Importing Libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import scipy.stats as stats
```

Data Acquisition and Cleaning

```
d1 = pd.read_csv("all_reviews.csv")
```

```
C:\Users\Saad Rashid\AppData\Local\Temp\
ipykernel_16184\152243813.py:1: DtypeWarning: Columns (5,9,10,11,12)
have mixed types. Specify dtype option on import or set
low_memory=False.
  d1 = pd.read_csv("all_reviews.csv")
```

```
d2 = pd.read_csv("glassdoor_reviews.csv")
```

```
d1
```

```
         rating                                                  title  \
0           5.0                                                   Good
1           4.0                                                   Good
2           4.0  Supervising the manufacturing the processes, e...
3           1.0                                               terrible
4           4.0                  It could be so good, but it isn’t
...         ...                                                    ...
9901884     4.0  I enjoyed working here.   The only thing I did...
9901885     2.0                                  Guest Service Agent
9901886     5.0                             Great Company and Staff
9901887     2.0                             Working at Victoria inn
9901888     2.0                                      banquet server

                                        status  \
0        Current Employee, more than 10 years
1           Former Employee, less than 1 year
2         Current Employee, more than 1 year
3         Current Employee, more than 1 year
4         Current Employee, more than 3 years
...                                       ...
9901884     Former Employee, less than 1 year
9901885                       Former Employee
9901886    Former Employee, more than 3 years
9901887     Former Employee, more than 1 year
9901888                     Current Employee
```

```
                                                      pros  \
0                     Knowledge gain of complete  project
1                  Good work,good work , flexible, support
2        This company is a best opportunity for me to l...
3                           I wish there were some to list
4        Fast Paced. Endless challenges. Inclusive envi...
...                                                    ...
9901884  The pros were that the staff were awesome.   V...
9901885  No Pros comment at all . Horrible manager ,Hor...
9901886  Great managers who really care about their emp...
9901887  Flexible schedule and free food, you can choos...
9901888  flexible hours  and know the setting of banque...

                                                      cons advice
Recommend  \
0                     Financial growth  and personal growth    NaN
v
1          Good,work, flexible,good support, good team work    NaN
v
2        Monthly Target work,Maintain production schedu...    NaN
v
3                                    too many to list here    NaN
x
4        The biggest perk of the job provides no value ...    NaN
o
...                                                    ...    ...
...
9901884  Some challenges were the policies and procedur...    NaN
v
9901885  The working environment is gossip. Girls like ...    NaN
x
9901886           Pay wasn't that great during slow season    NaN
v
9901887  Seasonal unstable job/hours and the management...    NaN
x
9901888                 not enough hour, lack of  training    NaN
x

        CEO Approval Business Outlook Career Opportunities  \
0                  o                v                     3
1                  o                o                     4
2                  o                v                     2
3                  x                x                   1.0
4                  o                o                   3.0
...              ...              ...                   ...
9901884            r                r                   4.0
9901885            o                x                   2.0
9901886            o                v                   3.0
9901887            o                x                   1.0
```

```
9901888            o                 o                   2.0

        Compensation and Benefits Senior Management Work/Life Balance
\
0                             3                 3                   3

1                             4                 4                   4

2                             3                 2                   2

3                           3.0               1.0                 3.0

4                           3.0               3.0                 1.0

...                         ...               ...                 ...

9901884                     4.0               2.0                 3.0

9901885                     2.0               2.0                 2.0

9901886                     2.0               5.0                 4.0

9901887                     3.0               2.0                 3.0

9901888                     3.0               2.0                 3.0


         Culture & Values  Diversity & Inclusion  \
0                     3.0                   3.0
1                     4.0                   4.0
2                     2.0                   2.0
3                     1.0                   NaN
4                     4.0                   5.0
...                   ...                   ...
9901884               3.0                   NaN
9901885               1.0                   NaN
9901886               4.0                   NaN
9901887               1.0                   NaN
9901888               2.0                   NaN

                                                 firm_link
date  \
0        Reviews/Baja-Steel-and-Fence-Reviews-E5462645.htm  Nov 19,
2022
1        Reviews/Baja-Steel-and-Fence-Reviews-E5462645.htm  Jan 29,
2022
2        Reviews/Baja-Steel-and-Fence-Reviews-E5462645.htm  Aug 12,
2021
3        https://www.glassdoor.com/Reviews/Calgary-Flam...  Sep 24,
2020
4        https://www.glassdoor.com/Reviews/Calgary-Flam...  Mar 25,
```

```
2023
...                                                          ...
...
9901884  Reviews/Victoria-Inn-Hotel-&-Convention-Centre...   Apr 9,
2016
9901885  Reviews/Victoria-Inn-Hotel-&-Convention-Centre...  Jul 13,
2016
9901886  Reviews/Victoria-Inn-Hotel-&-Convention-Centre...  Dec 17,
2014
9901887  Reviews/Victoria-Inn-Hotel-&-Convention-Centre...  Aug 13,
2019
9901888  Reviews/Victoria-Inn-Hotel-&-Convention-Centre...  Dec 10,
2015

                                job  index
0                   Manager Design    NaN
1               Anonymous Employee    NaN
2              Production Engineer    NaN
3         Senior Account Executive    NaN
4                Assistant Manager    NaN
...                            ...    ...
9901884                     Server    NaN
9901885          Anonymous Employee    NaN
9901886              Banquet Server    NaN
9901887            Server/Bartender    NaN
9901888                              NaN

[9901889 rows x 19 columns]

d2

                           firm date_review
job_title  \
0       AFH-Wealth-Management  2015-04-05

1       AFH-Wealth-Management  2015-12-11              Office
Administrator
2       AFH-Wealth-Management  2016-01-28              Office
Administrator
3       AFH-Wealth-Management  2016-04-16

4       AFH-Wealth-Management  2016-04-23              Office
Administrator
...                       ...         ...
...
838561         the-LEGO-Group  2021-06-02                 Marketing
Manager
838562         the-LEGO-Group  2021-06-03                   Sales
Associate
838563         the-LEGO-Group  2021-06-03
```

```
Strategist
838564         the-LEGO-Group  2021-06-04   Customer Service
Representative
838565         the-LEGO-Group  2021-06-04        Human Resources
Specialist

                                    current  \
0                          Current Employee
1        Current Employee, more than 1 year
2        Current Employee, less than 1 year
3                          Current Employee
4        Current Employee, more than 1 year
...                                      ...
838561  Current Employee, more than 5 years
838562   Current Employee, less than 1 year
838563                     Current Employee
838564   Current Employee, less than 1 year
838565  Current Employee, more than 3 years

                                          location  overall_rating  \
0                                              NaN               2
1                    Bromsgrove, England, England               2
2                    Bromsgrove, England, England               1
3                                              NaN               5
4                    Bromsgrove, England, England               1
...                                            ...             ...
838561                   München, Bavaria, Bavaria               5
838562                    London, England, England               3
838563                                         NaN               4
838564                                         NaN               5
838565  Kladno, Central Bohemian, Central Bohemian               5

        work_life_balance  culture_values  diversity_inclusion
career_opp  \
0                     4.0             3.0                  NaN
2.0
1                     3.0             1.0                  NaN
2.0
2                     1.0             1.0                  NaN
1.0
3                     2.0             3.0                  NaN
2.0
4                     2.0             1.0                  NaN
2.0
...                   ...             ...                  ...
...
838561                4.0             5.0                  4.0
4.0
838562                NaN             NaN                  NaN
NaN
```

```
838563                5.0             5.0                   5.0
3.0
838564                NaN             NaN                   NaN
NaN
838565                4.0             5.0                   4.0
4.0

        comp_benefits  senior_mgmt recommend ceo_approv outlook  \
0                 3.0          3.0         x          o       r
1                 1.0          4.0         x          o       r
2                 1.0          1.0         x          o       x
3                 2.0          3.0         x          o       r
4                 1.0          1.0         x          o       x
...               ...          ...       ...        ...     ...
838561            4.0          4.0         v          v       v
838562            NaN          NaN         o          o       o
838563            5.0          3.0         v          o       o
838564            NaN          NaN         o          o       o
838565            5.0          5.0         v          v       o

                                                 headline  \
0              Young colleagues, poor micro management
1                         Excellent staff, poor salary
2                       Low salary, bad micromanagement
3                         Over promised under delivered
4                               client reporting admin
...                                                   ...
838561           Just an awesome company to work for!!!
838562                                  working at lego
838563           not interested in growing their people
838564                                Great Place to Work
838565  I strongly recommend the LEGO Group as employer

                                                    pros  \
0       Very friendly and welcoming to new staff. Easy...
1           Friendly, helpful and hard-working colleagues
2       Easy to get the job even without experience in...
3                                Nice staff to work with
4                   Easy to get the job, Nice colleagues.
...                                                   ...
838561  Great company values, awesome product, smart c...
838562                      staff discount is really nice
838563                    loved brand for a lot of people
838564           Good wages, good hours, lots of resources
838565  The LEGO Group is company with many opportunit...

                                                    cons
0         Poor salaries, poor training and communication.
1       Poor salary which doesn't improve much with pr...
2       Very low salary, poor working conditions, very...
```

```
3                No career progression and salary is poor
4       Abysmal pay, around minimum wage. No actual tr...
...                                                   ...
838561       Not very easy to transfer to other locations
838562  micro managing is a hassle\r\ncan become menta...
838563  you can spend 6-10 years without any promotion...
838564  Working every other weekend, busy seasons can ...
838565  Many things are centralized in Denmark and rel...

