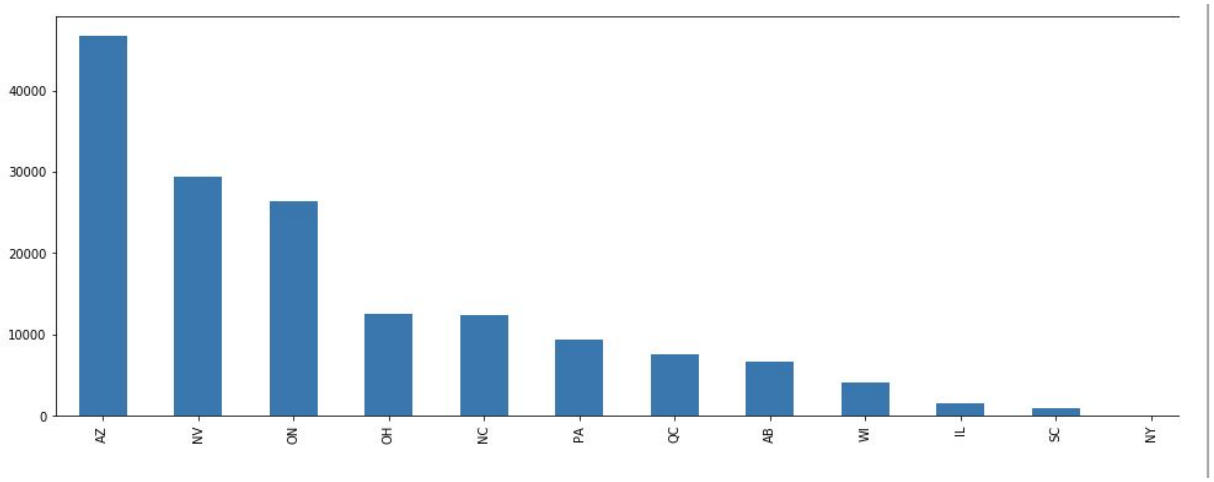
Presentation Report
Group Assignment #4

Data is the most important element of machine learning algorithms. As with good data you can use the best algorithm to further enhance the focal point. Data is the core of machine learning algorithms as it must be supplied supplied properly in order for the machine learning algorithm to understand the needs properly. Because the main purpose of machine learning algorithms is to unlock the concealed information/knowledge available in the data. The algorithms will end up providing incorrect, bogus insights if the data is available in a form not comprehended by the algorithm. We will be ensuring this happens with the tasks that we are completing with the data from the Yelp Dataset challenge.

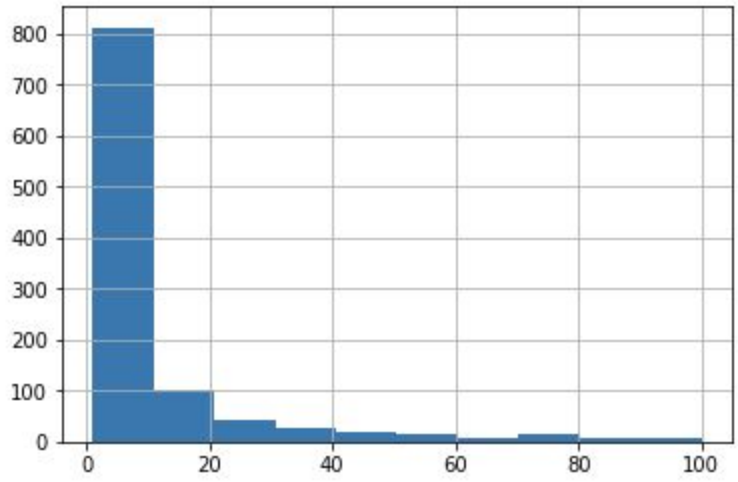
For the first task, we are trying to create recommendations for food places for users based on previous destinations they want to. For this task, we took into consideration the business table and the tips table. Within the business table, we first filtered all business's to solely business in Arizona. To insure the cities were in Arizona we found a dataset online with all Arizona cities that could help us determine if the cities given to us are actually there. We did this as Arizona was one of the top tier places in the data set that contained businesses.



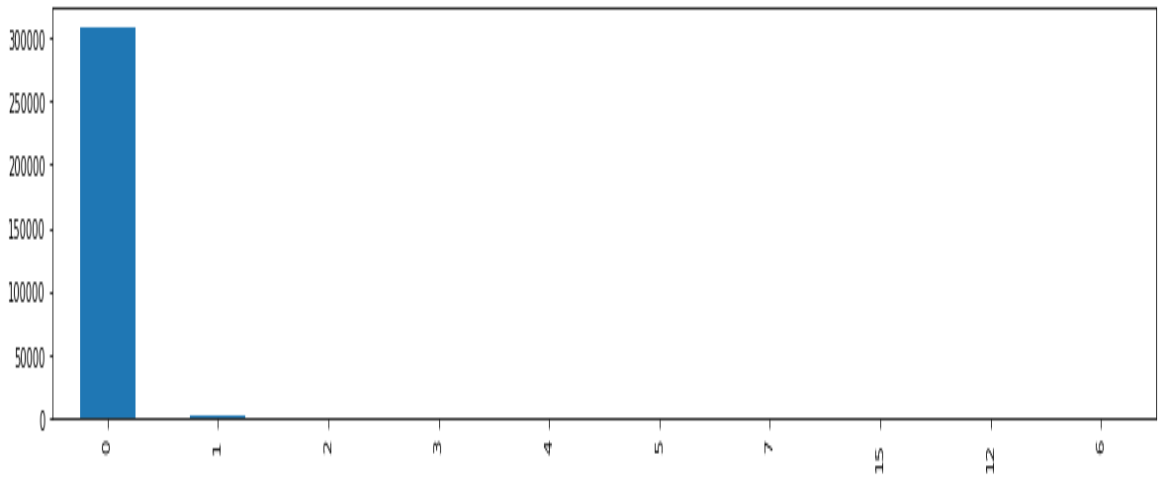
Another instance of cleaning we had to do is we wanted to look more in depth at the review distribution among business.

What we came to notice is that the average business had an average review count of around 136. So we are only taking business' into consideration that have those many reviews or more.

Review Count Distribution



We also filtered the categories column and remove categories that had no association with Food. (Did this with looking of the frequency of words that appear in a category and with the help of our word cloud) We also filledNa for businesses that have null categories or attributes. Even though these places have no metadata association to them, they could maybe a commodity in our dataset. Additionally we looked at the review count again as now we have filtered data to just food related places. Graphically, we found there were businesses with 1,2,3,or 4 reviews. These businesses were considered outliers. So we are taking business with reviews over 63. Also filled in 0.5 for values that were not listed as true or false. (Even though its false, how do we know if its actually false the attribute). Finally we are taking the text from the tips table into consideration. We will not accept compliments feature from that table in consideration as it is not used to such an extent.



To clean this field, we normalized the field and tokenized. We wanted to make sure the text is in a form that is predictable and analyzable for our task. With the tips feature, we are hoping to create a feature that will help us take into account the words and their frequency of occurrence. This could be a strong factor in helping give a recommendation to a user.

