# DonorsChoose

DonorsChoose.org receives hundreds of thousands of project proposals each year for classroom projects in need of funding. Right now, a large number of volunteers is needed to manually screen each submission before it's approved to be posted on the DonorsChoose.org website.

Next year, DonorsChoose.org expects to receive close to 500,000 project proposals. As a result, there are three main problems they need to solve:

- How to scale current manual processes and resources to screen 500,000 projects so that they can be posted as quickly and as efficiently as possible
- How to increase the consistency of project vetting across different volunteers to improve the experience for teachers
- How to focus volunteer time on the applications that need the most assistance

The goal of the competition is to predict whether or not a DonorsChoose.org project proposal submitted by a teacher will be approved, using the text of project descriptions as well as additional metadata about the project, teacher, and school. DonorsChoose.org can then use this information to identify projects most likely to need further review before approval.

## About the DonorsChoose Data Set

The `train.csv` data set provided by DonorsChoose contains the following features:

| Feature | Description |
| --- | --- |
| **`project_id`** | A unique identifier for the proposed project. **Example:** `p036502` |
| **`project_title`** | Title of the project. **Examples:**<br>• `Art Will Make You Happy!`<br>• `First Grade Fun` |
| **`project_grade_category`** | Grade level of students for which the project is targeted. One of the following enumerated values:<br>• `Grades PreK-2`<br>• `Grades 3-5`<br>• `Grades 6-8`<br>• `Grades 9-12` |
| **`project_subject_categories`** | One or more (comma-separated) subject categories for the project from the following enumerated list of values:<br>• `Applied Learning`<br>• `Care & Hunger`<br>• `Health & Sports`<br>• `History & Civics`<br>• `Literacy & Language`<br>• `Math & Science`<br>• `Music & The Arts`<br>• `Special Needs`<br>• `Warmth`<br>**Examples:**<br>• `Music & The Arts`<br>• `Literacy & Language, Math & Science` |
| **`school_state`** | State where school is located (Two-letter U.S. postal code). **Example:** `WY` |
| **`project_subject_subcategories`** | One or more (comma-separated) subject subcategories for the project. **Examples:**<br>• `Literacy` |

| Feature | Description |
| --- | --- |
| | • Literature & Writing, Social Sciences |
| **project_resource_summary** | An explanation of the resources needed for the project. **Example:**<br>• My students need hands on literacy materials to manage sensory needs! |
| **project_essay_1** | First application essay* |
| **project_essay_2** | Second application essay* |
| **project_essay_3** | Third application essay* |
| **project_essay_4** | Fourth application essay* |
| **project_submitted_datetime** | Datetime when project application was submitted. **Example:** 2016-04-28 12:43:56.245 |
| **teacher_id** | A unique identifier for the teacher of the proposed project. **Example:** bdf8baa8fedef6bfeec7ae4ff1c15c56 |
| **teacher_prefix** | Teacher's title. One of the following enumerated values:<br>• nan<br>• Dr.<br>• Mr.<br>• Mrs.<br>• Ms.<br>• Teacher. |
| **teacher_number_of_previously_posted_projects** | Number of project applications previously submitted by the same teacher. **Example:** 2 |

* See the section **Notes on the Essay Data** for more details about these features.

Additionally, the resources.csv data set provides more data about the resources required for each project. Each line in this file represents a resource required by a project:

| Feature | Description |
| --- | --- |
| **id** | A project_id value from the train.csv file. **Example:** p036502 |
| **description** | Desciption of the resource. **Example:** Tenor Saxophone Reeds, Box of 25 |
| **quantity** | Quantity of the resource required. **Example:** 3 |
| **price** | Price of the resource required. **Example:** 9.95 |

**Note:** Many projects require multiple resources. The id value corresponds to a project_id in train.csv, so you use it as a key to retrieve all resources needed for a project:

The data set contains the following label (the value you will attempt to predict):

| Label | Description |
| --- | --- |
| project_is_approved | A binary flag indicating whether DonorsChoose approved the project. A value of 0 indicates the project was not approved, and a value of 1 indicates the project was approved. |

## Notes on the Essay Data

Prior to May 17, 2016, the prompts for the essays were as follows:

- __project_essay_1:__ "Introduce us to your classroom"
- __project_essay_2:__ "Tell us more about your students"
- __project_essay_3:__ "Describe how your students will use the materials you're requesting"
- __project_essay_4:__ "Close by sharing why your project will make a difference"

Starting on May 17, 2016, the number of essays was reduced from 4 to 2, and the prompts for the first 2 essays were changed to the following:

- __project_essay_1:__ "Describe your students: What makes your students special? Specific details about their background, your neighborhood, and your school are all helpful."

your neighborhood, and your school are all helpful.

- __project_essay_2:__ "About your project: How will these materials make a difference in your students' learning and improve their school lives?"

  For all projects with project_submitted_datetime of 2016-05-17 and later, the values of project_essay_3 and project_essay_4 will be NaN.

In [124]:

```
import warnings
warnings.filterwarnings('ignore')

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

In [125]:

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

In [126]:

```
project_data.head(3)
```

Out[126]:

| | Unnamed: 0 | id | teacher_id | teacher_prefix | school_state | project_submitted_datetime | pro |
|---|---|---|---|---|---|---|---|
| **0** | 160221 | p253737 | c90749f5d961ff158d4b4d1e7dc665fc | Mrs. | IN | 2016-12-05 13:43:57 | Gra |
| **1** | 140945 | p258326 | 897464ce9ddc600bced1151f324dd63a | Mr. | FL | 2016-10-25 09:22:10 | Gra |
| **2** | 21895 | p182444 | 3465aaf82da834c0582ebd0ef8040ca0 | Ms. | AZ | 2016-08-31 12:03:56 | Gra |

In [127]:

```
project_data.describe()
```

Out[127]:

| | Unnamed: 0 | teacher_number_of_previously_posted_projects | project_is_approved |
|---|---|---|---|
| **count** | 109248.000000 | 109248.000000 | 109248.000000 |
| **mean** | 91183.786568 | 11.153165 | 0.848583 |
| **std** | 52548.095272 | 27.777154 | 0.358456 |
| **min** | 0.000000 | 0.000000 | 0.000000 |
| **25%** | 45743.500000 | 0.000000 | 1.000000 |
| **50%** | 91253.500000 | 2.000000 | 1.000000 |
| **75%** | 136712.500000 | 9.000000 | 1.000000 |

| | Unnamed: 0 | teacher_number_of_previously_posted_projects | project_is_approved |
|---|---|---|---|
| max | 182079.000000 | 451.000000 | 1.000000 |

In [128]:

```
project_data.shape
```

Out[128]:

```
(109248, 17)
```

In [129]:

```
resources_data.head()
```

Out[129]:

| | id | description | quantity | price |
|---|---|---|---|---|
| 0 | p233245 | LC652 - Lakeshore Double-Space Mobile Drying Rack | 1 | 149.00 |
| 1 | p069063 | Bouncy Bands for Desks (Blue support pipes) | 3 | 14.95 |
| 2 | p069063 | Cory Stories: A Kid's Book About Living With Adhd | 1 | 8.45 |
| 3 | p069063 | Dixon Ticonderoga Wood-Cased #2 HB Pencils, Bo... | 2 | 13.59 |
| 4 | p069063 | EDUCATIONAL INSIGHTS FLUORESCENT LIGHT FILTERS... | 3 | 24.95 |

In [130]:

```
value_counts = project_data['project_is_approved'].value_counts()
```

In [131]:

```
print('the percentage of projects approved', value_counts[1]/ (value_counts[0]+value_counts[1])*100
)
```

```
the percentage of projects approved 84.85830404217927
```

In [132]:

```
print('the percentage of projects not approved', value_counts[0]/(value_counts[0]+value_counts[1])
*100)
```

```
the percentage of projects not approved 15.141695957820739
```

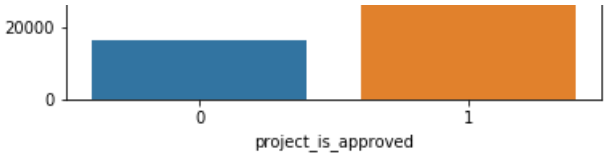
# 1.2 Data Analysis

In [133]:

```
sns.countplot(x='project_is_approved', data=project_data)
```

Out[133]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x14694135fd0>
```




In [134]:

```
project_data.head(5)
```

Out[134]:

| | Unnamed: 0 | id | teacher_id | teacher_prefix | school_state | project_submitted_datetime | pro |
|---|---|---|---|---|---|---|---|
| 0 | 160221 | p253737 | c90749f5d961ff158d4b4d1e7dc665fc | Mrs. | IN | 2016-12-05 13:43:57 | Gra |
| 1 | 140945 | p258326 | 897464ce9ddc600bced1151f324dd63a | Mr. | FL | 2016-10-25 09:22:10 | Gra |
| 2 | 21895 | p182444 | 3465aaf82da834c0582ebd0ef8040ca0 | Ms. | AZ | 2016-08-31 12:03:56 | Gra |
| 3 | 45 | p246581 | f3cb9bffbba169bef1a77b243e620b60 | Mrs. | KY | 2016-10-06 21:16:17 | Gra |
| 4 | 172407 | p104768 | be1f7507a41f8479dc06f047086a39ec | Mrs. | TX | 2016-07-11 01:10:09 | Gra |

### 1.2.1 Univariate Analysis: School State

In [135]:

```
#lowest percentage of projects approved by school states
temp = (pd.DataFrame(project_data.groupby(by='school_state')['project_is_approved'].apply(lambda x:
np.mean(x))).reset_index()).sort_values(by='project_is_approved')
```

In [136]:

```
temp.head()
```

Out[136]:

| | school_state | project_is_approved |
|---|---|---|
| 46 | VT | 0.800000 |
| 7 | DC | 0.802326 |
| 43 | TX | 0.813142 |
| 26 | MT | 0.816327 |
| 18 | LA | 0.831245 |

In [137]:

```
temp.columns
```

Out[137]:

```
Index(['school_state', 'project_is_approved'], dtype='object')
```

In [138]:

```
temp.columns = ['school_state', 'num_proposals']
```

In [139]:

```
temp.head()
```

Out[139]:

| | school_state | num_proposals |
|---|---|---|
| **46** | VT | 0.800000 |
| **7** | DC | 0.802326 |
| **43** | TX | 0.813142 |
| **26** | MT | 0.816327 |
| **18** | LA | 0.831245 |

In [140]:

```
#stacked bar plots matplotlib:
https://matplotlib.org/gallery/lines_bars_and_markers/bar_stacked.html
def stack_plot(data, xtick, col2='project_is_approved', col3='total'):
    ind = np.arange(data.shape[0])

    plt.figure(figsize=(20,5))
    p1 = plt.bar(ind, data[col3].values)
    p2 = plt.bar(ind, data[col2].values)

    plt.ylabel('Projects')
    plt.title('Number of projects aproved vs rejected')
    plt.xticks(ind, list(data[xtick].values))
    plt.legend((p1[0], p2[0]), ('total', 'accepted'))
    plt.show()
```

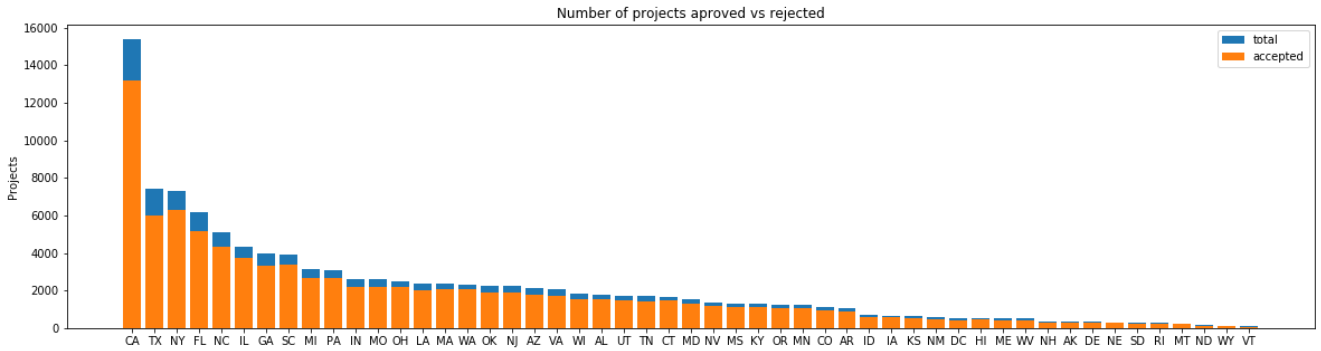
In [141]:

```
def univariate_barplots(data, col1, col2='project_is_approved', top=False):
    # Count number of zeros in dataframe python: https://stackoverflow.com/a/51540521/4084039
    temp = pd.DataFrame(project_data.groupby(col1)[col2].agg(lambda x: x.eq(1).sum())).reset_index(
)

    # Pandas dataframe grouby count: https://stackoverflow.com/a/19385591/4084039
    temp['total'] = pd.DataFrame(project_data.groupby(col1)
[col2].agg({'total':'count'})).reset_index()['total']
    temp['Avg'] = pd.DataFrame(project_data.groupby(col1)[col2].agg({'Avg':'mean'})).reset_index()[
'Avg']

    temp.sort_values(by=['total'],inplace=True, ascending=False)

    if top:
        temp = temp[0:top]

    stack_plot(temp, xtick=col1, col2=col2, col3='total')
    print(temp.head(5))
    print("="*50)
    print(temp.tail(5))
```

In [142]:

In [142]:

```
univariate_barplots(project_data, 'school_state', 'project_is_approved', False)
```



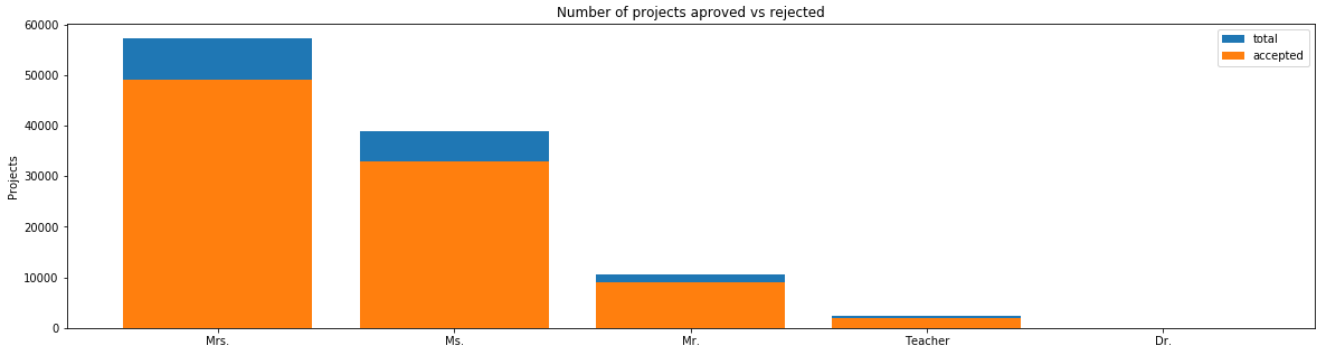

```
   school_state  project_is_approved  total       Avg
4            CA                13205  15388  0.858136
43           TX                 6014   7396  0.813142
34           NY                 6291   7318  0.859661
9            FL                 5144   6185  0.831690
27           NC                 4353   5091  0.855038
==================================================
   school_state  project_is_approved  total       Avg
39           RI                  243    285  0.852632
26           MT                  200    245  0.816327
28           ND                  127    143  0.888112
50           WY                   82     98  0.836735
46           VT                   64     80  0.800000
```

**SUMMARY: Every state has greater than 80% success rate in approval**

### 1.2.2 Univariate Analysis: teacher_prefix

In [143]:

```
univariate_barplots(project_data, 'teacher_prefix', 'project_is_approved', False)
```

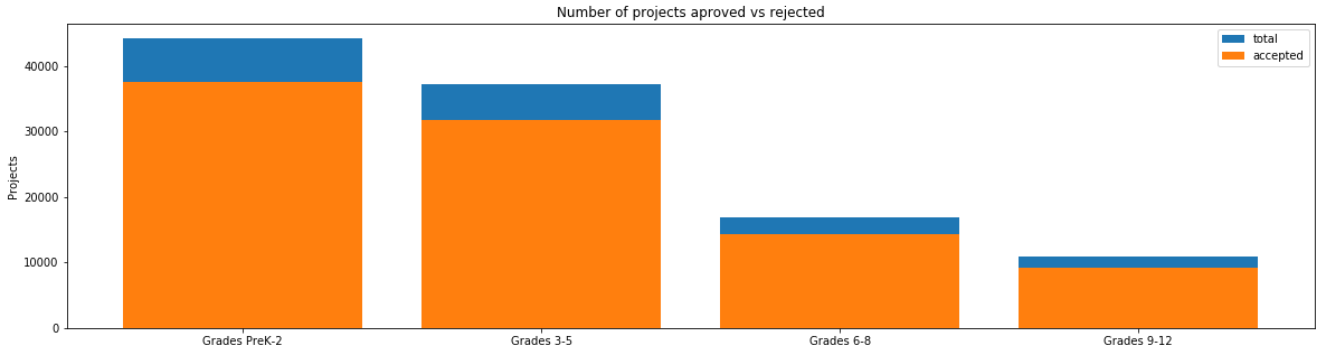



```
  teacher_prefix  project_is_approved  total       Avg
2           Mrs.                48997  57269  0.855559
3            Ms.                32860  38955  0.843537
1            Mr.                 8960  10648  0.841473
4        Teacher                 1877   2360  0.795339
0            Dr.                    9     13  0.692308
==================================================
  teacher_prefix  project_is_approved  total       Avg
2           Mrs.                48997  57269  0.855559
3            Ms.                32860  38955  0.843537
1            Mr.                 8960  10648  0.841473
4        Teacher                 1877   2360  0.795339
0            Dr.                    9     13  0.692308
```

### 1.2.3 Univariate Analysis: project_grade_category

In [144]:

```
univariate_barplots(project_data, 'project_grade_category', 'project_is_approved', top=False)
```




```
  project_grade_category  project_is_approved  total       Avg
3          Grades PreK-2                37536  44225  0.848751
0              Grades 3-5                31729  37137  0.854377
1              Grades 6-8                14258  16923  0.842522
2             Grades 9-12                 9183  10963  0.837636
==================================================
  project_grade_category  project_is_approved  total       Avg
3          Grades PreK-2                37536  44225  0.848751
0              Grades 3-5                31729  37137  0.854377
1              Grades 6-8                14258  16923  0.842522
2             Grades 9-12                 9183  10963  0.837636
```

### 1.2.4 Univariate Analysis: project_subject_categories

In [145]:

```
xxx = list(project_data['project_subject_categories'])
```

In [146]:

```
type(xxx)
```

Out[146]:

```
list
```

In [147]:

```
# remove special characters from list of strings python:
https://stackoverflow.com/a/47301924/4084039

# https://www.geeksforgeeks.org/removing-stop-words-nltk-python/
# https://stackoverflow.com/questions/23669024/how-to-strip-a-specific-word-from-a-string
# https://stackoverflow.com/questions/8270092/remove-all-whitespace-in-a-string-in-python

cat_list = []
for i in xxx:
    temp = ''
    for j in i.split(','):
        if 'The' in j.split():
            j = j.replace(' ', '')

        j = j.replace(' ','')
        temp +=j.strip()+' '
        temp = temp.replace('&', '_')

    cat_list.append(temp.strip())
```

