Solution

(2) Case Study: Source and investigate usable data sources

(GenAI Life Cycle Phase 2: Data Understanding self-practice)

Note on EDA

When performing Exploratory Data Analysis (EDA), the specific techniques used are less important than the overall goal: gaining a solid understanding of the data before transforming or modeling it. The primary objective of EDA is to uncover insights about the structure, quality, and relationships within the dataset to inform subsequent steps effectively.

- Load the head of each file to view the first few entries of each dataset.

In [26]:
```
import pandas as pd

# Load datasets
df_business = pd.read_csv('yelp_academic_dataset_business.csv')
df_review = pd.read_csv('yelp_academic_dataset_review.csv')
df_user = pd.read_csv('yelp_academic_dataset_user.csv')

# Perform EDA
print("Business Dataset Head:")
print(df_business.head())
print("\nReview Dataset Head:")
print(df_review.head())
print("\nUser Dataset Head:")
print(df_user.head())
```

```
Business Dataset Head:
              business_id                          name  \
0  MTSW4McQd7CbVtyjqoe9mw            St Honore Pastries
1  bBDDEgkFA1Otx9Lfe7BZUQ                Sonic Drive-In
2  WKMJwqnfZKsAae75RMP6jA  Roast Coffeehouse and Wine Bar
3  tMkwHmWFUEXrC9ZduonpTg            The Green Pheasant
4  aPNXGTDkf-4bjhyMBQxqpQ                   Craft Hall

                address          city state postal_code   latitude   longitude  \
0          935 Race St  Philadelphia    PA       19107  39.955505  -75.155564
1  2312 Dickerson Pike     Nashville    TN       37207  36.208102  -86.768170
2  10359 104 Street NW      Edmonton    AB     T5J 1B9  53.546045 -113.499169
3        215 1st Ave S     Nashville    TN       37201  36.159886  -86.773197
4   901 N Delaware Ave  Philadelphia    PA       19123  39.962582  -75.135657

   stars  review_count  is_open  \
0    4.0            10        1
1    1.5             1        1
2    4.0             5        0
3    4.0             4        0
4    3.5             1        1

                                      attributes  \
0  {'RestaurantsDelivery': 'False', 'OutdoorSeati...
1  {'RestaurantsAttire': "'casual'", 'Restaurants...
2  {'OutdoorSeating': 'False', 'Caters': 'True', ...
3  {'RestaurantsGoodForGroups': 'True', 'HappyHou...
4  {'OutdoorSeating': 'True', 'RestaurantsPriceRa...

                                      categories  \
0  Restaurants, Food, Bubble Tea, Coffee & Tea, B...
1  Ice Cream & Frozen Yogurt, Fast Food, Burgers,...
2  Coffee & Tea, Food, Cafes, Bars, Wine Bars, Re...
3                  Restaurants, Japanese, Seafood
4  Eatertainment, Arts & Entertainment, Brewpubs,...

                                               hours
0  {'Monday': '7:0-20:0', 'Tuesday': '7:0-20:0', ...
1  {'Monday': '0:0-0:0', 'Tuesday': '6:0-21:0', '...
2  {'Monday': '8:0-18:0', 'Tuesday': '8:0-18:0', ...
3  {'Wednesday': '16:0-22:0', 'Thursday': '16:0-2...
4  {'Monday': '0:0-0:0', 'Wednesday': '16:0-22:0'...

Review Dataset Head:
                review_id                 user_id             business_id  \
0  vBK79c3_1Ff_oqkh5VpfGg  Ohhrhu1RkqfVciIVx_W5HQ  nRKndeZLQ3eDL10UMwS2rQ
1  p198qZsKOMCUhgdtRWsOKQ  3MpDvy5gEdsbZh9-p92dHg  8QnuWGVNBhzyYXGSeRdi4g
2  -7LkjSPzfVgnVpuVuRuOow  uAu772KpSkb-tPFgZmU-lA  2GYg3liJ9-m6Z67L_4_BRQ
3  5M9fBQINWamyQPdJ7OQkeA  kHmXUEOAsIbguUrPoX6Tkg  iD25g7vUG3asELjXtMGF9w
4  onlgwy5qGDEzddsrnIvtWg  pYXeL0RCqus2IfhthYCOyA  W7NxQw8UYFR0HLPrI08tvw

   stars  useful  funny  cool  \
0    5.0     2.0    5.0   2.0
1    4.0     0.0    0.0   0.0
2    5.0     7.0    0.0   3.0
3    2.0     2.0    2.0   1.0
4    4.0     0.0    0.0   0.0

                                                text                 date
0  HOLY SMOKES!\n\nactual pumpkin pie mixed in wi...  2009-10-13 19:49:51
1  After my ROTD  yesterday of a different Sweet ...  2013-10-24 19:24:33
2  I LOVE Weaver's Way and really disagree with s...  2008-12-03 04:13:43
3  Meh is all I can say for TDS. The restaurant i...  2012-12-04 02:14:36
4  Don't know what it is but If my tummy's feelin...  2012-02-01 14:21:25

User Dataset Head:
                  user_id    name  review_count        yelping_since  useful  \
0  qVc8ODYU5SZjKXVBgXdI7w  Walker             2  2007-01-25 16:47:26    7217
1  j14WgRoU_-2ZE1aw1dXrJg  Daniel            11  2009-01-25 04:35:42   43091
2  2WnXYQFK0hXEoTxPtV2zvg   Steph             3  2008-07-25 10:41:00    2086
3  SZDeASXq7o05mMNLshsdIA    Gwen             3  2005-11-29 04:38:33     512
4  hA5lMy-EnncsH4JoR-hFGQ   Karen             3  2007-01-05 19:40:59      29

   funny   cool                                               elite  \
0   1259   5994                                                2007
1  13066  27281  2009,2010,2011,2012,2013,2014,2015,2016,2017,2...
2   1010   1003                            2009,2010,2011,2012,2013
3    330    299                                 2009,2010,2011
4     15      7                                                NaN
```

```
                                              friends  fans  ...  \
0  NSCy54eWehBJyZdG2iE84w, pe42u7DcCH2QmI81NX-8qA...   267  ...
1  ueRPE0CX75ePGMqOFVj6IQ, 52oH4DrRvzzl8wh5UXyU0A...  3138  ...
2  LuO3Bn4f3rlhyHIaNfTlnA, j9B4XdHUhDfTKVecyWQgyA...    52  ...
3  enx1vVPnfdNUdPho6PH_wg, 4wOcvMLtU6a9Lslggq74Vg...    28  ...
4  PBK4q9KEEBHhFvSXCUirIw, 3FWPpM7KU1gXeOM_ZbYMbA...     1  ...

   compliment_more  compliment_profile  compliment_cute  compliment_list  \
0               65                  55               56               18
1              264                 184              157              251
2               13                  10               17                3
3                4                   1                6                2
4                1                   0                0                0

   compliment_note  compliment_plain  compliment_cool  compliment_funny  \
0              232               844              467               467
1             1847              7054             3131              3131
2               66                96              119               119
3               12                16               26                26
4                1                 1                0                 0

   compliment_writer  compliment_photos
0                239                180
1               1521               1946
2                 35                 18
3                 10                  9
4                  0                  0

[5 rows x 22 columns]
```

