Capstone Project

Niranjan Dhomse

Food Demand Forecasting

Problem Statement

- Our client is a meal 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 us to help these centers with demand forecasting for upcoming weeks so that these centers will plan the stock of raw materials accordingly
- https://datahack.analyticsvidhya.com/contest/genpact-machine-learning-hackathon-1/

Importance

- 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

Data

- Analytics Vidhya runs hackathons for various ML problems. This data is sourced from their site.
- Data description: https://datahack.analyticsvidhya.com/contest/genpact-machine-learning-hackathon-1/#ProblemStatement
- The training data set has 456,548 rows, 9 columns

Techniques

- The base problem is a time series type problem with multi-series or multi-variate time series
- Used ***sktime*** package along with LinearRegression, XGBoost and RandForest with RandomizedSearchCV

Timeframe/Challenges

- Recommend also applying newer libraries and tools such as ***skforecast*** and Facebook's ***Prophet***. Particularly for the multi-series/multi-variate capabilities

EDA - Understanding the data

Datasets

1. **train** data (core training data including the target variable - num_orders) – 456k rows
2. **fulfilment center** info with features related to each center identified by center_id – 77 centers
3. **meal** info with features related to each meal identified by meal_id – 51 meal combos
4. **test** data – this *does not have the target variable*. – 32k rows
    a. This is for submission in this competition. Will be evaluated by the organizers and generate ranking for the submission

About the data

- We have 145 weeks of data for each center/meal combination - with a total of 77 x 51 = 3927 time series' 145 weeks each
- Goal is to predict next 10 weeks of orders - target variable: *num_orders*
- Data is fairly clean. No null values and no duplicate rows. Although from time series perspective, we do see many weeks missing data for a center/meal combo - we assume those are weeks where there were no orders for that combo
- We merge the reference data from center and meal datasets to create a dataset for EDA, feature engineering and modeling

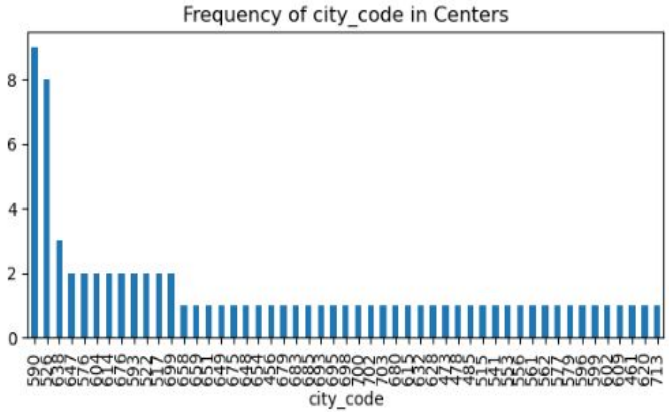
EDA - Key findings from the data

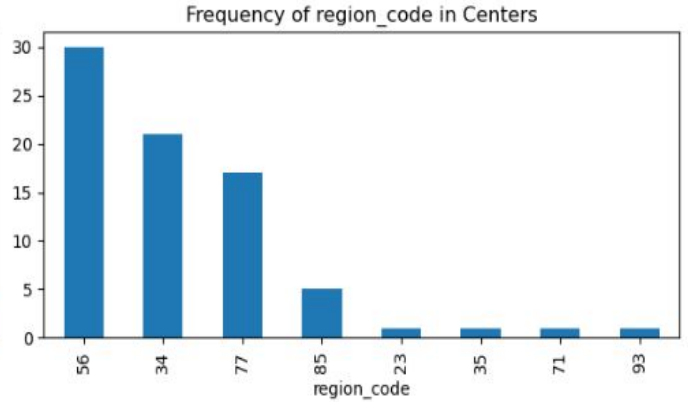
Key Findings

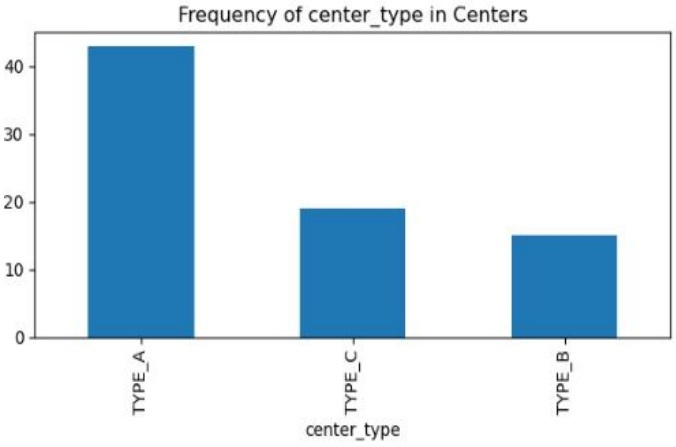
- Target variable *num_orders* distribution is very right skewed. Although the max is 24299, less than 0.75% of the instances have num_orders > 2000, the median value is only 136 and 75% values are 324 or less
- Beverages category orders comprises 28% of all occurrences and is at least 4x the other categories
- Even though 57% of the order occurrences are from TYPE_A centers, TYPE_B generates more orders on average
- Lower prices clearly attract more orders
- Thai cuisine has about 25% more meal options than other cuisines. However, Italian cuisine seems to be more popular and generates the highest number of orders on average
- Emailer and Homepage promos result in 3x orders on average
- Larger Op Areas generate larger number of orders on average

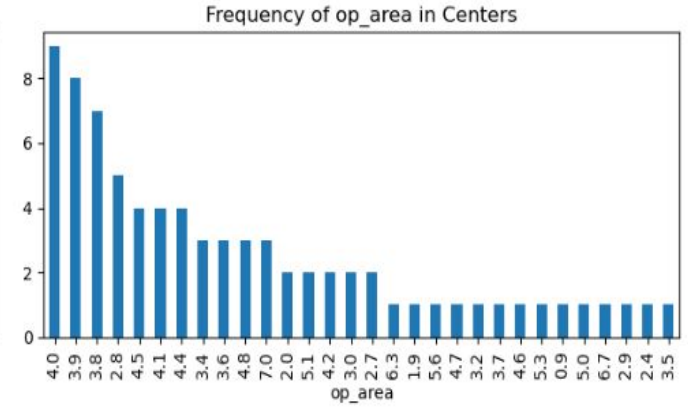
EDA (center info)

- Majority of the centers are in regions 56, 34, 77
- 56% of the centers are TYPE_A
- 30% of the centers have op_areas between 3.8 and 4.0
- City codes 590 and 526 have highest number of centers, 9 and 8 resp



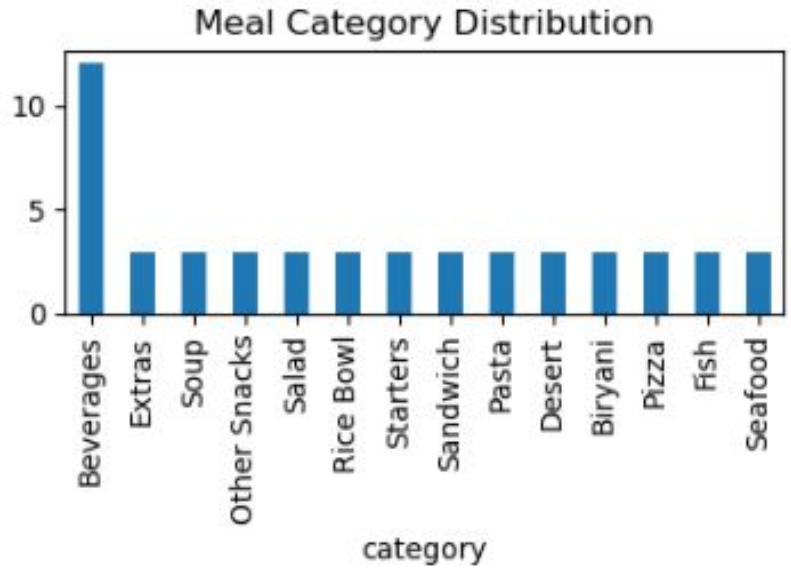


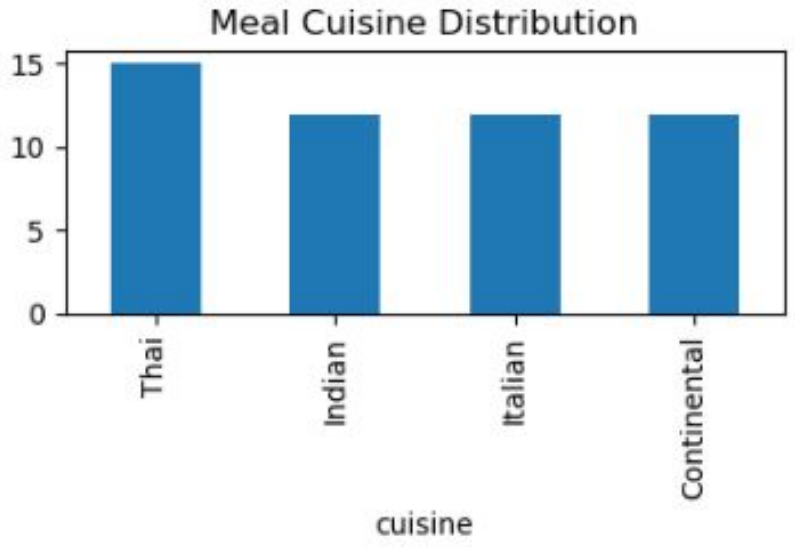




EDA (meal info)

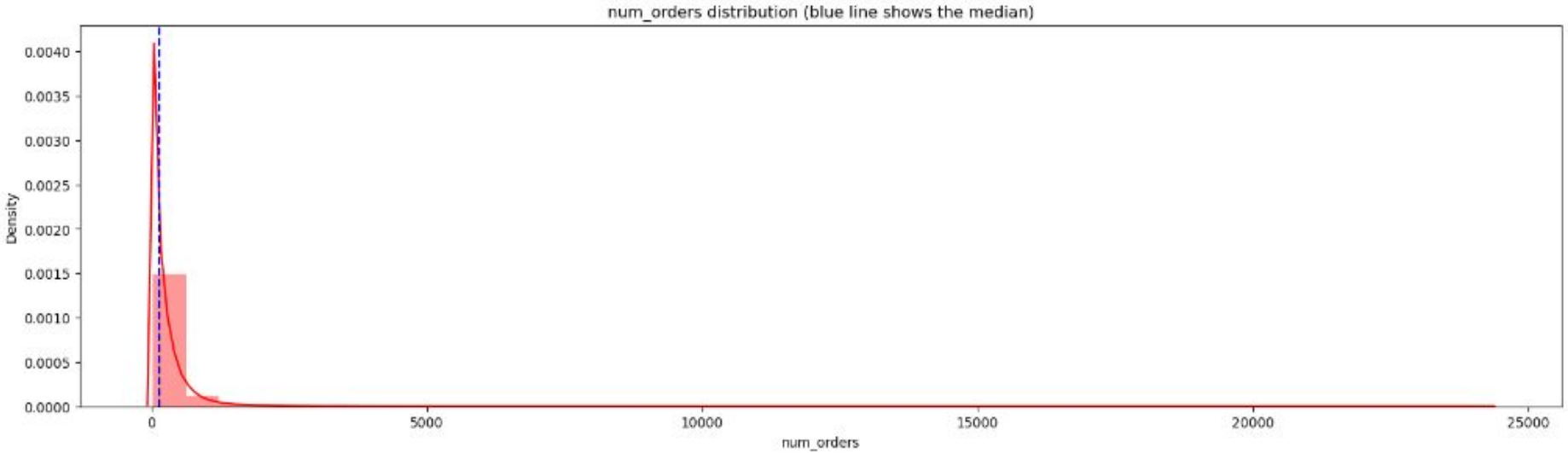
- Beverages category is the most prominent type of meals offered (24% of all meal categories)
- Cuisine type is almost evenly distributed in available meals, with Thai cuisine having about 25% more options