In [148]:

```
cat_list
```

Out[148]:

```
['Literacy_Language',
 'History_Civics Health_Sports',
 'Health_Sports',
 'Literacy_Language Math_Science',
 'Math_Science',
 'Literacy_Language SpecialNeeds',
 'Literacy_Language SpecialNeeds',
 'Math_Science',
 'Health_Sports',
 'Literacy_Language',
 'Literacy_Language',
 'Literacy_Language AppliedLearning',
 'Math_Science',
 'SpecialNeeds',
 'Literacy_Language',
 'Health_Sports',
 'Literacy_Language SpecialNeeds',
 'Math_Science Literacy_Language',
 'AppliedLearning',
 'Health_Sports',
 'Literacy_Language',
 'Math_Science SpecialNeeds',
 'Literacy_Language',
 'Music_TheArts',
 'Math_Science',
 'Math_Science',
 'Literacy_Language Math_Science',
 'Literacy_Language Math_Science',
 'Literacy_Language SpecialNeeds',
 'Literacy_Language AppliedLearning',
 'Literacy_Language',
 'SpecialNeeds',
 'Math_Science Literacy_Language',
 'History_Civics',
 'Literacy_Language',
 'Health_Sports',
 'Literacy_Language Math_Science',
 'Health_Sports Literacy_Language',
 'Health_Sports',
 'Literacy_Language',
 'Literacy_Language',
 'Literacy_Language',
 'Literacy_Language',
 'Literacy_Language Music_TheArts',
 'Math_Science',
 'Literacy_Language',
 'Literacy_Language',
 'Warmth Care_Hunger',
 'Literacy_Language Math_Science',
 'Health_Sports',
 'Health_Sports',
 'Literacy_Language',
 'Math_Science History_Civics',
 'Literacy_Language',
 'Health_Sports',
 'Math_Science',
 'SpecialNeeds',
 'Literacy_Language Math_Science',
 'Literacy_Language',
 'Literacy_Language',
 'Health_Sports',
 'Math_Science',
 'Literacy_Language',
 'Music_TheArts',
 'Music_TheArts',
 'SpecialNeeds',
 'Math_Science',
 'Literacy_Language',
 'Math_Science',
 'AppliedLearning Literacy_Language',
 'Math_Science',
 'Math_Science',
 'Literacy_Language',
 'AppliedLearning',
 'Math_Science',
```

```
'Music_TheArts',
'Literacy_Language Math_Science',
'AppliedLearning SpecialNeeds',
'Math_Science',
'Music_TheArts',
'Literacy_Language',
'Math_Science',
'Literacy_Language',
'Literacy_Language History_Civics',
'Math_Science',
'Literacy_Language',
'Health_Sports',
'Literacy_Language SpecialNeeds',
'SpecialNeeds',
'Literacy_Language',
'Literacy_Language Music_TheArts',
'Health_Sports',
'AppliedLearning Health_Sports',
'Music_TheArts',
'Literacy_Language',
'Math_Science',
'Literacy_Language',
'Literacy_Language',
'AppliedLearning',
'Math_Science',
'Literacy_Language Math_Science',
'Health_Sports',
'Literacy_Language Math_Science',
'Music_TheArts',
'Health_Sports',
'Music_TheArts',
'Literacy_Language',
'Literacy_Language',
'SpecialNeeds',
'SpecialNeeds',
'Literacy_Language Math_Science',
'Literacy_Language',
'Literacy_Language',
'Literacy_Language Math_Science',
'Literacy_Language',
'Literacy_Language Music_TheArts',
'Literacy_Language SpecialNeeds',
'History_Civics Literacy_Language',
'Math_Science',
'Math_Science',
'Literacy_Language Math_Science',
'Literacy_Language',
'Health_Sports SpecialNeeds',
'Literacy_Language SpecialNeeds',
'SpecialNeeds',
'Math_Science',
'AppliedLearning Literacy_Language',
'Music_TheArts',
'Music_TheArts',
'Literacy_Language SpecialNeeds',
'Literacy_Language',
'Literacy_Language',
'Literacy_Language',
'Health_Sports',
'Health_Sports SpecialNeeds',
'Music_TheArts',
'Literacy_Language SpecialNeeds',
'Literacy_Language',
'AppliedLearning Math_Science',
'Math_Science',
'Literacy_Language SpecialNeeds',
'Literacy_Language',
'Health_Sports',
'Math_Science',
'Literacy_Language',
'AppliedLearning',
'Literacy_Language Math_Science',
'Literacy_Language Math_Science',
'AppliedLearning Literacy_Language',
'Math_Science',
'Health_Sports',
'Math_Science',
```

```
'Math_Science',
'SpecialNeeds',
'Math_Science History_Civics',
'Literacy_Language Math_Science',
'Health_Sports',
'Math_Science',
'Math_Science',
'Math_Science Music_TheArts',
'Literacy_Language',
'History_Civics Literacy_Language',
'Health_Sports',
'Literacy_Language Music_TheArts',
'Math_Science',
'Health_Sports',
'Music_TheArts',
'Health_Sports',
'Math_Science',
'Math_Science SpecialNeeds',
'Literacy_Language SpecialNeeds',
'Literacy_Language Math_Science',
'Literacy_Language',
'AppliedLearning',
'Literacy_Language Math_Science',
'Literacy_Language',
'Math_Science Music_TheArts',
'Literacy_Language',
'History_Civics Literacy_Language',
'Literacy_Language Math_Science',
'Literacy_Language Math_Science',
'Math_Science',
'Literacy_Language Math_Science',
'Math_Science',
'Math_Science',
'Math_Science',
'Literacy_Language Math_Science',
'Music_TheArts',
'Math_Science',
'Literacy_Language Math_Science',
'Math_Science',
'SpecialNeeds',
'Literacy_Language Math_Science',
'Health_Sports',
'Literacy_Language Math_Science',
'Math_Science',
'Health_Sports AppliedLearning',
'Math_Science',
'Literacy_Language',
'Literacy_Language',
'AppliedLearning',
'Music_TheArts',
'Math_Science',
'Literacy_Language SpecialNeeds',
'Literacy_Language Math_Science',
'Literacy_Language',
'Math_Science',
'Math_Science',
'Music_TheArts',
'Math_Science',
'Literacy_Language',
'Music_TheArts',
'Health_Sports',
'AppliedLearning Literacy_Language',
'Math_Science',
'Literacy_Language',
'Math_Science',
'Literacy_Language',
'Music_TheArts',
'Math_Science',
'AppliedLearning',
'Literacy_Language Math_Science',
'Health_Sports',
'AppliedLearning Math_Science',
'Math_Science',
'Literacy_Language History_Civics',
'AppliedLearning SpecialNeeds',
'Literacy_Language',
'Music_TheArts',
```

```
'History_Civics Literacy_Language',
'Literacy_Language Math_Science',
'Literacy_Language SpecialNeeds',
'Literacy_Language Math_Science',
'Math_Science',
'SpecialNeeds',
'Literacy_Language SpecialNeeds',
'AppliedLearning Health_Sports',
'Math_Science',
'Literacy_Language History_Civics',
'Literacy_Language Math_Science',
'Literacy_Language',
'AppliedLearning SpecialNeeds',
'History_Civics Music_TheArts',
'AppliedLearning',
'AppliedLearning',
'AppliedLearning',
'Literacy_Language Math_Science',
'Math_Science AppliedLearning',
'Warmth Care_Hunger',
'Math_Science',
'Literacy_Language SpecialNeeds',
'Math_Science',
'Math_Science',
'Literacy_Language',
'Math_Science',
'Math_Science',
'Literacy_Language SpecialNeeds',
'Literacy_Language Math_Science',
'Health_Sports SpecialNeeds',
'Literacy_Language',
'Literacy_Language',
'Literacy_Language',
'Literacy_Language',
'History_Civics Literacy_Language',
'Literacy_Language Math_Science',
'Health_Sports',
'Literacy_Language',
'Literacy_Language Math_Science',
'Music_TheArts History_Civics',
'Literacy_Language',
'Math_Science',
'Literacy_Language History_Civics',
'Literacy_Language Math_Science',
'Math_Science',
'Math_Science',
'Health_Sports',
'Literacy_Language Math_Science',
'Health_Sports',
'Literacy_Language',
'AppliedLearning',
'Math_Science',
'Health_Sports SpecialNeeds',
'Literacy_Language',
'Music_TheArts',
'Math_Science SpecialNeeds',
'Literacy_Language Math_Science',
'Literacy_Language SpecialNeeds',
'Warmth Care_Hunger',
'Health_Sports',
'AppliedLearning SpecialNeeds',
'Literacy_Language',
'History_Civics',
'Literacy_Language Math_Science',
'History_Civics',
'Math_Science',
'Music_TheArts',
'Music_TheArts',
'Health_Sports',
'AppliedLearning Literacy_Language',
'Math_Science',
'Math_Science',
'Math_Science',
'Math_Science',
'Literacy_Language Math_Science',
'Literacy_Language Music_TheArts',
'Math_Science Literacy_Language',
```

```
'Health_Sports',
'Health_Sports',
'Math_Science SpecialNeeds',
'Math_Science',
'Health_Sports',
'Literacy_Language SpecialNeeds',
'Literacy_Language AppliedLearning',
'Health_Sports',
'AppliedLearning',
'Literacy_Language',
'Health_Sports',
'AppliedLearning Literacy_Language',
'Literacy_Language',
'Math_Science',
'Literacy_Language',
'Literacy_Language Math_Science',
'Literacy_Language Math_Science',
'AppliedLearning Literacy_Language',
'Literacy_Language Math_Science',
'Literacy_Language Math_Science',
'Math_Science AppliedLearning',
'Math_Science',
'History_Civics Literacy_Language',
'Health_Sports',
'Health_Sports',
'Literacy_Language History_Civics',
'Health_Sports',
'Math_Science',
'Literacy_Language Math_Science',
'Music_TheArts',
'Music_TheArts',
'AppliedLearning Literacy_Language',
'Literacy_Language',
'Health_Sports',
'Literacy_Language',
'Math_Science',
'Literacy_Language Math_Science',
'Literacy_Language',
'Math_Science',
'Literacy_Language',
'AppliedLearning Literacy_Language',
'Math_Science',
'AppliedLearning',
'Literacy_Language Math_Science',
'Literacy_Language',
'Literacy_Language',
'Literacy_Language',
'Health_Sports',
'AppliedLearning',
'Math_Science',
'AppliedLearning Health_Sports',
'Literacy_Language Music_TheArts',
'Health_Sports',
'Math_Science',
'Math_Science',
'Literacy_Language',
'Literacy_Language Math_Science',
'Music_TheArts',
'Literacy_Language',
'Health_Sports AppliedLearning',
'AppliedLearning',
'Literacy_Language',
'Literacy_Language Math_Science',
'Math_Science',
'Literacy_Language',
'Literacy_Language',
'Math_Science AppliedLearning',
'Literacy_Language Math_Science',
'Health_Sports',
'Literacy_Language Music_TheArts',
'Math_Science SpecialNeeds',
'Literacy_Language Music_TheArts',
'Literacy_Language',
'Literacy_Language',
'Literacy_Language',
'Health_Sports',
'Math_Science Literacy_Language',
```

```
'Literacy_Language Math_Science',
'Literacy_Language',
'AppliedLearning Math_Science',
'Math_Science',
'AppliedLearning',
'AppliedLearning Math_Science',
'Music_TheArts',
'Literacy_Language',
'Literacy_Language',
'Literacy_Language Math_Science',
'Health_Sports',
'AppliedLearning Math_Science',
'Literacy_Language SpecialNeeds',
'AppliedLearning Literacy_Language',
'Math_Science SpecialNeeds',
'Literacy_Language',
'Literacy_Language',
'Math_Science',
'Literacy_Language',
'Literacy_Language Math_Science',
'History_Civics',
'Math_Science',
'AppliedLearning SpecialNeeds',
'Literacy_Language',
'Music_TheArts',
'Health_Sports',
'Math_Science',
'Literacy_Language Math_Science',
'Literacy_Language Math_Science',
'Literacy_Language Math_Science',
'Literacy_Language SpecialNeeds',
'Literacy_Language',
'Literacy_Language Math_Science',
'Literacy_Language Music_TheArts',
'Music_TheArts',
'Music_TheArts',
'Math_Science SpecialNeeds',
'History_Civics Literacy_Language',
'History_Civics',
'Literacy_Language Math_Science',
'Literacy_Language Math_Science',
'Health_Sports Literacy_Language',
'AppliedLearning Music_TheArts',
'Health_Sports',
'Math_Science',
'Health_Sports SpecialNeeds',
'Math_Science SpecialNeeds',
'Health_Sports',
'Literacy_Language SpecialNeeds',
'Literacy_Language',
'Literacy_Language',
'Literacy_Language Math_Science',
'History_Civics Music_TheArts',
'Health_Sports',
'Math_Science',
'Literacy_Language',
'Literacy_Language Math_Science',
'Literacy_Language',
'Literacy_Language',
'Literacy_Language SpecialNeeds',
'Literacy_Language SpecialNeeds',
'Literacy_Language',
'Math_Science SpecialNeeds',
'Health_Sports',
'Health_Sports',
'Math_Science',
'Literacy_Language',
'Math_Science',
'Literacy_Language',
'Math_Science',
'Literacy_Language',
'Literacy_Language Math_Science',
'Literacy_Language Math_Science',
'History_Civics',
'Literacy_Language',
'Health_Sports',
'Literacy_Language',
```

```
'AppliedLearning Music_TheArts',
'Math_Science',
'Literacy_Language',
'Health_Sports',
'AppliedLearning',
'Literacy_Language',
'Health_Sports',
'Literacy_Language',
'Math_Science',
'Math_Science History_Civics',
'Health_Sports SpecialNeeds',
'Literacy_Language Math_Science',
'Math_Science',
'AppliedLearning',
'Literacy_Language',
'Literacy_Language',
'Math_Science',
'SpecialNeeds',
'Math_Science',
'Literacy_Language SpecialNeeds',
'Math_Science',
'Literacy_Language Math_Science',
'Literacy_Language Math_Science',
'Literacy_Language',
'Literacy_Language',
'Literacy_Language',
'Music_TheArts',
'Literacy_Language SpecialNeeds',
'SpecialNeeds',
'Health_Sports',
'SpecialNeeds',
'Literacy_Language Math_Science',
'Health_Sports',
'Math_Science',
'Math_Science',
'Math_Science',
'Math_Science',
'Health_Sports',
'Literacy_Language History_Civics',
'Literacy_Language',
'Literacy_Language',
'Math_Science',
'AppliedLearning',
'Warmth Care_Hunger',
'Literacy_Language SpecialNeeds',
'Literacy_Language',
'Health_Sports',
'Literacy_Language Math_Science',
'Literacy_Language',
'Math_Science SpecialNeeds',
'Health_Sports',
'Literacy_Language',
'Math_Science SpecialNeeds',
'AppliedLearning SpecialNeeds',
'Literacy_Language Math_Science',
'Math_Science',
'Health_Sports',
'Health_Sports',
'Health_Sports',
'Warmth Care_Hunger',
'Literacy_Language Math_Science',
'Math_Science',
'Music_TheArts',
'Health_Sports',
'Literacy_Language',
'Literacy_Language Math_Science',
'Literacy_Language Math_Science',
'Literacy_Language',
'Math_Science',
'AppliedLearning Literacy_Language',
'Math_Science',
'Music_TheArts',
'History_Civics Math_Science',
'AppliedLearning',
'Health_Sports',
'Math_Science',
'AppliedLearning Literacy_Language',
```

```
'Math_Science',
'Literacy_Language History_Civics',
'Literacy_Language SpecialNeeds',
'Math_Science',
'Literacy_Language',
'Math_Science',
'Literacy_Language Music_TheArts',
'Math_Science',
'Music_TheArts',
'Literacy_Language',
'Math_Science',
'AppliedLearning',
'History_Civics Music_TheArts',
'Literacy_Language',
'Literacy_Language SpecialNeeds',
'Literacy_Language Math_Science',
'Math_Science',
'History_Civics Literacy_Language',
'Health_Sports AppliedLearning',
'Literacy_Language Math_Science',
'Literacy_Language',
'Literacy_Language',
'Math_Science',
'Math_Science',
'Literacy_Language',
'Math_Science Literacy_Language',
'Health_Sports',
'Health_Sports',
'Math_Science',
'AppliedLearning SpecialNeeds',
'Literacy_Language Math_Science',
'Health_Sports',
'Math_Science',
'Health_Sports',
'Literacy_Language',
'Math_Science Literacy_Language',
'Literacy_Language Math_Science',
'Math_Science Music_TheArts',
'Literacy_Language Math_Science',
'Literacy_Language Math_Science',
'Literacy_Language Math_Science',
'Literacy_Language',
'Math_Science AppliedLearning',
'Math_Science',
'Literacy_Language',
'Literacy_Language Math_Science',
'Literacy_Language',
'Math_Science',
'Literacy_Language Math_Science',
'AppliedLearning',
'Literacy_Language',
'Literacy_Language',
'SpecialNeeds',
'Literacy_Language',
'AppliedLearning Math_Science',
'Math_Science',
'Math_Science',
'Math_Science',
'Literacy_Language',
'Literacy_Language',
'AppliedLearning',
'Math_Science',
'Literacy_Language',
'Math_Science',
'Math_Science History_Civics',
'Math_Science',
'Literacy_Language Math_Science',
'Literacy_Language',
'Literacy_Language',
'Literacy_Language',
'Health_Sports',
'Math_Science Music_TheArts',
'Literacy_Language',
'Health_Sports',
'History_Civics Literacy_Language',
'Literacy_Language Math_Science',
'AppliedLearning Literacy_Language',
```

```
'Literacy_Language Music_TheArts',
'Literacy_Language',
'AppliedLearning Music_TheArts',
'Literacy_Language SpecialNeeds',
'Math_Science',
'Music_TheArts AppliedLearning',
'AppliedLearning',
'Literacy_Language',
'Math_Science',
'SpecialNeeds',
'Literacy_Language',
'Health_Sports',
'Music_TheArts',
'Health_Sports SpecialNeeds',
'Music_TheArts',
'Music_TheArts',
'Literacy_Language',
'Health_Sports',
'Math_Science AppliedLearning',
'Warmth Care_Hunger',
'AppliedLearning',
'Warmth Care_Hunger',
'Literacy_Language',
'Math_Science',
'Literacy_Language',
'AppliedLearning Literacy_Language',
'Literacy_Language',
'Math_Science Literacy_Language',
'Literacy_Language Math_Science',
'Literacy_Language Math_Science',
'Health_Sports',
'Literacy_Language',
'AppliedLearning',
'Music_TheArts',
'Literacy_Language',
'Literacy_Language',
'Literacy_Language Math_Science',
'History_Civics Literacy_Language',
'Math_Science',
'Literacy_Language',
'Literacy_Language SpecialNeeds',
'Music_TheArts',
'Literacy_Language Math_Science',
'Music_TheArts',
'Literacy_Language Math_Science',
'Music_TheArts',
'Math_Science',
'Literacy_Language SpecialNeeds',
'Music_TheArts',
'Literacy_Language',
'Literacy_Language SpecialNeeds',
'AppliedLearning Literacy_Language',
'Health_Sports',
'Music_TheArts',
'Literacy_Language',
'Math_Science',
'Literacy_Language Math_Science',
'Music_TheArts',
'Music_TheArts',
'Health_Sports Music_TheArts',
'Math_Science',
'Math_Science Literacy_Language',
'AppliedLearning Music_TheArts',
'Warmth Care_Hunger',
'Math_Science Health_Sports',
'Math_Science Health_Sports',
'Literacy_Language',
'Health_Sports SpecialNeeds',
'Health_Sports',
'Music_TheArts',
'Literacy_Language Math_Science',
'Math_Science',
'Literacy_Language Music_TheArts',
'SpecialNeeds',
'Health_Sports',
'Math_Science',
'Math Science',
```

```
'Math_Science History_Civics',
'Math_Science',
'Math_Science SpecialNeeds',
'Literacy_Language Math_Science',
'Math_Science',
'Health_Sports',
'AppliedLearning',
'Literacy_Language',
'Math_Science',
'SpecialNeeds Health_Sports',
'Literacy_Language Math_Science',
'Literacy_Language',
'Literacy_Language Math_Science',
'Music_TheArts',
'AppliedLearning',
'Literacy_Language',
'Math_Science',
'Literacy_Language',
'History_Civics Music_TheArts',
'Health_Sports SpecialNeeds',
'Literacy_Language Math_Science',
'Health_Sports',
'Health_Sports',
'AppliedLearning Health_Sports',
'Math_Science',
'History_Civics',
'Music_TheArts',
'Math_Science',
'Literacy_Language Math_Science',
'Math_Science SpecialNeeds',
'Literacy_Language',
'Literacy_Language Math_Science',
'Literacy_Language Math_Science',
'Literacy_Language Math_Science',
'Literacy_Language SpecialNeeds',
'Warmth Care_Hunger',
'Health_Sports Literacy_Language',
'Music_TheArts',
'Math_Science',
'Literacy_Language Math_Science',
'Literacy_Language Math_Science',
'Literacy_Language Math_Science',
'Math_Science SpecialNeeds',
'Literacy_Language Math_Science',
'Literacy_Language Math_Science',
'Literacy_Language SpecialNeeds',
'Math_Science Music_TheArts',
'AppliedLearning SpecialNeeds',
'Math_Science',
'Health_Sports',
'Literacy_Language',
'Health_Sports',
'Literacy_Language Math_Science',
'Literacy_Language Math_Science',
'Literacy_Language Math_Science',
'AppliedLearning',
'Math_Science',
'Literacy_Language',
'Literacy_Language',
'Math_Science',
'Math_Science',
'Literacy_Language Math_Science',
'AppliedLearning Music_TheArts',
'Music_TheArts',
'Math_Science AppliedLearning',
'AppliedLearning SpecialNeeds',
'Literacy_Language',
'Literacy_Language Math_Science',
'Health_Sports',
'AppliedLearning',
'Math_Science SpecialNeeds',
'Math_Science',
'Literacy_Language Math_Science',
'Literacy_Language SpecialNeeds',
'Literacy_Language Math_Science',
'Literacy_Language',
'Literacy_Language',
```

```
'Literacy_Language',
'Literacy_Language',
'Math_Science',
'Math_Science Literacy_Language',
'Literacy_Language',
'Literacy_Language Math_Science',
'Literacy_Language SpecialNeeds',
'Math_Science Literacy_Language',
'Literacy_Language',
'Literacy_Language Math_Science',
'Literacy_Language Math_Science',
'Math_Science',
'Literacy_Language SpecialNeeds',
'Literacy_Language Math_Science',
'Math_Science Music_TheArts',
'Health_Sports',
'AppliedLearning',
'Health_Sports',
'SpecialNeeds',
'Music_TheArts',
'Literacy_Language',
'Music_TheArts',
'Math_Science',
'Literacy_Language',
'Math_Science Music_TheArts',
'Music_TheArts',
'Literacy_Language Math_Science',
'Math_Science',
'Math_Science SpecialNeeds',
'Literacy_Language',
'Health_Sports SpecialNeeds',
'Math_Science',
'Math_Science',
'Literacy_Language',
'Literacy_Language',
'Music_TheArts',
'Health_Sports Math_Science',
'Health_Sports',
'AppliedLearning SpecialNeeds',
'Literacy_Language Math_Science',
'AppliedLearning SpecialNeeds',
'Warmth Care_Hunger',
'Health_Sports',
'Literacy_Language',
'History_Civics Literacy_Language',
'Literacy_Language Math_Science',
'Literacy_Language Music_TheArts',
'Literacy_Language',
'Literacy_Language Math_Science',
'Literacy_Language Math_Science',
'Literacy_Language AppliedLearning',
'Literacy_Language Math_Science',
'Literacy_Language Math_Science',
'Literacy_Language',
'Literacy_Language SpecialNeeds',
'AppliedLearning Math_Science',
'Math_Science Literacy_Language',
'Literacy_Language',
'AppliedLearning Literacy_Language',
'Literacy_Language',
'Literacy_Language',
'Math_Science',
'Math_Science Literacy_Language',
'Literacy_Language AppliedLearning',
'Literacy_Language Math_Science',
'Warmth Care_Hunger',
'Literacy_Language',
'AppliedLearning Music_TheArts',
'Literacy_Language',
'Math_Science',
'Literacy_Language Math_Science',
'Literacy_Language',
'Math_Science',
'History_Civics Literacy_Language',
'Music_TheArts',
'Music_TheArts',
'Literacy_Language',
```

```
'Literacy_Language',
'AppliedLearning SpecialNeeds',
'AppliedLearning',
'History_Civics',
'Math_Science',
'Math_Science Music_TheArts',
'Math_Science Music_TheArts',
'Literacy_Language',
'Math_Science',
'Math_Science',
'Literacy_Language',
'Literacy_Language Music_TheArts',
'Math_Science',
'Health_Sports',
'History_Civics Math_Science',
'Health_Sports',
'Literacy_Language History_Civics',
'Literacy_Language',
'History_Civics Literacy_Language',
'Literacy_Language',
'Math_Science',
'Math_Science SpecialNeeds',
'Literacy_Language Math_Science',
'Literacy_Language',
'Literacy_Language',
'Literacy_Language AppliedLearning',
'Health_Sports',
'Math_Science',
'History_Civics Literacy_Language',
'SpecialNeeds Music_TheArts',
'Literacy_Language',
'Literacy_Language Math_Science',
'Health_Sports',
'Literacy_Language',
'Health_Sports Literacy_Language',
'AppliedLearning Literacy_Language',
'AppliedLearning Math_Science',
'Literacy_Language SpecialNeeds',
'Music_TheArts Warmth Care_Hunger',
'Math_Science',
'Literacy_Language Math_Science',
'Literacy_Language SpecialNeeds',
'Math_Science Literacy_Language',
'Music_TheArts',
'AppliedLearning Literacy_Language',
'Math_Science',
'Literacy_Language Math_Science',
'Math_Science AppliedLearning',
'Math_Science AppliedLearning',
'Literacy_Language Math_Science',
'Literacy_Language',
'Health_Sports',
'Literacy_Language',
'Health_Sports Literacy_Language',
'Health_Sports Literacy_Language',
'Health_Sports Math_Science',
'Literacy_Language',
'Music_TheArts',
'Math_Science',
'Math_Science',
'Health_Sports',
'Literacy_Language AppliedLearning',
'Health_Sports',
'Literacy_Language',
'AppliedLearning SpecialNeeds',
'Math_Science',
'Literacy_Language Math_Science',
'Math_Science',
'Literacy_Language',
'Health_Sports',
'Literacy_Language',
'Literacy_Language Math_Science',
'Literacy_Language Math_Science',
'Literacy_Language',
'Literacy_Language',
'Literacy_Language Math_Science',
'AppliedLearning',
'Math_Science',
```

```
'Math_Science',
'AppliedLearning',
'Literacy_Language',
'Literacy_Language Math_Science',
'Health_Sports',
'Literacy_Language',
'Literacy_Language',
'Literacy_Language',
'Literacy_Language Math_Science',
'Literacy_Language',
'Music_TheArts',
'Literacy_Language',
'Warmth Care_Hunger',
'Literacy_Language SpecialNeeds',
'Music_TheArts',
'Math_Science',
'Literacy_Language',
'Literacy_Language',
'Literacy_Language',
'Literacy_Language',
'Literacy_Language Math_Science',
'Literacy_Language Math_Science',
'Literacy_Language',
'Literacy_Language Math_Science',
'Math_Science',
'Literacy_Language',
'AppliedLearning',
'Literacy_Language',
'Math_Science',
'Literacy_Language Math_Science',
'Health_Sports',
'Math_Science',
'Literacy_Language Math_Science',
'AppliedLearning Literacy_Language',
'Math_Science',
'Math_Science',
'Math_Science',
'Literacy_Language',
'AppliedLearning',
'Health_Sports',
'Math_Science',
'Math_Science',
'Math_Science History_Civics',
'Literacy_Language',
'Math_Science',
'Health_Sports',
'SpecialNeeds',
'Math_Science',
'History_Civics Literacy_Language',
'Health_Sports SpecialNeeds',
'History_Civics',
'Math_Science',
'Math_Science',
'AppliedLearning',
'Literacy_Language Math_Science',
'Literacy_Language',
'Math_Science',
'Math_Science SpecialNeeds',
'Literacy_Language',
'Math_Science',
'AppliedLearning',
'Math_Science Literacy_Language',
'Literacy_Language Math_Science',
'Health_Sports',
'Health_Sports',
'Literacy_Language SpecialNeeds',
'Health_Sports',
'Health_Sports',
'Health_Sports Literacy_Language',
'Literacy_Language',
'SpecialNeeds',
'Math_Science Music_TheArts',
'SpecialNeeds',
'AppliedLearning Math_Science',
'Literacy_Language SpecialNeeds',
'Literacy_Language SpecialNeeds',
'Literacy_Language Math_Science',
```

```
 'Literacy_Language Math_Science',
 'Literacy_Language SpecialNeeds',
 ...]
```

In [149]:

```
project_data['clean_categories'] = cat_list
project_data.head()
```

Out[149]:

| | Unnamed: 0 | id | teacher_id | teacher_prefix | school_state | project_submitted_datetime | pro |
|---|---|---|---|---|---|---|---|
| 0 | 160221 | p253737 | c90749f5d961ff158d4b4d1e7dc665fc | Mrs. | IN | 2016-12-05 13:43:57 | Gra |
| 1 | 140945 | p258326 | 897464ce9ddc600bced1151f324dd63a | Mr. | FL | 2016-10-25 09:22:10 | Gra |
| 2 | 21895 | p182444 | 3465aaf82da834c0582ebd0ef8040ca0 | Ms. | AZ | 2016-08-31 12:03:56 | Gra |
| 3 | 45 | p246581 | f3cb9bffbba169bef1a77b243e620b60 | Mrs. | KY | 2016-10-06 21:16:17 | Gra |
| 4 | 172407 | p104768 | be1f7507a41f8479dc06f047086a39ec | Mrs. | TX | 2016-07-11 01:10:09 | Gra |

In [150]:

```
project_data.drop(labels='project_subject_categories',axis = 1,  inplace=True)
```

In [151]:

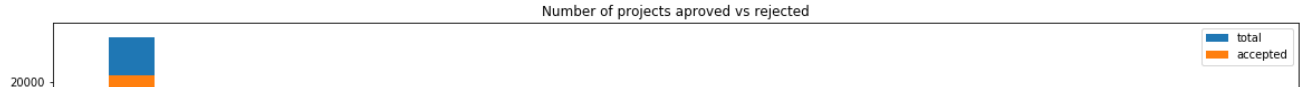
```
project_data.head(1)
```

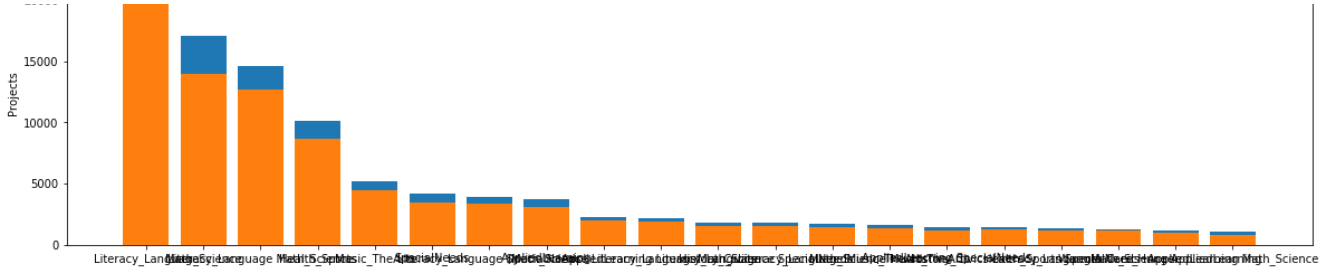
Out[151]:

| | Unnamed: 0 | id | teacher_id | teacher_prefix | school_state | project_submitted_datetime | proje |
|---|---|---|---|---|---|---|---|
| 0 | 160221 | p253737 | c90749f5d961ff158d4b4d1e7dc665fc | Mrs. | IN | 2016-12-05 13:43:57 | Grade |

In [152]:

```
univariate_barplots(project_data, 'clean_categories', 'project_is_approved', top=20)
```






```
                 clean_categories  project_is_approved  total       Avg
24                Literacy_Language                20520  23655  0.867470
32                    Math_Science                13991  17072  0.819529
28  Literacy_Language Math_Science                12725  14636  0.869432
8                    Health_Sports                 8640  10177  0.848973
40                   Music_TheArts                 4429   5180  0.855019
==================================================
                    clean_categories  project_is_approved  total       Avg
19  History_Civics Literacy_Language                 1271   1421  0.894441
14        Health_Sports SpecialNeeds                 1215   1391  0.873472
50                Warmth Care_Hunger                 1212   1309  0.925898
33      Math_Science AppliedLearning                 1019   1220  0.835246
4       AppliedLearning Math_Science                  855   1052  0.812738
```

In [153]:

```
#now we can see how many unique different project categories
#https://stackoverflow.com/a/22898595/4084039
from collections import Counter
my_counter = Counter()

for i in project_data['clean_categories'].values:
    my_counter.update(i.split())
```

In [154]:

```
my_counter
```

Out[154]:

```
Counter({'AppliedLearning': 12135,
         'Care_Hunger': 1388,
         'Health_Sports': 14223,
         'History_Civics': 5914,
         'Literacy_Language': 52239,
         'Math_Science': 41421,
         'Music_TheArts': 10293,
         'SpecialNeeds': 13642,
         'Warmth': 1388})
```

In [155]:

```
#converting it into python dict into dataframe to plot the bar plot
#https://stackoverflow.com/questions/18837262/convert-python-dict-into-a-dataframe
xxx_1 = pd.DataFrame.from_dict(my_counter.items())
xxx_1
```

Out[155]:

| | 0 | 1 |
|---|---|---|
| **0** | Literacy_Language | 52239 |
| **1** | History_Civics | 5914 |
| **2** | Health_Sports | 14223 |
| **3** | Math_Science | 41421 |
| **4** | SpecialNeeds | 13642 |
| **5** | AppliedLearning | 12135 |

| | | |
|---|---|---|
| 6 | Music_TheArts | 10293 |
| 7 | Warmth | 1388 |
| 8 | Care_Hunger | 1388 |

In [156]:

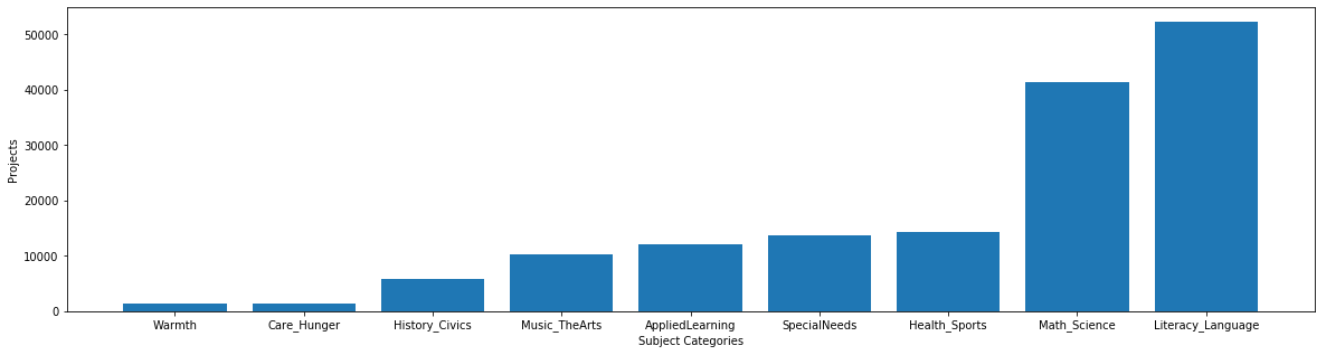
```
xxx_1.columns = ['unique_subject_categories', 'count']

#https://www.geeksforgeeks.org/python-pandas-dataframe-sort_values-set-1/
xxx_1.sort_values(by='count', axis=0, ascending=True, inplace=True)
xxx_1
```

Out[156]:

| | unique_subject_categories | count |
|---|---|---|
| 7 | Warmth | 1388 |
| 8 | Care_Hunger | 1388 |
| 1 | History_Civics | 5914 |
| 6 | Music_TheArts | 10293 |
| 5 | AppliedLearning | 12135 |
| 4 | SpecialNeeds | 13642 |
| 2 | Health_Sports | 14223 |
| 3 | Math_Science | 41421 |
| 0 | Literacy_Language | 52239 |

In [157]:

```
#plot it
plt.figure(figsize=(20,5))
plt.bar(xxx_1['unique_subject_categories'], xxx_1['count'])
plt.xlabel('Subject Categories')
plt.ylabel('Projects')
plt.show()
```




### 1.2.5 Univariate Analysis: project_subject_subcategories

In [158]:

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

Out[158]:

```
0                       ESL, Literacy
1    Civics & Government, Team Sports
2      Health & Wellness, Team Sports
3                Literacy, Mathematics
4                         Mathematics
Name: project_subject_subcategories, dtype: object
```

In [159]:

```
#Working on product_subject sub categories

sub_cat_list = []

for i in list(project_data['project_subject_subcategories'].values):

    temp = ''
    for j in i.split(','):
        if 'The' in j.split():
            j = j.replace('The', '')

        j = j.replace(' ','')
        temp += j.strip()+' '
        temp = temp.replace('&', '_')

    sub_cat_list.append(temp.strip())
```

In [160]:

```
sub_cat_list
```

Out[160]:

```
['ESL Literacy',
 'Civics_Government TeamSports',
 'Health_Wellness TeamSports',
 'Literacy Mathematics',
 'Mathematics',
 'Literature_Writing SpecialNeeds',
 'Literacy SpecialNeeds',
 'Mathematics',
 'Health_Wellness',
 'Literacy Literature_Writing',
 'Literacy',
 'Literacy ParentInvolvement',
 'EnvironmentalScience Health_LifeScience',
 'SpecialNeeds',
 'Literacy',
 'Health_Wellness',
 'Literacy SpecialNeeds',
 'AppliedSciences Literature_Writing',
 'EarlyDevelopment',
 'Health_Wellness',
 'Literacy',
 'Health_LifeScience SpecialNeeds',
 'Literacy',
 'Music',
 'AppliedSciences Mathematics',
 'Mathematics',
 'Literacy Mathematics',
 'ForeignLanguages Mathematics',
 'Literacy SpecialNeeds',
 'Literacy Other',
 'Literacy',
 'SpecialNeeds',
 'Health_LifeScience Literacy',
 'Economics FinancialLiteracy',
 'Literature_Writing',
 'TeamSports',
 'Literature_Writing Mathematics',
 'Health_Wellness Literacy',
 'Gym_Fitness Health_Wellness',
 'Literacy Literature_Writing',
 'Literacy Literature_Writing',
 'Literature_Writing',
 'Literacy Literature_Writing',
 'Literacy VisualArts',
 'Mathematics',
 'Literacy Literature_Writing',
 'Literature_Writing',
 'Warmth Care_Hunger',
 'Literature_Writing Mathematics',
```

```
'Gym_Fitness Health_Wellness',
'Gym_Fitness Health_Wellness',
'Literacy',
'Mathematics SocialSciences',
'Literacy',
'Gym_Fitness Health_Wellness',
'AppliedSciences Mathematics',
'SpecialNeeds',
'Literature_Writing Mathematics',
'Literature_Writing',
'Literacy Literature_Writing',
'Health_Wellness',
'AppliedSciences Mathematics',
'Literature_Writing',
'VisualArts',
'VisualArts',
'SpecialNeeds',
'AppliedSciences Mathematics',
'Literacy',
'AppliedSciences Mathematics',
'College_CareerPrep Literature_Writing',
'EnvironmentalScience',
'AppliedSciences Health_LifeScience',
'ESL Literature_Writing',
'College_CareerPrep',
'Health_LifeScience Mathematics',
'Music',
'Literature_Writing Mathematics',
'CharacterEducation SpecialNeeds',
'Mathematics',
'Music',
'Literacy Literature_Writing',
'EnvironmentalScience Mathematics',
'Literacy',
'Literacy SocialSciences',
'AppliedSciences Mathematics',
'Literature_Writing',
'Health_Wellness TeamSports',
'Literacy SpecialNeeds',
'SpecialNeeds',
'Literacy',
'Literature_Writing VisualArts',
'Health_Wellness',
'CharacterEducation Health_Wellness',
'Music PerformingArts',
'Literacy',
'Mathematics',
'Literacy',
'Literacy',
'CommunityService',
'Mathematics',
'Literature_Writing Mathematics',
'Gym_Fitness TeamSports',
'Literacy Mathematics',
'Music PerformingArts',
'Health_Wellness',
'VisualArts',
'Literacy',
'Literature_Writing',
'SpecialNeeds',
'SpecialNeeds',
'Literature_Writing Mathematics',
'Literature_Writing',
'Literature_Writing',
'Literacy Mathematics',
'Literature_Writing',
'Literature_Writing VisualArts',
'Literature_Writing SpecialNeeds',
'History_Geography Literature_Writing',
'Health_LifeScience',
'AppliedSciences Mathematics',
'Literature_Writing Mathematics',
'Literature_Writing',
'Gym_Fitness SpecialNeeds',
'Literacy SpecialNeeds',
'SpecialNeeds',
'Mathematics',
```

```
'EarlyDevelopment Literature_Writing',
'Music',
'PerformingArts',
'Literature_Writing SpecialNeeds',
'Literacy Literature_Writing',
'Literacy',
'ESL Literacy',
'Health_Wellness',
'Gym_Fitness SpecialNeeds',
'VisualArts',
'Literacy SpecialNeeds',
'Literacy Literature_Writing',
'EarlyDevelopment Mathematics',
'AppliedSciences',
'Literacy SpecialNeeds',
'Literacy Literature_Writing',
'Health_Wellness',
'EnvironmentalScience Mathematics',
'Literacy Literature_Writing',
'CommunityService',
'Literature_Writing Mathematics',
'Literacy Mathematics',
'College_CareerPrep Literacy',
'EnvironmentalScience',
'Health_Wellness',
'Mathematics',
'Mathematics',
'SpecialNeeds',
'Health_LifeScience History_Geography',
'Literature_Writing Mathematics',
'Gym_Fitness Health_Wellness',
'EnvironmentalScience Health_LifeScience',
'EnvironmentalScience Mathematics',
'AppliedSciences VisualArts',
'Literacy',
'History_Geography Literature_Writing',
'Health_Wellness NutritionEducation',
'Literacy PerformingArts',
'AppliedSciences Mathematics',
'Gym_Fitness Health_Wellness',
'Music PerformingArts',
'Health_Wellness',
'AppliedSciences Mathematics',
'Mathematics SpecialNeeds',
'Literature_Writing SpecialNeeds',
'Literacy Mathematics',
'Literacy Literature_Writing',
'College_CareerPrep',
'Literature_Writing Mathematics',
'Literacy Literature_Writing',
'AppliedSciences VisualArts',
'Literacy Literature_Writing',
'History_Geography Literacy',
'Literacy Mathematics',
'Literacy Mathematics',
'AppliedSciences',
'Literature_Writing Mathematics',
'Mathematics',
'Mathematics',
'Mathematics',
'Literature_Writing Mathematics',
'PerformingArts VisualArts',
'EnvironmentalScience Health_LifeScience',
'Literacy Mathematics',
'Mathematics',
'SpecialNeeds',
'Literacy Mathematics',
'Gym_Fitness Health_Wellness',
'Literature_Writing Mathematics',
'Mathematics',
'Gym_Fitness ParentInvolvement',
'Health_LifeScience',
'Literature_Writing',
'Literacy Literature_Writing',
'EarlyDevelopment Other',
'Music',
'AppliedSciences EnvironmentalScience',
```

```
'Literacy SpecialNeeds',
'Literature_Writing Mathematics',
'ForeignLanguages Literacy',
'AppliedSciences Mathematics',
'Health_LifeScience',
'PerformingArts',
'Mathematics',
'Literacy Literature_Writing',
'Music',
'NutritionEducation',
'EarlyDevelopment Literature_Writing',
'Mathematics',
'Literacy',
'AppliedSciences',
'Literacy',
'Music',
'AppliedSciences EnvironmentalScience',
'College_CareerPrep',
'Literacy Mathematics',
'Health_Wellness NutritionEducation',
'EarlyDevelopment Mathematics',
'AppliedSciences',
'Literature_Writing SocialSciences',
'EarlyDevelopment SpecialNeeds',
'ESL Literacy',
'Music PerformingArts',
'History_Geography Literacy',
'Literature_Writing Mathematics',
'Literacy SpecialNeeds',
'Literature_Writing Mathematics',
'EnvironmentalScience Mathematics',
'SpecialNeeds',
'Literacy SpecialNeeds',
'EarlyDevelopment NutritionEducation',
'AppliedSciences',
'Literature_Writing SocialSciences',
'Literature_Writing Mathematics',
'Literacy',
'College_CareerPrep SpecialNeeds',
'SocialSciences VisualArts',
'College_CareerPrep',
'EarlyDevelopment',
'Other',
'Literacy Mathematics',
'AppliedSciences CharacterEducation',
'Warmth Care_Hunger',
'Mathematics',
'Literature_Writing SpecialNeeds',
'Mathematics',
'AppliedSciences Mathematics',
'Literacy Literature_Writing',
'AppliedSciences',
'AppliedSciences Mathematics',
'Literature_Writing SpecialNeeds',
'Literature_Writing Mathematics',
'Health_Wellness SpecialNeeds',
'Literacy',
'Literacy Literature_Writing',
'Literature_Writing',
'Literacy',
'History_Geography Literacy',
'Literature_Writing Mathematics',
'Gym_Fitness Health_Wellness',
'Literacy Literature_Writing',
'Literature_Writing Mathematics',
'Music SocialSciences',
'ESL',
'Health_LifeScience',
'Literacy SocialSciences',
'Literacy Mathematics',
'AppliedSciences Health_LifeScience',
'EnvironmentalScience Health_LifeScience',
'Health_Wellness NutritionEducation',
'Literature_Writing Mathematics',
'Gym_Fitness Health_Wellness',
'Literacy',
'CharacterEducation CommunityService',
```

'Mathematics',
'Health_Wellness SpecialNeeds',
'ESL Literacy',
'VisualArts',
'AppliedSciences SpecialNeeds',
'Literacy Mathematics',
'Literacy SpecialNeeds',
'Warmth Care_Hunger',
'Gym_Fitness',
'College_CareerPrep SpecialNeeds',
'Literature_Writing',
'SocialSciences',
'Literacy Mathematics',
'Civics_Government SocialSciences',
'AppliedSciences',
'Music PerformingArts',
'Music',
'Health_Wellness',
'CharacterEducation Literacy',
'Mathematics',
'Health_LifeScience',
'AppliedSciences',
'EnvironmentalScience Health_LifeScience',
'Literature_Writing Mathematics',
'ESL VisualArts',
'AppliedSciences Literacy',
'Gym_Fitness Health_Wellness',
'Health_Wellness',
'Mathematics SpecialNeeds',
'AppliedSciences',
'TeamSports',
'Literacy SpecialNeeds',
'Literature_Writing Other',
'Health_Wellness',
'College_CareerPrep EarlyDevelopment',
'Literature_Writing',
'Gym_Fitness Health_Wellness',
'College_CareerPrep Literature_Writing',
'Literature_Writing',
'AppliedSciences Mathematics',
'Literacy',
'Literacy Mathematics',
'Literacy Mathematics',
'EarlyDevelopment Literacy',
'Literacy Mathematics',
'Literacy Mathematics',
'AppliedSciences EarlyDevelopment',
'AppliedSciences EnvironmentalScience',
'History_Geography Literature_Writing',
'Health_Wellness',
'Health_Wellness',
'Literacy SocialSciences',
'Health_Wellness',
'AppliedSciences',
'Literacy Mathematics',
'PerformingArts',
'VisualArts',
'EarlyDevelopment Literacy',
'Literature_Writing',
'Health_Wellness',
'Literacy',
'AppliedSciences',
'Literature_Writing Mathematics',
'Literature_Writing',
'Mathematics',
'Literacy',
'CharacterEducation Literature_Writing',
'AppliedSciences Mathematics',
'CharacterEducation College_CareerPrep',
'Literacy Mathematics',
'Literacy',
'Literacy',
'Literature_Writing',
'Health_Wellness',
'Other',
'Mathematics',
'CommunityService Health_Wellness',

```
'Literature_Writing VisualArts',
'Health_Wellness',
'EnvironmentalScience',
'EnvironmentalScience',
'Literacy',
'Literacy Mathematics',
'Music PerformingArts',
'Literacy',
'Health_Wellness Other',
'EarlyDevelopment',
'Literature_Writing',
'Literacy Mathematics',
'Mathematics',
'Literature_Writing',
'Literature_Writing',
'AppliedSciences College_CareerPrep',
'Literacy Mathematics',
'Gym_Fitness TeamSports',
'Literature_Writing VisualArts',
'Mathematics SpecialNeeds',
'Literature_Writing VisualArts',
'Literacy',
'Literacy',
'Literacy',
'Health_Wellness',
'Health_LifeScience Literacy',
'Literature_Writing Mathematics',
'Literacy',
'EarlyDevelopment Mathematics',
'Mathematics',
'CharacterEducation',
'College_CareerPrep Mathematics',
'Music',
'Literacy',
'ESL Literacy',
'Literature_Writing Mathematics',
'Health_Wellness',
'CommunityService EnvironmentalScience',
'Literacy SpecialNeeds',
'EarlyDevelopment Literacy',
'Mathematics SpecialNeeds',
'Literacy',
'Literacy Literature_Writing',
'Mathematics',
'Literature_Writing',
'Literacy Mathematics',
'History_Geography',
'Mathematics',
'Other SpecialNeeds',
'Literacy',
'Music PerformingArts',
'Health_Wellness',
'AppliedSciences',
'Literature_Writing Mathematics',
'Literacy Mathematics',
'Literature_Writing Mathematics',
'Literacy SpecialNeeds',
'ESL Literacy',
'Literature_Writing Mathematics',
'ESL Music',
'VisualArts',
'Music PerformingArts',
'Mathematics SpecialNeeds',
'History_Geography Literature_Writing',
'History_Geography',
'ESL Mathematics',
'Literature_Writing Mathematics',
'Health_Wellness Literature_Writing',
'EarlyDevelopment Music',
'Health_Wellness',
'Mathematics',
'Health_Wellness SpecialNeeds',
'Mathematics SpecialNeeds',
'Gym_Fitness TeamSports',
'Literacy SpecialNeeds',
'ESL Literacy',
'Literacy',
```

