
ANIMOTO
TRIAL


~60% failure rate
Beef
Ojo de Kobe $202
Lomo de Kobe $161
Cuadril de Kobe $82
Bife de Kobe $150
Cadus $182
Nieto M.H
Cabrnt $45
Septimo Dia Cab. o M.H
Promo
toma 1 LLeva 1 $35
Septima Malbec $58
COPA JOVEN $ 25
COPA BARRICA $ 30
COPA "ESPECIAL" $ 36
La Cabrera

Survival Kit for Startups

yelp


+ 5.94 Gb
+ 77,000 Businesses
+ 2,200,000 Reviews
+ 550,000 Users

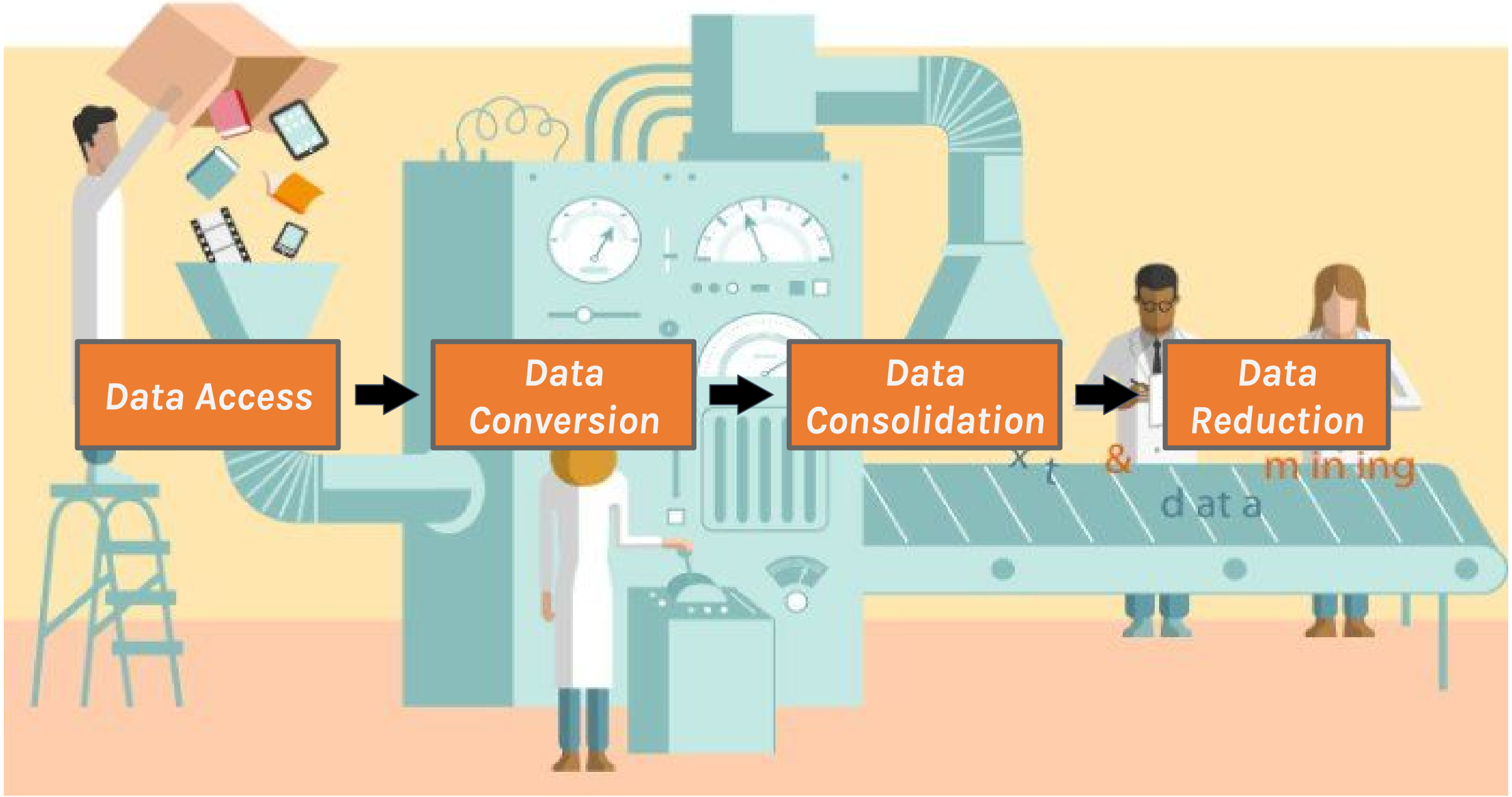

You Know Nothing, Jon Snow

Attributes

+ *70 Attributes*

What is the first thing that comes to your mind after hearing

Las Vegas?


Data Access
Data Conversion
Data Consolidation
Data Reduction
x t
&
m in ing
d at a

Data Access

- Data was made available on the link below https://www.yelp.com/dataset_challenge

- Click on "Get the Data" button to access the dataset

File_Name	Description	File Format	Size	Number of Records
yelp_academic_dataset_business	List of reviewed business	JSON	69 MB	77,000
yelp_academic_dataset_review	Review information on business	JSON	1.94 GB	2,200,000
yelp_academic_dataset_user	Information on Yelp users	JSON	236 MB	552,000
yelp_academic_dataset_checkin	Check ins at businesses	JSON	26 MB	45,166
yelp_academic_dataset_tip	Tips for each business	JSON	119 MB	591,000

Data Conversion

- Convert JSON file to usable format (csv)

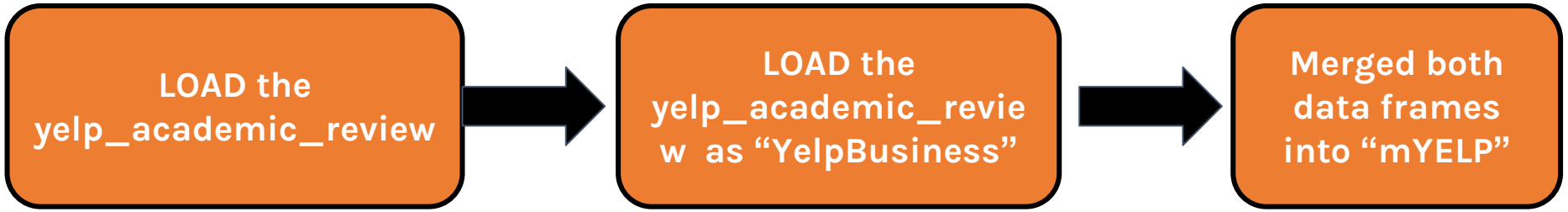
- Used Python script available from Yelp “json_to_csv_converter.py” to convert the files in csv format.

