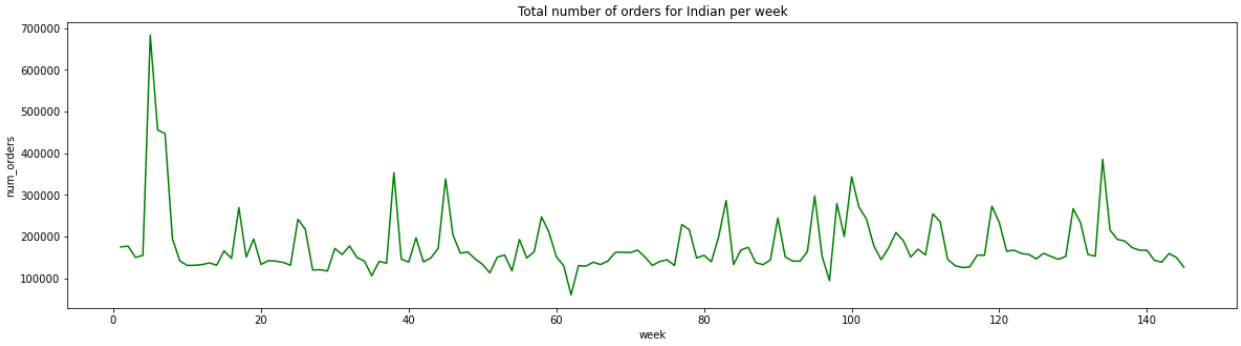
Number of orders based of cuisine:

The interplay between food and culture has huge global impact. The exchange of food is the exchange of culture, and thus understanding a place's food can be an incredible insight into their culture and lifestyle.

- The graph represents the number of orders based on the variety of cuisine.
- As per the graph it shows that Italian food is most popular among the customers.
- Indian and Thai is approximately same popular among the customers.

Below graphs represents Total number of orders based on different cuisine:

Indian:



Indian cuisine dates back over 5000 years. Each region has its own traditions, religions and culture that influence its food. By above graph we can see there was sudden upsurge in 5th – 10th week which was almost 7,00,000 after 10th week orders were fluctuating between 1,00,000 – 4,00,000. Lowest was recorded in 60th week which was Below 1,00,000 orders.

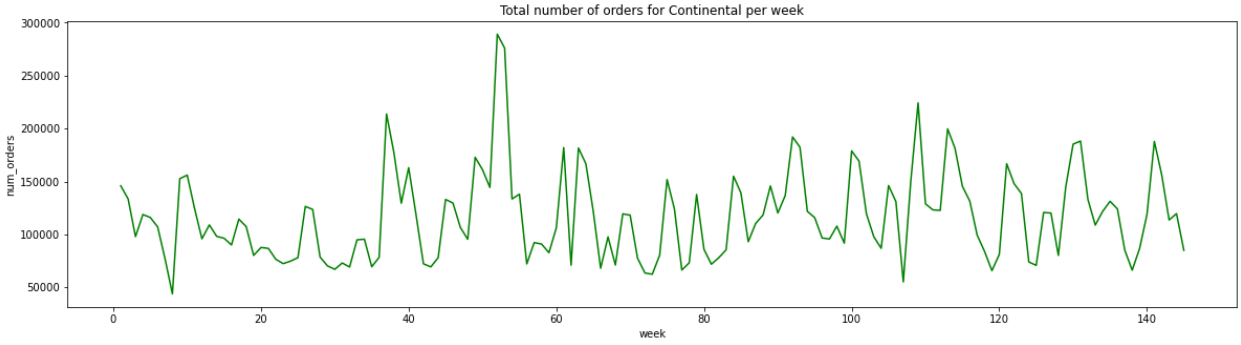
Thai:

With a diverse variety of flavors, Thai cuisine is one of the most popular cuisines in the world. From above line plot we can see that there was upward trend twice in between 45th week to 50th week and after which we also saw two down falls in orders in 61th week and 110th week.

Italian:

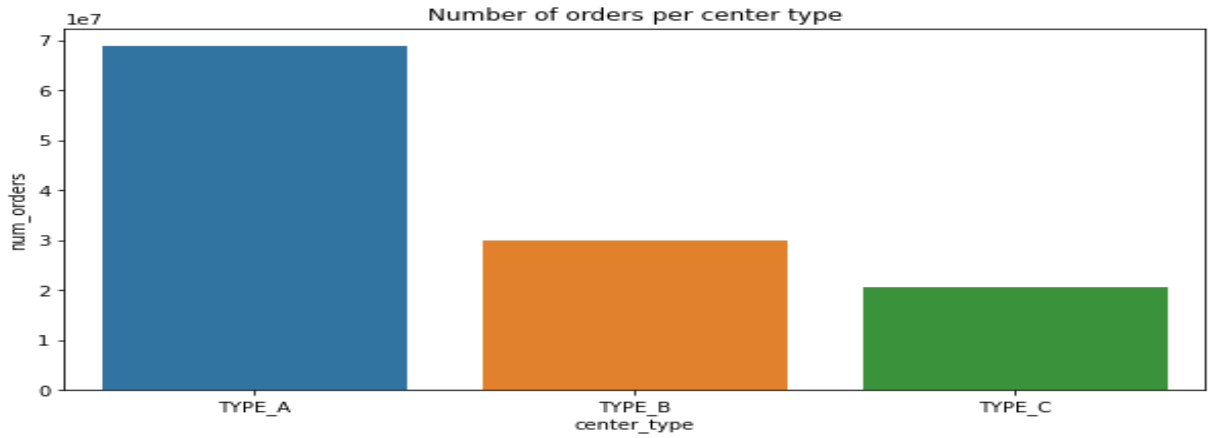
Italian food receives a huge fan base more than any other cuisines. From pasta to pizza Italian food is simply the best. In Italian cuisine we can observe many upsurges in different weeks and was recorded lowest in the 60th week.

Continental:



Continental cuisine has seen lowest order in 10th week which was below 50,000. Maximum number of orders was observed in 50th week which was approx. 3,00,000 orders.

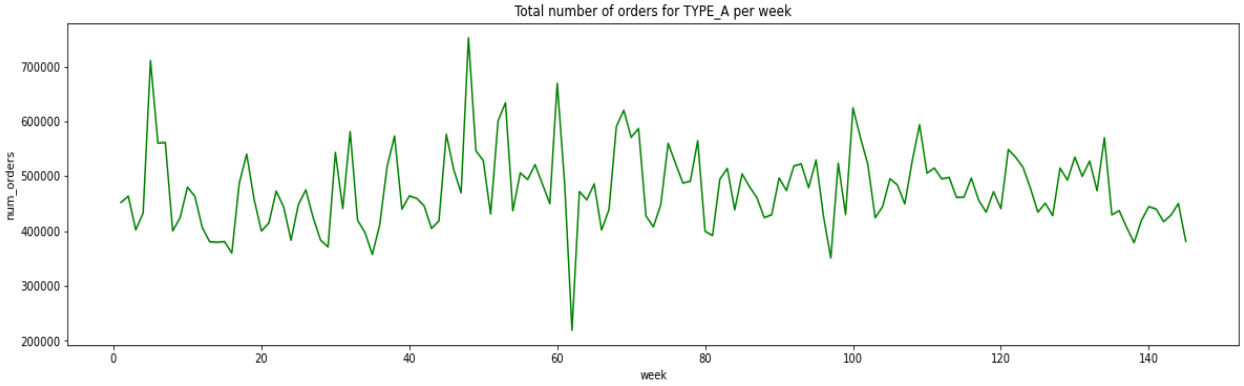
Total number of orders per center types:



- The given graph represents the order placed by the customer based on rating.
- This graph represents that from Type A restaurant is the most popular among the customers.
- It shows that customers' orders are dependent on the restaurant rating.

