

[Write a Review](#)[Events](#)[Talk](#)[Log In](#)[Sign Up](#)

Find burgers, barbers, spas, handymen...

Near Los Angeles, CA

[Restaurants](#)[Nightlife](#)[Home Services](#)[Delivery](#)

熊燒Bar

Photo by Kun Feng L.

YELP RESTAURANTS BIG DATA ANALYSIS

CIS5200 - GROUP 5

SUBMITTED TO DR. JONGWOOK WOO

Maria Boldina

Meena Muthiah

Prasanna Natuthurai

Julia Stachurska

TABLE OF CONTENTS

- OVERVIEW
- WORKFLOW CHART
- DATA SET DETAILS
- TECHNICAL SPECIFICATIONS
- VISUALIZATIONS
- INFERRED INSIGHTS
- GITHUB LINK
- REFERENCES

OVERVIEW

- The goal of this project is to take advantage of the Big Data analysis to capture some valuable insights about the restaurant industry from the 2017 Yelp challenge data set.
- What is Yelp? Yelp is a social networking site that lets users post reviews and rate businesses.
- Why Yelp? Since its inception in 2004, Yelp has collected a staggering 142 million reviews from users for local businesses. They have an average of 145 million unique visitors to their site every month.
- Why restaurants? Because it is the 2nd largest category on Yelp

WORKFLOW CHART

RAW DATA



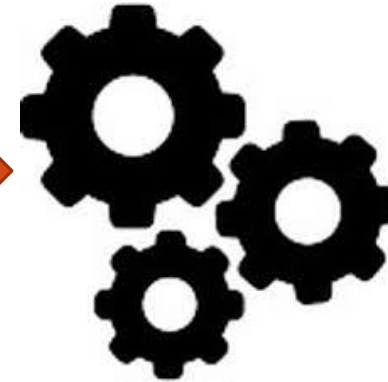
1. Download and extract data from Yelp

ETL & STORAGE



2. Upload JSON files to HDFS using Ambari
3. Install SerDe

ANALYSIS



4. Create tables and queries using HiveQL

REPORTING



5. Export results to Tableau for visualization

JSON FILE

QUERIES

EXPORT

2017 YELP CHALLENGE OPEN DATA SET

Data set URL: <https://www.yelp.com/dataset>

Data set size:

- 1 TAR file - 2.28 GB compressed
- 6 JSON files - 5.79 GB uncompressed
- Categories: business, reviews, user, check-in, tip and photos.

This set includes information about local businesses across 4 countries (USA, Canada, UK, Germany).

- 4.7M reviews
- 1.1M users
- 156K businesses
- 51K restaurants
- 12 metropolitan areas



RAW JSON FILE DATA

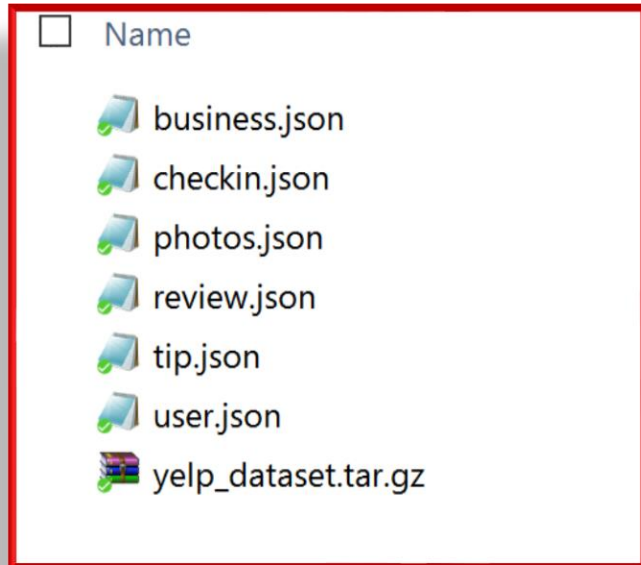
■ JavaScript Object Notation

business.json

```
1 {"business_id": "YDf95gJZaq05wvo7hTQbbQ", "name": "Richmond Town Square", "neighborhood": "", "address": "691 Richmond Rd
2 {"business_id": "mLwM-h2YhXl2NCgdS84_Bw", "name": "South Florida Style Chicken & Ribs", "neighborhood": "Eastland", "add
3 {"business_id": "v2WhjAB3PIBA8J8VxG3wEg", "name": "The Tea Emporium", "neighborhood": "Riverdale", "address": "337 Danfor
4 {"business_id": "CVtCbSB1zUcUWg-9TNGTuQ", "name": "TRUmatch", "neighborhood": "", "address": "7702 E Doubletree Ranch Rd
5 {"business_id": "duHFB87uNSXImQmvBh87Q", "name": "Blimpie", "neighborhood": "", "address": "4719 N 20Th St", "city": "Ph
6 {"business_id": "uUEMrhJiL1a1pCA_I1SU7Q", "name": "Baxter's Cigars", "neighborhood": "", "address": "2017 E Camelback Rd
7 {"business_id": "2eJEUJIP54tex7T9Y0cLSw", "name": "Back-Health Chiropractic", "neighborhood": "", "address": "4425 N 24th
8 {"business_id": "fEylCY3UEH8YJ0Xa7lu6lA", "name": "Auto Bathhouse", "neighborhood": "Lawrenceville", "address": "5770 Butl
9 {"business_id": "kFtuYklkAIImYw8RZAieGw", "name": "JAB Jewelry Designs", "neighborhood": "", "address": "3220 Washington
10 {"business_id": "NqiQdFa93wzUJGo29NbTPQ", "name": "Neighborhood Vision Center", "neighborhood": "", "address": "1425 S H
11 {"business_id": "N9BN9ldVl1FNzcB9_eAstw", "name": "Red Rock Bowling UYE Part 2", "neighborhood": "Summerlin", "address":
12 {"business_id": "6s3z3TlPHOIecuSyPEOp7A", "name": "Sq Cutz", "neighborhood": "", "address": "9393 N 90th St, Ste 112", "c
13 {"business_id": "n33Izvzk_z9_51H6NsQF-A", "name": "Safeway", "neighborhood": "", "address": "9101 E Baseline Rd", "city"
14 {"business_id": "m06OZRFTaKKi6U0omfLq4g", "name": "Artificial Grass Masters", "neighborhood": "", "address": "12417 W Pir
15 {"business_id": "SDMRxmCKPnt1AHPBKq064Q", "name": "Applebee's", "neighborhood": "", "address": "9616 E Independence Blvd
16 {"business_id": "iFEiMJoEqyB9080UNSDlZa", "name": "China Garden", "neighborhood": "", "address": "190 E Dallas Rd", "city
17 {"business_id": "qrAht4wWRYWj1sEjxq574A", "name": "Beach Ventures Roofing", "neighborhood": "", "address": "62 S Ctr", "c
18 {"business_id": "jqp5TibmgJZVFKHTp3XfqQ", "name": "Alpaul Automobile Wash", "neighborhood": "", "address": "2128 Warrens
19 {"business_id": "1eLYCXThDYZFHLK_KoPltka", "name": "Koko Bakery", "neighborhood": "Goodrich-Kintland", "address": "2710 D
```

