Final Project Documentation

Yelp-Database

The Yelp dataset is a subset of our businesses, reviews, and user data for use in personal, educational, and academic purposes. Available as JSON files.[“https://www.yelp.com/dataset”]. This is a big database with more than 6M rows of data so we need to make sure our scripts can handle such requirements.

Scope of the project:

a) The scope of this project is to design and create a clean relational database for the publicly available yelp database which is given in JSON format.
b) Querying this dataset to analyze service based businesses and calculate key metrics to further understand the market.

Example of Yelp Database JSON:

Each file is composed of a single object type, one JSON-object per-line.
Take a look at some examples to get you started: https://github.com/Yelp/dataset-examples.
Note: the follow examples contain inline comments, which are technically not valid JSON. This is done here to simplify the documentation and explaining the structure, the JSON files you download will not contain any comments and will be fully valid JSON.

business.json

Contains business data including location data, attributes, and categories.

```
{
   // string, 22 character unique string business id
   "business_id": "tnhfDv5Il8EaGSXZGiuQGg",

   // string, the business's name
   "name": "Garaje",

   // string, the full address of the business
   "address": "475 3rd St",

   // string, the city
   "city": "San Francisco",

   // string, 2 character state code, if applicable
   "state": "CA",

   // string, the postal code
   "postal code": "94107",

   // float, latitude
   "latitude": 37.7817529521,

   // float, longitude
   "longitude": -122.39612197,
```

```
    // float, star rating, rounded to half-stars
    "stars": 4.5,

    // integer, number of reviews
    "review_count": 1198,

    // integer, 0 or 1 for closed or open, respectively
    "is_open": 1,

    // object, business attributes to values. note: some attribute values might be objects
    "attributes": {
        "RestaurantsTakeOut": true,
        "BusinessParking": {
            "garage": false,
            "street": true,
            "validated": false,
            "lot": false,
            "valet": false
        },
    },

    // an array of strings of business categories
    "categories": [
        "Mexican",
        "Burgers",
        "Gastropubs"
    ],

    // an object of key day to value hours, hours are using a 24hr clock
    "hours": {
        "Monday": "10:00-21:00",
        "Tuesday": "10:00-21:00",
        "Friday": "10:00-21:00",
        "Wednesday": "10:00-21:00",
        "Thursday": "10:00-21:00",
        "Sunday": "11:00-18:00",
        "Saturday": "10:00-21:00"
    }
}
```

review.json

Contains full review text data including the user_id that wrote the review and the business_id the review is written for.

```
{
    // string, 22 character unique review id
    "review_id": "zdSx_SD6obEhz9VrW9uAWA",

    // string, 22 character unique user id, maps to the user in user.json
    "user_id": "Ha3iJu77CxlrFm-vQRs_8g",

    // string, 22 character business id, maps to business in business.json
    "business_id": "tnhfDv5Il8EaGSXZGiuQGg",
```

```
    // integer, star rating
    "stars": 4,

    // string, date formatted YYYY-MM-DD
    "date": "2016-03-09",

    // string, the review itself
    "text": "Great place to hang out after work: the prices are decent, and the ambience is fun. It's a bit loud, but very lively. The staff is friendly, and the food is good. They have a good selection of drinks.",

    // integer, number of useful votes received
    "useful": 0,

    // integer, number of funny votes received
    "funny": 0,

    // integer, number of cool votes received
    "cool": 0
}
```

user.json

User data including the user's friend mapping and all the metadata associated with the user.

```
{
    // string, 22 character unique user id, maps to the user in user.json
    "user_id": "Ha3iJu77CxlrFm-vQRs_8g",

    // string, the user's first name
    "name": "Sebastien",

    // integer, the number of reviews they've written
    "review_count": 56,

    // string, when the user joined Yelp, formatted like YYYY-MM-DD
    "yelping_since": "2011-01-01",

    // array of strings, an array of the user's friend as user_ids
    "friends": [
        "wqoXYLWmpkEH0YvTmHBsJQ",
        "KUXLLiJGrjtSsapmxmpvTA",
        "6e9rJKQC3n0RSKyHLViL-Q"
    ],

    // integer, number of useful votes sent by the user
    "useful": 21,

    // integer, number of funny votes sent by the user
    "funny": 88,

    // integer, number of cool votes sent by the user
    "cool": 15,

    // integer, number of fans the user has
    "fans": 1032,
```

```
    // array of integers, the years the user was elite
    "elite": [
        2012,
        2013
    ],

    // float, average rating of all reviews
    "average_stars": 4.31,

    // integer, number of hot compliments received by the user
    "compliment_hot": 339,

    // integer, number of more compliments received by the user
    "compliment_more": 668,

    // integer, number of profile compliments received by the user
    "compliment_profile": 42,

    // integer, number of cute compliments received by the user
    "compliment_cute": 62,

    // integer, number of list compliments received by the user
    "compliment_list": 37,

    // integer, number of note compliments received by the user
    "compliment_note": 356,

    // integer, number of plain compliments received by the user
    "compliment_plain": 68,

    // integer, number of cool compliments received by the user
    "compliment_cool": 91,

    // integer, number of funny compliments received by the user
    "compliment_funny": 99,

    // integer, number of writer compliments received by the user
    "compliment_writer": 95,

    // integer, number of photo compliments received by the user
    "compliment_photos": 50
}
```

checkin.json

Checkins on a business.

```
{
    // string, 22 character business id, maps to business in business.json
    "business_id": "tnhfDv5Il8EaGSXZGiuQGg"

    // string which is a comma-separated list of timestamps for each checkin, each with format
YYYY-MM-DD HH:MM:SS
    "date": "2016-04-26 19:49:16, 2016-08-30 18:36:57, 2016-10-15 02:45:18, 2016-11-18
01:54:50, 2017-04-20 18:39:06, 2017-05-03 17:58:02"
```

}

tip.json

Tips written by a user on a business. Tips are shorter than reviews and tend to convey quick suggestions.

```
{
   // string, text of the tip
   "text": "Secret menu - fried chicken sando is da bombbbbbb Their zapatos are good too.",

   // string, when the tip was written, formatted like YYYY-MM-DD
   "date": "2013-09-20",

   // integer, how many compliments it has
   "compliment_count": 172,

   // string, 22 character business id, maps to business in business.json
   "business_id": "tnhfDv5Il8EaGSXZGiuQGg",

   // string, 22 character unique user id, maps to the user in user.json
   "user_id": "49JhAJh8vSQ-vM4Aourl0g"
}
```

photo.json

Contains photo data including the caption and classification (one of "food", "drink", "menu", "inside" or "outside").

```
{
   // string, 22 character unique photo id
   "photo_id": "_nN_DhLXkfwEkwPNxne9hw",
   // string, 22 character business id, maps to business in business.json
   "business_id" : "tnhfDv5Il8EaGSXZGiuQGg",
   // string, the photo caption, if any
   "caption" : "carne asada fries",
   // string, the category the photo belongs to, if any
   "label" : "food"
}
```