Sample Code :

```
import argparse
import collections
import csv
import simplejson as json


def read_and_write_file(json_file_path, csv_file_path, column_names):
    """Read in the json dataset file and write it out to a csv file, given the column
names."""
    with open(csv_file_path, 'wb+') as fout:
        csv_file = csv.writer(fout)
        csv_file.writerow(list(column_names))
        with open(json_file_path) as fin:
            for line in fin:
                line_contents = json.loads(line)
                csv_file.writerow(get_row(line_contents, column_names))
```

Data Consolidation



Sample Code :

```
yelpBussEnd = subset(yelpBusiness, city == "Las Vegas")
View(yelpBussEnd)

## Merged both dataframes into "mYelp" by using their business_id
## Exported it to a csv file to work it on Tableau

mYelp=merge(yelp_academic_dataset_review,yelpBussEnd,by="business_id")
write.csv(mYelp, file = "mYelp.csv")
```

Sample Code :

```
## 5 different subsets of data to account for all asian restaurants
## Grepl is a function that searches text in a variable and returns if its true.
mYelp1 = subset(mYelp, grepl("Chinese", mYelp$categories))
mYelp2 = subset(mYelp, grepl("Japanese", mYelp$categories))
mYelp3 = subset(mYelp, grepl("Sushi", mYelp$categories))
mYelp4 = subset(mYelp, grepl("Thai", mYelp$categories))
mYelp5 = subset(mYelp, grepl("Vietnamese", mYelp$categories))

## Combining all the datasets
mYelpComb = rbind(mYelp1, mYelp2, mYelp3, mYelp4, mYelp5)
```

- “**mYelpAsian**” has **95,852** observations and **97** variables

- For further analysis we divided mYelpAsian into **Training Data**(70%) and **Testing Data**(30%)

Sample Code :

```
# Show the repeat entries
df[duplicated(df),]
# Show unique repeat entries (row names may differ, but values are the same)
unique(df[duplicated(df),])
# Original data with repeats removed. These do the same:
unique(df)
df[!duplicated(df),]

# Export it as a csv

mYelpFinal = unique(mYelpComb)
write.csv(mYelpFinal, file = "mYelpAsian.csv")
```

Data Reduction

- *Data was limited to restaurants in* ***Las Vegas***
- *The User, tip and check-ins file was omitted*
- *We ran the summary() on dataset mYelpAsian and found that lot of variables have NULL values*
- *Based on the results of Summary(),we selected the below variables for further analysis*

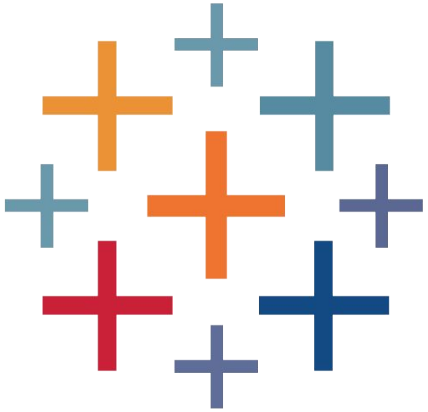
Wi-Fi	*Attire*	*Take Out*	*Credit Cards*
Parking	*TV*	*Noise*	*Ambience*
Waiter	*Delivery*	*Alcohol*	*Price Range*

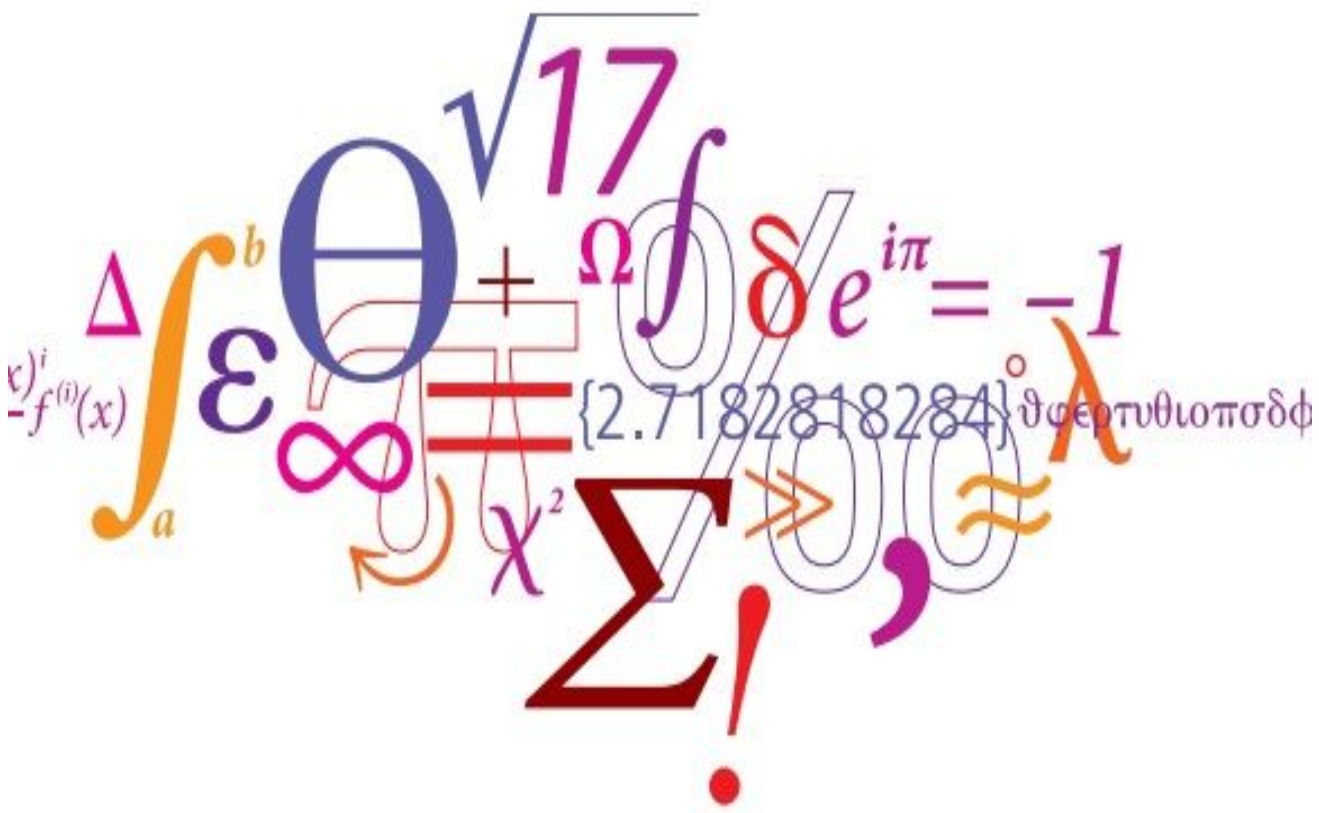
Tools we used

- Rstudio
- Rattle
- Shiny
- Microsoft Excel
- Tableau
- RAW (*http://raw.densitydesign.org*)
- Datawrangler (*http://vis.stanford.edu/wrangler/app*)

RAW

DataWrangler alpha





Featured Engineering :

