In [1]:

```
import pandas as pd # dataframe manipulation
import numpy as np # linear algebra

# data visualization
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
print('Seaborn verion', sns.__version__)
sns.set_style('whitegrid')

# text data
import string
import re
df = pd.read_csv('bestsellers with categories.csv')
```

```
Seaborn verion 0.12.2
```

In [ ]:

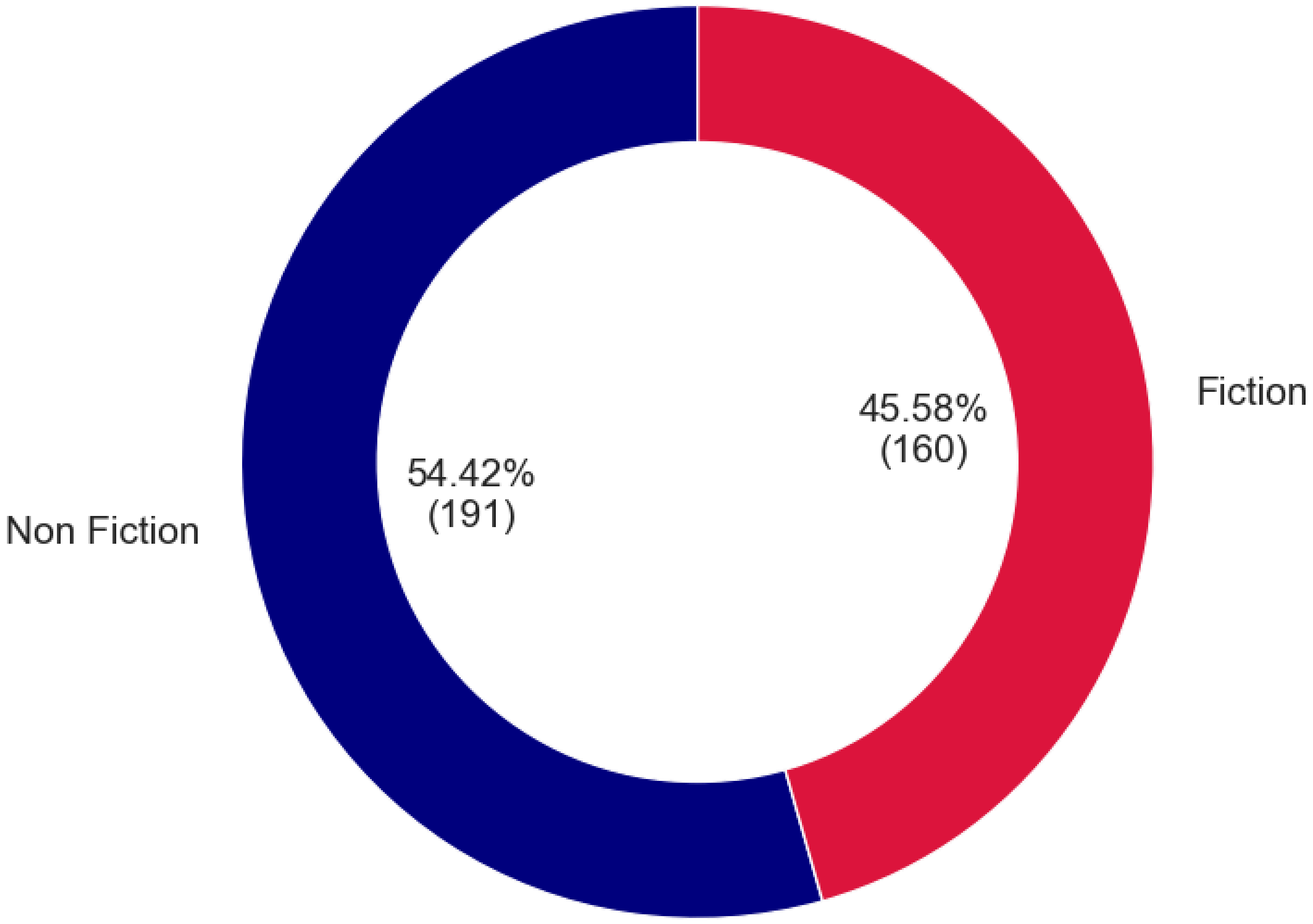
```
df.rename(columns={"User Rating": "User_Rating"}, inplace=True)
df[df.Author == 'J. K. Rowling']
df[df.Author == 'J.K. Rowling']
df.loc[df.Author == 'J. K. Rowling', 'Author'] = 'J.K. Rowling'
df['name_len'] = df['Name'].apply(lambda x: len(x) - x.count(" ")) # subtract whitespaces
punctuations = string.punctuation
print('list of punctuations : ', punctuations)

# percentage of punctuations
def count_punc(text):
    """This function counts the number of punctuations in a text"""
    count = sum(1 for char in text if char in punctuations)
    return round(count/(len(text) - text.count(" "))*100, 3)

# apply function
df['punc%'] = df['Name'].apply(lambda x: count_punc(x))
```