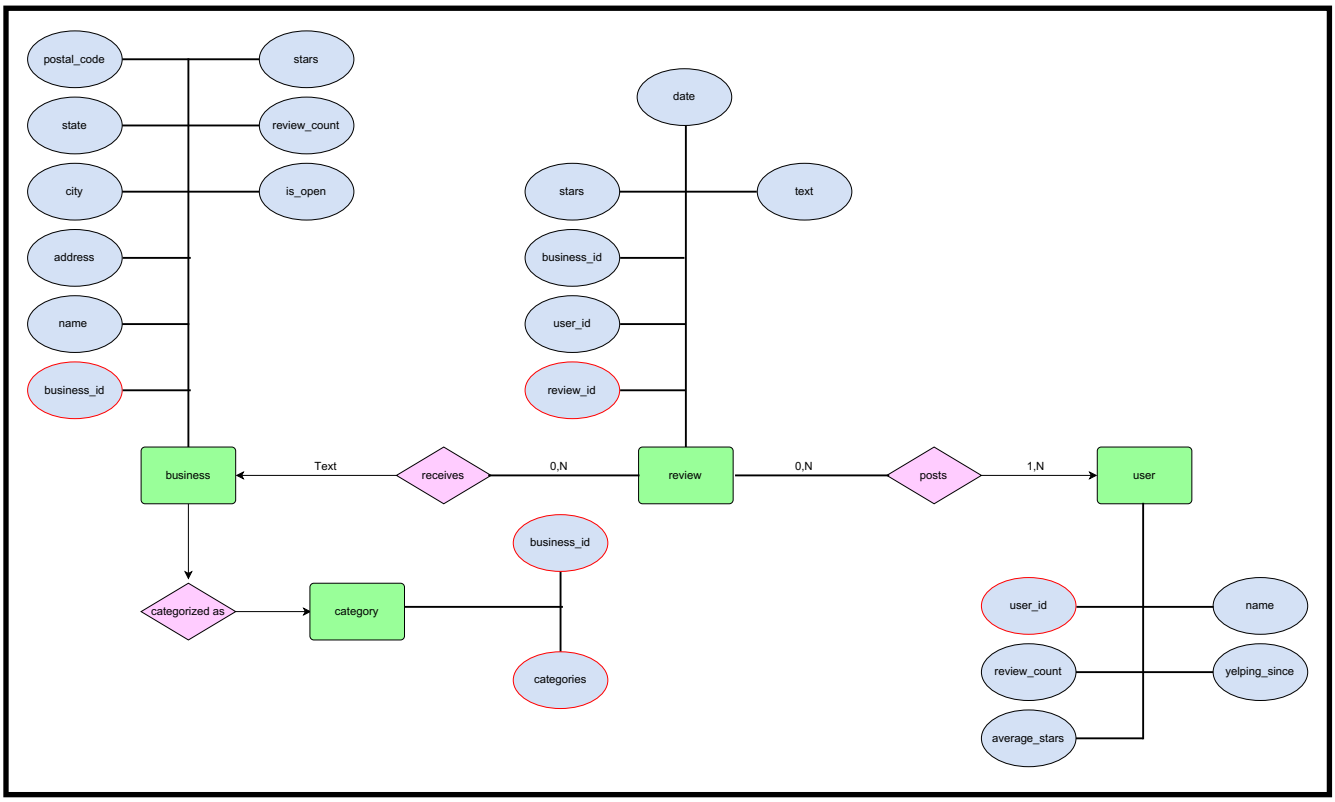
Questions we aim to answer:

1. Name of the user that has posted the most reviews.
2. Number of businesses in each category
3. Average stars of business in the state of FLORIDA in each postal code
4. Category with the most reviews
5. What percentage of existing Home service businesses on yelp that have happy customers ?
6. What is the number of reviews per year in the home services category ?
7. What is the average rate of reviews written per year ?
8. What is the number of happy customers in each category ?
9. How many businesses have shutdown per category?
10. What is the number of businesses per category in MA ?

