BUSINESS ANALYSIS FOR CAFÉ PROPRIETORS

PAUL RACHAPONG CHIRARATTANANON

PROPOSAL PART

FOR MILESTONE 4, PLEASE SKIP TO PAGE 11

PROPOSAL PART

- This presentation doubles for both Milestone 1 and Milestone 4.
- For graders of milestone 4, please skip to page 11.
- Graders of milestone 1, please begin grading from the next page.
- (If you're confused why I do it this way, please take note that there's no place to upload file in Milestone 4 Submission.)

DESCRIPTION

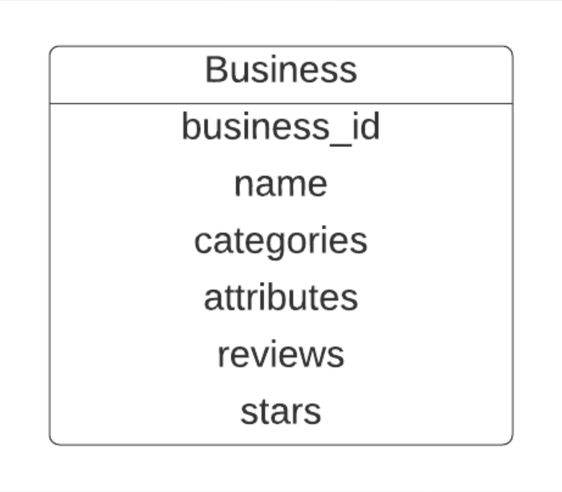
- Yelp Academic Dataset.
- Explore business dataset to uncover business insights for CoffeeKing.
- Chosen because we know little about lobbying and sports, and should have more ideas about a more ‘essential’ business.
- Current and would-be café proprietors as well as investors are our target audience.

STEPS TO IMPORT AND CLEAN THE DATA

- The business dataset subset of Yelp Academic Dataset was downloaded from Yelp website.
- A JSON file of business dataset was uploaded onto Databricks platform.
- A temporary view was created, which enables SQL queries.
- Then it’s about making the right queries to obtain the right information