In [27]:
```
print("Business Dataset Head:")
df_business.head()
```

```
Business Dataset Head:
```

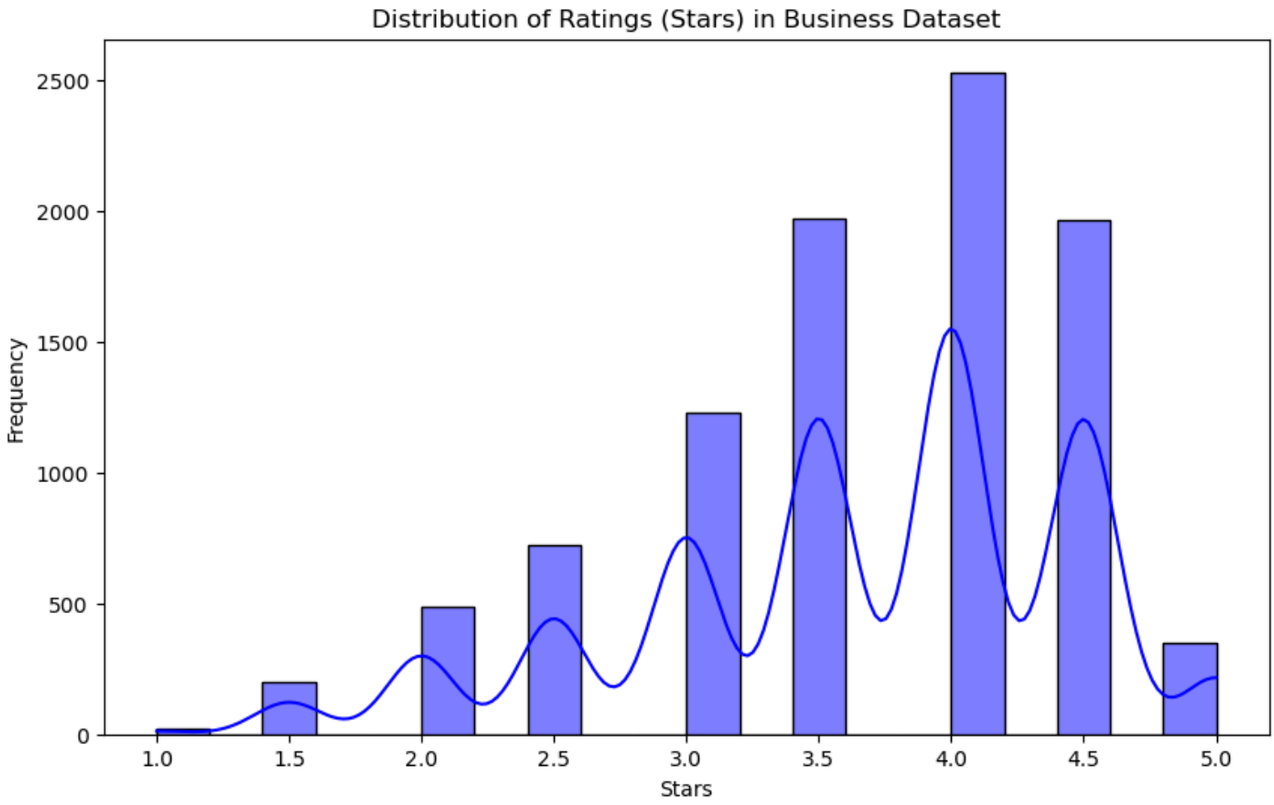
Out[27]:

	business_id	name	address	city	state	postal_code	latitude
0	MTSW4McQd7CbVtyjqoe9mw	St Honore Pastries	935 Race St	Philadelphia	PA	19107	39.955505
1	bBDDEgkFA1Otx9Lfe7BZUQ	Sonic Drive-In	2312 Dickerson Pike	Nashville	TN	37207	36.208102
2	WKMJwqnfZKsAae75RMP6jA	Roast Coffeehouse and Wine Bar	10359 104 Street NW	Edmonton	AB	T5J 1B9	53.546045
3	tMkwHmWFUEXrC9ZduonpTg	The Green Pheasant	215 1st Ave S	Nashville	TN	37201	36.159886
4	aPNXGTDkf-4bjhyMBQxqpQ	Craft Hall	901 N Delaware Ave	Philadelphia	PA	19123	39.962582

1. First file: `yelp_academic_dataset_business.csv`

In [28]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Load business dataset
df_business = pd.read_csv('yelp_academic_dataset_business.csv')