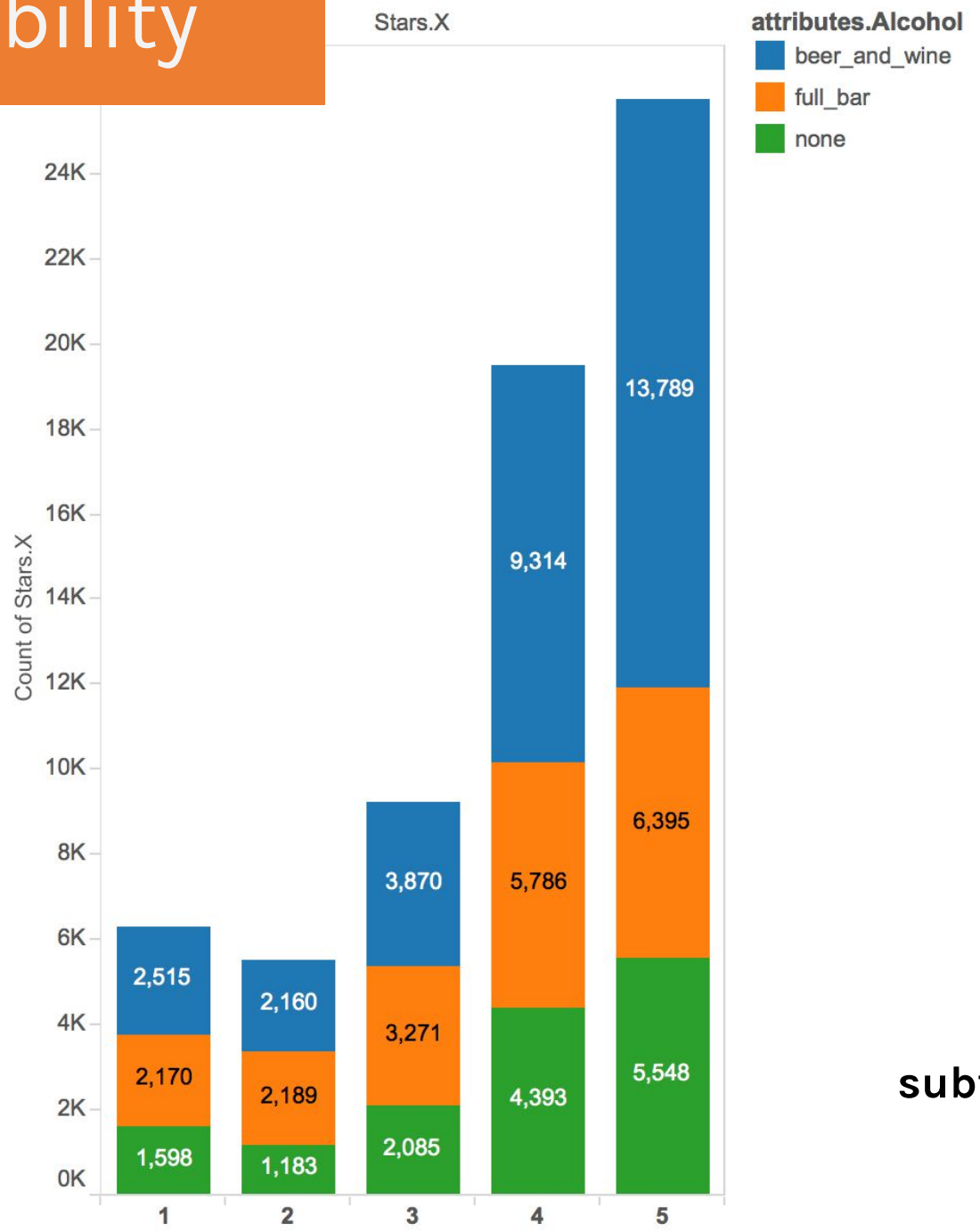
1. Combination of two or three attributes

2. Changed the multi-class variables to binary class variables

3. Classified the ratings into **GOOD** or **BAD** reviews from 1 to 5 stars

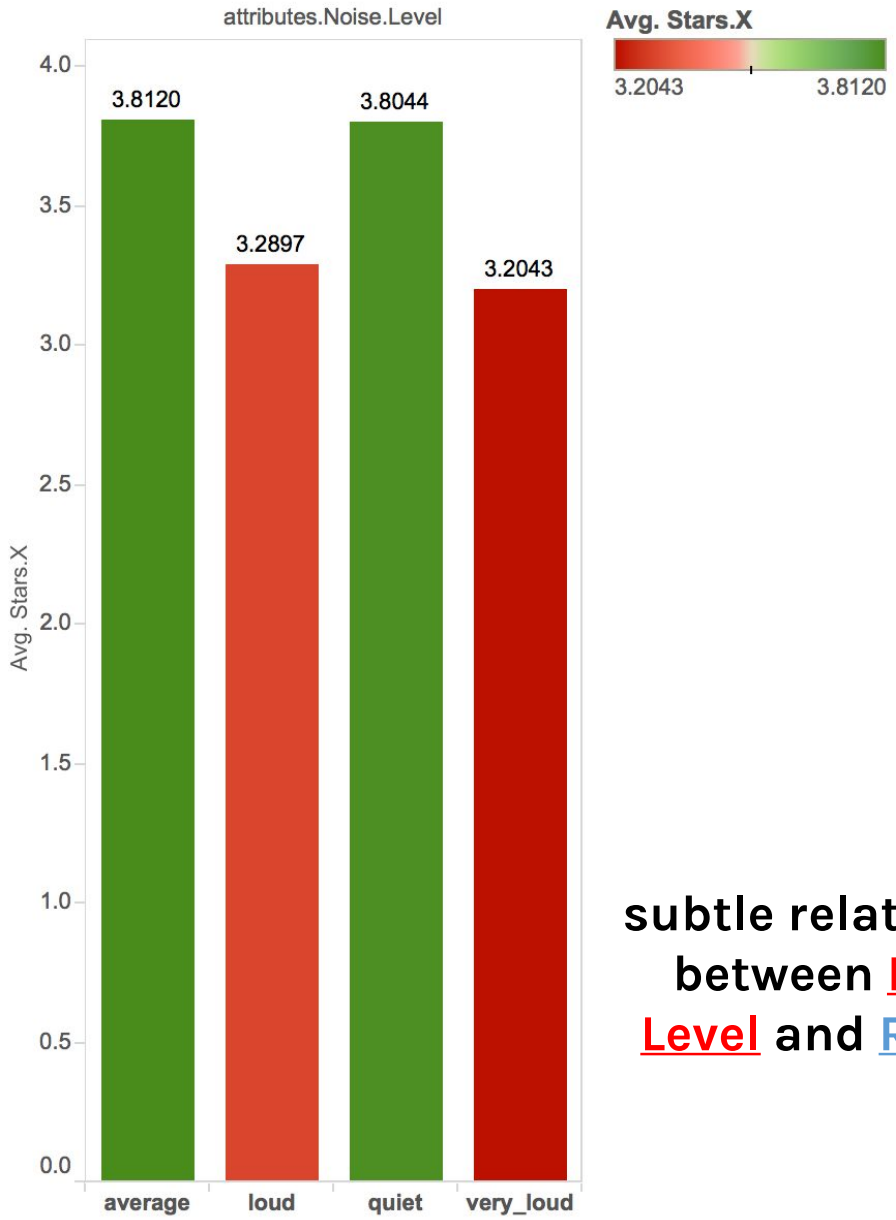
Findings...

