Untitled1

June 25, 2020

[2]:

```
#import all the required librariers
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
```

[3]:

```
#Load the CSV file

data= pd.read_csv('train.csv')
data.head(5)
```

[3]:

```
   index  beer/ABV  beer/beerId  beer/brewerId                  beer/name  \
0  40163       5.0        46634          14338                   Chiostro
1   8135      11.0         3003            395   Bearded Pat's Barleywine
2  10529       4.7          961            365       Naughty Nellie's Ale
3  44610       4.4          429              1            Pilsner Urquell
4  37062       4.4         4904           1417  Black Sheep Ale (Special)

                beer/style  review/appearance  review/aroma  review/overall  \
0     Herbed / Spiced Beer                4.0           4.0             4.0
1      American Barleywine                4.0           3.5             3.5
2  American Pale Ale (APA)                3.5           4.0             3.5
3           Czech Pilsener                3.0           3.0             2.5
4         English Pale Ale                4.0           3.0             3.0

   review/palate  review/taste  \
0            4.0           4.0
1            3.5           3.0
2            3.5           3.5
3            3.0           3.0
4            3.5           2.5

                                         review/text  \
0  Pours a clouded gold with a thin white head. N…
1  12oz bottle into 8oz snifter.\t\tDeep ruby red…
2  First enjoyed at the brewpub about 2 years ago…
3  First thing I noticed after pouring from green…
4  A: pours an amber with a one finger head but o…

                                 review/timeStruct  review/timeUnix  \
0  {'min': 38, 'hour': 3, 'mday': 16, 'sec': 10, …       1229398690
1  {'min': 38, 'hour': 23, 'mday': 8, 'sec': 58, …       1218238738
2  {'min': 7, 'hour': 18, 'mday': 26, 'sec': 2, '…       1101492422
3  {'min': 7, 'hour': 1, 'mday': 20, 'sec': 5, 'y…       1308532025
4  {'min': 51, 'hour': 6, 'mday': 12, 'sec': 48, …       1299912708

   user/ageInSeconds user/birthdayRaw  user/birthdayUnix user/gender  \
0                NaN              NaN                NaN         NaN
1                NaN              NaN                NaN         NaN
2                NaN              NaN                NaN        Male
3       1.209827e+09     Aug 10, 1976        208508400.0        Male
4                NaN              NaN                NaN         NaN

  user/profileName
0       RblWthACoz
1          BeerSox
2       mschofield
3        molegar76
4       Brewbro000
```

[4]:
```
#checking data from end
data.tail(5)
```

[4]:
```
       index  beer/ABV  beer/beerId  beer/brewerId  \
37495  35175      5.50        22450           3268
37496  23666      8.50         7463           1199
37497  47720      4.75         1154            394
37498  33233     11.20        19960           1199
37499  23758      8.50         7463           1199

                                      beer/name  \
37495                Blackberry Scottish-Style
37496                   Founders Dirty Bastard
37497                            Stoudt's Fest
37498  Founders KBS (Kentucky Breakfast Stout)
37499                   Founders Dirty Bastard

                        beer/style  review/appearance  review/aroma  \
37495         Fruit / Vegetable Beer                4.0           3.5
37496         Scotch Ale / Wee Heavy                4.5           4.0
```

```
37497             MÃ¤rzen / Oktoberfest                4.0           3.5
37498  American Double / Imperial Stout                4.0           4.0
37499            Scotch Ale / Wee Heavy                4.0           4.0

       review/overall  review/palate  review/taste  \
37495             3.5            3.5           3.5
37496             3.5            4.5           4.5
37497             4.0            4.5           4.0
37498             4.0            5.0           5.0
37499             4.0            4.5           4.0

                                             review/text  \
37495  12 oz brown longneck with no freshness dating…
37496  A - A bright red with a maroon-amber hue; mini…
37497  Sampled on tap at Redbones.\t\tThis marzen sty…
37498  Pours a black body with a brown head that very…
37499  A nice sweet, malty beer…nothing complex, ju…

                                       review/timeStruct  review/timeUnix  \
37495  {'min': 56, 'hour': 23, 'mday': 10, 'sec': 1, …       1207871761
37496  {'min': 45, 'hour': 5, 'mday': 10, 'sec': 14, …       1263102314
37497  {'min': 3, 'hour': 1, 'mday': 25, 'sec': 36, '…       1067043816
37498  {'min': 52, 'hour': 19, 'mday': 29, 'sec': 33,…       1296330753
37499  {'min': 40, 'hour': 18, 'mday': 4, 'sec': 28, …       1252089628

       user/ageInSeconds user/birthdayRaw  user/birthdayUnix user/gender  \
37495                NaN              NaN                NaN         NaN
37496                NaN              NaN                NaN         NaN
37497                NaN              NaN                NaN         NaN
37498                NaN              NaN                NaN         NaN
37499                NaN              NaN                NaN         NaN

      user/profileName
37495         Redrover
37496         jmerloni
37497       UncleJimbo
37498       Stockfan42
37499           JayQue
```

[5]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 37500 entries, 0 to 37499
Data columns (total 19 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   index              37500 non-null  int64
 1   beer/ABV           37500 non-null  float64
 2   beer/beerId        37500 non-null  int64
 3   beer/brewerId      37500 non-null  int64
 4   beer/name          37500 non-null  object
 5   beer/style         37500 non-null  object
 6   review/appearance  37500 non-null  float64
 7   review/aroma       37500 non-null  float64
 8   review/overall     37500 non-null  float64
 9   review/palate      37500 non-null  float64
 10  review/taste       37500 non-null  float64
 11  review/text        37490 non-null  object
 12  review/timeStruct  37500 non-null  object
 13  review/timeUnix    37500 non-null  int64
 14  user/ageInSeconds  7856 non-null   float64
 15  user/birthdayRaw   7856 non-null   object
 16  user/birthdayUnix  7856 non-null   float64
 17  user/gender        15314 non-null  object
 18  user/profileName   37495 non-null  object
dtypes: float64(8), int64(4), object(7)
memory usage: 5.4+ MB
```

[6]:
```
data.describe()
```

[6]:
```
              index      beer/ABV   beer/beerId  beer/brewerId  \
count  37500.000000  37500.000000  37500.000000   37500.000000
mean   24951.887573      7.403725  21861.152027    3036.595120
std    14434.009669      2.318145  18923.130832    5123.084675
min        0.000000      0.100000    175.000000       1.000000
25%    12422.500000      5.400000   5441.000000     395.000000
50%    24942.500000      6.900000  17538.000000    1199.000000
75%    37416.750000      9.400000  34146.000000    1315.000000
max    49999.000000     57.700000  77207.000000   27797.000000

       review/appearance  review/aroma  review/overall  review/palate  \
count       37500.000000  37500.000000     37500.00000   37500.000000
mean            3.900053      3.873240         3.88944       3.854867
std             0.588778      0.680865         0.70045       0.668068
min             0.000000      1.000000         0.00000       1.000000
25%             3.500000      3.500000         3.50000       3.500000
50%             4.000000      4.000000         4.00000       4.000000
75%             4.500000      4.500000         4.50000       4.500000
max             5.000000      5.000000         5.00000       5.000000

       review/taste  review/timeUnix  user/ageInSeconds  user/birthdayUnix
count  37500.000000     3.750000e+04       7.856000e+03       7.856000e+03
mean       3.922440     1.232794e+09       1.176705e+09       2.416303e+08
std        0.716504     7.190955e+07       3.375514e+08       3.375514e+08
```

```
min        1.000000    9.262944e+08     7.034366e+08      -2.208960e+09
25%        3.500000    1.189194e+09     9.794810e+08       1.433628e+08
50%        4.000000    1.248150e+09     1.100009e+09       3.183264e+08
75%        4.500000    1.291330e+09     1.274973e+09       4.388544e+08
max        5.000000    1.326267e+09     3.627295e+09       7.148988e+08
```

[7]:

```
#check Data Type
data.dtypes
```

[7]:

```
index                    int64
beer/ABV               float64
beer/beerId              int64
beer/brewerId            int64
beer/name               object
beer/style              object
review/appearance      float64
review/aroma           float64
review/overall         float64
review/palate          float64
review/taste           float64
review/text             object
review/timeStruct       object
review/timeUnix          int64
user/ageInSeconds      float64
user/birthdayRaw        object
user/birthdayUnix      float64
user/gender             object
user/profileName        object
dtype: object
```

[8]:

```
#Check for missing Value
data.isna().sum()
```

[8]:

```
index                     0
beer/ABV                  0
beer/beerId               0
beer/brewerId             0
beer/name                 0
beer/style                0
review/appearance         0
review/aroma              0
review/overall            0
review/palate             0
review/taste              0
review/text              10
review/timeStruct         0
review/timeUnix           0
```

```
user/ageInSeconds    29644
user/birthdayRaw     29644
user/birthdayUnix    29644
user/gender          22186
user/profileName         5
dtype: int64
```

[9]:

```
data.columns
```

[9]:

```
Index(['index', 'beer/ABV', 'beer/beerId', 'beer/brewerId', 'beer/name',
       'beer/style', 'review/appearance', 'review/aroma', 'review/overall',
       'review/palate', 'review/taste', 'review/text', 'review/timeStruct',
       'review/timeUnix', 'user/ageInSeconds', 'user/birthdayRaw',
       'user/birthdayUnix', 'user/gender', 'user/profileName'],
      dtype='object')
```

[10]:

```
#Remove unnecessary columns and deal with missing value

data = data.drop(["beer/brewerId"], axis=1)
data = data.drop(["beer/beerId"], axis=1)
data = data.drop(["review/timeUnix"], axis=1)
data = data.drop(["user/profileName"], axis=1)
data = data.dropna()
data.head()
```

[10]:

```
    index  beer/ABV                             beer/name      beer/style  \
3   44610       4.4                        Pilsner Urquell  Czech Pilsener
19  29757       7.2                Founders Centennial IPA    American IPA
22  35307       5.5                            Pumpkin Ale     Pumpkin Ale
32  12702       6.0                               La Goule         Witbier
39  42710       5.4  Aecht Schlenkerla Rauchbier MÃ¤rzen      Rauchbier

    review/appearance  review/aroma  review/overall  review/palate  \
3                 3.0           3.0             2.5            3.0
19                3.5           3.5             4.0            4.0
22                3.0           4.0             5.0            4.5
32                3.5           3.5             3.5            4.0
39                3.5           4.5             3.5            3.5

    review/taste                                        review/text  \
3            3.0  First thing I noticed after pouring from green…
19           3.5  The Centennial IPA pours a nicely carbonated r…
22           4.5  Pours a murky amber with a nice off-white head…
32           3.0  This one is only found in France where I got i…
39           4.5  Pours a caramel brown color. With a very subtl…

                                    review/timeStruct  user/ageInSeconds  \
3   {'min': 7, 'hour': 1, 'mday': 20, 'sec': 5, 'y…       1.209827e+09
19  {'min': 40, 'hour': 0, 'mday': 11, 'sec': 0, '…       1.203865e+09
22  {'min': 24, 'hour': 16, 'mday': 28, 'sec': 50,…       1.110294e+09
32  {'min': 23, 'hour': 3, 'mday': 22, 'sec': 9, '…       1.228831e+09
39  {'min': 0, 'hour': 21, 'mday': 28, 'sec': 4, '…       9.078554e+08