# Distribution of Ratings (Stars)
plt.figure(figsize=(10, 6))
sns.histplot(df_business['stars'], bins=20, kde=True, color='blue')
plt.title('Distribution of Ratings (Stars) in Business Dataset')
plt.xlabel('Stars')
```

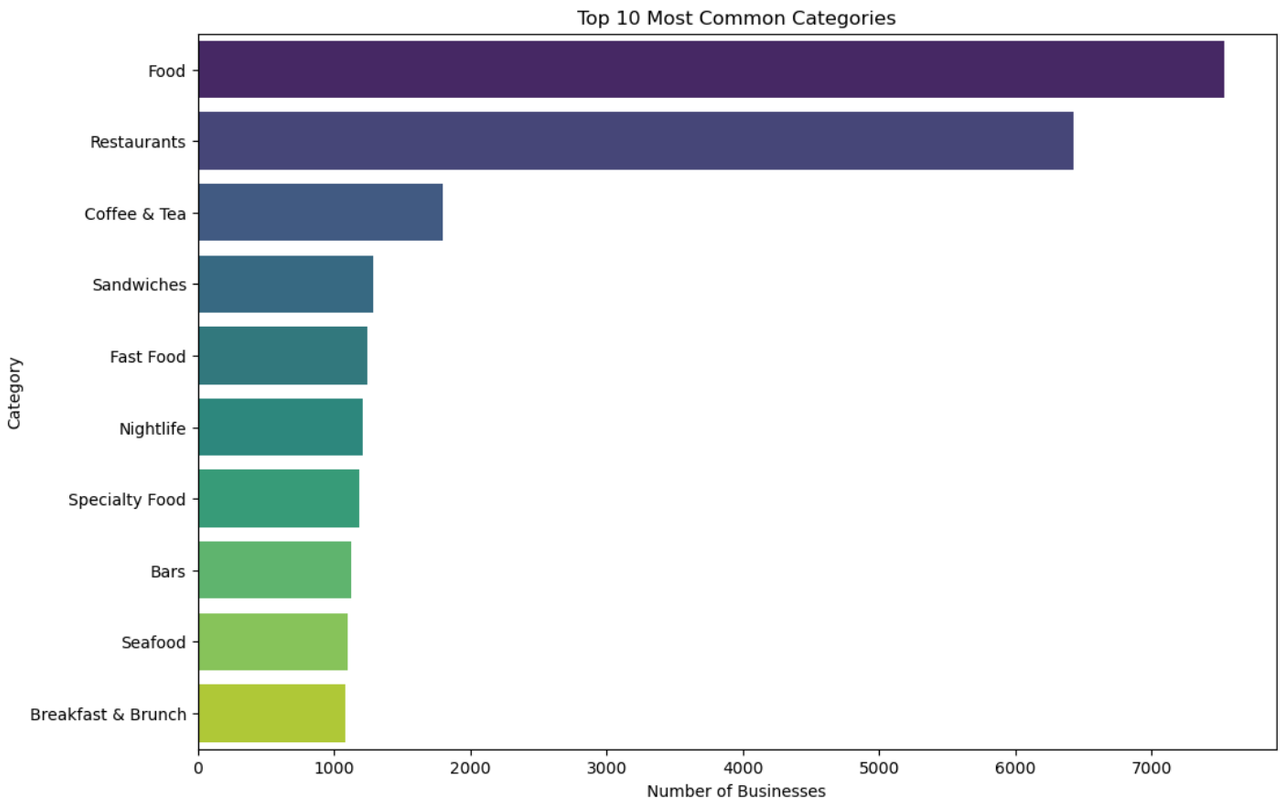
```
plt.ylabel('Frequency')
plt.show()
```



In [29]:
```
print("\nReview Dataset Head:")
print(df_review.head())
```

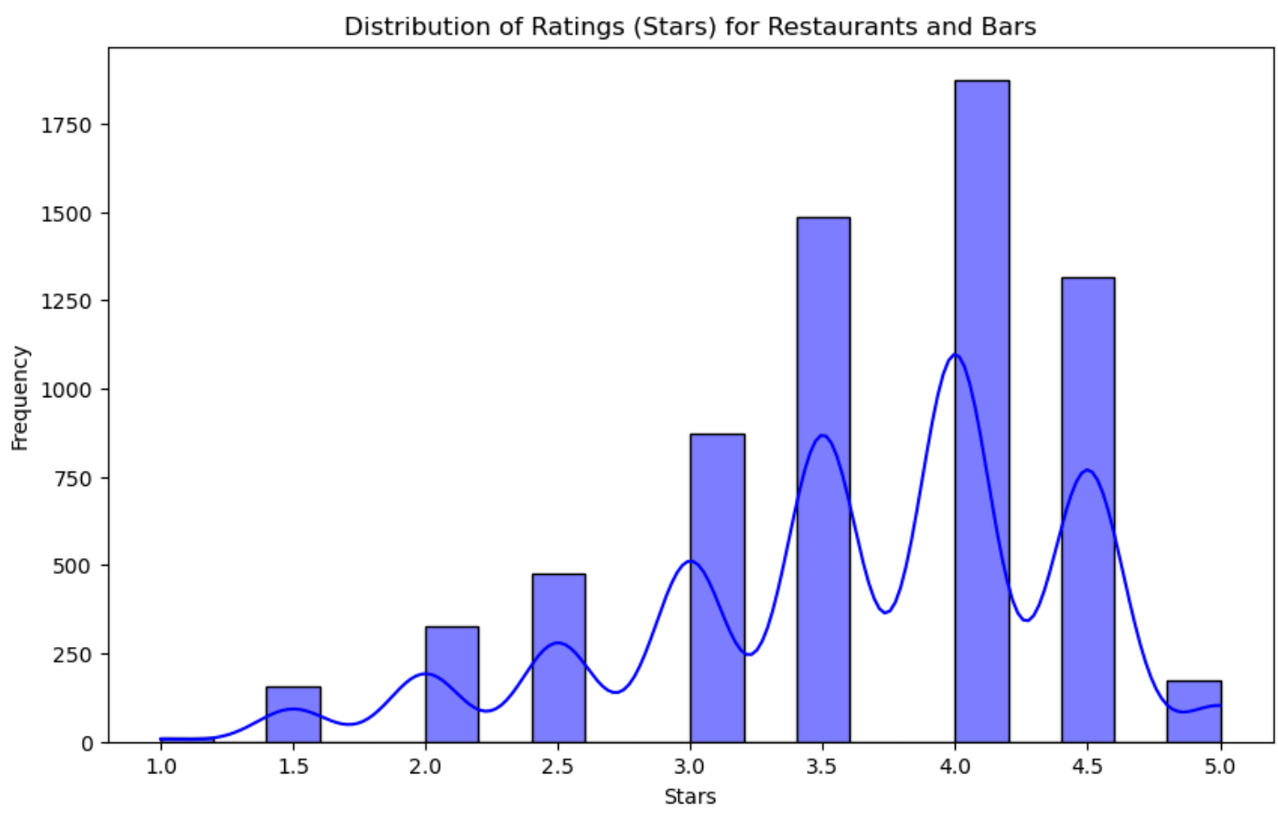
```
Review Dataset Head:
                review_id                 user_id             business_id  \
0  vBK79c3_1Ff_oqkh5VpfGg  Ohhrhu1RkqfVciIVx_W5HQ  nRKndeZLQ3eDL10UMwS2rQ
1  p198qZsKOMCUhgdtRWsOKQ  3MpDvy5gEdsbZh9-p92dHg  8QnuWGVNBhzyYXGSeRdi4g
2  -7LkjSPzfVgnVpuVuRuOow  uAu772KpSkb-tPFgZmU-lA  2GYg3liJ9-m6Z67L_4_BRQ
3  5M9fBQINWamyQPdJ7OQkeA  kHmXUEOAsIbguUrPoX6Tkg  iD25g7vUG3asELjXtMGF9w
4  onlgwy5qGDEzddsrnIvtWg  pYXeL0RCqus2IfhthYCOyA  W7NxQw8UYFR0HLPrI08tvw

   stars  useful  funny  cool  \
0    5.0     2.0    5.0   2.0
1    4.0     0.0    0.0   0.0
2    5.0     7.0    0.0   3.0
3    2.0     2.0    2.0   1.0
4    4.0     0.0    0.0   0.0

                                                text                 date
0  HOLY SMOKES!\n\nactual pumpkin pie mixed in wi...  2009-10-13 19:49:51
1  After my ROTD  yesterday of a different Sweet ...  2013-10-24 19:24:33
2  I LOVE Weaver's Way and really disagree with s...  2008-12-03 04:13:43
3  Meh is all I can say for TDS. The restaurant i...  2012-12-04 02:14:36
4  Don't know what it is but If my tummy's feelin...  2012-02-01 14:21:25
```

In [30]:
```
# Extract top categories from the categories column
df_business['categories'] = df_business['categories'].fillna('')
categories = df_business['categories'].str.split(',').explode().str.strip()

# Count the occurrences of each category
category_counts = categories.value_counts().head(10)

# Plot the top 10 categories
plt.figure(figsize=(12, 8))
sns.barplot(x=category_counts.values, y=category_counts.index, palette='viridis')
plt.title('Top 10 Most Common Categories')
plt.xlabel('Number of Businesses')
plt.ylabel('Category')
plt.show()
```



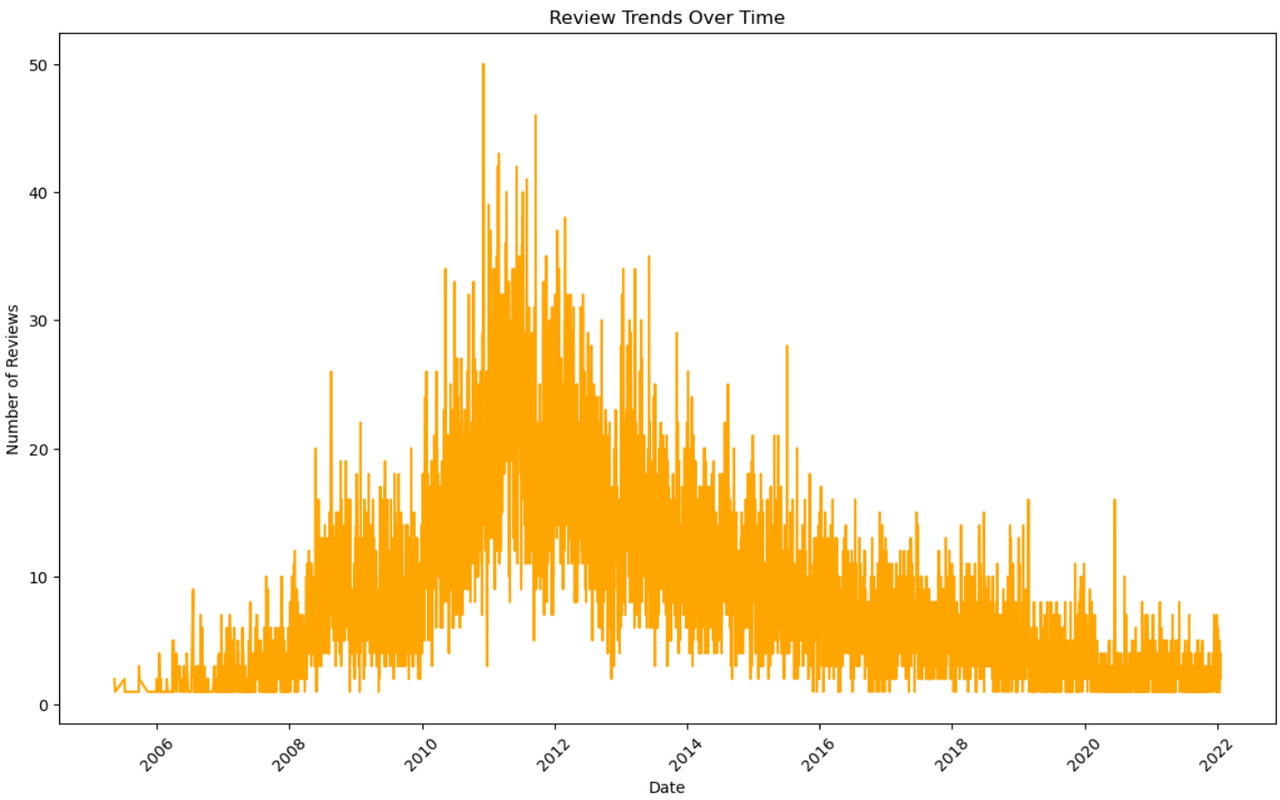
- The scope of this project encompasses restaurants and bars so let's revisit 'Distribution of Ratings (Stars) in Business Dataset' focused on those two categories

In [31]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Load business dataset
df_business = pd.read_csv('yelp_academic_dataset_business.csv')

# Filter for Restaurants and Bars
df_filtered = df_business[df_business['categories'].str.contains('Restaurants|Bars', na=

# Distribution of Ratings (Stars) for Restaurants and Bars
plt.figure(figsize=(10, 6))
sns.histplot(df_filtered['stars'], bins=20, kde=True, color='blue')
plt.title('Distribution of Ratings (Stars) for Restaurants and Bars')
plt.xlabel('Stars')
plt.ylabel('Frequency')
plt.show()
```