JSON FILE SCHEMA

business.json



```
{
  "business_id": "4bEjOyTaDG24SY5TxsaUNQ",
  "full_address": "3655 Las Vegas Blvd S\nThe Strip\nLas Vegas, NV 89109",
  "hours": {
    "Monday": {"close": "23:00", "open": "07:00"},
    "Tuesday": {"close": "23:00", "open": "07:00"},
    "Friday": {"close": "00:00", "open": "07:00"},
    "Wednesday": {"close": "23:00", "open": "07:00"},
    "Thursday": {"close": "23:00", "open": "07:00"},
    "Sunday": {"close": "23:00", "open": "07:00"},
    "Saturday": {"close": "00:00", "open": "07:00"}
  },
  "open": true,
  "categories": ["Breakfast & Brunch", "Steakhouses", "French", "Restaurants"],
  "city": "Las Vegas",
  "review_count": 4084,
  "name": "Mon Ami Gabi",
  "neighborhoods": ["The Strip"],
  "longitude": -115.172588519464,
  "state": "NV",
  "stars": 4.0,
  "attributes": {
    "Alcohol": "full_bar",
    "Noise Level": "average",
    "Has TV": false,
    "Attire": "casual",
    "Ambience": {
      "romantic": true,
      "intimate": false,
      "touristy": false,
      "hipster": false,
      "classy": true,
      "trendy": false,
      "casual": false
    }
  }
},
```



RCONGIU JSON SERDE

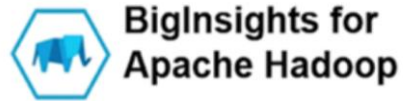
- <https://github.com/rcongiu/Hive-JSON-Serde>
- Read/write data stored in JSON format
- Support for JSON arrays, maps and nested structures
- HIVE > ADD JAR json-serde-1.3.8-jar-with-dependencies.jar;
- TABLE> ROW FORMAT SERDE 'org.openx.data.jsonserde.JsonSerDe'

```
-bash-4.1$ wget -O json-serde-1.3.8-jar-with-dependencies.jar www.congiu.net/hive-json-serde/1.3.8/hdp23/json-serde-1.3.8-jar-with-dependencies.jar;
--2017-10-26 01:08:08-- http://www.congiu.net/hive-json-serde/1.3.8/hdp23/json-serde-1.3.8-jar-with-dependencies.jar
Resolving www.congiu.net... 64.90.44.101
Connecting to www.congiu.net|64.90.44.101|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 85492 (83K) [application/java-archive]
Saving to: "json-serde-1.3.8-jar-with-dependencies.jar"

100%[=====>] 85,492          330K/s   in 0.3s

2017-10-26 01:08:08 (330 KB/s) - "json-serde-1.3.8-jar-with-dependencies.jar" saved [85492/85492]
```


TECHNICAL SPECIFICATIONS



- IBM Bluemix BigInsights cluster
- 1 Data Node | vCPU = 4 (24 GB RAM)
- 1 Management Node | vCPU = 12 (48 GB RAM)
- Data Disk - 1 TB SATA
- CPU Speed - 2.30 GHz
- Version - IOP 4.2
- Operating System - CentOS 6.6
- Data Centre - Washington DC

[Write a Review](#)

[Events](#)

[Talk](#)

[Log In](#)


[Sign Up](#)




Find burgers, barbers, spas, handymen...

Near [Current Location](#)



 [Restaurants](#)

 [Nightlife](#)