[838566 rows x 18 columns]
```

```
df1 = d1.copy()
df2 = d2.copy()

df1['firm_link']
```

```
0          Reviews/Baja-Steel-and-Fence-Reviews-E5462645.htm
1          Reviews/Baja-Steel-and-Fence-Reviews-E5462645.htm
2          Reviews/Baja-Steel-and-Fence-Reviews-E5462645.htm
3          https://www.glassdoor.com/Reviews/Calgary-Flam...
4          https://www.glassdoor.com/Reviews/Calgary-Flam...
                                 ...
9901884    Reviews/Victoria-Inn-Hotel-&-Convention-Centre...
9901885    Reviews/Victoria-Inn-Hotel-&-Convention-Centre...
9901886    Reviews/Victoria-Inn-Hotel-&-Convention-Centre...
9901887    Reviews/Victoria-Inn-Hotel-&-Convention-Centre...
9901888    Reviews/Victoria-Inn-Hotel-&-Convention-Centre...
Name: firm_link, Length: 9901889, dtype: object
```

Extracting firm name from the firm link to match with the dataset2 'df2'

```
import re

def extract_company_name(link):
    match =
re.search(r'(?:Reviews/|https://www.glassdoor.com/Reviews/)(.*?)(?:-
Reviews|&)', link)
    return match.group(1).replace('-', ' ') if match else None

df1['firm_link'] = df1['firm_link'].apply(extract_company_name)

df1['firm_link']
```

```
0          Baja Steel and Fence
1          Baja Steel and Fence
2          Baja Steel and Fence
3                Calgary Flames
4                Calgary Flames
                   ...
```

```
9901884     Victoria Inn Hotel
9901885     Victoria Inn Hotel
9901886     Victoria Inn Hotel
9901887     Victoria Inn Hotel
9901888     Victoria Inn Hotel
Name: firm_link, Length: 9901889, dtype: object
```

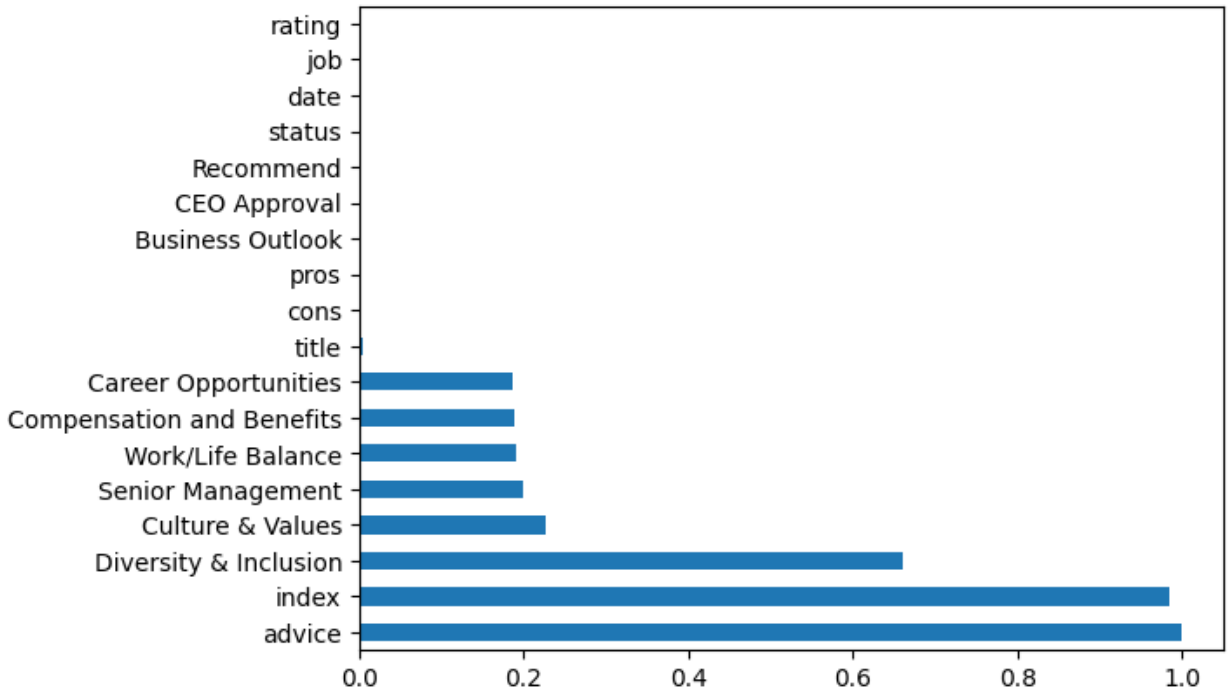
Missing values in both of the datasets

```
df1.isnull().sum()
```

```
rating                          171
title                         39424
status                          171
pros                            188
cons                            234
advice                      9901876
Recommend                       171
CEO Approval                    171
Business Outlook                171
Career Opportunities        1848262
Compensation and Benefits   1878091
Senior Management           1975466
Work/Life Balance           1894490
Culture & Values            2246773
Diversity & Inclusion       6544282
firm_link                         0
date                            171
job                             171
index                       9740269
dtype: int64
```

```
missing_percent1 = df1.isna().sum().sort_values(ascending=False) /
len(df1)
missing_percent1
missing_percent1[missing_percent1 != 0].plot(kind = 'barh')
```

```
<Axes: >
```

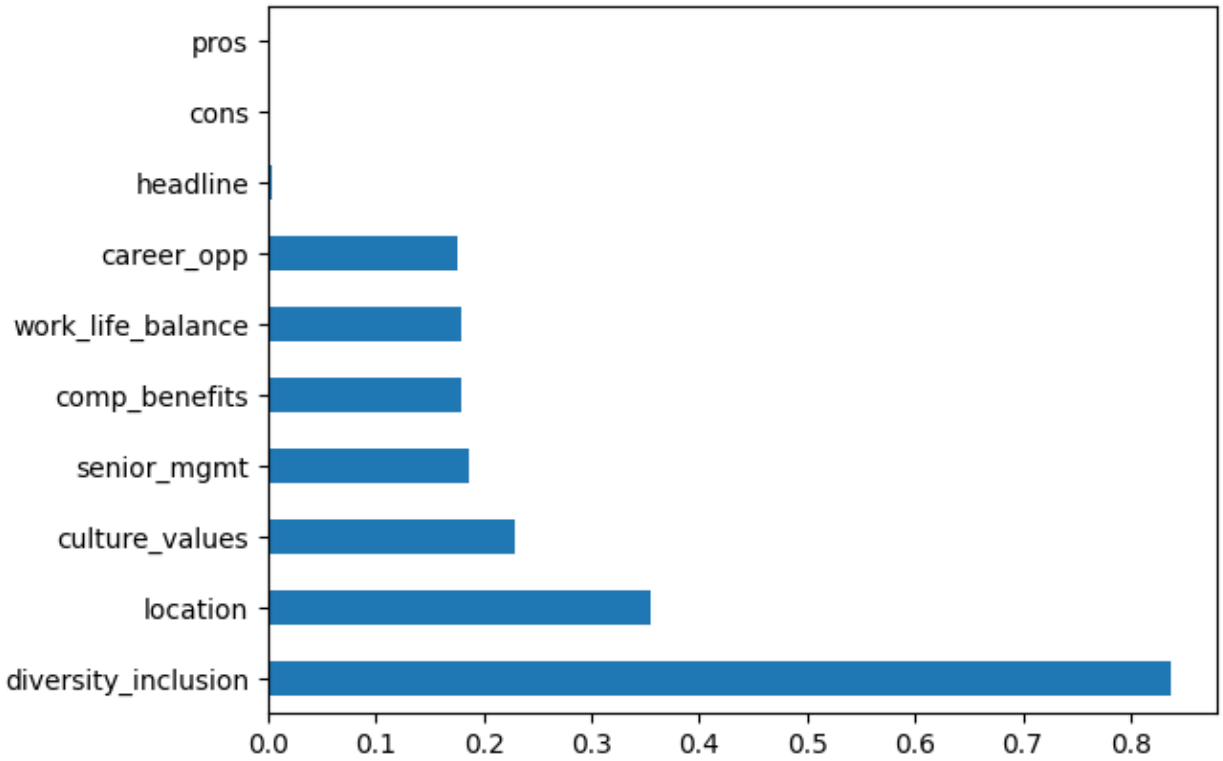



```
df2.isnull().sum()
```

```
firm                         0
date_review                  0
job_title                    0
current                      0
location                297343
overall_rating               0
work_life_balance       149894
culture_values          191373
diversity_inclusion     702500
career_opp              147501
comp_benefits           150082
senior_mgmt             155876
recommend                    0
ceo_approv                   0
outlook                      0
headline                  2590
pros                         2
cons                        13
dtype: int64
```

```
missing_percent2 = df2.isna().sum().sort_values(ascending=False) /
len(df2)
missing_percent2
missing_percent2[missing_percent2 != 0].plot(kind = 'barh')
```

```
<Axes: >
```



Dropping unwanted or unique columns to maintaing stability for further merge

```
df1.drop('advice', axis=1, inplace=True)
df1.drop('index', axis=1, inplace=True)
df2.drop('location', axis=1, inplace=True)
```

Dropping missing rows and duplicates

```
df1_cleaned = df1.drop_duplicates()
df2_cleaned = df2.drop_duplicates()

df1_cleaned = df1_cleaned.dropna()
df2_cleaned = df2_cleaned.dropna()
```

Rename columns in df2 and df2 to match with eachother

```
df1_cleaned.rename(columns={
    'title': 'headline',
    'firm_link': 'firm',
}, inplace=True)

df2_cleaned.rename(columns={
    'overall_rating': 'rating',
    'job_title': 'job',
```

```
    'current': 'status',
    'work_life_balance': 'Work/Life Balance',
    'comp_benefits': 'Compensation and Benefits',
    'recommend': 'Recommend',
    'ceo_approv': 'CEO Approval',
    'outlook': 'Business Outlook',
    'career_opp': 'Career Opportunities',
    'senior_mgmt': 'Senior Management',
    'culture_values': 'Culture & Values',
    'diversity_inclusion': 'Diversity & Inclusion',
    'date_review': 'date',
}, inplace=True)
```

Finally merging both datasets