In [2]:

```
no_dup = df.drop_duplicates('Name')
g_count = no_dup['Genre'].value_counts()

fig, ax = plt.subplots(figsize=(8, 8))

def make_autopct(values):
    def my_autopct(pct):
        total = sum(values)
        val = int(round(pct*total/100.0))
        return '{p:.2f}%\n({v:d})'.format(p=pct,v=val)
    return my_autopct

genre_col = ['navy','crimson']
#genre_col = ['khaki','plum']

center_circle = plt.Circle((0, 0), 0.7, color='white')
plt.pie(x=g_count.values, labels=g_count.index, autopct=make_autopct(g_count.values),
          startangle=90, textprops={'size': 15}, pctdistance=0.5, colors=genre_col)
ax.add_artist(center_circle)

fig.suptitle('Distribution of Genre for all unique books from 2009 to 2019', fontsize=20)
fig.show()
```

```
C:\Users\HP\AppData\Local\Temp\ipykernel_4992\1094739433.py:22: UserWarning: Matplotlib is currently using module://matplotlib_i
nline.backend_inline, which is a non-GUI backend, so cannot show the figure.
  fig.show()
```

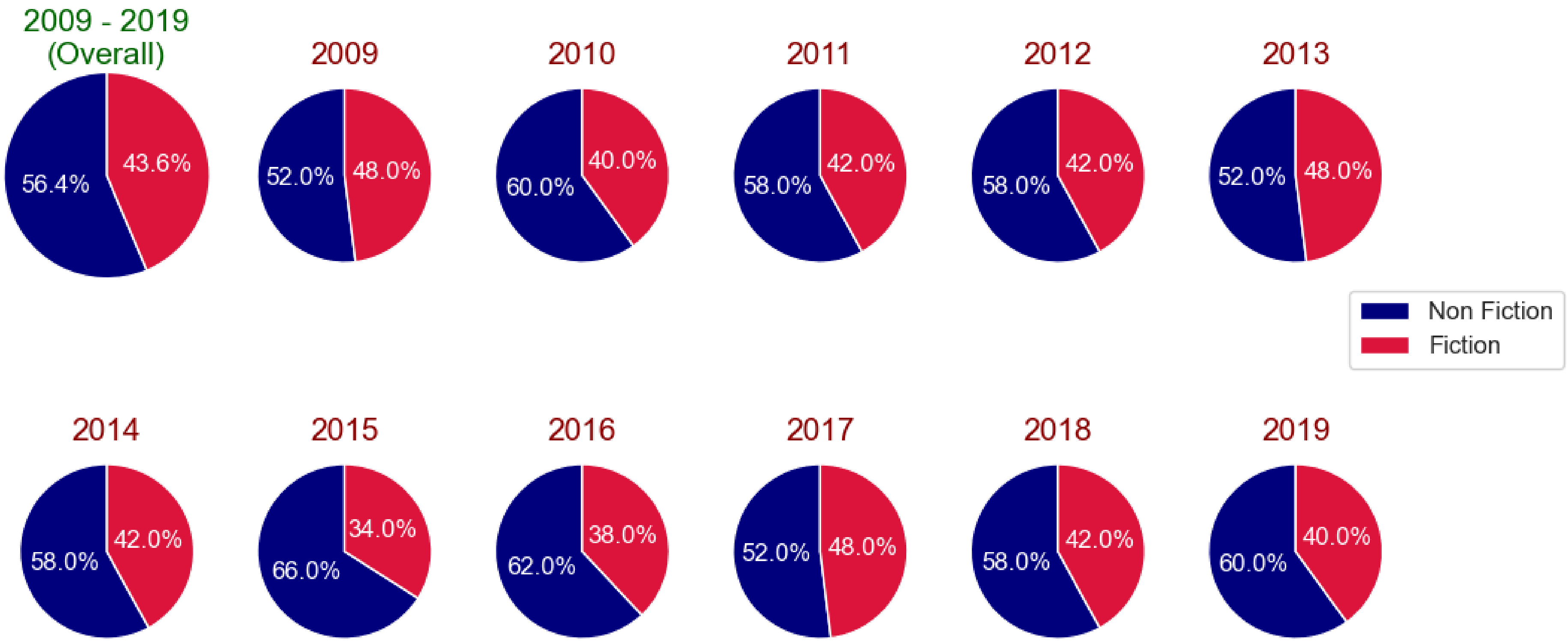
Distribution of Genre for all unique books from 2009 to 2019