EDA (Target Variable)

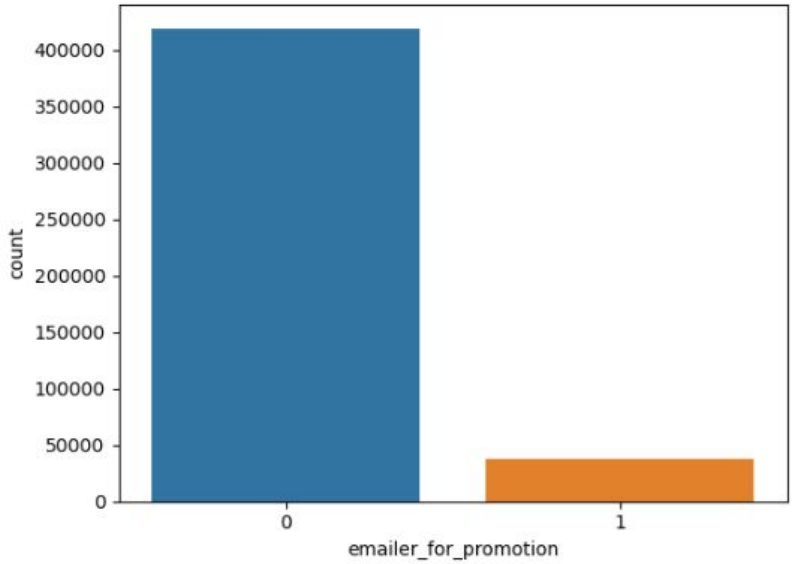
- Target variable *num_orders* distribution is very right skewed
- We see max value > 20,000
- However, less than 0.75% of the instances have *num_orders* > 2000
- The median value is only 136



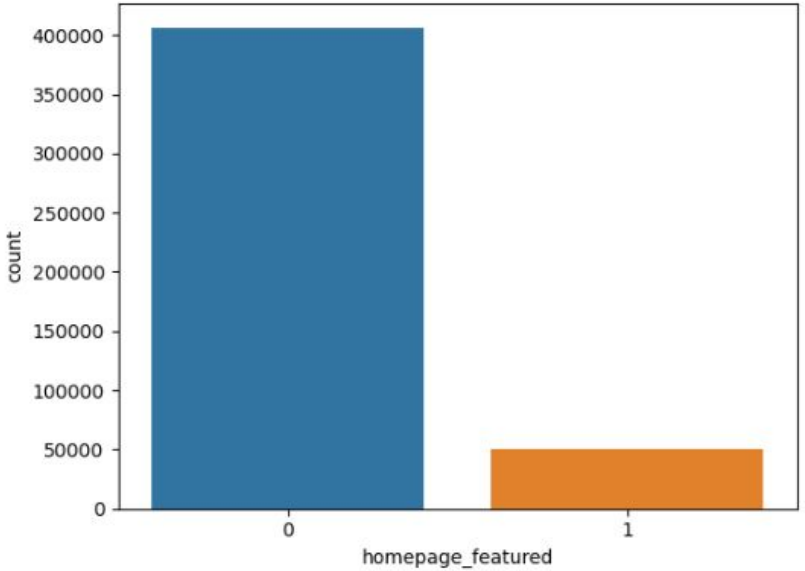
EDA (Train data: emailer and homepage promos)

- Emailer for promotion was sent about 8.1% of the time
- Featuring on Homepage happened about 11% of the time

emailer_for_promotion	emailer_for_promotion	Percentage
0	419498	91.884753
1	37050	8.115247

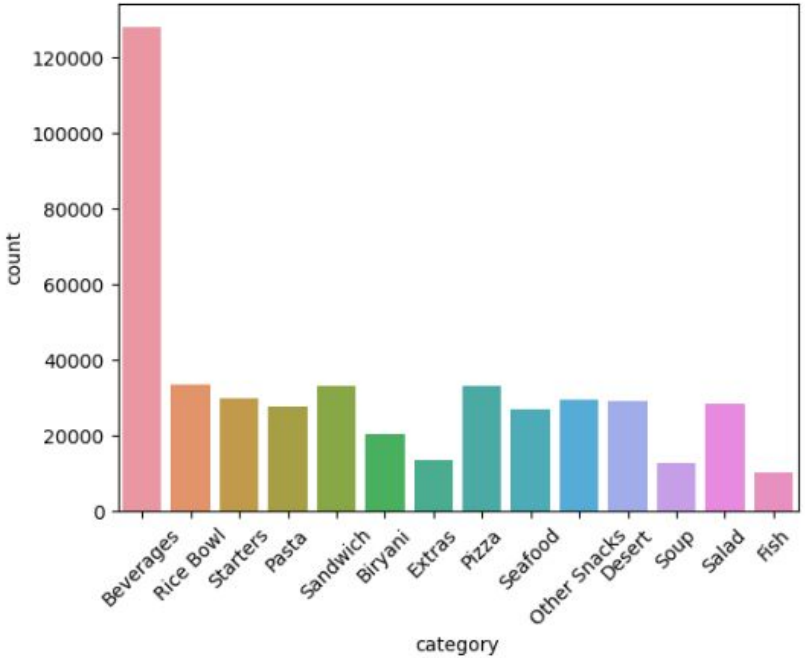


homepage_featured	homepage_featured	Percentage
0	406693	89.080009
1	49855	10.919991

