

This is the First part of my journey to become a data scientist to achieve that let's start exploring data set of good reads books dataset and this is obtained from kaggle link

<https://www.kaggle.com/jealousleopard/goodreadsbooks/downloads/goodreadsbooks.zip/6>
<https://www.kaggle.com/jealousleopard/goodreadsbooks/downloads/goodreadsbooks.zip/6>

```
In [0]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings("ignore")
```

```
In [9]: data = pd.read_csv("/content/drive/My Drive/Data_Scientist/EDA/books.csv", error_bad_lines=False)
data.head()
```

b'Skipping line 4012: expected 10 fields, saw 11\nSkipping line 5688: expected 10 fields, saw 11\nSkipping line 7056: expected 10 fields, saw 11\nSkipping line 10601: expected 10 fields, saw 11\nSkipping line 10668: expected 10 fields, saw 11\n'

```
Out[9]:
```

	bookID	title	authors	average_rating	isbn	isbn13	language_code	# num_pages	ratings_count	te
0	1	Harry Potter and the Half-Blood Prince (Harry ...	J.K. Rowling-Mary GrandPré	4.56	0439785960	9780439785969	eng	652	1944099	
1	2	Harry Potter and the Order of the Phoenix (Har...	J.K. Rowling-Mary GrandPré	4.49	0439358078	9780439358071	eng	870	1996446	
2	3	Harry Potter and the Sorcerer's Stone (Harry P...	J.K. Rowling-Mary GrandPré	4.47	0439554934	9780439554930	eng	320	5629932	
3	4	Harry Potter and the Chamber of Secrets (Harry...	J.K. Rowling	4.41	0439554896	9780439554893	eng	352	6267	
4	5	Harry Potter and the Prisoner of Azkaban (Harr...	J.K. Rowling-Mary GrandPré	4.55	043965548X	9780439655484	eng	435	2149872	

```
In [11]: print(data.shape)

(13714, 10)
```

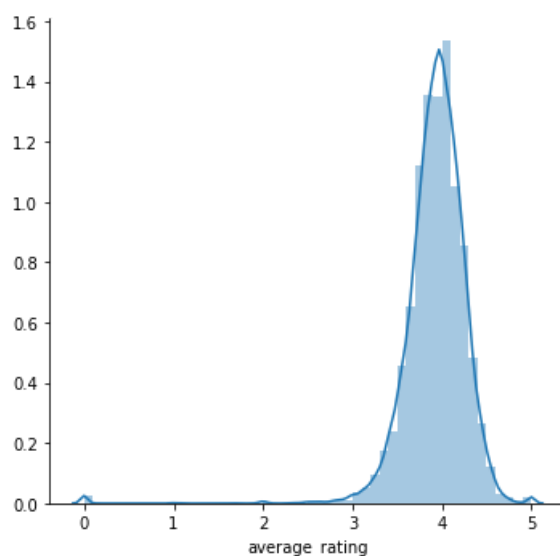
It is a dataset containing 13714 records with 10 features

```
In [26]: data.describe()
```

```
Out[26]:
```

	bookID	average_rating	isbn13	# num_pages	ratings_count	text_reviews_count
count	13714.000000	13714.000000	1.371400e+04	13714.000000	1.371400e+04	13714.000000
mean	22159.859195	3.930620	9.764017e+12	342.402727	1.776540e+04	533.632128
std	13700.926816	0.357893	3.987679e+11	252.650165	1.129572e+05	2529.006691
min	1.000000	0.000000	8.987060e+09	0.000000	0.000000e+00	0.000000
25%	10619.250000	3.770000	9.780345e+12	196.000000	8.300000e+01	7.000000
50%	21321.500000	3.960000	9.780613e+12	301.000000	6.305000e+02	40.000000
75%	33311.750000	4.130000	9.780940e+12	421.000000	4.742250e+03	222.000000
max	47709.000000	5.000000	9.790008e+12	6576.000000	5.629932e+06	93619.000000

```
In [16]: sns.FacetGrid(data,size=5)\
.map(sns.distplot,"average_rating")\
.add_legend();
plt.show();
```

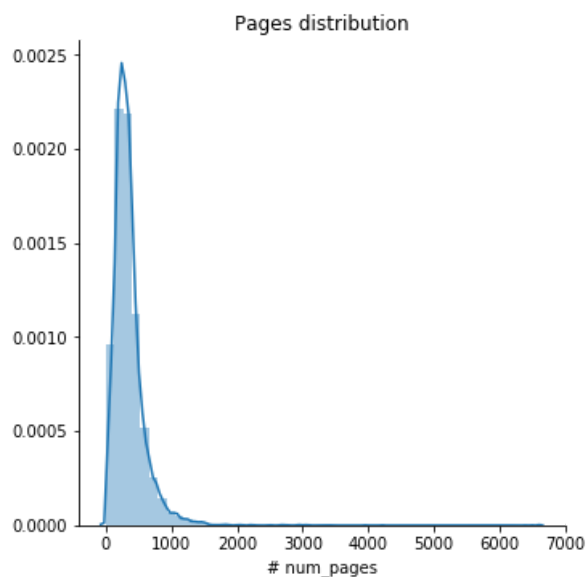


The most of the average rating presents between 3 - 5

```
In [20]: data.columns
```