The feature attributes we will be left so far for machine learning task 1 is as follows:

```
   Unnamed: 0             business_id                 date  \
0           0  5KheTjYPu1HcQzQFtm4_vw  2011-12-26 01:46:17
1           1  5KheTjYPu1HcQzQFtm4_vw  2010-09-26 02:05:03
2           2  5KheTjYPu1HcQzQFtm4_vw  2013-12-13 07:24:47
3           3  5KheTjYPu1HcQzQFtm4_vw  2012-03-29 22:15:28
4           4  5KheTjYPu1HcQzQFtm4_vw  2012-08-09 23:06:14

                  user_id  Unnamed: 0.1  \
0  jRyO2V1pA4CdVVqCIOPc1Q         64070
1  grQ3v_hGwyf7MfjAcpx-qg         64070
2  ISNbZn_lBKP7w1aGGaVyvw         64070
3  JiaJJzXFc6vBPke9v_V9Uw         64070
4  d6cLvMTHXdxhAXTdcjuNxQ         64070

                                          categories     city  \
0  Mexican, Restaurants, Bars, Nightlife, Breakfa...  Phoenix
1  Mexican, Restaurants, Bars, Nightlife, Breakfa...  Phoenix
2  Mexican, Restaurants, Bars, Nightlife, Breakfa...  Phoenix
3  Mexican, Restaurants, Bars, Nightlife, Breakfa...  Phoenix
4  Mexican, Restaurants, Bars, Nightlife, Breakfa...  Phoenix

                   name  review_count  stars  ... RestaurantsReservations  \
0  Arriba Mexican Grill           285    3.5  ...                      True
1  Arriba Mexican Grill           285    3.5  ...                      True
2  Arriba Mexican Grill           285    3.5  ...                      True
3  Arriba Mexican Grill           285    3.5  ...                      True
4  Arriba Mexican Grill           285    3.5  ...                      True

  RestaurantsTableService RestaurantsTakeOut Smoking WheelchairAccessible  \
0                    True                True       0                     0
1                    True                True       0                     0
2                    True                True       0                     0
3                    True                True       0                     0
4                    True                True       0                     0

  WiFi_'no' WiFi_u'free' WiFi_u'no' WiFi_u'paid'  \
0         0            0          1            0
1         0            0          1            0
2         0            0          1            0
3         0            0          1            0
4         0            0          1            0

                                        cleaned_tips
0  ['good chips and salsa loud at times good serv...
1                    ['queso fundido is the best']
2  ['fyi the green chilis are spicy i dont mind a...
3  ['this therapist is cheap 99cent margaritas ev...
4                  ['really great refried beans']

[5 rows x 52 columns]
Index(['Unnamed: 0', 'business_id', 'date', 'user_id', 'Unnamed: 0.1',
       'categories', 'city', 'name', 'review_count', 'stars', 'state',
       'AgesAllowed', 'Alcohol', 'Ambience', 'BYOB', 'BYOBCorkage',
       'BestNights', 'BikeParking', 'BusinessAcceptsBitcoin',
       'BusinessAcceptsCreditCards', 'BusinessParking', 'ByAppointmentOnly',
       'Caters', 'CoatCheck', 'Corkage', 'DietaryRestrictions', 'DogsAllowed',
       'DriveThru', 'GoodForDancing', 'GoodForKids', 'GoodForMeal',
       'HappyHour', 'HasTV', 'Music', 'NoiseLevel', 'Open24Hours',
       'OutdoorSeating', 'RestaurantsAttire', 'RestaurantsCounterService',
       'RestaurantsDelivery', 'RestaurantsGoodForGroups',
       'RestaurantsPriceRange2', 'RestaurantsReservations',
       'RestaurantsTableService', 'RestaurantsTakeOut', 'Smoking',
       'WheelchairAccessible', 'WiFi_'no'', 'WiFi_u'free'', 'WiFi_u'no'',
       'WiFi_u'paid'', 'cleaned_tips'],
      dtype='object')
```

The second machine learning task is we are taking the reviews table and considering only businesses in Arizona. Within here we are going to try to determine if we get. What concerns a customer and what makes a good restaurant

- Did a user like or not like a restaurant?

The datasets we will be using is the reviews table and the business table. We will be taking the reviews table only into consideration. Maybe the business table to help give us an insight of where was the business and what is?

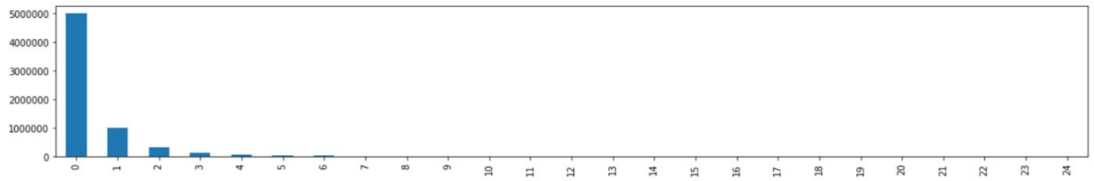
For the review table we looked into the features it has which are business_id, cool, date, funny, stars, text, useful, userid.

The features we were curious about is cool, funny, stars, and useful. These attributes we will not be factoring in as they are rarely used what so ever. Will give us no benefit

when we move onto feature engineering.

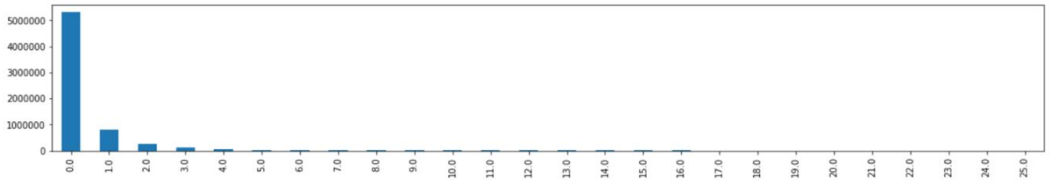
```
In [4]: cool=reviews['cool'].values;
        pd.Series(cool).value_counts()[:25].plot(kind='bar',figsize=(20,3))

Out[4]: <matplotlib.axes._subplots.AxesSubplot at 0x110105810>
```



