Project 9 Solutions

Kevin Choe

Collaborators: (Collaborators listed here. Include names, which part of the project you gave or sought help with, and how you helped or were helped.)

TA help: Summeth Guda

Online resources used: (List of links/resources (if any) here. Include web addresses, which part of the project the resource helped with, and how you were helped.)

Question 1

```
import pandas as pd
```

```
import pandas as pd
beers = pd.read_parquet("/class/datamine/data/beer/beers.parquet")
breweries = pd.read_parquet("/class/datamine/data/beer/breweries.parquet")
reviews = pd.read_parquet("/class/datamine/data/beer/reviews.parquet")
def prepare_data(myDF, min_num_donations):
  myDF = myDF.loc[myDF.loc[:, "score"].notna(), :]
  myDF = myDF.loc[myDF.loc[:, "username"].notna(), :]
  myDF = myDF.loc[myDF.loc[:, "beer_id"].notna(), :]
  myDF = myDF.reset_index(drop=True)

  goodusernames = myDF.loc[:, "username"].value_counts() >= min_num_donations
  goodusernames = goodusernames.loc[goodusernames].index.values.tolist()

  goodid = myDF.loc[:, "beer_id"].value_counts() >= min_num_donations
  goodid = goodid.loc[goodid].index.values.tolist()

  myDF = myDF.loc[myDF.loc[:, "username"].isin(goodusernames) & myDF.loc[:, "beer_id"].isin(goodid), :]
  return myDF

train = prepare_data(reviews, 1000)
print(train.shape) # (952105, 10)
```

```
(952105, 10)
```

Question 2

```
def summer(data):
  data['standardized_score'] = ((data['score'] - data['score'].mean())/data['score'].std())
  return data

myresults = train.groupby(["username"]).apply(summer)
print(myresults)
```

```
         beer_id         username  ... score standardized_score
3         125646  GratefulBeerGuy  ...  4.58           1.010719
4         125646         LukeGude  ...  4.31           1.119071
10        125646           jshusc  ...  4.56           1.020464
13        125646      jngrizzaffi  ...  4.53           0.802403
19        125646          Lucular  ...  4.25           0.747185
...          ...              ...  ...   ...                ...
9068902   125646        patre_tim  ...  4.60           1.356303
9068988   125646        GuyFawkes  ...  4.50           0.983790
9069122   125646        kingjohnh  ...  4.51           0.914368
9069124   125646         lemmy187  ...  4.50           1.010520
9069138   125646      mattreitz49  ...  4.40           0.735992

[952105 rows x 11 columns]
```

Question 3

```
score_matrix = pd.pivot_table(myresults,index = 'beer_id', columns = 'username', values = 'standardized

score_matrix.head()
```

```
username  1971bernat  1Sundown2C  22Blue  ...   zimm421  zonker17    zotzot
beer_id                                   ...
5                NaN         NaN     NaN  ... -1.889867       NaN       NaN
6          -0.472504   -0.240035     NaN  ...       NaN -1.005231       NaN
7          -1.675515         NaN     NaN  ...       NaN       NaN       NaN
10               NaN         NaN     NaN  ...       NaN -1.005231 -0.820021
17         -0.472504         NaN     NaN  ...       NaN       NaN       NaN

[5 rows x 1824 columns]
```

```
type(score_matrix)
```

```
<class 'pandas.core.frame.DataFrame'>
```

```
score_matrix.shape
```

```
(1469, 1824)
```

Question 4

```
myresults = score_matrix.mean(axis=0)
score_matrix.fillna(value=myresults)
```

```
username    1971bernat    1Sundown2C  ...      zonker17       zotzot
beer_id                               ...
5       -1.506138e-16  7.346075e-16  ...  1.761608e-15 -2.842171e-15
6       -4.725043e-01 -2.400347e-01  ... -1.005231e+00 -2.842171e-15
7       -1.675515e+00  7.346075e-16  ...  1.761608e-15 -2.842171e-15
10      -1.506138e-16  7.346075e-16  ... -1.005231e+00 -8.200206e-01
17      -4.725043e-01  7.346075e-16  ...  1.761608e-15 -2.842171e-15
...               ...           ...  ...           ...          ...
202078  -1.506138e-16  5.353896e-01  ...  1.761608e-15 -2.842171e-15
211516  -1.506138e-16  7.346075e-16  ...  1.761608e-15 -2.842171e-15
213281  -1.506138e-16  7.346075e-16  ...  1.761608e-15 -2.842171e-15
221289  -1.506138e-16  7.346075e-16  ...  1.761608e-15 -2.842171e-15
```

```
221843   -1.506138e-16  7.346075e-16  ...  1.761608e-15  5.333365e-01

[1469 rows x 1824 columns]
```

Pledge

By submitting this work I hereby pledge that this is my own, personal work. I've acknowledged in the designated place at the top of this file all sources that I used to complete said work, including but not limited to: online resources, books, and electronic communications. I've noted all collaboration with fellow students and/or TA's. I did not copy or plagiarize another's work.

> As a Boilermaker pursuing academic excellence, I pledge to be honest and true in all that I do. Accountable together – We are Purdue.