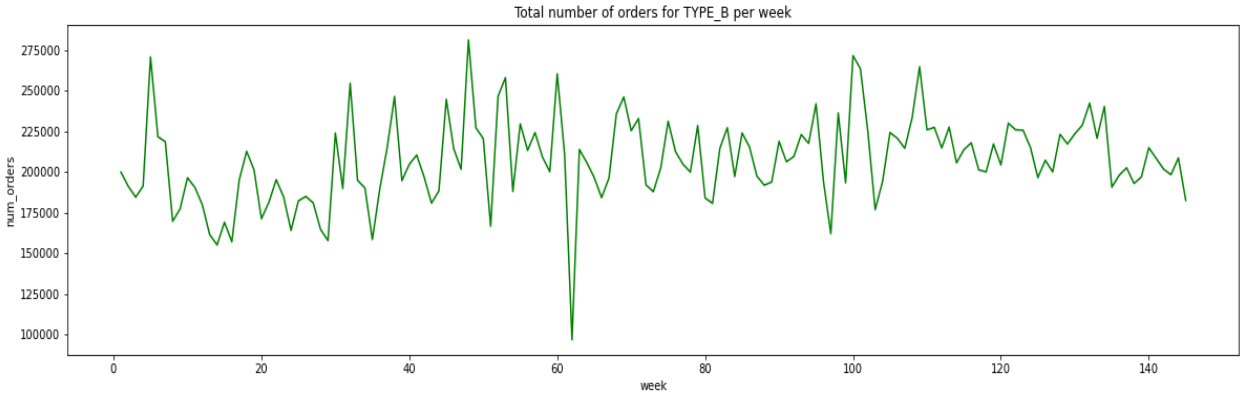
Total number of orders for different center types:

Type A:



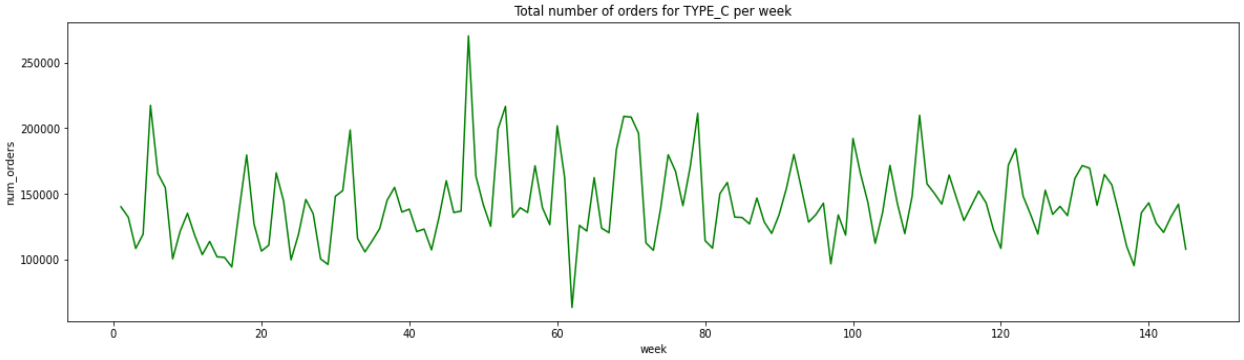
In 5th and 45th week we can see sudden spikes of orders. And in 61th week there was a deep-down fall. But most of orders are in 4,00,000 to 6,00,000.

Type B:



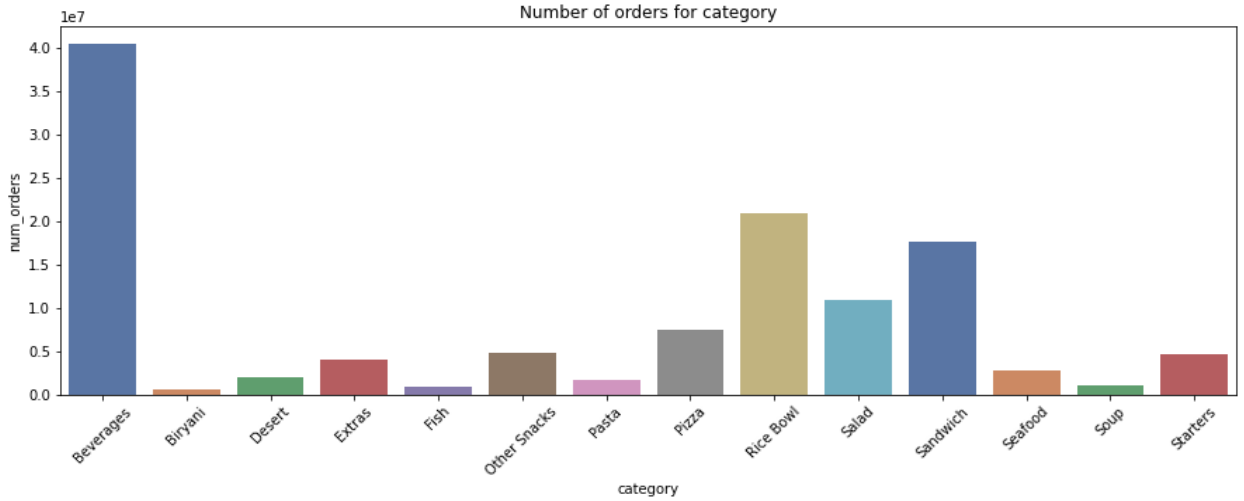
Total number of orders are mostly in 1,75,000 to 2,25,000. We can also see a reduced demand in 61th week.

Type C:



Almost total number of orders are in between 2,50,000 to 1,00,000. Highest among all was in 45th week and lowest was in 61th week.

Total number of orders for category:



Trying new thing is process of human psychology. Therefore, providing various categories of food is one of the important process to attract customers. If we get so many categories of food and that too from same restaurant, what could be better than this...

From the above bar graph, we can see that most ordered category was beverage and least ordered are biryani, fish, pasta, soup.

Total number of orders for different category:

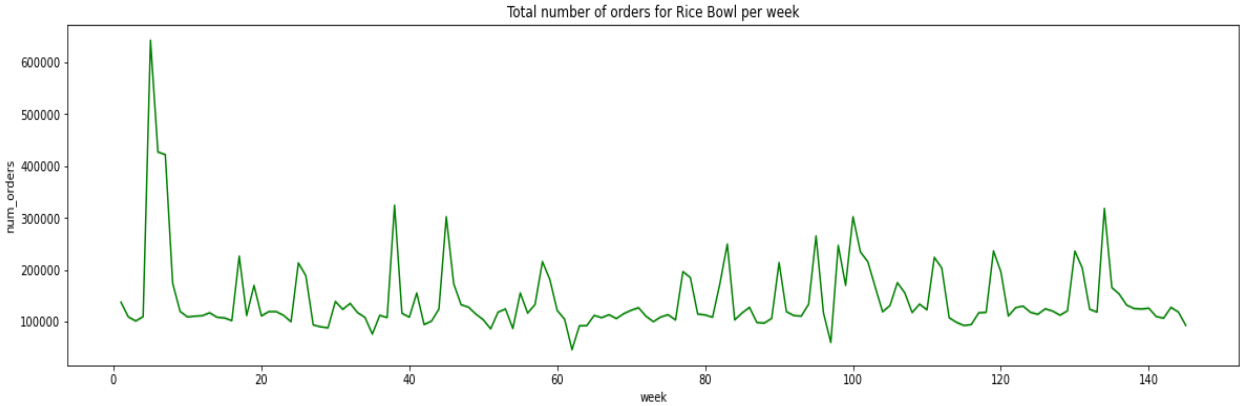
In every dining venue consumers are seeking excitement on the plate. If you're not challenging your customers' taste buds with international ideas, you're probably losing sales to someone who is

Beverages:

Beverages are an integral part of our lives and we select them for lots of different reasons – thirst, fun, taste, energy, hunger, boredom, to cool down and even warm up. Personal selection is based on your age, stage, lifestyle and the occasion. Many beverages are also part of our social occasions with family, friends and at special events and holidays.

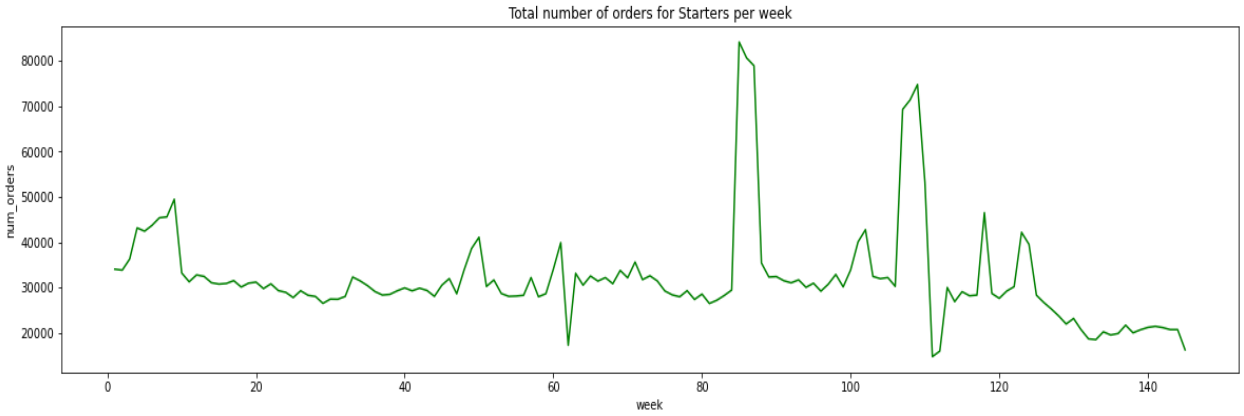
From above graph we can see that total number of Orders for beverage was mostly in between 2,50,000 to 3,50,000. But we can observe sudden hike from 45th to 50th week. And least orders were in 61th week.

Rice Bowl:



Rice is a revered staple in many cultures around the globe. Mastering the techniques of rice cookery can open the door to a rich assortment of dishes that are on trend with today's global tastes. Most of the orders for rice bowl can be seen in range of 1,00,000 to 3,00,000. There was a huge hike in orders in first ten weeks which was more than 6,00,000.

