ML1010

Independant Coding Project 2

Problem:
Examining numeric columns was becoming a frequent task and beginning to consume a great deal of time. In the Group Project it became frequent to determine the distribution of items in our dataset. For example, while examining the length of the review, or the number of tokens, the numbers were extremely skewed with very short reviews dominating the dataset, while long reviews were much less common. It was difficult to understand and visualize this information to make an informed decision about which data to include and exclude from different experiements

Solution:
Create a utility function to allow for seeing the data distribution in various capacities by zooming in on subranges of the data as well as changing the reporting detail (data grouping by numeric binning)

Configuration

```
#Parameters
PROJECT_NAME = 'ML1010_Weekly'
ENABLE_COLAB = True

#Root Machine Learning Directory. Projects appear underneath
GOOGLE_DRIVE_MOUNT = '/content/gdrive'
COLAB_ROOT_DIR = GOOGLE_DRIVE_MOUNT + '/MyDrive/Colab Notebooks'
COLAB_INIT_DIR = COLAB_ROOT_DIR + '/utility_files'

LOCAL_ROOT_DIR = '/home/magni/Documents/ML_Projects'
LOCAL_INIT_DIR = LOCAL_ROOT_DIR + '/utility_files'
```

Bootstrap Environment

```
#add in support for utility file directory and importing
import sys
import os

if ENABLE_COLAB:
  #Need access to drive
  from google.colab import drive
  drive.mount(GOOGLE_DRIVE_MOUNT, force_remount=True)

  #add in utility directory to syspath to import
  INIT_DIR = COLAB_INIT_DIR
  sys.path.append(os.path.abspath(INIT_DIR))

  #Config environment variables
  ROOT_DIR = COLAB_ROOT_DIR

else:
  #add in utility directory to syspath to import
  INIT_DIR = LOCAL_INIT_DIR
  sys.path.append(os.path.abspath(INIT_DIR))

  #Config environment variables
  ROOT_DIR = LOCAL_ROOT_DIR

#Import Utility Support
from jarvis import Jarvis
jarvis = Jarvis(ROOT_DIR, PROJECT_NAME)

import mv_python_utils as mvutils
```

```
Mounted at /content/gdrive
Wha...where am I?
I am awake now.

I have set your current working directory to /content/gdrive/MyDrive/Colab Notebooks/ML
The current time is 11:18
Hello sir. Extra caffeine may help.
```

Setup Runtime Environment

```
if ENABLE_COLAB:
  #!pip install scipy -q
  #!pip install scikit-learn -q
  #!pip install pycaret -q
  #!pip install matplotlib -q
  #!pip install joblib -q
  #!pip install pandasql -q

  display('Google Colab enabled')
else:
  display('Google Colab not enabled')
```

```
#Common imports
import json
import gzip
import pandas as pd
import numpy as np
import matplotlib
import re
import nltk
import matplotlib.pyplot as plt

pd.set_option('mode.chained_assignment', None)
nltk.download('stopwords')
%matplotlib inline
```

```
'Google Colab enabled'
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
```