```
merged_df = pd.concat([df2_cleaned, df1_cleaned], ignore_index=True)
```

Handling outliers

```
def remove_outliers_iqr(data):
    q1 = data.quantile(0.25)
    q3 = data.quantile(0.75)
    iqr = q3 - q1
    lowerB = q1 - 1.5 * iqr
    upperB = q3 + 1.5 * iqr
    return data[(data >= lowerB) & (data <= upperB)]

merged_df['rating'] = remove_outliers_iqr(merged_df['rating'])
merged_df['Culture & Values'] = remove_outliers_iqr(merged_df['Culture
& Values'])
merged_df['Diversity & Inclusion'] =
remove_outliers_iqr(merged_df['Diversity & Inclusion'])

merged_df.to_csv('merged_and_cleaned_data.csv', index=False)

# df = merged_df.copy()
df = pd.read_csv('merged_and_cleaned_data.csv')

df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3129490 entries, 0 to 3129489
Data columns (total 17 columns):
 #   Column                    Dtype
---  ------                    -----
 0   firm                      object
 1   date                      object
 2   job                       object
 3   status                    object
 4   rating                    float64
```

```
 5   Work/Life Balance          object
 6   Culture & Values           float64
 7   Diversity & Inclusion      float64
 8   Career Opportunities       object
 9   Compensation and Benefits  object
 10  Senior Management          object
 11  Recommend                  object
 12  CEO Approval               object
 13  Business Outlook           object
 14  headline                   object
 15  pros                       object
 16  cons                       object
dtypes: float64(3), object(14)
memory usage: 405.9+ MB
```

Exploratory Data Analysis (EDA)

```
df
```

```
                          firm           date  \
0        AFH-Wealth-Management     2020-10-01
1        AFH-Wealth-Management     2021-02-05
2        AFH-Wealth-Management     2021-02-07
3        AFH-Wealth-Management     2021-02-07
4        AFH-Wealth-Management     2021-03-23
...                        ...            ...
3129485    Victoria Inn Hotel    Dec 5, 2021
3129486    Victoria Inn Hotel    Sep 9, 2021
3129487    Victoria Inn Hotel    Jul 1, 2021
3129488    Victoria Inn Hotel   Nov 30, 2020
3129489    Victoria Inn Hotel    Oct 7, 2020

                                       job
status  \
0                     Office Administrator  Former Employee, more than
3 years
1                          Quality Control                       Former
Employee
2                        IFA Administrator   Former Employee, less than
1 year
3                     Investment Operations   Former Employee, more than
1 year
4                           Administrative                       Former
Employee
...                                    ...
...
3129485          Front Desk Night Auditor   Former Employee, less than
1 year
```

```
3129486                              Porter                     Current
Employee
3129487   Customer Service Representative  Current Employee, more than
1 year
3129488                   Front Desk Agent                     Current
Employee
3129489                     Banquet Server   Former Employee, less than
1 year

         rating Work/Life Balance  Culture & Values  Diversity &
Inclusion  \
0           2.0               1.0               3.0
1.0
1           1.0               3.0               1.0
2.0
2           4.0               3.0               3.0
4.0
3           3.0               5.0               5.0
4.0
4           1.0               5.0               1.0
2.0
...         ...               ...               ...
...
3129485     3.0               4.0               3.0
3.0
3129486     4.0               4.0               5.0
4.0
3129487     2.0               1.0               1.0
5.0
3129488     5.0               4.0               3.0
3.0
3129489     3.0               4.0               3.0
4.0

        Career Opportunities Compensation and Benefits Senior
Management  \
0                        1.0                       2.0
2.0
1                        1.0                       1.0
1.0
2                        4.0                       4.0
2.0
3                        3.0                       1.0
2.0
4                        1.0                       2.0
1.0
...                      ...                       ...                .
..
3129485                  3.0                       3.0
```

```
2.0
3129486                  4.0                         4.0
4.0
3129487                  1.0                         1.0
1.0
3129488                  5.0                         5.0
4.0
3129489                  3.0                         3.0
4.0

        Recommend CEO Approval Business Outlook  \
0               x            o                x
1               x            o                o
2               v            o                v
3               x            o                v
4               x            o                r
...           ...          ...              ...
3129485         v            o                v
3129486         v            o                v
3129487         x            o                x
3129488         v            o                v
3129489         v            o                v

                                           headline  \
0        The people both make and destroy this place
1                                Very low salaries
2                                             Good
3                                        AFH Review
4                    Great for people, not for work
...                                             ...
3129485                                 No Security
3129486                                Flexible job
3129487                                Event server
3129488             Good place to start your career
3129489                           Victoria Inn server

                                                    pros  \
0        Great people in some places, excellent Christm...
1        Majority of the people there are lovely, and t...
2         Nice environment, love people, not too stressful
3        -Great People\r\n-Heading in a good direction ...
4        You meet lovely people and make long life friends
...                                                    ...
3129485  Independent work , meeting new guests, networking
3129486   - Fits within your schedule\r\n- Bonus pay + tip
3129487                     Good team work and free food.
3129488  Free staff meals\r\nFlexible shift\r\nFriendly...
3129489                          Good tips and pay is good

                                                    cons
```

```
0        Poor pay, huge gap for pay between senior mana...
1        Salaries are much lower than market competitor...
2                        Management can be clicky at times
3        -Low Salary\r\n-Middle Management likes to mic...
4        It's not a good work environment and you're no...
...                                                    ...
3129485  Senior manager was harsh and sometime down rig...
3129486  - Fewer shifts depending on your availability ...
3129487          Management, less hours during the winter.
3129488  Work depends on customer flow\r\nCustomers som...
3129489  Varying shift hours lack of understanding betw...

[3129490 rows x 17 columns]
```

Descriptive Statistics of the Dataset

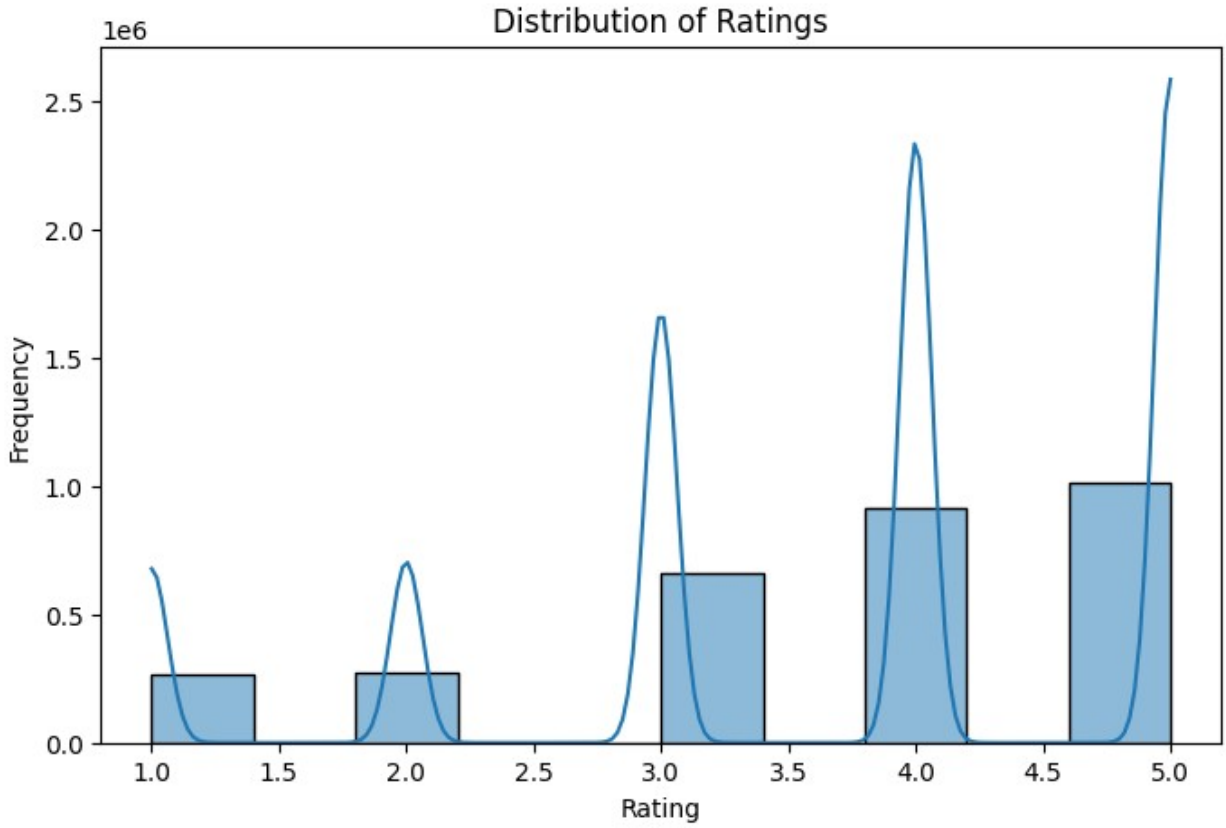
```
df.describe()
```

```
             rating  Work/Life Balance  Culture & Values  \
count  3.129490e+06       3.129490e+06      3.129490e+06
mean   3.682574e+00       3.463104e+00      3.622715e+00
std    1.244857e+00       1.373402e+00      1.355479e+00
min    1.000000e+00       1.000000e+00      1.000000e+00
25%    3.000000e+00       3.000000e+00      3.000000e+00
50%    4.000000e+00       4.000000e+00      4.000000e+00
75%    5.000000e+00       5.000000e+00      5.000000e+00
max    5.000000e+00       5.000000e+00      5.000000e+00

       Diversity & Inclusion  Career Opportunities  Compensation and
Benefits  \
count           3.129490e+06          3.129490e+06
3.129490e+06
mean            3.843587e+00          3.532660e+00
3.468274e+00
std             1.273234e+00          1.310190e+00
1.266458e+00
min             1.000000e+00          1.000000e+00
1.000000e+00
25%             3.000000e+00          3.000000e+00
3.000000e+00
50%             4.000000e+00          4.000000e+00
4.000000e+00
75%             5.000000e+00          5.000000e+00
5.000000e+00
max             5.000000e+00          5.000000e+00
5.000000e+00

       Senior Management
count       3.129490e+06
mean        3.297208e+00
```

```
std          1.398500e+00
min          1.000000e+00
25%          2.000000e+00
50%          3.000000e+00
75%          5.000000e+00
max          5.000000e+00
```

Histogram for Rating

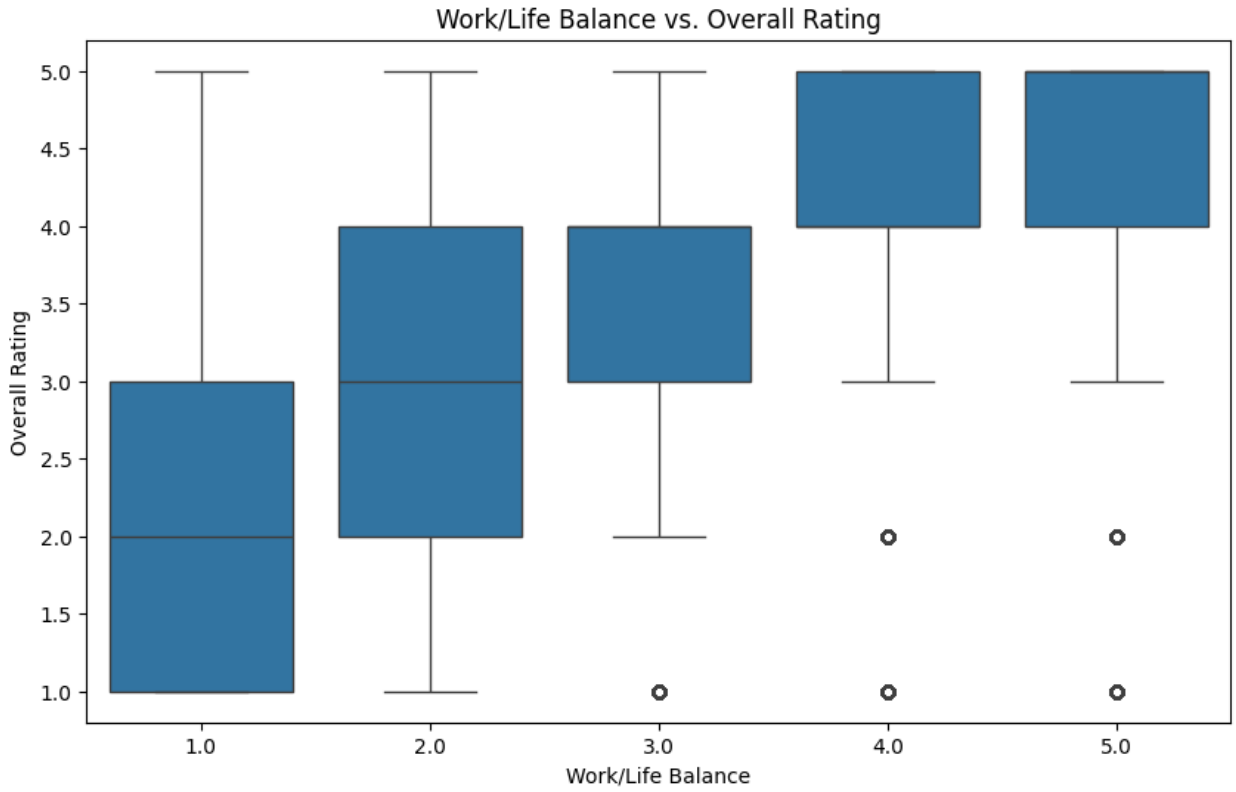
```
plt.figure(figsize=(8, 5))
sns.histplot(df['rating'], bins=10, kde=True)
plt.title('Distribution of Ratings')
plt.xlabel('Rating')
plt.ylabel('Frequency')
plt.show()
```



Why Chosen: A histogram is ideal for visualizing the distribution of a single continuous variable —in this case, employee ratings. It helps us understand the spread and frequency of different rating values, showing how many employees gave certain ratings.

Box plot of Work/Life Balance vs Rating

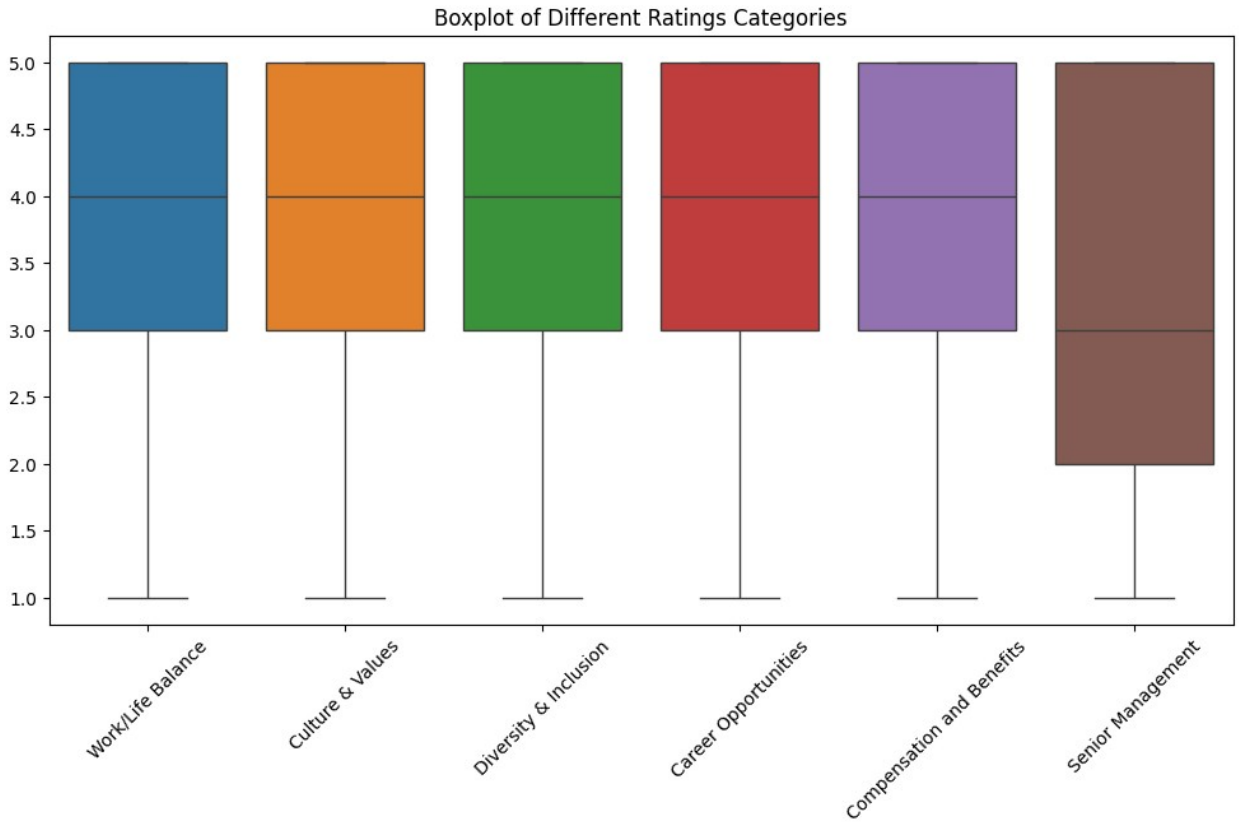
```
plt.figure(figsize=(10, 6))
sns.boxplot(x='Work/Life Balance', y='rating', data=df)
plt.title('Work/Life Balance vs. Overall Rating')
plt.xlabel('Work/Life Balance')
plt.ylabel('Overall Rating')
plt.show()
```



Why Chosen: Boxplots are great for comparing the distribution of a continuous variable (ratings) across different categories (work-life balance scores). It shows the spread, median, and outliers, making it easier to understand how work-life balance influences overall satisfaction.

Boxplot: Different Ratings Categories

```
plt.figure(figsize=(12, 6))
sns.boxplot(data=df[['Work/Life Balance', 'Culture & Values',
'Diversity & Inclusion',
                     'Career Opportunities', 'Compensation and
Benefits', 'Senior Management']])
plt.title('Boxplot of Different Ratings Categories')
plt.xticks(rotation=45)
plt.show()
```



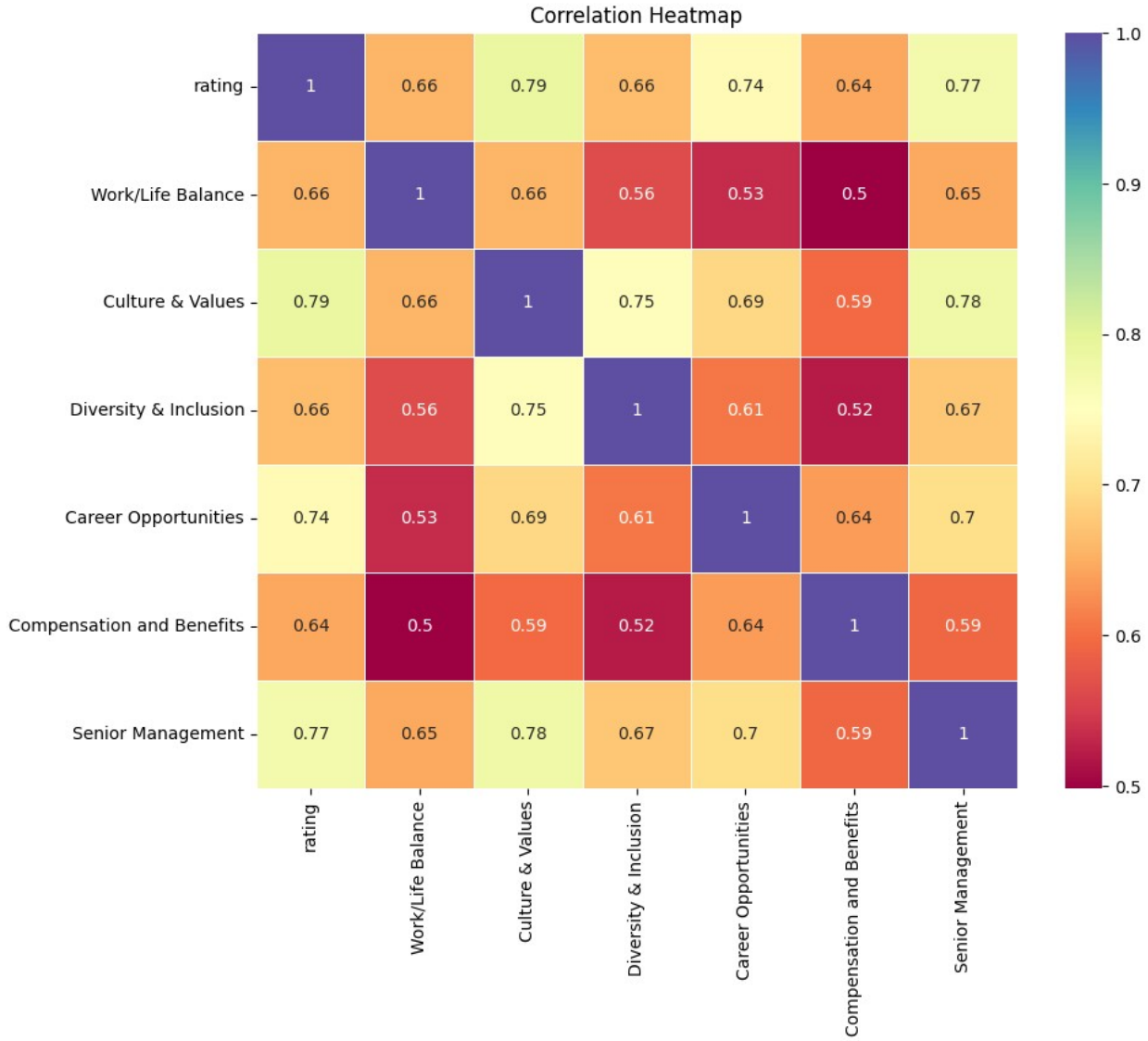
Why Chosen: A boxplot is ideal for comparing the distribution of multiple continuous variables (in this case, different ratings categories). It provides a clear summary of each category's distribution, highlighting key statistics such as the median, quartiles, and potential outliers.

In this visualization, we compare the spread of ratings for categories like Work/Life Balance, Culture & Values, Diversity & Inclusion, Career Opportunities, Compensation and Benefits, and Senior Management. The boxplot makes it easy to see which categories have a wider range of ratings (indicating more variability in employee perceptions) and which have more consistent ratings (smaller interquartile ranges).

Correlation heatmap

Selecting only numerical columns for correlation

```
Numericdf = df.select_dtypes(include=['float64', 'int64'])

plt.figure(figsize=(10, 8))
sns.heatmap(Numericdf.corr(), annot=True, cmap='Spectral',
linewidths=0.5)
plt.title('Correlation Heatmap')
plt.show()
```



Why Chosen: A heatmap is perfect for visualizing the strength of correlations between multiple numerical variables, such as work-life balance, culture, and compensation. The color coding makes it easy to identify strong or weak relationships.

Hypothesis Testing for Work-Life Balance

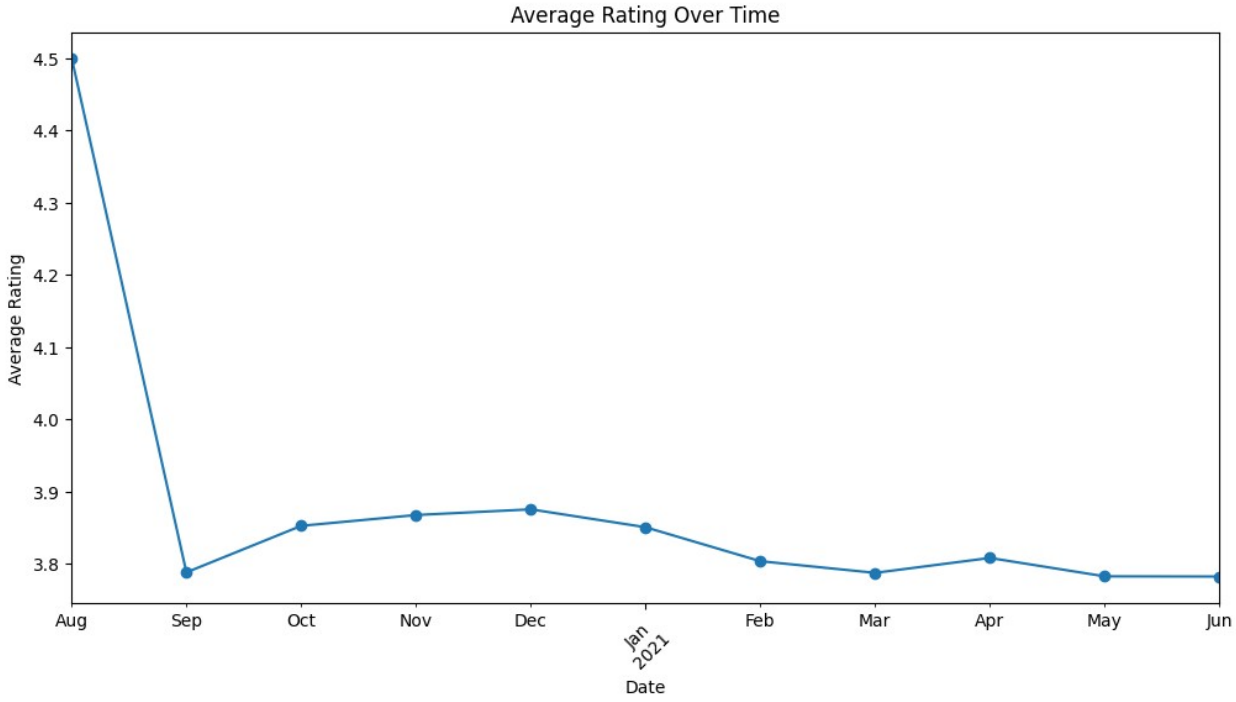
```
wlbM = df['Work/Life Balance'].dropna().mean()
tStat, pVal = stats.ttest_1samp(df['Work/Life Balance'].dropna(), 3)

print(f"Hypothesis Testing for 'Work/Life Balance' Rating:")
print(f"Mean: {wlbM}")
print(f"T-Statistic: {tStat}, P-Value: {pVal}")
```

```
Hypothesis Testing for 'Work/Life Balance' Rating:
Mean: 3.4631038923275037
T-Statistic: 596.5098082037315, P-Value: 0.0
```

Time-Based Rating Trend

Plotting time-based trend of 'rating'

```
df['date'] = pd.to_datetime(df['date'], errors='coerce')

plt.figure(figsize=(12, 6))
df.groupby(df['date'].dt.to_period('M'))
['rating'].mean().plot(kind='line', marker='o')
plt.title('Average Rating Over Time')
plt.xlabel('Date')
plt.ylabel('Average Rating')
plt.xticks(rotation=45)
plt.show()
```



Why Chosen: Line charts are the most effective way to visualize trends over time. This plot helps observe how average employee ratings fluctuate across different time periods, making it easy to spot patterns or trends.

```
data = df.copy()
```

Let's identify the companies with the highest number of reviews.

```
data['firm'].value_counts().head(10)
```

```
firm
Amazon                                  82390
Tata Consultancy Services               42841
Walmart                                 34310
Cognizant Technology Solutions          29655
Accenture                               28466
Deloitte                                26652
McDonald s                              25863
IBM                                     24074
PwC                                     22178
Target                                  20150
Name: count, dtype: int64
```

Now, we will find the top 10 firms with the highest average ratings.