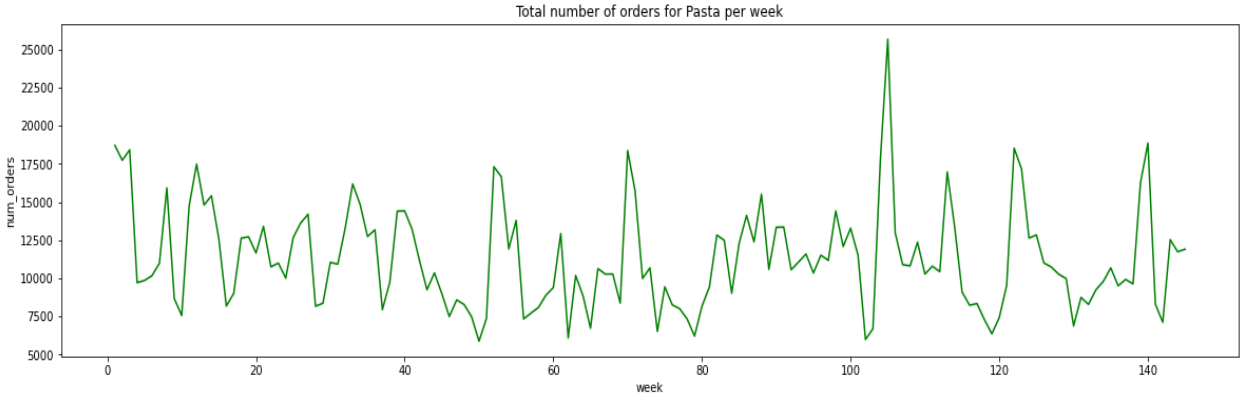
Starters:



Staters... also known as an appetizer, they are generally served at the beginning of the main course or served in cocktail parties. Different countries across the world have their own signature starters.

Orders for starters had seen spike in approx. 90th and 110th week. And least orders of starters were seen in 61th and 150th week.

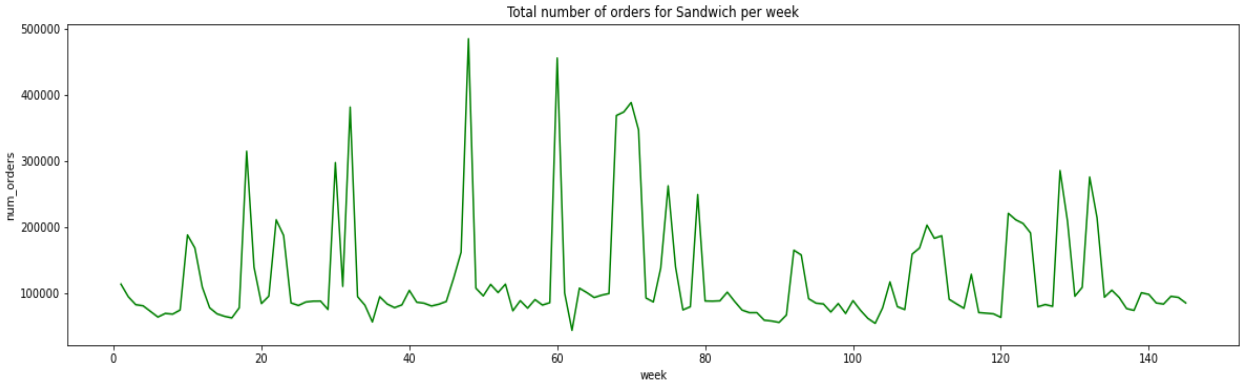
Pasta:



Pasta refers to the staple food of tradition Italian cuisine which is made using dough, water, eggs, vegetables, and oil.

Orders for pasta are in range of 7500 to 1,75,000 orders. Maximum number of orders were seen in 110th week. And most of the orders were in range 7,500 to 17,500.

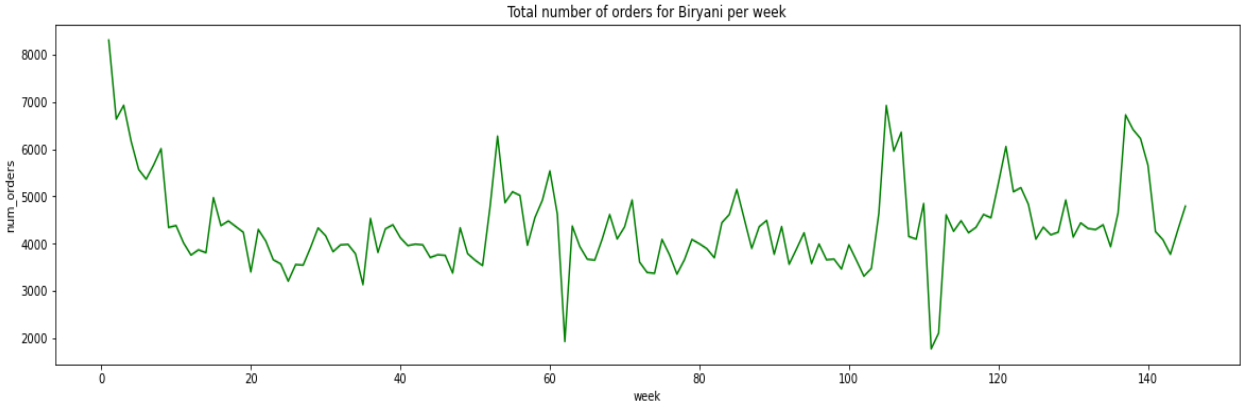
Sandwich:



Sandwich is easiest and cost wise cheapest product. Many people prefer sandwich in their breakfast. Even kids are given sandwich for their meals.

Orders for sandwich are hiked at various occasion but mostly it was in range 1,00,000 to 3,00,000. Highest number of orders were in 45th , 60th ,65th and 30th week.

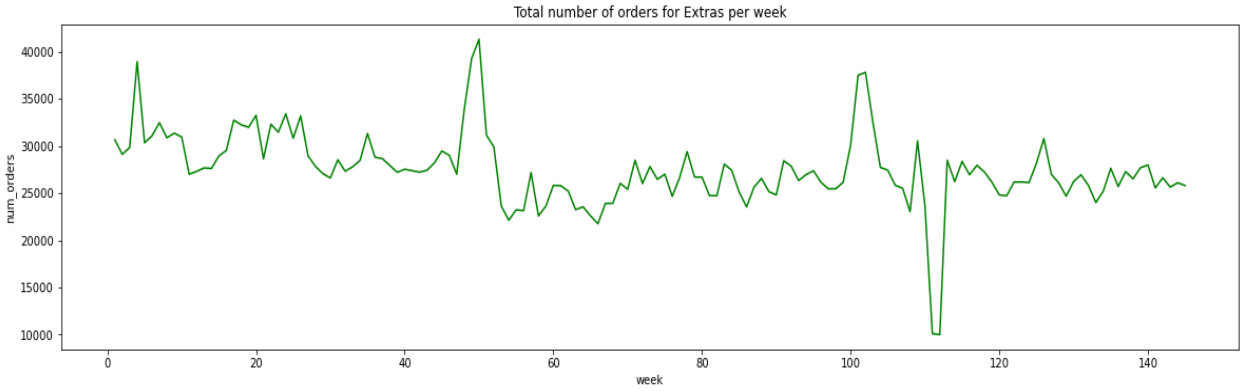
Biryani:



Different regions of India prepare their own versions of biryani, but its essence remains the same everywhere. What makes biryani such a special dish is its unassuming look, which actually is quite deceiving until you take a bite of it. A simple-looking plate of colored rich is a melting pot of a myriad of flavors from spices, vegetables and meats.

From the above line plot of biryani, we can clearly see how demand has decreased from 1st week. It has been stable for some weeks in the range 400 to 600 number of orders. Least number of orders were seen in 61th and 110th week.

Extras:

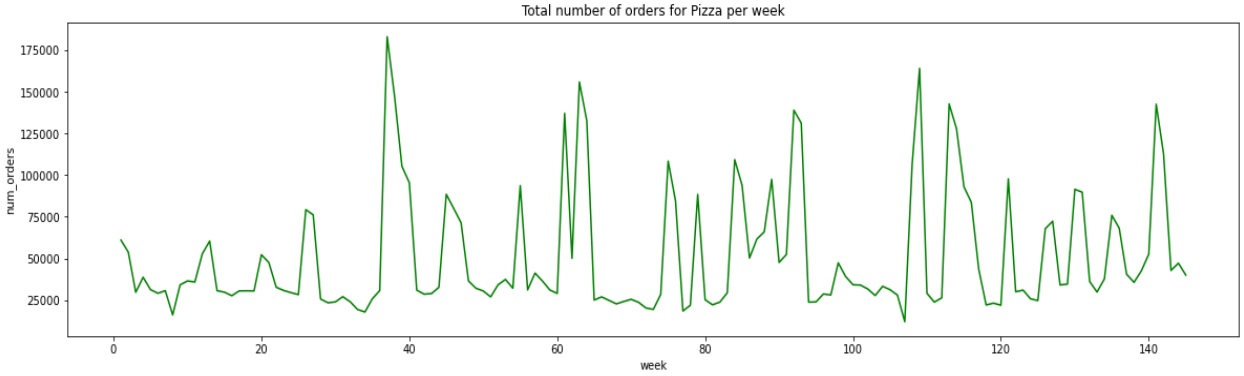


The extras are those food items that are not essential, but add spice to eating.