   user/birthdayRaw  user/birthdayUnix user/gender
3      Aug 10, 1976        208508400.0        Male
19     Oct 18, 1976        214470000.0        Male
22      Oct 6, 1979        308041200.0        Male
32      Jan 3, 1976        189504000.0        Male
39      Mar 6, 1986        510480000.0        Male
```

[11]:
```
#Review on available colums
data.columns
```

[11]:
```
Index(['index', 'beer/ABV', 'beer/name', 'beer/style', 'review/appearance',
       'review/aroma', 'review/overall', 'review/palate', 'review/taste',
       'review/text', 'review/timeStruct', 'user/ageInSeconds',
       'user/birthdayRaw', 'user/birthdayUnix', 'user/gender'],
      dtype='object')
```

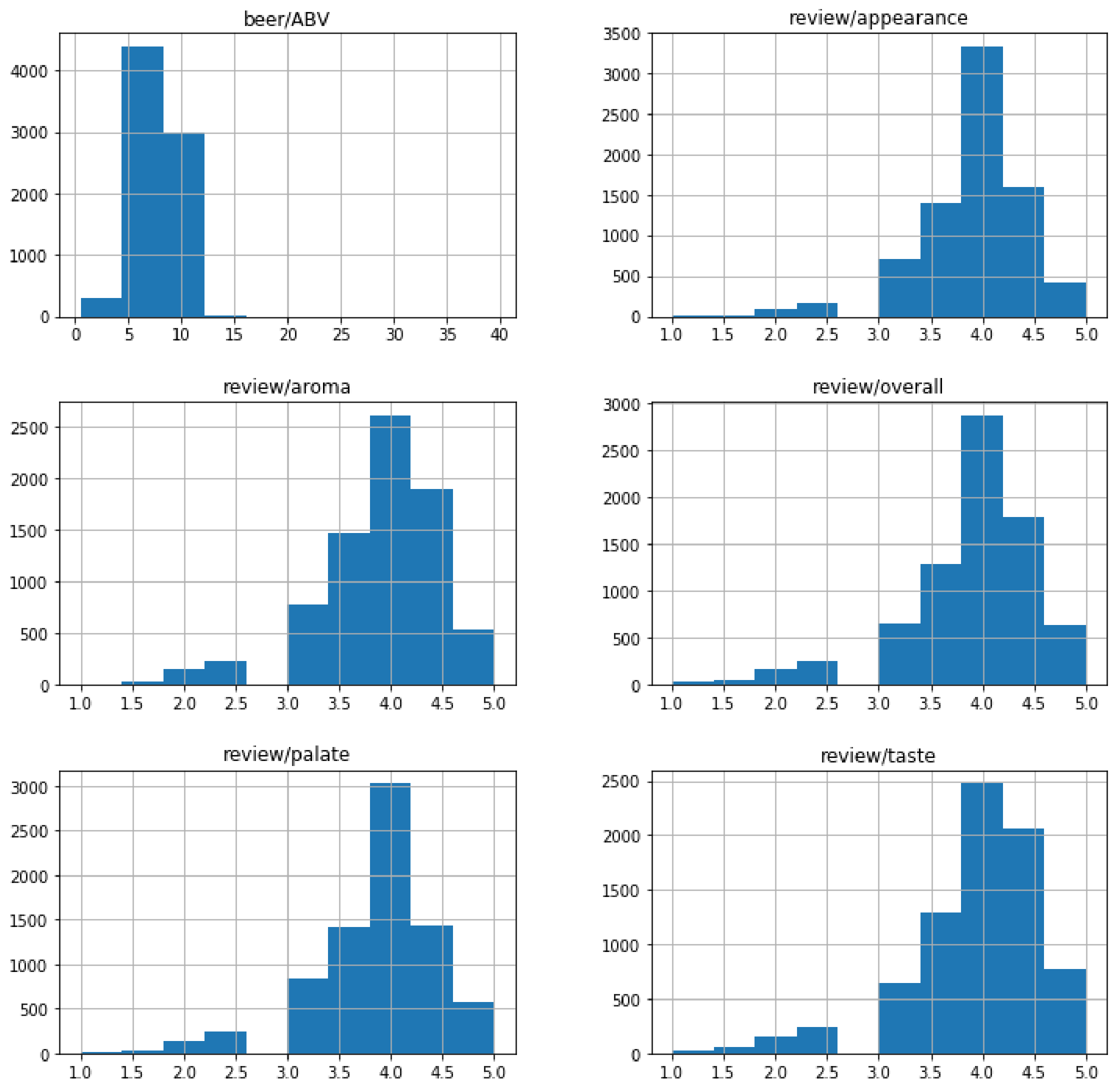
[12]:
```
#Sorting the data
data = data[['beer/ABV', 'beer/name', 'beer/style', 'review/appearance',
       'review/aroma', 'review/overall', 'review/palate', 'review/taste']]
data = data.sort_values(by=['beer/ABV', 'beer/name', 'beer/style','review/
 ↪overall'])
data.head(10)
```

[12]:
```
       beer/ABV                           beer/name        beer/style  \
29458       0.5        Bernard S &#269;istou Hlavou  Low Alcohol Beer
2376        2.2  Harboe Den Glada Danskens LÃ¤ttÃ¶l  Low Alcohol Beer
3180        2.4                     MÃ¸rkt HvidtÃ¸l  Low Alcohol Beer
5249        2.4                            SkibsÃ¸l         Rauchbier
4567        2.4                            SkibsÃ¸l         Rauchbier
22192       2.8             Harboe BjÃ¸rnebryg 2,8%   Euro Pale Lager
15916       2.8                 Harboe Classic 2,8%   Euro Pale Lager
29610       2.8                    Harboe KrÃ¤ftÃ¶l   Euro Pale Lager
35317       2.8                 Harboe Pilsner 2,8%   Euro Pale Lager
27503       2.8              Harboe PÃ¥skebryg 2,8%   Euro Pale Lager