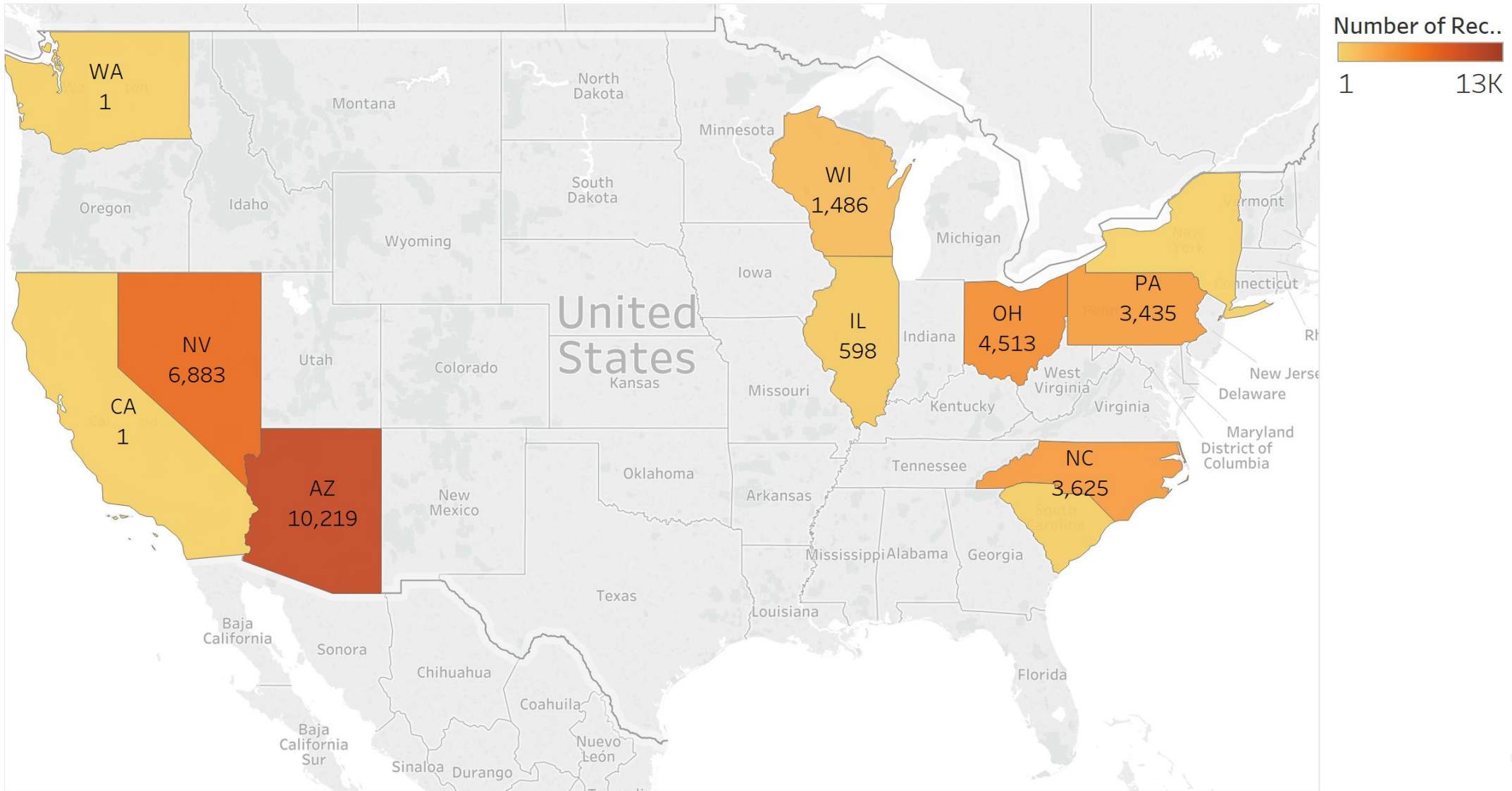
 [Home Services](#) ▾

 [Delivery](#)

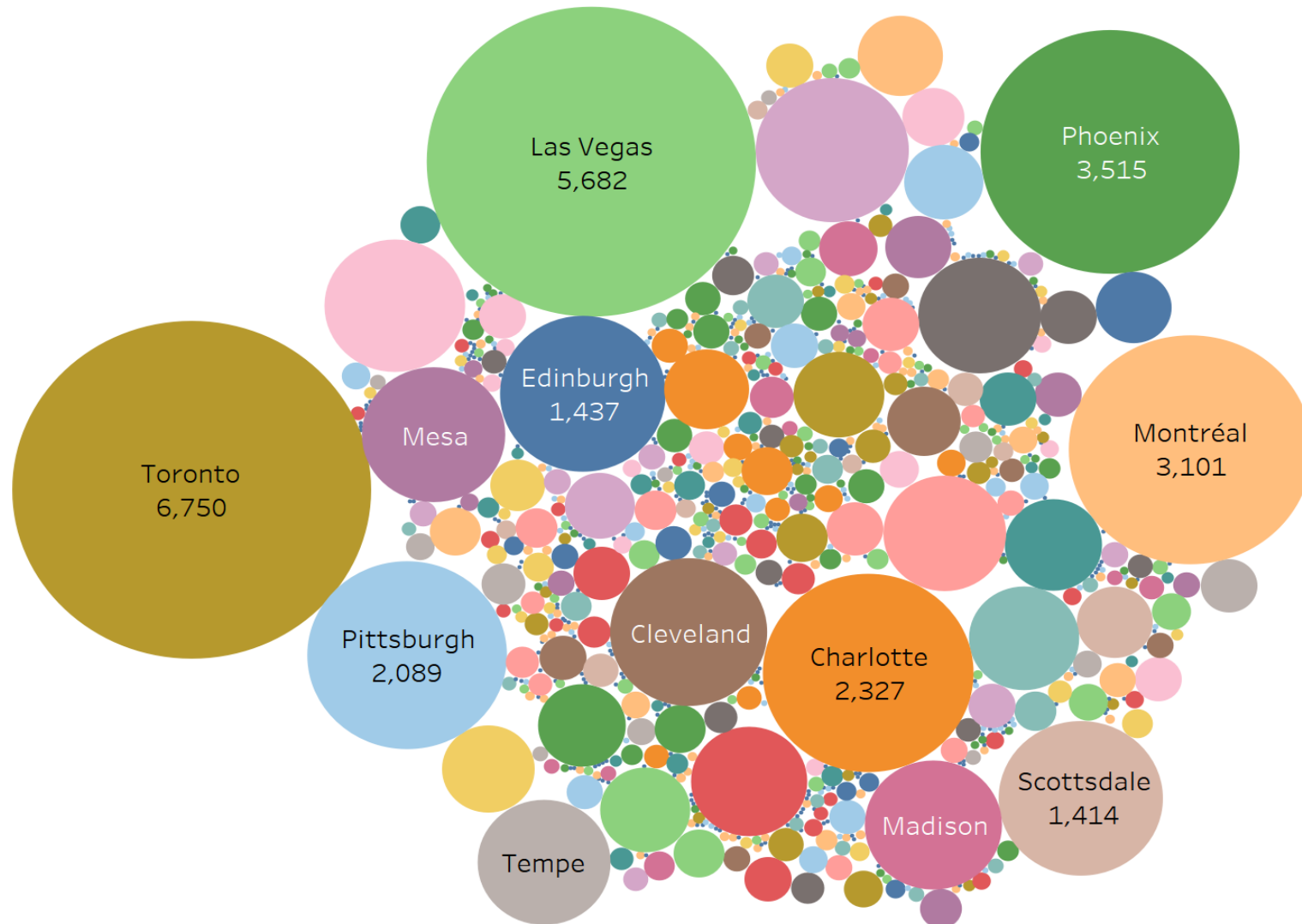
Cheshire
Photo by Natalia K.