'Literacy',
'Literature_Writing Mathematics',
'SocialSciences VisualArts',
'Health_Wellness',
'Mathematics',
'Literacy Literature_Writing',
'Literacy Mathematics',
'Literacy',
'Literacy',
'Literacy SpecialNeeds',
'ESL SpecialNeeds',
'Literature_Writing',
'Mathematics SpecialNeeds',
'Health_Wellness',
'Health_Wellness',
'Mathematics',
'Literature_Writing',
'AppliedSciences Health_LifeScience',
'Literacy',
'AppliedSciences EnvironmentalScience',
'Literacy Literature_Writing',
'Literature_Writing Mathematics',
'Literature_Writing Mathematics',
'Civics_Government History_Geography',
'Literacy Literature_Writing',
'NutritionEducation',
'Literature_Writing',
'CharacterEducation VisualArts',
'Mathematics',
'Literacy',
'Gym_Fitness',
'College_CareerPrep Other',
'Literacy',
'Health_Wellness',
'Literacy',
'AppliedSciences EnvironmentalScience',
'AppliedSciences SocialSciences',
'Health_Wellness SpecialNeeds',
'Literature_Writing Mathematics',
'AppliedSciences',
'EarlyDevelopment',
'Literacy',
'Literacy',
'Mathematics',
'SpecialNeeds',
'Mathematics',
'Literacy SpecialNeeds',
'Mathematics',
'Literacy Mathematics',
'Literacy Mathematics',
'Literacy Literature_Writing',
'Literacy Literature_Writing',
'Literacy',
'VisualArts',
'Literacy SpecialNeeds',
'SpecialNeeds',
'Health_Wellness TeamSports',
'SpecialNeeds',
'Literacy Mathematics',
'Gym_Fitness',
'AppliedSciences Mathematics',
'AppliedSciences EnvironmentalScience',
'Mathematics',
'Health_LifeScience Mathematics',
'Gym_Fitness Health_Wellness',
'Literature_Writing SocialSciences',
'Literacy Literature_Writing',
'Literacy Literature_Writing',
'AppliedSciences',
'Other',
'Warmth Care_Hunger',
'Literacy SpecialNeeds',
'ESL Literature_Writing',
'Health_Wellness',
'Literacy Mathematics',
'Literacy Literature_Writing',
'Health_LifeScience SpecialNeeds',
'Gym_Fitness TeamSports',

```
'Gym_Fitness TeamSports',
'Literacy Literature_Writing',
'AppliedSciences SpecialNeeds',
'Other SpecialNeeds',
'Literacy Mathematics',
'Mathematics',
'Health_Wellness',
'Health_Wellness NutritionEducation',
'Health_Wellness',
'Warmth Care_Hunger',
'Literacy Mathematics',
'AppliedSciences Mathematics',
'Music',
'Gym_Fitness',
'Literacy',
'Literature_Writing Mathematics',
'Literacy Mathematics',
'Literature_Writing',
'AppliedSciences Mathematics',
'College_CareerPrep Literature_Writing',
'AppliedSciences Mathematics',
'VisualArts',
'History_Geography Mathematics',
'EarlyDevelopment',
'Health_Wellness NutritionEducation',
'AppliedSciences Mathematics',
'EarlyDevelopment Literacy',
'AppliedSciences Mathematics',
'Literacy SocialSciences',
'Literacy SpecialNeeds',
'Mathematics',
'Literacy',
'Mathematics',
'Literacy Music',
'EnvironmentalScience Mathematics',
'Music PerformingArts',
'Literacy',
'AppliedSciences',
'College_CareerPrep',
'History_Geography Music',
'Literacy Literature_Writing',
'Literacy SpecialNeeds',
'Literature_Writing Mathematics',
'AppliedSciences',
'Civics_Government Literacy',
'NutritionEducation Other',
'Literacy Mathematics',
'Literacy Literature_Writing',
'Literature_Writing',
'AppliedSciences EnvironmentalScience',
'AppliedSciences Mathematics',
'Literacy Literature_Writing',
'Health_LifeScience Literacy',
'Gym_Fitness Health_Wellness',
'Gym_Fitness Health_Wellness',
'AppliedSciences Mathematics',
'Other SpecialNeeds',
'Literature_Writing Mathematics',
'Gym_Fitness Health_Wellness',
'AppliedSciences Mathematics',
'Health_Wellness',
'ESL Literature_Writing',
'EnvironmentalScience Literature_Writing',
'Literature_Writing Mathematics',
'EnvironmentalScience VisualArts',
'Literature_Writing Mathematics',
'Literacy Mathematics',
'Literacy Mathematics',
'Literacy Literature_Writing',
'AppliedSciences EarlyDevelopment',
'Mathematics',
'Literacy',
'ESL Mathematics',
'Literature_Writing',
'AppliedSciences',
'Literature_Writing Mathematics',
'Other',
'Literacy'
```

```
'Literacy',
'ESL Literature_Writing',
'SpecialNeeds',
'Literacy',
'CharacterEducation Mathematics',
'AppliedSciences Mathematics',
'AppliedSciences',
'EnvironmentalScience Health_LifeScience',
'Literacy Literature_Writing',
'Literacy Literature_Writing',
'EarlyDevelopment',
'EnvironmentalScience Health_LifeScience',
'Literacy Literature_Writing',
'AppliedSciences Health_LifeScience',
'EnvironmentalScience History_Geography',
'Mathematics',
'Literature_Writing Mathematics',
'Literacy',
'Literacy',
'Literature_Writing',
'TeamSports',
'Health_LifeScience VisualArts',
'Literacy',
'Gym_Fitness TeamSports',
'Civics_Government Literacy',
'Literature_Writing Mathematics',
'EarlyDevelopment Literature_Writing',
'Literacy VisualArts',
'ESL Literacy',
'Other VisualArts',
'Literacy SpecialNeeds',
'Mathematics',
'Music ParentInvolvement',
'College_CareerPrep EarlyDevelopment',
'Literacy',
'Mathematics',
'SpecialNeeds',
'Literacy Literature_Writing',
'Gym_Fitness NutritionEducation',
'Music PerformingArts',
'Health_Wellness SpecialNeeds',
'Music',
'VisualArts',
'ESL Literacy',
'Gym_Fitness TeamSports',
'AppliedSciences ParentInvolvement',
'Warmth Care_Hunger',
'Other',
'Warmth Care_Hunger',
'Literacy Literature_Writing',
'AppliedSciences',
'Literacy Literature_Writing',
'EarlyDevelopment Literacy',
'Literacy',
'AppliedSciences Literature_Writing',
'Literacy Mathematics',
'Literacy Mathematics',
'Health_Wellness TeamSports',
'Literacy',
'College_CareerPrep Other',
'VisualArts',
'Literacy',
'Literacy Literature_Writing',
'Literature_Writing Mathematics',
'History_Geography Literature_Writing',
'AppliedSciences Mathematics',
'Literacy',
'Literature_Writing SpecialNeeds',
'Music',
'Literature_Writing Mathematics',
'Music',
'Literacy Mathematics',
'Music PerformingArts',
'AppliedSciences EnvironmentalScience',
'Literature_Writing SpecialNeeds',
'VisualArts',
'Literacy',
'Literature_Writing SpecialNeeds'
```

```
'Literature_Writing SpecialNeeds',
'CharacterEducation Literature_Writing',
'Health_Wellness',
'VisualArts',
'Literature_Writing',
'AppliedSciences Mathematics',
'Literacy Mathematics',
'Music',
'VisualArts',
'Health_Wellness VisualArts',
'AppliedSciences',
'EnvironmentalScience Literature_Writing',
'CharacterEducation VisualArts',
'Warmth Care_Hunger',
'EnvironmentalScience Health_Wellness',
'Health_LifeScience Health_Wellness',
'Literature_Writing',
'Health_Wellness SpecialNeeds',
'Health_Wellness',
'Music',
'Literature_Writing Mathematics',
'Mathematics',
'Literature_Writing VisualArts',
'SpecialNeeds',
'Health_Wellness',
'Mathematics',
'AppliedSciences',
'AppliedSciences History_Geography',
'AppliedSciences EnvironmentalScience',
'Mathematics SpecialNeeds',
'Literature_Writing Mathematics',
'AppliedSciences',
'Gym_Fitness Health_Wellness',
'Other',
'Literacy',
'AppliedSciences Mathematics',
'SpecialNeeds TeamSports',
'Literacy Mathematics',
'Literacy',
'Literature_Writing Mathematics',
'VisualArts',
'Other',
'Literacy',
'Mathematics',
'Literacy',
'History_Geography VisualArts',
'Health_Wellness SpecialNeeds',
'Literacy Mathematics',
'Gym_Fitness Health_Wellness',
'Gym_Fitness',
'CharacterEducation Health_Wellness',
'Mathematics',
'SocialSciences',
'Music',
'EnvironmentalScience Mathematics',
'Literacy Mathematics',
'Mathematics SpecialNeeds',
'Literacy Literature_Writing',
'Literacy Mathematics',
'Literacy Mathematics',
'Literacy Mathematics',
'Literacy SpecialNeeds',
'Warmth Care_Hunger',
'Health_Wellness Literacy',
'Music',
'AppliedSciences Mathematics',
'Literature_Writing Mathematics',
'Literacy Mathematics',
'Literacy Mathematics',
'Mathematics SpecialNeeds',
'Literature_Writing Mathematics',
'Literacy Mathematics',
'Literacy SpecialNeeds',
'Mathematics VisualArts',
'Other SpecialNeeds',
'Mathematics',
'Health_Wellness',
'Literacy',
```

'Literacy',
'NutritionEducation',
'Literacy Mathematics',
'Literacy Mathematics',
'Literature_Writing Mathematics',
'Other',
'EnvironmentalScience Mathematics',
'Literacy',
'ESL Literature_Writing',
'EnvironmentalScience',
'Mathematics',
'Literature_Writing Mathematics',
'EarlyDevelopment VisualArts',
'Music PerformingArts',
'AppliedSciences College_CareerPrep',
'ParentInvolvement SpecialNeeds',
'Literacy',
'Literature_Writing Mathematics',
'Gym_Fitness Health_Wellness',
'EarlyDevelopment',
'Mathematics SpecialNeeds',
'AppliedSciences',
'Literacy Mathematics',
'Literacy SpecialNeeds',
'Literature_Writing Mathematics',
'Literacy',
'Literacy',
'Literature_Writing',
'ESL Literacy',
'AppliedSciences Health_LifeScience',
'EnvironmentalScience Literature_Writing',
'Literature_Writing',
'Literacy Mathematics',
'Literature_Writing SpecialNeeds',
'AppliedSciences Literacy',
'Literacy',
'Literature_Writing Mathematics',
'Literacy Mathematics',
'AppliedSciences Mathematics',
'Literature_Writing SpecialNeeds',
'Literature_Writing Mathematics',
'Mathematics VisualArts',
'Health_Wellness NutritionEducation',
'CharacterEducation Other',
'Gym_Fitness Health_Wellness',
'SpecialNeeds',
'Music PerformingArts',
'Literature_Writing',
'Music PerformingArts',
'AppliedSciences Mathematics',
'Literacy Literature_Writing',
'Health_LifeScience VisualArts',
'Music PerformingArts',
'Literacy Mathematics',
'AppliedSciences Mathematics',
'Mathematics SpecialNeeds',
'ESL Literacy',
'Health_Wellness SpecialNeeds',
'EnvironmentalScience',
'AppliedSciences Mathematics',
'ESL Literature_Writing',
'Literacy Literature_Writing',
'VisualArts',
'Health_Wellness Mathematics',
'Health_Wellness NutritionEducation',
'College_CareerPrep SpecialNeeds',
'Literacy Mathematics',
'CharacterEducation SpecialNeeds',
'Warmth Care_Hunger',
'Gym_Fitness',
'Literature_Writing',
'History_Geography Literature_Writing',
'Literature_Writing Mathematics',
'Literature_Writing VisualArts',
'Literacy Literature_Writing',
'Literacy Mathematics',
'Literacy Mathematics',

```
'Literacy ParentInvolvement',
'Literature_Writing Mathematics',
'Literacy Mathematics',
'Literacy',
'Literature_Writing SpecialNeeds',
'EarlyDevelopment EnvironmentalScience',
'Health_LifeScience Literature_Writing',
'Literacy',
'CharacterEducation ESL',
'Literature_Writing',
'Literacy',
'Mathematics',
'AppliedSciences Literature_Writing',
'ESL EarlyDevelopment',
'Literacy Mathematics',
'Warmth Care_Hunger',
'Literacy',
'College_CareerPrep Music',
'Literature_Writing',
'AppliedSciences',
'Literacy Mathematics',
'Literacy Literature_Writing',
'Mathematics',
'Civics_Government Literacy',
'VisualArts',
'VisualArts',
'ESL',
'EarlyDevelopment SpecialNeeds',
'College_CareerPrep',
'History_Geography',
'Health_LifeScience',
'Mathematics VisualArts',
'AppliedSciences VisualArts',
'Literature_Writing',
'EnvironmentalScience',
'AppliedSciences EnvironmentalScience',
'Literacy',
'Literature_Writing VisualArts',
'Mathematics',
'Health_Wellness',
'History_Geography Mathematics',
'Gym_Fitness Health_Wellness',
'Literature_Writing SocialSciences',
'ESL Literacy',
'Civics_Government Literature_Writing',
'Literacy',
'AppliedSciences Mathematics',
'Mathematics SpecialNeeds',
'Literature_Writing Mathematics',
'Literacy',
'Literacy Literature_Writing',
'Literacy Other',
'Health_Wellness',
'AppliedSciences Mathematics',
'Civics_Government Literacy',
'SpecialNeeds VisualArts',
'ForeignLanguages',
'Literacy Mathematics',
'Gym_Fitness Health_Wellness',
'Literacy',
'Health_Wellness Literacy',
'EarlyDevelopment Literacy',
'EarlyDevelopment Mathematics',
'Literature_Writing SpecialNeeds',
'VisualArts Warmth Care_Hunger',
'AppliedSciences',
'Literature_Writing Mathematics',
'Literacy SpecialNeeds',
'EnvironmentalScience Literacy',
'PerformingArts',
'EarlyDevelopment Literacy',
'EnvironmentalScience',
'Literacy Mathematics',
'AppliedSciences EarlyDevelopment',
'AppliedSciences Extracurricular',
'Literacy Mathematics',
'Literacy',
```

```
'Health_Wellness',
'Literacy',
'Health_Wellness Literacy',
'Health_Wellness Literacy',
'Health_Wellness Mathematics',
'Literacy Literature_Writing',
'Music',
'Mathematics',
'AppliedSciences EnvironmentalScience',
'Health_Wellness',
'ESL EarlyDevelopment',
'Health_Wellness',
'Literacy',
'Other SpecialNeeds',
'AppliedSciences',
'Literacy Mathematics',
'EnvironmentalScience Mathematics',
'Literature_Writing',
'Gym_Fitness TeamSports',
'ESL Literacy',
'Literature_Writing Mathematics',
'Literature_Writing Mathematics',
'Literacy',
'Literature_Writing',
'Literacy Mathematics',
'College_CareerPrep Other',
'AppliedSciences EnvironmentalScience',
'EnvironmentalScience Health_LifeScience',
'EarlyDevelopment Other',
'Literacy',
'Literacy Mathematics',
'Gym_Fitness TeamSports',
'Literacy',
'Literacy',
'Literacy',
'Literacy Mathematics',
'ESL Literacy',
'VisualArts',
'Literacy',
'Warmth Care_Hunger',
'Literacy SpecialNeeds',
'Music PerformingArts',
'AppliedSciences',
'Literacy Literature_Writing',
'Literacy Literature_Writing',
'Literacy',
'Literacy',
'Literacy Mathematics',
'Literacy Mathematics',
'Literacy',
'Literacy Mathematics',
'Health_LifeScience',
'Literacy',
'Other',
'Literacy',
'EnvironmentalScience Mathematics',
'Literature_Writing Mathematics',
'TeamSports',
'Mathematics',
'Literature_Writing Mathematics',
'College_CareerPrep Literature_Writing',
'AppliedSciences Mathematics',
'AppliedSciences Health_LifeScience',
'Mathematics',
'Literacy',
'EarlyDevelopment Other',
'Health_Wellness',
'EnvironmentalScience Mathematics',
'EnvironmentalScience',
'AppliedSciences History_Geography',
'Literacy Literature_Writing',
'EnvironmentalScience Health_LifeScience',
'Gym_Fitness Health_Wellness',
'SpecialNeeds',
'AppliedSciences Mathematics',
'Civics_Government Literacy',
'Health_Wellness SpecialNeeds',
```

```
 'Economics FinancialLiteracy',
 'Mathematics',
 'EnvironmentalScience Mathematics',
 'Extracurricular Other',
 'Literacy Mathematics',
 'Literacy Literature_Writing',
 'Mathematics',
 'Mathematics SpecialNeeds',
 'Literacy Literature_Writing',
 'EnvironmentalScience',
 'CharacterEducation',
 'Health_LifeScience Literature_Writing',
 'Literacy Mathematics',
 'Gym_Fitness',
 'Gym_Fitness Health_Wellness',
 'Literature_Writing SpecialNeeds',
 'Gym_Fitness',
 'TeamSports',
 'Health_Wellness Literacy',
 'Literature_Writing',
 'SpecialNeeds',
 'EnvironmentalScience VisualArts',
 'SpecialNeeds',
 'EarlyDevelopment Mathematics',
 'Literacy SpecialNeeds',
 'Literacy SpecialNeeds',
 'Literature_Writing Mathematics',
 'ESL SpecialNeeds',
 ...]
```

In [161]:

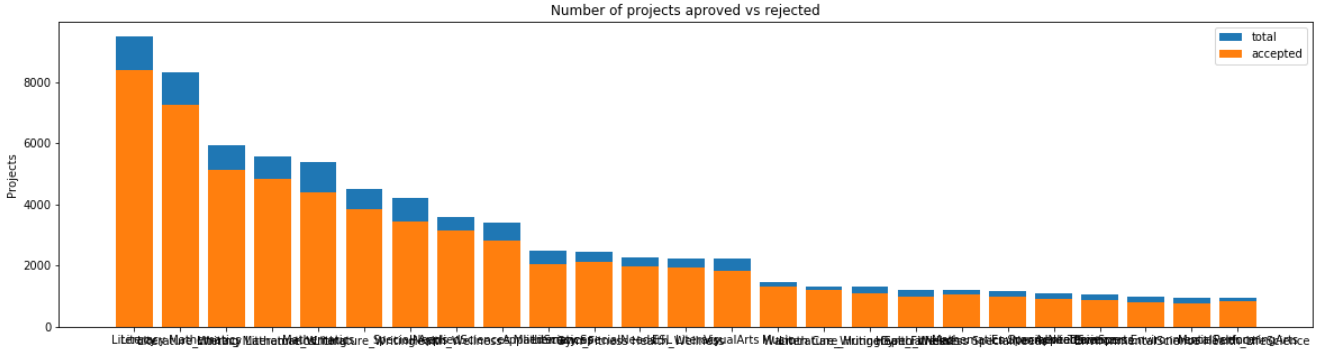
```
project_data['clean_subcategories'] = sub_cat_list
project_data.drop(labels='project_subject_subcategories', axis=1, inplace=True)
project_data.head(2)
```

Out[161]:

| | Unnamed: 0 | id | teacher_id | teacher_prefix | school_state | project_submitted_datetime | pro |
|---|---|---|---|---|---|---|---|
| **0** | 160221 | p253737 | c90749f5d961ff158d4b4d1e7dc665fc | Mrs. | IN | 2016-12-05 13:43:57 | Gra |
| **1** | 140945 | p258326 | 897464ce9ddc600bced1151f324dd63a | Mr. | FL | 2016-10-25 09:22:10 | Gra |

In [162]:

```
#plot it
univariate_barplots(project_data, 'clean_subcategories', 'project_is_approved', top=25)
```



```
                clean_subcategories  project_is_approved  total       Avg
317                         Literacy                 8371   9486  0.882458
319             Literacy Mathematics                 7260   8325  0.872072
331  Literature_Writing Mathematics                  5140   5923  0.867803
318     Literacy Literature_Writing                  4823   5571  0.865733
342                      Mathematics                 4385   5379  0.815207
==================================================
                         clean_subcategories  project_is_approved  total  \
188                      EnvironmentalScience                  894   1079
396                                TeamSports                  864   1061
8       AppliedSciences EnvironmentalScience                  785    984
193  EnvironmentalScience Health_LifeScience                  782    964
356                      Music PerformingArts                  840    948

          Avg
188  0.828545
396  0.814326
8    0.797764
193  0.811203
356  0.886076
```

In [163]:

```
# counting the project cleaned subcategories
from collections import Counter
my_counter_sub = Counter()

for i in project_data['clean_subcategories'].values:
    my_counter_sub.update(i.split())
```

In [164]:

```
my_counter_sub
```

Out[164]:

```
Counter({'AppliedSciences': 10816,
         'Care_Hunger': 1388,
         'CharacterEducation': 2065,
         'Civics_Government': 815,
         'College_CareerPrep': 2568,
         'CommunityService': 441,
         'ESL': 4367,
         'EarlyDevelopment': 4254,
         'Economics': 269,
         'EnvironmentalScience': 5591,
         'Extracurricular': 810,
         'FinancialLiteracy': 568,
         'ForeignLanguages': 890,
         'Gym_Fitness': 4509,
         'Health_LifeScience': 4235,
         'Health_Wellness': 10234,
         'History_Geography': 3171,
         'Literacy': 33700,
         'Literature_Writing': 22179,
         'Mathematics': 28074,
         'Music': 3145,
         'NutritionEducation': 1355,
         'Other': 2372,
         'ParentInvolvement': 677,
         'PerformingArts': 1961,
         'SocialSciences': 1920,
         'SpecialNeeds': 13642,
         'TeamSports': 2192,
         'VisualArts': 6278,
         'Warmth': 1388})
```

In [165]:

```
#convert it into dict
dict_sub = dict(my_counter_sub)
```

In [166]:

```
#convert it into python dataframe so that we can bar plot it with projects
xxx_2 = pd.DataFrame.from_dict(dict_sub.items())
xxx_2.head()
```

Out[166]:

| | 0 | 1 |
|---|---|---|
| **0** | ESL | 4367 |
| **1** | Literacy | 33700 |
| **2** | Civics_Government | 815 |
| **3** | TeamSports | 2192 |
| **4** | Health_Wellness | 10234 |

In [167]:

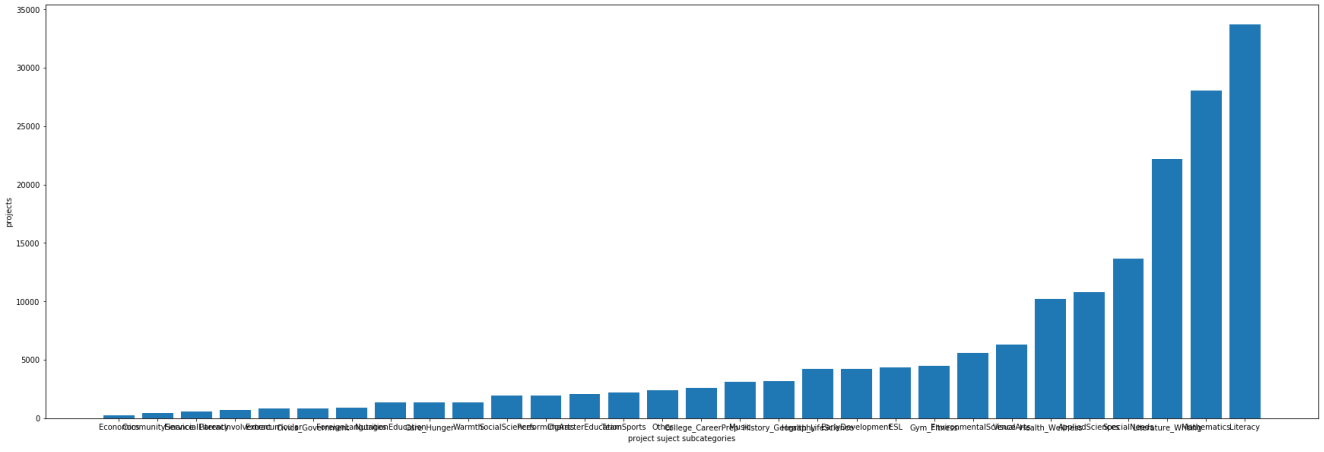
```
#sort it
xxx_2.columns = ['cleaned_sub_categories', 'count']
xxx_2.sort_values(by='count', axis=0, inplace=True)
xxx_2.head()
```

Out[167]:

| | cleaned_sub_categories | count |
|---|---|---|
| **16** | Economics | 269 |
| **26** | CommunityService | 441 |
| **17** | FinancialLiteracy | 568 |
| **8** | ParentInvolvement | 677 |
| **29** | Extracurricular | 810 |

In [168]:

```
#plot it
plt.figure(figsize=(30,10))
plt.bar(xxx_2['cleaned_sub_categories'], xxx_2['count'])
plt.xlabel('project suject subcategories')
plt.ylabel('projects')
plt.show()
```




### 1.2.6 Univariate Analysis: Text features (Title)

In [169]:

```
# Looking for how many words in project title for each project
```

```
word_count = project_data['project_title'].str.split().apply(len).value_counts()
word_dict = dict(word_count)
```

In [170]:

```
#convert into pandas dataframe so that we can plot
xxx_3 = pd.DataFrame.from_dict(word_dict.items())
xxx_3.columns = ['number of words in project title', 'count']
xxx_3.head(2)
```

Out[170]:

| | number of words in project title | count |
|---|---|---|
| **0** | 4 | 19979 |
| **1** | 5 | 19677 |

In [171]:

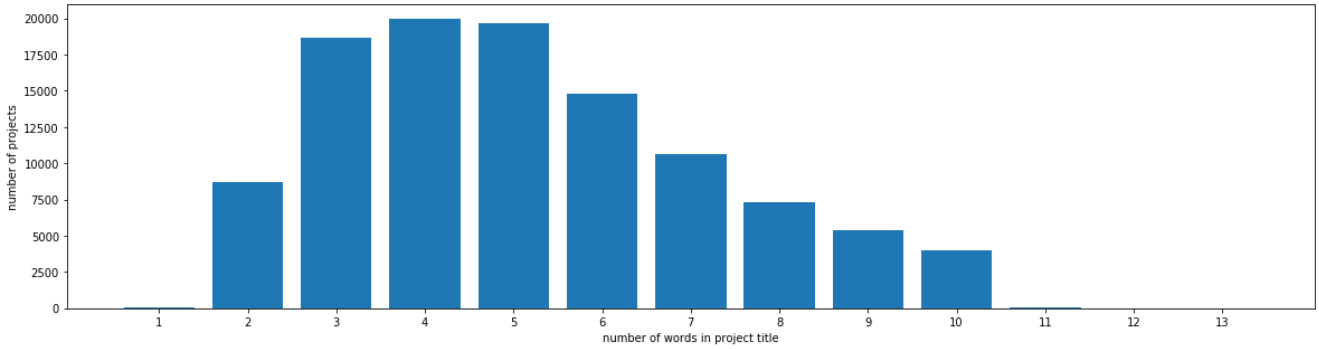
```
#sort it
xxx_3.sort_values(by='count', axis=0, ascending=True, inplace=True)
xxx_3.head()
```

Out[171]:

| | number of words in project title | count |
|---|---|---|
| **12** | 13 | 1 |
| **11** | 12 | 11 |
| **10** | 11 | 30 |
| **9** | 1 | 31 |
| **8** | 10 | 3968 |

In [172]:

```
plt.figure(figsize=(20,5))
plt.bar(xxx_3['number of words in project title'], xxx_3['count'])
plt.xlabel('number of words in project title')
plt.ylabel('number of projects')
plt.xticks(xxx_3['number of words in project title'])
plt.show()
```




In [173]:

```
#check how many words in the project where the project is approved
words_count_approved = project_data[project_data['project_is_approved']==1]
words_count_approved.head(2)
```

Out[173]:

| | Unnamed: | . . | | | | | |
|---|---|---|---|---|---|---|---|

| | Unnamed: 0 | id | teacher_id | teacher_prefix | school_state | project_submitted_datetime | pro |
|---|---|---|---|---|---|---|---|
| 1 | 140945 | p258326 | 897464ce9ddc600bced1151f324dd63a | Mr. | FL | 2016-10-25 09:22:10 | Gra |
| 3 | 45 | p246581 | f3cb9bffbba169bef1a77b243e620b60 | Mrs. | KY | 2016-10-06 21:16:17 | Gra |

In [174]:

```
#finding how many words in that
approved_word_count = words_count_approved['project_title'].str.split().apply(len)
approved_word_count = approved_word_count.values
approved_word_count
```

Out[174]:

```
array([5, 2, 3, ..., 6, 5, 7], dtype=int64)
```

In [175]:

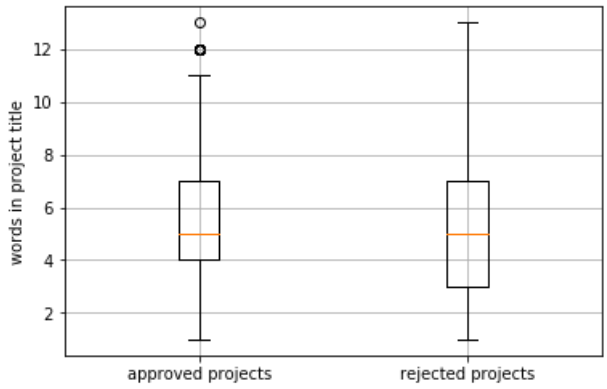
```
rejected_word_count = project_data['project_title'].str.split().apply(len)
rejected_word_count = rejected_word_count.values
rejected_word_count
```

Out[175]:

```
array([7, 5, 7, ..., 6, 5, 7], dtype=int64)
```

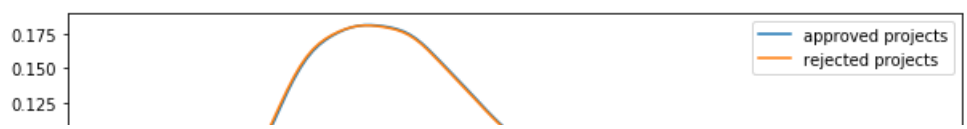
In [176]:

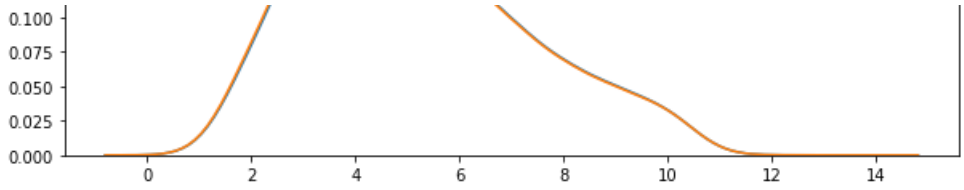
```
#Box plot
plt.boxplot(x=[approved_word_count, rejected_word_count])
plt.xticks([1,2], ['approved projects', 'rejected projects'])
plt.ylabel('words in project title')
plt.grid()
plt.show()
```




In [177]:

```
plt.figure(figsize=(10,3))
sns.kdeplot(approved_word_count, label='approved projects', bw=0.6)
sns.kdeplot(rejected_word_count, label='rejected projects', bw=0.6)
plt.legend()
plt.show()
```






### 1.2.7 Univariate Analysis: Text features (Project Essay's)

In [178]:

```
#merge all the essay columns
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)
```

In [179]:

```
project_data.head(2)
```

Out[179]:

| | Unnamed: 0 | id | teacher_id | teacher_prefix | school_state | project_submitted_datetime | pro |
|---|---|---|---|---|---|---|---|
| 0 | 160221 | p253737 | c90749f5d961ff158d4b4d1e7dc665fc | Mrs. | IN | 2016-12-05 13:43:57 | Gra |
| 1 | 140945 | p258326 | 897464ce9ddc600bced1151f324dd63a | Mr. | FL | 2016-10-25 09:22:10 | Gra |

In [180]:

```
approved_word_count_essay = project_data[project_data['project_is_approved']==1]
['essay'].str.split().apply(len)
approved_word_count_essay =  approved_word_count_essay.values
approved_word_count_essay
```

Out[180]:

```
array([221, 213, 234, ..., 181, 254, 263], dtype=int64)
```

In [181]:

```
rejected_word_count_essay = project_data[project_data['project_is_approved']==0]
['essay'].str.split().apply(len).values
rejected_word_count_essay
```

Out[181]:

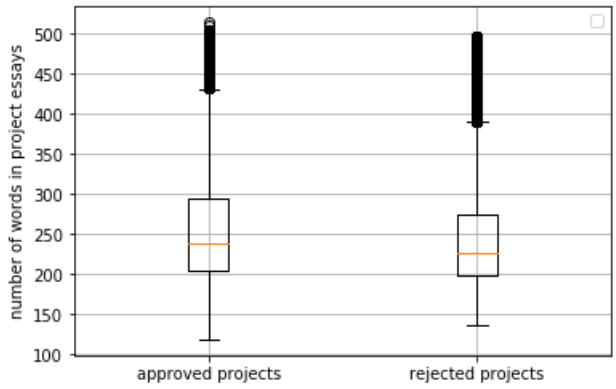
```
array([272, 361, 219, ..., 211, 298, 317], dtype=int64)
```

In [182]:

```
plt.boxplot(x=[approved_word_count_essay, rejected_word_count_essay])
```

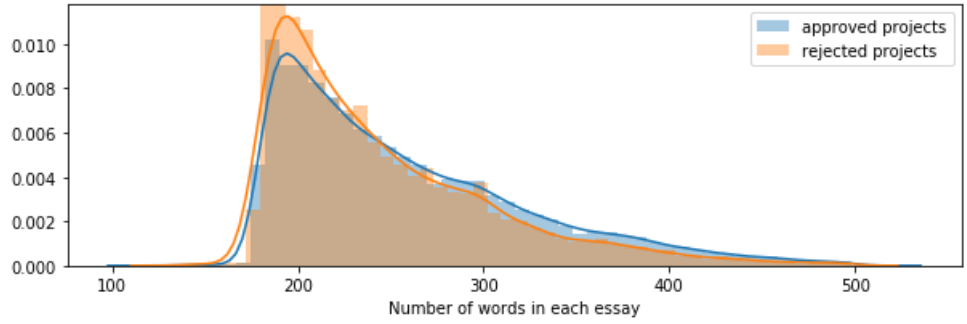
```
plt.xticks([1,2],['approved projects', 'rejected projects'])
plt.ylabel('number of words in project essays')
plt.legend()
plt.grid()
plt.show()
```

```
No handles with labels found to put in legend.
```




In [183]:

```
plt.figure(figsize=(10,3))
sns.distplot(approved_word_count_essay, label='approved projects')
sns.distplot(rejected_word_count_essay, label='rejected projects')
plt.xlabel('Number of words in each essay')
plt.legend()
plt.show()
```




### 1.2.8 Univariate Analysis: Cost per project

In [184]:

```
#Price only available on resource dataset
resources_data.head(2)
```

Out[184]:

| | id | description | quantity | price |
|---|---|---|---|---|
| **0** | p233245 | LC652 - Lakeshore Double-Space Mobile Drying Rack | 1 | 149.00 |
| **1** | p069063 | Bouncy Bands for Desks (Blue support pipes) | 3 | 14.95 |

In [185]:

```
#Since there are too many id's repeated here so we can group it together based on same id
# https://stackoverflow.com/questions/22407798/how-to-reset-a-dataframes-indexes-for-all-groups-in
-one-step
price = resources_data.groupby(by='id')['price'].agg({'price':'sum', 'quantity':'sum'}).reset_index
()
price.head(3)
```

Out[185]:

| | id | price | quantity |
|---|---|---|---|
| **0** | p000001 | 459.56 | 459.56 |
| **1** | p000002 | 515.89 | 515.89 |
| **2** | p000003 | 298.97 | 298.97 |

In [186]:

```
#Since Ids are the same in both the dataset and we can join them like in SQL
project_data = pd.merge(project_data, price, how='left')
project_data.head(2)
```

Out[186]:

| | Unnamed: 0 | id | teacher_id | teacher_prefix | school_state | project_submitted_datetime | pro |
|---|---|---|---|---|---|---|---|
| **0** | 160221 | p253737 | c90749f5d961ff158d4b4d1e7dc665fc | Mrs. | IN | 2016-12-05 13:43:57 | Gra |
| **1** | 140945 | p258326 | 897464ce9ddc600bced1151f324dd63a | Mr. | FL | 2016-10-25 09:22:10 | Gra |

In [187]:

```
#approved projects based on price
approved_price = project_data[project_data['project_is_approved']==1]['price'].values
approved_price
```

Out[187]:

```
array([299.  , 232.9 ,  67.98, ..., 239.96,  73.05, 109.9 ])
```

In [188]:

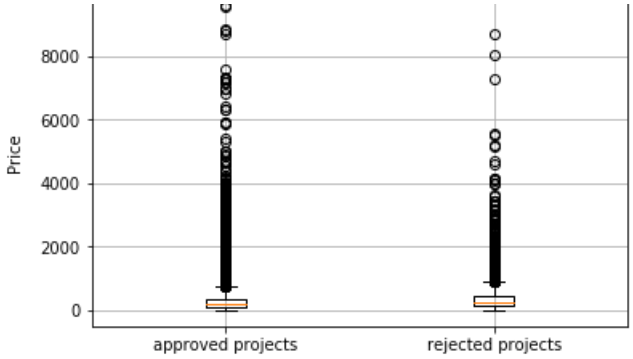
```
#rejected projects based on price
rejected_price = project_data[project_data['project_is_approved']==0]['price'].values
rejected_price
```

Out[188]:

```
array([154.6 , 516.85, 219.46, ..., 747.  , 300.18, 737.95])
```

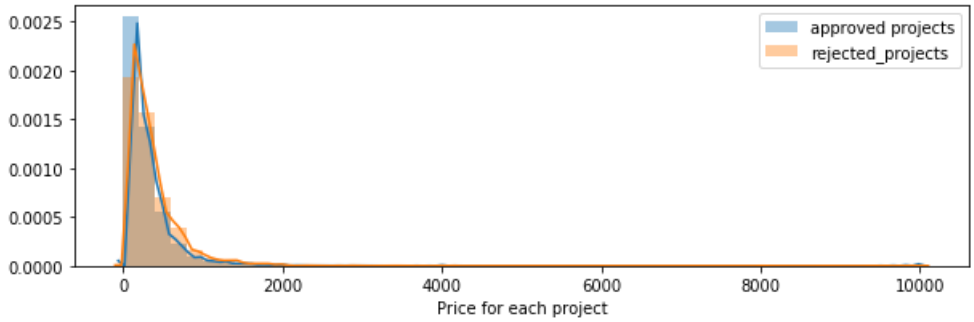
In [189]:

```
#Boxplot
plt.boxplot(x=[approved_price, rejected_price])
plt.xticks([1,2],['approved projects', 'rejected projects'])
plt.ylabel('Price')
plt.grid()
plt.show()
```




In [190]:

```
plt.figure(figsize=(10,3))
sns.distplot(approved_price, label='approved projects')
sns.distplot(rejected_price, label='rejected_projects')
plt.xlabel('Price for each project')
plt.legend()
plt.show()
```




In [191]:

```
project_data['price'].shape
```

Out[191]:

```
(109248,)
```

In [68]:

```
# http://zetcode.com/python/prettytable/
from prettytable import PrettyTable
x = PrettyTable()
x.field_names = ['percentile', 'approved projects', 'rejected projects']

for i in range(0,101,5):
    x.add_row([i, np.round(np.percentile(approved_price, i),3), np.round(np.percentile(rejected_pri
ce,i), 3)])

print(x)
```

```
+------------+--------------------+--------------------+
| percentile | approved projects | rejected projects |
+------------+--------------------+--------------------+
|      0     |        0.66        |        1.97        |
|      5     |       13.59        |        41.9        |
|     10     |       33.88        |       73.67        |
|     15     |        58.0        |       99.109       |
|     20     |       77.38        |       118.56       |
|     25     |       99.95        |       140.892      |
|     30     |       116.68       |       162.23       |
|     35     |      137.232       |       184.014      |
|     40     |       157.0        |       208.632      |
|     45     |      178.265       |       235.106      |
|     50     |       198.99       |       263.145      |
|     55     |       223.99       |       292.61       |
```

```
|      60      |       255.63       |       325.144      |
|      65      |       285.412      |        362.39      |
|      70      |       321.225      |        399.99      |
|      75      |       366.075      |       449.945      |
|      80      |        411.67      |       519.282      |
|      85      |        479.0       |       618.276      |
|      90      |        593.11      |       739.356      |
|      95      |       801.598      |       992.486      |
|     100      |        9999.0      |        9999.0      |
+--------------+--------------------+--------------------+
```

### 1.2.9 Univariate Analysis: teacher_number_of_previously_posted_projects

In [69]:

```
#Now we can do the same to the teacher_number_of_previously_posted_projects
project_data.head(2)
```

Out[69]:

| | Unnamed: 0 | id | teacher_id | teacher_prefix | school_state | project_submitted_datetime | pro |
|---|---|---|---|---|---|---|---|
| **0** | 160221 | p253737 | c90749f5d961ff158d4b4d1e7dc665fc | Mrs. | IN | 2016-12-05 13:43:57 | Gra |
| **1** | 140945 | p258326 | 897464ce9ddc600bced1151f324dd63a | Mr. | FL | 2016-10-25 09:22:10 | Gra |

In [70]:

```
previously_teacher_approved_projects = project_data[project_data['project_is_approved']==1]
['teacher_number_of_previously_posted_projects'].values
previously_teacher_approved_projects
```

Out[70]:

```
array([7, 4, 1, ..., 3, 0, 0], dtype=int64)
```

In [71]:

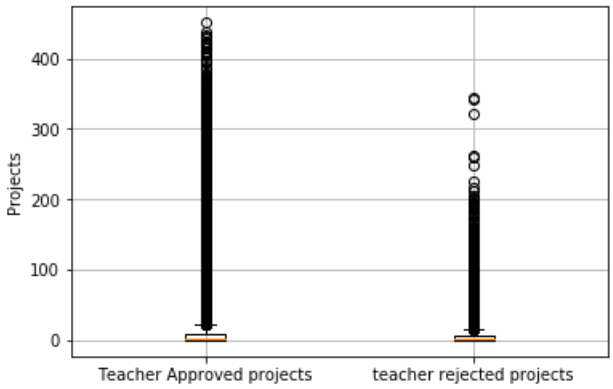
```
previously_teacher_rejected_projects = project_data[project_data['project_is_approved']==0]
['teacher_number_of_previously_posted_projects'].values
previously_teacher_rejected_projects
```

Out[71]:

```
array([0, 1, 5, ..., 4, 0, 1], dtype=int64)
```

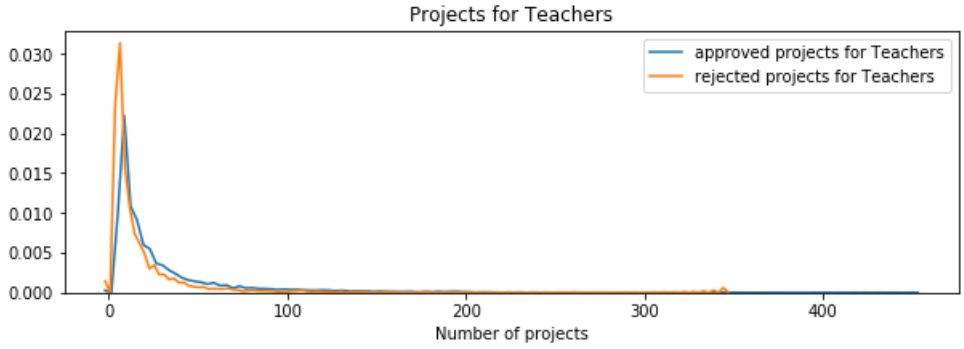
In [72]:

```
#Boxplot
plt.boxplot(x=[previously_teacher_approved_projects, previously_teacher_rejected_projects])
plt.xticks([1,2], ['Teacher Approved projects', 'teacher rejected projects'])
plt.ylabel('Projects')
plt.grid()
plt.show()
```




In [73]:

```
#Distplot
plt.figure(figsize=(10,3))
sns.distplot(previously_teacher_approved_projects, hist=False, label='approved projects for
Teachers')
sns.distplot(previously_teacher_rejected_projects, hist=False, label='rejected projects for
Teachers')
plt.title('Projects for Teachers')
plt.xlabel('Number of projects')
plt.legend()
plt.show()
```




## 1.2.10 Univariate Analysis: project_resource_summary

In [74]:

```
#Now we can look on the project_resource_summary
project_data.head(2)
```

Out[74]:

| | Unnamed: 0 | id | teacher_id | teacher_prefix | school_state | project_submitted_datetime | pro |
|---|---|---|---|---|---|---|---|
| **0** | 160221 | p253737 | c90749f5d961ff158d4b4d1e7dc665fc | Mrs. | IN | 2016-12-05 13:43:57 | Gra |
| **1** | 140945 | p258326 | 897464ce9ddc600bced1151f324dd63a | Mr. | FL | 2016-10-25 09:22:10 | Gra |

| Unnamed: | | | | | | | |
|---|---|---|---|---|---|---|---|

In [75]:

```
#Number of words on each summary when projecte is approved
word_count_summary_approved = project_data[project_data['project_is_approved']==1]
['project_resource_summary'].str.split().apply(len).values
word_count_summary_approved
```

Out[75]:

```
array([11, 20, 26, ..., 36, 15, 27], dtype=int64)
```

In [76]:

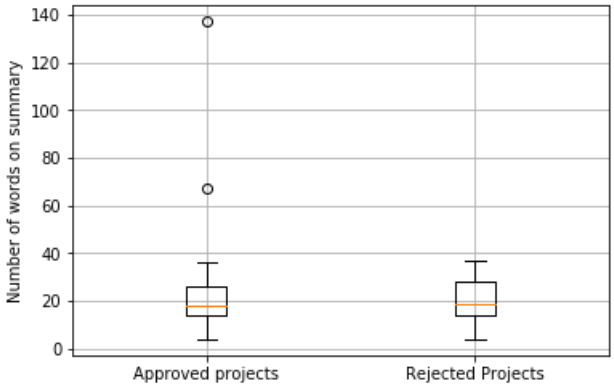
```
#Number of words on each summary when projecte is rejected
word_count_summary_rejected = project_data[project_data['project_is_approved']==0]
['project_resource_summary'].str.split().apply(len).values
word_count_summary_rejected
```

Out[76]:

```
array([13, 19, 32, ..., 19, 11, 18], dtype=int64)
```

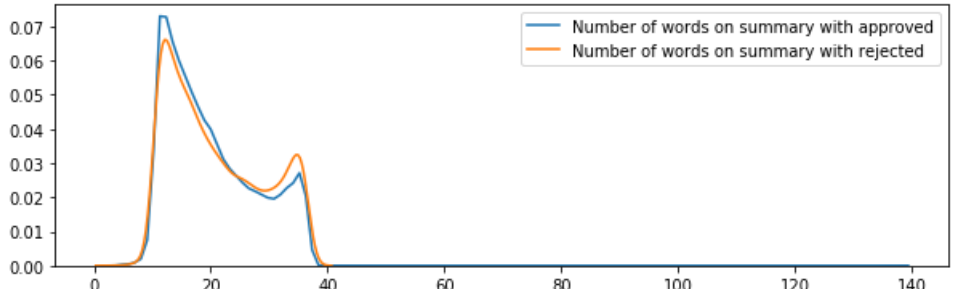
In [77]:

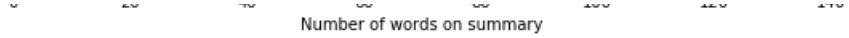
```
#boxplot
plt.boxplot([word_count_summary_approved, word_count_summary_rejected])
plt.xticks([1,2],['Approved projects', 'Rejected Projects'])
plt.ylabel('Number of words on summary')
plt.grid()
plt.show()
```




In [78]:

```
#Dist plot
plt.figure(figsize=(10,3))
sns.distplot(word_count_summary_approved, hist=False, label='Number of words on summary with
approved')
sns.distplot(word_count_summary_rejected, hist=False, label='Number of words on summary with
rejected')
plt.xlabel('Number of words on summary')
plt.legend()
plt.show()
```






## 1.3 Text preprocessing

### 1.3.1 Essay Text

In [79]:

```
project_data.head(2)
```

Out[79]:

| | Unnamed: 0 | id | teacher_id | teacher_prefix | school_state | project_submitted_datetime | pro |
|---|---|---|---|---|---|---|---|
| 0 | 160221 | p253737 | c90749f5d961ff158d4b4d1e7dc665fc | Mrs. | IN | 2016-12-05 13:43:57 | Gra |
| 1 | 140945 | p258326 | 897464ce9ddc600bced1151f324dd63a | Mr. | FL | 2016-10-25 09:22:10 | Gra |

In [80]:

```
project_data['essay'].values[2000]
```

Out[80]:

```
"Describing my students isn't an easy task.  Many would say that they are inspirational, creative, and hard-working.  They are all unique - unique in their interests, their learning, their abilities, and so much more.  What they all have in common is their desire to learn each day, despite difficulties that they encounter.  \\r\\nOur classroom is amazing - because we understand that everyone learns at their own pace.  As the teacher, I pride myself in making sure my students are always engaged, motivated, and inspired to create their own learning! \\r\\nThis project is to help my students choose seating that is more appropriate for them, developmentally.  Many students tire of sitting in chairs during lessons, and having different seats available helps to keep them engaged and learning.\\r\\nFlexible seating is important in our classroom, as many of our students struggle with attention, focus, and engagement.  We currently have stability balls for seating, as well as regular chairs, but these stools will help students who have trouble with balance, or find it difficult to sit on a stability ball for a long period of time.  We are excited to try these st ools as a part of our engaging classroom community!nannan"
```

In [81]:

```
#Remove the contracting words in that essay
# https://stackoverflow.com/a/47091490/4084039

def decontract(phrase):

    phrase = re.sub(r" won't", 'will not', phrase)
    phrase = re.sub(r"can't", 'can not', phrase)

    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 [83]:

```
import re
sent = decontract(project_data['essay'].values[20000])
print(sent)
```

```
My kindergarten students have varied disabilities ranging from speech and language delays, cogniti
ve delays, gross/fine motor delays, to autism. They are eager beavers and always strive to work th
eir hardest working past their limitations. \r\n\r\nThe materials we have are the ones I seek out
for my students. I teach in a Title I school where most of the students receive free or reduced pr
ice lunch.  Despite their disabilities and limitations, my students love coming to school and come
eager to learn and explore.Have you ever felt like you had ants in your pants and you needed to gr
oove and move as you were in a meeting? This is how my kids feel all the time. The want to be able
to move as they learn or so they say.Wobble chairs are the answer and I love then because they dev
elop their core, which enhances gross motor and in Turn fine motor skills. \r\nThey also want to l
earn through games, my kids donot want to sit and do worksheets. They want to learn to count by ju
mping and playing. Physical engagement is the key to our success. The number toss and color and sh
ape mats can make that happen. My students will forget they are doing work and just have the fun a
6 year old deserves.nannan
```

In [84]:

```
# removing the line breaks
# http://texthandler.com/info/remove-line-breaks-python/

sent = sent.replace("\\r", ' ')
sent = sent.replace("\\n", ' ')
sent = sent.replace('\\"', ' ')
print(sent)
```

```
My kindergarten students have varied disabilities ranging from speech and language delays, cogniti
ve delays, gross/fine motor delays, to autism. They are eager beavers and always strive to work th
eir hardest working past their limitations.     The materials we have are the ones I seek out for
my students. I teach in a Title I school where most of the students receive free or reduced price
lunch.  Despite their disabilities and limitations, my students love coming to school and come eag
er to learn and explore.Have you ever felt like you had ants in your pants and you needed to groov
e and move as you were in a meeting? This is how my kids feel all the time. The want to be able to
move as they learn or so they say.Wobble chairs are the answer and I love then because they develo
p their core, which enhances gross motor and in Turn fine motor skills.   They also want to learn t
hrough games, my kids donot want to sit and do worksheets. They want to learn to count by jumping
and playing. Physical engagement is the key to our success. The number toss and color and shape ma
ts can make that happen. My students will forget they are doing work and just have the fun a 6 yea
r old deserves.nannan
```

In [85]:

```
#remove spacial character: https://stackoverflow.com/a/5843547/4084039
sent = re.sub('[^A-Za-z0-9]+', ' ', sent)
print(sent)
```

```
My kindergarten students have varied disabilities ranging from speech and language delays cognitiv
e delays gross fine motor delays to autism They are eager beavers and always strive to work their
hardest working past their limitations The materials we have are the ones I seek out for my studen
ts I teach in a Title I school where most of the students receive free or reduced price lunch
Despite their disabilities and limitations my students love coming to school and come eager to lea
rn and explore Have you ever felt like you had ants in your pants and you needed to groove and mov
e as you were in a meeting This is how my kids feel all the time The want to be able to move as th
ey learn or so they say Wobble chairs are the answer and I love then because they develop their co
re which enhances gross motor and in Turn fine motor skills They also want to learn through games
my kids donot want to sit and do worksheets They want to learn to count by jumping and playing
Physical engagement is the key to our success The number toss and color and shape mats can make th
at happen My students will forget they are doing work and just have the fun a 6 year old deserves
nannan
```

In [86]:

```
# https://gist.github.com/sebleier/554280
# we are removing the words from the stop words list: 'no', 'nor', 'not'
```

```
# 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"]
```

In [87]:

```
#Combining all the above statements
from tqdm import tqdm

proocessed_essay = []
for i in tqdm(project_data['essay'].values):

    sent = decontract(i)
    sent = sent.replace("\\r", ' ')
    sent = sent.replace("\\n", ' ')
    sent = sent.replace("\\'", ' ')
    sent = re.sub('[^A-Za-z0-9]+', ' ', sent)

    sent = ' '.join(j for j in sent.split() if j not in stopwords)
    proocessed_essay.append(sent.lower().strip())
```

```
100%|██████████████████████████████████████████████████████████████| 109248/109248 [02:11<00:00, 829.03it/s]
```

### 1.3.2 Project title Text

In [88]:

```
Processed_title = []

for i in tqdm(project_data['project_title'].values):
    sent = decontract(i)
    sent = sent.replace("\\r", ' ')
    sent = sent.replace("\\n", ' ')
    sent = sent.replace("\\'", ' ')
    sent = re.sub('[^A-Za-z0-9]+', ' ', sent)

    sent = ' '.join(j for j in sent.split() if j not in stopwords)
    Processed_title.append(sent.lower().strip())
```

```
100%|██████████████████████████████████████████████████████████████| 109248/109248 [00:07<00:00, 13763.03it/s]
```

In [89]:

```
Processed_title
```

Out[89]:

```
['educational support english learners home',
 'wanted projector hungry learners',
 'soccer equipment awesome middle school students',
 'techie kindergarteners',
 'interactive math tools',
 'flexible seating mrs jarvis terrific third graders',
 'chromebooks special education reading program',
 'it 21st century',
 'targeting more success class',
 'just for love reading pure pleasure',
 'reading changes lives',
 'elevating academics parent rapports through technology',
 'building life science experiences',
 'everyone deserves heard',
 'tablets can show us the world',
 'making recess active',
 'making great leap with leapfrog',
 'technology teaches tomorrow talents today',
 'test time',
 'wiggling our way success',
 'magic carpet ride our library',
 'from sitting standing classroom',
 'books budding intellectuals',
 'instrumental power conquering steam',
 's t e a m challenges science technology engineering art math',
 'math masters',
 'techy teaching',
 '4th grade french immersion class ipads',
 'hands on language literacy',
 'basic classroom supplies needed',
 '2nd grade explores world charlotte web',
 'an all inclusive learning space',
 'learning facts from fiction',
 'computing our way financial literacy part 2',
 'have a ball',
 'put me in coach',
 'inquiry based discovery through laptop learning',
 'target our kids with a printer and ink',
 'kinders inspired target fitness part one',
 'engaging students technology',
 'leveling books multi age class',
 'a twist writing traits my first graders',
 'we need non fiction',
 'all hands tech',
 'pressing mastery after flood',
 'chromebooks create intrigue and motivation while filling in the gaps',
 'all out paper',
 'keep our closet open',
 'chromebook are gold',
 'rainy day run around',
 'be active be energized',
 'great books clean organized filing cabinets successful students',
 'stand learn',
 'reading together',
 'swim for life at ymca',
 'stem we need capitalize technology',
 'but first coffee',
 'the mouse hunt',
 'awesome authors need terrific table',
 'interactive assessments',
 'picnic table to make us able to do more',
 'forming magnificent minds',
 'adding interactive technology to the young writers toolbox',
 'we need more paper ink new year',
 'read about art',
 'keep computer lab nice',
 'science technology engineering math oh my kinder stem',
 'the magic reading',
 'stem made simple sensible integrated meaningful purposeful learning engaging',
 'chrome up my class',
 'immersion trip outdoor gear',
 'magnets electricity live on',
 'gotta catch chromebook',
 'college signing day rally prizes deserving students',
```

'student seating paradise',
'jump with music',
'a comfortable place learn',
'growth mindset future',
'stand up desks mrs brown class',
'make music make our year',
'crazy computers',
'i need seat',
'fall love reading',
'classroom books inspire',
'planes trains steam',
'technology bust',
'targeting love baseball hitting bulls eye',
'exploring graphic novels',
'read understand like everyone else',
'education through technology',
'publishers need printer',
'chicka chicka boom boom help us cool classroom',
'help keep us motivated',
'new music stands for benton middle high school band',
'new literacy unit books',
'ilearn igrow isucceed ipads',
'leveled books everyone',
'dr seuss others help us read',
'buttons bulldogs',
'teaching math with manipulatives',
'21st century learners 21st century technology',
'fun physically fit',
'project some light over here',
'buzzing with books',
'pe health technology',
'ceramics our history clay sculpture',
'louisiana flooded classroom',
'paper pencils markers oh my',
'easy 1 2 3',
'start year strong',
'super supplies',
'1st graders reaching for the stars in reading and writing',
'writer workshop 1st grade authors',
'technology classroom',
'doing it the write way',
'paper paper paper please',
'diminish digitial divide',
'global learners taking lead',
'magical morphing exploring wondrous life cycle butterflies',
'use your marbles stimulate brain',
'flexible seating',
'creating sense community',
'help us have ball pe',
'book em danno',
'easy eyes nose',
'student materials needed',
'ipad and supplies for our room',
'gone chopin bach five',
'without a string the world is silent',
'bring projects life with color',
'learning my words and listening to my weekly',
'who doesnot love lego books',
'deep heart texas',
'make move',
'music please motivation needed more vigorous physical activity',
'art 21st century',
'sand water word work fun',
'not all who wander are lost j r r tolkien',
'kindergarten stem stations',
'first grade cool coding',
'a classroom library exceptional students',
'learning through history holocaust',
'we want shake wobble bounce',
'ready to go with our macbook pro',
'book bins all',
'spreading love one card time',
'ditch white board get boogie board',
'who needs chromebook we do',
'steam through technology inquiry based learning',
'flexible seating',
'more movement hokki stools',

'we need a kleen slate',
'6th graders think like engineers',
'shhhhh i working independently in learning center',
'from national parks outer space',
'flexible seating optimal learning',
'getting our move on',
'safely searching critters',
'ocean bots deep sea explorers',
'lego robotics programming resources digital media mics',
'help us organize our classroom library',
'history comes life mrs butler 5th grade class',
'healthy bodies dirty hands let kids kids',
'can you hear me now help my students hear heard',
'hovering our hovergraft',
'pedaling way through our day',
'team practices make perfect teams',
'my students crazy apple watches',
'growing future programmers',
'scientific calculators',
'alternative seating comfy classrooms',
'tablets rescue',
'colorful learning environment',
'tech my knowledge',
'wobble while you work',
'reading together fun',
'the magic steam prek',
'my education my seating choice flexible seating classroom',
'globe gazing my students want see world',
'learning technology',
'pitter ipadder',
'teachers love video games too',
'testing point',
'calculators kids',
'building math life skills future',
'becoming architects engineers builders age 6',
'flexible seating third',
'ineeds support steam',
'endless possibilities',
'building grade level chromebooks',
'seating active learners',
'providing learning environment that kids need',
'headphones kiddos',
'operation 60',
'projecting way future',
'amazing anchor charts',
'behavior technology a match made heaven',
'sitting pretty science lab',
'foundations pre k writing',
'if you write you author',
'sensory toys make sense world',
'ring our bells all hear',
'chromebooks stem',
'balls bubbles birthday books',
'flexible seating flexible classroom',
'safe books bouncy balls',
'some like it hot some like it cold',
'5th grade life science owl pellet dissection project',
'perfect position 5th grade orchestra',
'an urban prek 5 technology classroom looking 3d printer',
'learning can be fun rewarding',
'music knowledge go hand hand part 2',
'help feed newington elementary school students',
'classroom projection 21st century way',
'calculating our success',
'i scope you can help us become well rounded readers',
'savvy stem start up using robust robotics',
'21st century technology needed',
'help us bring music home part 2',
'robots what would lives like without',
'sew happy',
'chevron we solve and create with ipads',
'taking hyper out hyperactivity',
'students need supplies',
'science for success',
'just little wiggle room',
'wiggle seating kinder kids',
'7th grade action researchers',

'pick us play tune',
'current events classroom',
'open window technology',
'help us develop reader workshop',
'ipad learning',
'rollin down river',
'fidget toys chairs middle school kids',
'guided reading resources',
'fueling our bodies minds right from start',
'enhancing minds research inquiry',
'amplifying student learning one ipad time',
'flexible seating',
'listening center my little learners',
'21st century students 21st century classroom',
'art enhances learning',
'the future bright with technology',
'technology environment responsibility',
'wiggle while they work',
'get recharged',
'let the children play',
'life essentials',
'cu l es la ecuaci n de esta l nea',
'classroom carpet',
'flexible seating creating 21st century learning environment',
'robotics future interactive minds create',
'board not bored',
'bearcat chem try',
'autodesk inventor comes alive with 8gb memory computers o high',
'writing towards success',
'see learning',
'help us hokki pokie',
'journey new exciting places',
'a ray cleanliness',
'sharpen pencils',
'becoming literate citizens',
'reading the classics in class before they disappear forever',
'go go google gadgets',
'lifting weights lifting spirits',
'book month program',
'technology makes learning meaningful',
'staying up date',
'amazing student work binders',
'document camera present dissections projects diagrams lessons',
'books books books',
'fire up learning with amazon fire tablet',
'hands on learning stems',
'classroom pets fish tadpoles turtles chameleons hermit crabs',
'healthy snack attacks',
'flexible seating fun',
'stay fit exercise with spark 2',
'optimizing reading growth accelerated reader',
'super star second graders',
'multiplying our efforts after flood',
'we like move it move it',
'we are oysters looking for our pearls',
'classroom essentials',
'prototyping help others',
'super scholars accelerating towards excellence',
'audio books students with visual impairments',
'furniture firsties',
'oodles outdoor fun',
'learning shred',
'ap literature success new novels',
'social studies first',
'chromebooks fantastic 1st graders',
'gamification learning',
'a kindergarten stem grow on',
'guitar tuners',
'band basics create music',
'mindfulness essential oils',
'reading learning about friendship',
'tablets inspire middle school math minds',
'biology interactive learning log',
'littlebits big learning',
'organization express train',
'ordinary finds extraordinary minds',
'special supplies bilingual students',

'special supplies bilingual students',
'creativity is intelligence having fun bobcats steam ahead',
'active recess let get moving',
'wobble while we work',
'what ya make 3d objects',
'ap chemistry prep books',
'take me out to the ballgame',
'learn comfort',
'we like move move flexible seating',
'moving robots',
'chromebook for learning',
'reading on the front lines in 7th grade',
'healthy lives archery',
'tablet tech',
'handwriting without tears',
'stem twenty first century students',
'getting comfy our classroom library',
'chromebooks here chromebooks there chromebooks everywhere',
'bring color',
'education sweet nice seat',
'teach students they remember engage students they learn',
'technology enhances learning',
'kids coding creativity',
'diving into the microscopic world',
'collaborate with chrome',
'grab seat go',
'help us rock learn',
'learning use technology one ipad time',
'wiggly worms',
'splendid science',
'balance balls balanced learning experience',
'strike band',
'ipad art room',
'leveled readers happy students',
'e books r us',
'flexible seating for focused students',
'pretty presentations',
'proficiency scientific presentations',
'walls wiggly students need wiggly seats',
'developing love reading part 3',
'graphing calculators higher mathematics',
'the phonics reading club',
'getting comfy engaged new carpet',
'computer science math class',
'no ordinary organizer',
'turn frowny faces upside',
'cleaning classroom library',
'mini ipad huge difference',
'a poetry celebration',
'active classroom',
'break tech learn cooperatively',
'seeking sensational supplies',
'stop safety patrol',
'supplies success',
'flexible seating our flexible learning space',
'hydroponic garden',
'a classroom rug ms clark class',
'starting with sounds words',
'tech savvy third graders need tablets',
'recording live music with a macbook pro',
'books grow below grade level readers',
'flexible seating',
'supplies needed',
'laughs learning through poetry',
'meeting individual needs one scribble at a time',
'algebra 1 supplies',
'how will world end a study dystopian literature',
'today reader tomorrow leader',
'stem learning brought life',
'1st grade wise owls',
'movement towards healthy lifestyle',
'the ties that bind custom built writing portfolios',
'full steam ahead complete our chromebook cart',
'help my journalism students go pro',
'reading takes you greatest adventure',
'face facts developing nonfiction classroom library',
'take your seat',
'stand up move',

'stand up move',
'hela cells',
'pencils notebook folders please',
'classroom rugs center learning first grade',
'kids program code dash dot robots csforall hourofcode',
'mind your math',
'creativity critical thinking interactive technology',
'math tools classroom',
'technology music classroom',
'building forever readers',
'chromebooks build confidence english language learners',
'boxing our way academic success',
'movement is freedom',
'urban garden grows interest environmental science',
'kindergarteners love wobble',
'books carriers kindergarten literacy centers',
'taking display student work next level',
'chapter books third graders',
'wireless tech developing journalists',
'leave it better than you found it',
'bean bag pod',
'criss cross applesauce',
'by the time i graduate will i need a textbook',
'calc kids need calculators',
'wobble away 2nd grade',
'read lead succeed',
'no more squeaks squawks woodwind mouthpieces needed',
'students need think their feet',
'stem kindergarten',
'one book two books red book blue book',
'letters numbers come to life',
'creating 3rd grade community learners',
'technology alternative classroom experience',
'the chrome needs polishing order sparkle',
'reading math helps mind bloom',
'a tune makes lesson a better class',
'operation graphic design',
'smart tv needed smart music students',
'louisiana flooded students growing giftedness',
'extra extra classroom supplies needed',
'landmark art',
'mini ipads awesome 2nd grade learners',
'explore tubs',
'comfort classroom success',
'fostering social emotional development multicultural pre k class',
'the fourth r recess',
'experience another dimension math 3d printer',
'silence golden',
'wobble my wiggles away',
'keeping newark fit',
'extraordinary students need technology',
'connecting beyond classroom',
'graphic novels reading',
'having fun school',
'our neighborhood work',
'get moving get cozy get learning',
'we need technology middle school',
'superhero literacy',
'hi ho hi ho we need osmo',
'making students feel home with cozy classroom',
'classroom library needs books',
'technology finger tips',
'keep calm use cromebook',
'desktops desktops desktops',
'too loud think printer without ink technology sink',
'flexible seating activity rug promote active healthy individuals',
'we need move it move it',
'math must haves',
'paramount technology integration',
'personalized science notebooks',
'bistro style library',
'exploring science stem experiments',
'sixth graders need book club books',
'a classroom students want',
'learning flexible so is our classroom',
'mrs esposito class loves learning current events',
'for love literacy',
'food fuel learning'

'food fuel learning',
'light kindle fire learning',
'my community is graffiti wordle',
'tablets third grade',
'center time',
'hula hoop moving groovin',
'goodbye desk chairs',
'get my kinders fired up reading',
'dress play',
'take seat learning neat',
'know your h20 groundwater quality testing',
'learn like 2099',
'bounce into learning',
'tablets individualized learning',
'chemists chrome books',
'flexible seating flexible learning',
'listening center 4 daily 5',
'decreasing reading gap level o',
'our math skills will keep getting hotter with hot dots',
'play time first step learning',
'math their fingertips',
'technology all mrs wahlberg class',
'mathematicians ahead',
'making students centered on learning',
'creative kinders',
'more than just our abcs kindergarten literacy materials',
'we need organized classroom',
'the world classroom',
'painting outside part ii',
'in need s amore speech therapy materials',
'fidgets help us focus',
'ifit going gold part ii',
'critical thinking through sensory play',
'learning beyond classroom',
'building community one recess game time',
'robots stem education san bernardino',
'super scientists',
'mind blowing math motivating young mathematicians',
'computers explore',
'staying active during indoor recess',
'apple harvest knowledge farmers',
'novels reach new levels',
'classroom chromebooks college bound seniors',
'stemulating lab phase ii',
'teaching daily living skills to special needs children',
'clean tidy ready learn',
'listen learn',
'help me with my wiggles',
'burn calories at your desk',
'pedaling proficiency pedal seats alternative seating options',
'inspiring readers writers through technology',
'flexible seating',
'fun pe equipment',
'making insiders outsiders',
'inspiring stem activities kindergarten',
'ilearn ipads',
'we innovative providing tools interactive engagement',
'demonstration tools learning fun',
'wiggle room flexible seating options small groups',
'transforming stationary learning active movement opportunities',
'moving grooving 5th grade',
'taking care our bodies is now one less concern',
'the chrome zone',
'coding kindergarten',
'let accessorize',
'teamwork preschool',
'the other side fairytale',
'organization and planning are the keys to success',
'math reading is what we are needing',
'read teach repeat',
'angry birds physics',
'google apps helps us create',
'stem k 2',
'making pens pencils ourselves others',
'stand up ipads',
'daily road maps children',
'getting staying healthy',
'modeling multiple learning styles'

'modeling multiple learning styles',
'creative technology',
'hands math science tools superhero class',
'adventurous amazing books our library',
'ict class needs a chromebook',
'coding sphero',
'imagination digital storytelling',
'supplies should not a limiting factor',
'using music teach reading',
'exploring earth through seismicity',
'we rise above it all',
'our fairy tales folktales falling apart',
'ambitious science teaching why will alaskan way viaduct collapse',
'unleashing potential',
'get gullah with us',
'texts for all',
'language reading intervention',
'bridging gap',
'can hear',
'election fall 2016 materials',
'full tummies full hearts full minds',
'help young learners access technology',
'give them possibilities read their favorite books',
'notebooks young writers',
'keep everything weighing same',
'let use math understand world',
'loving literacy',
'i all ears',
'tetherball courts health exercise',
'carnival indoor recess fun',
'building bots',
'shredding through oldies',
'taking closer look through modeling independent learning',
'getting fit with ozo pedometers',
'full stem ahead',
'move music',
'put your listening ears',
'mom dad did you see my work my portfolio',
'21st century technology 21st century learners',
'science art together no way',
'classroom supplies',
'back basics school supplies classroom',
'empower young minds flexible seating classroom',
'loud proud',
'picking up steam kindergarten',
'time saved learning maximized',
'graphic novels rescue',
'all fun games while making academic gaines',
'the best seat class',
'chromebook robotics stem part 2',
'print world color',
'our classroom wish list this year',
'exploring enjoying life through great book',
'new year resolution become amazing readers',
'meeting students fine motor sensory needs special education',
'books all reading levels',
'engaging ourselves with technology',
'stem readers',
'kindergarten stations full steam ahead',
'starbuck goals',
'books hand adventures school',
'digital classroom library',
'a just right red chair pre k',
'that what in an owl pellet',
'ipads to motivate engage my students love reading',
'microscopes engage elementary students scientific investigation',
'students deserve the best',
'let make calender math possible',
'in sight in mind',
'extra extra read all about it',
'books our nook',
'tables fit needs little bodies',
'soccer equipment',
'bringing insects life 3d',
'book read alouds catapulting our students success',
'tools build lifetime skills',
'lockdown drills not annoyance',
'wobble chairs keep moving',

'wobble chairs keep moving',
'closing gap apps',
'reading using inference skills painting our ocean friends',
'library lacking literacy',
'fun 3d doodle set',
'reading rugs',
'apple pi',
'making music a family affair',
'make my students tech savvy',
'dear santa philadelphia 8th graders want books for christmas',
'magical math literature',
'more more equal access for all',
'hooked books',
'moving is our target',
'to hear the music pound let beef up our sound',
'focused learning',
'our students rock',
'perceiving patterns painting',
'magazines make learning fun',
'let play hockey',
'innovation nation creating learning space student exploration',
'tummies rumble when empty',
'flexible seating active seating active learners',
'on right track backpacks',
'find your colored square',
'help 5th grade scientists learn with technology',
'dusting off soul',
'taking learning scholastic let find out',
'scholastic news',
'learning science through hands approach',
'osmo ipad stem centers',
'keep chrome books safe fully charged every day',
'there nothing do end recess boredom get fit',
'organizing guiding future readers',
'oh baby parenting f a c s',
'creativity crayola',
'graphic novels library',
'we take what we value granted',
'listen love learning headphones needed',
'fidgeting students need fidgets',
'ipad minis many learners',
'macbook pro for my computer pros',
'engage students flexible seating',
'crazy ukulele',
'wiggly bottoms need special seats',
'we got beat we need drums',
'life after hurricane matthew',
'making magical music',
'steam and stem growing together',
'the printing press 2 0',
'clay glaze storage new kiln 05 02 16',
'curing autism mrs carter class',
'reading table',
'need reach our virtual mentors',
'walk on',
'let paint',
'carts computers',
'the great bridge project',
'flexible comfortable seating',
'let strings sing',
'pottery club',
'the art teaching kids need zen art school',
'stem kits maker space',
'stem books animal reports',
'creative sticky murals',
'feed our minds hungry students need snacks',
'keep school garden alive thriving',
'walking playing purpose',
'class library lacking chapter books',
'stand up success',
'wiggle while work',
'music books new musicians',
'flexible seating focus',
'math manipulatives eq3',
'help us put our supply shelf back together again',
'special education students need work station desks chairs',
'comfy chairs will help us become scholars',

'it happy day pre k',
'learn science lost wax jewelry',
'green screen projects help wanted',
'hands on exploration problem solving stem',
'binder finder',
'googlify our classroom',
'from abstract reality',
'active bodies engaged minds',
'beautiful copies',
'learning through listening a new literacy center',
'classroom manipulatives my amazing second graders',
'help us play adapted sports',
'technology technology we all about it',
'reaching reading goals',
'ipad minis kindergarten minis',
'technology art oh my',
'building print rich classroom',
'listen while you work',
'math center activities',
'you canot do required reading without required book',
'weaving through history',
'we got wiggles',
'bamboo pads differentiated learning',
'we want fitbits share please',
'physical education move',
'a carpet the heart our classroom community',
'organized manipulatives my motivated mathmeticians',
'the art collaborative working',
'help me teach',
'kill watt energy',
'hot dots learning',
'math tools create success',
'read together learn together',
'touch lives with touchtronic technology',
'sturdy shelving',
'addition way life',
'let calm read',
'burlington backpacks win',
'wiggle while you work flexible seating options',
'teaching pitch during critical period auditory development',
'science technology math yes please',
'kindergarten makeover',
'balance discs allow brain readiness learn',
'ipad myclass',
'rockin school chairs students autism spectrum',
'creating digital learners',
'mrs newsome',
'desktop computers will support inclusion special education students',
'it mathterpiece',
'school wide mindfulness',
'focus movement',
'kinesthetic kinders like move it move it',
'plant seed read',
'backpacks class',
'technology today transcendence tomorrow',
'supplies needed growing minds',
'flexible seating project',
'help our room got flooded',
'creative comfortable stem projects',
'extra extra read all about it reading in kindergarten',
'highlight this',
'look me grow',
'building student knowledge with geometric shape building sets',
'an organized classroom happy classroom',
'help immerse our art class watercolors',
'multiple mallet mania',
'robotics 3d printing our urban makerspace classroom',
'family engagement stem',
'media center makeover bringing school library inviting students',
'make learning permanent',
'stand up swing success',
'tiles not comfy',
'vivid visuals math reading',
'full steam ahead',
'teaching triumphantly tablets',
'raved readers',
'middle school supplies smiles',

```
'variety spice literature',
'book boxes clipboards mrs chen',
'discovering phantom language the phantom tollbooth',
'extra extra storage that is',
'scientific calculators science',
'charging our chrome',
'getting comfy cozy reading rug',
'harnessing wiggles with hokki stools',
'we like move it move it',
'steming ahead with folktales',
'act books',
'chromebooks classroom',
'we crazy coding',
'controlling robots one code time',
'starbucks classroom',
'empowering students through art creativity comes alive',
'check out playosmo com',
'food soul',
'miss luce classroom mailbox',
'never too young to be healthy',
'in living color',
'keep music alive',
'chromebooks classroom',
'hear music see music',
'first grade is full steam ahead',
'chromebooks my third grade class',
'in living color digital too',
'can you hear me now',
'1 2 books from you 3 4 we thank you even more',
'3d printer young designers innovators',
'show me why money matters',
'we want learn english',
'reaching new goals fitness mindfulness',
'science is so much fun',
'hands on minds on',
'more technology please',
'the read',
'painting supplies talented 4th graders',
'movin groovin workin part 2',
'healthier happier students',
'watch tech watch learn learn',
'supplies starting second grade',
'let get rid desks',
'backpacks organized scholars',
'1st graders move groove with technology',
'literacy centers 2 0',
'creative critical thinking technology literacy chromebooks',
'ipads wanted cooperative learning environment',
'puppets performance',
'goldilocks trespasses understanding plot through adaptation examinations',
'folder frenzy',
'a tidy area better area learn in',
'student led conferences',
'share learning love',
'chromebooks curious minds',
'relaxing reading nook',
'technology research',
'materials for our learning centers sound like a winner',
'life cycles unit hatching chicks',
'we rhyme we repeat we learn read',
'organization collaborative space',
'dear diary help students express',
'project read part 2',
'chromebook math',
'bridging technology gap',
'technology kindergarten',
'crazy kindles',
'keeping our teeth clean our stomachs full',
'cubbies please',
'piano project producing proud performers',
'digital magazine',
'starting year off right foot',
'leaders techchology',
'reading classics today',
'digitalize my classroom',
'helping my students become upfront learners',
'let connect steam',
```

```
'identity the self portrait',
'a apple',
'the touch the feel shapes learning our lives',
'21st century skills technology optimized improve our world',
'the alamo supplemental reading',
'the future health medicine',
'empowering students through art moving full steam ahead',
'learning photography early age',
'happiness seeing hearing students read',
'4th graders need understand importance enviromental science',
'scientist need journals',
'books ahoy',
'blue seat sacks engaging books esol classroom',
'tools success',
'wiggle wiggle wiggle learn',
'seeking knowledge through technology',
'let hit target being active classroom',
'reads around world',
'learning through technology',
'never underestimate importance enough room work',
'extra extra third graders read all about it',
'osmo save day',
'math mania learn math better path',
'future mathematicians scientists',
'reading chairs',
'weighty word wizards',
'flexible seating classroom flexible minds control',
'i like move move',
'making makerspace part two',
'extra extra read all about it social justice readers',
'classroom supplies needed',
'through eyes doc cam',
'listen learn',
'yoga exercise',
'help us hear our tasks',
'basic needs keep 3rd graders healthy organized',
'technology tubergen tigers',
'i can',
'white boards supplies students with special needs',
'calligraphy no agenda',
'fly us moon astronomy lab supplies',
'it would be nice to see',
'survival resilience redemption',
'stem inspiration through literature',
'come along listen to the lullaby east la',
'listening center extraordinaire',
'bees flowers planets yippee',
'ipads titus talented team',
'initiate ipads',
'3d printing innovation lab',
'listening working wiping away workshop',
'technology reading please',
'recess relief',
'a kidney table small group instruction',
'bouncing off walls first grade',
'leap learn',
'stand deliver',
'supplies school year',
'student instruments',
'magnificent math',
'now showing scientific minds',
'balancing acts',
'story acting ells',
'flexible seating',
'comfy cozy reading bags',
'learning overcome sensory deficits through different textures',
'hands on science for tiny hands',
'ipad accessories multiage',
'bring learning life',
'organize our supplies please',
'slap shot sports',
'engaging bilingual learners maximizing classroom space',
'flexible seating flexible brains',
'technology today learners',
'books build brilliant brains',
'finding truth fiction',
'flexible seating working wonders 2nd graders',
```

```
'technology future',
'chromebooks enhance our learning',
'third graders protecting our environment',
'complete core complete kids',
'early chapter books more',
'kinders class needs safe place technology part 2',
'please help our students fulfill their need for speed',
'the last lecture middle school mantras',
'fostering love literature',
'making reading exciting technology',
'flexible seating first graders',
'seamlessly integrating technology esol curriculum',
'start right with art',
'book tastings book clubs',
'essential snack for hungry learners',
'plop down read',
'all that jazz',
'coding fun part 1',
'listening to books helps us learn understand',
'magazines assist fluency comprehension',
'readers live thousand lives turing 5th graders into bookworms',
'picture books that pop',
'a place learn grow',
'learning better reader ipads',
'teaching social justice through read alouds',
'increasing engagement technology',
'we need bullfrogs dissect please arcf sims',
'reading essentials',
'equitable access collaborate communicate chromebooks',
'centers needed pre k',
'21st century learners need chromebooks',
'1 2 3 eyes me',
'lady lancers basketball',
'everyday counts especially math',
'expanding learning',
'ipads library media center part ii',
'touch screen tablets computer science mathematics',
'shine light biology',
'to mars 2030',
'reading fun',
'seating success super heroes',
'healthy bodies healthy minds',
'engineering kindergarten',
'chromebooks 21st century classroom',
'virtual field trips kg kids',
'creating lifelong readers learners thinkers',
'technology for all the stars are the limit',
'staying indoor active with gonoodle',
'lego work',
'project leopard cub coding club part iv',
'wired sound',
'fidget cubes fidgety',
'economics to market to market to learn about our economy',
'hands math redesign',
'technology sets us free',
'technology pre k',
'ipad literacy math stations',
'a seat one seat',
'creative coding',
'project high',
'today readers tomorrow leaders',
'bee aware environment',
'beautiful you project',
'a new home growing turbo',
'kindle excitement',
'we want omnikin ball',
'shaping up new year',
'flexible seating',
'no weighting for fitness',
'start something great',
'excited about active learning',
'ilearn ipads',
'hands learning through technology',
'gopro cameras going green environmental filmmaking',
'growing garden',
'kindergarten learners on ipads',
'a new look for new year',
```

```
 'wiggle n read',
 'super students need super supplies success second grade',
 'focus pocus',
 ...]
```

## 1. 4 Preparing data for models

In [90]:

```
project_data.columns
```

Out[90]:

```
Index(['Unnamed: 0', 'id', 'teacher_id', 'teacher_prefix', 'school_state',
       'project_submitted_datetime', 'project_grade_category', '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',
       'clean_categories', 'clean_subcategories', 'essay', 'price',
       'quantity'],
      dtype='object')
```

we are going to consider

```
       - school_state : categorical data
       - clean_categories : categorical data
       - clean_subcategories : categorical data
       - project_grade_category : categorical data
       - teacher_prefix : categorical data

       - project_title : text data
       - text : text data
       - project_resource_summary: text data

       - quantity : numerical
       - teacher_number_of_previously_posted_projects : numerical
       - price : numerical
```

### 1.4.1 Vectorizing Categorical data

In [90]:

```
# to check all the defined variables
# https://stackoverflow.com/questions/633127/viewing-all-defined-variables
%whos
```

```
Variable                                Type           Data/Info
-------------------------------------------------------------------
Counter                                 type           <class 'collections.Counter'>
PrettyTable                             type           <class 'prettytable.PrettyTable'>
Processed_title                         list           n=109248
approved_price                          ndarray        92706: 92706 elems, type `float64`, 741648 by
es (724.265625 kb)
approved_word_count                     ndarray        92706: 92706 elems, type `int64`, 741648
bytes (724.265625 kb)
approved_word_count_essay               ndarray        92706: 92706 elems, type `int64`, 741648 byte
s (724.265625 kb)
cat_list                                list           n=109248
decontract                              function       <function decontract at 0x0000025D98824D08>
dict_sub                                dict           n=30
i                                       str            Classroom Tech to Develop 21st Century Leader
j                                       str            Mathematics
my_counter                              Counter        Counter({'Literacy_Langua<...>88,
'Care_Hunger': 1388})
my_counter_sub                          Counter        Counter({'Literacy': 3370<...>: 441, 'Economi
s': 269})
np                                      module         <module 'numpy' from 'C:\
<...>ges\\numpy\\__init__.py'>
```

```
pd                                    module          <module 'pandas' from 'C:
<...>es\\pandas\\__init__.py'>
plt                                   module          <module
'matplotlib.pyplo<...>\\matplotlib\\pyplot.py'>
previously_teacher_approved_projects  ndarray         92706: 92706 elems, type `int64`, 741648 byte
s (724.265625 kb)
previously_teacher_rejected_projects  ndarray         16542: 16542 elems, type `int64`, 132336 byte
s (129.234375 kb)
price                                 DataFrame                 id    price <...>[260115 rows x
columns]
project_data                          DataFrame            Unnamed: 0       <...>109248 rows x 2
columns]
proocessed_essay                      list            n=109248
re                                    module          <module 're' from 'C:\\Anaconda\\lib\\re.py'>
rejected_price                        ndarray         16542: 16542 elems, type `float64`, 132336 by
es (129.234375 kb)
rejected_word_count                   ndarray         109248: 109248 elems, type `int64`, 873984
bytes (853.5 kb)
rejected_word_count_essay             ndarray         16542: 16542 elems, type `int64`, 132336 byte
s (129.234375 kb)
resources_data                        DataFrame                id         <...>1541272 rows x
columns]
sent                                  str             Classroom Tech Develop 21st Century Leaders
sns                                   module          <module 'seaborn' from
'C<...>s\\seaborn\\__init__.py'>
stack_plot                            function        <function stack_plot at 0x0000025DF358AD08>
stopwords                             list            n=176
sub_cat_list                          list            n=109248
temp                                  str             College_CareerPrep Mathematics
tqdm                                  type            <class 'tqdm._tqdm.tqdm'>
univariate_barplots                   function        <function univariate_barp<...>ts at
0x0000025DF358AEA0>
value_counts                          Series          1    92706\n0    16542\nN<...>is_approved, dt
pe: int64
warnings                              module          <module 'warnings' from
'<...>conda\\lib\\warnings.py'>
word_count                            Series          4     19979\n5     19677\<...>oject_title, dt
pe: int64
word_count_summary_approved           ndarray         92706: 92706 elems, type `int64`, 741648 byte
s (724.265625 kb)
word_count_summary_rejected           ndarray         16542: 16542 elems, type `int64`, 132336 byte
s (129.234375 kb)
word_dict                             dict            n=13
words_count_approved                  DataFrame            Unnamed: 0       <...>[92706 rows x 1
columns]
x                                     PrettyTable     +------------+-----------<...>----+----------
------+
xxx                                   list            n=109248
xxx_1                                 DataFrame         unique_subject_categori<...>
Literacy_Language  52239
xxx_2                                 DataFrame          cleaned_sub_categories<...>          Liter
y  33700
xxx_3                                 DataFrame           number of words in pr<...>
4  19979
```

In [91]:

```
# we use count vectorizer to convert the values into one hot encoded features
from sklearn.feature_extraction.text import CountVectorizer
vectorizer = CountVectorizer(cat_list, lowercase=False, binary=True)
vectorizer.fit(project_data['clean_categories'].values)
print(vectorizer.get_feature_names)
```

```
<bound method CountVectorizer.get_feature_names of CountVectorizer(analyzer='word', binary=True, d
ecode_error='strict',
        dtype=<class 'numpy.int64'>, encoding='utf-8',
        input=['Literacy_Language', 'History_Civics Health_Sports', 'Health_Sports', 'Literacy_Lang
uage Math_Science', 'Math_Science', 'Literacy_Language SpecialNeeds', 'Literacy_Language
SpecialNeeds', 'Math_Science', 'Health_Sports', 'Literacy_Language', 'Literacy_Language',
'Literacy_Language AppliedLear...ce', 'Literacy_Language Math_Science', 'Health_Sports
SpecialNeeds', 'AppliedLearning Math_Science'],
        lowercase=False, max_df=1.0, max_features=None, min_df=1,
        ngram_range=(1, 1), preprocessor=None, stop_words=None,
        strip_accents=None, token_pattern='(?u)\\b\\w\\w+\\b',
        tokenizer=None, vocabulary=None)>
```

In [92]:

```
categories_one_hot = vectorizer.transform(project_data['clean_categories'].values)
print('Shape after one hot encoding of features', categories_one_hot.shape)
```

```
Shape after one hot encoding of features (109248, 9)
```

In [93]:

```
#For subcategories
vectorizer = CountVectorizer(sub_cat_list, lowercase=False, binary=True)
vectorizer.fit(project_data['clean_subcategories'].values)
print(vectorizer.get_feature_names)
```

```
<bound method CountVectorizer.get_feature_names of CountVectorizer(analyzer='word', binary=True, d
ecode_error='strict',
        dtype=<class 'numpy.int64'>, encoding='utf-8',
        input=['ESL Literacy', 'Civics_Government TeamSports', 'Health_Wellness TeamSports', 'Liter
acy Mathematics', 'Mathematics', 'Literature_Writing SpecialNeeds', 'Literacy SpecialNeeds',
'Mathematics', 'Health_Wellness', 'Literacy Literature_Writing', 'Literacy', 'Literacy
ParentInvolvement', 'Environm...hematics', 'Literacy Mathematics', 'Health_Wellness SpecialNeeds',
'College_CareerPrep Mathematics'],
        lowercase=False, max_df=1.0, max_features=None, min_df=1,
        ngram_range=(1, 1), preprocessor=None, stop_words=None,
        strip_accents=None, token_pattern='(?u)\\b\\w\\w+\\b',
        tokenizer=None, vocabulary=None)>
```

In [94]:

```
sub_categories_one_hot = vectorizer.transform(project_data['clean_subcategories'].values)
print('the shape of the sub categories after one hot encoding', sub_categories_one_hot.shape)
```

```
the shape of the sub categories after one hot encoding (109248, 30)
```

## For School State

In [95]:

```
project_data.head(2)
```

Out[95]:

| | Unnamed: 0 | id | teacher_id | teacher_prefix | school_state | project_submitted_datetime | pro |
|---|---|---|---|---|---|---|---|
| 0 | 160221 | p253737 | c90749f5d961ff158d4b4d1e7dc665fc | Mrs. | IN | 2016-12-05 13:43:57 | Gra |
| 1 | 140945 | p258326 | 897464ce9ddc600bced1151f324dd63a | Mr. | FL | 2016-10-25 09:22:10 | Gra |

In [96]:

```
vectorizer = CountVectorizer(list(project_data['school_state'].values), lowercase=False, binary=Tru
e)
```

```
vectorizer.fit(project_data['school_state'].values)
print(vectorizer.get_feature_names)

school_state_onehot_encoded = vectorizer.transform(project_data['school_state'].values)
print('the shape of the state after onehot encoded', school_state_onehot_encoded.shape)
```

```
<bound method CountVectorizer.get_feature_names of CountVectorizer(analyzer='word', binary=True, d
ecode_error='strict',
        dtype=<class 'numpy.int64'>, encoding='utf-8',
        input=['IN', 'FL', 'AZ', 'KY', 'TX', 'FL', 'CT', 'GA', 'SC', 'NC', 'CA', 'CA', 'NY', 'OK',
'MA', 'TX', 'FL', 'NV', 'GA', 'OH', 'PA', 'NC', 'CA', 'AL', 'FL', 'AL', 'TX', 'LA', 'GA', 'VA', 'IN
', 'NC', 'NC', 'AR', 'CA', 'NY', 'WA', 'TX', 'CA', 'FL', 'CA', 'OK', 'WV', 'NV', 'LA', 'ID', 'TX',
'TN', 'CT',...AZ', 'MD', 'AZ', 'NY', 'TX', 'OH', 'IN', 'WI', 'MN', 'MD', 'MD', 'SC', 'MO', 'NJ', 'N
J', 'NY', 'VA'],
        lowercase=False, max_df=1.0, max_features=None, min_df=1,
        ngram_range=(1, 1), preprocessor=None, stop_words=None,
        strip_accents=None, token_pattern='(?u)\\b\\w\\w+\\b',
        tokenizer=None, vocabulary=None)>
the shape of the state after onehot encoded (109248, 51)
```

## For teacher prefix

In [97]:

```
project_data.head(2)
```

Out[97]:

| | Unnamed: 0 | id | teacher_id | teacher_prefix | school_state | project_submitted_datetime | pro |
|---|---|---|---|---|---|---|---|
| **0** | 160221 | p253737 | c90749f5d961ff158d4b4d1e7dc665fc | Mrs. | IN | 2016-12-05 13:43:57 | Gra |
| **1** | 140945 | p258326 | 897464ce9ddc600bced1151f324dd63a | Mr. | FL | 2016-10-25 09:22:10 | Gra |

In [98]:

```
# https://www.geeksforgeeks.org/working-with-missing-data-in-pandas/
project_data[project_data['teacher_prefix'].isnull()]
```

Out[98]:

| | Unnamed: 0 | id | teacher_id | teacher_prefix | school_state | project_submitted_datetime |
|---|---|---|---|---|---|---|
| **7820** | 17809 | p180947 | 834f75f1b5e24bd10abe9c3dbf7ba12f | NaN | CA | 2016-11-04 00:15:45 |
| **30368** | 22174 | p002730 | 339bd5a9e445d68a74d65b99cd325397 | NaN | SC | 2016-05-09 09:38:40 |

| | Unnamed: 0 | id | teacher_id | teacher_prefix | school_state | project_submitted_datetime |
|---|---|---|---|---|---|---|
| **57654** | 158692 | p197901 | e4be6aaaa887d4202df2b647fbfc82bb | NaN | PA | 2016-06-03 10:15:05 |

We can see that there three NaN values in teacher prefix and we can remove that because there is only three rows in it and it is insignificant when we compare with total number of rows

In [99]:

```
project_data.dropna(axis=0, inplace=True)
```

In [100]:

```
project_data.isnull().sum()
```

Out[100]:

```
Unnamed: 0                                      0
id                                              0
teacher_id                                      0
teacher_prefix                                  0
school_state                                    0
project_submitted_datetime                      0
project_grade_category                          0
project_title                                   0
project_essay_1                                 0
project_essay_2                                 0
project_essay_3                                 0
project_essay_4                                 0
project_resource_summary                        0
teacher_number_of_previously_posted_projects    0
project_is_approved                             0
clean_categories                                0
clean_subcategories                             0
essay                                           0
price                                           0
quantity                                        0
dtype: int64
```

We can see that there is no more missing values in any columns

In [101]:

```
vectorizer = CountVectorizer(list(project_data['teacher_prefix'].values), lowercase=False, binary=True)
vectorizer.fit(project_data['teacher_prefix'].values)
print(vectorizer.get_feature_names)

teacher_prefix_onehot_encoded = vectorizer.transform(project_data['teacher_prefix'].values)
print('='*50)
print('the shape of the teacher prefix after one hot encoded', teacher_prefix_onehot_encoded.shape)
```

```
<bound method CountVectorizer.get_feature_names of CountVectorizer(analyzer='word', binary=True, decode_error='strict',
        dtype=<class 'numpy.int64'>, encoding='utf-8',
        input=['Mrs.', 'Ms.', 'Ms.', 'Mr.', 'Ms.', 'Mrs.', 'Ms.', 'Mrs.', 'Ms.', 'Ms.', 'Ms.',
'Mrs.', 'Mr.', 'Mrs.', 'Mrs.', 'Mrs.', 'Mrs.', 'Mr.', 'Ms.', 'Ms.', 'Mrs.', 'Mrs.', 'Mrs.',
'Mrs.', 'Mrs.', 'Mrs.', 'Ms.', 'Ms.', 'Mrs.', 'Mrs.', 'Mrs.', 'Mrs.', 'Mrs.', 'Mrs.', 'Mr.',
'Mrs.', 'Ms.', 'Mrs.', 'Mr.'..., 'Mrs.', 'Ms.', 'Mr.', 'Ms.', 'Mrs.', 'Mrs.', 'Mrs.', 'Ms.', 'Mrs.
', 'Mrs.', 'Ms.', 'Mrs.', 'Mrs.'],
        lowercase=False, max_df=1.0, max_features=None, min_df=1,
        ngram_range=(1, 1), preprocessor=None, stop_words=None,
        strip_accents=None, token_pattern='(?u)\\b\\w\\w+\\b',
        tokenizer=None, vocabulary=None)>
==================================================
the shape of the teacher prefix after one hot encoded (3757, 4)
```

the shape of the teacher prefix after one hot encoded (5757, 4)

## For project grade category

In [104]:

```
project_data.head(2)
```

Out[104]:

| | Unnamed: 0 | id | teacher_id | teacher_prefix | school_state | project_submitted_datetime | p |
|---|---|---|---|---|---|---|---|
| **61** | 175822 | p046845 | 97f18c16914244c16db8a02260b2b488 | Mrs. | SC | 2016-05-03 18:39:03 | G |
| **92** | 167398 | p002482 | c27352b1817b956e2bd722897d9a6552 | Ms. | HI | 2016-04-29 16:21:09 | G |

In [105]:

```
project_data['project_grade_category'].unique()
```

Out[105]:

```
array(['Grades PreK-2', 'Grades 3-5', 'Grades 9-12', 'Grades 6-8'],
      dtype=object)
```

We can see that there are only 4 categories in project grade category

In [106]:

```
vectorizer = CountVectorizer(list(project_data['project_grade_category'].values), lowercase=False,
binary=True)
vectorizer.fit(project_data['project_grade_category'].values)
print(vectorizer.get_feature_names)

print('='*50)
grade_onehot_encoded = vectorizer.transform(project_data['project_grade_category'].values)
print('the shape of the matrix after one hot encoding of project grade
category',grade_onehot_encoded)
```

```
<bound method CountVectorizer.get_feature_names of CountVectorizer(analyzer='word', binary=True, d
ecode_error='strict',
        dtype=<class 'numpy.int64'>, encoding='utf-8',
        input=['Grades PreK-2', 'Grades 3-5', 'Grades 9-12', 'Grades 9-12', 'Grades 3-5', 'Grades P
reK-2', 'Grades 3-5', 'Grades PreK-2', 'Grades PreK-2', 'Grades PreK-2', 'Grades 9-12', 'Grades
PreK-2', 'Grades 6-8', 'Grades 3-5', 'Grades PreK-2', 'Grades 9-12', 'Grades 3-5', 'Grades 6-8', 'G
rades 6-8', 'G...ades 6-8', 'Grades PreK-2', 'Grades PreK-2', 'Grades 3-5', 'Grades 3-5', 'Grades 6
-8', 'Grades 3-5'],
        lowercase=False, max_df=1.0, max_features=None, min_df=1,
        ngram_range=(1, 1), preprocessor=None, stop_words=None,
        strip_accents=None, token_pattern='(?u)\\b\\w\\w+\\b',
        tokenizer=None, vocabulary=None)>
==================================================
the shape of the matrix after one hot encoding of project grade category   (0, 1) 1
  (0, 2) 1
  (1, 1) 1
  (2, 0) 1
  (2, 1) 1
  (3, 0) 1
  (3, 1) 1
  (4, 1) 1
  (5, 1) 1
```

```
  (5, 2) 1
  (6, 1) 1
  (7, 1) 1
  (7, 2) 1
  (8, 1) 1
  (8, 2) 1
  (9, 1) 1
  (9, 2) 1
  (10, 0) 1
  (10, 1) 1
  (11, 1) 1
  (11, 2) 1
  (12, 1) 1
  (13, 1) 1
  (14, 1) 1
  (14, 2) 1
  : :
  (3736, 1) 1
  (3737, 1) 1
  (3738, 1) 1
  (3739, 1) 1
  (3740, 1) 1
  (3741, 1) 1
  (3742, 1) 1
  (3743, 1) 1
  (3743, 2) 1
  (3744, 1) 1
  (3744, 2) 1
  (3745, 1) 1
  (3746, 1) 1
  (3747, 1) 1
  (3748, 1) 1
  (3749, 1) 1
  (3750, 1) 1
  (3751, 1) 1
  (3751, 2) 1
  (3752, 1) 1
  (3752, 2) 1
  (3753, 1) 1
  (3754, 1) 1
  (3755, 1) 1
  (3756, 1) 1
```

## 1.4.2 Vectorizing Text data

#### 1.4.2.1 Bag of words

In [105]:

```
%whos
```

```
Variable                                Type               Data/Info
------------------------------------------------------------------------
CountVectorizer                         type               <class
'sklearn.feature_e<...>on.text.CountVectorizer'>
Counter                                 type               <class 'collections.Counter'>
PrettyTable                             type               <class 'prettytable.PrettyTable'>
Processed_title                         list               n=109248
approved_price                          ndarray            92706: 92706 elems, type `float64`, 74164
bytes (724.265625 kb)
approved_word_count                     ndarray            92706: 92706 elems, type `int64`, 741648
ytes (724.265625 kb)
approved_word_count_essay               ndarray            92706: 92706 elems, type `int64`, 741648
bytes (724.265625 kb)
cat_list                                list               n=109248
categories_one_hot                      csr_matrix           (0, 4) 1\n  (1, 2) 1\n <...>47, 0) 1\n
(109247, 5) 1
decontract                              function           <function decontract at
0x0000025D98824D08>
dict_sub                                dict               n=30
grade_onehot_encoded                    csr_matrix           (0, 1) 1\n  (0, 2) 1\n <...>3755, 1)
1\n  (3756, 1) 1
i                                       str                Classroom Tech to Develop 21st Century Le
```