Demand for extras has also seen slight down fall from 1st week. It was least in 110th week.

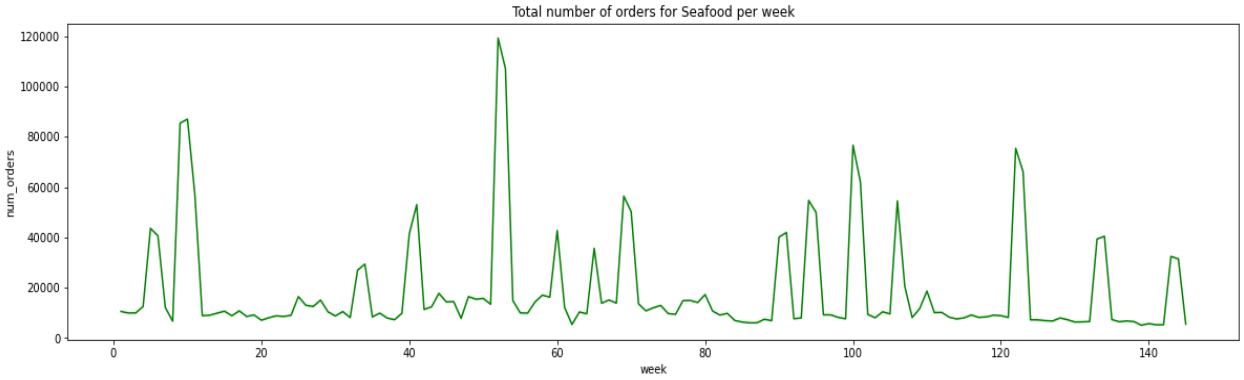
Highest number of orders were recorded in 45th week.

Pizzas:



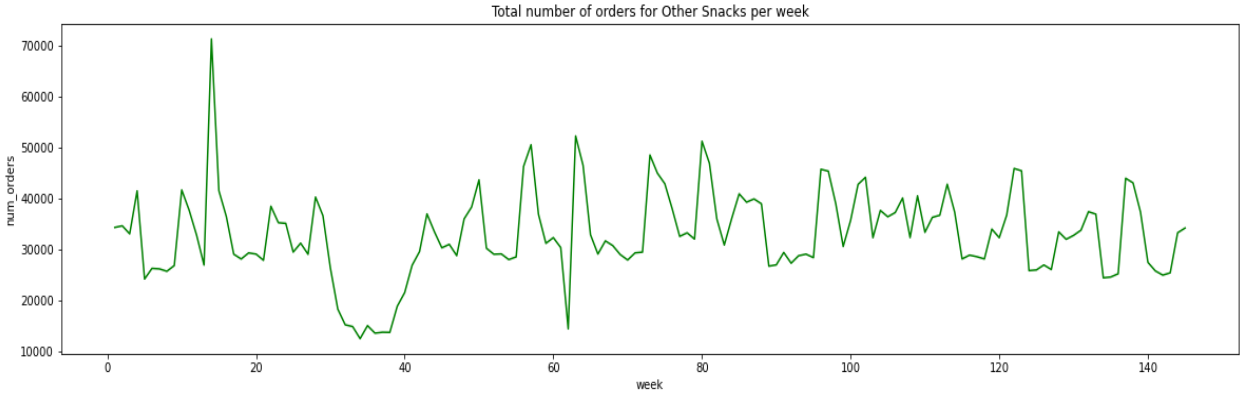
Pizza is one of the few foods that almost everyone knows and loves around the world. It's cheesy, delicious, and just about everywhere. Pizzas were in high demand on many occasions. Highest number of orders were approx. 1,75,000 orders. Although lowest number of orders was 25,000.

Seafood:



Seafood is a high-protein food that is low in calories, total fat, and saturated fat. High in vitamins and minerals, seafood has been shown to have numerous health benefits. For example, recent studies have shown that eating seafood can decrease the risk of heart attack, stroke, obesity, and hypertension. Seafood also provides essential nutrients for developing infants and children. Seafood was also seen hikes in several weeks. Highest orders were observed in 50th week which was nearly 1,20,000.

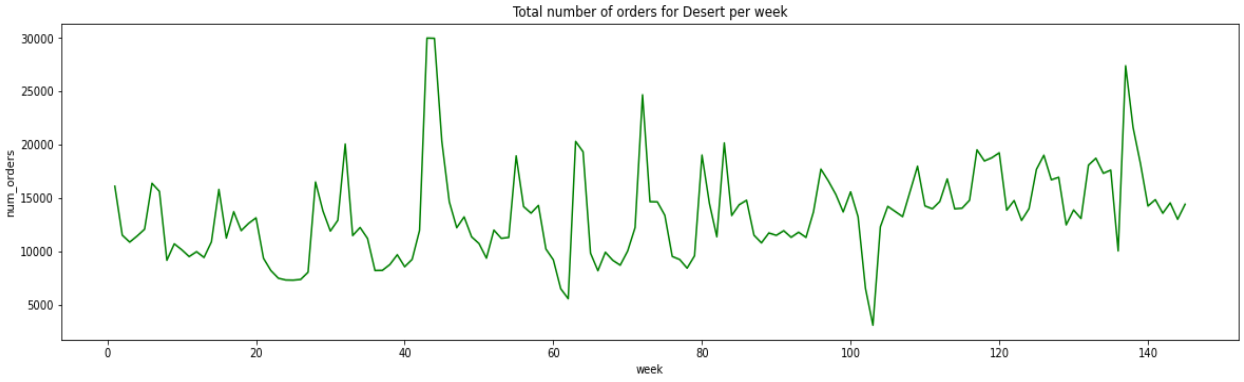
Snacks:



A snack is a portion of food often smaller than a regular meal, generally eaten between meals. Snacks come in a variety of forms including packaged and processed foods and items made from fresh ingredients at home.

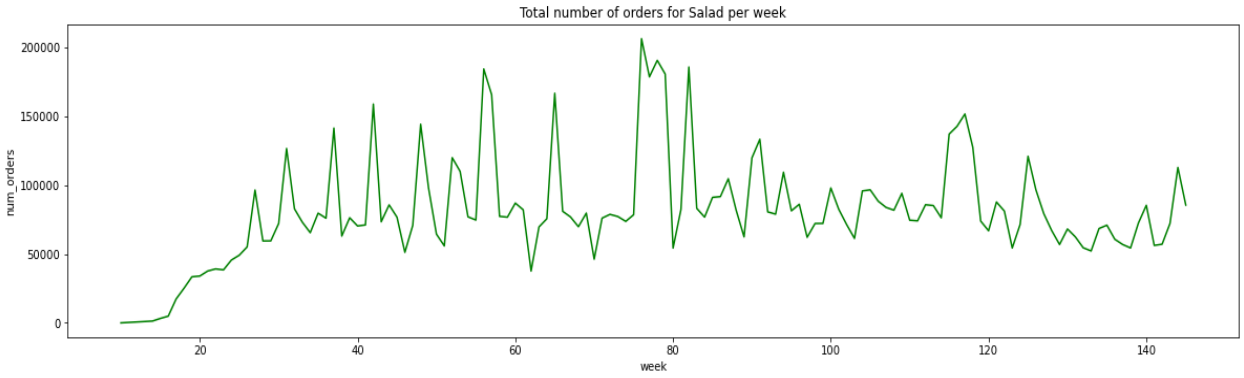
Total number of snacks order are in range from 30,000 to 45,000. Demand was hiked in 15th to 20th week. And were dropped in 30th to 40th week.

Decent:



Decent were having order by approx. 10,000 to 20,000 people. Although highest were in 42th to 43th week which was around 30,000.

Salad:

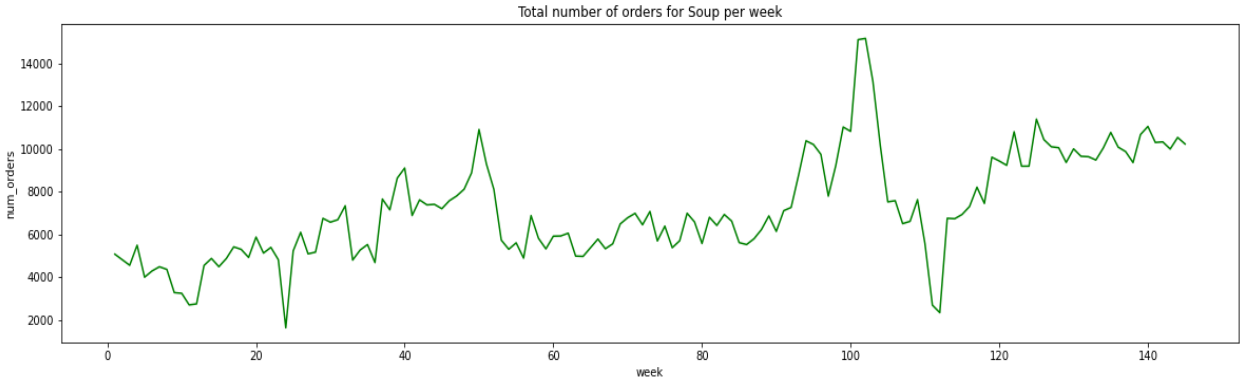


Salad Ready is a consumer food product fresh, washed, bite-sized, detached leaves of leaf lettuce, contained in a sealed, transparent, polypropylene bag. As most of the people are now health conscious. We can see how demand of salad increased over time from zero order at beginning to 2,00,000 in 75th to 80th week. And continues to be nearby 1,00,000.

