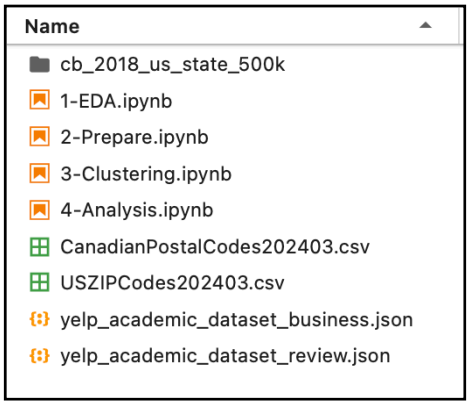
Yelp Analysis Documentation

Example Running 2-Prepare.ipynb, 3-Clustering.ipynb, 4-Analysis.ipynb with Different Category of Businesses - Gastropubs

1. Before starting, ensure all requirements are installed and all required source files are loaded into local folder, in addition to the ipynb notebooks:
   - yelp_academic_dataset_business.json
   - yelp_academic_dataset_review.json
   - USZIPCodes202403.csv
   - CanadianPostalCodes202403.csv
   - Folder called “cb_2018_us_state_500k” with 7 files inside all starting with “cb_2018_us_state_500k”



   (See README for more details about these requirements and links to data sources.)

Preparing Data: Open 2-Prepare.ipynb

2. Change filter to match desired category or subset of businesses.
   In this example, we are filtering for open Gatropubs instead of open Restaurants.

2.2.1. Select subset of businesses

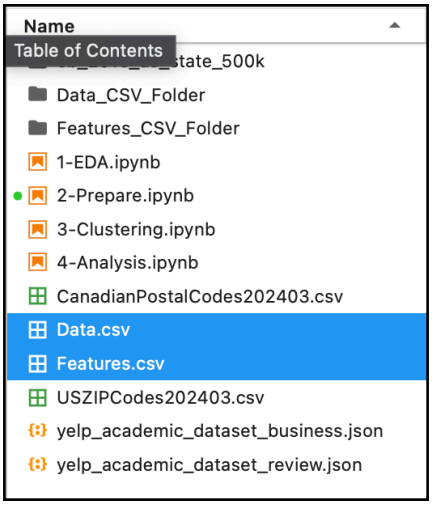
UPDATE HERE:

[7]:
```
# SELECT DESIRED SUBSET OF BUSINESSES

# all open businesses (119698):
#filteredDF = businessDF.where(businessDF.is_open == 1)

# open Restaurants (34987):
#filteredDF = businessDF.filter(businessDF.categories.contains('Restaurants')).where(businessDF.is_open == 1)

# open Gastropubs (331):
filteredDF = businessDF.filter(businessDF.categories.contains('Gastropubs')).where(businessDF.is_open == 1)
```

3. Run notebook.

4. Find the new file named “part-00000-…” in folder named “Features_CSV_Folder” Rename the “part-…” file as “Features.csv” and move it into local folder.

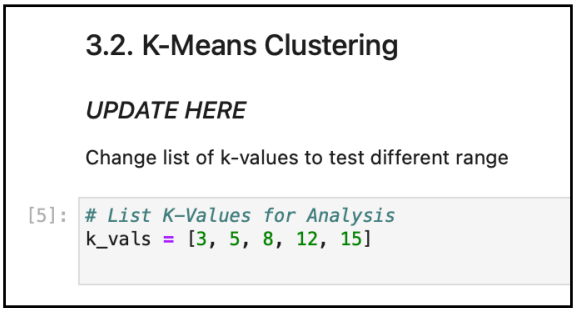
5. Similarly, find the new file named “part-00000-…” in folder named “Data_CSV_Folder” Rename the “part-…” file as “Data.csv” and move it into local folder.



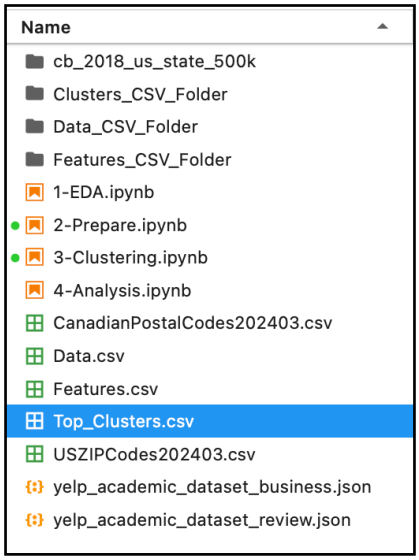
Clustering: Open 3-Clustering.ipynb

6. Update list of k-means values as needed.
   In this example, we are testing 3, 5, 8, 12, 15

7. Run notebook.



8. Find the new file named “part-00000-…” in folder named “Clusters_CSV_Folder” Rename the “part-…” file as “Features.csv” and move it into local folder.



Results displayed in Clustering Notebook:

Highest Minimum Stars

K_Value	Cluster	Business_Count	Min_Stars	Max_Stars	Mean_Stars	Median_Stars	Stars_Std
3	1	106	3.0	5.0	3.943396226415094	4.0	0.3672338992352546
5	1	61	3.0	5.0	4.0	4.0	0.4082482904638632
8	4	58	3.0	4.5	3.9655172413793105	4.0	0.30867670257903096
5	2	56	3.0	4.5	3.8482142857142856	4.0	0.3682840166001192
3	0	127	2.5	4.5	3.7755905511811023	4.0	0.4396456970762142
5	0	104	2.5	4.5	3.8461538461538463	4.0	0.4067209813669471
8	3	69	2.5	4.5	3.7028985507246377	4.0	0.47215744336121407
5	4	60	2.5	5.0	3.8916666666666666	4.0	0.7253968327555279
3	2	95	2.5	5.0	3.8842105263157896	4.0	0.6500925939734137

Lowest Maximum Stars

K_Value	Cluster	Business_Count	Min_Stars	Max_Stars	Mean_Stars	Median_Stars	Stars_Std
5	0	104	2.5	4.5	3.8461538461538463	4.0	0.4067209813669471
8	3	69	2.5	4.5	3.7028985507246377	4.0	0.47215744336121407
3	0	127	2.5	4.5	3.7755905511811023	4.0	0.4396456970762142
8	4	58	3.0	4.5	3.9655172413793105	4.0	0.30867670257903096
5	2	56	3.0	4.5	3.8482142857142856	4.0	0.3682840166001192
3	1	106	3.0	5.0	3.943396226415094	4.0	0.3672338992352546
5	1	61	3.0	5.0	4.0	4.0	0.4082482904638632
3	2	95	2.5	5.0	3.8842105263157896	4.0	0.6500925939734137
5	4	60	2.5	5.0	3.8916666666666666	4.0	0.7253968327555279

Highest Mean Stars

K_Value	Cluster	Business_Count	Min_Stars	Max_Stars	Mean_Stars	Median_Stars	Stars_Std
5	1	61	3.0	5.0	4.0	4.0	0.4082482904638632
8	4	58	3.0	4.5	3.9655172413793105	4.0	0.30867670257903096
3	1	106	3.0	5.0	3.943396226415094	4.0	0.3672338992352546
5	4	60	2.5	5.0	3.8916666666666666	4.0	0.7253968327555279
3	2	95	2.5	5.0	3.8842105263157896	4.0	0.6500925939734137
5	2	56	3.0	4.5	3.8482142857142856	4.0	0.3682840166001192
5	0	104	2.5	4.5	3.8461538461538463	4.0	0.4067209813669471
3	0	127	2.5	4.5	3.7755905511811023	4.0	0.4396456970762142
8	3	69	2.5	4.5	3.7028985507246377	4.0	0.47215744336121407

Lowest Mean Stars

K_Value	Cluster	Business_Count	Min_Stars	Max_Stars	Mean_Stars	Median_Stars	Stars_Std
8	3	69	2.5	4.5	3.7028985507246377	4.0	0.47215744336121407
3	0	127	2.5	4.5	3.7755905511811023	4.0	0.4396456970762142
5	0	104	2.5	4.5	3.8461538461538463	4.0	0.4067209813669471
5	2	56	3.0	4.5	3.8482142857142856	4.0	0.3682840166001192
3	2	95	2.5	5.0	3.8842105263157896	4.0	0.6500925939734137
5	4	60	2.5	5.0	3.8916666666666666	4.0	0.7253968327555279
3	1	106	3.0	5.0	3.943396226415094	4.0	0.3672338992352546
8	4	58	3.0	4.5	3.9655172413793105	4.0	0.30867670257903096
5	1	61	3.0	5.0	4.0	4.0	0.4082482904638632