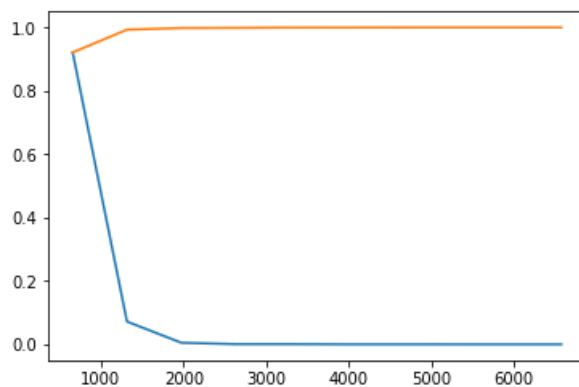
```
Out[20]: Index(['bookID', 'title', 'authors', 'average_rating', 'isbn', 'isbn13',
               'language_code', '# num_pages', 'ratings_count', 'text_reviews_count'],
              dtype='object')
```

```
In [24]: sns.FacetGrid(data,size=5)\
.map(sns.distplot,"# num_pages")\
.add_legend();
plt.title("Pages distribution");
plt.show();
```



```
In [87]: counts,bin_edges = np.histogram(data['# num_pages'],bins=10,density=True)
pdf = counts/(sum(counts))
cdf = np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf)
plt.plot(bin_edges[1:],cdf)
plt.show();
```

```
[9.21175441e-01 7.18973312e-02 5.10427301e-03 6.56263672e-04
 7.29181858e-04 1.45836372e-04 1.45836372e-04 7.29181858e-05
 0.00000000e+00 7.29181858e-05]
[ 0.    657.6 1315.2 1972.8 2630.4 3288.   3945.6 4603.2 5260.8 5918.4
 6576. ]
```

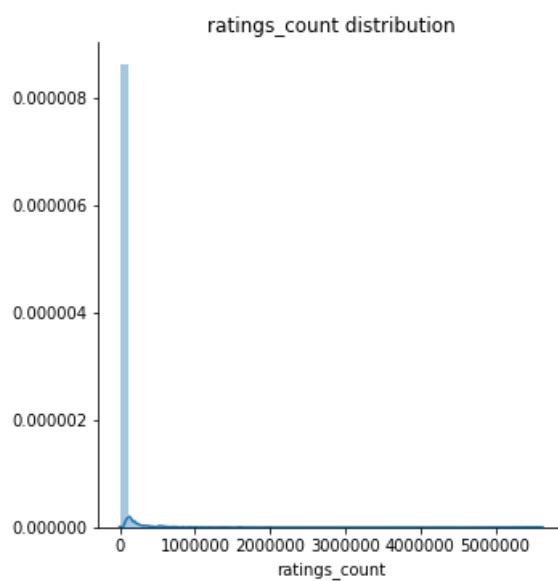


```
In [91]: num_pages_list = list(data['# num_pages'])
num_pages_list.sort()
print("99 th percentile value is",np.percentile(data['# num_pages'],99))
```

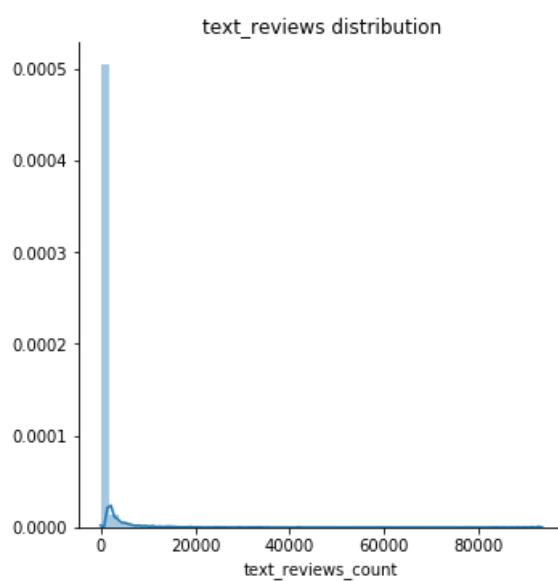
99 th percentile value is 1200.0

1% of the books are more than 1200 pages

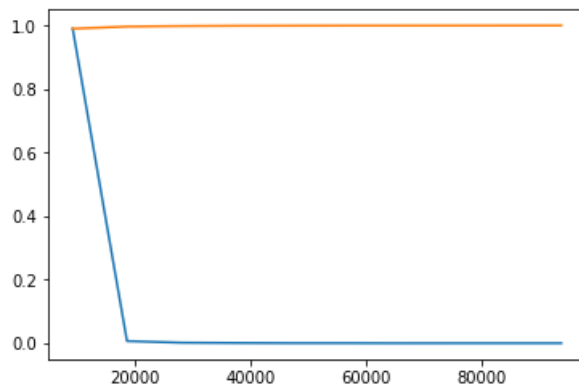
```
In [31]: sns.FacetGrid(data,size=5)\
.map(sns.distplot,"ratings_count")\
.add_legend();
plt.title("ratings_count distribution");
plt.show();
```



```
In [29]: sns.FacetGrid(data,size=5)\
.map(sns.distplot,"text_reviews_count")\
.add_legend();
plt.title("text_reviews distribution");
plt.show();
```



```
In [92]: counts,bin_edges = np.histogram(data['text_reviews_count'],bins=10,density=True)
pdf = counts/(sum(counts))
cdf = np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf)
plt.plot(bin_edges[1:],cdf)
plt.show();
```



```
In [97]: num_pages_list = list(data['text_reviews_count'])
num_pages_list.sort()
print("99.5 th percentile value is",np.percentile(data['text_reviews_count'],99.5))
```

