dataset_stats

November 9, 2019

0.1 Descriptive Statistics

```
In [13]: # import necessary libraries
         import numpy as np
         import json
         import re
         import matplotlib.pyplot as plt
         % matplotlib inline
```

```
UsageError: Line magic function `%` not found.
```

```
In [20]: # read the first 500,000 yelp reviews
         lines=open('yelp_dataset/review.json',encoding="utf8").readlines()[:500000]
```

```
In [4]: # An example of the format in which the review is available
        lines[0]
```

```
Out[4]: '{"review_id":"Q1sbwvVQXV2734tPgoKj4Q","user_id":"hG7b0MtEbXx5QzbzE6C_VA","business_id'
```

```
In [7]: # Create a list of the json dictionary containing reviews for all reviews in the datas
        review_list = [json.loads(line) for line in lines]
        # Create a list of the 'labels' for all reviews in the dataset
        labels_list = [review['stars']for review in review_list]
```

```
In [11]: # An example of the dictionary of review data
         review_list[0]
```

```
Out[11]: {'review_id': 'Q1sbwvVQXV2734tPgoKj4Q',
          'user_id': 'hG7b0MtEbXx5QzbzE6C_VA',
          'business_id': 'ujmEBvifdJM6h6RLv4wQIg',
          'stars': 1.0,
          'useful': 6,
          'funny': 1,
          'cool': 0,
          'text': 'Total bill for this horrible service? Over $8Gs. These crooks actually had t
          'date': '2013-05-07 04:34:36'}
```

0.1.1 Statistics Concerning Word Count of Documents

```
In [9]: #### Number of documents
        print("Number of documents in the dataset is : ", len(review_list))
        print("")
        #### Statistics describing number of words across documents
        wrd_num_list = [len(re.findall(r'[a-zA-Z]+',review_list[i]['text'])) for i in range(le
        min_wrd_num = min(wrd_num_list)
        pct_25_wrd_num = np.percentile(wrd_num_list,25)
        avg_wrd_num = np.mean(wrd_num_list)
        median_wrd_num = np.median(wrd_num_list)
        pct_75_wrd_num = np.percentile(wrd_num_list,75)
        max_wrd_num = max(wrd_num_list)
        print("Minimum number of words across documents in the dataset is : ", min_wrd_num )
        print("")
        print("25th percentile of number of words across documents in the dataset is : ", pct_
        print("")
        print("Average number of words across documents in the dataset is : ", avg_wrd_num )
        print("")
        print("Median number of words across documents in the dataset is : ", median_wrd_num)
        print("")
        print("75th percentile of number of words across documents in the dataset is : ", pct_
        print("")
        print("Maximum number of words across documents in the dataset is : ", max_wrd_num )
```

```
Number of documents in the dataset is :  500000

Minimum number of words across documents in the dataset is :  0

25th percentile of number of words across documents in the dataset is :  43.0

Average number of words across documents in the dataset is :  110.884286

Median number of words across documents in the dataset is :  79.0

75th percentile of number of words across documents in the dataset is :  141.0

Maximum number of words across documents in the dataset is :  1031
```

```
In [27]: wrd_len_list.count(0)

Out[27]: 52

In [29]: new_wrd_len_list = [len(re.findall(r'\w+',review_list[i]['text'])) for i in range(len

In [30]: new_wrd_len_list.count(0)

Out[30]: 10
```