Highest Median Stars

K_Value	Cluster	Business_Count	Min_Stars	Max_Stars	Mean_Stars	Median_Stars	Stars_Std
3	0	127	2.5	4.5	3.7755905511811023	4.0	0.4396456970762142
8	3	69	2.5	4.5	3.7028985507246377	4.0	0.47215744336121407
5	0	104	2.5	4.5	3.8461538461538463	4.0	0.4067209813669471
8	4	58	3.0	4.5	3.9655172413793105	4.0	0.30867670257903096
3	1	106	3.0	5.0	3.943396226415094	4.0	0.3672338992352546
5	1	61	3.0	5.0	4.0	4.0	0.4082482904638632
3	2	95	2.5	5.0	3.8842105263157896	4.0	0.6500925939734137
5	4	60	2.5	5.0	3.8916666666666666	4.0	0.7253968327555279
5	2	56	3.0	4.5	3.8482142857142856	4.0	0.3682840166001192

Lowest Median Stars

K_Value	Cluster	Business_Count	Min_Stars	Max_Stars	Mean_Stars	Median_Stars	Stars_Std
3	0	127	2.5	4.5	3.7755905511811023	4.0	0.4396456970762142
8	3	69	2.5	4.5	3.7028985507246377	4.0	0.47215744336121407
5	0	104	2.5	4.5	3.8461538461538463	4.0	0.4067209813669471
8	4	58	3.0	4.5	3.9655172413793105	4.0	0.30867670257903096
3	1	106	3.0	5.0	3.943396226415094	4.0	0.3672338992352546
5	1	61	3.0	5.0	4.0	4.0	0.4082482904638632
3	2	95	2.5	5.0	3.8842105263157896	4.0	0.6500925939734137
5	4	60	2.5	5.0	3.8916666666666666	4.0	0.7253968327555279
5	2	56	3.0	4.5	3.8482142857142856	4.0	0.3682840166001192

Lowest Standard Deviation of Stars

K_Value	Cluster	Business_Count	Min_Stars	Max_Stars	Mean_Stars	Median_Stars	Stars_Std
8	4	58	3.0	4.5	3.9655172413793105	4.0	0.30867670257903096
3	1	106	3.0	5.0	3.943396226415094	4.0	0.3672338992352546
5	2	56	3.0	4.5	3.8482142857142856	4.0	0.3682840166001192
5	0	104	2.5	4.5	3.8461538461538463	4.0	0.4067209813669471
5	1	61	3.0	5.0	4.0	4.0	0.4082482904638632
3	0	127	2.5	4.5	3.7755905511811023	4.0	0.4396456970762142
8	3	69	2.5	4.5	3.7028985507246377	4.0	0.47215744336121407
3	2	95	2.5	5.0	3.8842105263157896	4.0	0.6500925939734137
5	4	60	2.5	5.0	3.8916666666666666	4.0	0.7253968327555279

Analysis: Open 4-Analysis.ipynb

9. Run notebook.

10. Results are displayed in notebook, ready to be interpreted, with clusters available for additional analysis.

Results displayed in Analysis Notebook:

Clusters Overview: Stars

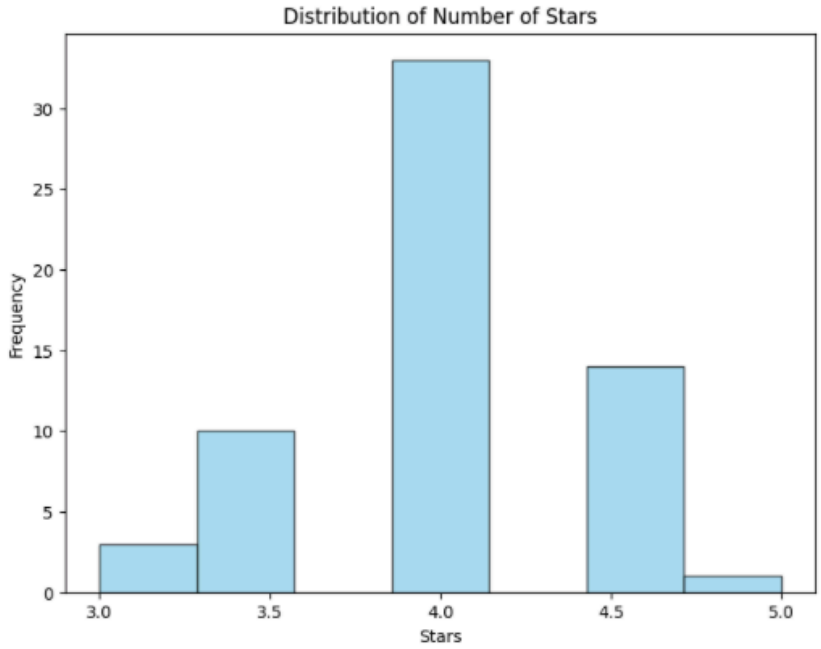
Source	Variance	Min	Max	Mean
Best_1	0.167	3.0	5.0	4.0
Best_2	0.095	3.0	4.5	3.966
Best_3	0.135	3.0	5.0	3.943
Best_4	0.526	2.5	5.0	3.892
Worst_1	0.223	2.5	4.5	3.703
Worst_2	0.193	2.5	4.5	3.776
Worst_3	0.165	2.5	4.5	3.846
Worst_4	0.136	3.0	4.5	3.848

Clusters Overview: Review Count

Source	Variance	Min	Max	Mean
Best_1	203234.6	10.0	2497.0	368.262
Best_2	48880.254	21.0	1245.0	332.534
Best_3	143542.86	36.0	2497.0	398.057
Best_4	5093.249	5.0	455.0	55.65
Worst_1	45072.645	19.0	1210.0	282.275
Worst_2	121536.22	16.0	3260.0	323.685
Worst_3	38302.047	10.0	1210.0	234.76
Worst_4	222602.42	36.0	3260.0	465.786

Best Cluster from Gastropubs:

```
Cluster_Name Best_1
```



Top Categories for Cluster Best_1:

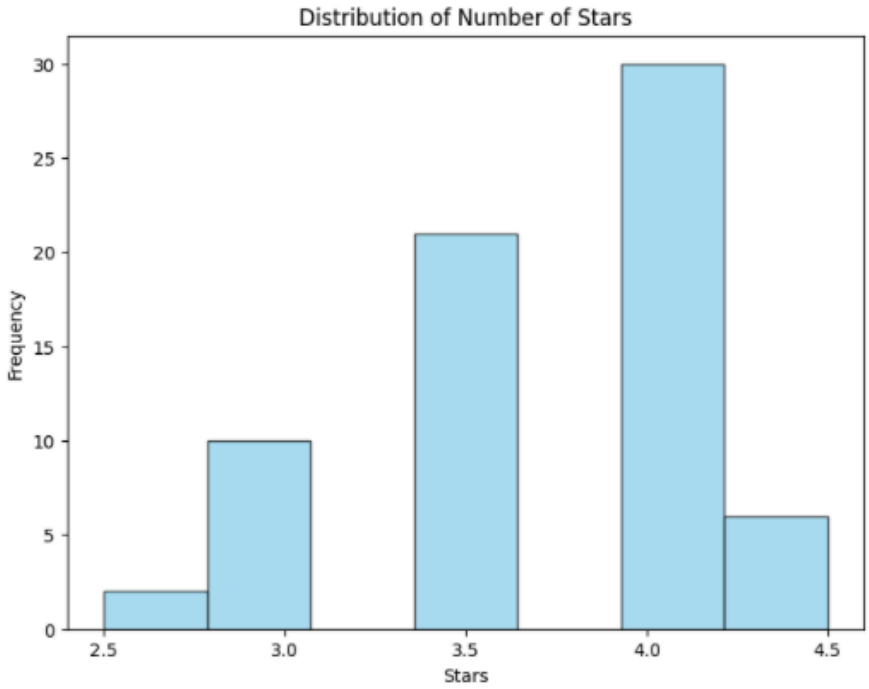
Category	count	Percent
Gastropubs	61	100.0
Restaurants	61	100.0
Nightlife	53	86.885
Bars	52	85.246
Cocktail Bars	19	31.148
American (New)	19	31.148
American (Traditional)	16	26.23
Pubs	16	26.23
Beer Bar	13	21.311
Food	13	21.311

Feature	Variance	Min	Max	Mean
BusinessParking_street	0.0	1.0	1.0	1.0
Alcohol	0.031	0.0	1.0	0.951
BusinessAcceptsCreditCards	0.048	0.0	1.0	0.951
RestaurantsGoodForGroups	0.103	0.0	1.0	0.885
OutdoorSeating	0.128	0.0	1.0	0.852
BikeParking	0.139	0.0	1.0	0.836
HappyHour	0.17	0.0	1.0	0.787
RestaurantsTakeOut	0.189	0.0	1.0	0.754
HasTV	0.197	0.0	1.0	0.738
Ambience_classy	0.218	0.0	1.0	0.689
RestaurantsTableService	0.23	0.0	1.0	0.656
WiFi	0.246	0.0	1.0	0.59
GoodForMeal_dinner	0.249	0.0	1.0	0.574
Ambience_casual	0.251	0.0	1.0	0.557
RestaurantsDelivery	0.252	0.0	1.0	0.459
BestNights_saturday	0.246	0.0	1.0	0.41
BestNights_friday	0.246	0.0	1.0	0.41
WheelchairAccessible	0.246	0.0	1.0	0.41
RestaurantsReservations	0.239	0.0	1.0	0.377
NoiseLevel	0.019	0.0	0.67	0.369
Ambience_trendy	0.234	0.0	1.0	0.361
RestaurantsPriceRange2	0.008	0.0	0.67	0.336
GoodForMeal_latenight	0.224	0.0	1.0	0.328
Caters	0.161	0.0	1.0	0.197
DogsAllowed	0.161	0.0	1.0	0.197
BusinessParking_garage	0.15	0.0	1.0	0.18
Ambience_hipster	0.139	0.0	1.0	0.164
BusinessParking_lot	0.139	0.0	1.0	0.164
BestNights_wednesday	0.139	0.0	1.0	0.164
BestNights_thursday	0.139	0.0	1.0	0.164
BestNights_monday	0.103	0.0	1.0	0.115
Smoking	0.051	0.0	1.0	0.107
Corkage	0.09	0.0	1.0	0.098
BestNights_tuesday	0.077	0.0	1.0	0.082
GoodForMeal_lunch	0.062	0.0	1.0	0.066
Music_live	0.048	0.0	1.0	0.049
GoodForKids	0.032	0.0	1.0	0.033
GoodForMeal_brunch	0.032	0.0	1.0	0.033
BusinessParking_valet	0.0	0.0	0.0	0.0
GoodForMeal_dessert	0.0	0.0	0.0	0.0

```
[Stage 956:>
Average words per review: 110.91697885716746
```

Worst Cluster from Gastropubs:

```
Cluster_Name Worst_1
```



Top Categories for Cluster Worst_1:

Category	count	Percent
Gastropubs	69	100.0
Restaurants	69	100.0
Nightlife	57	82.609
Bars	56	81.159
American (New)	41	59.42
American (Traditional)	35	50.725
Pubs	29	42.029
Food	25	36.232
Burgers	22	31.884
Sports Bars	19	27.536

Feature	Variance	Min	Max	Mean
RestaurantsGoodForGroups	0.0	1.0	1.0	1.0
HasTV	0.0	1.0	1.0	1.0
BusinessAcceptsCreditCards	0.014	0.0	1.0	0.986
RestaurantsTakeOut	0.029	0.0	1.0	0.971
Ambience_casual	0.029	0.0	1.0	0.971
GoodForMeal_dinner	0.042	0.0	1.0	0.957
WiFi	0.055	0.0	1.0	0.942
Alcohol	0.036	0.5	1.0	0.913
OutdoorSeating	0.126	0.0	1.0	0.855
BusinessParking_lot	0.126	0.0	1.0	0.855
RestaurantsTableService	0.126	0.0	1.0	0.855
GoodForKids	0.136	0.0	1.0	0.841
GoodForMeal_lunch	0.136	0.0	1.0	0.841
BikeParking	0.173	0.0	1.0	0.783
HappyHour	0.181	0.0	1.0	0.768
WheelchairAccessible	0.247	0.0	1.0	0.58
BestNights_friday	0.247	0.0	1.0	0.58
BestNights_saturday	0.253	0.0	1.0	0.522
RestaurantsDelivery	0.253	0.0	1.0	0.478
Caters	0.253	0.0	1.0	0.478
RestaurantsReservations	0.242	0.0	1.0	0.391
NoiseLevel	0.022	0.0	0.67	0.384
RestaurantsPriceRange2	0.002	0.0	0.33	0.325
GoodForMeal_latenight	0.22	0.0	1.0	0.319
BestNights_tuesday	0.188	0.0	1.0	0.246
Ambience_classy	0.181	0.0	1.0	0.232
BusinessParking_street	0.164	0.0	1.0	0.203
Music_live	0.146	0.0	1.0	0.174
BestNights_thursday	0.126	0.0	1.0	0.145
DogsAllowed	0.126	0.0	1.0	0.145
BestNights_monday	0.115	0.0	1.0	0.13
BestNights_wednesday	0.092	0.0	1.0	0.101
Smoking	0.053	0.0	1.0	0.094
GoodForMeal_dessert	0.068	0.0	1.0	0.072
Ambience_trendy	0.055	0.0	1.0	0.058
BusinessParking_garage	0.042	0.0	1.0	0.043
GoodForMeal_brunch	0.029	0.0	1.0	0.029
Ambience_hipster	0.014	0.0	1.0	0.014
BusinessParking_valet	0.014	0.0	1.0	0.014
Corkage	0.014	0.0	1.0	0.014

```
[Stage 2314:============================>
Average words per review: 99.29320321694782
```