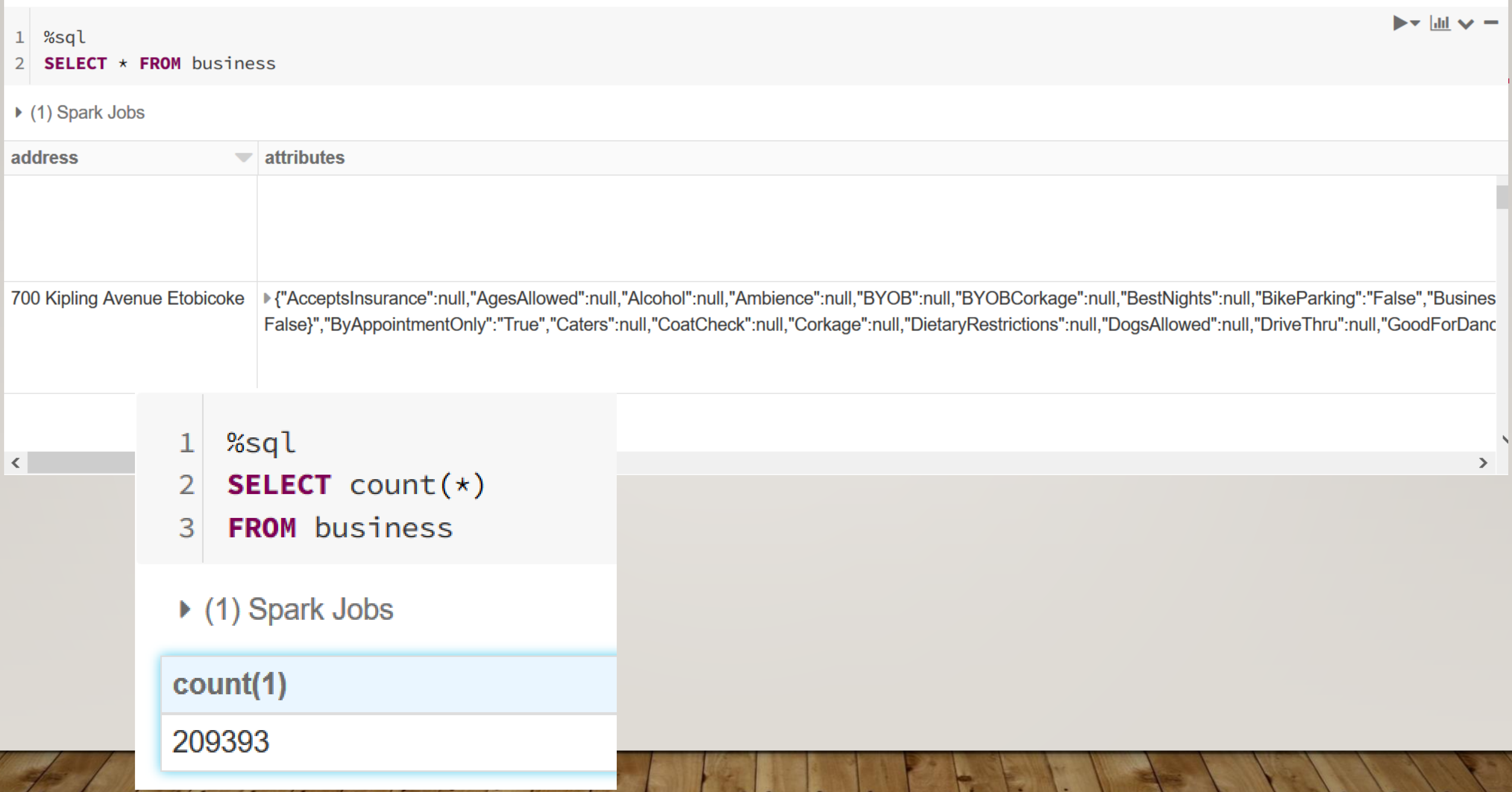
ER DIAGRAM

- Only using one dataset, so technically no entity relationship.
- These are the columns we plan to use (there are more).



Business
business_id
name
categories
attributes
reviews
stars

SOME INITIAL VIEWS OF THE DATA



```
%sql
SELECT * FROM business
```

▸ (1) Spark Jobs

address	attributes
700 Kipling Avenue Etobicoke	▸{"AcceptsInsurance":null,"AgesAllowed":null,"Alcohol":null,"Ambience":null,"BYOB":null,"BYOBCorkage":null,"BestNights":null,"BikeParking":"False","Busines False}","ByAppointmentOnly":"True","Caters":null,"CoatCheck":null,"Corkage":null,"DietaryRestrictions":null,"DogsAllowed":null,"DriveThru":null,"GoodForDanc

```
%sql
SELECT count(*)
FROM business
```

▸ (1) Spark Jobs

count(1)
209393

QUESTIONS

- Can we predict a café’s rating by some of its attributes such as wifi availability?
- Can we predict a café’s rating by type of (table vs counter) service?
- Can we predict a café’s rating by its price range?
- Can we predict a café’s rating by presence of outdoor seating?

HYPOTHESIS

- A café’s rating should be positively correlated with wifi availability.
- A café’s rating should be positively correlated with outdoor seating availability.
- A café’s rating should be positively correlated with suitability for groups.

APPROACH

- Will look at the aforementioned features as well as ratings first.
- Might add more variables later if we fail to find any form of correlation from our initial assumption.
- Will perform multivariate regression on the data, and may even resort to neural networks if it seems needed.

FINDINGS PART

FOR MILESTONE 1, PLEASE RETURN TO PAGE 3

WHO YOU ARE, AND WHY YOU ARE HERE

- You're here because you're café proprietors, restaurant proprietors, or investors.
- You want to know what makes up a successful café, so that you can improve your business or make the right investment.

HYPOTHESIS

- A café's rating should be positively correlated with wifi availability.
- A café's rating should be positively correlated with outdoor seating availability.
- A café's rating should be positively correlated with suitability for groups.
- Other findings were added to the questions later.

WHAT WE DID

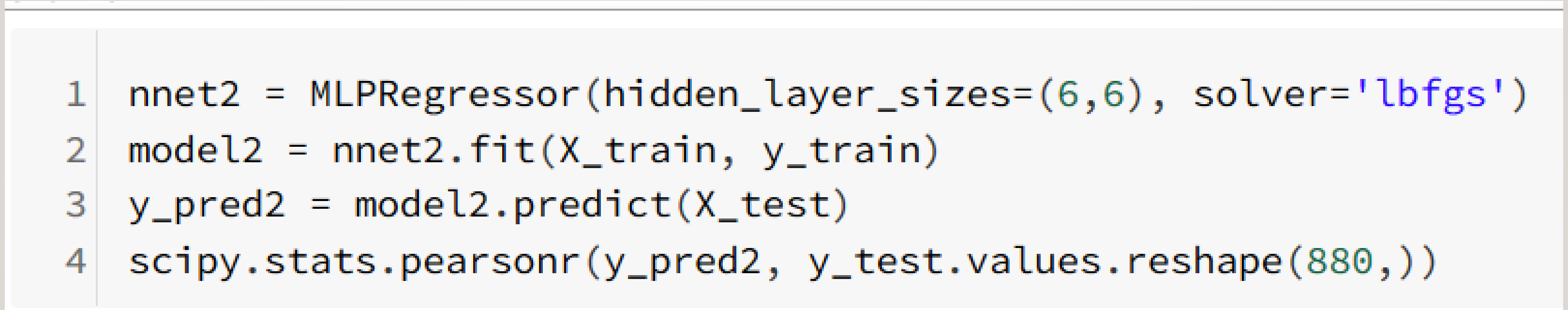
- We looked at some of the attributes of existing cafés, and the corresponding ratings (stars) they receive.
- 6 attributes were found to be correlated with café rating, with high statistical significance.

6 ATTRIBUTES MAKING A SUCCESSFUL CAFÉ

- WiFi – correlation -0.085, no WiFi is actually better, but not strong identifier.
- Alcohol – correlation 0.185, selling alcohol raises the place's popularity.
- Price Range – correlation 0.212, budget price not preferred by customers, mid-range better.
- Outdoor Seating – correlation 0.253, café patrons prefer outdoor seating.
- Brunch Service – correlation 0.203, brunch time is a good time for cafés, …
- Lunch Service – correlation 0.181, as well as lunch. Other meals and dessert not correlated.
- 1 is total correlation and -1 is total negative correlation, 0 is absolute no correlation.

PREDICTING RATINGS WITH 6 ATTRIBUTES

- Using 4 layers neural network, as below.
- Correlation between predicted ratings and actual ratings are 0.411
- P-value of correlation is 4.0e-37, virtually 100% certain of the correlation.



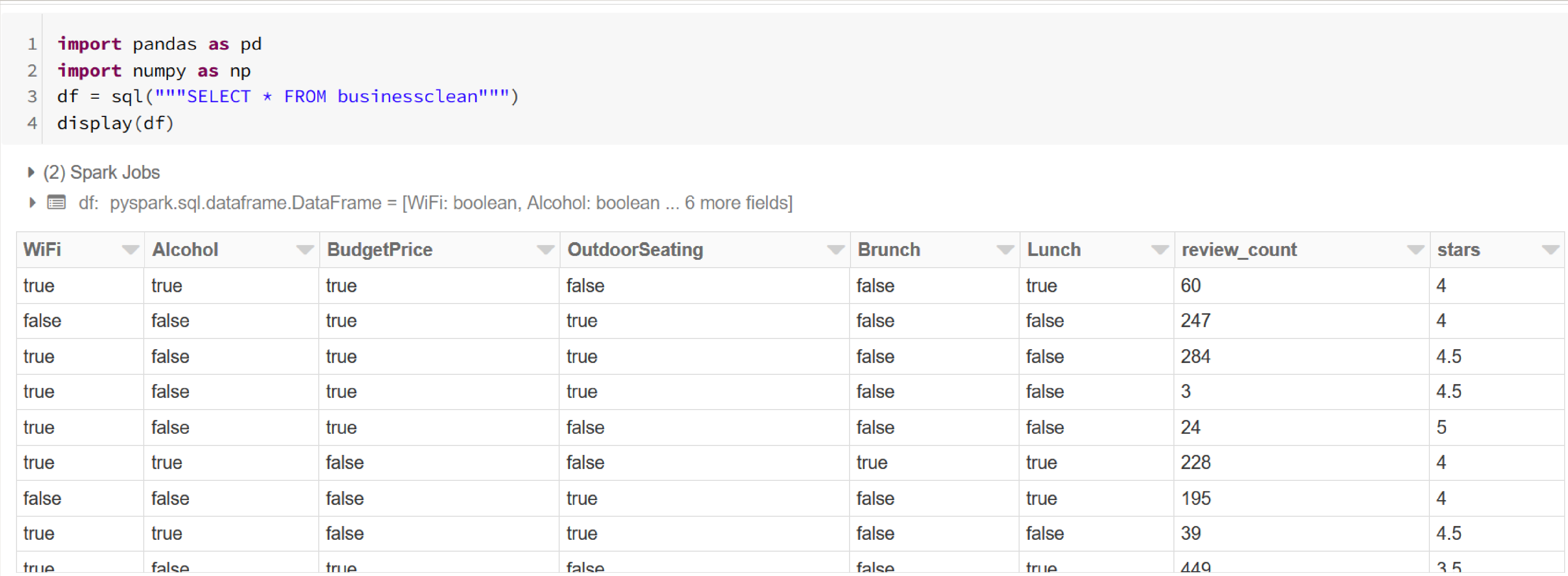
```
nnet2 = MLPRegressor(hidden_layer_sizes=(6,6), solver='lbfgs')
model2 = nnet2.fit(X_train, y_train)
y_pred2 = model2.predict(X_test)
scipy.stats.pearsonr(y_pred2, y_test.values.reshape(880,))
```

SQL PART

NEXT 2 PAGES FOR SQL LEARNERS.

PROPRIETORS PLEASE SKIP TO THE FINAL FEW PAGES OF PRESENTATION.

DATA TABLE USED TO ARRIVE AT CONCLUSION



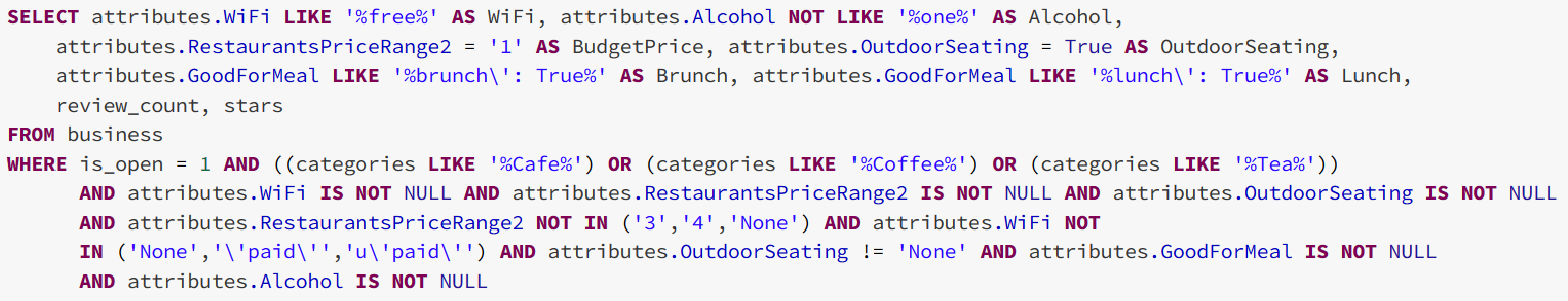
```
import pandas as pd
import numpy as np
df = sql("""SELECT * FROM businessclean""")
display(df)
```

▸ (2) Spark Jobs

▸ df: pyspark.sql.dataframe.DataFrame = [WiFi: boolean, Alcohol: boolean ... 6 more fields]

WiFi	Alcohol	BudgetPrice	OutdoorSeating	Brunch	Lunch	review_count	stars
true	true	true	false	false	true	60	4
false	false	true	true	false	false	247	4
true	false	true	true	false	false	284	4.5
true	false	true	true	false	false	3	4.5
true	false	true	false	false	false	24	5
true	true	false	false	true	true	228	4
false	false	false	true	false	true	195	4
true	true	false	true	false	false	39	4.5
true	false	true	false	false	true	449	3.5

SQL QUERY USED TO OBTAIN TABLE



```sql
SELECT attributes.WiFi LIKE '%free%' AS WiFi, attributes.Alcohol NOT LIKE '%one%' AS Alcohol,
    attributes.RestaurantsPriceRange2 = '1' AS BudgetPrice, attributes.OutdoorSeating = True AS OutdoorSeating,
    attributes.GoodForMeal LIKE '%brunch\': True%' AS Brunch, attributes.GoodForMeal LIKE '%lunch\': True%' AS Lunch,
    review_count, stars
FROM business
WHERE is_open = 1 AND ((categories LIKE '%Cafe%') OR (categories LIKE '%Coffee%') OR (categories LIKE '%Tea%'))
      AND attributes.WiFi IS NOT NULL AND attributes.RestaurantsPriceRange2 IS NOT NULL AND attributes.OutdoorSeating IS NOT NULL
      AND attributes.RestaurantsPriceRange2 NOT IN ('3','4','None') AND attributes.WiFi NOT
      IN ('None','\'paid\'','u\'paid\'') AND attributes.OutdoorSeating != 'None' AND attributes.GoodForMeal IS NOT NULL
      AND attributes.Alcohol IS NOT NULL
```