#1: Alcohol Availability



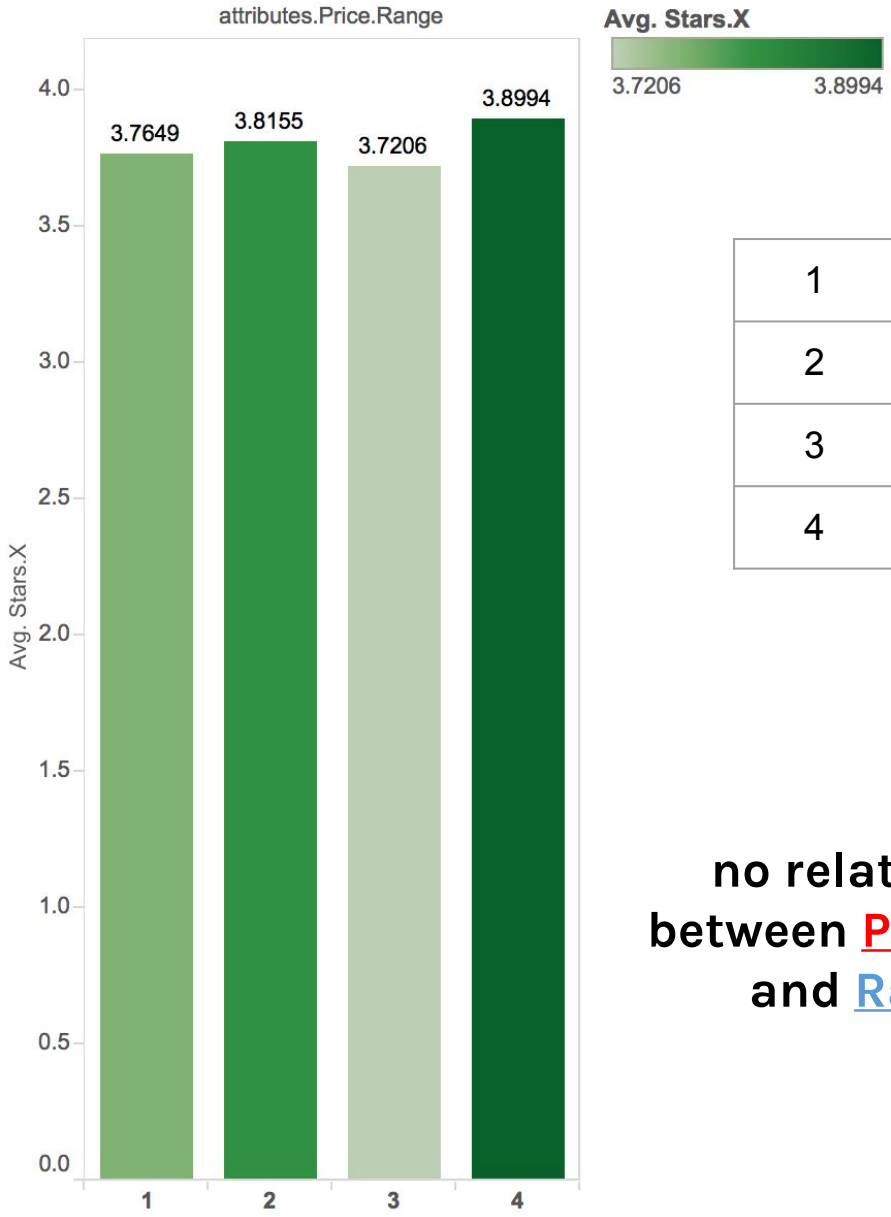
subtle relationship between Alcohol and Ratings

#2: Noise level



subtle relationship between Noise. Level and Ratings

#3: Price Range



1	<10 $
2	10-25$
3	25-45$
4	>45$

no relationship between Price.Range and Ratings

Linear Regression Model

Target Variable : **Stars.x**

Started off with **97** variables and iterated till finding relevant attributes

Final model looked at only **13** Variables

R-Square : **0.02**, Very Low value

Difference between **PREDICTED & REAL** values are **VERY HIGH**

Relevant attributes : **Order at counter, TV, Delivery, Price Range, Noise level, and others.**

```
Call:
lm(formula = stars.x ~ ., data = Yelp_train)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2912 -0.7889  0.2687  1.1008  2.7348

Coefficients:
                                          Estimate   Std. Error t value Pr
(>|t|)
(Intercept)                                3.40450    0.07905  43.069  <
2e-16 ***
attributes.Order.at.CounterFalse           0.79944    0.03657  21.860  < 2e-16 ***
attributes.Order.at.CounterTrue            0.12156    0.02781   4.372   1.24e-05 ***
……..
attributes.Noise.Levelaverage             -0.09554    0.06376  -1.498 0.134023
attributes.Noise.Levelloud                -0.49255    0.07181  -6.860 6.97
e-12 ***
attributes.Noise.Levelquiet               -0.06826    0.06426  -1.062
0.288100
attributes.Noise.Levelvery_loud           -0.51128    0.14853  -3.442 0.000577 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.278 on 67016 degrees of freedom
  (54 observations deleted due to missingness)
Multiple R-squared:  0.0269,        Adjusted R-squared:  0.02654
F-statistic: 74.11 on 25 and 67016 DF,  p-value: < 2.2e-16
```

Linear Regression Model

(using Rattle to cross-check)

Target Variable : Stars.x

Ignored all categorical variables with more than 10 unique values

Started with 54 variables and iterated till found the relevant attributes

Output : **7** statistically significant attributes

R-Squared : **0.03**

High std errors

```
==== ANOVA ====

Analysis of Variance Table

Response: stars.x
                              Df Sum Sq Mean Sq F value    Pr(>F)
attributes.Alcohol             2    996  498.14 307.864 < 2.2e-16 ***
attributes.Parking.lot         1    298  297.76 184.025 < 2.2e-16 ***
attributes.Good.For.brunch     1     61   61.22  37.837 7.757e-10 ***
attributes.Waiter.Service      1     62   62.17  38.425 5.738e-10 ***
attributes.Price.Range         1    291  291.14 179.934 < 2.2e-16 ***
attributes.Takes.Reservations  1    426  426.18 263.392 < 2.2e-16 ***
attributes.Caters              1    143  143.17  88.481 < 2.2e-16 ***
Residuals                  45260  73232    1.62
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
[1] "\n"
```

Classification Model

Target Variable : Good review (binary)