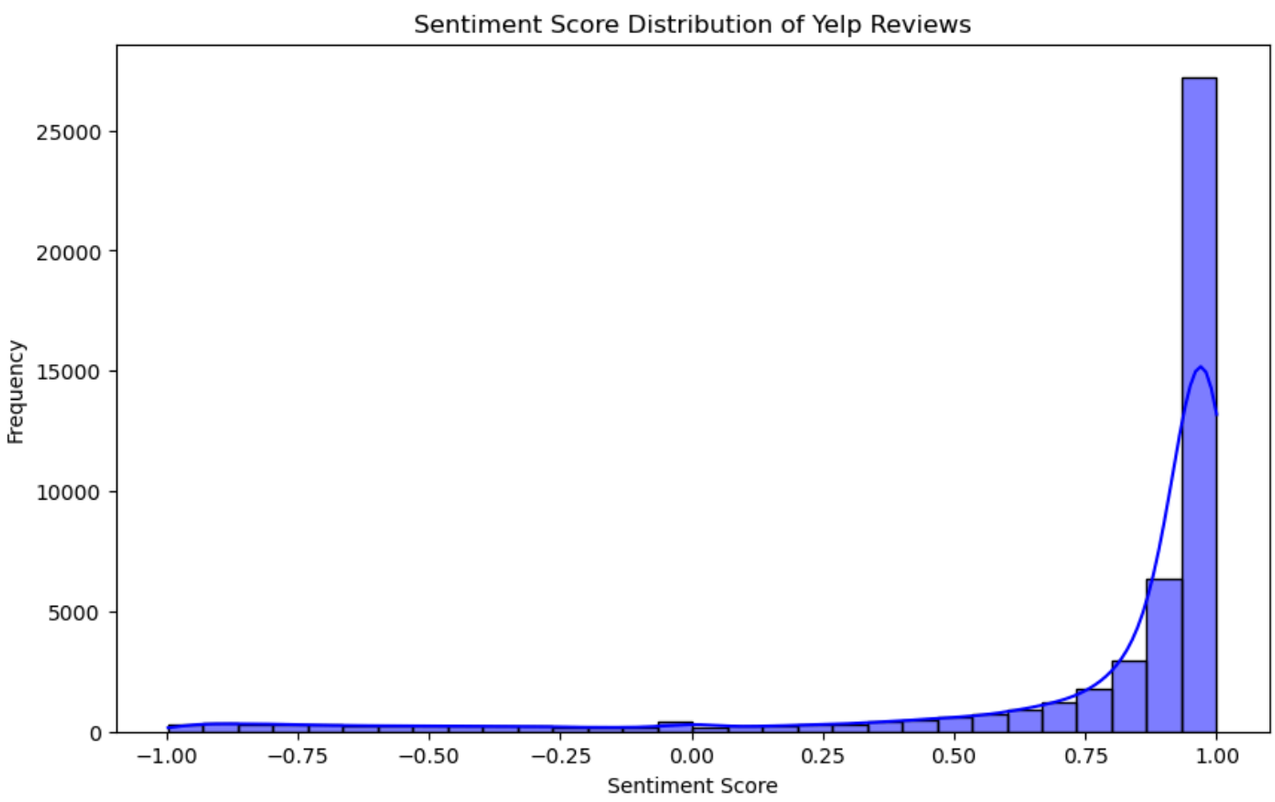
2. Second file: `yelp_academic_dataset_review.csvv`

In [32]:
```
# Load review dataset
df_reviews = pd.read_csv('yelp_academic_dataset_review.csv')

# Convert date to datetime format
df_reviews['date'] = pd.to_datetime(df_reviews['date'])

# Group by date and count reviews
reviews_by_date = df_reviews.groupby(df_reviews['date'].dt.date).size()

# Line plot of reviews over time
plt.figure(figsize=(14, 8))
reviews_by_date.plot(kind='line', color='orange')
plt.title('Review Trends Over Time')
plt.xlabel('Date')
plt.ylabel('Number of Reviews')
plt.xticks(rotation=45)
plt.show()
```



- We can look at the sentiment scores of the reviews

In [33]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import nltk
from nltk.sentiment.vader import SentimentIntensityAnalyzer

# Download VADER Lexicon
nltk.download('vader_lexicon', quiet=True)

# Initialize Sentiment Analyzer
sid = SentimentIntensityAnalyzer()

# Load Reviews Dataset
df_reviews = pd.read_csv('yelp_academic_dataset_review.csv')

# Sentiment Analysis on Review Text
df_reviews['sentiment_score'] = df_reviews['text'].apply(lambda x: sid.polarity_scores(x