Load Data

```
jarvis.showAllDataFiles()
```

```
[D] /content/gdrive/MyDrive/Colab Notebooks/data/Jarvis/04_test
---[   gz][  csv]--> pima-indians-diabetes.csv.gz (8.53 KB)
---[   gz][  csv]--> wk3_task_data.csv.gz (33.47 KB)

[D] /content/gdrive/MyDrive/Colab Notebooks/data/ML1010-Group-Project [Empty director

[D] /content/gdrive/MyDrive/Colab Notebooks/data/ML1010-Group-Project/01_original
---[   gz][ json]--> Cell_Phones_and_Accessories_5.json.gz (161.24 MB)
---[   gz][ json]--> meta_Cell_Phones_and_Accessories.json.gz (343.33 MB)


[D] /content/gdrive/MyDrive/Colab Notebooks/data/ML1010-Group-Project/02_working
[*][  pkl]---------> 01_Cellphone_small.pkl (45.46 MB)
---[   gz][  pkl]--> 01_NLP_ReviewText_Narrow_1.pkl.gz (6.88 MB)
---[   gz][  pkl]--> 01_NLP_ReviewText_Narrow_2.pkl.gz (170.55 MB)
---[   gz][  pkl]--> 01_NLP_ReviewText_Narrow_3.pkl.gz (295.59 MB)
[*][  pkl]---------> 01_NLP_ReviewText_small.pkl (28.94 MB)
[*][  pkl]---------> 01_NLP_Summary_small.pkl (3.82 MB)
[*][  pkl]---------> 01_NLP_Title_small.pkl (2.73 MB)
---[   gz][  pkl]--> 01_NL_ReviewText_All(new).pkl.gz (593.23 MB)
---[   gz][  pkl]--> 01_NL_ReviewText_All.pkl.gz (592.92 MB)
---[   gz][  pkl]--> 01_NL_ReviewText_textSplit.pkl.gz (15.78 MB)
[*][  pkl]---------> 02_Cellphone.pkl (46.32 MB)
[*][  pkl]---------> 02_NLP_ReviewTextData.pkl (87.00 MB)
[*][  pkl]---------> 02_NLP_SummaryData.pkl (8.32 MB)
[*][  pkl]---------> 02_NLP_TitleData.pkl (16.71 MB)
[*][  pkl]---------> 03_Cellphone.pkl (46.31 MB)
[*][  pkl]---------> 03_NLP_ReviewTextData.pkl (28.94 MB)
[*][  pkl]---------> 03_NLP_ReviewText_Narrow.pkl (17.13 MB)
[*][  pkl]---------> 03_NLP_SummaryData.pkl (3.82 MB)
[*][  pkl]---------> 03_NLP_TitleData.pkl (2.73 MB)
```

```
[*][  pkl]---------> 03_NLP_TitleData.pkl (2.73 MB)
[*][  pkl]---------> 04_NLP_ReviewText_Narrow.pkl (16.95 MB)
[*][  pkl]---------> 05_NLP_ReviewText_Narrow.pkl (66.15 MB)
[*][  pkl]---------> 05_NLP_ReviewText_Narrow_full.pkl (207.91 MB)

[D] /content/gdrive/MyDrive/Colab Notebooks/data/ML1010-Group-Project/03_train [Empty

[D] /content/gdrive/MyDrive/Colab Notebooks/data/ML1010-Group-Project/04_test [Empty 

[D] /content/gdrive/MyDrive/Colab Notebooks/data/ML1010_Weekly
---[   gz][  csv]--> complaints.csv.gz (370.67 MB)
[*][  csv]---------> movie_reviews_cleaned.csv (38.37 MB)
[*][  csv]---------> pima-indians-diabetes.csv (22.73 KB)
---[   gz][  tsv]--> rspct.tsv.gz (347.13 MB)
---[   gz][  csv]--> subreddit_info.csv.gz (37.80 KB)
[*][  csv]---------> wk3_task_data.csv (81.31 KB)

[D] /content/gdrive/MyDrive/Colab Notebooks/data/ML1010_Weekly/01_original [Empty dir

[D] /content/gdrive/MyDrive/Colab Notebooks/data/ML1010_Weekly/02_working [Empty dire

[D] /content/gdrive/MyDrive/Colab Notebooks/data/ML1010_Weekly/03_train [Empty direct

[D] /content/gdrive/MyDrive/Colab Notebooks/data/ML1010_Weekly/04_test [Empty directo

[D] /content/gdrive/MyDrive/Colab Notebooks/data/test_compress
[*][  pkl]---------> 02_NLP_SummaryData.pkl (8.32 MB)
[*][  pkl]---------> 02_NLP_TitleData.pkl (16.71 MB)
```

```
df = pd.read_pickle('/content/gdrive/MyDrive/Colab Notebooks/data/ML1010-Group-Project/02_wor
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 63413 entries, 0 to 63412
Data columns (total 49 columns):
 #   Column                              Non-Null Count  Dtype  
---  ------                              --------------  -----  
 0   uuid                                63413 non-null  object 
 1   reviewText                          63413 non-null  object 
 2   overall                             63413 non-null  float64
 3   reviewText_lemma                    63413 non-null  object 
 4   reviewText_nouns                    63413 non-null  object 
 5   reviewText_adjectives               63413 non-null  object 
 6   reviewText_verbs                    63413 non-null  object 
 7   reviewText_nav                      63413 non-null  object 
 8   reviewText_lemma_tb_pol             63310 non-null  float64
 9   reviewText_lemma_tb_subj            63310 non-null  float64
 10  reviewText_lemma_tb_tokens          63310 non-null  float64
 11  reviewText_lemma_tb_length          63310 non-null  float64
 12  reviewText_lemma_bert               63413 non-null  object 
 13  reviewText_lemma_flairSent          63310 non-null  float64
 14  reviewText_adjectives_tb_pol        50732 non-null  float64
 15  reviewText_adjectives_tb_subj       50732 non-null  float64
 16  reviewText_adjectives_tb_tokens         50732 non-null  float64
 17  reviewText_adjectives_tb_length         50732 non-null  float64
 18  reviewText_adjectives_bert              63413 non-null  object
 19  reviewText_adjectives_flairSent         50732 non-null  float64
 20  reviewText_verbs_tb_pol                 43234 non-null  float64
 21  reviewText_verbs_tb_subj                43234 non-null  float64
 22  reviewText_verbs_tb_tokens              43234 non-null  float64
 23  reviewText_verbs_tb_length              43234 non-null  float64
 24  reviewText_verbs_bert                   63413 non-null  object
 25  reviewText_verbs_flairSent              43234 non-null  float64
 26  reviewText_nav_tb_pol                   62332 non-null  float64
 27  reviewText_nav_tb_subj                  62332 non-null  float64
 28  reviewText_nav_tb_tokens                62332 non-null  float64
 29  reviewText_nav_tb_length                62332 non-null  float64
 30  reviewText_nav_bert                     63413 non-null  object
 31  reviewText_nav_flairSent                62332 non-null  float64
 32  overall_posneg                          63413 non-null  int64
 33  reviewText_lemma_flairSent_norm         63310 non-null  float64
 34  reviewText_lemma_flairSent_posneg       63310 non-null  float64
 35  reviewText_adjectives_flairSent_norm    50732 non-null  float64
 36  reviewText_adjectives_flairSent_posneg  50732 non-null  float64
 37  reviewText_verbs_flairSent_norm         43234 non-null  float64
 38  reviewText_verbs_flairSent_posneg       43234 non-null  float64
 39  reviewText_nav_flairSent_norm           62332 non-null  float64
 40  reviewText_nav_flairSent_posneg         62332 non-null  float64
 41  reviewText_lemma_tb_pol_norm            63310 non-null  float64
 42  reviewText_lemma_tb_pol_posneg          63310 non-null  float64
 43  reviewText_adjectives_tb_pol_norm       50732 non-null  float64
 44  reviewText_adjectives_tb_pol_posneg     50732 non-null  float64
 45  reviewText_verbs_tb_pol_norm            43234 non-null  float64
 46  reviewText_verbs_tb_pol_posneg          43234 non-null  float64
 47  reviewText_nav_tb_pol_norm              62332 non-null  float64
 48  reviewText_nav_tb_pol_posneg            62332 non-null  float64
dtypes: float64(37), int64(1), object(11)
memory usage: 23.7+ MB
```