In [3]:

```
y1 = np.arange(2009, 2014)
y2 = np.arange(2014, 2020)
g_count = df['Genre'].value_counts()

fig, ax = plt.subplots(2, 6, figsize=(12,6))

ax[0,0].pie(x=g_count.values, labels=None, autopct='%1.1f%%',
            startangle=90, textprops={'size': 12, 'color': 'white'},
            pctdistance=0.5, radius=1.3, colors=genre_col)
ax[0,0].set_title('2009 - 2019\n(Overall)', color='darkgreen', fontdict={'fontsize': 15})

for i, year in enumerate(y1):
    counts = df[df['Year'] == year]['Genre'].value_counts()
    ax[0,i+1].set_title(year, color='darkred', fontdict={'fontsize': 15})
    ax[0,i+1].pie(x=counts.values, labels=None, autopct='%1.1f%%',
                  startangle=90, textprops={'size': 12,'color': 'white'},
                  pctdistance=0.5, colors=genre_col, radius=1.1)

for i, year in enumerate(y2):
    counts = df[df['Year'] == year]['Genre'].value_counts()
    ax[1,i].pie(x=counts.values, labels=None, autopct='%1.1f%%',
                startangle=90, textprops={'size': 12,'color': 'white'},
                pctdistance=0.5, colors=genre_col, radius=1.1)
    ax[1,i].set_title(year, color='darkred', fontdict={'fontsize': 15})

#plt.suptitle('Distribution of Fiction and Non-Fiction books for every year from 2009 to 2019',
             #fontsize=25)
fig.legend(g_count.index, loc='center right', fontsize=12)
fig.show()
```

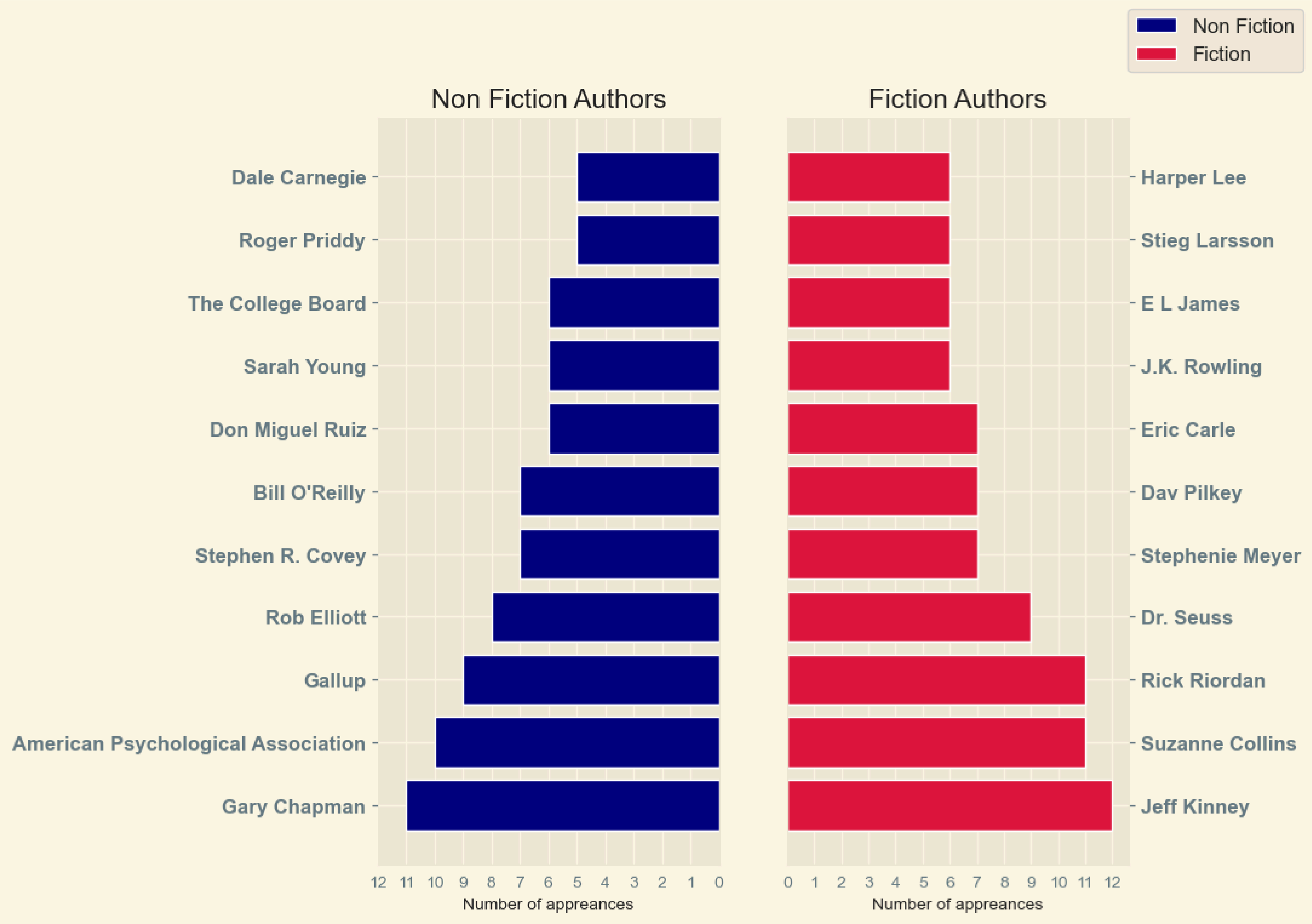
```
C:\Users\HP\AppData\Local\Temp\ipykernel_4992\413210449.py:29: UserWarning: Matplotlib is currently using module://matplotlib_in
line.backend_inline, which is a non-GUI backend, so cannot show the figure.
  fig.show()
```



In [4]:

```
best_nf_authors = df.groupby(['Author', 'Genre']).agg({'Name': 'count'}).unstack()['Name', 'Non Fiction'].sort_values(ascending=F
best_f_authors = df.groupby(['Author', 'Genre']).agg({'Name': 'count'}).unstack()['Name', 'Fiction'].sort_values(ascending=False)

with plt.style.context('Solarize_Light2'):
    fig, ax = plt.subplots(1, 2, figsize=(8,8))

    ax[0].barh(y=best_nf_authors.index, width=best_nf_authors.values,
           color=genre_col[0])
    ax[0].invert_xaxis()
    ax[0].yaxis.tick_left()
    ax[0].set_xticks(np.arange(max(best_f_authors.values)+1))
    ax[0].set_yticklabels(best_nf_authors.index, fontsize=12, fontweight='semibold')
    ax[0].set_xlabel('Number of appreances')
    ax[0].set_title('Non Fiction Authors')

    ax[1].barh(y=best_f_authors.index, width=best_f_authors.values,
           color=genre_col[1])
    ax[1].yaxis.tick_right()
    ax[1].set_xticks(np.arange(max(best_f_authors.values)+1))
    ax[1].set_yticklabels(best_f_authors.index, fontsize=12, fontweight='semibold')
    ax[1].set_title('Fiction Authors')
    ax[1].set_xlabel('Number of appreances')

    fig.legend(['Non Fiction', 'Fiction'], fontsize=12)

plt.show()
```

