ABH Internship report

Data Analytics: Month 1

After flattening the original business.json file, 100 attributes were generated. They are divided into 3 categories. The significant attributes are identity and address attributes since they are providing essential information about each business. All additional attributes are in category “Richness”.

**CATEGORIES OF ATTRIBUTES**

* **Identity attributes**

'name', 'business\_id', categories

* **Address attributes**

'address', 'city', 'postal\_code', 'state', 'latitude',

'longitude'

* **Richness attributes**

'attributes.AcceptsInsurance', 'attributes.AgesAllowed', 'attributes.Alcohol','attributes.Ambience',

'attributes.Ambience.casual', 'attributes.Ambience.classy',

'attributes.Ambience.divey', 'attributes.Ambience.hipster',

'attributes.Ambience.intimate', 'attributes.Ambience.romantic',

'attributes.Ambience.touristy', 'attributes.Ambience.trendy',

'attributes.Ambience.upscale', 'attributes.BYOB',

'attributes.BYOBCorkage', 'attributes.BestNights',

'attributes.BestNights.friday', 'attributes.BestNights.monday',

'attributes.BestNights.saturday', 'attributes.BestNights.sunday',

'attributes.BestNights.thursday', 'attributes.BestNights.tuesday',

'attributes.BestNights.wednesday', 'attributes.BikeParking',

'attributes.BusinessAcceptsBitcoin', 'attributes.BusinessAcceptsCreditCards',

'attributes.BusinessParking', 'attributes.BusinessParking.garage',

'attributes.BusinessParking.lot',

'attributes.BusinessParking.street', 'attributes.BusinessParking.valet',

'attributes.BusinessParking.validated', 'attributes.ByAppointmentOnly', 'attributes.Caters',

'attributes.CoatCheck', 'attributes.Corkage', 'attributes.DietaryRestrictions',

'attributes.DietaryRestrictions.dairy-free','attributes.DietaryRestrictions.gluten-free',

'attributes.DietaryRestrictions.halal', 'attributes.DietaryRestrictions.kosher',

'attributes.DietaryRestrictions.soy-free', 'attributes.DietaryRestrictions.vegan',

'attributes.DietaryRestrictions.vegetarian', 'attributes.DogsAllowed', 'attributes.DriveThru',

'attributes.GoodForDancing', 'attributes.GoodForKids',

'attributes.GoodForMeal', 'attributes.GoodForMeal.breakfast',

'attributes.GoodForMeal.brunch', 'attributes.GoodForMeal.dessert',

'attributes.GoodForMeal.dinner', 'attributes.GoodForMeal.latenight',

'attributes.GoodForMeal.lunch', 'attributes.HairSpecializesIn',

'attributes.HairSpecializesIn.africanamerican', 'attributes.HairSpecializesIn.asian',

'attributes.HairSpecializesIn.coloring',

'attributes.HairSpecializesIn.curly', 'attributes.HairSpecializesIn.extensions',

'attributes.HairSpecializesIn.kids', 'attributes.HairSpecializesIn.perms',

'attributes.HairSpecializesIn.straightperms', 'attributes.HappyHour', 'attributes.HasTV',

'attributes.Music', 'attributes.Music.background\_music', 'attributes.Music.dj',

'attributes.Music.jukebox', 'attributes.Music.karaoke',

'attributes.Music.live', 'attributes.Music.no\_music',

'attributes.Music.video', 'attributes.NoiseLevel',

'attributes.Open24Hours', 'attributes.OutdoorSeating',

'attributes.RestaurantsAttire', 'attributes.RestaurantsCounterService',

'attributes.RestaurantsDelivery', 'attributes.RestaurantsGoodForGroups',

'attributes.RestaurantsPriceRange2', 'attributes.RestaurantsReservations',

'attributes.RestaurantsTableService', 'attributes.RestaurantsTakeOut', 'attributes.Smoking',