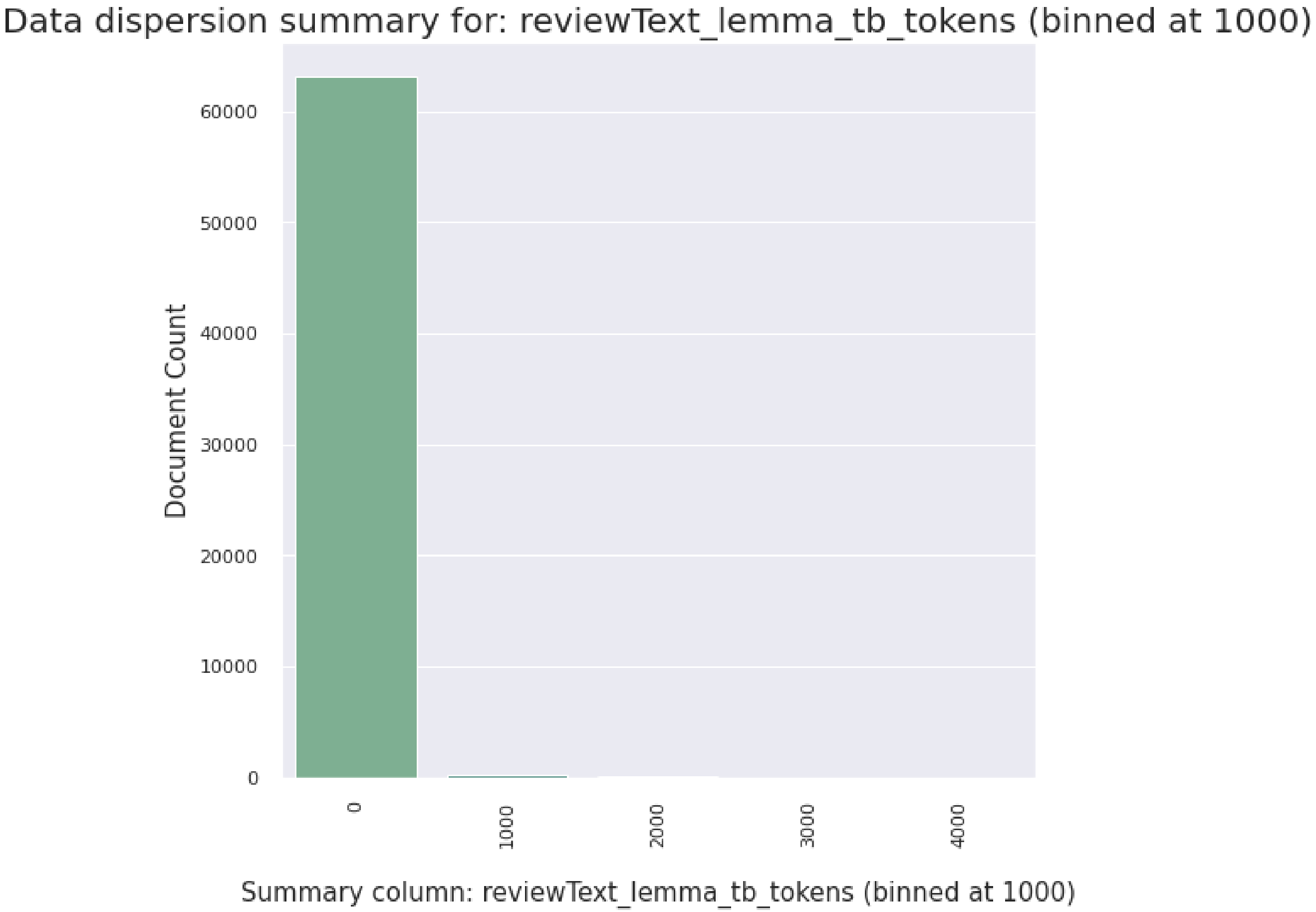
Independant Code Exploration

```
mvutils.showColumnSummary(df, 'reviewText_lemma_tb_tokens')
```

```
Dataframe shape (63413, 49)
Analysis column: reviewText_lemma_tb_tokens
Distinct values (incl. null): 1014
Number of na   values: 103
Number of null values: 103
Total documents in corpus: 63413
```