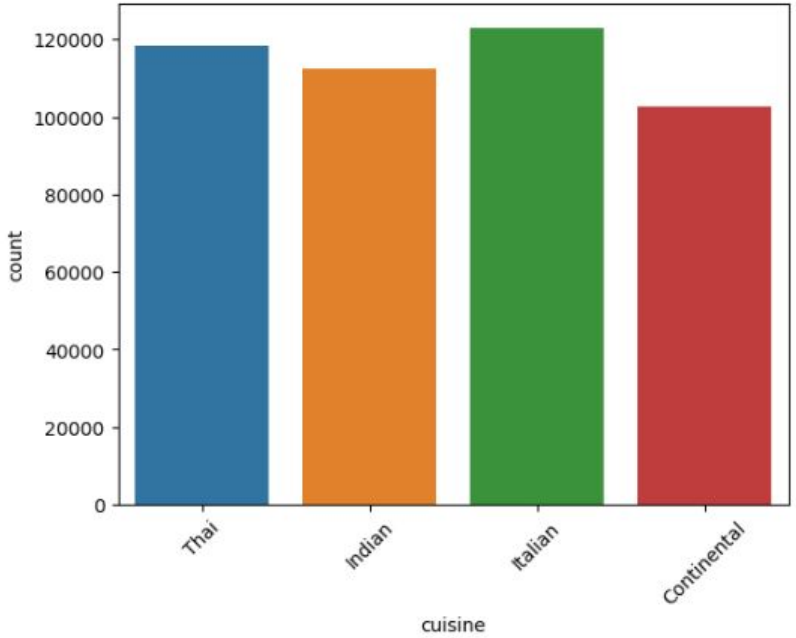


EDA (Train data: meal category and cuisine)

- *Beverages* category orders comprises 28% of all occurrences
- *Fish*, *Soup* and *Extras* are among the least frequently ordered between 2% and 3% of the times
- Rest of the categories are fairly evenly ordered about 5-7% of the times



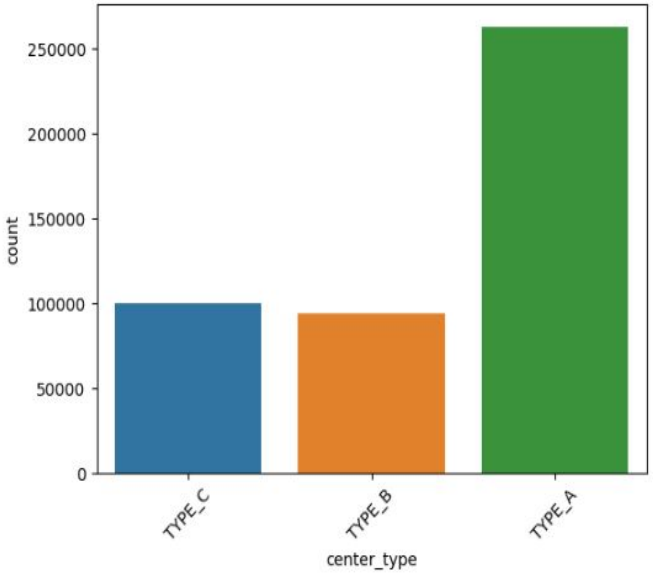
- Orders are fairly evenly distributed across *cuisine* category



EDA (center_type, op_area)

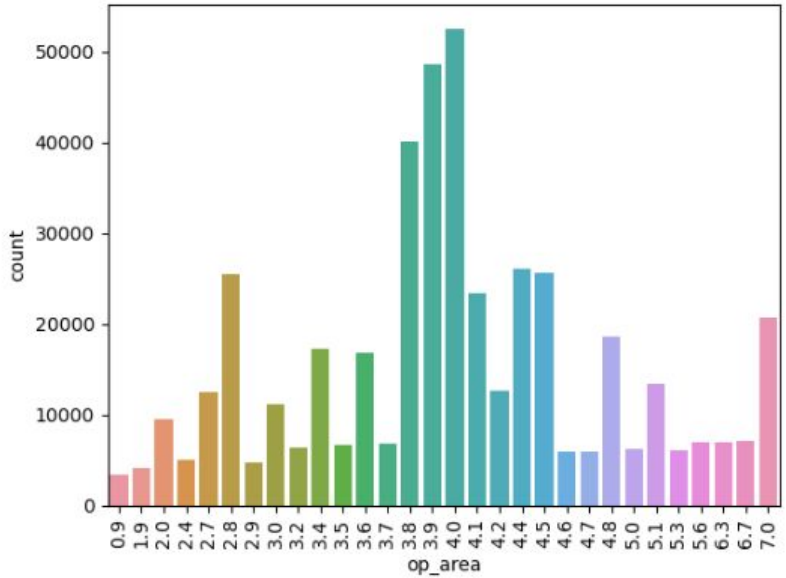
Order occurrences by center_type

- 57% of the orders come from *TYPE_A* centers - which aligns with the ratio of the TYPE_A centers in the company