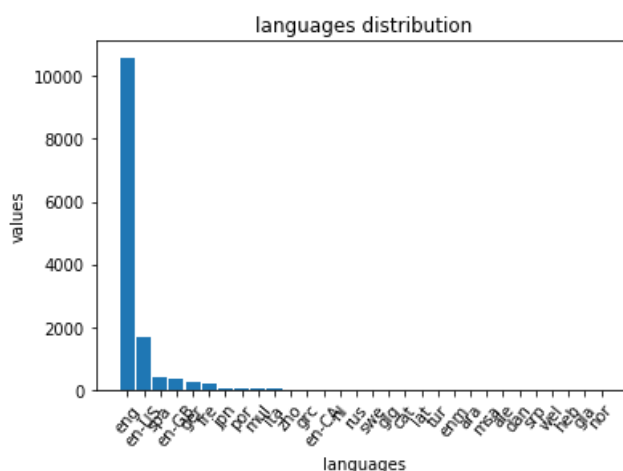
99.5 th percentile value is 14443.609999999997

0.5 percentile of the text_reviews_count is greater than 15000

```
In [42]: lang_list = data['language_code'].value_counts()
lang_list_index = list(lang_list.index)
lang_lis_values = []
for i in lang_list :
    lang_lis_values.append(i)
print(lang_list_index)
print(lang_lis_values)
```

```
['eng', 'en-US', 'spa', 'en-GB', 'ger', 'fre', 'jpn', 'por', 'mul', 'ita', 'zho', 'grc',
'en-CA', 'nl', 'rus', 'swe', 'glg', 'cat', 'lat', 'tur', 'enm', 'ara', 'msa', 'ale', 'da
n', 'srp', 'wel', 'heb', 'gla', 'nor']
[10594, 1699, 419, 341, 238, 209, 64, 27, 21, 19, 16, 12, 9, 7, 7, 6, 4, 3, 3, 3, 3, 2, 1,
1, 1, 1, 1, 1, 1]
```

```
In [65]: plt.bar(lang_list_index,lang_lis_values,width=0.9)
plt.xlabel('values', fontsize=10)
plt.ylabel('languages', fontsize=10)
plt.xticks(lang_list_index, fontsize=10, rotation=50)
plt.title("languages distribution");
plt.show();
```

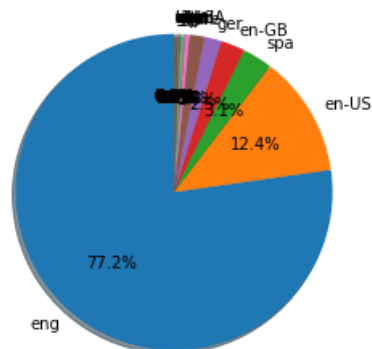


```
In [86]: # pie plot : https://pythonspot.com/matplotlib-pie-chart/

import matplotlib.pyplot as plt

# Plot
plt.pie(lang_lis_values, labels=lang_list_index,
autopct='%0.1f%%', shadow=True, startangle=90)

plt.axis('equal')
plt.show()
```



77.2% of the books are english language based

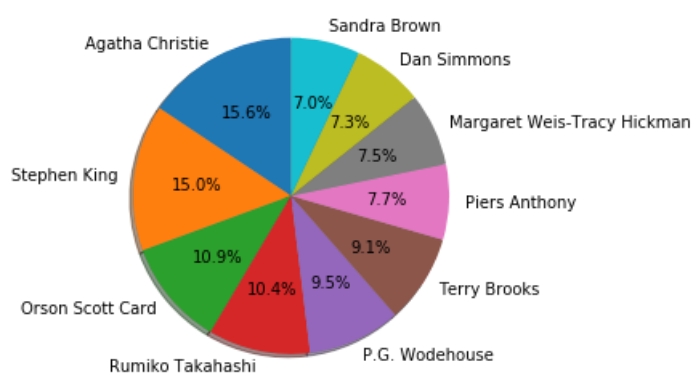
```
In [0]: authors_list = data['authors'].value_counts()
authors_list = authors_list[:10]
authors_list_index = list(authors_list.index)
authors_lis_values = []
for i in authors_list :
    authors_lis_values.append(i)
```

```
In [107]: # pie plot : https://pythonspot.com/matplotlib-pie-chart/

import matplotlib.pyplot as plt

# Plot
plt.pie(authors_lis_values, labels=authors_list_index,
autopct='%0.1f%%', shadow=True, startangle=90)

plt.axis('equal')
plt.show()
```



The above picture gives us the top 10 authors of the given data set

```
In [115]: !pip install tqdm

Requirement already satisfied: tqdm in /usr/local/lib/python3.6/dist-packages (4.28.1)
```

```
In [0]: from tqdm import tqdm
```

```
In [0]: # https://stackoverflow.com/a/47091490/4084039
import re

def decontracted(phrase):
    # specific
    phrase = re.sub(r"won't", "will not", phrase)
    phrase = re.sub(r"can't", "can not", phrase)

    # general
    phrase = re.sub(r"\n't", " not", phrase)
    phrase = re.sub(r"\ 're", " are", phrase)
    phrase = re.sub(r"\ 's", " is", phrase)
    phrase = re.sub(r"\ 'd", " would", phrase)
    phrase = re.sub(r"\ 'll", " will", phrase)
    phrase = re.sub(r"\ 't", " not", phrase)
    phrase = re.sub(r"\ 've", " have", phrase)
    phrase = re.sub(r"\ 'm", " am", phrase)
    return phrase
```

```
In [122]: preprocessed_text_values = []

for sentence in tqdm(data['title'].values) :
    term = decontracted(sentence)
    term = term.replace(")", "")
    term = term.replace("(", "")
    term = term.replace("&", "_")
    term = term.replace("#", "")
    preprocessed_text_values.append(term.lower().strip())

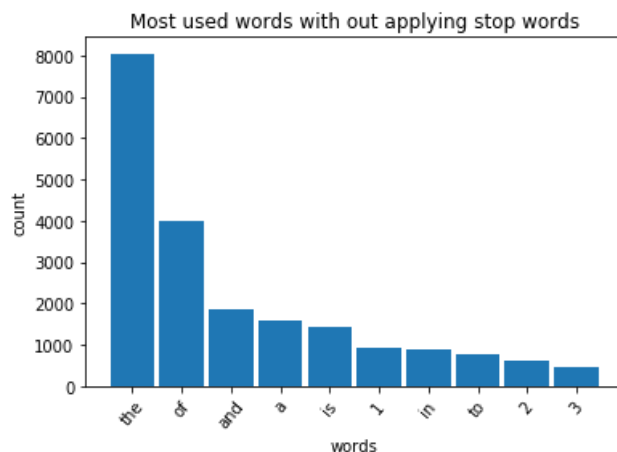
100%|██████████| 13714/13714 [00:00<00:00, 73761.77it/s]
```

```
In [0]: from collections import Counter
my_counter = Counter()
for word in preprocessed_text_values:
    my_counter.update(word.split())
```

```
In [0]: sorted_my_counter = sorted(my_counter.items(), key=lambda kv: kv[1], reverse=True)
```

```
In [0]: for i in sorted_my_counter[0:10] :
    words.append(i[0])
    word_count.append(i[1])
```

```
In [149]: plt.bar(words[0:10], word_count[0:10], width=0.9)
plt.xlabel('words', fontsize=10)
plt.ylabel('count', fontsize=10)
plt.xticks(words[0:10], fontsize=10, rotation=50)
plt.title("Most used words with out applying stop words");
plt.show();
```



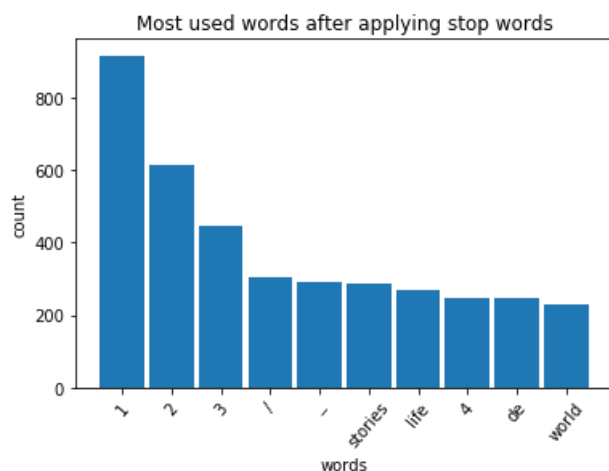
```
In [0]: # https://gist.github.com/sebleier/554280
# we are removing the words from the stop words list: 'no', 'nor', 'not'
stopwords= ['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "
you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his',
'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself', 'they',
'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'l
'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had',
'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'u
'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through', 'd
'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over',
'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', '
'most', 'other', 'some', 'such', 'only', 'own', 'same', 'so', 'than', 'too', 'v
's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now',
've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'do
'hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "isn't", 'ma', 'mightn',
'mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't", 'wasn'
'won', "won't", 'wouldn', "wouldn't"]
```

```
In [154]: print(sorted_my_counter)

[('a', 1578), ('1', 917), ('2', 615), ('3', 444), ('/', 306), ('_', 290), ('stories', 28
7), ('life', 269), ('4', 248), ('de', 247), ('world', 231), ('vol.', 229), ('guide', 22
3), ('history', 205), ('la', 197), ('book', 189), ('volume', 189), ('love', 176), ('ne
w', 170), ('5', 167), ('tales', 167), ('story', 165), ('complete', 161), ('time', 160),
('man', 153), ('war', 150), ('6', 147), ('american', 136), ('one', 131), ('el', 129),
('house', 125), ('trilogy', 123), ('death', 116), ('great', 115), ('black', 113), ('litt
le', 113), ('night', 111), ('star', 111), ('7', 109), ('america', 105), ('dark', 103),
('lost', 102), ('last', 102), ('lord', 98), ('harry', 93), ('not', 93), ('art', 93), ('t
hree', 93), ('short', 88), ('chronicles', 88), ('women', 88), ('king', 88), ('8', 82),
('no', 78), ('poems', 78), ('girl', 77), ('die', 75), ('magic', 75), ('vampire', 74),
('children', 73), ('like', 73), ('adventures', 72), ('first', 72), ('secret', 72), ('mys
tery', 72), ('family', 71), ('best', 70), ('essays', 70), ('writings', 69), ('god', 69),
('philosophy', 68), ('murder', 68), ('potter', 67), ('saga', 65), ('city', 64), ('whit
e', 64), ('der', 64), ('collected', 63), ('tale', 63), ('1:', 63), ('science', 62), ('se
lected', 62), ('woman', 61), ('five', 60), ('rings', 60), ('true', 60), ('1-3', 59), ('m
odern', 59), ('10', 59), ('people', 59), ('novels', 58), ('journey', 58), ('two', 58),
('jack', 58), ('john', 57), ('dead', 56), ('9', 56), ('sea', 55), ('shadow', 55), ('hear
t', 55), ('good', 54), ('long', 54), ('red', 54), ('years', 53), ('-', 53), ('big', 53),
('power', 53), ('books', 52), ('moon', 52), ('way', 51), ('letters', 51), ('fire', 51),
('light', 51), ('11', 51), ('dead', 50), ('1-2', 50), ('1-3', 50), ('1-4', 50), ('1-5', 50),
('1-6', 50), ('1-7', 50), ('1-8', 50), ('1-9', 50), ('1-10', 50), ('1-11', 50), ('1-12', 50),
('1-13', 50), ('1-14', 50), ('1-15', 50), ('1-16', 50), ('1-17', 50), ('1-18', 50), ('1-19', 50),
('1-20', 50), ('1-21', 50), ('1-22', 50), ('1-23', 50), ('1-24', 50), ('1-25', 50), ('1-26', 50),
('1-27', 50), ('1-28', 50), ('1-29', 50), ('1-30', 50), ('1-31', 50), ('1-32', 50), ('1-33', 50),
('1-34', 50), ('1-35', 50), ('1-36', 50), ('1-37', 50), ('1-38', 50), ('1-39', 50), ('1-40', 50),
('1-41', 50), ('1-42', 50), ('1-43', 50), ('1-44', 50), ('1-45', 50), ('1-46', 50), ('1-47', 50),
('1-48', 50), ('1-49', 50), ('1-50', 50), ('1-51', 50), ('1-52', 50), ('1-53', 50), ('1-54', 50),
('1-55', 50), ('1-56', 50), ('1-57', 50), ('1-58', 50), ('1-59', 50), ('1-60', 50), ('1-61', 50),
('1-62', 50), ('1-63', 50), ('1-64', 50), ('1-65', 50), ('1-66', 50), ('1-67', 50), ('1-68', 50),
('1-69', 50), ('1-70', 50), ('1-71', 50), ('1-72', 50), ('1-73', 50), ('1-74', 50), ('1-75', 50),
('1-76', 50), ('1-77', 50), ('1-78', 50), ('1-79', 50), ('1-80', 50), ('1-81', 50), ('1-82', 50),
('1-83', 50), ('1-84', 50), ('1-85', 50), ('1-86', 50), ('1-87', 50), ('1-88', 50), ('1-89', 50),
('1-90', 50), ('1-91', 50), ('1-92', 50), ('1-93', 50), ('1-94', 50), ('1-95', 50), ('1-96', 50),
('1-97', 50), ('1-98', 50), ('1-99', 50), ('1-100', 50), ('1-101', 50), ('1-102', 50), ('1-103', 50),
('1-104', 50), ('1-105', 50), ('1-106', 50), ('1-107', 50), ('1-108', 50), ('1-109', 50), ('1-110', 50),
('1-111', 50), ('1-112', 50), ('1-113', 50), ('1-114', 50), ('1-115', 50), ('1-116', 50), ('1-117', 50),
('1-118', 50), ('1-119', 50), ('1-120', 50), ('1-121', 50), ('1-122', 50), ('1-123', 50), ('1-124', 50),
('1-125', 50), ('1-126', 50), ('1-127', 50), ('1-128', 50), ('1-129', 50), ('1-130', 50), ('1-131', 50),
('1-132', 50), ('1-133', 50), ('1-134', 50), ('1-135', 50), ('1-136', 50), ('1-137', 50), ('1-138', 50),
('1-139', 50), ('1-140', 50), ('1-141', 50), ('1-142', 50), ('1-143', 50), ('1-144', 50), ('1-145', 50),
('1-146', 50), ('1-147', 50), ('1-148', 50), ('1-149', 50), ('1-150', 50), ('1-151', 50), ('1-152', 50),
('1-153', 50), ('1-154', 50), ('1-155', 50), ('1-156', 50), ('1-157', 50), ('1-158', 50), ('1-159', 50),
('1-160', 50), ('1-161', 50), ('1-162', 50), ('1-163', 50), ('1-164', 50), ('1-165', 50), ('1-166', 50),
('1-167', 50), ('1-168', 50), ('1-169', 50), ('1-170', 50), ('1-171', 50), ('1-172', 50), ('1-173', 50),
('1-174', 50), ('1-175', 50), ('1-176', 50), ('1-177', 50), ('1-178', 50), ('1-179', 50), ('1-180', 50),
('1-181', 50), ('1-182', 50), ('1-183', 50), ('1-184', 50), ('1-185', 50), ('1-186', 50), ('1-187', 50),
('1-188', 50), ('1-189', 50), ('1-190', 50), ('1-191', 50), ('1-192', 50), ('1-193', 50), ('1-194', 50),
('1-195', 50), ('1-196', 50), ('1-197', 50), ('1-198', 50), ('1-199', 50), ('1-200', 50), ('1-201', 50),
('1-202', 50), ('1-203', 50), ('1-204', 50), ('1-205', 50), ('1-206', 50), ('1-207', 50), ('1-208', 50),
('1-209', 50), ('1-210', 50), ('1-211', 50), ('1-212', 50), ('1-213', 50), ('1-214', 50), ('1-215', 50),
('1-216', 50), ('1-217', 50), ('1-218', 50), ('1-219', 50), ('1-220', 50), ('1-221', 50), ('1-222', 50),
('1-223', 50), ('1-224', 50), ('1-225', 50), ('1-226', 50), ('1-227', 50), ('1-228', 50), ('1-229', 50),
('1-230', 50), ('1-231', 50), ('1-232', 50), ('1-233', 50), ('1-234', 50), ('1-235', 50), ('1-236', 50),
('1-237', 50), ('1-238', 50), ('1-239', 50), ('1-240', 50), ('1-241', 50), ('1-242', 50), ('1-243', 50),
('1-244', 50), ('1-245', 50), ('1-246', 50), ('1-247', 50), ('1-248', 50), ('1-249', 50), ('1-250', 50),
('1-251', 50), ('1-252', 50), ('1-253', 50), ('1-254', 50), ('1-255', 50), ('1-256', 50), ('1-257', 50),
('1-258', 50), ('1-259', 50), ('1-260', 50), ('1-261', 50), ('1-262', 50), ('1-263', 50), ('1-264', 50),
('1-265', 50), ('1-266', 50), ('1-267', 50), ('1-268', 50), ('1-269', 50), ('1-270', 50), ('1-271', 50),
('1-272', 50), ('1-273', 50), ('1-274', 50), ('1-275', 50), ('1-276', 50), ('1-277', 50), ('1-278', 50),
('1-279', 50), ('1-280', 50), ('1-281', 50), ('1-282', 50), ('1-283', 50), ('1-284', 50), ('1-285', 50),
('1-286', 50), ('1-287', 50), ('1-288', 50), ('1-289', 50), ('1-290', 50), ('1-291', 50), ('1-292', 50),
('1-293', 50), ('1-294', 50), ('1-295', 50), ('1-296', 50), ('1-297', 50), ('1-298', 50), ('1-299', 50),
('1-300', 50), ('1-301', 50), ('1-302', 50), ('1-303', 50), ('1-304', 50), ('1-305', 50), ('1-306', 50),
('1-307', 50), ('1-308', 50), ('1-309', 50), ('1-310', 50), ('1-311', 50), ('1-312', 50), ('1-313', 50),
('1-314', 50), ('1-315', 50), ('1-316', 50), ('1-317', 50), ('1-318', 50), ('1-319', 50), ('1-320', 50),
('1-321', 50), ('1-322', 50), ('1-323', 50), ('1-324', 50), ('1-325', 50), ('1-326', 50), ('1-327', 50),
('1-328', 50), ('1-329', 50), ('1-330', 50), ('1-331', 50), ('1-332', 50), ('1-333', 50), ('1-334', 50),
('1-335', 50), ('1-336', 50), ('1-337', 50), ('1-338', 50), ('1-339', 50), ('1-340', 50), ('1-341', 50),
('1-342', 50), ('1-343', 50), ('1-344', 50), ('1-345', 50), ('1-346', 50), ('1-347', 50), ('1-348', 50),
('1-349', 50), ('1-350', 50), ('1-351', 50), ('1-352', 50), ('1-353', 50), ('1-354', 50), ('1-355', 50),
('1-356', 50), ('1-357', 50), ('1-358', 50), ('1-359', 50), ('1-360', 50), ('1-361', 50), ('1-362', 50),
('1-363', 50), ('1-364', 50), ('1-365', 50), ('1-366', 50), ('1-367', 50), ('1-368', 50), ('1-369', 50),
('1-370', 50), ('1-371', 50), ('1-372', 50), ('1-373', 50), ('1-374', 50), ('1-375', 50), ('1-376', 50),
('1-377', 50), ('1-378', 50), ('1-379', 50), ('1-380', 50), ('1-381', 50), ('1-382', 50), ('1-383', 50),
('1-384', 50), ('1-385', 50), ('1-386', 50), ('1-387', 50), ('1-388', 50), ('1-389', 50), ('1-390', 50),
('1-391', 50), ('1-392', 50), ('1-393', 50), ('1-394', 50), ('1-395', 50), ('1-396', 50), ('1-397', 50),
('1-398', 50), ('1-399', 50), ('1-400', 50), ('1-401', 50), ('1-402', 50), ('1-403', 50), ('1-404', 50),
('1-405', 50), ('1-406', 50), ('1-407', 50), ('1-408', 50), ('1-409', 50), ('1-410', 50), ('1-411', 50),