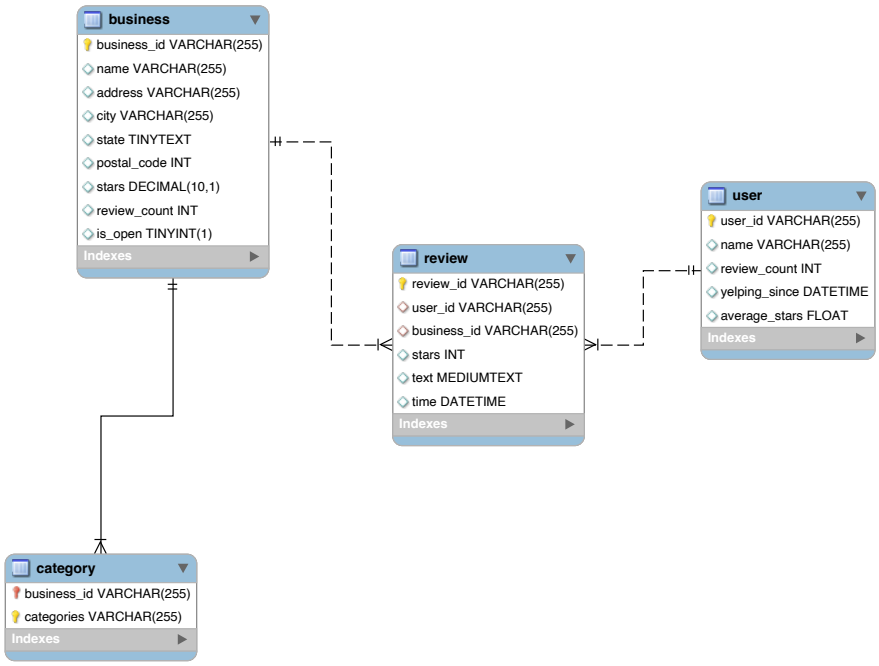
Designing the SQL Database

Keeping in mind the scope of the project and the analysis we aim to perform, we structured the data fulfill these objectives.

Entity-Relationship Diagram



UML Diagram



Note:
As is evident from the ERD and UML the unstructured JSON data available to us is converted into a clean structured form by dropping redundant and irrelevant data. All tables made are in 3rd Normal Form that means:
a) All data is atomic.
b) There are no partial dependencies.
c) There are no transitive dependencies.

SQL Table Creation and Data Insertion

data_insertion.ipynb: script that converts the json dataset into structured data and inserts the data into the designed mysql database.

This way of inserting data takes a lot of time and isn't efficient to handle big datasets such as ours so we dig deeper and optimized the process and reduced the run time significantly.

optimized_data_insertion.ipynb: script that is highly optimized to convert yelp JSON dataset which contains more 6M rows into structured and inserting that data into the database much faster than data_insertion.ipynb.

Output:
business Table

100% 1:1

Result Grid Filter Rows: Search Edit: Export/Import: Fetch rows:

business_id	name	address	city	state	postal_c...	stars	review_count	is_open
__2LkeVwTuKar5AE95ldzw	McDonald's	11007 Bloomingdale Ave	Riverview	FL	33569	1.5	13	1
__4gkf_0UJW78rkRzFm6Gw	Cafe Baladi	235 Lancaster Ave	Malvern	PA	19355	5.0	54	1
__93GWaEEcUf9TSyyJbkmQ	Sunwest Park	17362 Old Dixie Hwy	Hudson	FL	34667	3.5	48	1
__AB6oytCW4hD2Jca02kRA	Chick-fil-A	184 W County Ctr N Spc, Space N	Des Peres	MO	63131	3.5	15	1
__aP-gQsROtMFNWR51oMYA	Wildcat House	1801 N Stone Ave	Tucson	AZ	85705	3.0	13	0
__bDOPMbPcYaC9WsCyAiWQ	Fantastic Sams Cut and Color	1345 Easton Rd	Warrington	PA	18976	4.0	5	1
__bSsYCu2KDn8pb31Epgrg	American Seafoods	3034 George Nick Connor Dr	New Orleans	LA	70119	5.0	5	1
__CYdei4W4pVb4SThJ-HYg	Raider Axe Lodge	5018 N Coolidge Ave, Bldg 2 Ste D	Tampa	FL	33614	5.0	35	0
__Dbc84VjNmVLNgyTUP4iw	Vampire Penguin	10789 Park Blvd	Seminole	FL	33772	4.5	79	1
__dp19BldlU5bGy87LJ0KA	Revibed	1402 N Sheldon St	Indianapolis	IN	46201	5.0	5	1
__FDUGj1N-HrvAmFriy2Vg	Arizona Steel & Ornamental...	3150 E Benson Hwy	Tucson	AZ	85706	3.0	6	1
__I81oH6jmYwh1YXjVWsWA	The Corkscroo Bar & Pizzeria	10 E 9th St	Reno	NV	89501	4.5	14	0
__Jhyl5p8bfnnYOGjwPtsg	Al-Sham Restaurant	201 W Boise Ave	Boise	ID	83706	4.0	19	0
__Ld2XsuPMHjo5uzLF0sXQ	Hand & Stone Massage and...	2500 Grant Ave	Philadelphia	PA	19114	3.0	93	1
__LrkbeqnfmprmB98uLZ6A	Dollar or Two	1305 W Chester Pike	Havertown	PA	19083	2.0	6	1
__NA1bpnfNRvjlQkNo5_HA	Elliston Place Smoke Shop	2204 Elliston Pl, Ste H	Nashville	TN	37203	3.5	15	1
__q2lqxbh1tnug8xyzhYLg	E.Sierra Media	1855 Sullivan Ln, Ste 215,	Sparks	NV	89431	4.0	5	1
__qax_vxKX9FjNNyye9z9A	Fast Freddie's Diner	624 W Main St	Henderson...	TN	37075	4.0	6	0
__QqLzVxwrrdBuc-5evfqA	Windermere Animal Hospital	9785 Olympia Dr	Fishers	IN	46037	4.0	13	1