Order occurrences by op_area

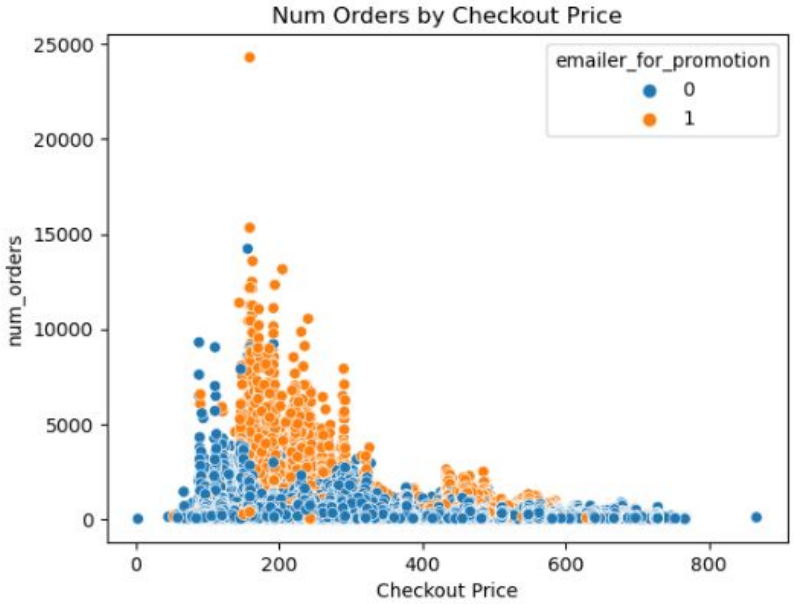
- 30% of the orders come from centers with op_area between 3.8 and 4.0 - which aligns with the ratio of the op_area of the centers in the company

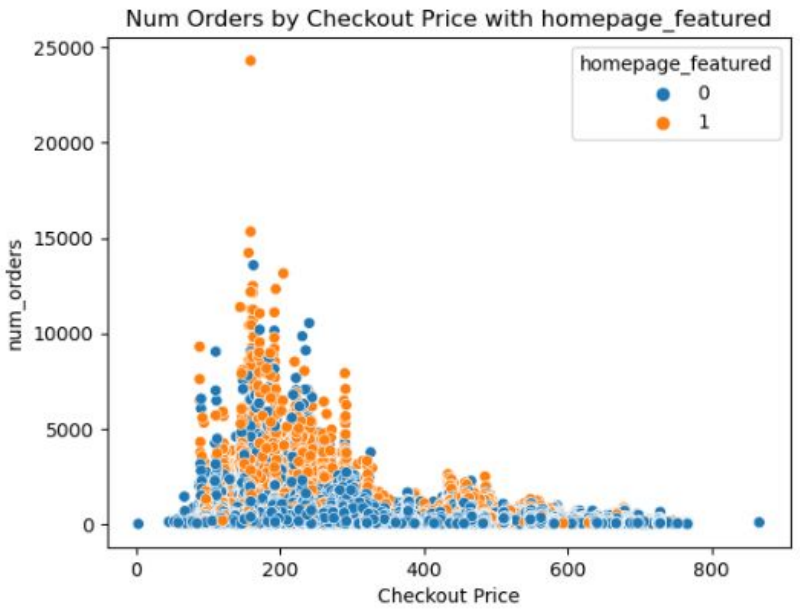


EDA (impact of checkout price and promos)

Correlation

- Lower prices clearly attract more orders
- And homepage_featured, emailer_for_promotion also have a big impact in higher number of orders on average

