Preprocessing of the Data

```
%matplotlib inline

import warnings
warnings.filterwarnings("ignore")

import pandas as pd
import numpy as np
import pickle
import re
import os
import nltk
# Tutorial about Python regular expressions: https://pymotw.com/2/re/

from nltk.corpus import stopwords
from tqdm import tqdm
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.feature_extraction.text import CountVectorizer
from nltk.sentiment.vader import SentimentIntensityAnalyzer
nltk.download('vader_lexicon')
from tqdm import tqdm


[nltk_data] Downloading package vader_lexicon to /root/nltk_data...
```

1. Reading Data

```
project_data = pd.read_csv('train_data.csv')
resource_data = pd.read_csv('resources.csv')

print("Number of data points in train data", project_data.shape)
print('-'*50)
print("The attributes of data :", project_data.columns.values)

Number of data points in train data (109248, 17)
--------------------------------------------------
The attributes of data : ['Unnamed: 0' 'id' 'teacher_id'
'teacher_prefix' 'school_state'
 'project_submitted_datetime' 'project_grade_category'
 'project_subject_categories' 'project_subject_subcategories'
 'project_title' 'project_essay_1' 'project_essay_2' 'project_essay_3'
 'project_essay_4' 'project_resource_summary'
 'teacher_number_of_previously_posted_projects' 'project_is_approved']

print("Number of data points in train data", resource_data.shape)
print(resource_data.columns.values)
resource_data.head(2)

Number of data points in train data (1541272, 4)
['id' 'description' 'quantity' 'price']
```

```
        id                                           description
quantity  \
0  p233245  LC652 - Lakeshore Double-Space Mobile Drying Rack
1
1  p069063         Bouncy Bands for Desks (Blue support pipes)
3

    price
0  149.00
1   14.95
```

2. Preprocessing Categorical Features: project_grade_category

```
project_data['project_grade_category'].value_counts() # we have only 4
different types of grade categories available
```

```
Grades PreK-2    44225
Grades 3-5       37137
Grades 6-8       16923
Grades 9-12      10963
Name: project_grade_category, dtype: int64
```

we need to remove the spaces, replace the '-' with '_' and convert all the letters to small

```
# https://stackoverflow.com/questions/36383821/pandas-dataframe-apply-
function-to-column-strings-based-on-other-column-value
# replacing space with  '_'
project_data['project_grade_category'] =
project_data['project_grade_category'].str.replace(' ','_')
project_data['project_grade_category'] =
project_data['project_grade_category'].str.replace('-','_')
project_data['project_grade_category'] =
project_data['project_grade_category'].str.lower()
project_data['project_grade_category'].value_counts()
```

```
grades_prek_2    44225
grades_3_5       37137
grades_6_8       16923
grades_9_12      10963
Name: project_grade_category, dtype: int64
```

3. Preprocessing Categorical Features: project_subject_categories

```
print("Number of categories for each type:")
print("="*40)
project_data['project_subject_categories'].value_counts()
```

```
Number of categories for each type:
========================================
```

Literacy & Language	23655
Math & Science	17072
Literacy & Language, Math & Science	14636
Health & Sports	10177
Music & The Arts	5180
Special Needs	4226
Literacy & Language, Special Needs	3961
Applied Learning	3771
Math & Science, Literacy & Language	2289
Applied Learning, Literacy & Language	2191
History & Civics	1851
Math & Science, Special Needs	1840
Literacy & Language, Music & The Arts	1757
Math & Science, Music & The Arts	1642
Applied Learning, Special Needs	1467
History & Civics, Literacy & Language	1421
Health & Sports, Special Needs	1391
Warmth, Care & Hunger	1309
Math & Science, Applied Learning	1220
Applied Learning, Math & Science	1052
Literacy & Language, History & Civics	809
Health & Sports, Literacy & Language	803
Applied Learning, Music & The Arts	758
Math & Science, History & Civics	652
Literacy & Language, Applied Learning	636
Applied Learning, Health & Sports	608
Math & Science, Health & Sports	414
History & Civics, Math & Science	322
History & Civics, Music & The Arts	312
Special Needs, Music & The Arts	302
Health & Sports, Math & Science	271
History & Civics, Special Needs	252
Health & Sports, Applied Learning	192
Applied Learning, History & Civics	178
Health & Sports, Music & The Arts	155
Music & The Arts, Special Needs	138
Literacy & Language, Health & Sports	72
Health & Sports, History & Civics	43
History & Civics, Applied Learning	42
Special Needs, Health & Sports	42
Special Needs, Warmth, Care & Hunger	23
Health & Sports, Warmth, Care & Hunger	23
Music & The Arts, Health & Sports	19
Music & The Arts, History & Civics	18
History & Civics, Health & Sports	13
Math & Science, Warmth, Care & Hunger	11
Music & The Arts, Applied Learning	10
Applied Learning, Warmth, Care & Hunger	10
Literacy & Language, Warmth, Care & Hunger	9
Music & The Arts, Warmth, Care & Hunger	2

```
History & Civics, Warmth, Care & Hunger           1
Name: project_subject_categories, dtype: int64
```