('1-412', 50), ('1-413', 50), ('1-414', 50), ('1-415', 50), ('1-416', 50), ('1-417', 50), ('1-418', 50),
('1-419', 50), ('1-420', 50), ('1-421', 50), ('1-422', 50), ('1-423', 50), ('1-424', 50), ('1-425', 50),
('1-426', 50), ('1-427', 50), ('1-428', 50), ('1-429', 50), ('1-430', 50), ('1-431', 50), ('1-432', 50),
('1-433', 50), ('1-434', 50), ('1-435', 50), ('1-436', 50), ('1-437', 50), ('1-438', 50), ('1-439', 50),
('1-440', 50), ('1-441', 50), ('1-442', 50), ('1-443', 50), ('1-444', 50), ('1-445', 50), ('1-446', 50),
('1-447', 50), ('1-448', 50), ('1-449', 50), ('1-450', 50), ('1-451', 50), ('1-452', 50), ('1-453', 50),
('1-454', 50), ('1-455', 50), ('1-456', 50), ('1-457', 50), ('1-458', 50), ('1-459', 50), ('1-460', 50),
('1-461', 50), ('1-462', 50), ('1-463', 50), ('1-464', 50), ('1-465', 50), ('1-466', 50), ('1-467', 50),
('1-468', 50), ('1-469', 50), ('1-470', 50), ('1-471', 50), ('1-472', 50), ('1-473', 50), ('1-474', 50),
('1-475', 50), ('1-476', 50), ('1-477', 50), ('1-478', 50), ('1-479', 50), ('1-480', 50), ('1-481', 50),
('1-482', 50), ('1-483', 50), ('1-484', 50), ('1-485', 50), ('1-486', 50), ('1-487', 50), ('1-488', 50),
('1-489', 50), ('1-490', 50), ('1-491', 50), ('1-492', 50), ('1-493', 50), ('1-494', 50), ('1-495', 50),
('1-496', 50), ('1-497', 50), ('1-498', 50), ('1-499', 50), ('1-500', 50), ('1-501', 50), ('1-502', 50),
('1-503', 50), ('1-504', 50), ('1-505', 50), ('1-506', 50), ('1-507', 50), ('1-508', 50), ('1-509', 50),
('1-510', 50), ('1-511', 50), ('1-512', 50), ('1-513', 50), ('1-514', 50), ('1-515', 50), ('1-516', 50),
('1-517', 50), ('1-518', 50), ('1-519', 50), ('1-520', 50), ('1-521', 50), ('1-522', 50), ('1-523', 50),
('1-524', 50), ('1-525', 50), ('1-526', 50), ('1-527', 50), ('1-528', 50), ('1-529', 50), ('1-530', 50),
('1-531', 50), ('1-532', 50), ('1-533', 50), ('1-534', 50), ('1-535', 50), ('1-536', 50), ('1-537', 50),
('1-538', 50), ('1-539', 50), ('1-540', 50), ('1-541', 50), ('1-542', 50), ('1-543', 50), ('1-544', 50),
('1-545', 50), ('1-546', 50), ('1-547', 50), ('1-548', 50), ('1-549', 50), ('1-550', 50), ('1-551', 50),
('1-552', 50), ('1-553', 50), ('1-554', 50), ('1-555', 50), ('1-556', 50), ('1-557', 50), ('1-558', 50),
('1-559', 50), ('1-560', 50), ('1-561', 50), ('1-562', 50), ('1-563', 50), ('1-564', 50), ('1-565', 50),
('1-566', 50), ('1-567', 50), ('1-568', 50), ('1-569', 50), ('1-570', 50), ('1-571', 50), ('1-572', 50),
('1-573', 50), ('1-574', 50), ('1-575', 50), ('1-576', 50), ('1-577', 50), ('1-578', 50), ('1-579', 50),
('1-580', 50), ('1-581', 50), ('1-582', 50), ('1-583', 50), ('1-584', 50), ('1-585', 50), ('1-586', 50),
('1-587', 50), ('1-588', 50), ('1-589', 50), ('1-590', 50), ('1-591', 50), ('1-592', 50), ('1-593', 50),
('1-594', 50), ('1-595', 50), ('1-596', 50), ('1-597', 50), ('1-598', 50), ('1-599', 50), ('1-600', 50),
('1-601', 50), ('1-602', 50), ('1-603', 50), ('1-604', 50), ('1-605', 50), ('1-606', 50), ('1-607', 50),
('1-608', 50), ('1-609', 50), ('1-610', 50), ('1-611', 50), ('1-612', 50), ('1-613', 50), ('1-614', 50),
('1-615', 50), ('1-616', 50), ('1-617', 50), ('1-618', 50), ('1-619', 50), ('1-620', 50), ('1-621', 50),
('1-622', 50), ('1-623', 50), ('1-624', 50), ('1-625', 50), ('1-626', 50), ('1-627', 50), ('1-628', 50),
('1-629', 50), ('1-630', 50), ('1-631', 50), ('1-632', 50), ('1-633', 50), ('1-634', 50), ('1-635', 50),
('1-636', 50), ('1-637', 50), ('1-638', 50), ('1-639', 50), ('1-640', 50), ('1-641', 50), ('1-642', 50),
('1-643', 50), ('1-644', 50), ('1-645', 50), ('1-646', 50), ('1-647', 50), ('1-648', 50), ('1-649', 50),
('1-650', 50), ('1-651', 50), ('1-652', 50), ('1-653', 50), ('1-654', 50), ('1-655', 50), ('1-656', 50),
('1-657', 50), ('1-658', 50), ('1-659', 50), ('1-660', 50), ('1-661', 50), ('1-662', 50), ('1-663', 50),
('1-664', 50), ('1-665', 50), ('1-666', 50), ('1-667', 50), ('1-668', 50), ('1-669', 50), ('1-670', 50),
('1-671', 50), ('1-672', 50), ('1-673', 50), ('1-674', 50), ('1-675', 50), ('1-676', 50), ('1-677', 50),
('1-678', 50), ('1-679', 50), ('1-680', 50), ('1-681', 50), ('1-682', 50), ('1-683', 50), ('1-684', 50),
('1-685', 50), ('1-686', 50), ('1-687', 50), ('1-688', 50), ('1-689', 50), ('1-690', 50), ('1-691', 50),
('1-692', 50), ('1-693', 50), ('1-694', 50), ('1-695', 50), ('1-696', 50), ('1-697', 50), ('1-698', 50),
('1-699', 50), ('1-700', 50), ('1-701', 50), ('1-702', 50), ('1-703', 50), ('1-704', 50), ('1-705', 50),
('1-706', 50), ('1-707', 50), ('1-708', 50), ('1-709', 50), ('1-710', 50), ('1-711', 50), ('1-712', 50),