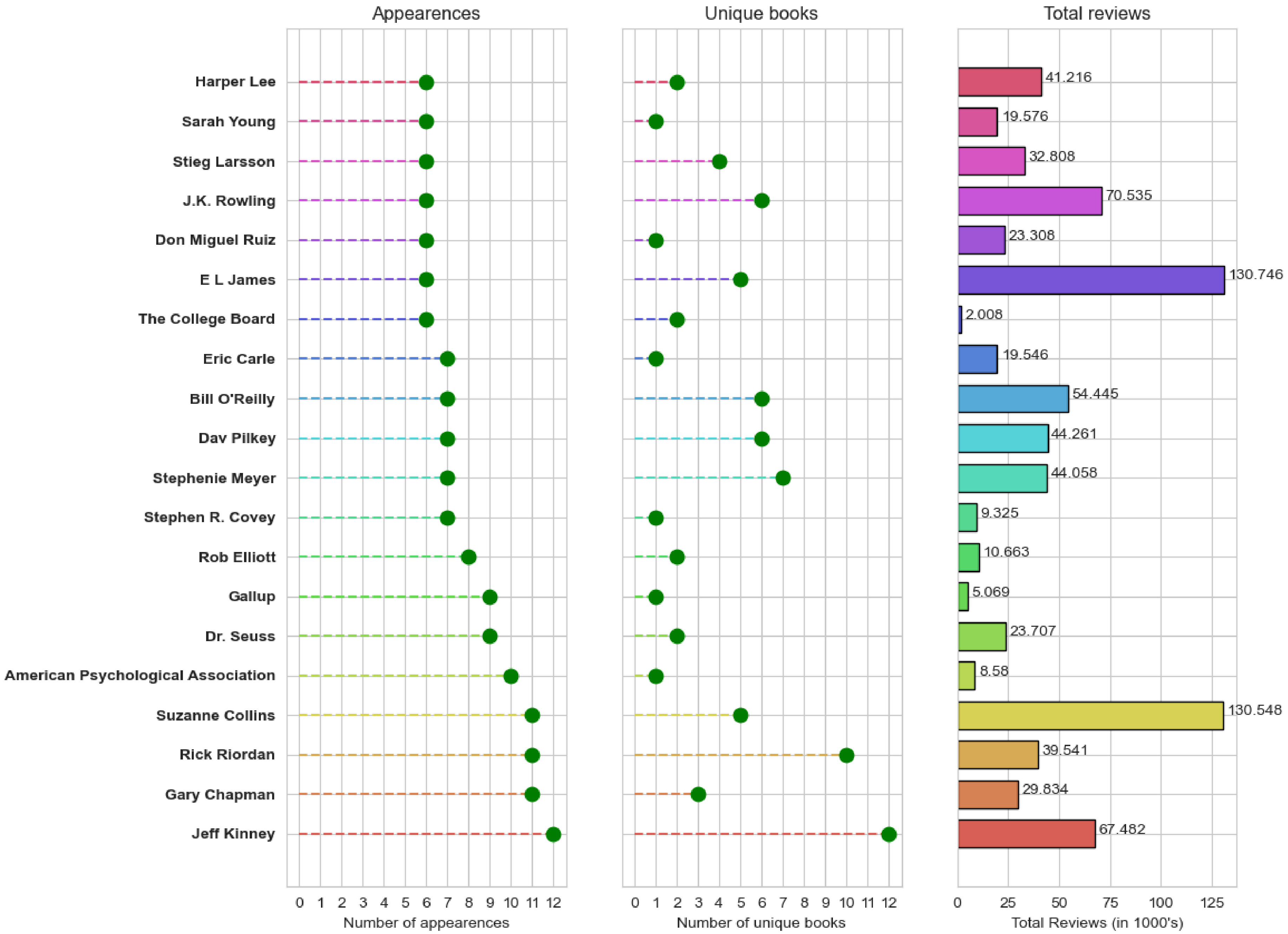
```
C:\Users\HP\AppData\Local\Temp\ipykernel_4992\2335528297.py:12: UserWarning: FixedFormatter should only be used together with Fi
xedLocator
  ax[0].set_yticklabels(best_nf_authors.index, fontsize=12, fontweight='semibold')
C:\Users\HP\AppData\Local\Temp\ipykernel_4992\2335528297.py:20: UserWarning: FixedFormatter should only be used together with Fi
xedLocator
  ax[1].set_yticklabels(best_f_authors.index, fontsize=12, fontweight='semibold')
```