DISCUSSION AND RECOMMENDATIONS

DISCUSSION

- From 3 hypothesis, one was rejected, one was verified, and one was discarded due to incompatible information.
- 4 more hypothesis was later created and verified, totally 6 verified hypothesis.
- Didn’t create a new metric per se, but mined text from the nested layer of JSON that was not readily extracted by SQL and Python.
- We found correlations which could improve a café business.

RECOMMENDATIONS

- WiFi is not needed in a café, but times are changing and correlations are low, so this may change in the future.
- Do serve alcohol and provide outdoor seating.
- Cafés should be open for brunch and lunch times.
- If capital permits, do cater for mid-ranged patrons. Budget places do not rank well.
- All being said, these insights assume popularity takes precedence over other metrics.
- Profit could be a different matter entirely, will need to be investigated.

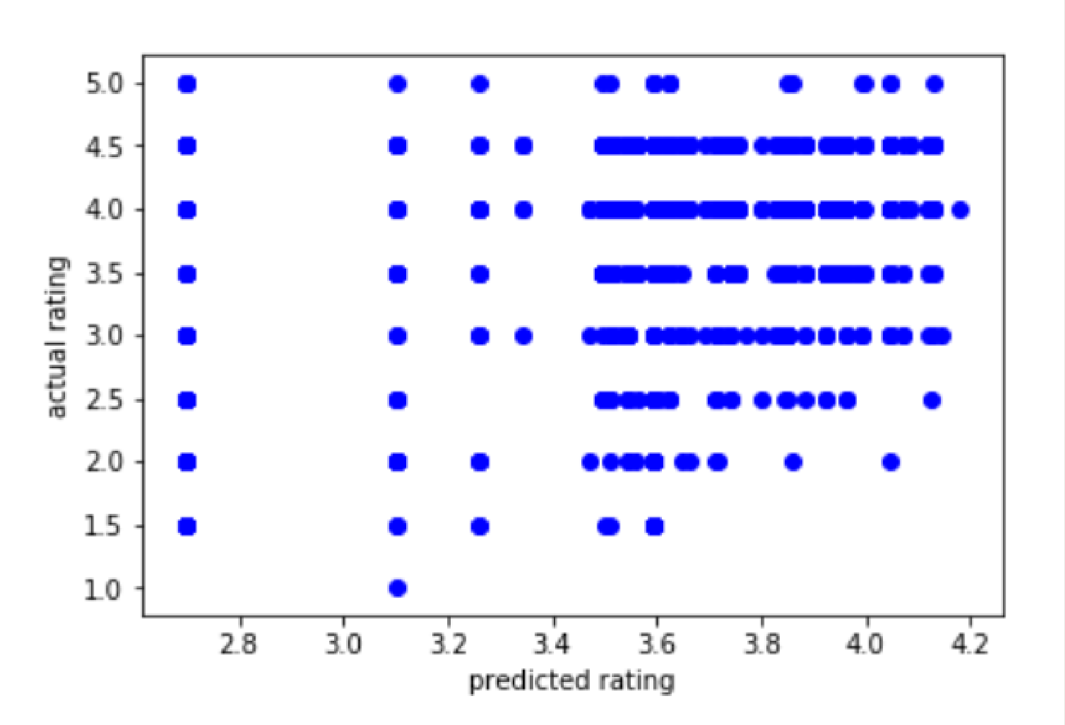
APPENDIX

A LITTLE FURTHER TECHNICAL INFORMATION

SAMPLE SIZE

- Total 209393 businesses in the dataset, but not all are cafés or restaurants.
- 2200 businesses left after filtering out non-related businesses and those with not enough information.
- Train-Test split was 60-40, putting 1320 in training and 880 in testing set.

PLOT OF PREDICTED VERSUS ACTUAL RATINGS



QUESTIONS?

(THIS IS JUST TO MIMIC A REAL PRESENTATION, THE ASSIGNMENT ENDS HERE)