Fitness Studio Preference Analysis

Kushal Shah
ks5574@nyu.edu

Rupesh Bagwe
rpb337@nyu.edu

Vinayak Natarajan
vn686@nyu.edu

Abstract:—Fitness Studios play an important role in the lifestyle of every individual. These Fitness Studios can be Gyms, Yoga centres, Martial Arts centres, Dance centres, clubs for swimming and so on. It is observed that the choice of the people for choosing the fitness studios for their respective workouts are generally influenced by the quality of some specific fitness studio in their vicinity. So proper business knowledge is required to implement any specific decision regarding marketing or even setting up a new business. This Analysis work is aimed at securing investment capital from Business developers and planning the marketing strategies accordingly.

1. Introduction

The Fitness Studios in any locality can sometimes play a crucial role in the fitness choices of the people living in that locality. As people prefer to workout on a daily basis to maintain a healthy lifestyle, the Fitness Studios need to be close to their homes and if not people prefer some different kind of workout mode rather than travelling to some distant fitness studio. The main theme of this study is to analyse the preferences of the people on what kind of Fitness Studio do they use and thus secure private business setups in that locality based on the insights received from this analysis. This can also help other local businesses to give targeted advertisements and change their business plans accordingly in that area. So a detailed analysis on people's choices for Fitness Studios and the kind of ratings received by the existing Fitness Studios is the primary goal of this study.

2. Problem Statement & Formulation

2.1 Problem Statement

The formal problem statement being tackled is, Leveraging Fitness Studio Preferences patterns of people to attract private investments in a specific locality and for existing businesses to adopt newer plans based on the insights of this analysis.

2.2 Formulation

This process can be performed by aid of external data or corporate investors who can choose to "Promote Certain Regions". An example if a locality in New York has great reviews for a gym and many people are using that gym then secondary businesses like gym supplies can be promoted in this locality. And if this region has bad reviews for gym, then newer business opportunities for setting up gym can be promoted in this region.

3. Datasets and Tools

3.1 Datasets

We are primarily looking at the Yelp data but have possible plans for other datasets. Following is a complete list:-

- Yelp business data
- Yelp review data
- Yelp User data

3.2 Tools

Major tools that have been used are:

- PySpark

- SparkML
- ElasticSearch
- Kibana

4. Architecture & Proposed Approach

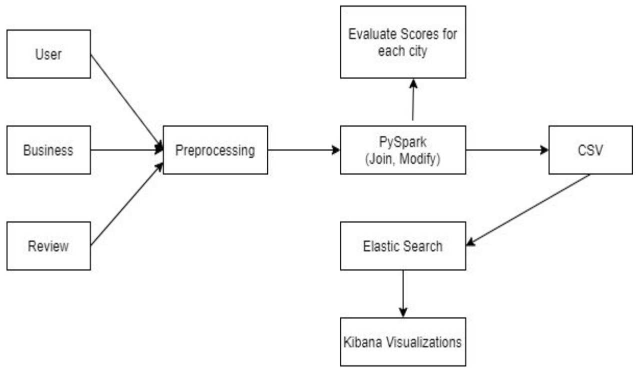

Figure 1: Architecture

This work leverages Pyspark and SparkML to guide the data flow to calculate scores for each Fitness Studio in each locality. ElasticSearch and Kibana are used for visualization of the Analysis. Figure 1 shows the layout of the architecture being developed.

4.1 Data Cleaning

The data we get is a dump from Yelp Business data. Figure 2 shows the raw data. We have around 10GB of data which is unordered, un-grouped, and full of issues. Figure 3 shows the data post cleaning. The Figure 4 shows the exact flow of the score generation model in our project. For this microanalysis we performed the following operations:

- Broke the data apart into separate lines
- Joined the datasets. (a)Yelp Business Dataset and (b)Yelp Review dataset
- Filtered based on the Fitness Studio types
- Filtered according to the positive and negative reviews.
- Used RegexTokenizer, StopWordRemover and Word count functions to find the occurences of Fitness Studio names.
- Took top 1000 words and found the points/score for each word.
- For required city and fitness studio type, these steps are repeated to find score of that studio.

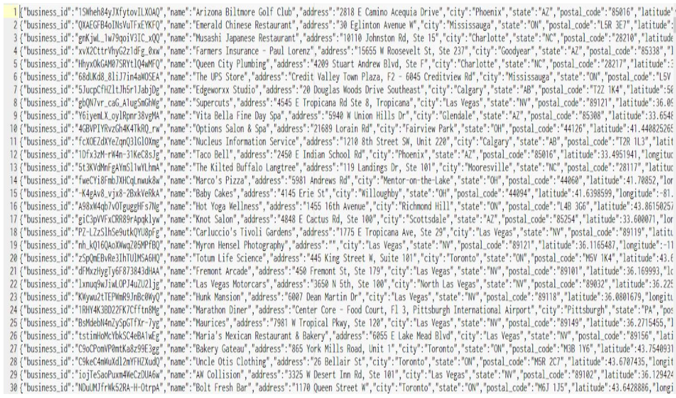
Figure 2: Raw Data

business_id	name	city	state	categories	stars	text
zzRgQ-Y1PKRMHVzXS...	Phoenix BJJ & MMA...	Goodyear	AZ	Fitness & Instruc...	4.0	My hubby knows on...
zzRgQ-Y1PKRMHVzXS...	Phoenix BJJ & MMA...	Goodyear	AZ	Fitness & Instruc...	5.0	Excellent school ...
zzRgQ-Y1PKRMHVzXS...	Phoenix BJJ & MMA...	Goodyear	AZ	Fitness & Instruc...	5.0	Great place to le...
zzRgQ-Y1PKRMHVzXS...	Phoenix BJJ & MMA...	Goodyear	AZ	Fitness & Instruc...	5.0	This is easily th...
zzRgQ-Y1PKRMHVzXS...	Phoenix BJJ & MMA...	Goodyear	AZ	Fitness & Instruc...	5.0	I have been train...
zzRgQ-Y1PKRMHVzXS...	Phoenix BJJ & MMA...	Goodyear	AZ	Fitness & Instruc...	1.0	Be sure you don't...
zymuRbVoLoJSSNdBD...	Hawk's Gym	Las Vegas	NV	Fitness & Instruc...	4.0	It was really cle...
zymuRbVoLoJSSNdBD...	Hawk's Gym	Las Vegas	NV	Fitness & Instruc...	5.0	Was there last ni...
zymuRbVoLoJSSNdBD...	Hawk's Gym	Las Vegas	NV	Fitness & Instruc...	4.0	This is a grittie...
zymuRbVoLoJSSNdBD...	Hawk's Gym	Las Vegas	NV	Fitness & Instruc...	5.0	Great place to re...
zymuRbVoLoJSSNdBD...	Hawk's Gym	Las Vegas	NV	Fitness & Instruc...	4.0	hawks gym is a "M...
zymuRbVoLoJSSNdBD...	Hawk's Gym	Las Vegas	NV	Fitness & Instruc...	3.0	located in a comm...
zymuRbVoLoJSSNdBD...	Hawk's Gym	Las Vegas	NV	Fitness & Instruc...	1.0	Before going to H...
zymuRbVoLoJSSNdBD...	Hawk's Gym	Las Vegas	NV	Fitness & Instruc...	5.0	Do you like eerie...
zymuRbVoLoJSSNdBD...	Hawk's Gym	Las Vegas	NV	Fitness & Instruc...	5.0	I went at an odd ...
zymuRbVoLoJSSNdBD...	Hawk's Gym	Las Vegas	NV	Fitness & Instruc...	1.0	They allow smokin...
zymuRbVoLoJSSNdBD...	Hawk's Gym	Las Vegas	NV	Fitness & Instruc...	1.0	I tried a lot of ...
zymuRbVoLoJSSNdBD...	Hawk's Gym	Las Vegas	NV	Fitness & Instruc...	4.0	I had a good time...
zymuRbVoLoJSSNdBD...	Hawk's Gym	Las Vegas	NV	Fitness & Instruc...	5.0	Since the Blue Mo...
zymuRbVoLoJSSNdBD...	Hawk's Gym	Las Vegas	NV	Fitness & Instruc...	5.0	I didn't see much...
zymuRbVoLoJSSNdBD...	Hawk's Gym	Las Vegas	NV	Fitness & Instruc...	4.0	This was my first...
zymuRbVoLoJSSNdBD...	Hawk's Gym	Las Vegas	NV	Fitness & Instruc...	5.0	Was there again l...
zymuRbVoLoJSSNdBD...	Hawk's Gym	Las Vegas	NV	Fitness & Instruc...	2.0	A private men's c...
zyT1smf2oJwimXKAS...	F45 Henderson	Henderson	NV	Interval Training...	5.0	It's like having ...
zyT1smf2oJwimXKAS...	F45 Henderson	Henderson	NV	Interval Training...	5.0	I love this place...
zyT1smf2oJwimXKAS...	F45 Henderson	Henderson	NV	Interval Training...	5.0	I rarely write re...
zyT1smf2oJwimXKAS...	F45 Henderson	Henderson	NV	Interval Training...	5.0	These workouts ar...
zyT1smf2oJwimXKAS...	F45 Henderson	Henderson	NV	Interval Training...	5.0	This workout roya...
zyT1smf2oJwimXKAS...	F45 Henderson	Henderson	NV	Interval Training...	5.0	Full disclosure- ...
zyT1smf2oJwimXKAS...	F45 Henderson	Henderson	NV	Interval Training...	5.0	My daughter told ...
zyT1smf2oJwimXKAS...	F45 Henderson	Henderson	NV	Interval Training...	5.0	Joining F45 on a ...
zyT1smf2oJwimXKAS...	F45 Henderson	Henderson	NV	Interval Training...	5.0	This is by far my...
zyT1smf2oJwimXKAS...	F45 Henderson	Henderson	NV	Interval Training...	5.0	F45 is my new fav...
zxLnc3U7SJANsjKv9...	Bikram Yoga North...	Scottsdale	AZ	Yoga, Fitness & I...	5.0	Can't say enough ...