VISUALIZATIONS

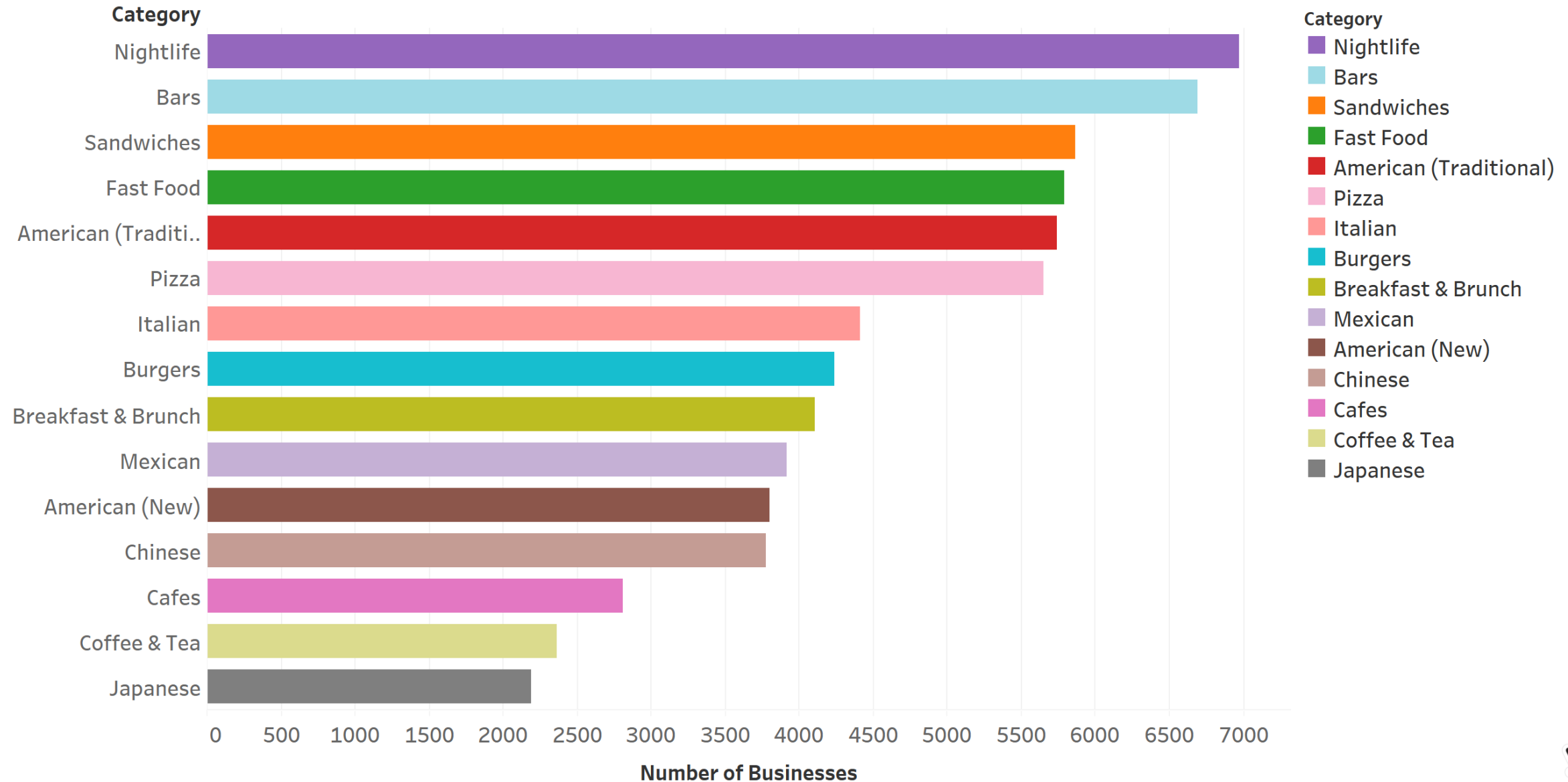
MAP OF RESTAURANTS ACROSS UNITED STATES



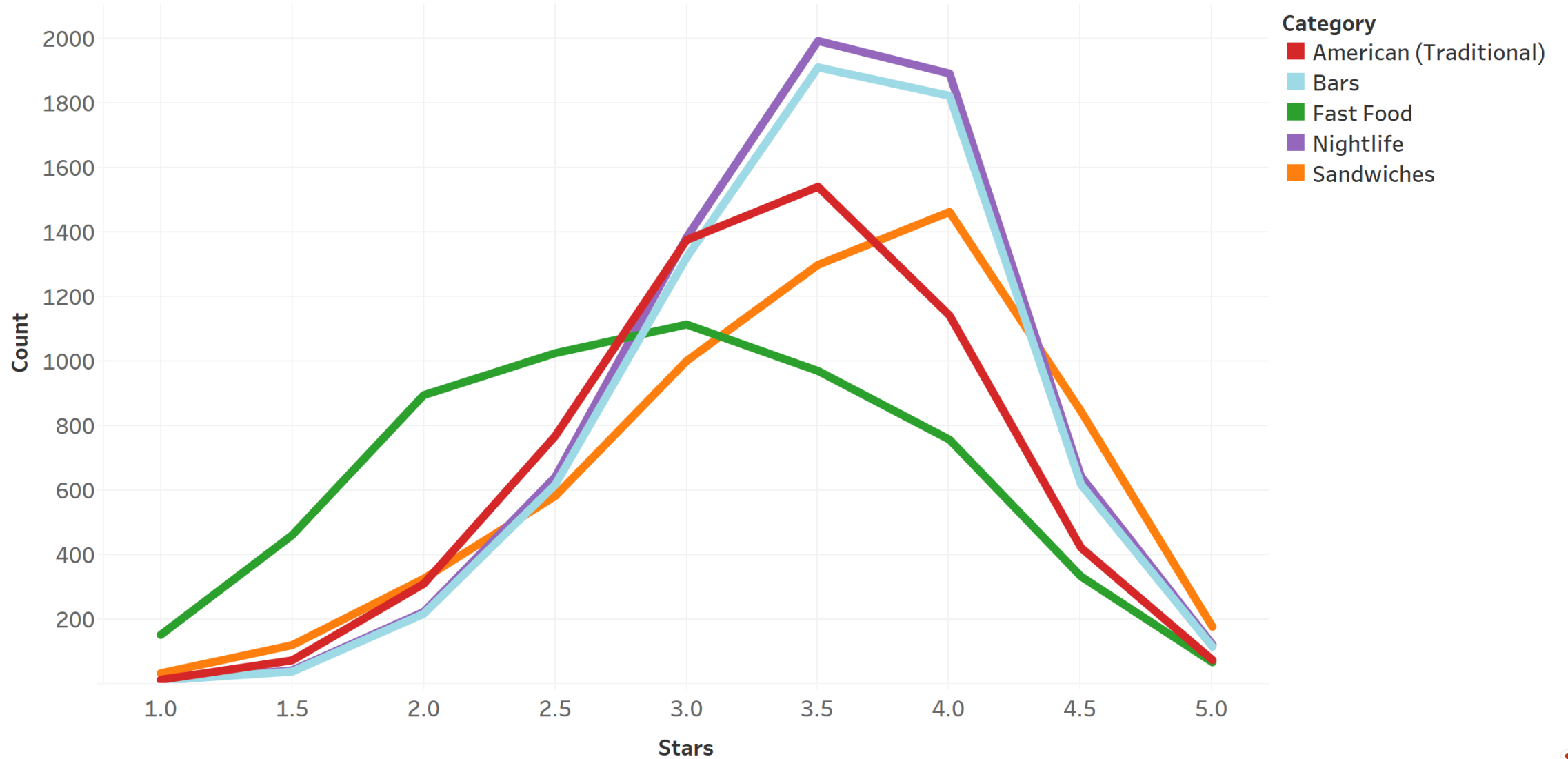
WHICH CITIES HAVE THE HIGHEST NUMBER OF RESTAURANTS?



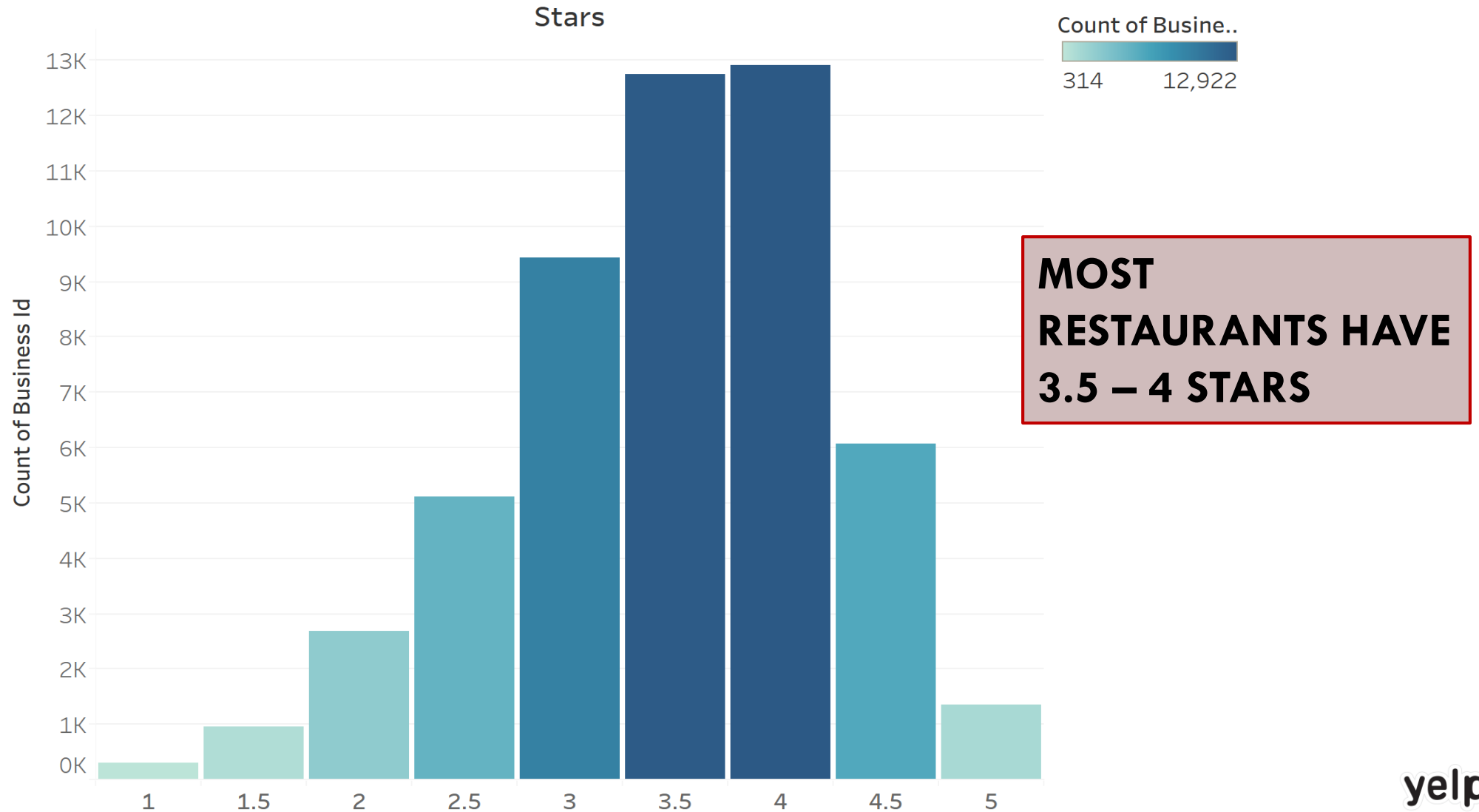
TOP 15 SUB-CATEGORIES OF RESTAURANTS



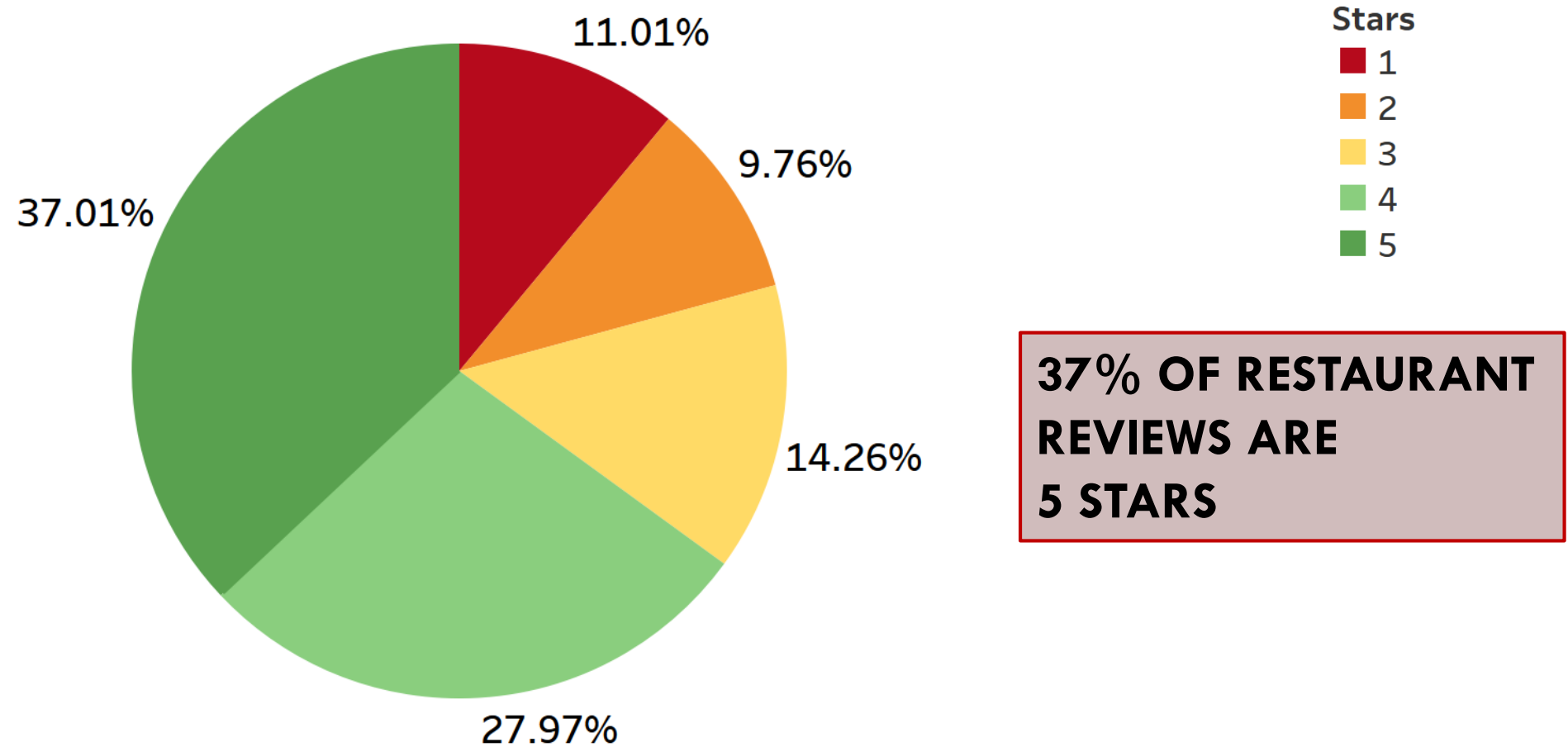
DISTRIBUTION OF RATINGS VS CATEGORIES



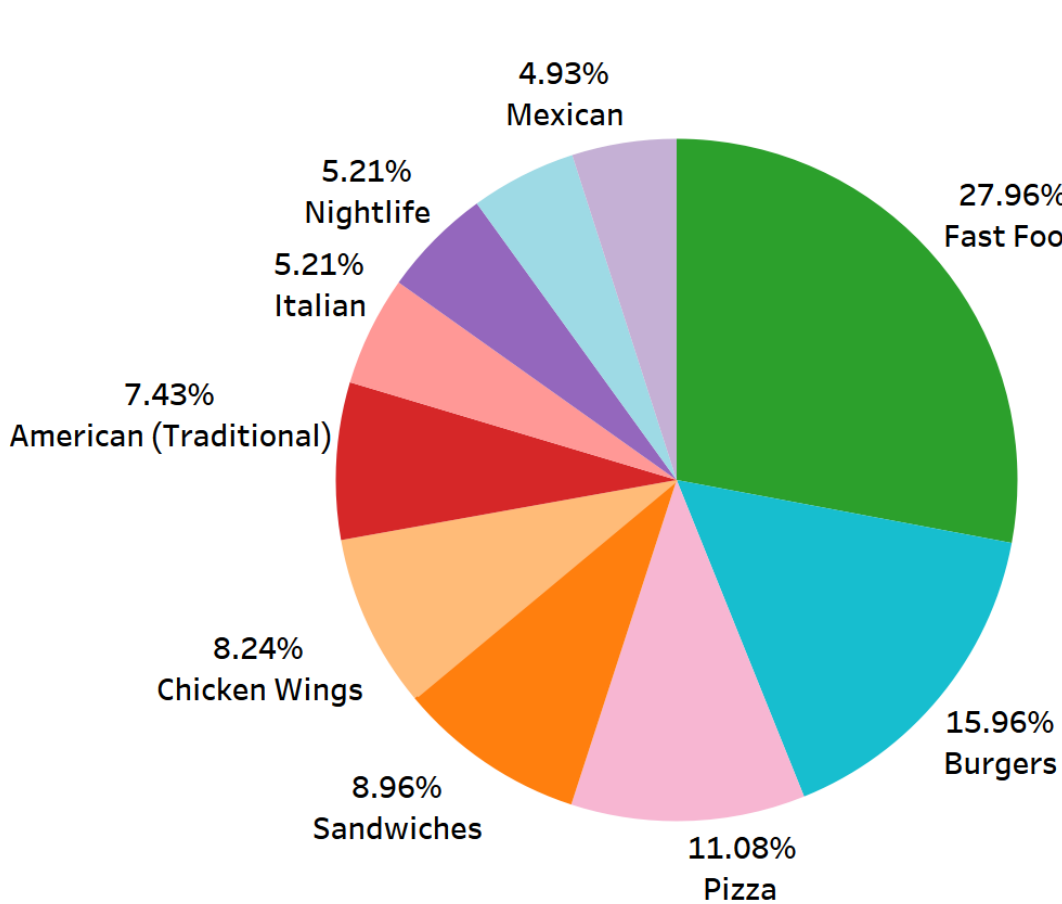
WHAT RATINGS DO THE MAJORITY OF RESTAURANTS HAVE?



