Restaurant Revenue Prediction

By Jay Li
Irene Izere

Background

With over 1,200 quick service restaurants across the globe, TFI is the company behind some of the world's most well-known brands: Burger King, Sbarro, Popeyes, Usta Donerci, and Arby’s. They employ over 20,000 people in Europe and Asia and make significant daily investments in developing new restaurant sites.

Problem Statement

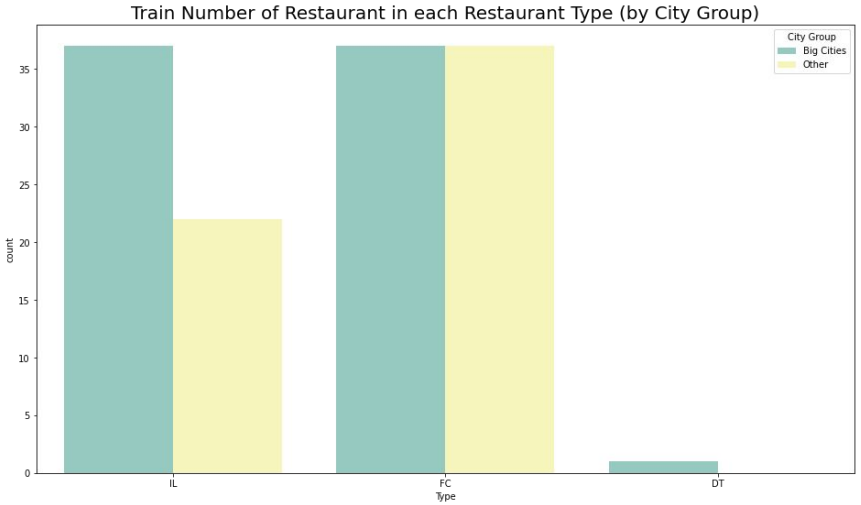
- What should the TFI company take into consideration when investing on a new restaurant for highly profitability?