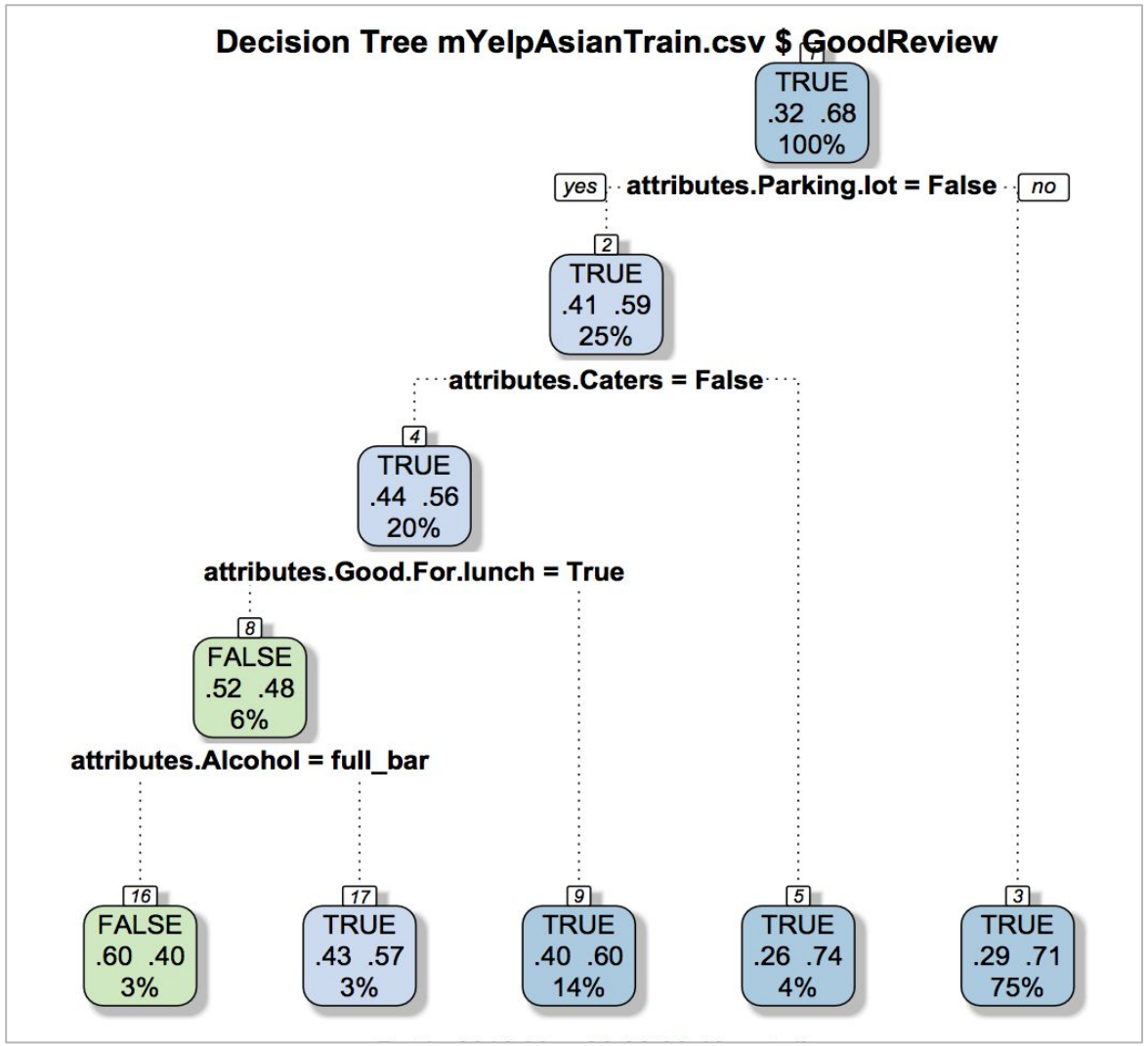
Ignored all categorical variables with more than 25 unique values

Started with 78 variables and iterated till found the relevant attributes

Output : **7** statistically significant attributes

Results are consistent with **"LINEAR MODEL"**

Relevant attributes : **Parking.lot, Caters, Good.For. Lunch and Alcohol**



Key Findings from Exploratory analysis and Analytical Models

Parking

Noise Level: AVG. or Low

Waiter Service

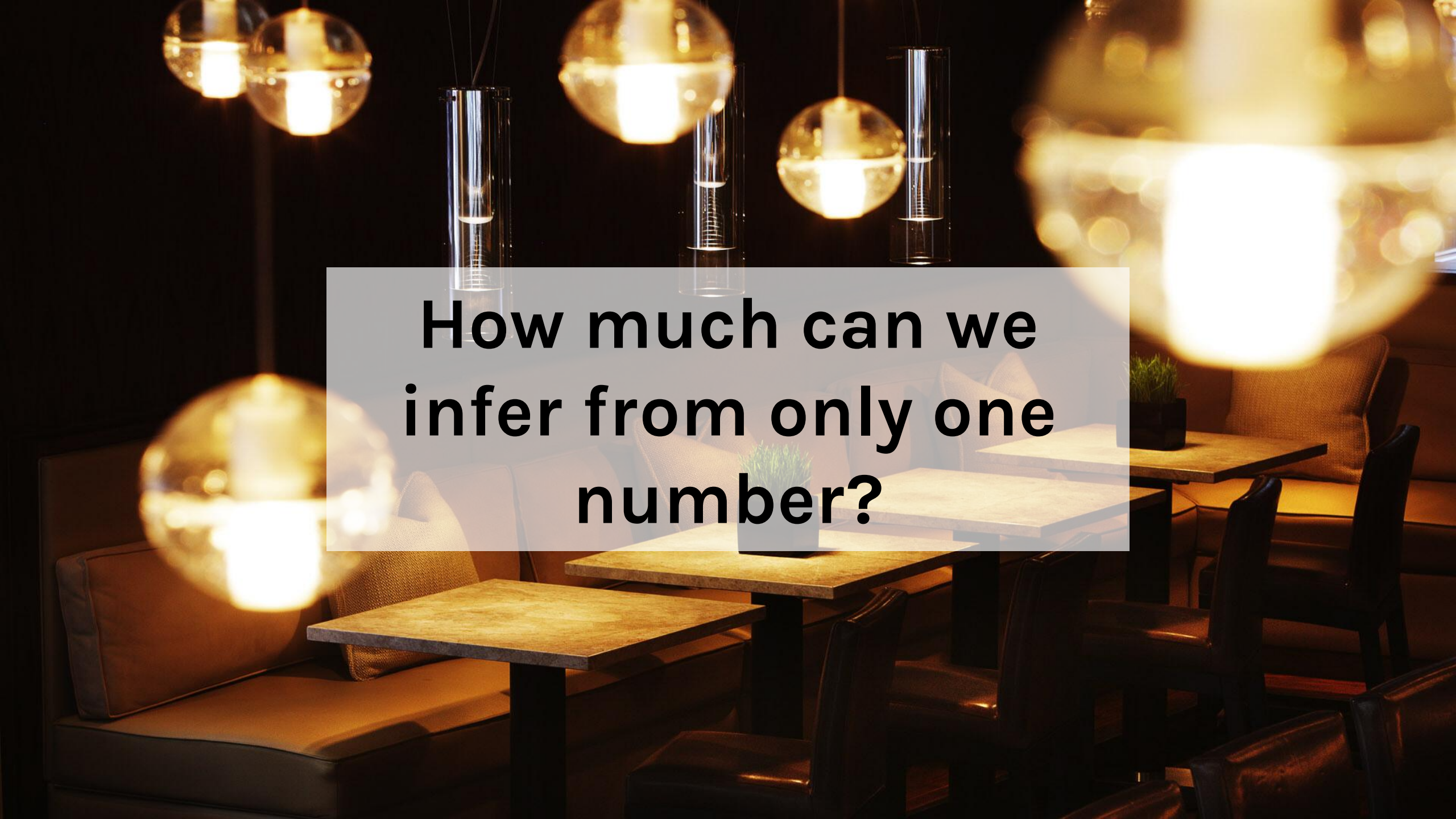
Ambience

Good for brunch

Takes Reservation

Take Out Delivery

Alcohol


How much can we infer from only one number?


How does a
review looks?

Bryan R.
New York, NY
Elite '14
115 friends
295 reviews

★★★★★ 9/2/2014

What a great place for a random night.

I stumbled upon this place on Yelp one night when I was out alone on a work trip to SF. I stopped in for a drink and stayed for a few hours. I love the charm of this place and the welcoming nature. The bar kind of naturally selects for people willing to have a conversation with the person next to them. Also, there's a great burger place nearby that you can go and bring food back from (grab a burger for the bartender too).

Chuck k.
Menlo Park, CA
134 friends
346 reviews

★☆☆☆☆ 9/28/2007

My pork chop with pomegranate sauce was actually quite tasty. So why a 1 star review? Rude, poor service.

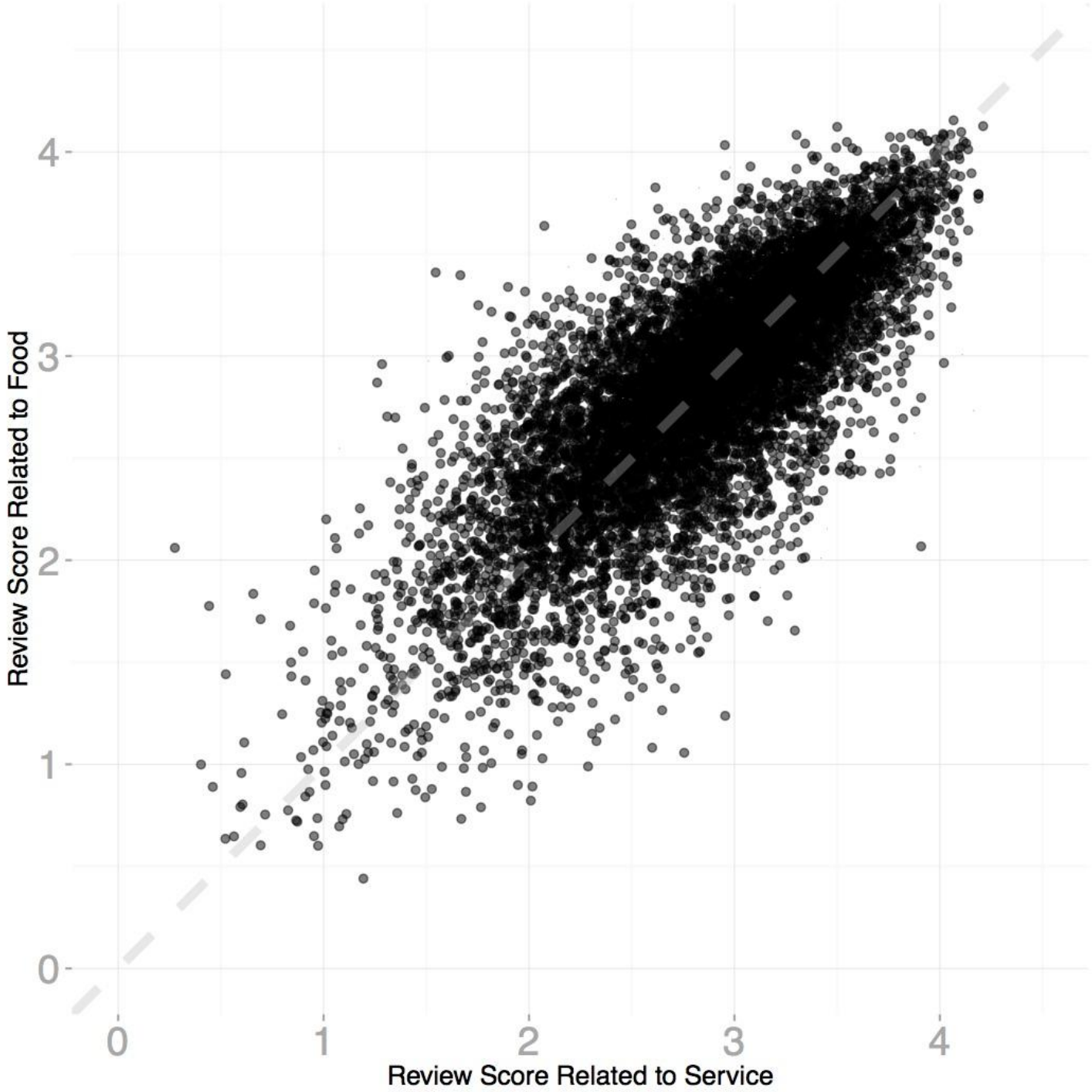
Our server was impatient, combative, and argumentative. Twice she came to the table and we were talking and didn't snap to immediately, she left and once gave us the "Wrap it up" signal with her finger.