remove spaces, 'the' replace '&' with '_', and ',' with '_'

```
project_data['project_subject_categories'] =
project_data['project_subject_categories'].str.replace(' The ','')
project_data['project_subject_categories'] =
project_data['project_subject_categories'].str.replace(' ','')
project_data['project_subject_categories'] =
project_data['project_subject_categories'].str.replace('&','_')
project_data['project_subject_categories'] =
project_data['project_subject_categories'].str.replace(',','_')
project_data['project_subject_categories'] =
project_data['project_subject_categories'].str.lower()
project_data['project_subject_categories'].value_counts()
```

```
literacy_language                          23655
math_science                               17072
literacy_language_math_science             14636
health_sports                              10177
music_arts                                  5180
specialneeds                                4226
literacy_language_specialneeds              3961
appliedlearning                             3771
math_science_literacy_language              2289
appliedlearning_literacy_language           2191
history_civics                              1851
math_science_specialneeds                   1840
literacy_language_music_arts                1757
math_science_music_arts                     1642
appliedlearning_specialneeds                1467
history_civics_literacy_language            1421
health_sports_specialneeds                  1391
warmth_care_hunger                          1309
math_science_appliedlearning                1220
appliedlearning_math_science                1052
literacy_language_history_civics             809
health_sports_literacy_language              803
appliedlearning_music_arts                   758
math_science_history_civics                  652
literacy_language_appliedlearning            636
appliedlearning_health_sports                608
math_science_health_sports                   414
history_civics_math_science                  322
history_civics_music_arts                    312
specialneeds_music_arts                      302
health_sports_math_science                   271
history_civics_specialneeds                  252
health_sports_appliedlearning                192
appliedlearning_history_civics               178
```

```
health_sports_music_arts                   155
music_arts_specialneeds                    138
literacy_language_health_sports             72
health_sports_history_civics                43
history_civics_appliedlearning              42
specialneeds_health_sports                  42
specialneeds_warmth_care_hunger             23
health_sports_warmth_care_hunger            23
music_arts_health_sports                    19
music_arts_history_civics                   18
history_civics_health_sports                13
math_science_warmth_care_hunger             11
music_arts_appliedlearning                  10
appliedlearning_warmth_care_hunger          10
literacy_language_warmth_care_hunger         9
music_arts_warmth_care_hunger                2
history_civics_warmth_care_hunger            1
Name: project_subject_categories, dtype: int64
```

4. Preprocessing Categorical Features: teacher_prefix

```
project_data['teacher_prefix'].value_counts()
```

```
Mrs.       57269
Ms.        38955
Mr.        10648
Teacher     2360
Dr.           13
Name: teacher_prefix, dtype: int64
```

```
# check if we have any nan values are there
print(project_data['teacher_prefix'].isnull().values.any())
print("number of nan
values",project_data['teacher_prefix'].isnull().values.sum())
```

```
True
number of nan values 3
```

numebr of missing values are very less in number, we can replace it with Mrs. as most of the projects are submitted by Mrs.

```
project_data['teacher_prefix']=project_data['teacher_prefix'].fillna('
Mrs.')
```

```
project_data['teacher_prefix'].value_counts()
```

```
Mrs.       57272
Ms.        38955
Mr.        10648
Teacher     2360
```

```
Dr.           13
Name: teacher_prefix, dtype: int64
```

Remove '.' convert all the chars to small

```
project_data['teacher_prefix'] =
project_data['teacher_prefix'].str.replace('.','')
project_data['teacher_prefix'] =
project_data['teacher_prefix'].str.lower()
project_data['teacher_prefix'].value_counts()
```

```
mrs        57272
ms         38955
mr         10648
teacher     2360
dr            13
Name: teacher_prefix, dtype: int64
```

5. Preprocessing Categorical Features: project_subject_subcategories

```
project_data['project_subject_subcategories'].value_counts()
```

```
Literacy                                 9486
Literacy, Mathematics                    8325
Literature & Writing, Mathematics        5923
Literacy, Literature & Writing           5571
Mathematics                              5379
                                         ...
Community Service, Gym & Fitness            1
Parent Involvement, Team Sports             1
Gym & Fitness, Social Sciences              1
Community Service, Music                    1
Economics, Foreign Languages                1
Name: project_subject_subcategories, Length: 401, dtype: int64
```

same process we did in project_subject_categories

```
project_data['project_subject_subcategories'] =
project_data['project_subject_subcategories'].str.replace(' The ','')
project_data['project_subject_subcategories'] =
project_data['project_subject_subcategories'].str.replace(' ','')
project_data['project_subject_subcategories'] =
project_data['project_subject_subcategories'].str.replace('&','_')
project_data['project_subject_subcategories'] =
project_data['project_subject_subcategories'].str.replace(',','_')
project_data['project_subject_subcategories'] =
project_data['project_subject_subcategories'].str.lower()
project_data['project_subject_subcategories'].value_counts()
```

```
literacy                          9486
literacy_mathematics              8325
literature_writing_mathematics    5923
```

```
literacy_literature_writing       5571
mathematics                       5379
                                  ...
communityservice_gym_fitness         1
parentinvolvement_teamsports         1
gym_fitness_socialsciences           1
communityservice_music               1
economics_foreignlanguages           1
Name: project_subject_subcategories, Length: 401, dtype: int64
```

6. Preprocessing Categorical Features: school_state

```
print("Number of projects submitted from each state:")
print("="*45)
project_data['school_state'].value_counts()
```

```
Number of projects submitted from each state:
=============================================

CA    15388
TX     7396
NY     7318
FL     6185
NC     5091
IL     4350
GA     3963
SC     3936
MI     3161
PA     3109
IN     2620
MO     2576
OH     2467
LA     2394
MA     2389
WA     2334
OK     2276
NJ     2237
AZ     2147
VA     2045
WI     1827
AL     1762
UT     1731
TN     1688
CT     1663
MD     1514
NV     1367
MS     1323
KY     1304
OR     1242
MN     1208
```

```
CO     1111
AR     1049
ID      693
IA      666
KS      634
NM      557
DC      516
HI      507
ME      505
WV      503
NH      348
AK      345
DE      343
NE      309
SD      300
RI      285
MT      245
ND      143
WY       98
VT       80
Name: school_state, dtype: int64
```

convert all of them into small letters

```
# converting all the state names into lower case
project_data['school_state'] =
project_data['school_state'].str.lower()
project_data['school_state'].value_counts()
```

```
ca    15388
tx     7396
ny     7318
fl     6185
nc     5091
il     4350
ga     3963
sc     3936
mi     3161
pa     3109
in     2620
mo     2576
oh     2467
la     2394
ma     2389
wa     2334
ok     2276
nj     2237
az     2147
va     2045
wi     1827
al     1762
```

```
ut     1731
tn     1688
ct     1663
md     1514
nv     1367
ms     1323
ky     1304
or     1242
mn     1208
co     1111
ar     1049
id      693
ia      666
ks      634
nm      557
dc      516
hi      507
me      505
wv      503
nh      348
ak      345
de      343
ne      309
sd      300
ri      285
mt      245
nd      143
wy       98
vt       80
Name: school_state, dtype: int64
```

7. Preprocessing Categorical Features: project_title

```
# 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

# 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've",\
            "you'll", "you'd", 'your', 'yours', 'yourself',
'yourselves', 'he', 'him', 'his', 'himself', \
            'she', "she's", 'her', 'hers', 'herself', 'it', "it's",
'its', 'itself', 'they', 'them', 'their',\
            'theirs', 'themselves', 'what', 'which', 'who', 'whom',
'this', 'that', "that'll", 'these', 'those', \
            'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being',
'have', 'has', 'had', 'having', 'do', 'does', \
            'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if',
'or', 'because', 'as', 'until', 'while', 'of', \
            'at', 'by', 'for', 'with', 'about', 'against', 'between',
'into', 'through', 'during', 'before', 'after',\
            'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out',
'on', 'off', 'over', 'under', 'again', 'further',\
            'then', 'once', 'here', 'there', 'when', 'where', 'why',
'how', 'all', 'any', 'both', 'each', 'few', 'more',\
            'most', 'other', 'some', 'such', 'only', 'own', 'same',
'so', 'than', 'too', 'very', \
            's', 't', 'can', 'will', 'just', 'don', "don't", 'should',
"should've", 'now', 'd', 'll', 'm', 'o', 're', \
            've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't",
'didn', "didn't", 'doesn', "doesn't", 'hadn',\
            "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn',
"isn't", 'ma', 'mightn', "mightn't", 'mustn',\
            "mustn't", 'needn', "needn't", 'shan', "shan't",
'shouldn', "shouldn't", 'wasn', "wasn't", 'weren', "weren't", \
            'won', "won't", 'wouldn', "wouldn't"]

project_data['project_title'].head(5)

0     Educational Support for English Learners at Home
1                Wanted: Projector for Hungry Learners
2    Soccer Equipment for AWESOME Middle School Stu...
3                                Techie Kindergarteners
4                                Interactive Math Tools
Name: project_title, dtype: object

print("printing some random reviews")
print(9, project_data['project_title'].values[9])
print(34, project_data['project_title'].values[34])
print(147, project_data['project_title'].values[147])
```

```
printing some random reviews
9 Just For the Love of Reading--\r\nPure Pleasure
34 \"Have A Ball!!!\"
147 Who needs a Chromebook?\r\nWE DO!!
```

```
# Combining all the above statements
from tqdm import tqdm
def preprocess_text(text_data):
    preprocessed_text = []
    # tqdm is for printing the status bar
    for sentance in tqdm(text_data):
        sent = decontracted(sentance)
        sent = sent.replace('\\r', ' ')
        sent = sent.replace('\\n', ' ')
        sent = sent.replace('\\"', ' ')
        sent = re.sub('[^A-Za-z0-9]+', ' ', sent)
        # https://gist.github.com/sebleier/554280
        sent = ' '.join(e for e in sent.split() if e.lower() not in
stopwords)
        preprocessed_text.append(sent.lower().strip())
    return preprocessed_text

project_data['project_title'] =
preprocess_text(project_data['project_title'].values)
```

```
100%|██████████| 109248/109248 [00:03<00:00, 31681.43it/s]
```

8. Preprocessing Categorical Features: essay

```
# merge two column text dataframe:
project_data["essay"] = project_data["project_essay_1"].map(str) +\
                        project_data["project_essay_2"].map(str) + \
                        project_data["project_essay_3"].map(str) + \
                        project_data["project_essay_4"].map(str)

print("printing some random essays")
print(9, project_data['essay'].values[9])
print('-'*50)
print(34, project_data['essay'].values[34])
print('-'*50)
print(147, project_data['essay'].values[147])
```

```
printing some random essay
9 Over 95% of my students are on free or reduced lunch.  I have a few
who are homeless, but despite that, they come to school with an
eagerness to learn.  My students are inquisitive eager learners who
embrace the challenge of not having great books and other resources
every day.  Many of them are not afforded the opportunity to engage
with these big colorful pages of a book on a regular basis at home and
they don't travel to the public library.  \r\nIt is my duty as a
teacher to do all I can to provide each student an opportunity to
```