100% 1:1

Result Grid Filter Rows: Search Edit: Export/Import:

business_id	categories
___UdvaxCnwsQ7nA1eKZAQ	Hotels & Travel
___UdvaxCnwsQ7nA1eKZAQ	Transportation
___UdvaxCnwsQ7nA1eKZAQ	Buses
__2LkeVwTuKar5AE95Idzw	Burgers
__2LkeVwTuKar5AE95Idzw	Coffee & Tea
__2LkeVwTuKar5AE95Idzw	Food
__2LkeVwTuKar5AE95Idzw	Restaurants
__2LkeVwTuKar5AE95Idzw	Fast Food
__4gkf_0UJW78rkRzFm6Gw	Halal
__4gkf_0UJW78rkRzFm6Gw	Lebanese
__4gkf_0UJW78rkRzFm6Gw	Middle Eastern
__4gkf_0UJW78rkRzFm6Gw	Restaurants
__4gkf_0UJW78rkRzFm6Gw	Mediterranean
__93GWaEEcUf9TSyyJbkmQ	Active Life
__93GWaEEcUf9TSyyJbkmQ	Beaches
__93GWaEEcUf9TSyyJbkmQ	Parks
__93GWaEEcUf9TSyyJbkmQ	Rafting/Kayak...

user_id	name	review_count	yelping_since	average_stars
___6aix-XvFcQz3GauAPpw	John	7	2011-09-26 18:10:05	2.75
___9Jl-8aF7z58JdTWlyjw	Erin	10	2011-04-23 23:47:24	3.9
___aO4EhZULBsJPvJ4rMhA	David	14	2010-12-04 22:21:55	1.69
___DjEwDb9e7Ny-NOiezZw	Patama	10	2013-10-28 19:29:04	4.8
___hQj63mwgFDhehNwv3ZQ	Taleena	11	2014-10-25 21:26:56	4.09
___I9ZYdYGkZ6dMYxwJEIQ	Jim	239	2011-08-17 17:35:52	3.99
___Pul-lcE0y9hjAer8GrQ	DeAnna	1	2015-02-15 15:49:36	1
___tX0MgAQYPaWssEjSxKw	Mary	2	2016-02-22 00:12:38	3
__--6h87PZrkaT0SAuQc-w	Jiban	2	2016-10-03 03:57:50	5
__-2EyiraNzLPrq6o1QDIA	Dave	2	2018-07-06 17:09:08	4.5
__-6iHRGqL_K9oM9KNs6pg	Morgan	22	2014-12-21 19:58:02	2.39
__-8WEg7xD0CwCDu04MzBA	Maria	39	2010-05-03 16:50:07	4
__-Kt26YrtJxGdWs8FqKCg	Shane	11	2012-08-06 04:44:19	4
__-LqRflgxlTHOxsGMBt_Q	Craig	9	2012-11-20 22:57:48	3.67
__-uwr6nLywqDa8d_QD4-A	Rose...	18	2010-03-28 14:46:49	3.5
__-w2mUbDTlC6u1lklEWwg	Zach	13	2012-12-24 04:54:21	4.46
__-WWzo7OEEd01pOA-rhPw	Ron	8	2011-11-21 13:53:41	1
__-YOsZp7ilfYVwD8Wdszg	David	2	2014-10-03 14:11:46	3
__-zs2o-TGiRzK4Wqmgl2A	John	2	2017-04-12 04:12:03	1
__05rytNjsye9MBhqB0DMA	Marck	1247	2006-02-12 20:20:32	3.65
__0cufkRcZf12oTT04xGLg	Mat	36	2014-10-31 20:18:36	3.65
__0D94KGQl7dBCcA2MmH...	Chris	48	2012-07-13 11:32:20	3.5
__0TajwzDW-qJ2301ooWUg	B	11	2015-02-25 21:07:52	3.17