Fish:

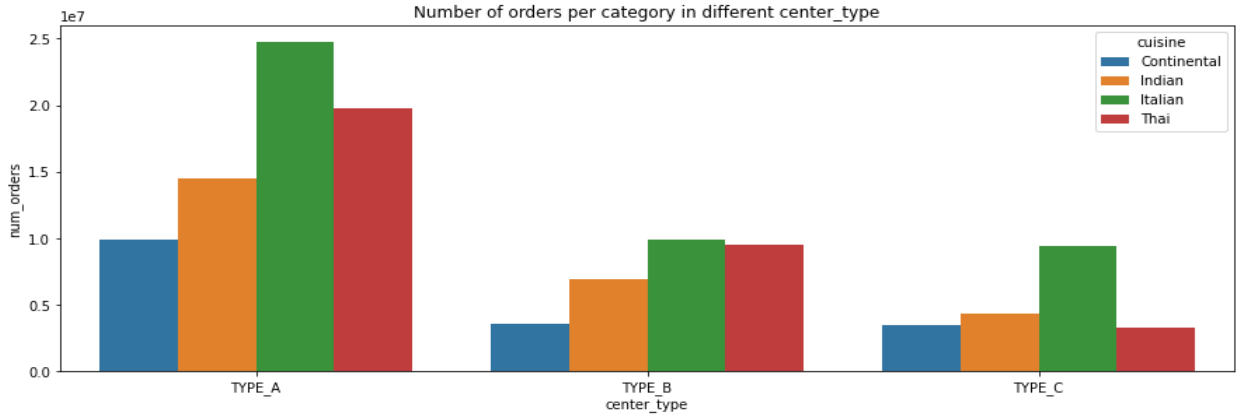
Fish is also one among the health food preferred by many people. Same as salad, demand of fish has also grown over a period of time it reached its peak in 130th to 135th week. And still continues to be in demand with over 15,000 ordering it.

Soup:



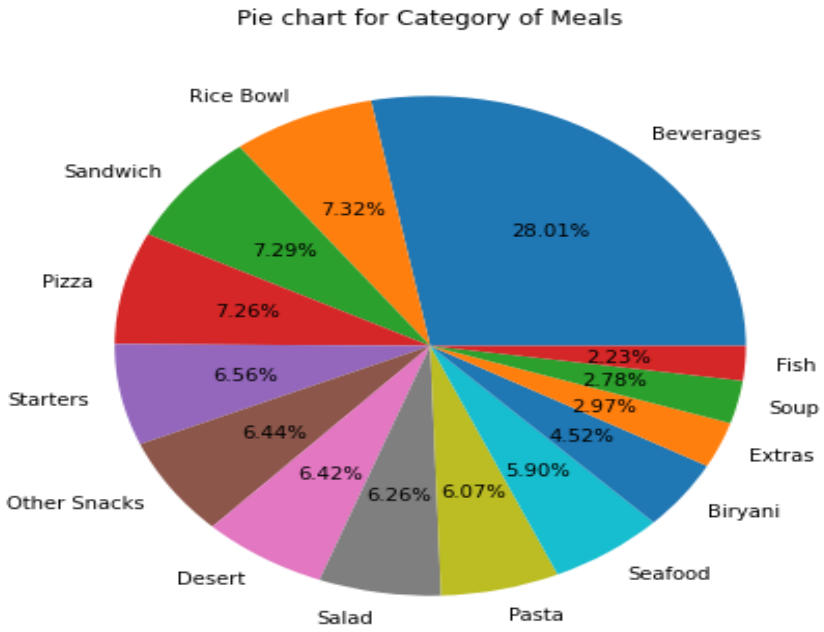
Combinations of meats, vegetables and grains are popular soup ingredients, and dairy products likewise play a critical role in many recipes. Soups with cream, milk or cheese bases, as well as those garnished with cheese or sour cream, make up a large segment of menu offerings in restaurants. Demand of soup have seen increase according to time from 5000 to 15000 its shows peoples love towards soup.

Number of orders based on center type and category:



- The data of cuisine like by the customer on the restaurant based on their rating.
- Italian food is most popular from Type A restaurant.
- Type C restaurant is not liked by the customers very much.

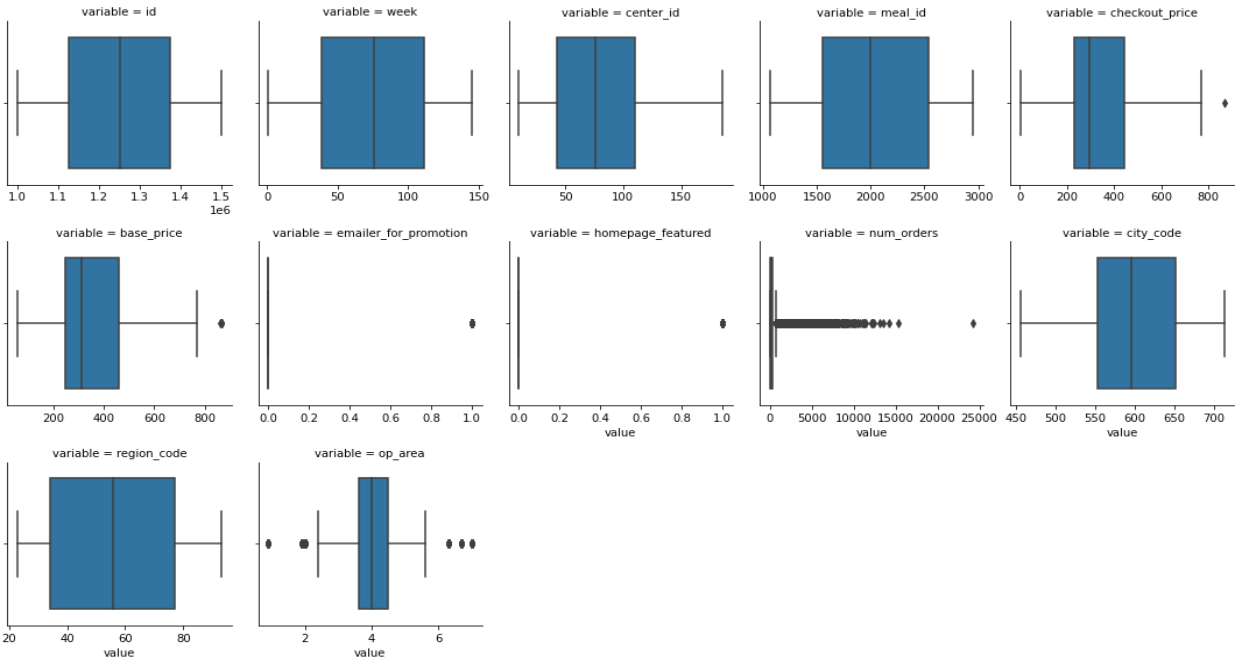
Pie chart for category of meals:



From the pie chart we found out that hold of Beverages is more than quarter that is 28.01%. Rice bowl, Sandwich and Pizza has almost equal share.

We found that Fish is the least sold item among others.

Checking Outliers:



- The box plot diagram used to check the outlier present in each of the value.

- Only Base Price, Checkout Price, Op area are having outliers.

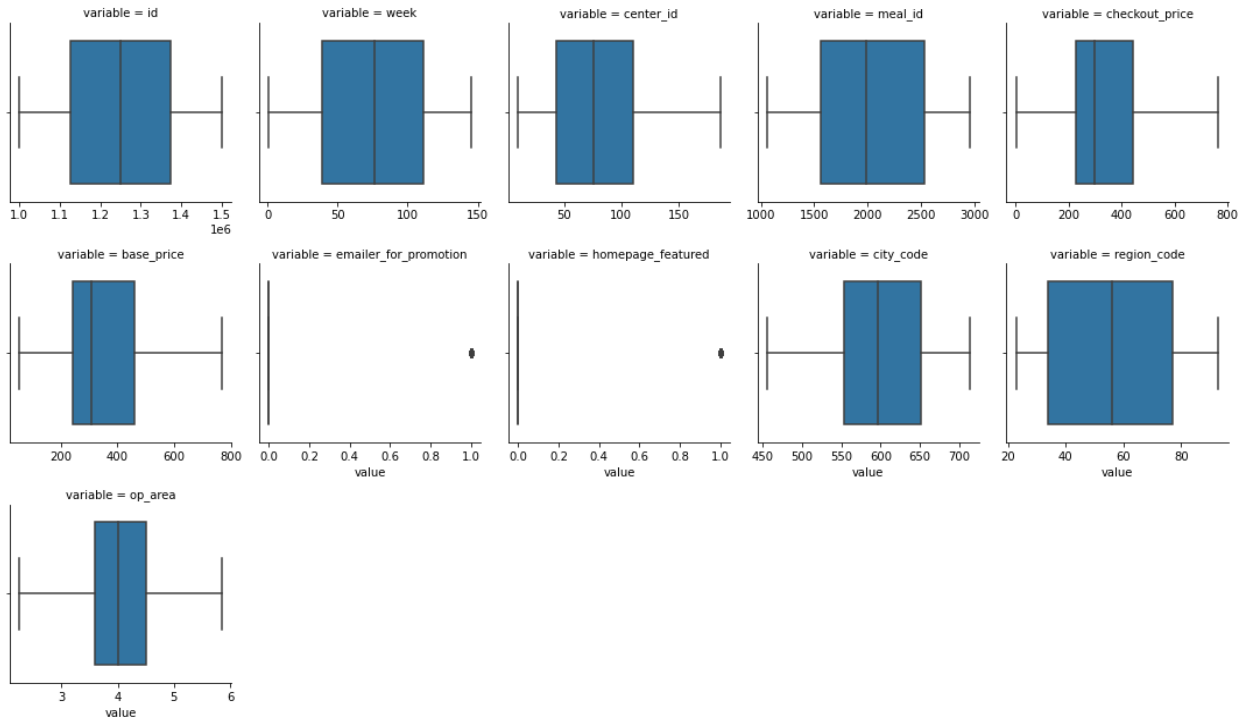
Treating Outliers:

Outlier Analysis is a process that involves identifying the anomalous observation in the dataset. Outliers are caused due to the incorrect entry or computational error, is-reporting, sampling error, Exceptional but true value error. bad data, wrong calculation, these can be identified as Outliers and should be dropped but at the same time you might want to correct them too, as they change the level of data i.e. mean which cause issues when you model your data.

Using IQR method for removing the outlier:

The interquartile range (IQR), also called the mid-spread or middle 50%, or technically H-spread, is a measure of statistical dispersion, being equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles, IQR = Q3 − Q1. In other words, the IQR is the first quartile subtracted from the third quartile; these quartiles can be clearly seen on a box plot on the data. It is a measure of the dispersion similar to standard deviation or variance, but is much more robust against outliers.

After removing the outliers:



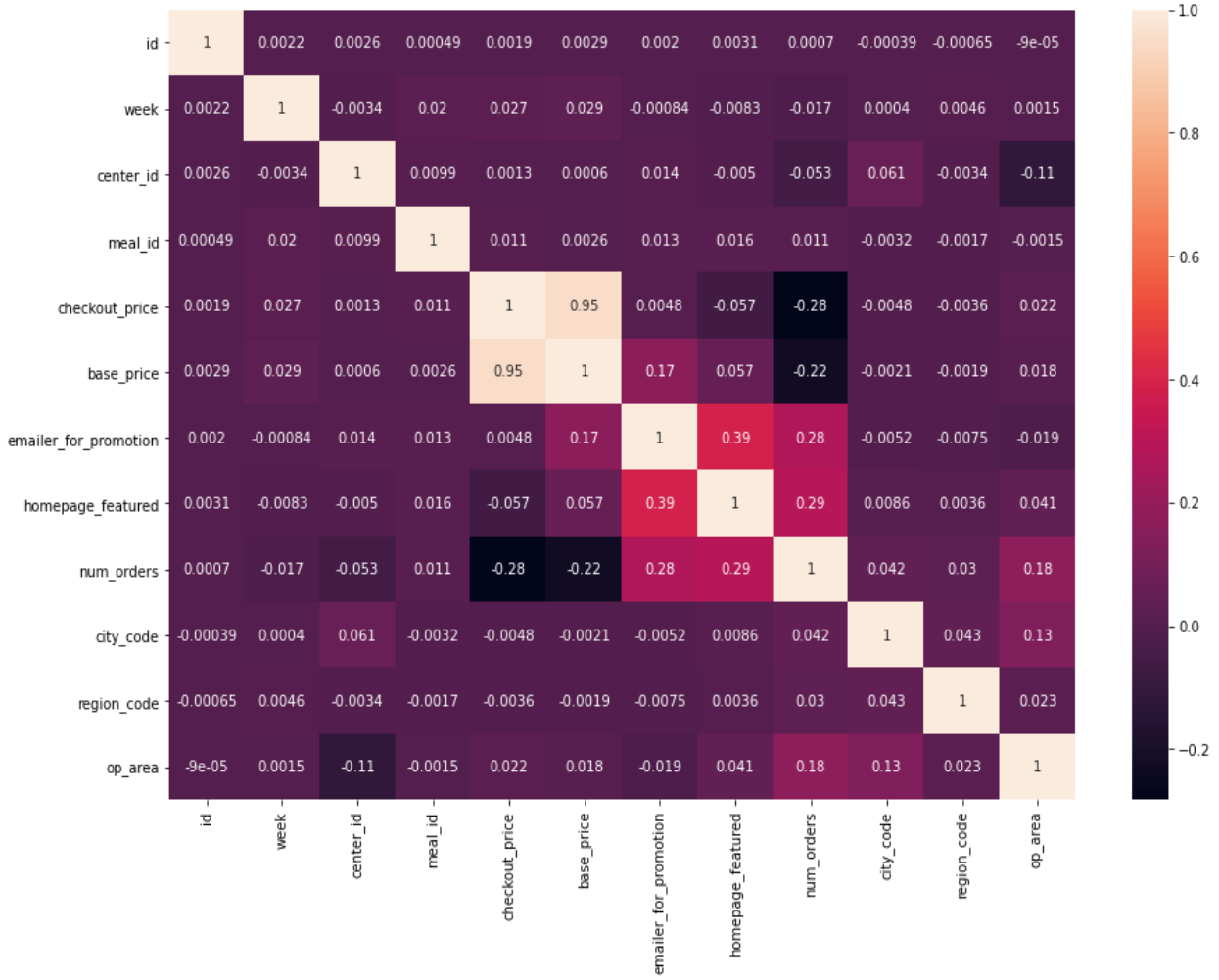
Heatmap Visualizing Correlation:

Multicollinearity occurs when independent variables in a regression model are correlated. This correlation is a problem because independent variables should be independent. If the degree of correlation between variables is high enough, it cause problems when you fit the model and interpret the result.

Multicollinearity causes the following two basic types of problems:

- The coefficient becomes very sensitive to small changes in the model.
- Multicollinearity reduces the precision of the estimate coefficients, which weakens the statistical power of your regression model.

Now you can't trust the p-value to select the independent variables to include in the model.



When checking for multicollinearity among the independent variables, we can first take a look at the correlation matrix to understand if there is any high correlation between continuous variables. From above heatmap we can see that base price and checkout price are highly correlated to each other which is not good for our model as it will lead to multicollinearity which will down grade the performance of our model.

Statistical test:

We performed some hypothesis testing on our data to

Shapirp-Wilk test:

The Shapiro-Wilk test examines if a variable is normally distributed in some population.

H0: population is normally distributed

H1: population is not normally distributed

from SciPy. Stats import shapiro

for i in data:

print(i,'\t',shapiro(data[i])

P_value is too small so we reject H0.

Conclusion: data are not normally distributed

Chi2 test:

Chi-squared tests of independence determine whether a relationship exists between two categorical variables. Do the values of one categorical variable depend on the value of the other categorical variable? If the two variables are independent, knowing the value of one variable provides no information about the value of the other variable.

HO= features are independent(category, cuisine)

H1= features are dependent

i) cat_cuis=pd.crosstab(data['category'],data['cuisine'])

import scipy.stats as st

print(st.chi2_contingency(cat_cuis))

P_value is too small so we reject H0

Conclusion: data is dependent i.e. data is statistically significant.

HO= features are independent(center type, category)

H1= features are dependent

ii) type_cat=pd.crosstab(data['center_type'],data['category'])

print(st.chi2_contingency(type_cat))

P_value is too small so we reject H0

Conclusion: data is dependent i.e. data is statistically significant.

HO= features are independent(center type, cuisine)

H1= features are dependent

iii)type_cat=pd.crosstab(data['center_type'],data['cuisine'])

print(st.chi2_contingency(type_cat))

P_value is too small so we reject H0

Conclusion: data is dependent i.e. data is statistically significant.

Kruskal-Wallis test:

- First pool all the data across the groups.
- Rank the data from 1 for the smallest value of the dependent variable and next smallest variable rank 2 and so on… (if any value ties, in that case it is advised to use mid-point), N being the highest variable.
- Compute the test statistic
- Determine critical value from Chi-Square distribution table
- Finally, formulate decision and conclusion

Calculation of the Kruskal-Wallis Non Parametric Hypothesis Test:

The Kruskal–Wallis Non Parametric Hypothesis Test is to compare medians among k groups (k > 2). The null and alternative hypotheses for the Kruskal-Wallis test are as follows:

H0: Population medians are equal

H1: Population medians are not all equal

As explained above, the procedure for Kruskal-Wallis test pools the observations from the k groups into one combined sample, and then rank from lowest to highest value (1 to N), where N is the total number of values in all the groups.