# Plot Sentiment Score Distribution
plt.figure(figsize=(10, 6))
sns.histplot(df_reviews['sentiment_score'], bins=30, kde=True, color='blue')
plt.title("Sentiment Score Distribution of Yelp Reviews")
plt.xlabel("Sentiment Score")
plt.ylabel("Frequency")
plt.show()
```

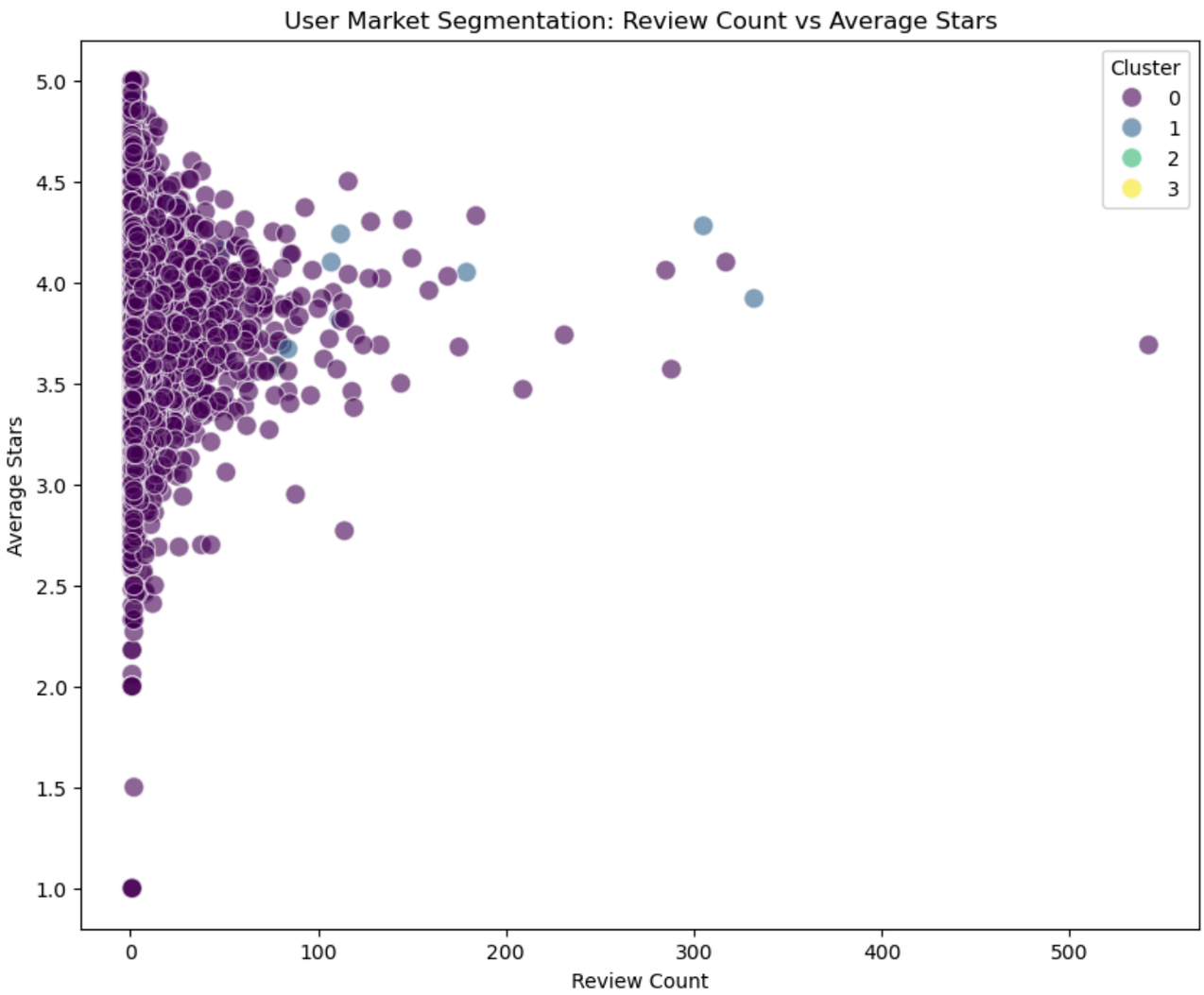


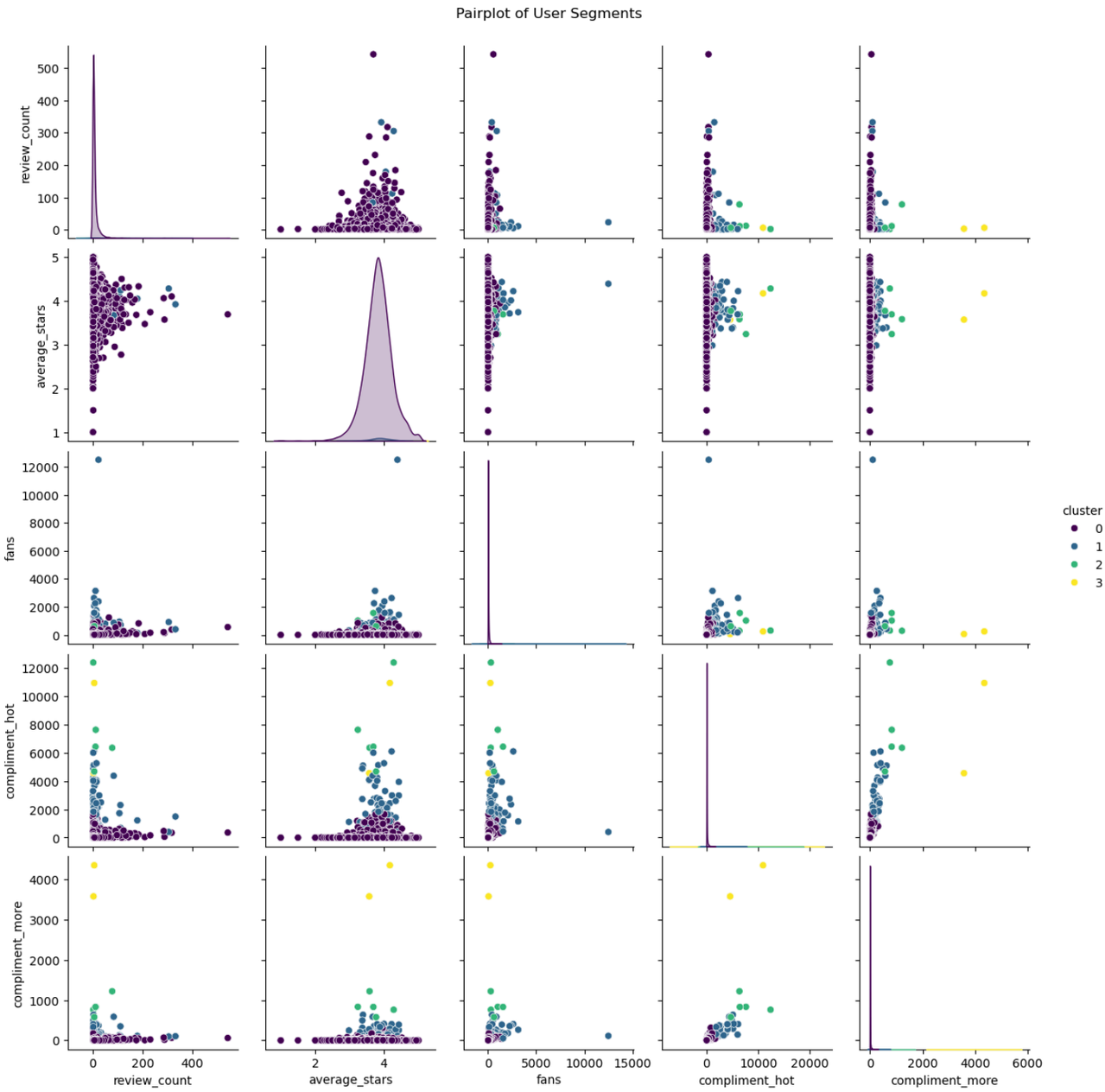
3. Third file: `yelp_academic_dataset_review.csvv`

We hadn't dealt with users in the Practice Learning Activity of Chapter 2. One useful way to get an overview of users is to segment into groups them by their different information.

In [34]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans

# Load user dataset
df_users = pd.read_csv('yelp_academic_dataset_user.csv')

# Feature selection: Choose relevant features for market segmentation
features = ['review_count', 'average_stars', 'fans',
            'compliment_hot', 'compliment_more', 'compliment_profile',
            'compliment_cute', 'compliment_list', 'compliment_note',
            'compliment_plain', 'compliment_cool', 'compliment_funny',
            'compliment_writer', 'compliment_photos']

# Fill missing values (if any)
df_users[features] = df_users[features].fillna(0)