'attributes.WheelchairAccessible', 'attributes.WiFi', 'hours.Friday',

'hours.Monday', 'hours.Saturday', 'hours.Sunday', 'hours.Thursday',

'hours.Tuesday', 'hours.Wednesday', 'is\_open' , 'review\_count', 'stars’

Task 1ANALYSIS OF SIGNIFICANT ATTRIBUTES

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Business\_id | name | address | state | Postal\_code | city | latitude | longitude | categories |
| Completeness | 100% | 100% | 96% | 100% | 99.65% | 99.99% | 100% | 100% | 99.74% |
| Empty values | 0% | 0% | 4% | 0% | 0.35% | 0.01% | 0% | 0% | 0.26% |
| Dummy values | - | - | - | <1% | - | - | - | - | - |

Frequency of column “categories”

|  |  |
| --- | --- |
| **Category** | **Frequency** (% of total not null values) |
| "Restaurants, Pizza" | 0.54 |
| "Nail Salons, Beauty & Spas" | 0.53 |
| "Pizza, Restaurants" | 0.51 |
| "Beauty & Spas, Nail Salons" | 0.49 |
| "Food, Coffee & Tea" | 0.46 |
| "Mexican, Restaurants" | 0.46 |
| "Coffee & Tea, Food" | 0.45 |
| "Restaurants, Mexican" | 0.44 |
| "Chinese, Restaurants" | 0.43 |
| "Hair Salons, Beauty & Spas" | 0.43 |

Frequency of column “state”

|  |  |
| --- | --- |
| **State** | **Frequency** (% of total not null values) |
| "AZ" | 29.43 |
| "NV" | 18.85 |
| "ON" | 17.34 |
| "NC" | 7.64 |
| "OH" | 7.63 |

Frequency of column “postal\_code”

|  |  |
| --- | --- |
| **Postal\_code** | **Frequency** (% of total not null values) |
| "89109" | 1.66 |
| "85251" | 1.12 |
| "85260" | 0.98 |
| "85281" | 0.98 |
| "89119" | 0.97 |

Frequency of column “city”

|  |  |
| --- | --- |
| **city** | **Frequency** (% of total not null values) |
| "Las Vegas" | 15.24 |
| "Toronto" | 9.81 |
| "Phoenix" | 9.74 |
| "Charlotte" | 4.93 |
| "Scottsdale" | 4.58 |
| "Calgary" | 4.01 |
| "Pittsburgh" | 3.64 |
| "Montréal" | 3.34 |
| "Mesa" | 3.15 |
| "Henderson" | 2.53 |

Frequency of column “latitude”

|  |  |
| --- | --- |
| **latitude** | **Frequency** (% of total not null values) |
| "36.1697096" | 0.08 |
| "36.128561" | 0.06 |
| "33.4483771" | 0.06 |
| "36.175" | 0.04 |
| "33.4986286" | 0.04 |

Frequency of column “longitude”

|  |  |
| --- | --- |
| **longitude** | **Frequency** (% of total not null values) |
| "-115.1236952" | 0.0872 |
| "-111.9403254" | 0.0867 |
| "-115.1711298" | 0.082 |
| "-111.8210866" | 0.0815 |
| "-115.2244853" | 0.0809 |

Frequency of column “address”

|  |  |
| --- | --- |
| **Address** | **Frequency** (% of total not null values) |
| null | 4.15 |
| "5757 Wayne Newton Blvd" | 0.05 |
| "7014 E Camelback Rd" | 0.04 |
| "3200 Las Vegas Blvd S" | 0.04 |
| "3111 W Chandler Blvd" | 0.04 |

Frequency of column “stars”

|  |  |
| --- | --- |
| **stars** | **Frequency** (% of total not null values) |
| 4 | 18.67 |
| 3.5 | 18.17 |
| 5 | 14.64 |
| 4.5 | 14.17 |
| 3 | 13.49 |
| 2.5 | 9.78 |
| 2 | 5.93 |
| 1.5 | 2.58 |
| 1 | 2.53 |