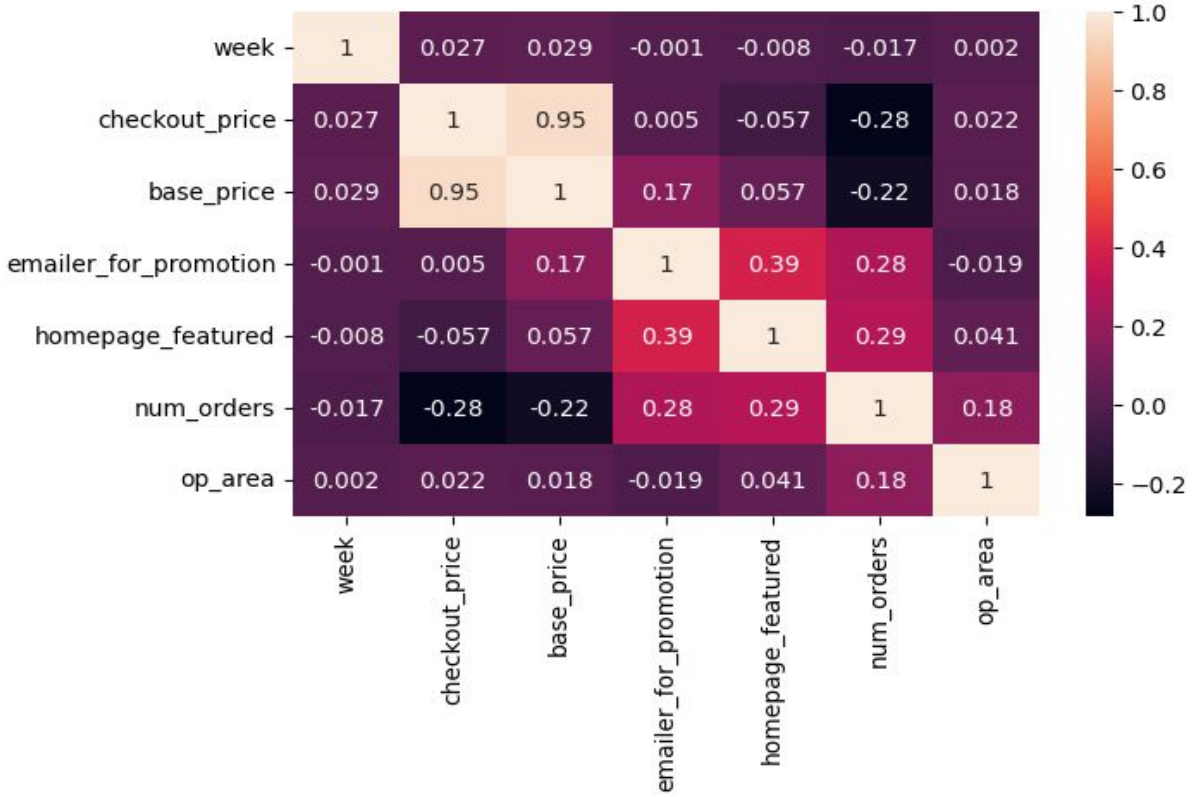




EDA (correlation matrix)

Independent variable correlation

- *checkout_price* and *base_price* are highly positively correlated
- *emailer_for_promotion* and *homepage_featured* promos occur together fairly often
- *base_price* seems to be higher when *emailer_for_promotion* is offered. Perhaps in many cases, the *base_price* is already high whe promotions are being offered?



EDA (correlation matrix - target variable)

Correlation to target variable

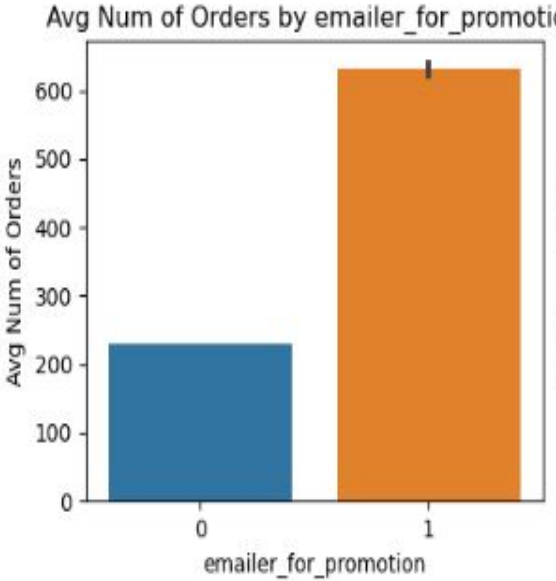
- We see some independent variables correlation to the target variable num_orders
- Higher number of orders when emailer_for_promotion, homepage_featured offered
- More orders come from centers with higher op_area
- Less orders with higher checkout_price and base_price

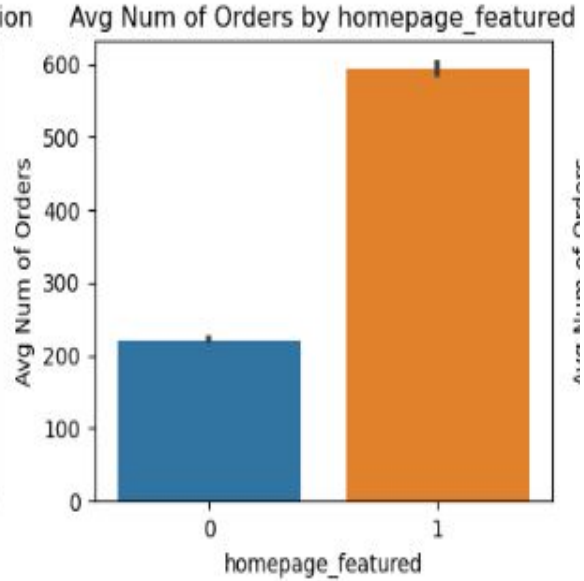
	num_orders
checkout_price	-0.282
base_price	-0.222
center_id	-0.053
week	-0.017
meal_id	0.011
region_code	0.030
city_code	0.042
op_area	0.177
emailer_for_promotion	0.277
homepage_featured	0.294
num_orders	1.000

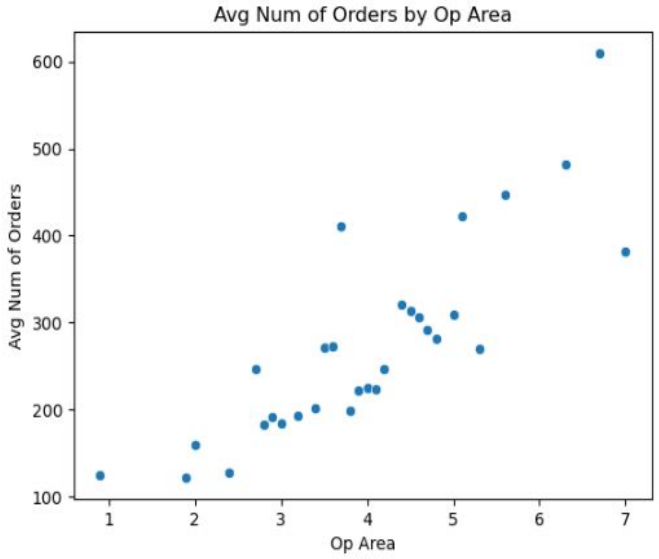
EDA (Impact of Op Area, emailer and homepage promos)

Correlation

- We see a very significant impact of both emailer promotions and homepage featuring on *num_orders*
- Average number of orders is roughly linearly proportional to the *op_area*