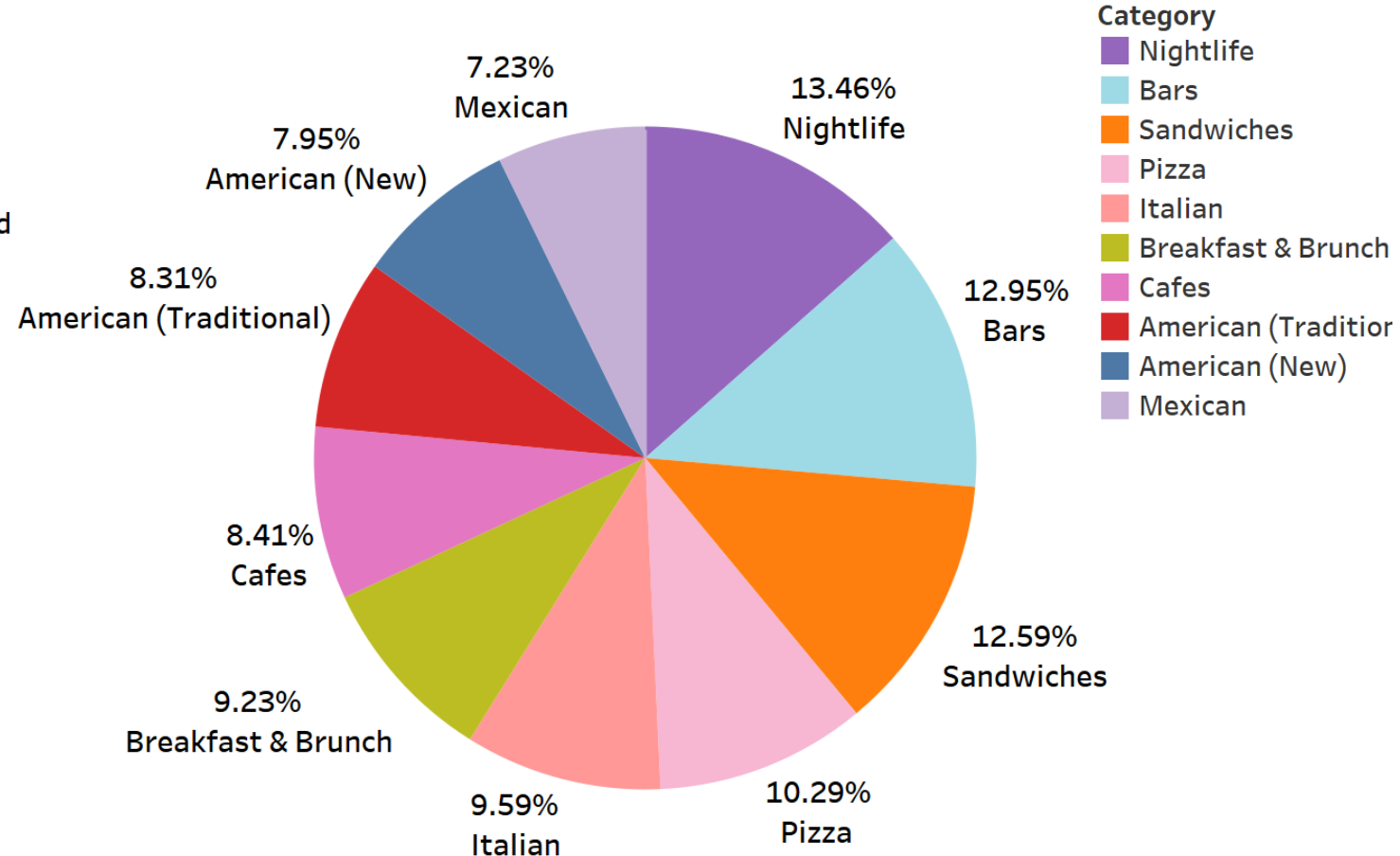
RATING DISTRIBUTION IN RESTAURANT REVIEWS



WHICH RESTAURANTS GET BAD VS GOOD REVIEWS?