```
data.groupby('firm')['rating'].mean().nlargest(10)
```

```
firm
"Instituto Brasileiro de Petróleo, Gás e Biocombustíveis (IBP)"    5.0
10th Degree                                                        5.0
12 Hour Massage                                                    5.0
121 Marketing                                                      5.0
1Leisure                                                           5.0
21 Comunicação                                                     5.0
24e Health Clubs                                                   5.0
3B Dienstleistungen                                                5.0
3rd Eye Technologies                                               5.0
4IT                                                                5.0
Name: rating, dtype: float64
```

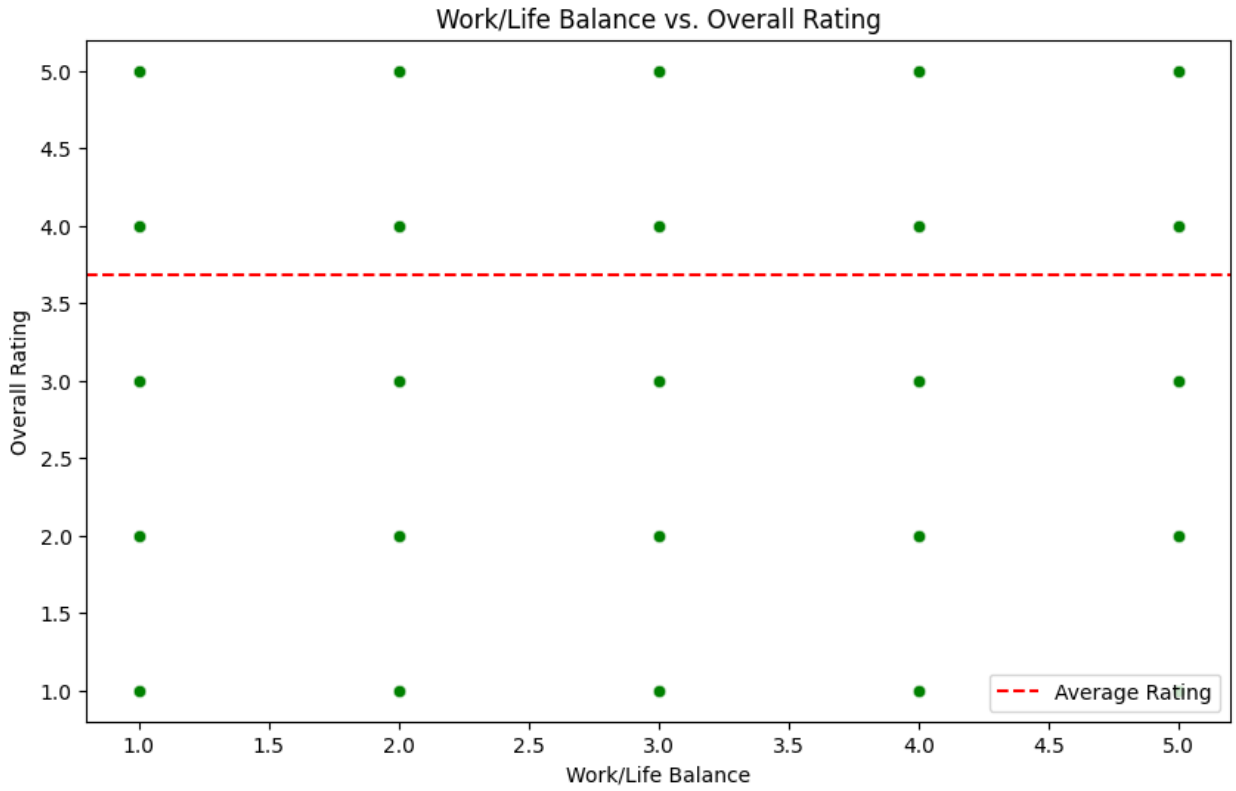
Now, let's explore how career opportunities and work-life balance relate to the overall rating.

Scatter plot for Work/Life Balance vs Rating

```
plt.figure(figsize=(10, 6))
sns.scatterplot(x='Work/Life Balance', y='rating', data=data,
color='green', alpha=0.6)
plt.title('Work/Life Balance vs. Overall Rating')
plt.xlabel('Work/Life Balance')
plt.ylabel('Overall Rating')
plt.axhline(y=data['rating'].mean(), color='red', linestyle='--',
label='Average Rating')
plt.legend()
plt.show()
```

```
# Calculate correlation
data[['Work/Life Balance', 'rating']].corr()
```

```
C:\Users\Saad Rashid\AppData\Local\Programs\Python\Python312\Lib\site-
packages\IPython\core\pylabtools.py:170: UserWarning: Creating legend
with loc="best" can be slow with large amounts of data.
  fig.canvas.print_figure(bytes_io, **kw)
```




```
                    Work/Life Balance    rating
Work/Life Balance            1.000000  0.655784
rating                       0.655784  1.000000
```

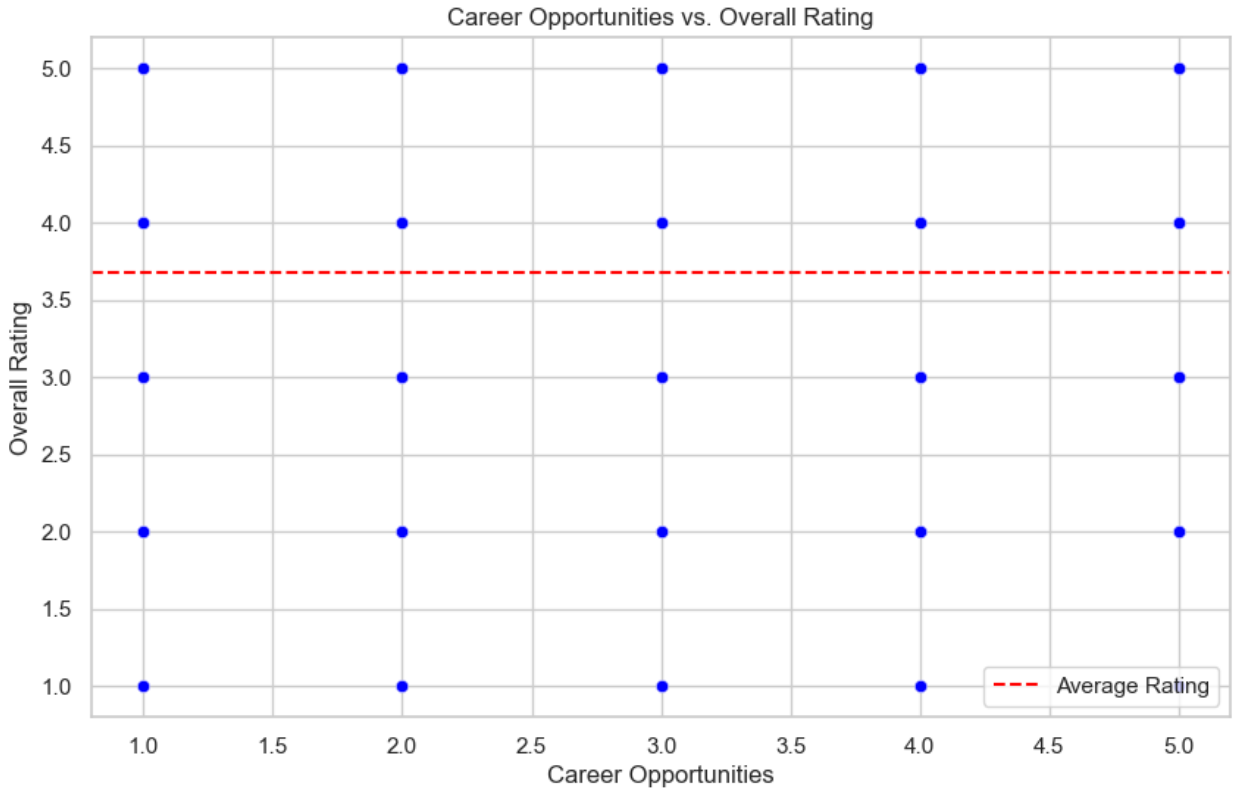
Identifying firms based on their ratings for career opportunities.

Scatter plot for Career Opportunities vs Rating

```
sns.set(style="whitegrid")

plt.figure(figsize=(10, 6))
sns.scatterplot(x='Career Opportunities', y='rating', data=data,
color='blue', alpha=0.6)
plt.title('Career Opportunities vs. Overall Rating')
plt.xlabel('Career Opportunities')
```

```
plt.ylabel('Overall Rating')
plt.axhline(y=data['rating'].mean(), color='red', linestyle='--',
label='Average Rating')
plt.legend()
plt.show()

# Calculate correlation
data[['Career Opportunities', 'rating']].corr()
```

```
C:\Users\Saad Rashid\AppData\Local\Programs\Python\Python312\Lib\site-
packages\IPython\core\pylabtools.py:170: UserWarning: Creating legend
with loc="best" can be slow with large amounts of data.
  fig.canvas.print_figure(bytes_io, **kw)
```

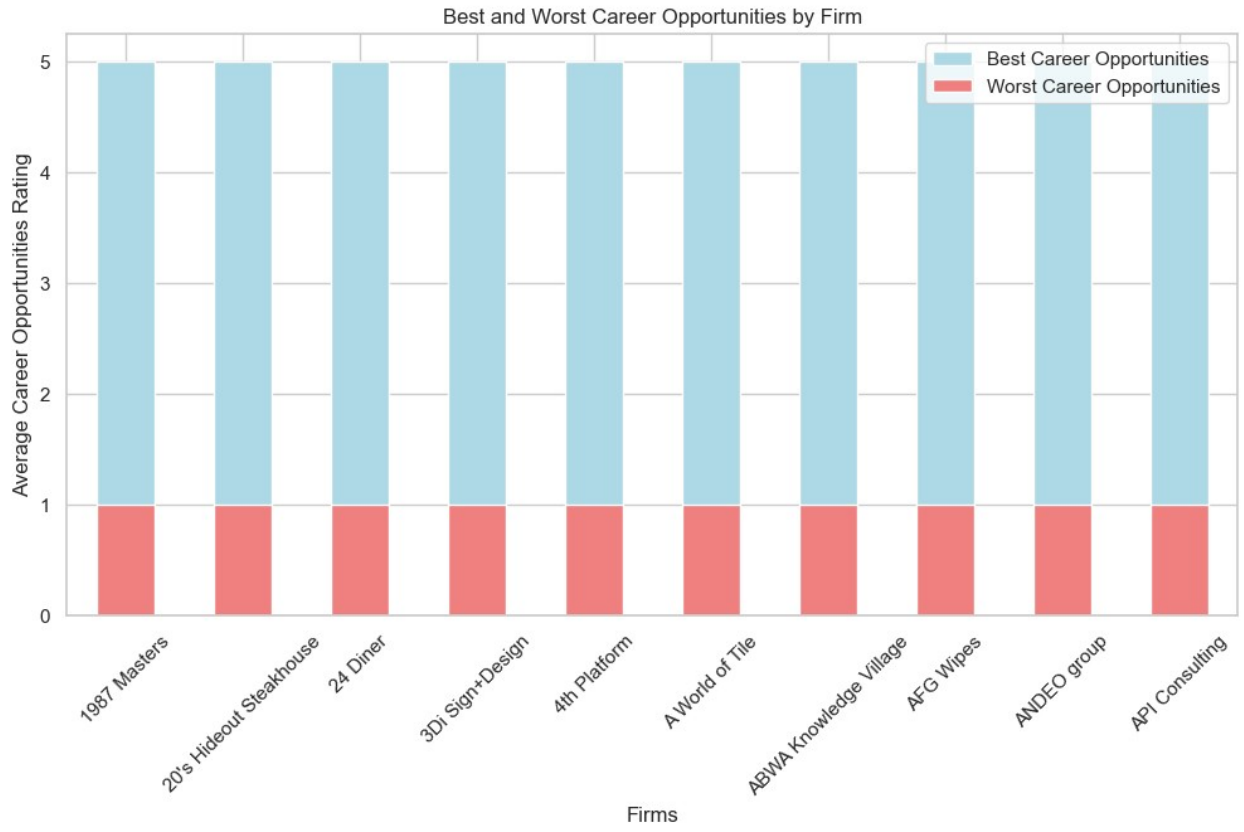