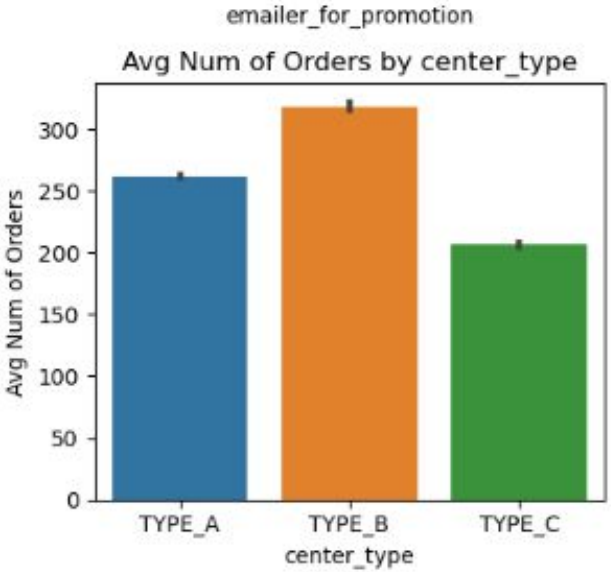




EDA (correlation - average number of orders)

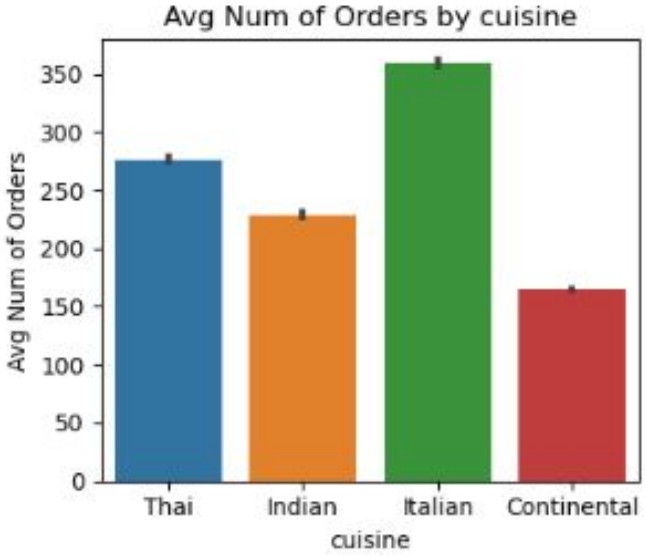
Correlation - center_type

- Even though TYPE_A comprises over 57% of the centers, TYPE_B generates more orders on average

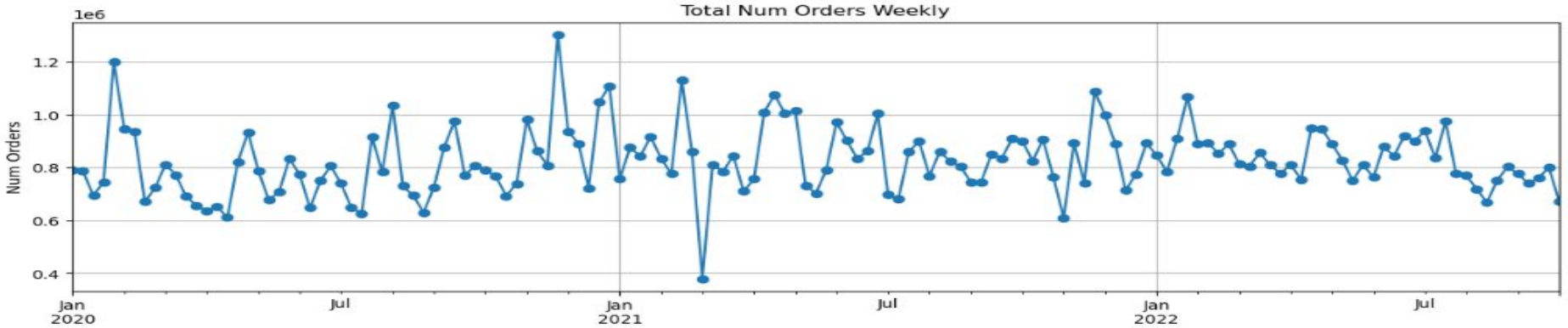


Correlation - cuisine

- The weekly order occurrences are evenly distributed by cuisine. However, Italian cuisine generates the highest number of orders on average