Bad reviews (1-2 stars)



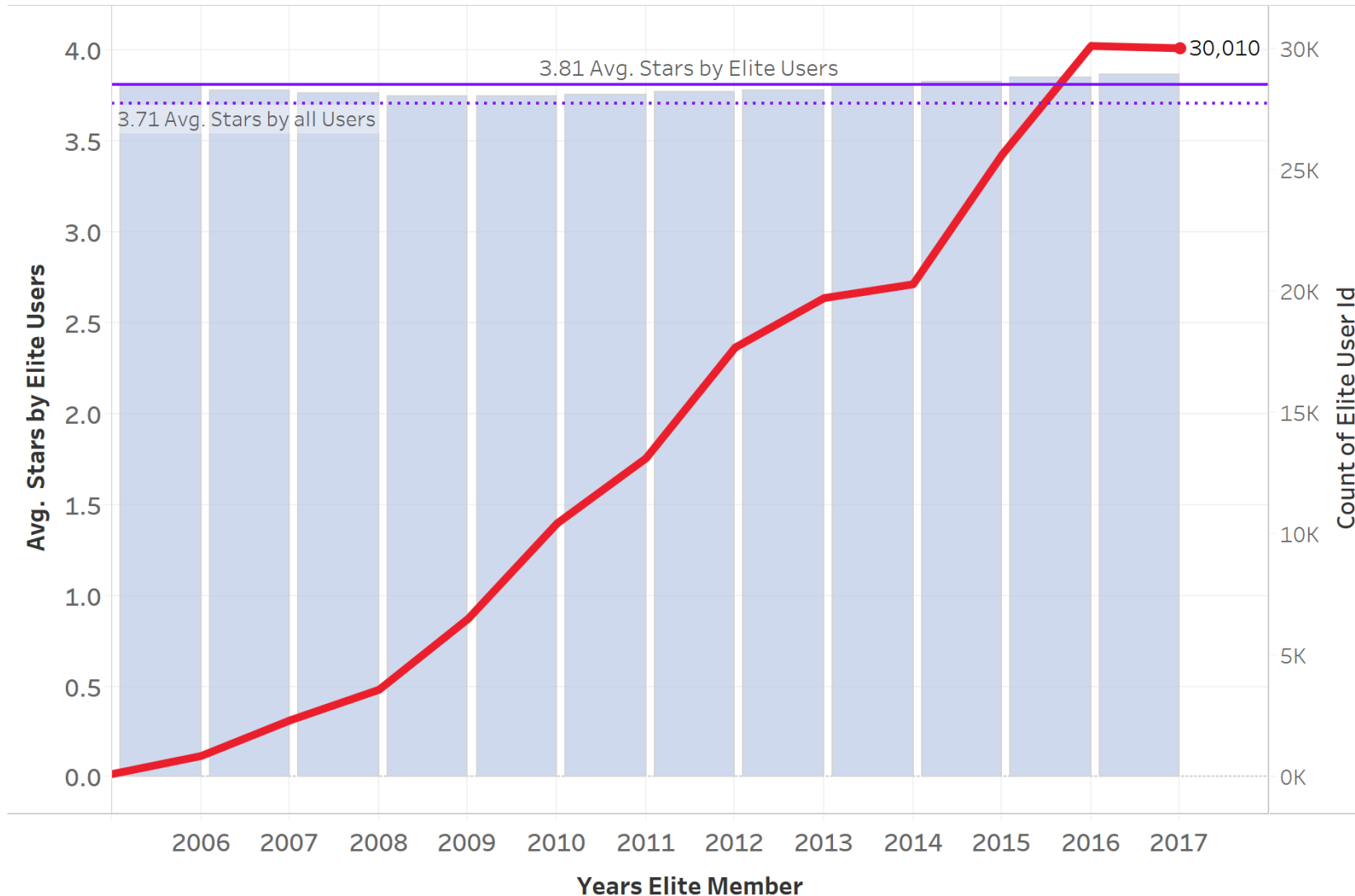
Good reviews (4-5 stars)

WHICH RESTAURANTS HAVE THE MOST REVIEWS?

Name	City	Review Count	Stars	Price \$
Mon Ami Gabi	Las Vegas	6,979	4	2
Bacchanal Buffet	Las Vegas	6,417	4	3
Wicked Spoon	Las Vegas	5,632	4	3
Gordon Ramsay BurGR	Las Vegas	5,429	4	2
Earl of Sandwich	Las Vegas	4,789	5	1
Hash House A Go Go	Las Vegas	4,371	4	2
Serendipity 3	Las Vegas	3,913	3	2
The Buffet	Las Vegas	3,873	4	3
Lotus of Siam	Las Vegas	3,838	4	2
The Buffet at Bellagio	Las Vegas	3,700	4	3
Secret Pizza	Las Vegas	3,542	4	1
Bouchon at the Venezia Tower	Las Vegas	3,439	4	3
MGM Grand Hotel	Las Vegas	3,285	3	2
Gangnam Asian BBQ Dining	Las Vegas	3,180	5	2
Hash House A Go Go 2	Las Vegas	2,963	4	2

RESTAURANTS WITH THE MOST REVIEWS ARE IN LAS VEGAS !!!

WHAT NUMBER OF YELP USERS ARE ELITE? DO THEY RATE DIFFERENTLY THAN NON-ELITE USERS?



**ONLY ~3% OF YELP
USERS ARE ELITE
MEMBERS**

**AVERAGE RATING
IS 3.71 VS 3.81
BY ELITE USERS**

INFERRED INSIGHTS

- Most popular categories of restaurants on Yelp are nightlife, bars, sandwiches and fast food
- Majority of restaurants on Yelp have ratings b/w 3.5 and 4 stars
- 37% of restaurant reviews are 5 stars
- Almost 28% of bad (1-2 star) reviews are for fast food restaurants
- Restaurants with the most reviews are located in Las Vegas
- Only 3% of Yelp users are Elite members (30K out of 1.1M)
- Elite and non-elite users give restaurants very similar star ratings

REFERENCES

- Data set URL <https://www.yelp.com/dataset>
- Serde source <https://github.com/rcongiu/Hive-JSON-Serde>
- IBM Bluemix: <https://console.bluemix.net/data/bic/>
- Tableau <https://www.tableau.com>
- JSON with Tableau
https://onlinehelp.tableau.com/current/pro/desktop/en-us/examples_json.html



QUESTIONS?