Figure 3: Filtered Data

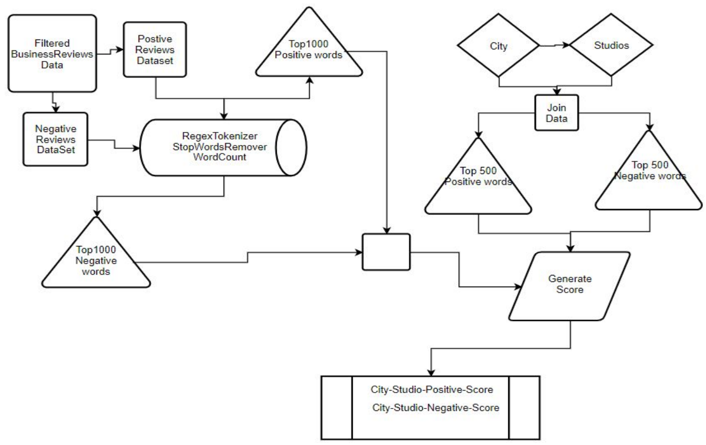

Figure 4: Studio Preference Microanalysis

4.2 Producer - Consumer Model

The Producer Consumer model relies on a Producer, a Consumer and Multiple topics. Currently there is a single producer and a single consumer. This might scale because the producer for the model is the data that is city wide. Later on the producer model can be expanded by having precise data from each zip code. The Consumer can then perform the data cleaning and manipulation part and then the data can be passed to the analysis engine.

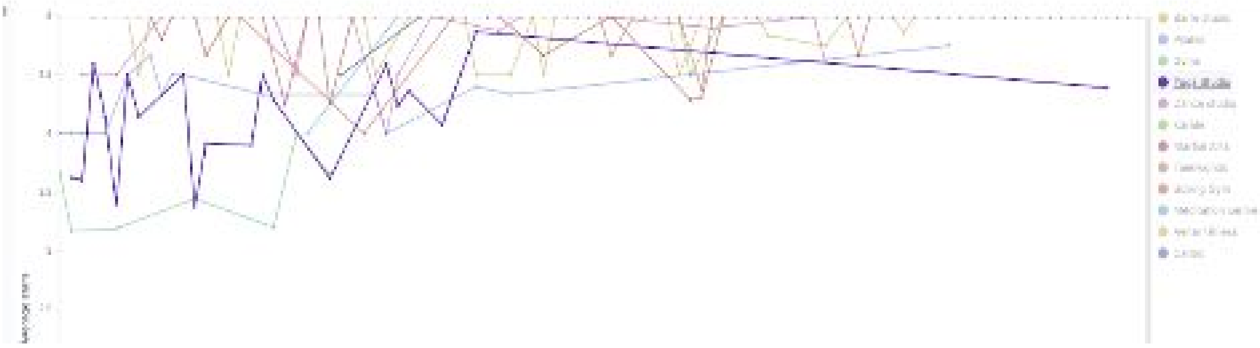

Figure 5: Line chart for Average Stars received by each Fitness Studio type

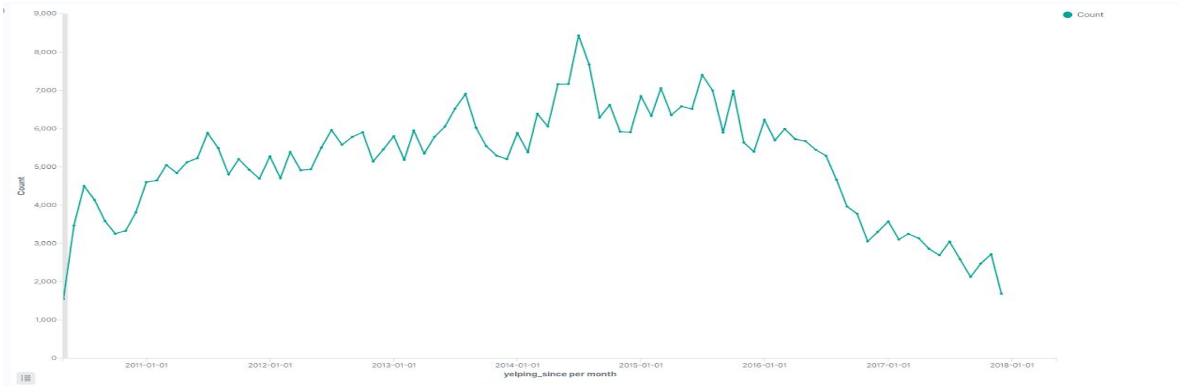

Figure 6: User trends

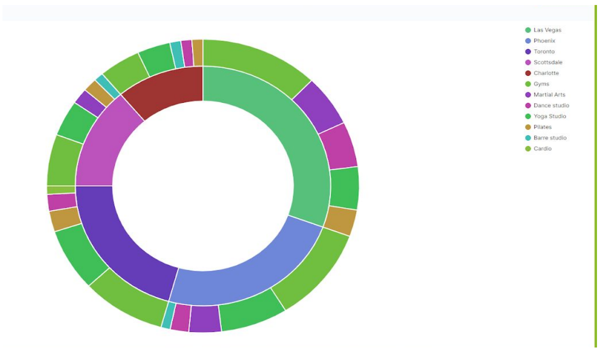

Figure 7: Donut Chart for Fitness Studio types per city

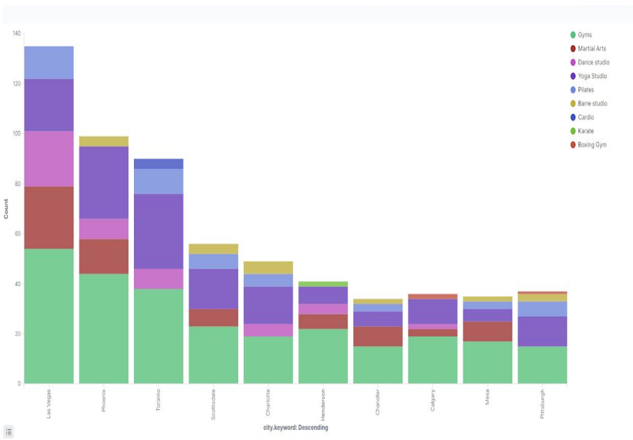

Figure 8: Stacked bar chart for Fitness Studio types per city

4.3 Analysis

ElasticSearch is used to make the dataset ready for visualization using Kibana. The Steps used in ElasticSearch are:

- Created es mapping to create a document for every row
- Inserted each document into elasticsearch
- Created an es index for data aggregation and visualization

The visualizations created using Kibana helps us understand the trends of preferences for Fitness Studios. Figure 8 is a stacked bar chart and each bar represents a city. Each city again shows the distribution of various Fitness Studios in that city. Figure 7 is a donut chart where inner donut represents cities and outer donut represents Fitness Studio types. Figure 5 in a multiple line chart. Here y-axis represents the Average score for each of the Fitness Studio and x-axis represents time, The line chart shows the trend of changing average scores over time for each Fitness Studio type.

Figure 6 shows the User signup trends. It is an indicator that new users want to join Fitness Studios or are looking for a new studio. So by looking at the signup trends business can get insight on when to roll out new membership offers or ad/marketing campaigns.

5. Achieved Goals

As a result of the Microanalysis (figure 4) done, we have the keywords that are derived from positive reviews of the customers and keywords that are derived from negative reviews of the customer.

word	csptotalocc	cspwordpoints
gym	1071	0.024875159679479734
great	752	0.017466031819765415
like	524	0.012170479619091859
place	523	0.012147253512948554
class	519	0.012054349088375334
fitness	515	0.011961444663802114
get	509	0.011822088026942283
classes	506	0.011752409708512369
time	455	0.010567878295203809
one	428	0.009940773429334571
workout	426	0.009894321217047962
really	423	0.009824642898618046
love	419	0.009731738474044826
staff	392	0.009104633608175589
always	347	0.008059458831726861
work	345	0.008013006619440025
people	322	0.007478806178144234
friendly	312	0.007246545116711183
good	310	0.007200092904424573
also	308	0.007153640692137...
best	294	0.006828475206131692
clean	271	0.006294274764835676
fun	270	0.00627104865869237
feel	270	0.00627104865869237
training	252	0.005852978748112879
nice	249	0.005783300429682964
well	246	0.005713622111253048
first	246	0.005713622111253048