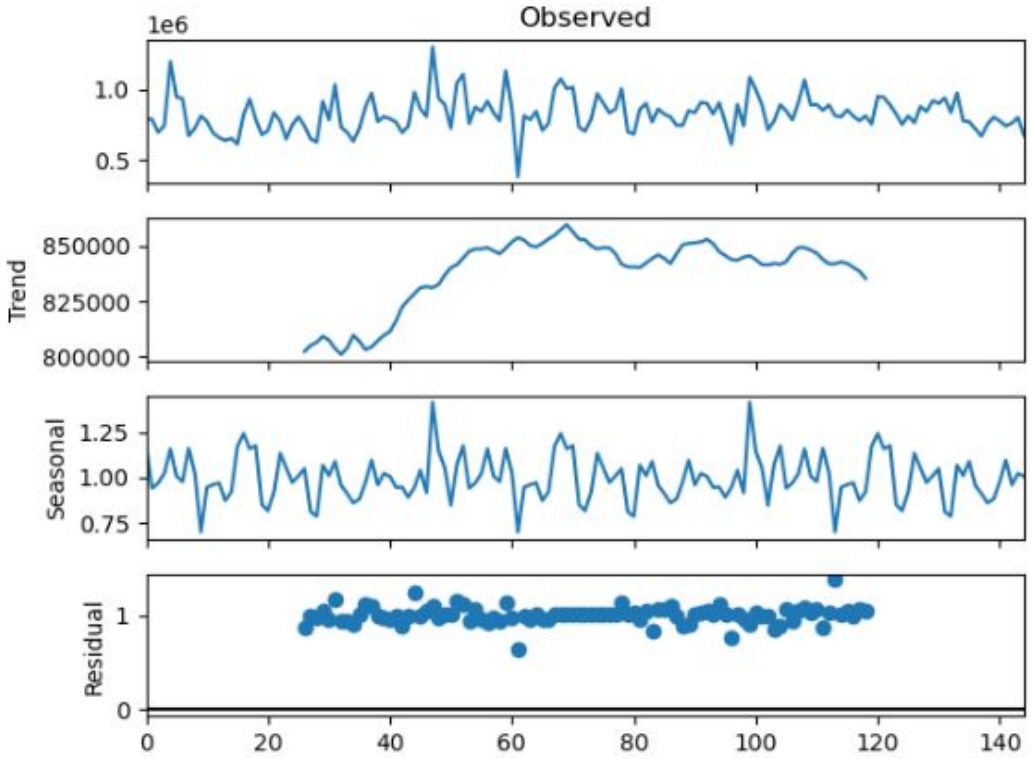


Time Series: *Total number of orders and revenue by week*



Decomposing the Time Series

- We can see some repeating seasonal pattern
- And increasing trend during the 2nd quartile of the 145 week period, but then it plateaued and slightly declined



Data Cleaning / Pre-processing

Observations and steps taken

1. The available data did not have any null values or duplicate rows
2. From a time series perspective, some rows were missing, presumably when there were no orders for any given week for a given center+meal combination
3. Few columns which were categorical had numeric values, so these had to be converted to category type so that some of the pandas and other functions would handle them accordingly
4. The dataset provided had train data and two separate reference data files for center and meal info. This reference data was merged with the train data to create the working dataframe
5. The time series provided was weekly data by center and meal id. Created date column from the week starting with Jan 1, 2020

Modeling and Forecasting

Observations and steps taken

1. The task was to forecast *num_orders* for each of the sub-series identified by *center_id* + *meal_id* for the next 10 weeks
2. Decided to break it down into two levels:
    a. Overall Aggregate level series - with sum of all orders by week
    b. One of the center_id/meal_id level time series - picked one with the highest number of orders total orders
    c. Forecasting for the whole set would then involve running through building the model and forecasting for each combination of center and meal ids. (This step was not undertaken because of the huge processing resources and time needed)
3. Decided to use MAPE as the evaluation metric, which seems reasonable for the given problem. However, did implement calculating and comparing RMSE and RMSLE metrics.
4. Steps and models compared:
    a. Created baseline using DummyRegressor
    b. LinearRegression
    c. XGBoostRegressor
    d. XGBoost with RandomizedSearchCV
    e. RandomForest with RandomizedSearchCV
5. Leveraged the *sktime* framework with the above forecasters for this task

Models and Forecasting Results

Overall Aggregate Level:

	Model	MAPE	RMSE	RMSLE	Exec Time
0	DummyRegressor	0.1435	113217.7234	0.1459	0.0000
1	LinearRegression	0.1167	130233.1366	0.1586	1.0000
2	XGBRegressor	0.1895	155043.3794	0.1922	1.0000
3	XGBRegressor with RandomizedSearchCV	0.1471	121734.9449	0.1548	78.2462
4	RandomForestRegressor with RandomizedSearchCV	0.1388	122989.3392	0.1565	20.8403

- For the Overall Aggregate series, the LinearRegression model came up with the best MAPE score(0.1167), followed by RandomForest (0.1388)

Models and Forecasting Results

For center_id=13, meal_id=1885:

	Model	MAPE	RMSE	RMSLE	Exec Time
0	DummyRegressor	0.1237	532.3073	0.1877	0.0000
1	LinearRegression	0.3750	1157.7575	0.4318	15.6234
2	XGBRegressor with RandomizedSearchCV	0.1013	451.7821	0.1573	35.2223
3	RandomForestRegressor with RandomizedSearchCV	0.0939	475.8379	0.1657	11.8819

- For the center_id=13, meal_id=1885 series, the tuned RandomForest model came up with the best MAPE score (0.0930), followed by tuned XGBoost model (0.1013)

Modeling Findings

Modeling Findings

- For the Overall Aggregate series, the LinearRegression model came up with the best MAPE score, followed by RandomForest
- For the center_id=13, meal_id=1885 series, the LinearRegression model came up with the best MAPE score, followed by RandomForest
- Decided to use MAPE as the evaluation metric, which seems reasonable for the given problem. However, did implement calculating and comparing RMSE and RMSLE metrics.
- Steps and models compared:
  - Created baseline using DummyRegressor
  - LinearRegression
  - XGBoostRegressor
  - XGBoost with RandomizedSearchCV
  - RandomForest with RandomizedSearchCV
- Leveraged the *sktime* framework with the above forecasters for this task

Modeling Conclusions and Next Steps

Conclusions & Next Steps

- Recommend looking at ***skforecast*** and ***sktime*** frameworks and perhaps the ***Prophet*** model to figure out how the models could be trained on a hierarchical multi-series dataset and forecast using all the available features in the dataset
- From business perspective, following recommendations could help the business grow its orders
  - Leveraging the emailer and homepage promos
  - Further including looking at other modes/channels of advertising
  - Review and publicize any discounts and promotions that will reduce the prices for the consumers
  - Review what is making the TYPE_B centers generate higher orders and implement for other center types
  - Review what is making the Italian cuisine offerings generate higher orders and see if some elements could be enhanced for other cuisines. Also, review the possibility of enhancing the Italian menu further