100% 1:1

Result Grid Filter Rows: Search Edit: Export/Import: Fetch rows:

review_id	user_id	business_id	stars	text	time
___7mnNcwBXBN76xXXsYyw	gexqRmOpGM7S3XMpx1LySA	l2YZuyG1KjiTbtwfNFHaOQ	5	Such great food and service ..not to mention AM…	2019-08-23 19:27:18
___881evFAzgFZxMEvHv-w	Nb0t9ikn5Y14d_ZtGyfF7w	YqSW6uIkrF9a2h6iLgRw2w	5	Excellent. Juicy, well-seasoned salmon. Even th…	2018-04-08 17:40:51
___B0ROhOE7Y_krltls_bQ	1g_7tmVotExGY4Q_1mY6uA	9n-1LQLX3ntBfBtMwgSpig	5	We are originally from northern New Jersey acr…	2012-02-08 16:15:55
___ELF3-3EiCmiWHBQ5Wvw	-kef2nlu6Chnf5drLRO7Sw	89SD5fNDDnJj-lTB40hLsQ	5	When I first entered King and I, I wondered, "wh…	2011-06-30 21:26:36
___eRprRS06mfmhTrpD1MA	YQnL9LCvc0-dTTqAH2qw1Q	lwCJ4KkAsRymBJJT3KCauQ	4	Have eaten here twice for lunch. Food is tasty t…	2015-01-26 17:23:53
___fHoCxKRTMm6mJkJ3u6A	qjkOmSEFFEeWVTtsIl0aAw	PwiFfUAwzDXbMGt0OnrB-g	4	Delicious. Harvest quite possibly has the best f…	2006-11-14 00:40:00
___GE2uU_3CmbvjX5yfwgg	3uvDTPwqoh8gQ3yVYvs_FA	qujpfyCm_e2l05MPP9zO4A	5	Amazing experience here. Will definitely buy mo…	2015-10-11 03:45:52
___gu3aPXpLzQ-l8b1680Q	Z90K8RRx5jgAtCHvzckJFA	Uky0DD3LU4C7eyNDhpmOXg	4	I love brunch at Day By Day :) We go all the tim…	2010-10-19 04:00:42
___hjNQv6lBHfndwG4Z4Ug	p4dved0OUoh-lKaXSTdCXA	WXxlJiG1Qlae0Dasbf-fzA	3	Drunchies from the friends required me to make…	2017-08-31 18:19:30
___ijx4svckVgMCp873dTQ	EGrX9p9bD1s8z1uj1L_vDQ	hzqRzJ6AuiZl_WbOXvpTmA	1	I should have taken Matthew's comment more s…	2014-07-01 22:56:16
___ld1ela-jAtpmCCNXwgg	qtdV6LwNXsCbOTa7lC4P6w	ZYXmUzLH8Ie21j3e42bUfw	1	Cara was at an event the ended at a certain tim…	2016-09-19 17:15:12
___q88Zr8oXrTXder9gQ5A	ctjrE57NKQM2Ba1HAKPdOQ	3iUCCf1FWmjlFbGYvBgf9w	5	This post is a year over due so in the summer of…	2020-08-20 23:25:16
___qbda8FYlGlXUSWcproA	J-TaT86tsYzz39KpeysgtQ	u_6LhNGXBt5bYIn-ue70Ag	5	After buying our home in South Philly, we broug…	2018-04-14 10:28:44
___QLf4ph-5Eq506Ypoqcg	uD4ssusutbH1PgnadpZsSQ	3a10MqoYLLUH8B1uEm02Pw	5	I'm a fan for life after what happened with an or…	2015-12-28 19:34:37
___RvufxpVXgbg7zMEHKRA	3lkCuUlrqnCl19jiKBz5LA	aFspsMR_MbxaOOS58YUGOA	2	I eat at Wasabi frequently for dinner... though it'…	2012-06-15 19:40:28
___S17UEpEbotmf9bsbl1Q	JE472SCEPg4H1p6dNLwYpA	Bl6QNjzmtG779fI5Hm4gfA	5	The BEST Mexican food around, and prices. Ev…	2021-07-15 14:45:10
___WGp7dCQpco8nY0OY0eg	YyLL5zQVNl_wbHkdrd9GMQ	l4leGBHXV2E8S8COd9rx9A	4	Great food and drinks and atmosphere. Had br…	2021-02-26 19:11:05
___wXTHo9ak3AFYcR1_w8w	Sk1utVLqifxGBYpQRJhALw	jrSC1bwN2lz9dVBrNTejSg	5	Well Jack I'm in love with you once again! I'm n…	2018-11-25 20:29:06