```
ers
j                                       str                Mathematics
my_counter                              Counter            Counter({'Literacy_Langua<...>88,
'Care_Hunger': 1388})
my_counter_sub                          Counter            Counter({'Literacy': 3370<...>: 441, 'Ecc
omics': 269})
np                                      module             <module 'numpy' from 'C:\
<...>ges\\numpy\\__init__.py'>
pd                                      module             <module 'pandas' from 'C:
<...>es\\pandas\\__init__.py'>
plt                                     module             <module
'matplotlib.pyplo<...>\\matplotlib\\pyplot.py'>
previously_teacher_approved_projects    ndarray            92706: 92706 elems, type `int64`, 741648
bytes (724.265625 kb)
previously_teacher_rejected_projects    ndarray            16542: 16542 elems, type `int64`, 132336
bytes (129.234375 kb)
price                                   DataFrame                     id    price <...>[260115 row
x 3 columns]
project_data                            DataFrame                Unnamed: 0       <...>n[3757 rows
x 20 columns]
proocessed_essay                        list               n=109248
re                                      module             <module 're' from
'C:\\Anaconda\\lib\\re.py'>
rejected_price                          ndarray            16542: 16542 elems, type `float64`, 13233
bytes (129.234375 kb)
rejected_word_count                     ndarray            109248: 109248 elems, type `int64`, 87398
bytes (853.5 kb)
rejected_word_count_essay               ndarray            16542: 16542 elems, type `int64`, 132336
bytes (129.234375 kb)
resources_data                          DataFrame                      id         <...>1541272 row
x 4 columns]
school_state_onehot_encoded             csr_matrix           (0, 15) 1\n  (1, 9) 1\n<...>, 34) 1\n
(109247, 45) 1
sent                                    str                Classroom Tech Develop 21st Century
Leaders
sns                                     module             <module 'seaborn' from
'C<...>s\\seaborn\\__init__.py'>
stack_plot                              function           <function stack_plot at
0x0000025DF358AD08>
stopwords                               list               n=176
sub_cat_list                            list               n=109248
sub_categories_one_hot                  csr_matrix           (0, 6) 1\n  (0, 17) 1\n<...>7, 4) 1\n
(109247, 19) 1
teacher_prefix_onehot_encoded           csr_matrix           (0, 1) 1\n  (1, 2) 1\n <...>3755, 1) 1\
n  (3756, 1) 1
temp                                    str                College_CareerPrep Mathematics
tqdm                                    type               <class 'tqdm._tqdm.tqdm'>
univariate_barplots                     function           <function univariate_barp<...>ts at 0x000
025DF358AEA0>
value_counts                            Series             1    92706\n0    16542\nN<...>is_approvec
dtype: int64
vectorizer                              CountVectorizer    CountVectorizer(analyzer=<...>er=None, vc
abulary=None)
warnings                                module             <module 'warnings' from
'<...>conda\\lib\\warnings.py'>
word_count                              Series             4     19979\n5     19677\<...>oject_title
dtype: int64
word_count_summary_approved             ndarray            92706: 92706 elems, type `int64`, 741648
bytes (724.265625 kb)
word_count_summary_rejected             ndarray            16542: 16542 elems, type `int64`, 132336
bytes (129.234375 kb)
word_dict                               dict               n=13
words_count_approved                    DataFrame                Unnamed: 0       <...>[92706 rows
x 17 columns]
x                                       PrettyTable        +------------+-----------<...>----+------
----------+
xxx                                     list               n=109248
xxx_1                                   DataFrame            unique_subject_categori<...>
Literacy_Language  52239
xxx_2                                   DataFrame             cleaned_sub_categories<...>          I
eracy  33700
xxx_3                                   DataFrame              number of words in pr<...>
4  19979
```

In [107]:

```
# We are considering only the words which appeared in at least 10 documents(rows or projects).
vectorizer = CountVectorizer(min_df=10)
text_bow_essay = vectorizer.fit_transform(proocessed_essay)
print('Shape of the matrix after one hot encoding', text_bow_essay.shape)
```

```
Shape of the matrix after one hot encoding (109248, 16647)
```

#### 1.4.2.2 Bag of Words on `project_title`

In [108]:

```
vectorizer = CountVectorizer(min_df=10)
text_bow_title = vectorizer.fit_transform(Processed_title)
print('The shape of the matrix after one hot encoding', text_bow_title.shape)
```

```
The shape of the matrix after one hot encoding (109248, 3335)
```

#### 1.4.2.3 TFIDF vectorizer

In [109]:

```
#For preprocessed essay
from sklearn.feature_extraction.text import TfidfVectorizer
vectorizer = TfidfVectorizer(min_df=10)
text_tfidf= vectorizer.fit_transform(proocessed_essay)
print('Shape of the matrix after one hot encoding', text_tfidf.shape)
```

```
Shape of the matrix after one hot encoding (109248, 16647)
```

#### 1.4.2.4 TFIDF Vectorizer on `project_title`

In [110]:

```
vectorizer = TfidfVectorizer(min_df=10)
text_tfidf = vectorizer.fit_transform(Processed_title)
print('Shape of the matrix after TFIDF', text_tfidf.shape)
```

```
Shape of the matrix after TFIDF (109248, 3335)
```

#### 1.4.2.5 Using Pretrained Models: Avg W2V

In [112]:

```
# Reading glove vectors in python: https://stackoverflow.com/a/38230349/4084039
def loadGloveModel(gloveFile):
    print ("Loading Glove Model")
    f = open(gloveFile,'r', encoding="utf8")
    model = {}
    for line in tqdm(f):
        splitLine = line.split()
        word = splitLine[0]
        embedding = np.array([float(val) for val in splitLine[1:]])
        model[word] = embedding
    print ("Done.",len(model)," words loaded!")
    return model
model = loadGloveModel(r'D:\Others\Pictures\model downloads\Glove\glove.42B.300d.txt')
```

```
Loading Glove Model
```

```
1917494it [05:59, 5337.79it/s]
```

```
Done. 1917494  words loaded!
```

In [113]:

```
words = []
for i in proocessed_essay:
    words.extend(i.split(' '))

for i in Processed_title:
    words.extend(i.split(' '))

print('the length of the corpus', len(words))
words = set(words)
print('The unique words in the corpus', len(words))
```

```
the length of the corpus 17012070
The unique words in the corpus 59180
```

In [114]:

```
#Intersected words
inter_words = set(model.keys()).intersection(words)
print("The number of words that are present in both glove vectors and our coupus", \
      len(inter_words),"(" ,np.round(len(inter_words)/len(words)*100, 3),"%)")
```

```
The number of words that are present in both glove vectors and our coupus 51533 ( 87.078 %)
```

In [115]:

```
word_corpus = {}
word_glove = set(model.keys())

for i in words:
    if i in word_glove:
        word_corpus[i] = model[i]
print('the word2vec length', len(word_corpus))
```

```
the word2vec length 51533
```

In [116]:

```
#Pickling the model http://www.jessicayung.com/how-to-use-pickle-to-save-and-load-variables-in-pyt
hon/
import pickle
with open('glove_vectors', 'wb') as f:
    pickle.dump(word_corpus, f)
```

In [117]:

```
#Unpickling
with open('glove_vectors', 'rb') as f:
    model = pickle.load(f)
    glove_words = set(model.keys())
```

In [118]:

```
#Avgw2V vector for preprcessed essay

avg_w2v_vector = []

for i in tqdm(proocessed_essay):
    vector = np.zeros(300)
    cnt_words = 0

    for word in i.split():
        if word in glove_words:
            vector += model[word]
            cnt_words +=1

    if cnt_words !=0:
            vector /= cnt_words

    avg_w2v_vector.append(vector)
```

```
print(len(avg_w2v_vector))
print(len(avg_w2v_vector[0]))
```

```
100%|████████████████████████████████████████████████████████████████| 109248/109248
[01:06<00:00, 1640.91it/s]
```

```
109248
300
```

#### 1.4.2.6 Using Pretrained Models: AVG W2V on `project_title`

In [119]:

```
#avg_w2v for processed title

avg_W2V_vectors = [];
for sentence in tqdm(Processed_title):
    vector = np.zeros(300)
    cnt_words =0;
    for word in sentence.split():
        if word in glove_words:
            vector += model[word]
            cnt_words += 1
    if cnt_words != 0:
        vector /= cnt_words
    avg_W2V_vectors.append(vector)

print(len(avg_W2V_vectors))
print(len(avg_W2V_vectors[0]))
```

```
100%|████████████████████████████████████████████████████████████████| 109248/109248
[00:03<00:00, 28745.03it/s]
```

```
109248
300
```

#### 1.4.2.7 Using Pretrained Models: TFIDF weighted W2V

In [120]:

```
Tf_idf_model = TfidfVectorizer()
Tf_idf_model.fit(proocessed_essay)

dictionary = dict(zip(Tf_idf_model.get_feature_names(), Tf_idf_model.idf_))
tf_idf_words = set(Tf_idf_model.get_feature_names())
```

In [121]:

```
# TFIDF Word2Vec
# compute TFIDF word2vec for each review.
tfidf_w2v_vectors = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(proocessed_essay): # for each review/sentence
    vector = np.zeros(300) # as word vectors are of zero length
    tf_idf_weight =0; # num of words with a valid vector in the sentence/review
    for word in sentence.split(): # for each word in a review/sentence
        if (word in glove_words) and (word in tf_idf_words):
            vec = model[word] # getting the vector for each word
            # here we are multiplying idf value(dictionary[word]) and the tf
value((sentence.count(word)/len(sentence.split())))
            tf_idf = dictionary[word]*(sentence.count(word)/len(sentence.split())) # getting the tf
idf value for each word
            vector += (vec * tf_idf) # calculating tfidf weighted w2v
            tf_idf_weight += tf_idf
    if tf_idf_weight != 0:
        vector /= tf_idf_weight
    tfidf_w2v_vectors.append(vector)

print(len(tfidf_w2v_vectors))
```

```
print(len(tfidf_w2v_vectors[0]))
```

```
100%|████████████████████████████████████████████████████████████| 109248/109248 [05:06<00:00, 356.08it/s]
```

```
109248
300
```

#### 1.4.2.9 Using Pretrained Models: TFIDF weighted W2V on `project_title`

In [122]:

```
# TFIDF Word2Vec
# compute TFIDF word2vec for each review.
tfidf_W2V_vectors = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(Processed_title): # for each review/sentence
    vector = np.zeros(300) # as word vectors are of zero length
    tf_idf_weight =0; # num of words with a valid vector in the sentence/review
    for word in sentence.split(): # for each word in a review/sentence
        if (word in glove_words) and (word in tf_idf_words):
            vec = model[word] # getting the vector for each word
            # here we are multiplying idf value(dictionary[word]) and the tf
value((sentence.count(word)/len(sentence.split())))
            tf_idf = dictionary[word]*(sentence.count(word)/len(sentence.split())) # getting the tf
idf value for each word
            vector += (vec * tf_idf) # calculating tfidf weighted w2v
            tf_idf_weight += tf_idf
    if tf_idf_weight != 0:
        vector /= tf_idf_weight
    tfidf_W2V_vectors.append(vector)

print(len(tfidf_W2V_vectors))
print(len(tfidf_W2V_vectors[0]))
```

```
100%|████████████████████████████████████████████████████████████| 109248/109248 [00:06<00:00, 16855.18it/s]
```

```
109248
300
```

### 1.4.3 Vectorizing Numerical features

In [192]:

```
project_data['price'].shape
```

Out[192]:

```
(109248,)
```

In [193]:

```
# check this one: https://www.youtube.com/watch?v=0HOqOcln3Z4&t=530s
# standardization sklearn: https://scikit-
learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html
from sklearn.preprocessing import StandardScaler

# price_standardized = standardScalar.fit(project_data['price'].values)
# this will rise the error
# ValueError: Expected 2D array, got 1D array instead: array=[725.05 213.03 329.   ... 399.   287.
73   5.5 ].
# Reshape your data either using array.reshape(-1, 1)

price_scalar = StandardScaler()
price_scalar.fit(project_data['price'].values.reshape(-1,1)) # finding the mean and standard
deviation of this data
print(f"Mean : {price_scalar.mean_[0]}, Standard deviation : {np.sqrt(price_scalar.var_[0])}")

# Now standardize the data with above maen and variance.
price_standardized = price_scalar.transform(project_data['price'].values.reshape(-1, 1))
```

```
price_standardized = price_scalar.transform(project_data['price'].values.reshape(-1, 1))
```

```
Mean : 298.1193425966608, Standard deviation : 367.49634838483496
```

In [194]:

```
price_standardised = price_scalar.transform(project_data['price'].values.reshape(-1,1))
```

In [195]:

```
price_standardised
```

Out[195]:

```
array([[-0.3905327 ],
       [ 0.00239637],
       [ 0.59519138],
       ...,
       [-0.15825829],
       [-0.61243967],
       [-0.51216657]])
```

### 1.4.4 Merging all the above features

- we need to merge all the numerical vectors i.e catogorical, text, numerical vectors

In [196]:

```
print(categories_one_hot.shape)
print(sub_categories_one_hot.shape)
print(text_bow_essay.shape)
print(price_standardised.shape)
```

```
(109248, 9)
(109248, 30)
(109248, 16647)
(109248, 1)
```

### Merging the BOW for processed essays

In [246]:

```
# merge two sparse matrices: https://stackoverflow.com/a/19710648/4084039
from scipy.sparse import hstack
X = hstack((categories_one_hot, sub_categories_one_hot, text_bow_essay, price_standardised))
```

In [247]:

```
print(X.shape)
```

```
(109248, 16687)
```

# Assignment 2: Apply TSNE

## 2.1 TSNE with `BOW` encoding of `project_title` feature

In [248]:

```
print(categories_one_hot.shape)
print(sub_categories_one_hot.shape)
print(text_bow_title.shape)
print(price_standardised.shape)
```

```
(109248, 9)
(109248, 30)
(109248, 3335)
(109248, 1)
```

In [285]:

```
#Merging all features
from scipy.sparse import hstack
X = hstack((categories_one_hot, sub_categories_one_hot, text_bow_title, price_standardised))
print(X.shape)
```

```
(109248, 3375)
```

In [286]:

```
type(X)
```

Out[286]:

```
scipy.sparse.coo.coo_matrix
```

In [288]:

```
#converting sparse to dense matrix using X.toarray()
aaa = X.toarray()
```

In [289]:

```
aaa.shape
```

Out[289]:

```
(109248, 3375)
```

In [293]:

```
#Considering top 5k points
X = aaa[0:5000,:]
```

In [333]:

```
X.shape
```

Out[333]:

```
(5000, 3375)
```

In [332]:

```
X
```

Out[332]:

```
array([[ 0.        ,  0.        ,  0.        , ...,  0.        ,
         0.        , -0.3905327 ],
       [ 0.        ,  0.        ,  1.        , ...,  0.        ,
         0.        ,  0.00239637],
       [ 0.        ,  0.        ,  1.        , ...,  0.        ,
         0.        ,  0.59519138],
       ...,
       [ 0.        ,  0.        ,  0.        , ...,  0.        ,
         0.        , -0.49749975],
       [ 0.        ,  0.        ,  0.        , ...,  0.        ,
         0.        , -0.34707649],
       [ 0.        ,  0.        ,  1.        , ...,  0.        ,
         0.        , -0.70245417]])
```

In [335]:

```
#Pickling X for future purpose
import pickle
with open('X', 'wb') as f:
    pickle.dump(X,f)
```

In [295]:

```
#Building TSNE
from sklearn.manifold import TSNE
tsne = TSNE(n_components=2, perplexity=30, learning_rate=200)

X_embedding = tsne.fit_transform(X)
```

In [213]:

```
#Taking project is approved as y
y = project_data['project_is_approved']
```

In [214]:

```
#considering only 5k points
y = y[:5000]
```

In [215]:

```
#since y is in series we can't reshape it so we are converting it into array
y = np.asarray(y)
```

In [216]:

```
y.shape
```

Out[216]:

```
(5000,)
```

In [307]:

```
#concatinating X_emedding and y
for_tnse = np.hstack((X_embedding, y.reshape(-1,1)))
```

In [308]:

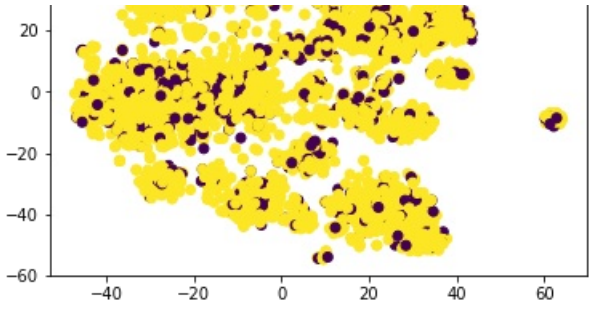
```
for_tsne
```

Out[308]:

```
array([[-34.17054367,   8.79829693,   0.        ],
       [ 36.10246277, -50.48965454,   1.        ],
       [ 36.50696564, -46.2404213 ,   0.        ],
       ...,
       [-27.52528763,  -1.39240873,   0.        ],
       [-29.89862061, -26.57378197,   1.        ],
       [ 33.15214157, -50.2638092 ,   1.        ]])
```

In [310]:

```
for_tsne_df = pd.DataFrame(for_tnse, columns=['Dimension_x', 'Dimension_y', 'project_is_approved'])
plt.scatter(for_tsne_df['Dimension_x'], for_tsne_df['Dimension_y'],
c=for_tsne_df['project_is_approved'])
plt.show()
```




## 2.2 TSNE with `TFIDF` encoding of `project_title` feature

In [251]:

```
print(categories_one_hot.shape)
print(sub_categories_one_hot.shape)
print(text_tfidf.shape)
print(price_standardised.shape)
```

```
(109248, 9)
(109248, 30)
(109248, 3335)
(109248, 1)
```

In [311]:

```
from scipy.sparse import hstack
X_1 = hstack((categories_one_hot, sub_categories_one_hot, text_tfidf, price_standardised))
print(X_1.shape)
```

```
(109248, 3375)
```

In [313]:

```
#converting sparse to dense matrix using X.toarray()
bbb = X_1.toarray()
```

In [314]:

```
bbb.shape
```

Out[314]:

```
(109248, 3375)
```

In [315]:

```
X_1 = bbb[:5000, :]
print(X_1.shape)
```

```
(5000, 3375)
```

In [336]:

```
#Pickling the X_1
with open('X_1', 'wb') as f:
    pickle.dump(X_1, f)
```

In [316]:

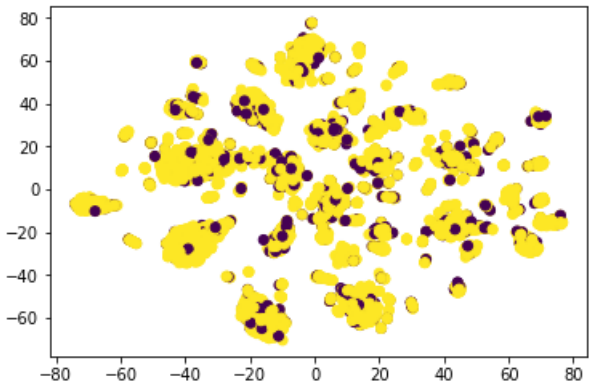
```
#Building TsNE
tsne_1 = TSNE(n_components=2, perplexity=30, learning_rate=200)
X_embedding_1 = tsne_1.fit_transform(X_1)
```

In [317]:

```
#concatinating X_emedding and y
for_tnse_1 = np.hstack((X_embedding_1, y.reshape(-1,1)))
```

In [318]:

```
for_tsne_df_1 = pd.DataFrame(for_tnse_1, columns=['Dimension_x', 'Dimension_y', 'project_is_approve
d'])
plt.scatter(for_tsne_df_1['Dimension_x'], for_tsne_df_1['Dimension_y'],
c=for_tsne_df_1['project_is_approved'])
plt.show()
```




## 2.3 TSNE with `AVG W2V` encoding of `project_title` feature

In [260]:

```
print(categories_one_hot.shape)
print(sub_categories_one_hot.shape)
print(len(avg_W2V_vectors))
print(price_standardised.shape)
```

```
(109248, 9)
(109248, 30)
109248
(109248, 1)
```

In [270]:

```
#Converting the avg_W2V vector into array
avg_W2V_vectors = (np.asarray(avg_W2V_vectors))
```

In [272]:

```
print(categories_one_hot.shape)
print(sub_categories_one_hot.shape)
print(avg_W2V_vectors.shape)
print(price_standardised.shape)
```

```
(109248, 9)
(109248, 30)
(109248, 300)
(109248, 1)
```

In [199]:

```
#Merging all the features
from scipy.sparse import hstack
X_3 = hstack((categories_one_hot, sub_categories_one_hot, avg_W2V_vectors, price_standardised))
print(X_3.shape)
```

```
(109248, 340)
```

In [200]:

```
X_3 = X_3.toarray()
```

In [201]:

```
X_3 = X_3[:5000,:]
print(X_3.shape)
```

```
(5000, 340)
```

In [202]:

```
#dumping it
with open('X_3', 'wb') as f:
    pickle.dump(X_3, f)
```

In [236]:

```
#Building TSNE
from sklearn.manifold import TSNE
tsne_3 = TSNE(n_components=2, perplexity=500, learning_rate=750)
X_embedding_3 = tsne_3.fit_transform(X_3)
```

In [218]:

```
#concatination X_embedding and y
#for_tnse_1 = np.hstack((X_embedding_1, y.reshape(-1,1)))
for_tsne_3 = np.hstack((X_embedding_3, y.reshape(-1,1)))
```

In [220]:

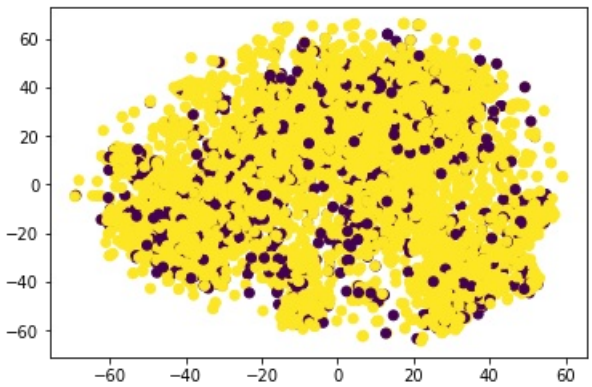
```
for_tsne_3.shape
```

Out[220]:

```
(5000, 3)
```

In [221]:

```
for_tsne_df_3 = pd.DataFrame(for_tsne_3, columns=['Dimension_x', 'Dimension_y', 'project_is_approve
d'])
plt.scatter(for_tsne_df_3['Dimension_x'], for_tsne_df_3['Dimension_y'],
c=for_tsne_df_3['project_is_approved'])
plt.show()
```




## 2.4 TSNE with `TFIDF Weighted W2V` encoding of `project_title` feature

In [203]:

```
print(categories_one_hot.shape)
print(sub_categories_one_hot.shape)
print(len(tfidf_W2V_vectors))
print(price_standardised.shape)
```

```
(109248, 9)
(109248, 30)
109248
(109248, 1)
```

In [204]:

```
#Converting the tfidf_W2V vector into array
tfidf_W2V_vectors = (np.asarray(tfidf_W2V_vectors))
```

In [205]:

```
print(categories_one_hot.shape)
print(sub_categories_one_hot.shape)
print(tfidf_W2V_vectors.shape)
print(price_standardised.shape)
```

```
(109248, 9)
(109248, 30)
(109248, 300)
(109248, 1)
```

In [206]:

```
X_4 = hstack((categories_one_hot,sub_categories_one_hot, tfidf_W2V_vectors, price_standardised))
```

In [207]:

```
X_4.shape
```

Out[207]:

```
(109248, 340)
```

In [209]:

```
X_4 = X_4.toarray()
```

In [210]:

```
X_4 = X_4[:5000, :]
print(X_4.shape)
```

```
(5000, 340)
```

In [211]:

```
#Pickling it
with open('X_4', 'wb') as f:
    pickle.dump(X_4, f)
```

In [233]:

```
tsne_4 = TSNE(n_components=2, perplexity=100, learning_rate=1000)
X_embedding_4 = tsne_4.fit_transform(X_4)
```

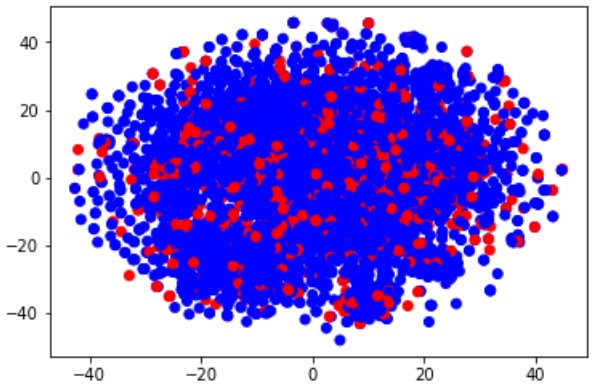
In [234]:

```
for_tsne_4 = np.hstack((X_embedding_4, y.reshape(-1,1)))
print(for_tsne_4.shape)
```

```
(5000, 3)
```

In [235]:

```
for_tsne_df_4 = pd.DataFrame(data=for_tsne_4,
columns=['Dimension_x','Dimension_y','project_is_approved'])
colors = {0:'red', 1:'blue'}
plt.scatter(for_tsne_df_4['Dimension_x'], for_tsne_df_4['Dimension_y'], c=
for_tsne_df_4['project_is_approved'].apply(lambda x: colors[x]))
plt.show()
```




# Summary

The plot tell us that the data isn't much changing even after we tried the different perplexity and learning rate. So it tells us that this is the originality of the data in 2 dimensions.

**Note:**

- For the memory constraint i took only 5000 points. If we have good computational resourse then we can use more points in the plot.