```
                      Career Opportunities    rating
Career Opportunities              1.000000  0.740302
rating                            0.740302  1.000000
```

Best and worst firms for career opportunities

```
bestCareer = data.groupby('firm')['Career
Opportunities'].mean().nlargest(10)
worstCareer = data.groupby('firm')['Career
Opportunities'].mean().nsmallest(10)
```

```
# Plotting the best and worst career opportunities
plt.figure(figsize=(12, 6))
bestCareer.plot(kind='bar', color='lightblue', label='Best Career
Opportunities')
worstCareer.plot(kind='bar', color='lightcoral', label='Worst Career
Opportunities')
plt.title('Best and Worst Career Opportunities by Firm')
plt.xlabel('Firms')
plt.ylabel('Average Career Opportunities Rating')
plt.xticks(rotation=45)
plt.legend()
plt.show()

bestCareer, worstCareer
```



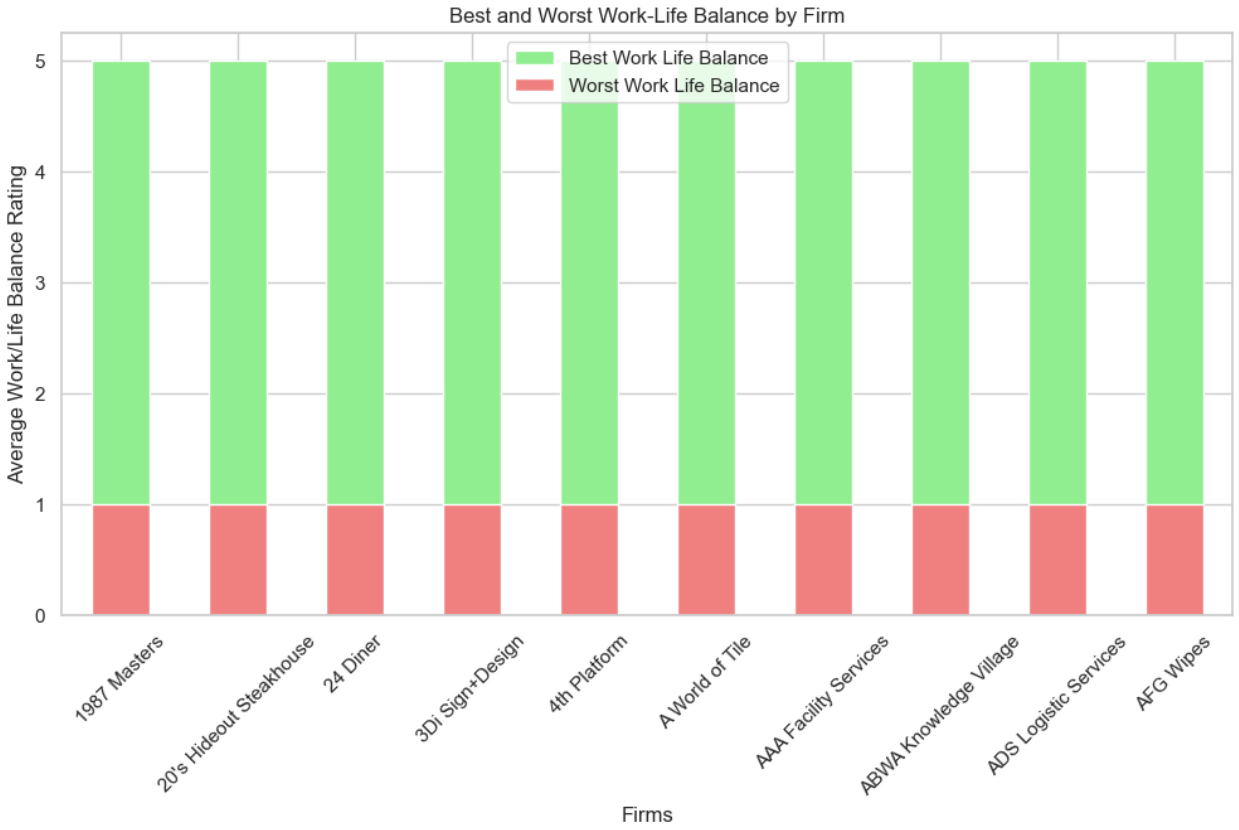

```
(firm
 121 Marketing            5.0
 1Leisure                 5.0
 1Pv6                     5.0
 21 Comunicação           5.0
 247 Labs                 5.0
 360 Sports               5.0
 3rd Eye Technologies     5.0
```

```
4IT                      5.0
5k Network               5.0
7 Arts Construction      5.0
Name: Career Opportunities, dtype: float64,
firm
1987 Masters               1.0
20's Hideout Steakhouse    1.0
24 Diner                   1.0
3Di Sign+Design            1.0
4th Platform               1.0
A World of Tile            1.0
ABWA Knowledge Village     1.0
AFG Wipes                  1.0
ANDEO group                1.0
API Consulting             1.0
Name: Career Opportunities, dtype: float64)
```

Best and worst firms for work-life balance

Plotting the best and worst work-life balance

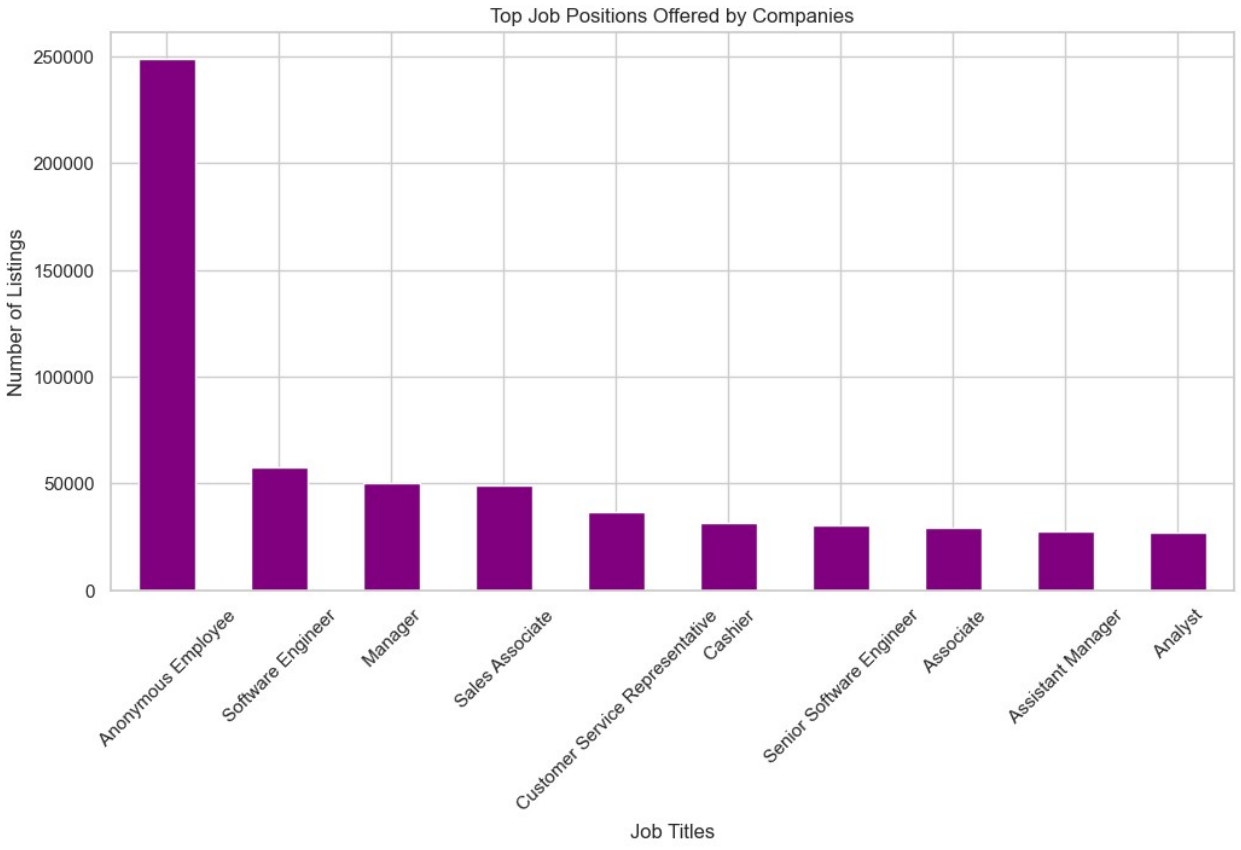
```
bestWLB = data.groupby('firm')['Work/Life
Balance'].mean().nlargest(10)
worstWLB = data.groupby('firm')['Work/Life
Balance'].mean().nsmallest(10)

plt.figure(figsize=(12, 6))
bestWLB.plot(kind='bar', color='lightgreen', label='Best Work Life
Balance')
worstWLB.plot(kind='bar', color='lightcoral', label='Worst Work Life
Balance')
plt.title('Best and Worst Work-Life Balance by Firm')
plt.xlabel('Firms')
plt.ylabel('Average Work/Life Balance Rating')
plt.xticks(rotation=45)
plt.legend()
plt.show()

bestWLB, worstWLB
```




```
(firm
 12 Hour Massage          5.0
 121 Marketing            5.0
 1Leisure                 5.0
 21 Comunicação           5.0
 24e Health Clubs         5.0
 360 Sports               5.0
 3B Dienstleistungen      5.0
 5k Network               5.0
 7 Arts Construction      5.0
 7SIGNAL                  5.0
 Name: Work/Life Balance, dtype: float64,
 firm
 1987 Masters                1.0
 20's Hideout Steakhouse     1.0
 24 Diner                    1.0
 3Di Sign+Design             1.0
 4th Platform                1.0
 A World of Tile             1.0
 AAA Facility Services       1.0
 ABWA Knowledge Village      1.0
 ADS Logistic Services       1.0
 AFG Wipes                   1.0
 Name: Work/Life Balance, dtype: float64)
```

Most common job titles