succeed in every aspect of life. \r\nReading is Fundamental! My students will read these books over and over again while boosting their comprehension skills. These books will be used for read alouds, partner reading and for Independent reading. \r\nThey will engage in reading to build their \"Love for Reading\" by reading for pure enjoyment. They will be introduced to some new authors as well as some old favorites. I want my students to be ready for the 21st Century and know the pleasure of holding a good hard back book in hand. There's nothing like a good book to read! \r\nMy students will soar in Reading, and more because of your consideration and generous funding contribution. This will help build stamina and prepare for 3rd grade. Thank you so much for reading our proposal!nannan
---
34 My students mainly come from extremely low-income families, and the majority of them come from homes where both parents work full time. Most of my students are at school from 7:30 am to 6:00 pm (2:30 to 6:00 pm in the after-school program), and they all receive free and reduced meals for breakfast and lunch. \r\n\r\n\r\nI want my students to feel as comfortable in my classroom as they do at home. Many of my students take on multiple roles both at home as well as in school. They are sometimes the caretakers of younger siblings, cooks, babysitters, academics, friends, and most of all, they are developing who they are going to become as adults. I consider it an essential part of my job to model helping others gain knowledge in a positive manner. As a result, I have a community of students who love helping each other in and outside of the classroom. They consistently look for opportunities to support each other's learning in a kind and helpful way.I am excited to be experimenting with alternative seating in my classroom this school year. Studies have shown that giving students the option of where they sit in a classroom increases focus as well as motivation. \r\n\r\nBy allowing students choice in the classroom, they are able to explore and create in a welcoming environment. Alternative classroom seating has been experimented with more frequently in recent years. I believe (along with many others), that every child learns differently. This does not only apply to how multiplication is memorized, or a paper is written, but applies to the space in which they are asked to work. I have had students in the past ask \"Can I work in the library? Can I work on the carpet?\" My answer was always, \"As long as you're learning, you can work wherever you want!\" \r\n\r\nWith the yoga balls and the lap-desks, I will be able to increase the options for seating in my classroom and expand its imaginable space.nannan
---
147 My students are eager to learn and make their mark on the world.\r\n\r\nThey come from a Title 1 school and need extra love.\r\n\r\nMy fourth grade students are in a high poverty area and still come to school every day to get their education. I am trying to make it fun and educational for them so they can get the most out of their schooling. I created a caring environment for the students to bloom! They deserve the best.\r\nThank you!\r\nI am requesting 1 Chromebook

to access online interventions, differentiate instruction, and get extra practice. The Chromebook will be used to supplement ELA and math instruction. Students will play ELA and math games that are engaging and fun, as well as participate in assignments online. This in turn will help my students improve their skills. Having a Chromebook in the classroom would not only allow students to use the programs at their own pace, but would ensure more students are getting adequate time to use the programs. The online programs have been especially beneficial to my students with special needs. They are able to work at their level as well as be challenged with some different materials. This is making these students more confident in their abilities.\r\n\r\nThe Chromebook would allow my students to have daily access to computers and increase their computing skills.\r\nThis will change their lives for the better as they become more successful in school. Having access to technology in the classroom would help bridge the achievement gap.nannan

```
project_data['essay'] = preprocess_text(project_data['essay'].values)
```

```
100%|██████████| 109248/109248 [01:27<00:00, 1242.28it/s]
```

```
print("printing some random essays after pre-processing")
print(9, project_data['essay'].values[9])
print('-'*50)
print(34, project_data['essay'].values[34])
print('-'*50)
print(147, project_data['essay'].values[147])
```

```
printing some random essays after pre-processing
9 95 students free reduced lunch homeless despite come school eagerness learn students inquisitive eager learners embrace challenge not great books resources every day many not afforded opportunity engage big colorful pages book regular basis home not travel public library duty teacher provide student opportunity succeed every aspect life reading fundamental students read books boosting comprehension skills books used read alouds partner reading independent reading engage reading build love reading reading pure enjoyment introduced new authors well old favorites want students ready 21st century know pleasure holding good hard back book hand nothing like good book read students soar reading consideration generous funding contribution help build stamina prepare 3rd grade thank much reading proposal nannan
--------------------------------------------------
34 students mainly come extremely low income families majority come homes parents work full time students school 7 30 6 00 pm 2 30 6 00 pm school program receive free reduced meals breakfast lunch want students feel comfortable classroom home many students take multiple roles home well school sometimes caretakers younger siblings cooks babysitters academics friends developing going become adults consider essential part job model helping others gain knowledge positive manner result community students love helping outside classroom consistently look opportunities support learning kind helpful way excited
```

```
experimenting alternative seating classroom school year studies shown
giving students option sit classroom increases focus well motivation
allowing students choice classroom able explore create welcoming
environment alternative classroom seating experimented frequently
recent years believe along many others every child learns differently
not apply multiplication memorized paper written applies space asked
work students past ask work library work carpet answer always long
learning work wherever want yoga balls lap desks able increase options
seating classroom expand imaginable space nannan
--------------------------------------------------
147 students eager learn make mark world come title 1 school need
extra love fourth grade students high poverty area still come school
every day get education trying make fun educational get schooling
created caring environment students bloom deserve best thank
requesting 1 chromebook access online interventions differentiate
instruction get extra practice chromebook used supplement ela math
instruction students play ela math games engaging fun well participate
assignments online turn help students improve skills chromebook
classroom would not allow students use programs pace would ensure
students getting adequate time use programs online programs especially
beneficial students special needs able work level well challenged
different materials making students confident abilities chromebook
would allow students daily access computers increase computing skills
change lives better become successful school access technology
classroom would help bridge achievement gap nannan
```

```
# preprocessing project_resource_summary column
project_data['project_resource_summary'] =
preprocess_text(project_data['project_resource_summary'].values)
```

```
100%|██████████| 109248/109248 [00:08<00:00, 12982.65it/s]
```

```
project_data.head()
```

```
   Unnamed: 0       id                          teacher_id
teacher_prefix  \
0      160221  p253737  c90749f5d961ff158d4b4d1e7dc665fc
Mrs.
1      140945  p258326  897464ce9ddc600bced1151f324dd63a
Mr.
2       21895  p182444  3465aaf82da834c0582ebd0ef8040ca0
Ms.
3          45  p246581  f3cb9bffbba169bef1a77b243e620b60
Mrs.
4      172407  p104768  be1f7507a41f8479dc06f047086a39ec
Mrs.

  school_state project_submitted_datetime project_grade_category  \
0           in        2016-12-05 13:43:57          grades_prek_2
1           fl        2016-10-25 09:22:10             grades_6_8
2           az        2016-08-31 12:03:56             grades_6_8
```

```
3           ky        2016-10-06 21:16:17          grades_prek_2
4           tx        2016-07-11 01:10:09          grades_prek_2

            project_subject_categories
project_subject_subcategories  \
0                  Literacy & Language                     ESL,
Literacy
1    History & Civics, Health & Sports  Civics & Government, Team
Sports
2                      Health & Sports    Health & Wellness, Team
Sports
3  Literacy & Language, Math & Science             Literacy,
Mathematics
4                       Math & Science
Mathematics

                                        project_title  \
0   Educational Support for English Learners at Home
1              Wanted: Projector for Hungry Learners
2  Soccer Equipment for AWESOME Middle School Stu...
3                            Techie Kindergarteners
4                            Interactive Math Tools

                                       project_essay_1  \
0  My students are English learners that are work...
1  Our students arrive to our school eager to lea...
2  \r\n\"True champions aren't always the ones th...
3  I work at a unique school filled with both ESL...
4  Our second grade classroom next year will be m...

                                       project_essay_2
project_essay_3  \
0  \"The limits of your language are the limits o...             NaN

1  The projector we need for our school is very c...             NaN

2  The students on the campus come to school know...             NaN

3  My students live in high poverty conditions wi...             NaN

4  For many students, math is a subject that does...             NaN


  project_essay_4
project_resource_summary  \
0             NaN  students need opportunities practice beginning...

1             NaN  students need projector help viewing education...
```

```
2             NaN  students need shine guards athletic socks socc...

3             NaN  students need engage reading math way inspire ...

4             NaN  students need hands practice mathematics fun p...


   teacher_number_of_previously_posted_projects
project_is_approved  \
0                                             0                    0

1                                             7                    1

2                                             1                    0

3                                             4                    1

4                                             1                    1


                                                 essay
0  students english learners working english seco...
1  students arrive school eager learn polite gene...
2  true champions not always ones win guts mia ha...
3  work unique school filled esl english second l...
4  second grade classroom next year made around 2...
```

8. Preprocessing Numerical Values: price

```
# https://stackoverflow.com/questions/22407798/how-to-reset-a-
dataframes-indexes-for-all-groups-in-one-step
price_data = resource_data.groupby('id').agg({'price':'sum',
'quantity':'sum'}).reset_index()
price_data.head(2)
```

```
        id   price  quantity
0  p000001  459.56         7
1  p000002  515.89        21
```

```
# join two dataframes in python:
project_data = pd.merge(project_data, price_data, on='id', how='left')

project_data.head()
```

```
   Unnamed: 0       id                          teacher_id
teacher_prefix  \
0      160221  p253737  c90749f5d961ff158d4b4d1e7dc665fc
Mrs.
1      140945  p258326  897464ce9ddc600bced1151f324dd63a
Mr.
```

```
2       21895  p182444  3465aaf82da834c0582ebd0ef8040ca0
Ms.
3          45  p246581  f3cb9bffbba169bef1a77b243e620b60
Mrs.
4      172407  p104768  be1f7507a41f8479dc06f047086a39ec
Mrs.

  school_state project_submitted_datetime project_grade_category  \
0           in        2016-12-05 13:43:57          grades_prek_2
1           fl        2016-10-25 09:22:10             grades_6_8
2           az        2016-08-31 12:03:56             grades_6_8
3           ky        2016-10-06 21:16:17          grades_prek_2
4           tx        2016-07-11 01:10:09          grades_prek_2

            project_subject_categories
project_subject_subcategories  \
0                  Literacy & Language                     ESL,
Literacy
1    History & Civics, Health & Sports  Civics & Government, Team
Sports
2                      Health & Sports    Health & Wellness, Team
Sports
3  Literacy & Language, Math & Science             Literacy,
Mathematics
4                       Math & Science
Mathematics

                                      project_title  \
0   Educational Support for English Learners at Home
1              Wanted: Projector for Hungry Learners
2  Soccer Equipment for AWESOME Middle School Stu...
3                           Techie Kindergarteners
4                           Interactive Math Tools

                                     project_essay_1  \
0  My students are English learners that are work...
1  Our students arrive to our school eager to lea...
2  \r\n\"True champions aren't always the ones th...
3  I work at a unique school filled with both ESL...
4  Our second grade classroom next year will be m...

                                     project_essay_2
project_essay_3  \
0  \"The limits of your language are the limits o...             NaN

1  The projector we need for our school is very c...             NaN

2  The students on the campus come to school know...             NaN
```

```
3  My students live in high poverty conditions wi...             NaN

4  For many students, math is a subject that does...             NaN


  project_essay_4
project_resource_summary  \
0             NaN  students need opportunities practice beginning...

1             NaN  students need projector help viewing education...

2             NaN  students need shine guards athletic socks socc...

3             NaN  students need engage reading math way inspire ...

4             NaN  students need hands practice mathematics fun p...


   teacher_number_of_previously_posted_projects
project_is_approved  \
0                                             0                    0

1                                             7                    1

2                                             1                    0

3                                             4                    1

4                                             1                    1


                                              essay   price  quantity

0  students english learners working english seco...  154.60        23

1  students arrive school eager learn polite gene...  299.00         1

2  true champions not always ones win guts mia ha...  516.85        22

3  work unique school filled esl english second l...  232.90         4

4  second grade classroom next year made around 2...   67.98         4
```

```
# dropping the columns which dont add value to our Model

project_data = project_data.drop(['Unnamed: 0', 'id', 'teacher_id',
```

```
'project_submitted_datetime', 'project_essay_1', 'project_essay_2',
                   'project_essay_3', 'project_essay_4'], axis=1)

project_data.shape
```

```
(109248, 12)
```

```
project_data.head()
```

```
  teacher_prefix school_state project_grade_category  \
0           Mrs.           in          grades_prek_2
1            Mr.           fl             grades_6_8
2            Ms.           az             grades_6_8
3           Mrs.           ky          grades_prek_2
4           Mrs.           tx          grades_prek_2

            project_subject_categories
project_subject_subcategories  \
0                  Literacy & Language                     ESL,
Literacy
1    History & Civics, Health & Sports  Civics & Government, Team
Sports
2                     Health & Sports    Health & Wellness, Team
Sports
3  Literacy & Language, Math & Science             Literacy,
Mathematics
4                      Math & Science
Mathematics

                                      project_title  \
0   Educational Support for English Learners at Home
1              Wanted: Projector for Hungry Learners
2  Soccer Equipment for AWESOME Middle School Stu...
3                           Techie Kindergarteners
4                           Interactive Math Tools

                           project_resource_summary  \
0  students need opportunities practice beginning...
1  students need projector help viewing education...
2  students need shine guards athletic socks socc...
3  students need engage reading math way inspire ...
4  students need hands practice mathematics fun p...

   teacher_number_of_previously_posted_projects
project_is_approved  \
0                                             0                    0

1                                             7                    1

2                                             1                    0
```

```
3                                                 4                     1

4                                                 1                     1
```

```
                                                 essay   price  quantity
0  students english learners working english seco...  154.60        23

1  students arrive school eager learn polite gene...  299.00         1

2  true champions not always ones win guts mia ha...  516.85        22

3  work unique school filled esl english second l...  232.90         4

4  second grade classroom next year made around 2...   67.98         4
```

Performing sentiment analysis

```
# examples for sentiment analysis
# this gives us the sentiment score based on the type of sentance

# below are some of the examples

sid = SentimentIntensityAnalyzer()

sample_sentence_1='you went off!!!'
ss_1 = sid.polarity_scores(sample_sentence_1)
print('sentiment score for sentence 1',ss_1)

sample_sentence_2='I am sad.'
ss_2 = sid.polarity_scores(sample_sentence_2)
print('sentiment score for sentence 2',ss_2)

sample_sentence_3='I am going to New Delhi tommorow.'
ss_3 = sid.polarity_scores(sample_sentence_3)
print('sentiment score for sentence 3',ss_3)
```

```
sentiment score for sentence 1 {'neg': 0.0, 'neu': 1.0, 'pos': 0.0,
'compound': 0.0}
sentiment score for sentence 2 {'neg': 0.756, 'neu': 0.244, 'pos':
0.0, 'compound': -0.4767}
sentiment score for sentence 3 {'neg': 0.0, 'neu': 1.0, 'pos': 0.0,
'compound': 0.0}
```

```
print("Number of columns in our Data before adding Sentiment Analysis
featues are:", project_data.shape[1])
```

```
Number of columns in our Data before adding Sentiment Analysis featues
are: 12
```

```
project_data.shape
```

```
(109248, 12)
```

```
# adding required columns to our Data to add the features of Sentiment
Analysis
project_data['negative_score']=''
project_data['positive_score']=''
project_data['neutral_score']=''
project_data['compound_score']=''

print("Number of columns in our Data after adding necessary columns
for Sentiment Analysis featues are:", project_data.shape[1])
```

```
Number of columns in our Data after adding necessary columns for
Sentiment Analysis featues are: 16
```

```
sid = SentimentIntensityAnalyzer()

for k in tqdm(range(project_data.shape[0])):
  #print(k)
  ss = sid.polarity_scores(project_data['essay'][k])
  #print(ss)
  #print(ss['neg'])
  project_data['negative_score'][k] = ss['neg']
  project_data['neutral_score'][k] = ss['neu']
  project_data['positive_score'][k] = ss['pos']
  project_data['compound_score'][k] = ss['compound']
```

```
100%|██████████| 109248/109248 [15:04<00:00, 120.80it/s]
```

```
# columns that has null values
project_data.isna().sum()
```

```
teacher_prefix                                  0
school_state                                    0
project_grade_category                          0
project_subject_categories                      0
project_subject_subcategories                   0
project_title                                   0
project_resource_summary                        0
teacher_number_of_previously_posted_projects    0
project_is_approved                             0
essay                                           0
price                                           0
quantity                                        0
negative_score                                  0
positive_score                                  0
neutral_score                                   0
```

```
compound_score                                    0
dtype: int64
```

```
print("Printing the indexs that of the rows that have Null value in
'project_title' column")
project_data[project_data['project_title'].isnull()].index.tolist()
```

```
Printing the indexs that of the rows that have Null value in
'project_title' column
```

```
[]
```

```
preprocessed_project_data = project_data
```

```
# converting into a new csv file without storing index values but with
headers (column names)
# index = False --- to not include row numbers
```

```
preprocessed_project_data.to_csv('preprocessed_project_data.csv',
index = False)
```

```
preprocessed_project_data.shape
```

```
(109248, 16)
```