Figure 9: Positive Keywords

The above table shows the positive keywords table and similarly there is a table with negative keywords. This gives us the sense of the number of occurrences of these words in the positive reviews and the last column is the score of that word with respect to the total number of occurrences of that word in the entire text.

```
('Tempe', 'Gym', 'pos', 29.397)
('Tempe', 'Yoga', 'pos', 28.576)
('Tempe', 'Martial Arts', 'pos', 26.482)
('Tempe', 'Gym', 'neg', 24.71)
('Tempe', 'Yoga', 'neg', 23.186)
('Tempe', 'Martial Arts', 'neg', 22.002)
-----
('Pittsburg', 'Gym', 'neg', 27.594)
('Pittsburg', 'Yoga', 'neg', 27.143)
('Pittsburg', 'Gym', 'pos', 26.74)
('Pittsburg', 'Yoga', 'pos', 26.689)
-----
('Phoenix', 'Gym', 'neg', 28.371)
('Phoenix', 'Yoga', 'pos', 28.234)
('Phoenix', 'Yoga', 'neg', 28.132)
('Phoenix', 'Gym', 'pos', 26.213)
-----
('Las Vegas', 'Yoga', 'pos', 29.448)
('Las Vegas', 'Gym', 'pos', 28.627)
('Las Vegas', 'Martial Arts', 'pos', 28.153)
('Las Vegas', 'Gym', 'neg', 23.707)
('Las Vegas', 'Yoga', 'neg', 22.373)
('Las Vegas', 'Martial Arts', 'neg', 22.176)
-----
('Cleveland', 'Yoga', 'pos', 29.293)
('Cleveland', 'Gym', 'pos', 26.55)
('Cleveland', 'Martial Arts', 'pos', 23.446)
('Cleveland', 'Gym', 'neg', 23.009)
('Cleveland', 'Yoga', 'neg', 19.474)
('Cleveland', 'Martial Arts', 'neg', 19.01)
```

Figure 10: Final scores

The second part of microanalysis is majorly for the generation of score pertaining to a specific fitness studio in a specific city. So figure 9 shows us the positive and negative scores of fitness studios in some cities. Looking at the first 2 results we see that yoga has a more positive score than gym in Phoenix which implies that people in Phoenix prefer Yoga centres over gyms.

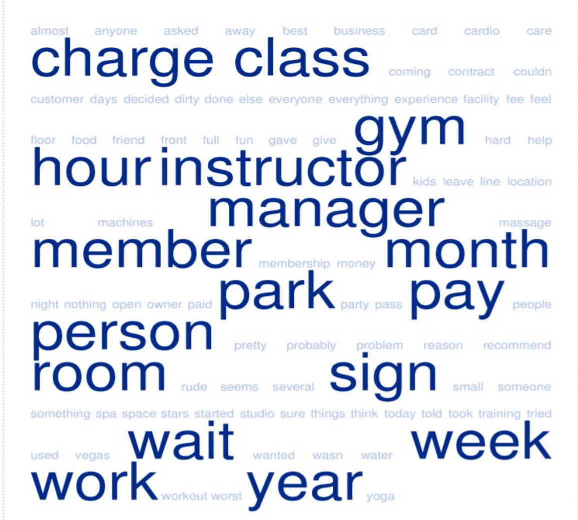

Figure 11: Negative Keywords

Figure 10 shows the Negative Keywords which are visualized in this format to make it easy for the user to understand.

Thus using the outcomes shown in figure 8, 9 and 10, we come to know the keywords present in positive/negative reviews and thus existing fitness studio managements can work on these factors to improve the quality of their service.

The score based on city and fitness studio makes it very easy for the entrepreneurs to study the trend in some specific city before and analyse the business requirements.

6. Conclusion

A model for finding the scores for Fitness Studios in different cities by analysing the user reviews of that specific fitness studio in that city. Our main goal was to provide the entrepreneurs with an analysis study of the preferences of people when it comes to choosing fitness studios and that we have achieved using Kibana visualizations. The extraction of positive and negative keywords from the user reviews is also done which would help the franchises to improve the service quality and design marketing plans.

7. Future Scope

Scalability needs to be achieved in the production environment and the plans for its implementation is:

- For each city filter business by category Active Life/Fitness Instruction(Upto 200k business, 20M+ reviews).
- Create indexes on elasticsearch based on metric to be evaluated
- Display both macro and micro level metric on Kibana to derive insight.
- We have used only Unigram but the feature will be more robust and useful after we use bigram and trigrams