```
topJobs = data['job'].value_counts().head(10)

# Plotting the top job titles
plt.figure(figsize=(12, 6))
topJobs.plot(kind='bar', color='purple')
plt.title('Top Job Positions Offered by Companies')
plt.xlabel('Job Titles')
plt.ylabel('Number of Listings')
plt.xticks(rotation=45)
plt.show()

topJobs
```



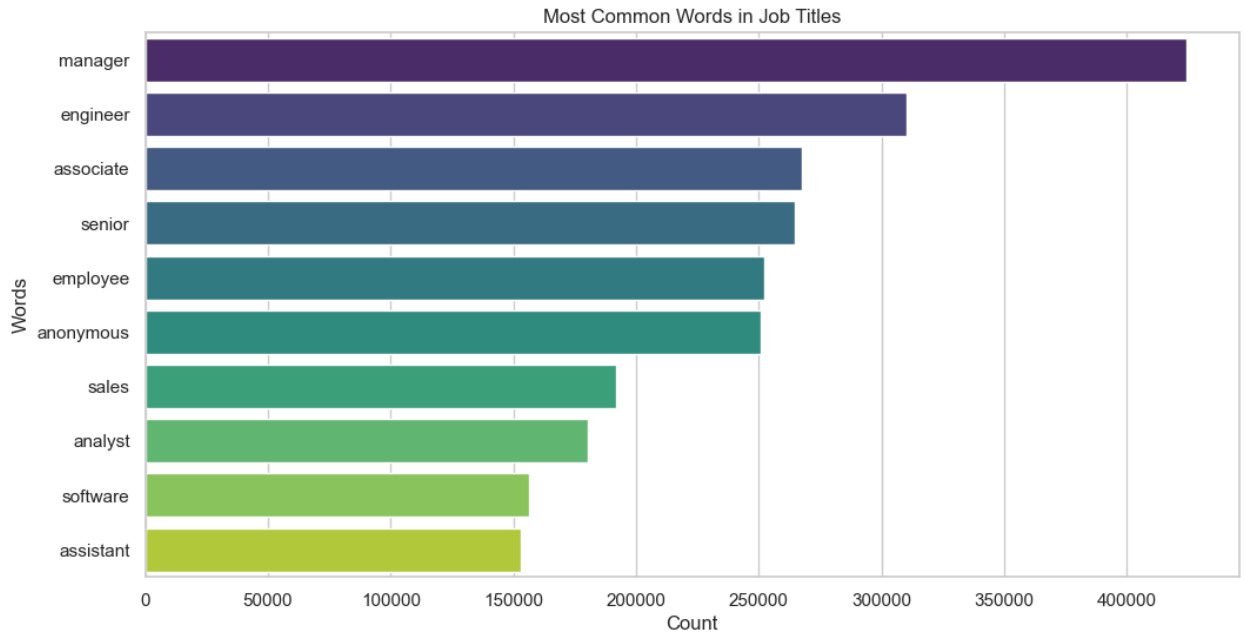

```
job
Anonymous Employee                 248721
Software Engineer                   57551
Manager                             50269
Sales Associate                     49219
Customer Service Representative     36438
Cashier                             31668
Senior Software Engineer            30298
```

```
Associate                           29474
Assistant Manager                   27668
Analyst                             26833
Name: count, dtype: int64
```

Most common words in job titles

```
from collections import Counter
import re

words = ' '.join(data['job']).lower()
wordsList = re.findall(r'\w+', words)
wordCounts = Counter(wordsList)

mostCommonWords = wordCounts.most_common(10)

# Plotting the most common words
plt.figure(figsize=(12, 6))
sns.barplot(x='Count', y='Word', data=pd.DataFrame(mostCommonWords,
columns=['Word', 'Count']), palette='viridis')
plt.title('Most Common Words in Job Titles')
plt.xlabel('Count')
plt.ylabel('Words')
plt.show()

mostCommonWords
```

```
C:\Users\Saad Rashid\AppData\Local\Temp\
ipykernel_16184\3993943361.py:12: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be
removed in v0.14.0. Assign the `y` variable to `hue` and set
`legend=False` for the same effect.

  sns.barplot(x='Count', y='Word', data=pd.DataFrame(mostCommonWords,
columns=['Word', 'Count']), palette='viridis')
```




```
[('manager', 424604),
 ('engineer', 310105),
 ('associate', 267404),
 ('senior', 264552),
 ('employee', 252460),
 ('anonymous', 251046),
 ('sales', 191968),
 ('analyst', 180320),
 ('software', 156363),
 ('assistant', 153023)]
```

Business Insights and Recommendations Based on Glassdoor Reviews

1. Correlation Between Work-Life Balance and Ratings

- **Insight**: There is a strong positive correlation (0.655) between work-life balance and overall ratings. Employees who rate work-life balance highly tend to give higher overall ratings to the company.
- **Recommendation**: Companies should prioritize work-life balance initiatives such as flexible hours, remote working options, and promoting a healthy work-life culture to improve overall employee satisfaction and retention.

2. Career Opportunities and Overall Ratings

- **Insight**: Career opportunities have an even stronger correlation (0.740) with overall ratings compared to work-life balance. Employees value career growth highly, and it directly impacts their perception of the company.
- **Recommendation**: To attract and retain top talent, companies should focus on creating clear career paths, offering mentorship programs, and facilitating internal promotions. Transparent communication regarding career development can further boost employee morale.

3. Top Companies by Reviews

- **Insight**: Amazon, Tata Consultancy Services, and Walmart are the firms with the highest number of reviews. However, the sheer number of reviews doesn't necessarily correlate with high average ratings.
- **Recommendation**: Companies with a large workforce (like Amazon) should carefully analyze their feedback data, especially from locations or departments with low satisfaction, and make targeted improvements.

4. Best and Worst Firms for Work-Life Balance and Career Opportunities

- **Insight**: Firms like "12 Hour Massage," "121 Marketing," and "1Leisure" are rated highly for both work-life balance and career opportunities, whereas firms like "1987 Masters" and "20's Hideout Steakhouse" score poorly.
- **Recommendation**: Firms struggling with work-life balance and career development should conduct employee surveys to understand the underlying issues and implement strategic changes. Providing job flexibility and career training can significantly improve employee satisfaction.

5. Rating Trends Over Time

- **Insight**: There is variability in ratings over time, with some firms showing declining trends in overall ratings. This could be a result of changing policies, management, or external factors.
- **Recommendation**: Companies should monitor ratings and feedback regularly. Declining trends in employee reviews may indicate the need for intervention, such as policy changes, improving management practices, or addressing any dissatisfaction within teams.

6. Diversity and Inclusion

- **Insight**: The average rating for diversity and inclusion is relatively high (mean: 3.84), indicating that employees generally perceive diversity efforts positively. However, there is room for improvement, especially in firms with lower scores in this area.
- **Recommendation**: Firms should continue to strengthen their diversity and inclusion programs, focusing on training, equal opportunity policies, and fostering an inclusive culture. Ensuring representation at all levels of the company will further improve perceptions.

7. Top Job Titles and Common Words

- **Insight**: The most common job titles include "Software Engineer," "Manager," and "Sales Associate." The most frequent words in job titles are "manager," "engineer," and "associate."
- **Recommendation**: Companies should consider offering specific career development programs tailored to these common roles, as well as targeted recruitment strategies. Highlighting growth opportunities for these roles could increase job satisfaction and attract new talent.

8. Senior Management Ratings

- **Insight**: Senior management ratings are relatively low compared to other categories (mean: 3.29). This suggests a gap in leadership effectiveness and employee satisfaction with management.
- **Recommendation**: Companies should focus on improving leadership training and communication between senior management and employees. Leadership development programs, regular feedback loops, and transparent decision-making processes can help improve perceptions of management.

9. Pay and Compensation

- **Insight**: Compensation and benefits have a moderate correlation with overall ratings (0.74), indicating that while important, it is not the sole driver of satisfaction.
- **Recommendation**: Companies should ensure competitive compensation but also focus on non-monetary benefits like recognition, career development, and work-life balance, which have a strong influence on overall satisfaction.

Conclusion:

By addressing the key areas highlighted, such as work-life balance, career growth, senior management, and diversity, companies can significantly improve employee satisfaction and retention. Implementing changes based on these insights will not only boost overall ratings but also enhance the company's reputation on platforms like Glassdoor, attracting both talent and customers.