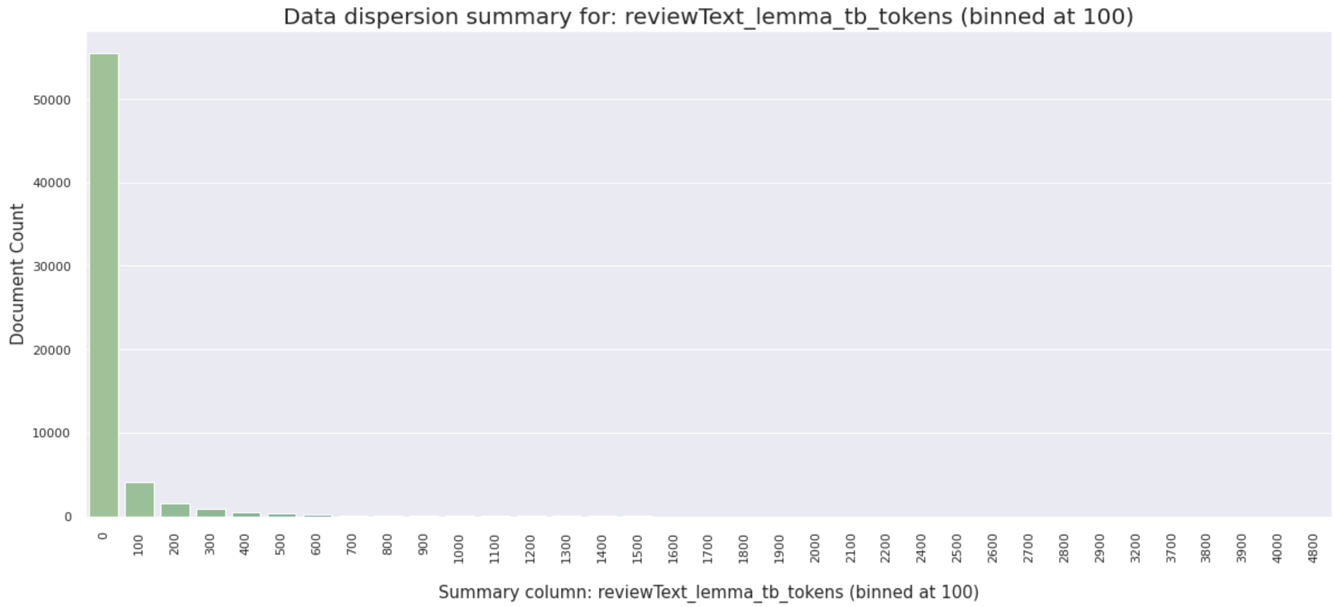
```
#Examine numeric column for distribution
mvutils.examineColumnNumeric(df,
                             'reviewText_lemma_tb_tokens'
                             )
```

```
Warning: 103 null values detected in column. Removing for analysis
```



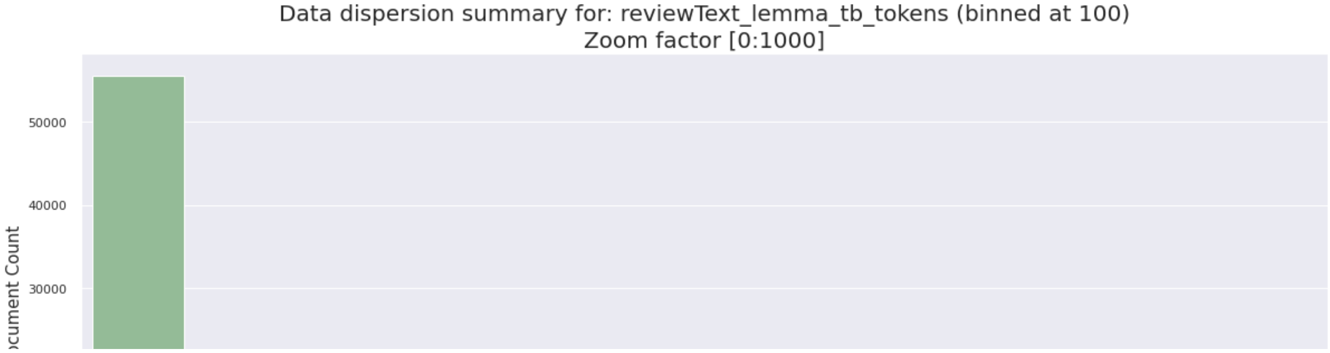
```
#Increase plotsize for better viewing,
#change binning size to 100 from default 1000
mvutils.examineColumnNumeric(df,
                             'reviewText_lemma_tb_tokens',
                             binsize=100,
                             plotsize=5
                             )
```

```
Warning: 103 null values detected in column. Removing for analysis
```



```
#Enable zoom to examine range 0-1000 binned at 100 for better viewing
mvutils.examineColumnNumeric(df,
                             'reviewText_lemma_tb_tokens',
                             binsize=100,
                             zoom=True,
                             minZoomLevel=0,
                             maxZoomLevel=1000,
                             plotsize=5)
```

```
Warning: 103 null values detected in column. Removing for analysis
```



```
#Data still dominated by 0-100
#Zoom to 0-200 with bin size of 10
mvutils.examineColumnNumeric(df,
                             'reviewText_lemma_tb_tokens',
                             binsize=10,
                             zoom=True,
                             minZoomLevel=0,
                             maxZoomLevel=200,
                             plotsize=5)
```

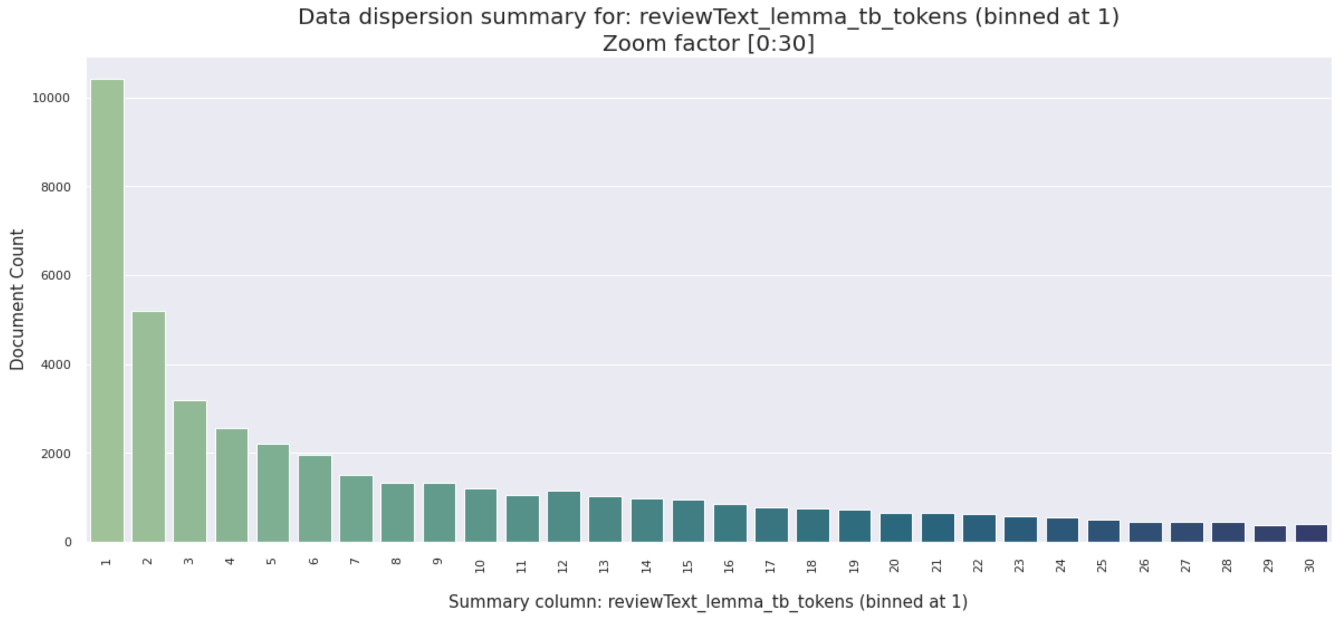
```
     Warning: 103 null values detected in column. Removing for analysis
```

```
import importlib
importlib.reload(mvutils)
```

```
     <module 'mv_python_utils' from '/content/gdrive/MyDrive/Colab Notebooks/utility_files/m
```

```
#Still not quite enough detail to determine where to cutoff
#Zoom with binsize 1, range 0-30
mvutils.examineColumnNumeric(df,
                             'reviewText_lemma_tb_tokens',
                             binsize=1,
                             zoom=True,
                             minZoomLevel=0,
                             maxZoomLevel=30,
                             plotsize=5,
                             verbose=True,
                             numRecords=5)
```

```
Warning: 103 null values detected in column. Removing for analysis
```

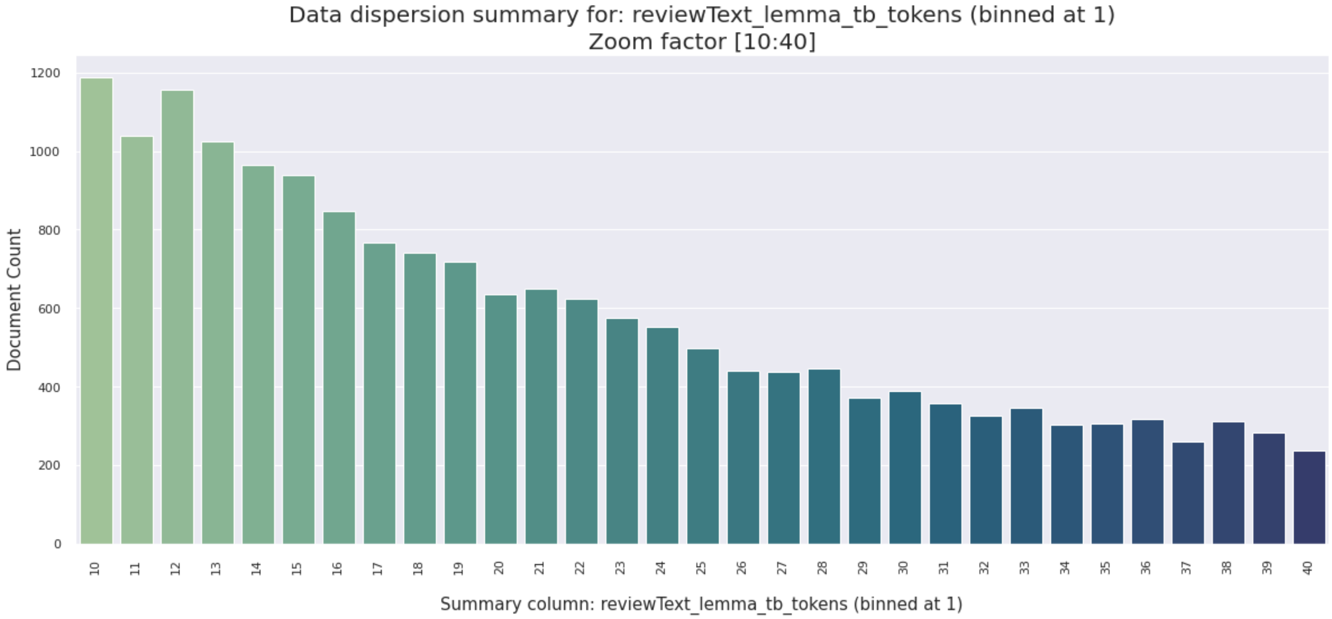


```
dataframe shape: (30, 2)

dataframe info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30 entries, 0 to 29
Data columns (total 2 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   bin_at_1     30 non-null     int64
 1   binnedCount  30 non-null     int64
dtypes: int64(2)
memory usage: 608.0 bytes
None

Top 5 in dataframe
```

	bin_at_1	binnedCount
0	30	388

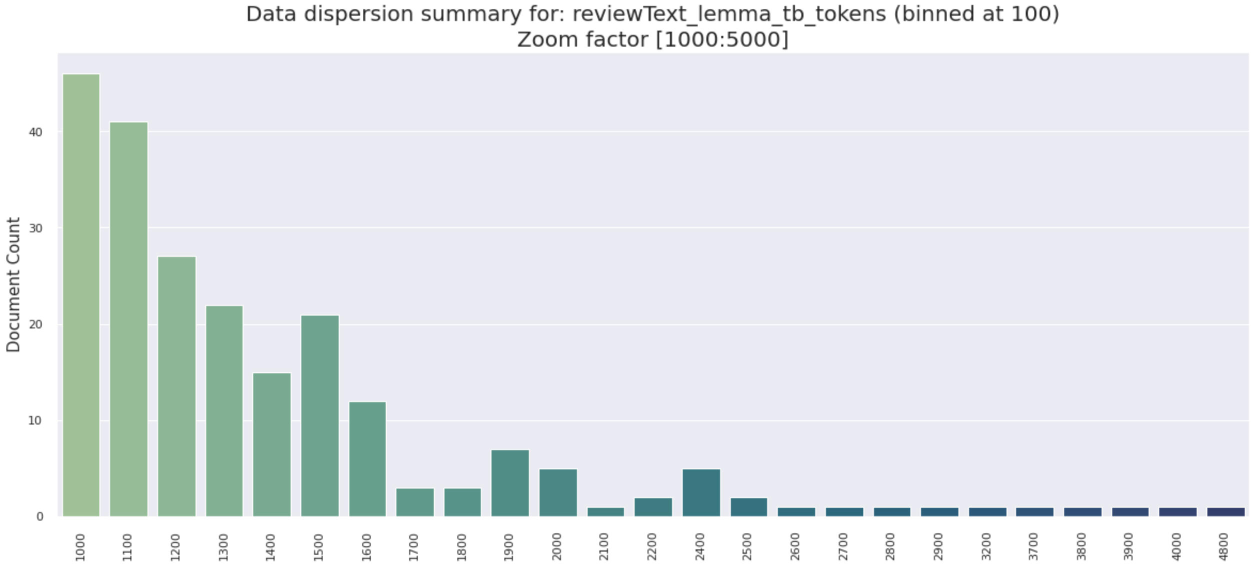
```
#Need a better view and scale for mid range numbers
#Zoom 10:40 with binsize=1
mvutils.examineColumnNumeric(df,
                             'reviewText_lemma_tb_tokens',
                             binsize=1,
                             zoom=True,
                             minZoomLevel=10,
                             maxZoomLevel=40,
                             plotsize=5)
```

```
Warning: 103 null values detected in column. Removing for analysis
```