Non Fiction
Fiction
Non Fiction Authors
Fiction Authors
Dale Carnegie
Roger Priddy
The College Board
Sarah Young
Don Miguel Ruiz
Bill O'Reilly
Stephen R. Covey
Rob Elliott
Gallup
American Psychological Association
Gary Chapman
Harper Lee
Stieg Larsson
E L James
J.K. Rowling
Eric Carle
Dav Pilkey
Stephenie Meyer
Dr. Seuss
Rick Riordan
Suzanne Collins
Jeff Kinney
12 11 10 9 8 7 6 5 4 3 2 1 0
0 1 2 3 4 5 6 7 8 9 10 11 12
Number of appreances
Number of appreances

```
In [5]: n_best = 20
```

```
top_authors = df.Author.value_counts().nlargest(n_best)
no_dup = df.drop_duplicates('Name') # removes all rows with duplicate book names

fig, ax = plt.subplots(1, 3, figsize=(11,10), sharey=True)

color = sns.color_palette("hls", n_best)

ax[0].hlines(y=top_authors.index , xmin=0, xmax=top_authors.values, color=color, linestyles='dashed')
ax[0].plot(top_authors.values, top_authors.index, 'go', markersize=9)
ax[0].set_xlabel('Number of appearences')
ax[0].set_xticks(np.arange(top_authors.values.max()+1))
ax[0].set_yticklabels(top_authors.index, fontweight='semibold')
ax[0].set_title('Appearences')

book_count = []
total_reviews = []
for name, col in zip(top_authors.index, color):
    book_count.append(len(no_dup[no_dup.Author == name]['Name']))
    total_reviews.append(no_dup[no_dup.Author == name]['Reviews'].sum()/1000)
ax[1].hlines(y=top_authors.index , xmin=0, xmax=book_count, color=color, linestyles='dashed')
ax[1].plot(book_count, top_authors.index, 'go', markersize=9)
ax[1].set_xlabel('Number of unique books')
ax[1].set_xticks(np.arange(max(book_count)+1))
ax[1].set_title('Unique books')

ax[2].barh(y=top_authors.index, width=total_reviews, color=color, edgecolor='black', height=0.7)
for name, val in zip(top_authors.index, total_reviews):
    ax[2].text(val+2, name, val)
ax[2].set_xlabel("Total Reviews (in 1000's)")
ax[2].set_title('Total reviews')

#plt.suptitle('Top 20 best selling Authors (from 2009 to 2019) details', fontsize=15)
plt.show()
```

```
C:\Users\HP\AppData\Local\Temp\ipykernel_4992\1934722200.py:14: UserWarning: FixedFormatter should only be used together with Fi
xedLocator
  ax[0].set_yticklabels(top_authors.index, fontweight='semibold')
```


Appearences
Unique books
Total reviews
Harper Lee
Sarah Young
Stieg Larsson
J.K. Rowling
Don Miguel Ruiz
E L James
The College Board
Eric Carle
Bill O'Reilly
Dav Pilkey
Stephenie Meyer
Stephen R. Covey
Rob Elliott
Gallup
Dr. Seuss
American Psychological Association
Suzanne Collins
Rick Riordan
Gary Chapman
Jeff Kinney
41.216
19.576
32.808
70.535
23.308
130.746
2.008
19.546
54.445
44.261
44.058
9.325
10.663
5.069
23.707
8.58
130.548
39.541
29.834
67.482
0 1 2 3 4 5 6 7 8 9 10 11 12
Number of appearences
0 1 2 3 4 5 6 7 8 9 10 11 12
Number of unique books
0 25 50 75 100 125
Total Reviews (in 1000's)

In [ ]:

```
Author Jeff Kinney is the best-selling author with 12 appearances in best-selling books from 2009 to 2019
```