       review/appearance  review/aroma  review/overall  review/palate  \
29458                4.0           3.0             2.0            3.5
2376                 3.5           2.0             2.0            2.5
3180                 3.5           4.0             3.5            4.0
5249                 4.0           4.0             4.5            3.5
4567                 3.5           4.0             5.0            4.0
```

```
22192                3.0           3.0            2.5            2.5
15916                3.0           2.5            3.0            3.0
29610                4.0           3.5            2.0            2.0
35317                2.0           1.5            1.5            2.5
27503                4.0           3.0            3.5            3.5

       review/taste
29458           2.0
2376            2.0
3180            3.5
5249            3.5
4567            4.0
22192           2.0
15916           2.5
29610           1.5
35317           1.5
27503           4.0
```

[13]:

```
#Visualisation
data.hist(figsize=(12,12))
plt.show()
```



[14]:

```
#Let's check the rating part.

data = data[(data['review/overall'] >= 1) | (data['review/appearance'] >= 1)]

# Check it out
data.info
```

[14]:

```
<bound method DataFrame.info of        beer/ABV
beer/name           beer/style  \
29458      0.50        Bernard S &#269;istou Hlavou     Low Alcohol Beer
2376       2.20  Harboe Den Glada Danskens LÃ¤ttÃ¶l     Low Alcohol Beer
3180       2.40                     MÃ¸rkt HvidtÃ¸l     Low Alcohol Beer
5249       2.40                            SkibsÃ¸l            Rauchbier
4567       2.40                            SkibsÃ¸l            Rauchbier
```

```
…         …                                 …                  …
19953     14.50                       Enrico's Cure   English Barleywine
34959     15.00                 Trafalgar Korruptor  American Strong Ale
18800     15.00                 Trafalgar Korruptor  American Strong Ale
4958      15.00                 Trafalgar Korruptor  American Strong Ale
6436      39.44      SchorschbrÃ¤u Schorschbock 40%              Eisbock

       review/appearance  review/aroma  review/overall  review/palate  \
29458                4.0           3.0             2.0            3.5
2376                 3.5           2.0             2.0            2.5
3180                 3.5           4.0             3.5            4.0
5249                 4.0           4.0             4.5            3.5
4567                 3.5           4.0             5.0            4.0
…                  …           …             …            …
19953                3.0           3.5             3.5            3.5
34959                2.0           2.5             2.0            2.0
18800                2.0           2.5             2.5            3.0
4958                 2.5           2.5             3.0            3.0
6436                 3.5           3.5             3.0            3.5

       review/taste
29458           2.0
2376            2.0
3180            3.5
5249            3.5
4567            4.0
…             …
19953           3.5
34959           2.5
18800           3.0
4958            3.5
6436            3.5

[7709 rows x 8 columns]>
```

[15]:

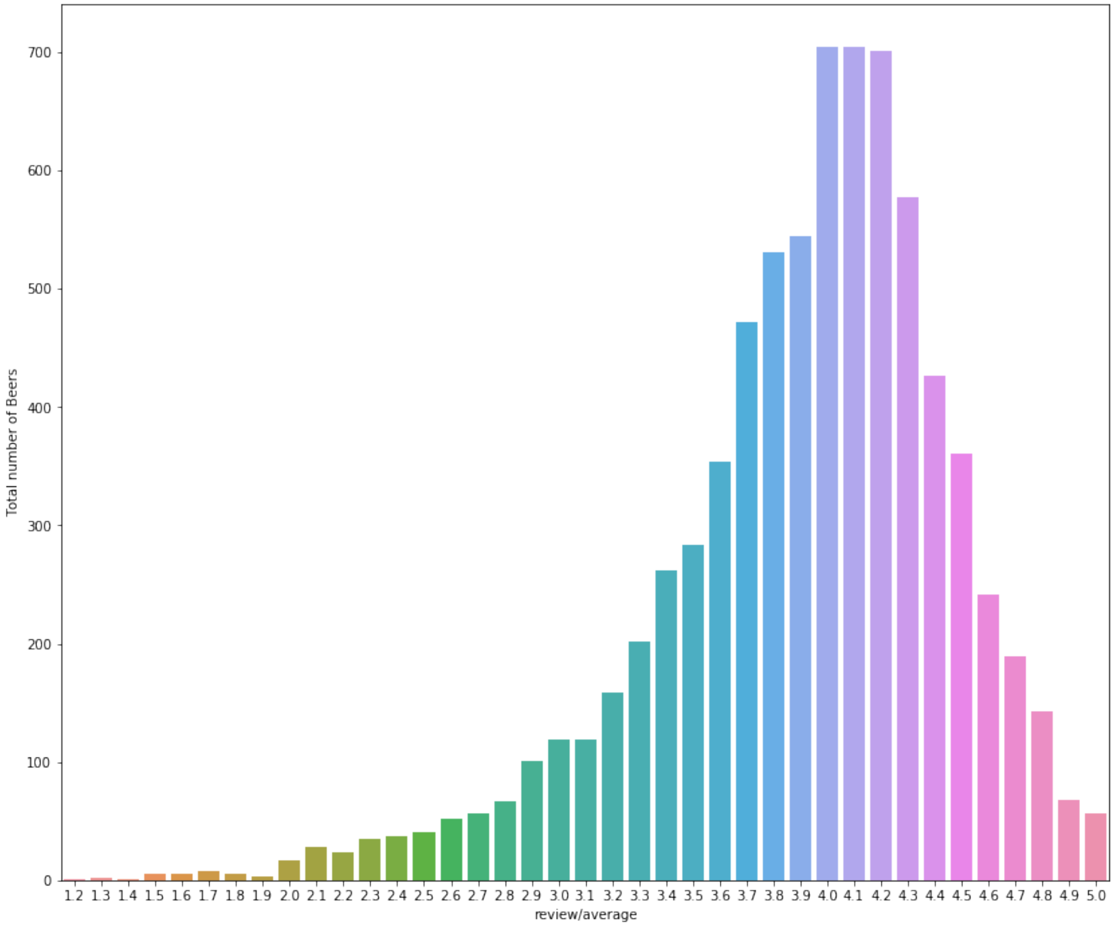
```
#Let's take avaerage review
data['review/average'] = data.apply(lambda row: (row["review/overall"] +␣
 ↪row["review/aroma"] +
                                                 row["review/appearance"] +␣
 ↪row["review/palate"] +
                                                 row["review/taste"]) / 5,␣
 ↪axis=1)

data = data.drop(data[(data["review/average"] < 1) | (data["review/average"] >␣
 ↪5)].index)
data.head()
```

[15]:

```
       beer/ABV                           beer/name        beer/style  \
29458       0.5        Bernard S &#269;istou Hlavou  Low Alcohol Beer
2376        2.2  Harboe Den Glada Danskens LÃ¤ttÃ¶l  Low Alcohol Beer
3180        2.4                     MÃ¸rkt HvidtÃ¸l  Low Alcohol Beer
5249        2.4                            SkibsÃ¸l         Rauchbier
4567        2.4                            SkibsÃ¸l         Rauchbier

       review/appearance  review/aroma  review/overall  review/palate  \
29458                4.0           3.0             2.0            3.5
2376                 3.5           2.0             2.0            2.5
3180                 3.5           4.0             3.5            4.0
5249                 4.0           4.0             4.5            3.5
4567                 3.5           4.0             5.0            4.0

       review/taste  review/average
29458           2.0             2.9
2376            2.0             2.4
3180            3.5             3.7
5249            3.5             3.9
4567            4.0             4.1
```

[16]:

```
#plot average rating graph

plt.figure(figsize=[14, 12])
sns.countplot(x='review/average', data=data, saturation=0.8)
plt.xlabel("review/average")
plt.ylabel("Total number of Beers");
```



[17]:
```
beer_style_taste_abv = data.loc[:,['beer/style','review/taste','review/
 ↪overall','beer/ABV']]

beer_style_taste_abv = beer_style_taste_abv.groupby('beer/style')['review/
 ↪taste','review/overall','beer/ABV'].mean()

beer_style_taste_abv = pd.DataFrame(data=beer_style_taste_abv)

beer_style_taste_abv = beer_style_taste_abv.sort_values(by=['review/
 ↪taste'],ascending=False).reset_index()

beer_style_taste_abv
```

```
/usr/local/lib/python3.7/site-packages/ipykernel_launcher.py:3: FutureWarning:

Indexing with multiple keys (implicitly converted to a tuple of keys) will be
deprecated, use a list instead.
```

```
[17]:                          beer/style  review/taste  review/overall   beer/ABV
      0   American Double / Imperial Stout      4.523220        4.359133   9.557276
      1               English Dark Mild Ale      4.500000        4.500000   3.000000
      2                          Chile Beer      4.500000        4.000000   4.400000
      3                  English Barleywine      4.392857        4.285714  10.475000
      4                   American Wild Ale      4.350000        4.050000   9.155000
      ..                                …             …              …         …
      83                    Euro Pale Lager      2.781818        3.081818   5.010000
      84                  Euro Strong Lager      2.712766        2.797872   8.574468
      85                        Light Lager      2.526316        2.776316   3.530263
      86                    Low Alcohol Beer      2.500000        2.500000   1.700000
      87                 American Malt Liquor      1.916667        2.166667   8.000000

      [88 rows x 4 columns]
```

```
[18]: data.corr()
```

```
[18]:                     beer/ABV  review/appearance  review/aroma  review/overall  \
      beer/ABV            1.000000           0.302264      0.399062        0.201879
      review/appearance   0.302264           1.000000      0.524204        0.467243
      review/aroma        0.399062           0.524204      1.000000        0.597857
      review/overall      0.201879           0.467243      0.597857        1.000000
      review/palate       0.369990           0.539034      0.594771        0.676774
      review/taste        0.369167           0.511796      0.702411        0.770928
      review/average      0.394425           0.716445      0.825123        0.854226

                         review/palate  review/taste  review/average
      beer/ABV                0.369990      0.369167        0.394425
      review/appearance       0.539034      0.511796        0.716445
      review/aroma            0.594771      0.702411        0.825123
      review/overall          0.676774      0.770928        0.854226
      review/palate           1.000000      0.714942        0.850818
      review/taste            0.714942      1.000000        0.899275
      review/average          0.850818      0.899275        1.000000
```

```
[60]: fig = px.scatter(beer_style_taste_abv,x="review/overall",y="beer/
       ↪ABV",trendline="ols")
      fig.show()
```

```
[21]: #Use Linear Model (Model #1)
      linear_model = LinearRegression( normalize = True )
```

```
[22]: #Here review/overall is dependent variable and needs to be predict
      linear_model.fit( X = data[ [ 'review/aroma', 'review/appearance', 'review/
       ↪palate', 'review/taste' ] ], y = data[ 'review/overall' ] )
      preds = linear_model.predict( data[ [ 'review/aroma', 'review/appearance',␣
       ↪'review/palate', 'review/taste' ] ] )
```

[23]:

```
# Coeffifients for each feature (aroma, appearance, palate, taste)
linear_model.coef_
```

[23]:

```
array([0.06079252, 0.03579335, 0.24593985, 0.53487576])
```

[24]:

```
#Accuracy Matrix
np.sqrt( mean_squared_error( data[ 'review/overall' ], preds ) )
```

[24]:

```
0.4256680741781897
```