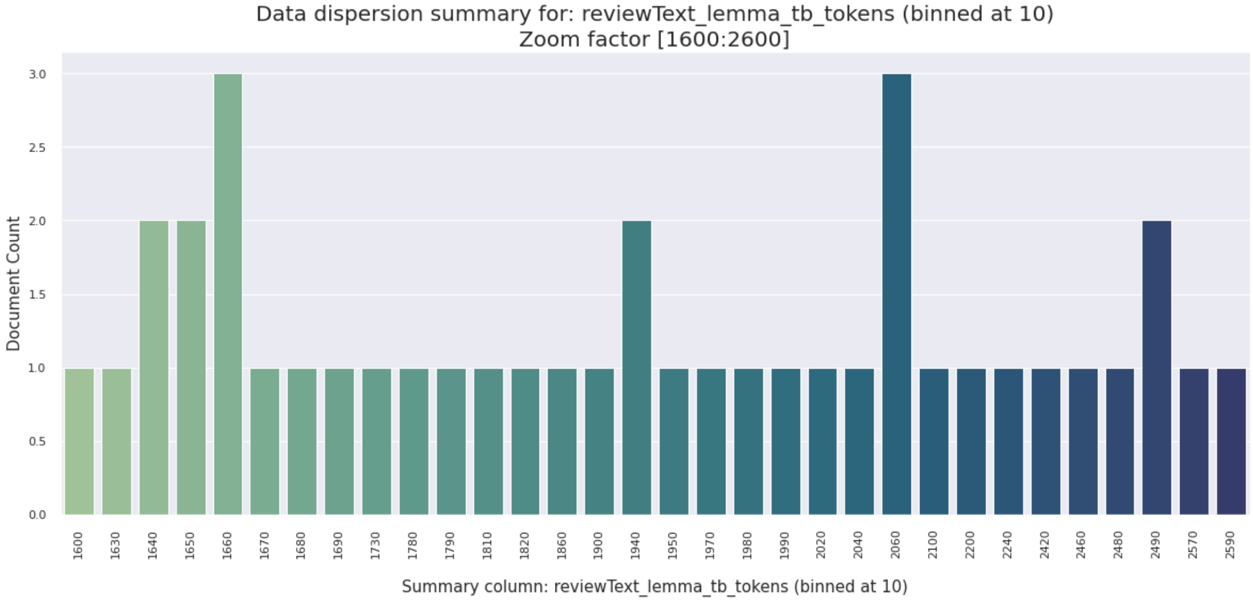
```
#Examine tail end of data (larger # of tokens)
#Zoom 1000:5000 binsize 100
mvutils.examineColumnNumeric(df,
                             'reviewText_lemma_tb_tokens',
                             binsize=100,
                             zoom=True,
                             minZoomLevel=1000,
                             maxZoomLevel=5000,
                             plotsize=5)
```

```
Warning: 103 null values detected in column. Removing for analysis
```



```
#Examine dropoff near the 1700 range
#Zoom 1600:2600, binsize 10
mvutils.examineColumnNumeric(df,
                             'reviewText_lemma_tb_tokens',
                             binsize=10,
                             zoom=True,
                             minZoomLevel=1600,
                             maxZoomLevel=2600,
                             plotsize=5)
```

```
Warning: 103 null values detected in column. Removing for analysis
```



✓ 1s completed at 11:48 AM