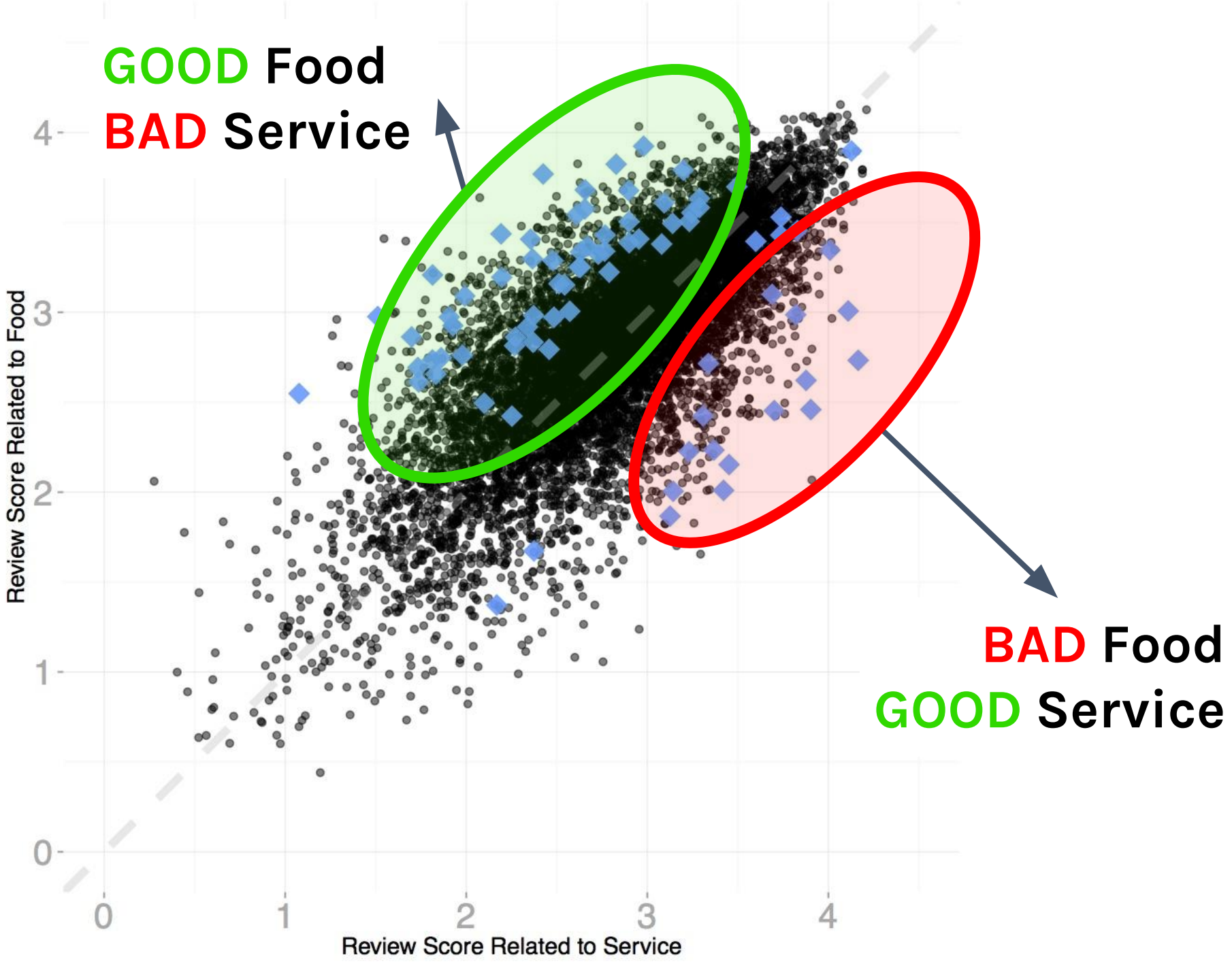
GOOD
food…
BAD
service
??

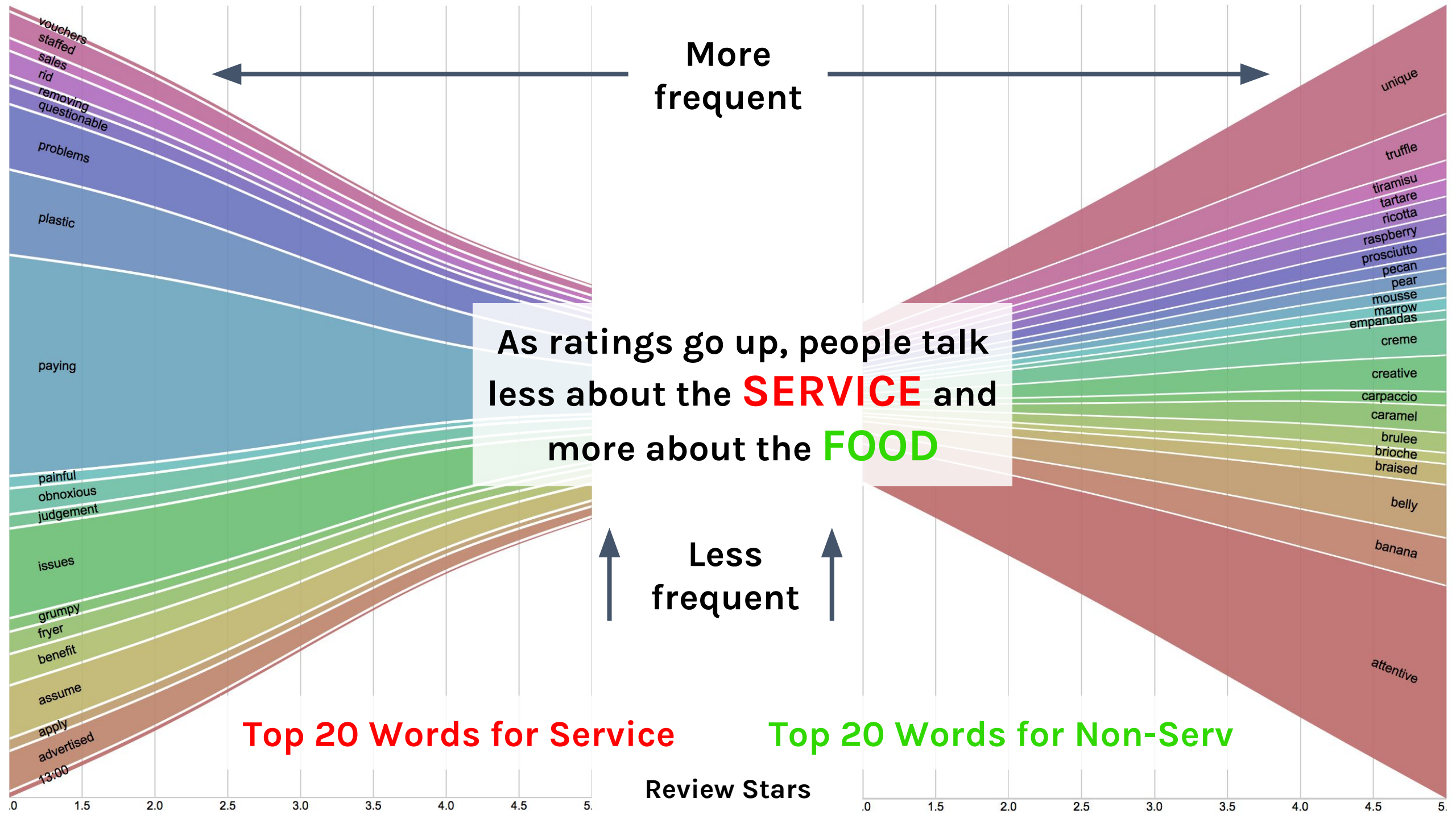

This leaves us with
inaccurate information


What if we only focused on:
Food
&
Service

Example


Review Score Related to Food
Review Score Related to Service
0
1
2
3
4


GOOD Food
BAD Service
BAD Food
GOOD Service
Review Score Related to Food
Review Score Related to Service
0
1
2
3
4


More frequent
vouchers
staffed
sales
rid
removing
questionable
problems
plastic
paying
painful
obnoxious
judgement
issues
grumpy
fryer
benefit
assume
apply
advertised
13:00
unique
truffle
tiramisu
tartare
ricotta
raspberry
prosciutto
pecan
pear
mousse
marrow
empanadas
creme
creative
carpaccio
caramel
brulee
brioche
braised
belly
banana
attentive
As ratings go up, people talk less about the SERVICE and more about the FOOD
Less frequent
Top 20 Words for Service
Top 20 Words for Non-Serv
Review Stars
1.5
2.0
2.5
3.0
3.5
4.0
4.5

More than 20 restaurants with bad service in Las Vegas

Example

NV
nds
206 reviews
Elite '16

Share review
Compliment
Send message
Follow Dana D.

4/12/2015

1 check-in

I am pissed.