The test statistic for the Kruskal Wallis test (mostly denoted as H) is defined as follows:

Where Ti = rank sum for the ith sample i = 1, 2,…,k

In Kruskal-Wallis test, the H value will not have any impact for any two groups in which the data values have same ranks. Either increasing the largest value or decreasing the smallest value will have zero effect on H. Hence, the extreme outliers (higher and lower side) will not impact this test.

i)

H0: Population means are equal

H1: Population means are not equal

st.kruskal(typeA['num_orders'],typeB['num_orders'],typeC['num_orders'])

as p value is too small we reject H0

Conclusion: Population means are not equal.

ii) H0: Population means are equal

H1: Population means are not equal

st.kruskal(Indain['num_orders'],Italian['num_orders'],Thai['num_orders'],Continental['num_orders'])

as p value is too small we reject H0

Conclusion: Population means are not equal.

iii)

H0: Population means are equal

H1: Population means are not equal

st.kruskal(Beverages['num_orders'],Rice_Bowl['num_orders'],Sandwich['num_orders'],Pizza['num_orders'],Starters['num_orders'],Other_Snacks['num_orders'],Desert['num_orders'],Salad['num_orders'],Pasta['num_orders'],Seafood['num_orders'],Biryani['num_orders'],Extras['num_orders'],Soup['num_orders'],Fish['num_orders'])

as p value is too small, we reject H0

Conclusion: Population means are not equal.

Feature Engineering:

In Feature Engineering we must select appropriate input data comprising features which is in the form of structured columns. Machine learning algorithms require features with specific characteristics to work properly. Here, the need for feature engineering arises.

- Preparing the proper input dataset, compatible with machine learning algorithm requirements.
- It improves the performance of machine learning algorithms.
- Few variables may not be in the form that we expect. Those variables will be transformed using various techniques.
- Non-numerical categorical variables will be transformed using One Hot Encoding or Label Encoding as appropriate
- Split the dataset into two parts: Training dataset and testing dataset.

Feature encoding:

Our machine learning algorithm can only read numerical values. It is essential to encoding categorical features into numerical values. In our dataset we have 3 categorical features (category, cuisine, center type).

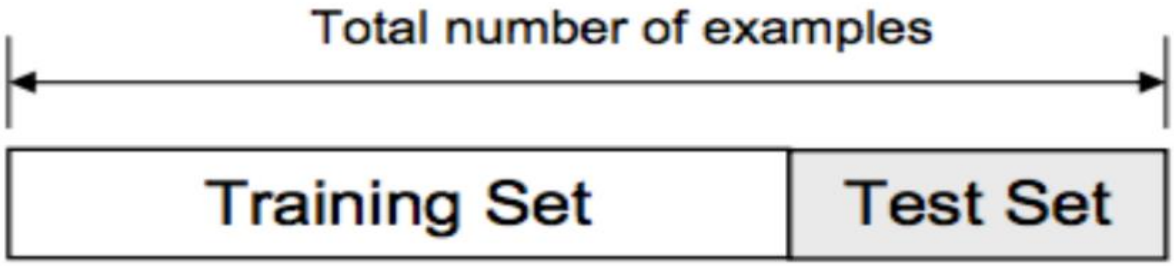
Category and cuisine are nominal data we used pandas get dummies method to encode categorical data to numeric.

Since center type is based on rating of restaurant and food. It's an ordinal data where Type A means very good, Type B means good. And Type C means average.

So, to encode ordinal data (center Type) we used label encoding method.

Train test split:

The data we use is usually split into training data and test data. The training set contains a known output and the model learns on this data in order to be generalized to other data later on. We have the test dataset (or subset) in order to test our model's prediction on this subset.



X = data.Drop('num_orders',axis=1) # Independent data

y = data['num_orders'] # dependent data

from sklearn.model_selection import train_test_split

X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.3,random_state = 1)

import scipy.stats as stats

H0: Both train test represents data

H1: Both train test does not represent data

stats.f_oneway(y_train,y_test,y)

as p_value > alpha we accept H0. i.e., both train test represents data

Transformation:

Some machine learning algorithms like linear regression and logistic regression explicitly assume the real-valued variables have a Gaussian distribution. Other nonlinear algorithms may not have this assumption, yet often perform better when variables have a Gaussian distribution. There are data preparation techniques that can be used to transform each variable to make the distribution Gaussian, or if not Gaussian, then more Gaussian like. These transforms are most effective when the data distribution is nearly-Gaussian to begin with and is afflicted with a skew or outliers. Another common reason for transformations is to remove distributional skewness. An un-skewed distribution is one that is roughly symmetric. This means that the probability of falling on either side of the distribution's mean is roughly equal.

Then a *Power Transformer* is used to make the data distribution more-Gaussian and standardize the result, centering the values on the mean value of 0 and a standard deviation of 1.0. After train test split. As our data is not normal, we tried to make is near to normal by using power transformation.

Scaling dataset

Machine learning algorithm just sees number — if there is a vast difference in the range say few ranging in thousands and few ranging in the tens, and it makes the underlying assumption that higher ranging numbers have superiority of some sort. So these more significant number starts playing a more decisive role while training the model.The machine learning algorithm works on numbers and does not know what that number represents. A weight of 10 grams and a price of 10 dollars represents completely two different things — which is a no brainer for humans, but for a model as a feature, it treats both as same.

Another reason why feature scaling is applied is that few algorithms like Neural network gradient descent converge much faster with feature scaling than without it.

Transform features by scaling each feature to a given range. This estimator scales and translates each feature individually such that it is in the given range on the training set, e.g., between zero and one. This Scaler shrinks the data within the range of -1 to 1 if there are negative values. We can set the range like [0,1] or [0,5] or [-1,1].

This Scaler responds well if the standard deviation is small and when a distribution is not Gaussian. This Scaler is sensitive to outliers.

Model testing (OLS base model):

Ordinary least squares (OLS) regression is a statistical method of analysis that estimates the relationship between one or more independent variables and a dependent variable. The method estimates the relationship by minimizing the sum of the squares of the difference between the observed and predicted values of the dependent variable configured as a straight line. OLS regression is used in bivariate model, that is, a model in which there is only one independent variable (X) predicting a dependent variable (Y). However, the logic of OLS regression can also be used in multivariate model in which there are two or more independent variables.

Linear regression is a method of finding the best straight line fitting to the given data, i.e. finding the best linear relationship between the independent and dependent variables.
In technical terms, linear regression is a machine learning algorithm that finds the best linear-fit relationship on any given data, between independent and dependent variables. It is mostly done by the Sum of Squared Residuals Method.

We choose linear model as its the basic and easy model. So its better to start with linear regression.

The important assumptions in linear regression analysis are:

- There should be a linear and additive relationship between dependent (response) variable and independent (predictor) variable(s). A linear relationship suggests that a change in response Y due to one unit change in X is constant, regardless of the value of X. An additive relationship suggests that the effect of X on Y is independent of other variables.
- There should be no correlation between the residual (error) terms.
- The independent variables should not be correlated.
- The error terms must have constant variance. This phenomenon is known as homoskedasticity.
- The error terms must be normally distributed.

OLS Regression Results

Dep. Variable:	num_orders	R-squared:	0.414
Model:	OLS	Adj. R-squared:	0.414
Method:	Least Squares	F-statistic:	8074.
Date:	Wed, 30 Jun 2021	Prob (F-statistic):	0.00
Time:	22:56:52	Log-Likelihood:	-2.2807e+06
No. Observations:	319583	AIC:	4.561e+06
Df Residuals:	319554	BIC:	4.562e+06
Df Model:	28		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	262.2379	0.538	487.409	0.000	261.183	263.292
id	0.0513	0.538	0.095	0.924	-1.003	1.106
week	-4.0012	0.543	-7.370	0.000	-5.065	-2.937
center_id	-19.8001	0.551	-35.939	0.000	-20.880	-18.720
meal_id	6.4563	0.629	10.259	0.000	5.223	7.690
checkout_price	-157.7769	2.284	-69.079	0.000	-162.254	-153.300
base_price	104.3678	2.690	38.801	0.000	99.096	109.640
emailer_for_promotion	65.5073	0.667	98.237	0.000	64.200	66.814
homepage_featured	71.4668	0.611	116.979	0.000	70.269	72.664
city_code	7.7104	0.548	14.082	0.000	6.637	8.784
region_code	11.7953	0.540	21.844	0.000	10.737	12.854
op_area	77.2377	0.557	138.670	0.000	76.146	78.329
category_Beverages	11.3135	0.904	12.519	0.000	9.542	13.085
category_Biryani	24.5322	0.710	34.553	0.000	23.141	25.924
category_Desert	26.7734	0.748	35.817	0.000	25.308	28.238
category_Extras	-34.6641	0.638	-54.333	0.000	-35.915	-33.414
category_Fish	-0.0508	0.599	-0.085	0.932	-1.224	1.123
category_Other Snacks	-74.5778	0.684	-108.962	0.000	-75.919	-73.236
category_Pasta	-74.9368	0.708	-105.863	0.000	-76.324	-73.549
category_Pizza	9.6468	0.712	13.548	0.000	8.251	11.042
category_Rice Bowl	166.5558	0.705	236.327	0.000	165.174	167.937
category_Salad	3.1385	0.672	4.667	0.000	1.821	4.456
category_Sandwich	35.0611	0.668	52.507	0.000	33.752	36.370
category_Seafood	-11.7994	0.694	-17.002	0.000	-13.160	-10.439
category_Soup	-63.9186	0.603	-106.067	0.000	-65.100	-62.737
category_Starters	-69.4553	0.700	-99.194	0.000	-70.828	-68.083
cuisine_Continental	-30.4345	1.161	-26.214	0.000	-32.710	-28.159
cuisine_Indian	-111.3769	0.805	-138.363	0.000	-112.955	-109.799
cuisine_Italian	56.2645	0.773	72.753	0.000	54.749	57.780
cuisine_Thai	81.6734	1.041	78.485	0.000	79.634	83.713
center_type	-2.8379	0.552	-5.139	0.000	-3.920	-1.756

Omnibus:	481828.670	Durbin-Watson:	1.996
Prob(Omnibus):	0.000	Jarque-Bera (JB):	820012561.087
Skew:	8.860	Prob(JB):	0.00
Kurtosis:	250.522	Cond. No.	6.15e+15

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 3.14e-26. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

As from the above ols model we can see our model is very poor with r square value only 0.414.