('1-713', 50), ('1-714', 50), ('1-715', 50), ('1-716', 50), ('1-717', 50), ('1-718', 50), ('1-719', 50),
('1-720', 50), ('1-721', 50), ('1-722', 50), ('1-723', 50), ('1-724', 50), ('1-725', 50), ('1-726', 50),
('1-727', 50), ('1-728', 50), ('1-729', 50), ('1-730', 50), ('1-731', 50), ('1-732', 50), ('1-733', 50),
('1-734', 50), ('1-735', 50), ('1-736', 50), ('1-737', 50), ('1-738', 50), ('1-739', 50), ('1-740', 50),
('1-741', 50), ('1-742', 50), ('1-743', 50), ('1-744', 50), ('1-745', 50), ('1-746', 50), ('1-747', 50),
('1-748', 50), ('1-749', 50), ('1-750', 50), ('1-751', 50), ('1-752', 50), ('1-753', 50), ('1-754', 50),
('1-755', 50), ('1-756', 50), ('1-757', 50), ('1-758', 50), ('1-759', 50), ('1-760', 50), ('1-761', 50),
('1-762', 50), ('1-763', 50), ('1-764', 50), ('1-765', 50), ('1-766', 50), ('1-767', 50), ('1-768', 50),
('1-769', 50), ('1-770', 50), ('1-771', 50), ('1-772', 50), ('1-773', 50), ('1-774', 50), ('1-775', 50),
('1-776', 50), ('1-777', 50), ('1-778', 50), ('1-779', 50), ('1-780', 50), ('1-781', 50), ('1-782', 50),
('1-783', 50), ('1-784', 50), ('1-785', 50), ('1-786', 50), ('1-787', 50), ('1-788', 50), ('1-789', 50),
('1-790', 50), ('1-791', 50), ('1-792', 50), ('1-793', 50), ('1-794', 50), ('1-795', 50), ('1-796', 50),
('1-797', 50), ('1-798', 50), ('1-799', 50), ('1-800', 50), ('1-801', 50), ('1-802', 50), ('1-803', 50),
('1-804', 50), ('1-805', 50), ('1-806', 50), ('1-807', 50), ('1-808', 50), ('1-809', 50), ('1-810', 50),
('1-811', 50), ('1-812', 50), ('1-813', 50), ('1-814', 50), ('1-815', 50), ('1-816', 50), ('1-817', 50),
('1-818', 50), ('1-819', 50), ('1-820', 50), ('1-821', 50), ('1-822', 50), ('1-823', 50), ('1-824', 50),
('1-825', 50), ('1-826', 50), ('1-827', 50), ('1-828', 50), ('1-829', 50), ('1-830', 50), ('1-831', 50),
('1-832', 50), ('1-833', 50), ('1-834', 50), ('1-835', 50), ('1-836', 50), ('1-837', 50), ('1-838', 50),
('1-839', 50), ('1-840', 50), ('1-841', 50), ('1-842', 50), ('1-843', 50), ('1-844', 50), ('1-845', 50),
('1-846', 50), ('1-847', 50), ('1-848', 50), ('1-849', 50), ('1-850', 50), ('1-851', 50), ('1-852', 50),
('1-853', 50), ('1-854', 50), ('1-855', 50), ('1-856', 50), ('1-857', 50), ('1-858', 50), ('1-859', 50),
('1-860', 50), ('1-861', 50), ('1-862', 50), ('1-863', 50), ('1-864', 50), ('1-865', 50), ('1-866', 50),
('1-867', 50), ('1-868', 50), ('1-869', 50), ('1-870', 50), ('1-871', 50), ('1-872', 50), ('1-873', 50),
('1-874', 50), ('1-875', 50), ('1-876', 50), ('1-877', 50), ('1-878', 50), ('1-879', 50), ('1-880', 50),
('1-881', 50), ('1-882', 50), ('1-883', 50), ('1-884', 50), ('1-885', 50), ('1-886', 50), ('1-887', 50),
('1-888', 50), ('1-889', 50), ('1-890', 50), ('1-891', 50), ('1-892', 50), ('1-893', 50), ('1-894', 50),
('1-895', 50), ('1-896', 50), ('1-897', 50), ('1-898', 50), ('1-899', 50), ('1-900', 50), ('1-901', 50),
('1-902', 50), ('1-903', 50), ('1-904', 50), ('1-905', 50), ('1-906', 50), ('1-907', 50), ('1-908', 50),
('1-909', 50), ('1-910', 50), ('1-911', 50), ('1-912', 50), ('1-913', 50), ('1-914', 50), ('1-915', 50),
('1-916', 50), ('1-917', 50), ('1-918', 50), ('1-919', 50), ('1-920', 50), ('1-921', 50), ('1-922', 50),
('1-923', 50), ('1-924', 50), ('1-925', 50), ('1-926', 50), ('1-927', 50), ('1-928', 50), ('1-929', 50),
('1-930', 50), ('1-931', 50), ('1-932', 50), ('1-933', 50), ('1-934', 50), ('1-935', 50), ('1-936', 50),
('1-937', 50), ('1-938', 50), ('1-939', 50), ('1-940', 50), ('1-941', 50), ('1-942', 50), ('1-943', 50),
('1-944', 50), ('1-945', 50), ('1-946', 50), ('1-947', 50), ('1-948', 50), ('1-949', 50), ('1-950', 50),
('1-951', 50), ('1-952', 50), ('1-953', 50), ('1-954', 50), ('1-955', 50), ('1-956', 50), ('1-957', 50),
('1-958', 50), ('1-959', 50), ('1-960', 50), ('1-961', 50), ('1-962', 50), ('1-963', 50), ('1-964', 50),
('1-965', 50), ('1-966', 50), ('1-967', 50), ('1-968', 50), ('1-969', 50), ('1-970', 50), ('1-971', 50),
('1-972', 50), ('1-973', 50), ('1-974', 50), ('1-975', 50), ('1-976', 50), ('1-977', 50), ('1-978', 50),
('1-979', 50), ('1-980', 50
```



```
In [162]: plt.bar(words[0:10],word_count[0:10],width=0.9)
plt.xlabel('words', fontsize=10)
plt.ylabel('count', fontsize=10)
plt.xticks(words[0:10], fontsize=10, rotation=50)
plt.title("Most used words after applying stop words");
plt.show();
```



Continue: *This is the EDA part and we need to do data preprocessing and data cleaning and machine learning models a lot more to go so let's keep learning*