Duplicates

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **business\_id** | **name** | **address** | **city** | **latitude** | **longitude** |
| "ZJKvmjBinGXEagy7l1na7Q" | "AMC Ahwatukee 24" | "4915 East Ray Rd" | "Phoenix" | "33.3167894" | -111.9773681 |
| "UYZ\_LzFRuWTv\_\_VsXUXhsg" | "AMC Ahwatukee 24" | "4915 East Ray Rd" | "Phoenix" | "33.3167894" | -111.9773681 |
| "f4mh1Y0rnvbJRfQ3jPkqzQ" | "AMC Concord Mills 24" | "8421 Concord Mills Blvd" | "Concord" | "35.3705892" | -80.7261556 |
| "Oj5l\_BG9S2H-RSujDjmaTQ" | "AMC Concord Mills 24" | "8421 Concord Mills Blvd" | "Concord" | "35.3705892" | -80.7261556 |
| "LjwVx0P37xXSNZaVUQkqxQ" | "AMC Mesa Grand 14" | "1647 S Stapley" | "Mesa" | "33.3836435" | -111.8029423 |
| "sR0GYpPKs9CzixSLei03Vw" | "AMC Mesa Grand 14" | "1647 S Stapley" | "Mesa" | "33.3836435" | -111.8029423 |
| "0XMOCF8omuTYfhkNRz1xMw" | "Arby"s" | "119 E Baseline Rd" | "Phoenix" | "33.3773552" | -112.0708618 |
| "x57\_qCiL58xcFOzvwqbWpQ" | "Arby"s" | "119 E Baseline Rd" | "Phoenix" | "33.3773552" | -112.0708618 |
| "-XI7MQo\_EVhQ8EtmhIn-BA" | "Bank of America" | "4741 E Ray Rd" | "Phoenix" | "33.3192875" | -111.9809869 |
| "1uU4Y-NZPjORVj1\_\_VI5hQ" | "Bank of America" | "4741 E Ray Rd" | "Chandler" | "33.3192875" | -111.9809869 |
| "7Fi61ncAOLfqcdWjj24w4Q" | "Bank of America" | "9260 W Northern Ave" | "Peoria" | "33.5521655" | -112.2590326 |
| "h9B-VivFS\_8BRX7MKF1bJw" | "Bank of America" | "9260 W Northern Ave" | "Glendale" | "33.5521655" | -112.2590326 |

-less than 1% duplicate data

Task 2 and Task 3

Range for latitude is -90:90 and for longitude is: -180:180

* All latitudes and longitudes are in proper ranges

States

* 36 distinct states
* Some states were suspicious but were actually from Canada
* "XGM", "XWY", "DUR", "CON", "DOW", "XGL", "BAS" : invalid states

|  |  |  |  |
| --- | --- | --- | --- |
| **address** | **postal\_code** | **city** | **state** |
| "759A The Queensway" | "M8Z 1N1" | "Toronto" | "OH" |
| "3625 Shaganappi Trail NW, Suite Z011" | "T3A 5C4" | "Calgary" | "AL" |

State does not match other data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **state** | **city** | **postal\_code** | **latitude** | **longitude** |
| "XWY" | "Leeds" | "LS27 8EB" | "43.6528212" | "-79.3763454" |
| "CON" | "Church Cove" | "TR12 6SX" | "35.532021" | "-80.8516825" |
| “NC” | “Charlotte” | "282010" | "35.1531965825" | "-80.8382177353" |
| “NV” | “Las Vegas” | "891118" | "36.0810279" | "-115.2419005" |

Postal\_code and address belong to Leeds, UK but coordinates are in Toronto, Canada ("Desi Masala")

Postal\_code and address belong to Helston, UK but coordinates are in North Carolina, USA ("The Old Lifeboat House")

6-digit postal\_code in US city Charlotte (should be 5-digit)