# Normalize/Scale the data
scaler = StandardScaler()
df_users_scaled = scaler.fit_transform(df_users[features])

# Perform K-means clustering
kmeans = KMeans(n_clusters=4, random_state=42)  # Change the number of clusters based on
df_users['cluster'] = kmeans.fit_predict(df_users_scaled)

# Visualize the segmentation
plt.figure(figsize=(10, 8))
sns.scatterplot(x=df_users['review_count'], y=df_users['average_stars'], hue=df_users['c
plt.title('User Market Segmentation: Review Count vs Average Stars')
plt.xlabel('Review Count')
plt.ylabel('Average Stars')
plt.legend(title='Cluster')
plt.show()

# Pairplot to visualize clusters
sns.pairplot(df_users[['review_count', 'average_stars', 'fans', 'compliment_hot', 'compl
plt.suptitle("Pairplot of User Segments", y=1.02)
plt.show()

# Display cluster centers
centers = pd.DataFrame(kmeans.cluster_centers_, columns=features)
print("Cluster Centers (User Segments):")
```

```
print(centers)

# Segment Analysis
for i in range(4):  # Change the number of clusters here if necessary
    print(f"\nCluster {i} Summary:")
    segment = df_users[df_users['cluster'] == i]
    print(segment[features].describe())
```