Use Cases

use-cases.sql : SQL file that contains 5 use cases and its SQL statements

Example output:

Limit to 1000 rows

```sql
-- use case 3: number of businesses in each state by category
SELECT
categories,
state,
COUNT(business.business_id) AS num_business
FROM
business
JOIN category ON business.business_id = category.business_id
GROUP BY categories,state
ORDER BY num_business desc;
```

100% 28:30

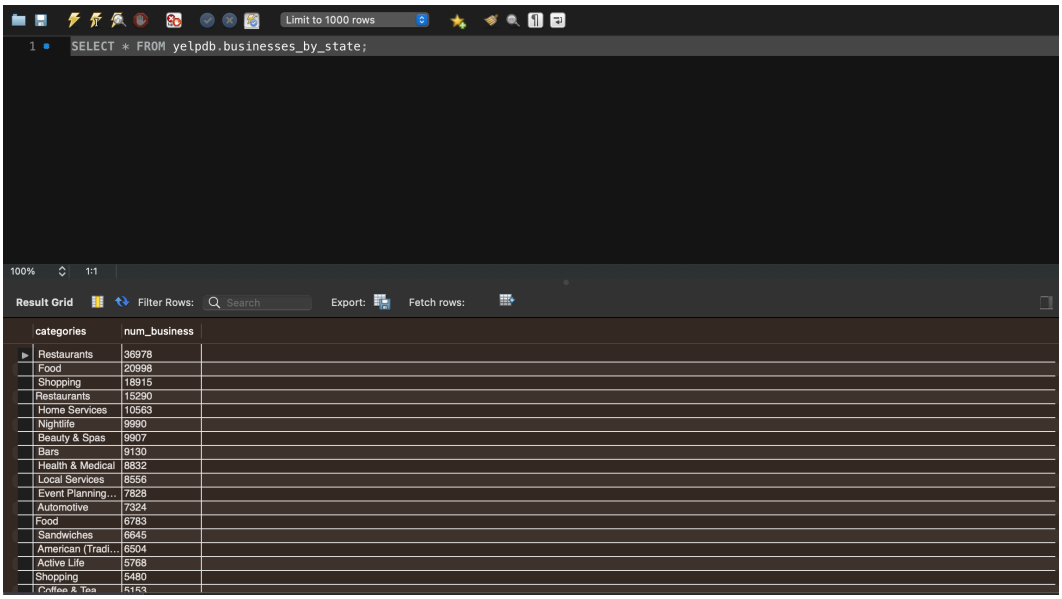
Result Grid Filter Rows: Search Export: Fetch rows:

categories	num_business
Restaurants	36978
Food	20998
Shopping	18915
Restaurants	15290
Home Services	10563
Nightlife	9990
Beauty & Spas	9907
Bars	9130
Health & Medical	8832
Local Services	8556
Event Planning…	7828
Automotive	7324
Food	6783
Sandwiches	6645
American (Tradi…	6504
Active Life	5768
Shopping	5480
Coffee & Tea	5153

Views

model_views.sql: This file creates views for all the use cases for our database.

Example Output:



categories	num_business
Restaurants	36978
Food	20998
Shopping	18915
Restaurants	15290
Home Services	10563
Nightlife	9990
Beauty & Spas	9907
Bars	9130
Health & Medical	8832
Local Services	8556
Event Planning...	7828
Automotive	7324
Food	6783
Sandwiches	6645
American (Tradi...	6504
Active Life	5768
Shopping	5480
Coffee & Tea	5153

Caption