By using statistical test we have seen that our data is not normally distributed. So its a voilation of assuptions of linear model. But from the above ols summary we can figure out that some of the features are insignificant.

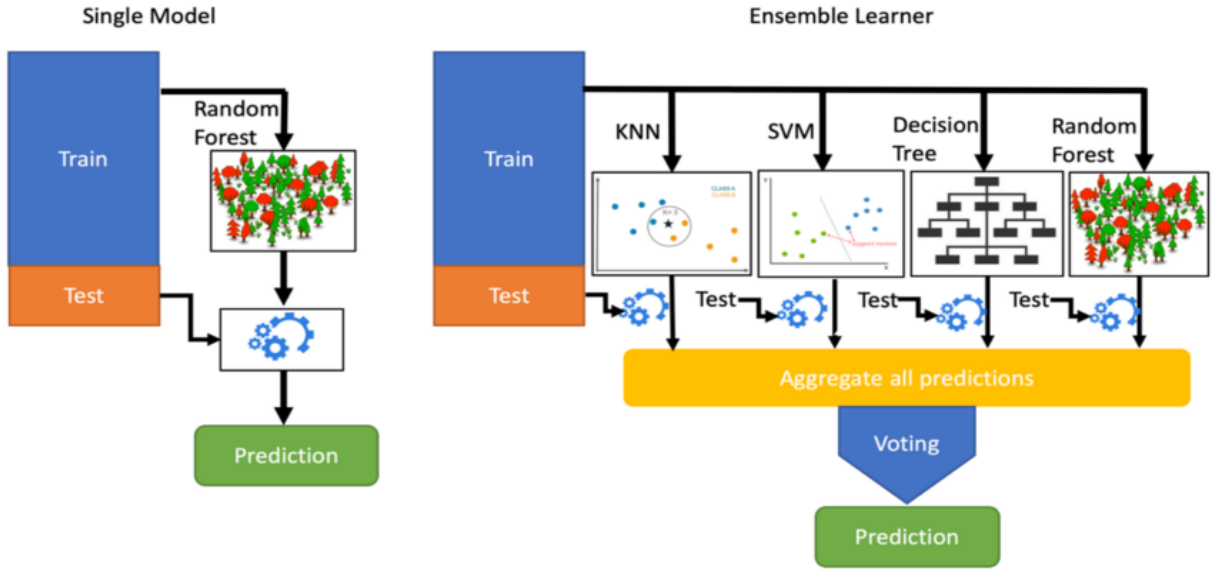
Removing insignificant features:

Base price is highly correlated with checkout price. The purpose of base price is served by checkout price so we delete base price.

id, category Fish, category Salad. Were the statistically insignificant features.

Non linear model:

Ensemble learning is a process in which decisions from multiple machine learning models are combined to reduce errors and improve prediction when compared to a Single ML model. Then the maximum voting technique is used on aggregated decisions to deduce the final prediction.



Types of Ensemble learning:

Ensemble learning methods can be performed in two ways:
- Bagging (parallel ensemble)
- Boosting (sequential ensemble)

Working of boosting algorithm:

The boosting algorithm creates new weak learners (models) and sequentially combines their predictions to improve the overall performance of the model. For any incorrect prediction, larger weights are assigned to misclassified samples and lower ones to samples that are

correctly classified. Weak learner models that perform better have higher weights in the final ensemble model. Boosting never changes the previous predictor and only corrects the next predictor by learning from mistakes. Since Boosting is greedy, it is recommended to set a stopping criterion such as model performance (early stopping) or several stages (e.g. depth of tree in tree-based learners) to prevent overfitting of training data. The first implementation of boosting was named AdaBoost (Adaptive Boosting).

Extreme Gradient Boosting:

Gradient boosting refers to a class of ensemble machine learning algorithms that can be used for classification or regression predictive modeling problems.

XGBRegressor:

```
from xgboost import XGBRegressor
model_xg = XGBRegressor()
```

R square value for xg boosting model train 0.876

R square value for xg boosting model test 0.847

From above you can see how XGBRegressor gave us a very good accuracy as compare to linear model. But we will not stop here as there is still a scope of improvement.

Hyperparameter:

- Hyperparameters are certain values or weights that determine the learning process of an algorithm.
- As stated earlier, XGBoost provides large range of hyperparameters. We can leverage the maximum power of XGBoost by tuning its hyperparameters.
- The most powerful ML algorithm like XGBoost is famous for picking up patterns and regularities in the data by automatically tuning thousands of learnable parameters.
- In tree-based models, like XGBoost the learnable parameters are the choice of decision variables at each node.
- XGBoost is a very powerful algorithm. So, it will have more design decisions and hence large hyperparameters. These are parameters specified by hand to the algo and fixed throughout a training phase.
- In tree-based models, hyperparameters include things like the maximum depth of the tree, the number of trees to grow, the number of variables to consider when building each tree, the minimum number of samples on a leaf and the fraction of observations used to build a tree.

r2 value for xgb_model model train 0.950

r2 value for xgb_model model test 0.890

Comparison and selection of model:

We first build the OLS model which gave us the R square value 0.41.Which is not good as model was underperforming. We observed that data is not normal.

So we chose to go for non linear algorithm, we got R square value 0.876 for train and 0.847 for XGBRegressor. As the model was giving good R Square value so we worked XGBoost hyper parameter.

After tuning we got R square value for train 0.95 and test 0.89, this gave is very good value. Performance of our model has changed drastically after tuning. This makes XGBoost best model for our dataset.

Assumptions:

- The food delivery company which operates in multiple cities. They have various fulfillment centers in these cities for dispatching meal orders to their customers.
- They provide various categories of food of different cuisines.
- Company was facing issues with managing employees to various locations and wastage of food on weekly basis.
- It can be prevented if they know rough idea of orders they will receive in coming weeks.

Conclusion:

A food delivery company has their centers in different locations in country. It provides various categories of food and cuisine like Indian,Thia, Italian,and Continental.Providing such services to customers upto their door step. And to make customers feels satisfied is a huge challenge in this world of competition. Attracting them with exiciting offers and promising them to with provide best services plays crucial role. But food lovers are very moody there might be sudden upsurge or downfall in number of orders, even climate and geographical condition have their impact on buisness.

Along with this there were more issues noticed by the company. Organisation were finding it challenging to manage their employees at right time in right location as it was very uncertian to say when their will be hike in orders and at which location. And there might be opposite condition that orders may get reduce which will lead to food wastage.

So here comes the role of machine learning to analyis data and predict the number of orders a company may recieve in coming weeks. From the avaiable information, We studied various buisness aspect such as the products they delivered, service they provided, their stratergies adapted to attract customers,etc.

We have also studied the technical aspects of the company by using visualisation techniques, exploratory data analysis, statistical test. Then we tried machine learning models to predict total number of orders. In our final model we have used extreme gradient boosting algorithm which gave us 95% R Square value in training data and 89% in test data. So now it easy task for the firm to handle thier employees at right location. And also preventing wastage of food

as they will be known to most accurate number of orders prior to a week. For any organisation preventing losses is equally important as making profits.

Limitations:

- This dataset does not have insights about days of order in week. So it's difficult to analysis whether orders were more in weekend or working days.
- It does not have timing of order. It would be helpful for us to know in which time people prefer to order food in a day.
- If dates were known we could have found out the impact of festival's, government holidays on food orders.

Reference: https://www.kaggle.com/ghoshsaptarshi/av-genpact-hack-dec2018

https://medium.com/analytics-vidhya/preparing-for-interview-on-machine-learning-3145caeea06b

https://medium.com/@gabrieltseng/gradient-boosting-and-xgboost-c306c1bcfaf5

https://statistics.laerd.com/spss-tutorials/kruskal-wallis-h-test-using-spss-statistics.php#:~:text=The%20Kruskal%2DWallis%20H%20test,continuous%20or%20ordinal%20dependent%20variable.