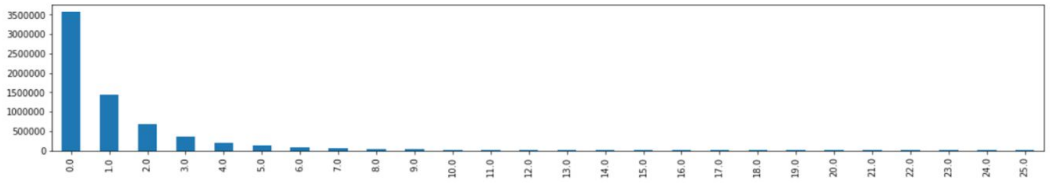

```
In [5]: funny=reviews['funny'].values;
        pd.Series(funny).value_counts()[:25].plot(kind='bar',figsize=(20,3))

Out[5]: <matplotlib.axes._subplots.AxesSubplot at 0x1105380d0>
```

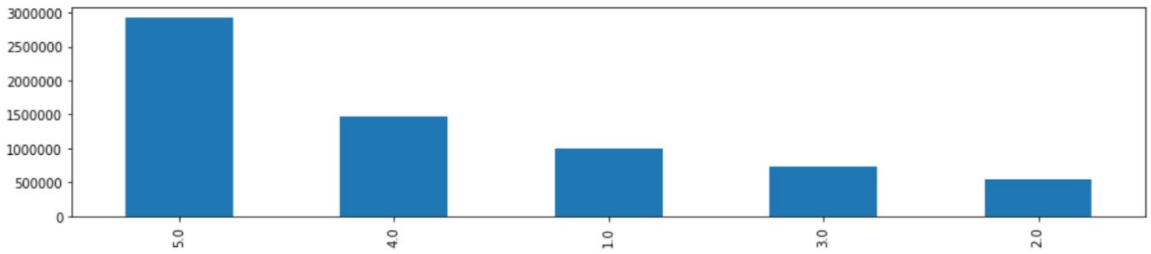



```
In [6]: useful=reviews['useful'].values;
        pd.Series(useful).value_counts()[:25].plot(kind='bar',figsize=(20,3))

Out[6]: <matplotlib.axes._subplots.AxesSubplot at 0x11058bd50>
```



```
Out[7]: <matplotlib.axes._subplots.AxesSubplot at 0x10eab8490>
```



The distribution for rating for stars makes sense as there is a wide range of ratings.

For text, we will be taking a basic approach for preprocessing. We will be applying the same preprocessing approach as we did for tips.

The attributes we will be factoring in for our second machine learning task is as follows:

```
In [2]: import pandas as pd

reviews = pd.read_csv('../Data/review.csv')
del reviews['cool']
del reviews['useful']
del reviews['funny']
print(reviews.head())
```

```
              business_id                 date               review_id  stars  \
0  ujmEBvifdJM6h6RLv4wQIg  2013-05-07 04:34:36  Q1sbwvVQXV2734tPgoKj4Q    1.0
1  NZnhc2sEQy3RmzKTZnqtwQ  2017-01-14 21:30:33  GJXCdrto3ASJOqKeVWPi6Q    5.0
2  WTqjgwHlXbSFevF32_DJVw  2016-11-09 20:09:03  2TzJjDVDEuAW6MR5Vuc1ug    5.0
3  ikCg8xy5JIg_NGPx-MSIDA  2018-01-09 20:56:38  yi0R0Ugj_xUx_Nek0-_Qig    5.0
4  b1b1eb3uo-w561D0ZfCEiQ  2018-01-30 23:07:38  11a8sVPMUFtaC7_ABRkmtw    1.0

                                                text                 user_id
0  Total bill for this horrible service? Over $8G...  hG7b0MtEbXx5QzbzE6C_VA
1  I *adore* Travis at the Hard Rock's new Kelly ...  yXQM5uF2jS6es16SJzNHfg
2  I have to say that this office really has it t...  n6-Gk65cPZL6Uz8qRm3NYw
3  Went in for a lunch. Steak sandwich was delici...  dacAIZ6fTM6mqwW5uxkskg
4  Today was my second out of three sessions I ha...  ssoyf2_x0EQMed6fgHeMyQ
```

```
In [3]: print(reviews.columns)
```

```
Index(['business_id', 'date', 'review_id', 'stars', 'text', 'user_id'], dtype='object')
```

I do not believe so far with how we are condensing our data we will have issues with our machine learning results. We are taking the appropriate considerations from filling in null values and handling outliers. The only thing, we are worried about is the text attributes. Text is not an easy thing to preprocess and it is hard to tell if taking this

approach will be okay or not.