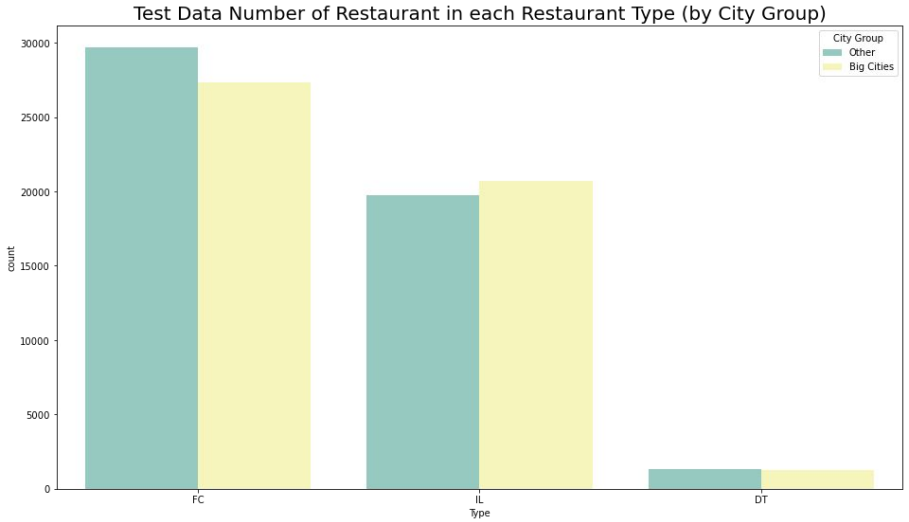
DATA COLLECTION

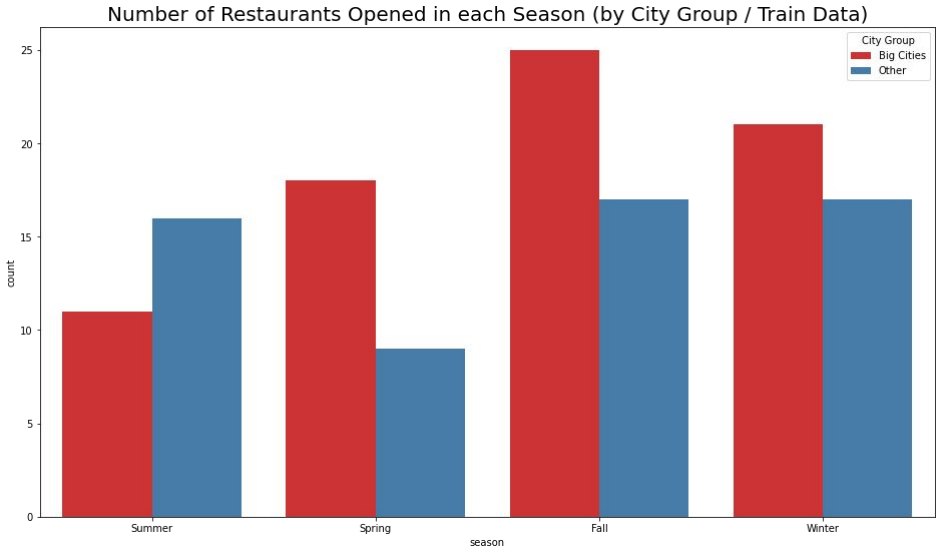
- The dataset is from kaggle competition.It has a train data set and a test dataset. The features are : id,city,city group,type , P1to P37 and Revenue.

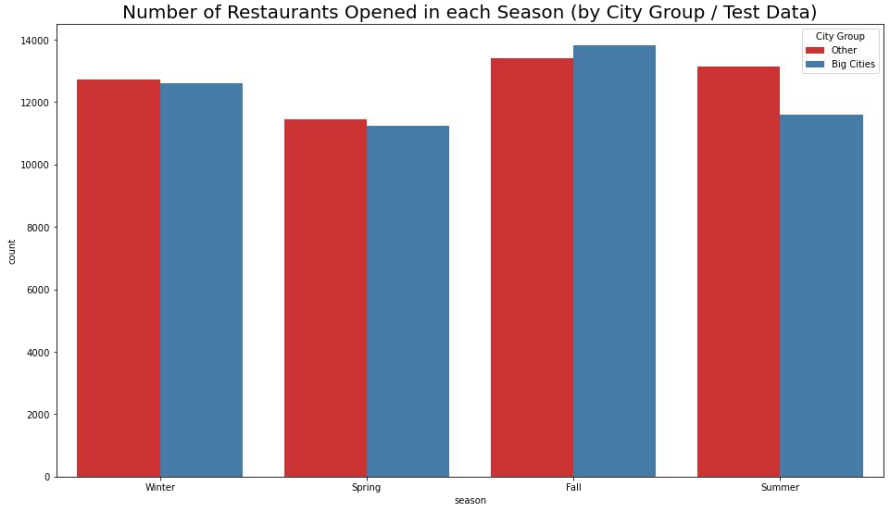
Data Cleaning and EDA

- Add new column from Open Date , extracting the month to make a Season column and year column to understand
- Drop the City column because of the high difference of unique value in the test dataset.


Train Number of Restaurant in each Restaurant Type (by City Group)
City Group
Big Cities
Other
count
Type
IL
FC
DT
0
5
10
15
20
25
30
35


Test Data Number of Restaurant in each Restaurant Type (by City Group)
City Group
Other
Big Cities
count
Type
FC
IL
DT
0
5000
10000
15000
20000
25000
30000


Number of Restaurants Opened in each Season (by City Group / Train Data)
City Group
Big Cities
Other
count
season
Summer
Spring
Fall
Winter
0
5
10
15
20
25


Number of Restaurants Opened in each Season (by City Group / Test Data)
City Group
Other
Big Cities
count
season
Winter
Spring
Fall
Summer
0
2000
4000
6000
8000
10000
12000
14000

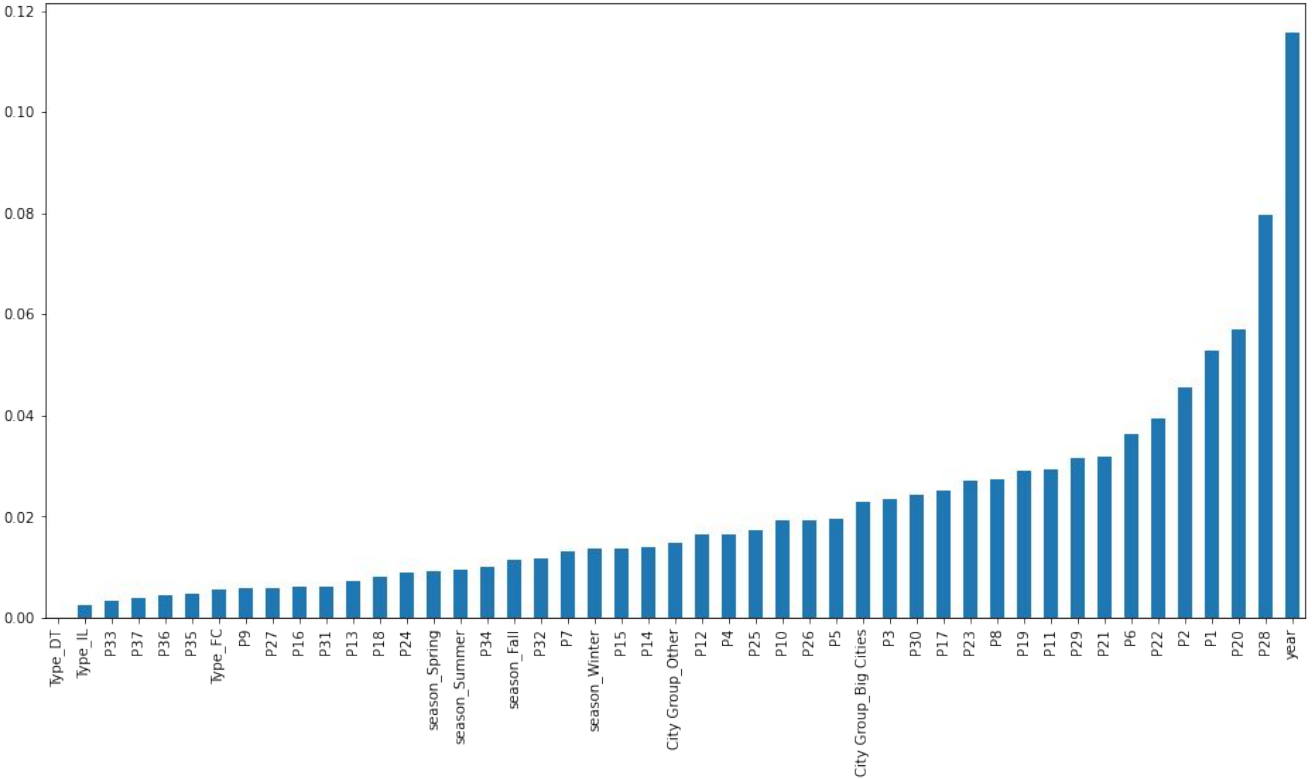
Modeling and Evaluate

We built and evaluated 6 different models. With the simple Linear Regression model as the baseline model, We are comparing all other models to the baseline for evaluations.

- Linear Regression
- Lasso Regression
- Ridge Regression
- Random Forest
- KNN
- XGboost

	Method	Linear Regression	Ridge Regression	Lasso Regression	Random Forest	KNN	XGBoost
0	Mean Absolute Error	0.488084	0.377203	0.344108	0.337494	0.317287	0.342634
1	Mean Squared Error	0.386681	0.217550	0.201000	0.170350	0.162181	0.183407
2	RMSE	0.621837	0.466422	0.448330	0.412734	0.402717	0.428260
3	R^2	-0.925315	-0.083196	-0.000794	0.151816	0.192488	0.086805

The KNN has the smallest RMSE ,the highest R2 and a small relative score variance between training and testing and it got the best scores on the kaggle competition.



This graph shows the features importance using the random forest model

Reflection

Looking back on this project , we didn't get a good score on the models because the dataset we have needed more cleaning and feature engineering.Looking at the values of P1-p37 features , there are some values that are equal to zero that may represent missing values rather than 0 as a value. One way to work on that is to use **knn imputer** to resolve some of those issues, we could also find a way to keep the city column because location is an important feature in predicting the profitability of a restaurant.