6-digit postal\_code in US city Las Vegas (should be 5-digit)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **name** | **address** | **city** | **Postal\_code** | **state** |
| "Walker"s Fish Market" | "6531 Mississauga Road" | "Mississauga" | "L5N" | "ON" |
| "Paprika House" | "555 11 Avenue SW, Suite 125" | "Calgary" | "T2R" | "AB" |
| "Sushi House" | "135 Main Street N" | "Brampton" | "L6X" | "ON" |
| "Zmaya" | "7880 Chemin de la Côte de Liesse" | "Saint-Laurent" | "H4T" | "QC" |

Unstandardized but valid records of postal\_code (3-character long instead of 6 characters) in Canada

Business\_id

* All entries are of length 22 and unique as required

Name

* Names starting with a number: "640 Toronto", "702 Pools", "9920 Apartments", "33 1/3 Record Store", "808 Sushi"
* 13 names consisted only of numbers: "3734", "217", "1602", "935", "76” (usually coming from house number in address)-check on map
* Bad words: "The Mole Hole", "Hole "n One", "Art Hole Noda", "DICK"S Sporting Goods", "Dick"s Diner" (<1%, 300 records)

Address

* 888 addresses without house number: "Phoenix Premium Outlets", "Métro Guy-Concordia", "Hawthorn Pkwy", "Millers Run Rd", "Grandview Ave", "University Center Ii Lot", "Carlton St"
* "RR 910", "1340 E", "9:00 & L", "PO 97622", "I-279"

City

* Suspicious cities: London, Manchester, Leeds (fit coordinates in North America, not in Europe)
* Suspicious: "Rocky View No. 44" and "Division No. 6"-invalid for city
* "Las Vegas," or "North Olmsted," or "Harrisburg," -comma not needed
* "Urbana, Illinois" or "Markham, ON" or "Scottsdale, AZ" or "Las Vegas, NV"-state included in city
* "110 Las Vegas"-number included
* Total unstandardized data<1%

Hours

* -maybe unnecessary to divide working hours by day, might be one string Mo-Fri, Sat…

Postal\_code

* Two types: US and Canadian formats (*A1A 1A1* format)
* US: 5 digits
* Canada: 6 characters (alphanumeric separated by space between third and fourth char)
* Unstandardized postal codes <1%

Categories

* 99.74% completeness
* 0.26% empty values
* "93385" distinct in db but after preprocessing only 1300 appeared
* Examples of unstandardized categories before processing: "Restaurants, Pizza" and "Pizza, Restaurants", "Nail Salons, Beauty & Spas" and "Beauty & Spas, Nail Salons", "Automotive, Auto Repair" and "Auto Repair, Automotive", "Restaurants, Chinese" and "Chinese, Restaurants", "Food, Grocery" and "Grocery, Food" (the same category but reversed order of words)
* Examples of standardized categories: fast food, coffeeshops, ski resorts, electronics repair, duty-free shops, sushi bars

Stars

-all of them are in range 1-5 and have 0.5 step

Is\_open

* 17.69% closed businesses
* For 10 manually checked businesses: 1 opened and 9 permanently closed according to Google maps

Categorization of data

Rules for categorization:

**Low-quality data**: records that have missing value in at least one of the significant columns, all records that have invalid patterns

**Medium-quality data**: records that do not have missing values but some values are not standardized

**High-quality data**: the rest of data that is standardized and have all significant attributes

There was another criteria for categorization where certain weights were assigned to some of the attributes, including significant attributes. The assigned weights were as follows: name = 20, address = 15, city = 10, postal\_code = 5, state = 5, latitude = 10, longitude = 10, stars = 10, categories = 15.

Since results showed there is 4.28% medium quality data and all rest is high quality, we have concluded it is not good categorization since there was not low quality data at all even though we have some missing values in significant attributes. That is the reason why we changed the criteria and switched to counting the percentage of invalid and standardized data as well as percentage of missing data as a measurement of quality.

The final categorization is as follows:

**Low-quality data**: 4.3%

**Medium-quality data**: 1.5%

**High-quality data**: 94.2%