1. Introduction

Background

In the last few years there was a huge jump in numbers of café in Saudi Arabia due to increased demand for it. And there will be a huge competition on this sector, and to overcome this from affecting your business you should make a better judgment of the market, and using data is the most powerful technique for making a better chances for success.

Problem

Due to increased competition on café businesses there will be a huge demand for data associated with it and analysis of it which will make a better chance for business success. There are multiple criteria that will make you a better place for customers such as location, price level, etc. and in this project we will determine which factor has the most impact for customer satisfaction and it will be gathered from customers reviews on Google Maps , and determine if it associated with any of the criteria.

Interest

Of course, there will be an interest from business owners to gather data and analysis of café shops in Saudi Arabia.

2.Data acquisition and cleaning

Data sources

Most of our data is taken from Google Places API which will allow us to access more data than foursquare because of a huge usage of Google Maps in Saudi Arabia

Data cleaning

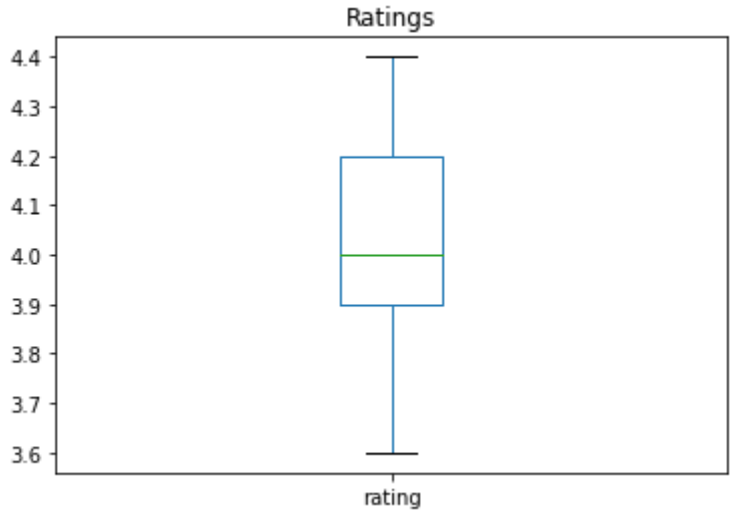
Data is collected from Google Places API while it makes it data separated 20 per request there was multiple request to increase our subject numbers and it was joined together in a single DataFrame to make it easily accessible, it was cleaned from unnecessary objects in our project. Any values of NaN is removed to make better estimations. Any column name with an ambiguity changed for simplification.

Feature selection

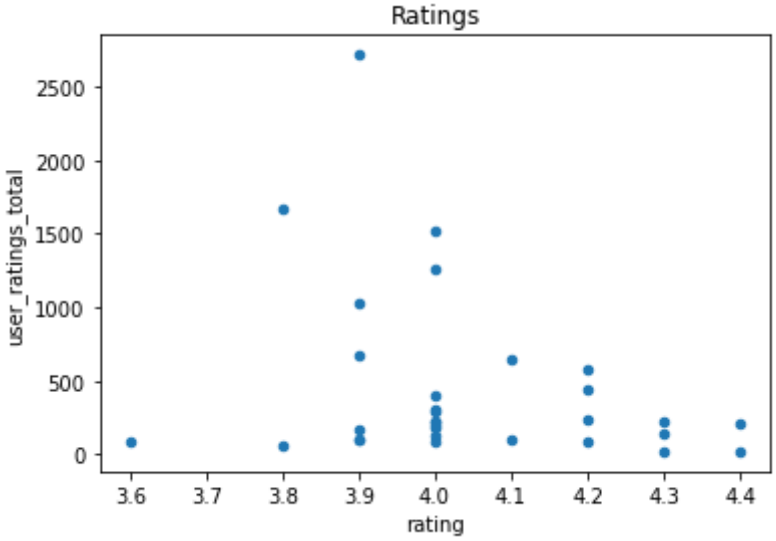
After data cleaning there was 41 samples and 13 features in the data.