0.1.2 Statistics Concerning value of label for documents

```
In [32]: #### Descriptive Statistics for labels
         #### Reviews are labeled using the 'stars' attribute
         num_labels = len(labels_list)
         avg_label = np.mean(labels_list)
         pct_25_label = np.percentile(labels_list,25)
         median_label = np.percentile(labels_list,50)
         pct_75_label = np.percentile(labels_list,75)
         unique_labels = set(labels_list)
         range_labels = len(unique_labels)
         min_label_value = min(unique_labels)
         max_label_value = max(unique_labels)
         print("Unique label values are : ", unique_labels)
         print("")
         print("Minimum value for labels is : ", min_label_value )
         print("")
         print("25th percentile of labels for documents in the dataset is : ", pct_25_label )
         print("")
         print("Average label value for documents in the dataset is : ", avg_label )
         print("")
         print("Median label value for documents in the dataset is : ", median_label)
         print("")
         print("75th percentile of labels for documents in the dataset is : ", pct_75_label)
         print("")
         print("Maximum value for labels is: ", max_label_value )
```

```
Unique label values are :  {1.0, 2.0, 3.0, 4.0, 5.0}

Minimum value for labels is :  1.0

25th percentile of labels for documents in the dataset is :  3.0

Average label value for documents in the dataset is :  3.729382

Median label value for documents in the dataset is :  4.0

75th percentile of labels for documents in the dataset is :  5.0

Maximum value for labels is:  5.0
```

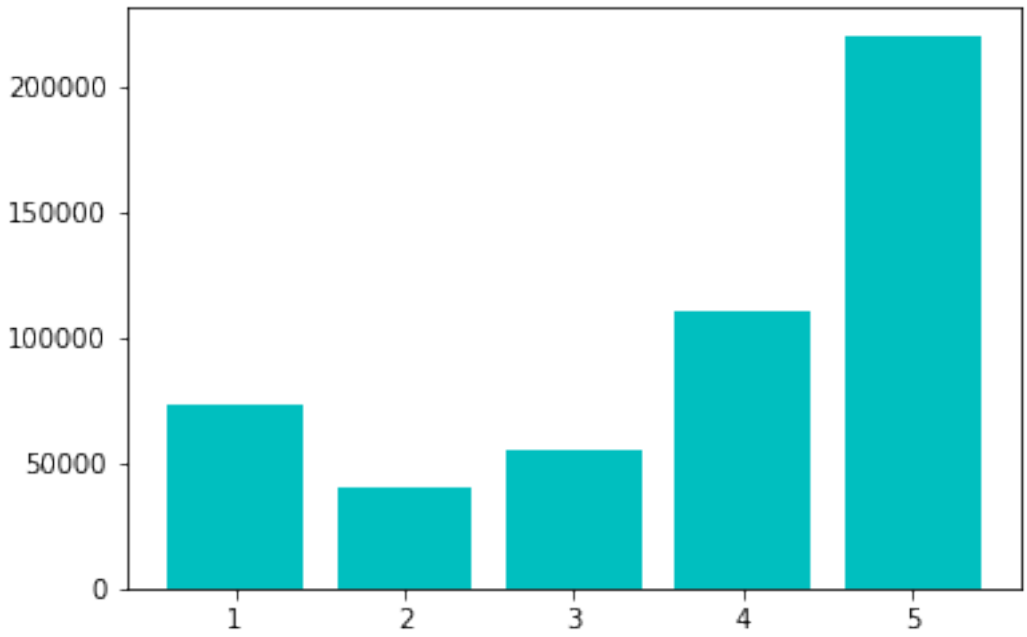
0.1.3 Label Distribution

```
In [8]: labels_distribution = {x:labels_list.count(x) for x in unique_labels}

In [9]: labels_distribution

Out[9]: {1.0: 72981, 2.0: 40636, 3.0: 55446, 4.0: 110585, 5.0: 220352}
```

In [18]:
```
plt.bar(list(labels_distribution.keys()), labels_distribution.values(), color='c')
plt.show()
```



0.1.4 Avergage word length of documents

In [34]:
```
#Calculate word lengths, word counts across documents
word_lengths = []
word_count = []
for i in range(len(review_list)):
    for word in (re.findall(r'\w+',review_list[i]['text'])):
        word_lengths.append(len(word))
        word_count.append(1)
# Calculate average
avg_word_len = sum(word_lengths)/sum(word_count)
print("Average word length across all documents is : ",avg_word_len)
```

```
Average word length across all documents is :  4.132913546329486
```

In [ ]: