In [386]:
```
import pandas as pd
import numpy as mp

%matplotlib inline
import matplotlib.pyplot as plt
import seaborn as sns

df = pd.read_csv("books.csv")
```

Датафрейм содержит информацию о книгах, опубликованных с 1900 года. Источник: https://www.kaggle.com/jealousleopard/goodreadsbooks

Вид:

In [387]:
```
df.head(1)
```

Out[387]:

	bookID	title	authors	average_rating	isbn	isbn13	language_code	pages	ratings_count
0	1	Harry Potter and the Half-Blood Prince (Harry ...	J.K. Rowling/Mary GrandPré	4.57	0439785960	9780439785969	eng	652	2095690

Тип данных:

In [388]:
```
df.dtypes
```

Out[388]:
```
bookID                  int64
title                  object
authors                object
average_rating        float64
isbn                   object
isbn13                  int64
language_code          object
pages                   int64
ratings_count           int64
text_reviews_count      int64
publication_date       object
publisher              object
dtype: object
```

Количество строк и столбцов:

In [389]:
```
df.shape
```

Out[389]:
```
(11125, 12)
```

1: Предположим, что количество выпущенных книг увеличивается с каждым годом.

Создадим новую колонку, где будет храниться только год издания. Для этого в каждой дате из колонки publication_date возьмем только 4 последних символа. Далее группируем по новой колоке, считаем количество книг, которые были изданы, сохраняем результат в новую переменную и рисуем график.

In [390]:
```
df['year'] = [int(str(i)[-4:]) for i in df['publication_date']]
years = df.groupby('year').agg({'year':'count'})
years.plot(title="Зависимость количества книг от года издания")
```

Out[390]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f7a82f8f9b0>
```

По полученному графику видно, что количество книг увеличивалось до начала 2000-х, а затем резко уменьшилось. Гипотеза частично верна.

2: Предположим, что старым (относительного этого датафрейма) книгам не ставили очень низкие оценки.

Для проверки отобразим зависимость среднего рейтинга от года издания.

In [391]:
```
df.plot.scatter(x='year', y='average_rating', c='year', colormap='viridis', title="Зависимость среднего рейтинга от года издания")
```

Out[391]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f7a82f8f080>
```

На графике внизу расположены точки, соответвующие 1980-2020 году и низкому рейтингу, близкому к 0. Самый низкий рейтинг более ранних книг - около 2.8. Гипотеза 2 верна.

3: Предположим, что все книги Джоан Роулинг имеют достаточно высокие оценки.

Для наглядности используем тот же тип графика, что в предыдущей гипотезе. Создадим новую колонку Rowling_books, где True указывает на то, что автор книги Джоан Роулинг.

In [392]:
```
df = df.assign(Rowling_books = df.authors=='J.K. Rowling')
sns.lmplot(x='year', y='average_rating', data = df, hue = 'Rowling_books', fit_reg=False)
```

Out[392]:
```
<seaborn.axisgrid.FacetGrid at 0x7f7a82f85898>
```

Как и ожидалось, книги Джоан Роулинг находятся сверху графика. Нет ни одной книги с плохой оценкой.

4: Предположим, что издательства, которые публикуют книги чаще всего, имеют плохой средний рейтинг книг. Также предположим, что данные издательства публикуют относительно небольшие книги.

Сгруппируем данные по колонке publisher, укажем среднее количество страниц, количество книг и средний рейтинг. Отсортируем по количеству книг по убыванию и возьмем только первые 5 издательств. Отобразим данные на графике.

In [393]:
```
publishers = df.groupby('publisher', as_index = False).agg({'pages':'mean', 'average_rating':'mean', 'bookID':'count'}).rename(columns={'bookID':'count'}).sort_values('count', ascending = False).head(5)
publishers
```

Out[393]:

	publisher	pages	average_rating	count
2129	Vintage	351.402516	3.894182	318
1485	Penguin Books	370.363985	3.920383	261
1502	Penguin Classics	412.875000	3.944565	184
1226	Mariner Books	384.393333	3.933533	150
189	Ballantine Books	394.201389	3.875000	144

На графике видно, что оценка книг издательств от 3.88 до 3.94, что является неплохим рейтингом, гипотеза ложна. Однако гипотеза про количество страниц у книг верна, значение не превышает 500.

In [394]:
```
publishers.iloc[:, :3].plot.line(x='publisher', subplots=True, title="Количество страниц и средний рейтинг в самых популярных издательствах")
```

Out[394]:
```
array([<matplotlib.axes._subplots.AxesSubplot object at 0x7f7a82bb3240>,
       <matplotlib.axes._subplots.AxesSubplot object at 0x7f7a82b40208>], dtype=object)
```