Pairplot of User Segments
review_count
average_stars
fans
compliment_hot
compliment_more
cluster
0
1
2
3

```
Cluster Centers (User Segments):
   review_count  average_stars      fans  compliment_hot  compliment_more  \
0     -0.011927      -0.001725 -0.057776       -0.087571        -0.056704
1      0.824785       0.140029  3.961726        4.773181         2.283638
2      0.725652      -0.280178  3.451124       17.692338         9.992213
3     -0.197268       0.115912  0.529222       18.294417        47.386677

   compliment_profile  compliment_cute  compliment_list  compliment_note  \
0           -0.045901        -0.067744        -0.048124        -0.089477
1            1.529439         3.299715         1.683253         5.071195
2            8.407126        11.785566         9.484763        10.785302
3           50.168645        34.420760        47.769957        29.256047

   compliment_plain  compliment_cool  compliment_funny  compliment_writer  \
0         -0.087325        -0.091380         -0.091380          -0.081437
1          4.977924         5.001773          5.001773           4.257068
2         17.922934        19.834496         19.834496          16.429109
3          8.925274        14.830815         14.830815          24.252477

   compliment_photos
0          -0.064030
1           2.972530
2          24.822127
3           4.104824

Cluster 0 Summary:
       review_count  average_stars         fans  compliment_hot  \
count   5734.000000    5734.000000  5734.000000     5734.000000
mean       7.974189       3.823075    31.576212       47.013777
std       17.597136       0.400220    67.981285      134.926837
min        1.000000       1.000000     0.000000        0.000000
25%        1.000000       3.600000     3.000000        1.000000
50%        3.000000       3.820000    10.000000        6.000000
75%        7.000000       4.050000    32.000000       28.000000
max      542.000000       5.000000  1377.000000     1649.000000

       compliment_more  compliment_profile  compliment_cute  compliment_list  \
count      5734.000000         5734.000000      5734.000000      5734.000000
mean          6.630450            4.049878         4.223404         2.453261
std          14.272741           11.729551        15.146029         8.199128
min           0.000000            0.000000         0.000000         0.000000
25%           0.000000            0.000000         0.000000         0.000000
50%           2.000000            1.000000         0.000000         0.000000
75%           7.000000            3.000000         2.000000         1.000000
max         314.000000          186.000000       455.000000       231.000000

       compliment_note  compliment_plain  compliment_cool  compliment_funny  \
count      5734.000000       5734.000000      5734.000000       5734.000000
mean         32.179805         67.080223        66.837461         66.837461
std          81.123413        214.360801       173.861142        173.861142
min           0.000000          0.000000         0.000000          0.000000
25%           2.000000          2.000000         2.000000          2.000000
50%           7.000000          9.000000        10.000000         10.000000
75%          25.000000         38.000000        45.000000         45.000000
max        1180.000000       4084.000000      2222.000000       2222.000000

       compliment_writer  compliment_photos
count        5734.000000        5734.000000
mean           26.630624          14.023195
std            66.624955          59.689203
min             0.000000           0.000000
25%             1.000000           0.000000
50%             5.000000           1.000000
75%            21.000000           6.000000
max           964.000000        1434.000000

Cluster 1 Summary:
       review_count  average_stars          fans  compliment_hot  \
count     79.000000      79.000000     79.000000       79.000000
mean      23.658228       3.879620    866.455696     2084.708861
std       56.108157       0.293849   1461.057115     1373.036193
min        1.000000       2.980000    106.000000      400.000000
25%        3.000000       3.715000    313.500000     1153.000000
50%        7.000000       3.870000    498.000000     1646.000000
75%       15.000000       4.055000    929.000000     2659.000000
max      332.000000       4.470000  12497.000000     6104.000000

       compliment_more  compliment_profile  compliment_cute  compliment_list  \
count        79.000000           79.000000        79.000000        79.000000
```

```
mean        201.696203          203.151899       168.139241        90.493671
std         118.140871          154.189368       158.522564        74.920620
min          51.000000           28.000000         7.000000         1.000000
25%         117.000000           92.500000        71.000000        33.000000
50%         169.000000          156.000000       118.000000        75.000000
75%         264.000000          285.500000       215.500000       130.500000
max         636.000000          827.000000       963.000000       360.000000

       compliment_note  compliment_plain  compliment_cool  compliment_funny  \
count        79.000000         79.000000        79.000000         79.000000
mean       1215.063291       3905.341772      2900.822785       2900.822785
std         628.790462       2580.459132      1565.999736       1565.999736
min         261.000000        685.000000       836.000000        836.000000
25%         802.000000       2222.000000      1800.500000       1800.500000
50%        1073.000000       3088.000000      2505.000000       2505.000000
75%        1479.500000       4926.000000      3417.500000       3417.500000
max        3190.000000      16674.000000      9371.000000       9371.000000

       compliment_writer  compliment_photos
count          79.000000          79.000000
mean         1031.987342         963.101266
std           793.907885         944.995694
min           102.000000           9.000000
25%           540.000000         282.500000
50%           844.000000         563.000000
75%          1198.500000        1445.500000
max          5811.000000        4824.000000

Cluster 2 Summary:
       review_count  average_stars         fans  compliment_hot  \
count      5.000000       5.000000     5.000000        5.000000
mean      21.800000       3.712000   760.400000     7500.600000
std       31.673333       0.376391   537.990056     2928.346598
min        2.000000       3.240000   290.000000     4684.000000
25%        6.000000       3.580000   308.000000     6360.000000
50%       11.000000       3.690000   620.000000     6436.000000
75%       12.000000       3.770000  1021.000000     7632.000000
max       78.000000       4.280000  1563.000000    12391.000000

       compliment_more  compliment_profile  compliment_cute  compliment_list  \
count         5.000000            5.000000         5.000000         5.000000
mean        844.200000         1072.400000       581.200000       487.200000
std         234.451914          210.691243       535.434123       383.589885
min         577.000000          873.000000       106.000000       255.000000
25%         761.000000          936.000000       212.000000       278.000000
50%         830.000000          983.000000       263.000000       327.000000
75%         833.000000         1183.000000      1101.000000       411.000000
max        1220.000000         1387.000000      1224.000000      1165.000000

       compliment_note  compliment_plain  compliment_cool  compliment_funny  \
count         5.000000            5.0000         5.000000          5.000000
mean       2524.800000        13714.6000     11154.200000      11154.200000
std         405.671296         9339.5609      2073.045151       2073.045151
min        2236.000000         5906.0000      8258.000000       8258.000000
25%        2238.000000         6790.0000     10235.000000      10235.000000
50%        2340.000000        11581.0000     10994.000000      10994.000000
75%        2614.000000        15322.0000     13022.000000      13022.000000
max        3196.000000        28974.0000     13262.000000      13262.000000

       compliment_writer  compliment_photos
count           5.000000           5.000000
mean         3852.600000        7792.200000
std          2118.432274        5092.532837
min          1849.000000        2296.000000
25%          2555.000000        3804.000000
50%          2613.000000        6829.000000
75%          5672.000000       11987.000000
max          6574.000000       14045.000000

Cluster 3 Summary:
       review_count  average_stars        fans  compliment_hot  \
count       2.00000       2.000000    2.000000        2.000000
mean        4.50000       3.870000  153.500000     7753.000000
std         2.12132       0.424264  132.228968     4512.755478
min         3.00000       3.570000   60.000000     4562.000000
25%         3.75000       3.720000  106.750000     6157.500000
50%         4.50000       3.870000  153.500000     7753.000000
75%         5.25000       4.020000  200.250000     9348.500000
max         6.00000       4.170000  247.000000    10944.000000
```

```
       compliment_more  compliment_profile  compliment_cute  compliment_list  \
count         2.000000            2.000000         2.000000         2.000000
mean       3961.000000         6350.500000      1683.000000      2434.000000
std         545.886435          973.686038        86.267027       244.658946
min        3575.000000         5662.000000      1622.000000      2261.000000
25%        3768.000000         6006.250000      1652.500000      2347.500000
50%        3961.000000         6350.500000      1683.000000      2434.000000
75%        4154.000000         6694.750000      1713.500000      2520.500000
max        4347.000000         7039.000000      1744.000000      2607.000000

       compliment_note  compliment_plain  compliment_cool  compliment_funny  \
count         2.000000          2.000000         2.000000          2.000000
mean       6758.500000       6896.500000      8370.000000       8370.000000
std        2626.901692       6872.370806      6943.788591       6943.788591
min        4901.000000       2037.000000      3460.000000       3460.000000
25%        5829.750000       4466.750000      5915.000000       5915.000000
50%        6758.500000       6896.500000      8370.000000       8370.000000
75%        7687.250000       9326.250000     10825.000000      10825.000000
max        8616.000000      11756.000000     13280.000000      13280.000000

       compliment_writer  compliment_photos
count            2.00000           2.000000
mean          5665.50000        1317.000000
std           2324.25999          46.669048
min           4022.00000        1284.000000
25%           4843.75000        1300.500000
50%           5665.50000        1317.000000
75%           6487.25000        1333.500000
max           7309.00000        1350.000000
```

- You may customize the code above to try and group the users by different segments.