3.Exploratory data analysis

Rating of the customer is a very important to determine if your business is going to continue or fail. As we can see in the next box graph our data collection gave us that the most of our rating is between 3.9 and 4.2 and the median is almost 4.0.



Also, there was a normal distribution of our data with the ratings and total user rating and as we can see in scatter graph below.



Relationship between customers reviews with location

	index	price_level	rating	user_ratings_total	geometry.location.lat	geometry.location.lng	geometry.viewport.northeast.lat	g
index	1.000000	0.211042	0.143825	-0.323397	0.415160	-0.136344	0.414904	
price_level	0.211042	1.000000	-0.361198	0.133406	0.038548	-0.098063	0.038387	
rating	0.143825	-0.361198	1.000000	-0.272730	0.159580	-0.036244	0.159678	
user_ratings_total	-0.323397	0.133406	-0.272730	1.000000	-0.042166	0.230122	-0.041824	
geometry.location.lat	0.415160	0.038548	0.159580	-0.042166	1.000000	0.043768	0.999998	
geometry.location.lng	-0.136344	-0.098063	-0.036244	0.230122	0.043768	1.000000	0.043689	
geometry.viewport.northeast.lat	0.414904	0.038387	0.159678	-0.041824	0.999998	0.043689	1.000000	
geometry.viewport.northeast.lng	-0.137105	-0.098203	-0.036011	0.230325	0.043192	0.999998	0.043113	
geometry.viewport.southwest.lat	0.414904	0.038387	0.159678	-0.041824	0.999998	0.043689	1.000000	
geometry.viewport.southwest.lng	-0.137105	-0.098203	-0.036011	0.230325	0.043192	0.999998	0.043113	
Cluster Labels	-0.023498	-0.047946	0.060847	0.305654	0.113719	0.820393	0.113457	

As can wee see there almost no correlation between location and customer reviews due to correlation factor was almost to 0.06 between Cluster Labels which was out location clusters and ratings which means that there is no correlation.

Relationship between customers reviews with price level.

	index	price_level	rating	user_ratings_total	geometry.location.lat	geometry.location.lng	geometry.viewport.northeast.lat
index	1.000000	0.274589	0.139980	-0.308604	0.450442	-0.357099	0.450481
price_level	0.274589	1.000000	-0.496614	0.275428	0.347273	-0.287945	0.346541
rating	0.139980	-0.496614	1.000000	-0.469295	-0.353727	0.516701	-0.352112
user_ratings_total	-0.308604	0.275428	-0.469295	1.000000	0.183608	-0.339262	0.183390
geometry.location.lat	0.450442	0.347273	-0.353727	0.183608	1.000000	-0.829718	0.999995
geometry.location.lng	-0.357099	-0.287945	0.516701	-0.339262	-0.829718	1.000000	-0.828853
geometry.viewport.northeast.lat	0.450481	0.346541	-0.352112	0.183390	0.999995	-0.828853	1.000000
geometry.viewport.northeast.lng	-0.357178	-0.287690	0.515876	-0.338975	-0.830402	0.999994	-0.829544
geometry.viewport.southwest.lat	0.450481	0.346541	-0.352112	0.183390	0.999995	-0.828853	1.000000
geometry.viewport.southwest.lng	-0.357178	-0.287690	0.515876	-0.338975	-0.830402	0.999994	-0.829544
Cluster Labels	NaN	NaN	NaN	NaN	NaN	NaN	NaN

As can we see in our data from cluster red there is a moderate correlation between price level and rating of the customers which was a inversely correlation which means that the customer prefer café shops that gave them the best price for there coffee which give us our first hint for our business.

Solutions

For better chances for a customer pleasing from our data there was a moderate inverse correlation between price level and customer ratings which means that customers are pleased with low prices of coffee , which gave business owners to give the customer what he/she needs.

Conclusion

Data was gathered and analyzed and gave us a small hint to the price and customer pleasing , but it was not enough data to determine all the factors to gave success of the business, I wish in the future data become more available and gathered more often because it will leads our future to a better life.