I walked into goldilocks at 10:30am. They have their $7 breakfast buffet 8am-11am and they had a long line. Well because no one was at the front where it said to wait to be seated (I waited a good 5min) and since someone else just went in the line, I decided to just fall in line too. It was a pretty long line and I was in line for a good 15-20. Finally when I got to the food, the servers were like "where is your wristband?" And I didn't have one so I asked if I can pay now since no one told me. She said no then I was actually willing to get back in line since I was already there and hungry but THEN she said, "yeah then no more food for you because now it's 11" WHAT? I was obviously there way before it closed! HORRIBLE CUSTOMER SERVICE I am so insulted. No one was there to help me, they saw me fall in line and didn't say anything, then they tell me I can't have food. PLUS they still had plenty of breakfast food left. Terrible. Never coming back again. I left and went across the street to Nanay Gloria's where they actually knew how to give hospitality to their customers.

Was this review ...?

Useful 7 Funny 2 Cool 1

Frita W.
Sacramento, CA
0 friends
2 reviews

12/30/2015

I visit this place every time I'm in town. Love the food! The sotanghon is the best I've ever had.

I'm not a huge pastry fan but their cakes are not too sweet which make them perfect for my taste.

Name of Business	Abs (AvgScoreServVsAvg ScoreNonServ)
Little Italy Pizza	1.7173
Goldilocks	1.4715
MIX 94.1's Bite Of Las Vegas	1.4319
Deli Pizza Cafe	1.3833
New Grand China	1.2793
The SLS Buffet	1.2453
Pyaar India Restaurant	1.2443
Arisra Thai Seafood & Steaks	1.2104
Quickee Burgers	1.1573
Liu's Kitchen	1.1497
Hogs Heaven Barbecue	1.1457
Oriental House	1.1258
Ray's Asian Cuisine and Ho..	1.1140
Pizzalicious	1.1019
Panini Cafe	1.0998
Bomas Bar & Grill	1.0933
Pho Lan	1.0905
Sunset Pizzeria	1.0904
Shalimar Restaurant	1.0895
Cafe Heidelberg German Ma..	1.0887
Shaanxi Gourmet	1.0744
The Village Mediterranean G..	1.0667
Pho Old Saigon	1.0595
Myxx Hookah Lounge	1.0585
El Regio	1.0534

Shiny

by RStudio

A web application framework for R

Turn your analyses into interactive web applications

No HTML, CSS, or JavaScript knowledge required

Recommendations to Kim & Lim

- Users mostly share positive comments about **food** and complain about the **service**
- There's **no** strong correlation between star ratings and most attributes
- An unique numerical star rating doesn't say anything about food, service, ambience, deals and worthiness of the business.

Info about competition

Customer service

Parking

Noise level

Take Out and Delivery

Alcohol

Caters

Challenges

Large datasets require lots of computational processing power

Data interpretation is not easy and sometimes we don’t arrive at any conclusions

Not to lose focus throughout the project when using different tools and scope

Lessons

It’s important to go through all the steps in the data preparation

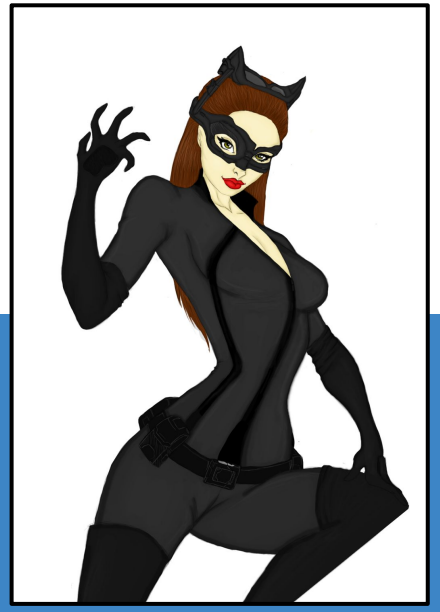
In linear regression pay attention to types of variables and content

In classification tree avoid constant variables and check consistency

Rattle is faster and more friendly to variable handling and building the decision tree

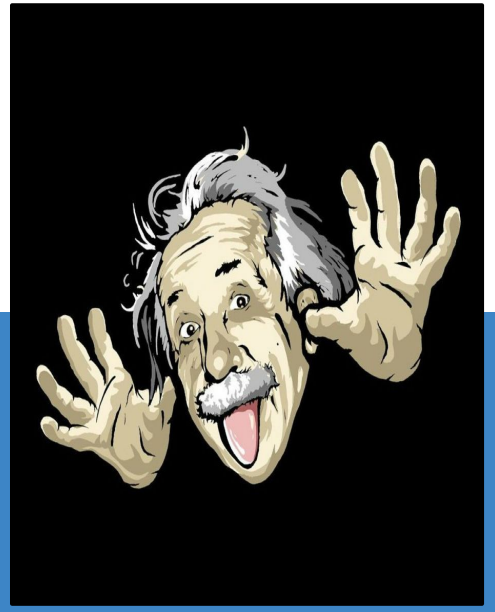
Raw and Datawrangler are great tools for visualizing spreadsheets in a different way

team Aryabhatta

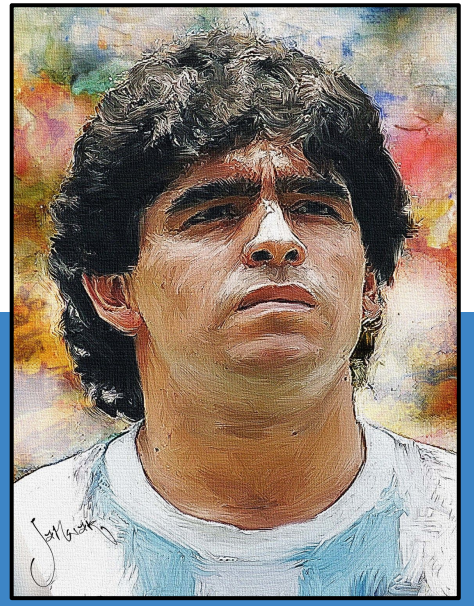


Shruti Khandelwal

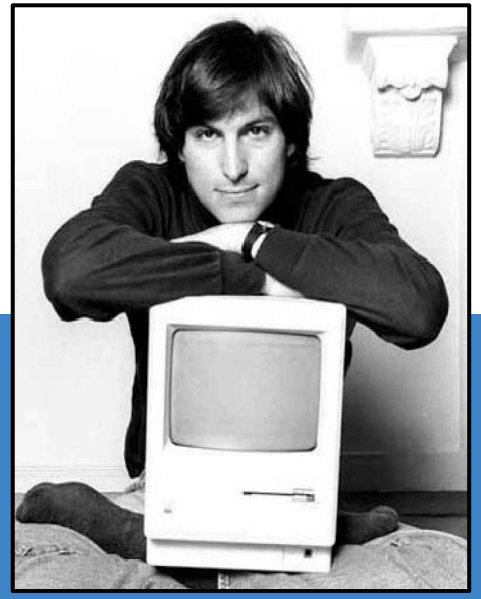
José Collado



Karan Sawant



Nicolás Metallo



Karthik Vannela