URBAN PLANNING FOR SUSTAINABLE CITIES

BY: Irene Izere

Introduction

- As the world is developing , where the population , the infrastructures are increasing , we are looking on new ways and better ways to improve our quality of life .There is a need to improve it socially, economically and in regards of the environment.

- This project is presented to the city officials and urban planners who are looking for a way to improve the sustainability of the cities.

- Sustainability need to be a great key for a good urban plan.

Problem statement

- This project is aiming to help urban planners design and plan not only for this generation but also for the future by understanding the area that need to be more focused on in order to achieve sustainability.

Definition

- ***Urban Planning*** is the design and regulation of the uses of spaces that focus on the physical form, economic functions and social impacts of the urban environment and on the location of different activities within it.

- ***Sustainability*** is the development that meets the needs of the present without compromising the ability of the future generations to meet their own needs.

Data collection

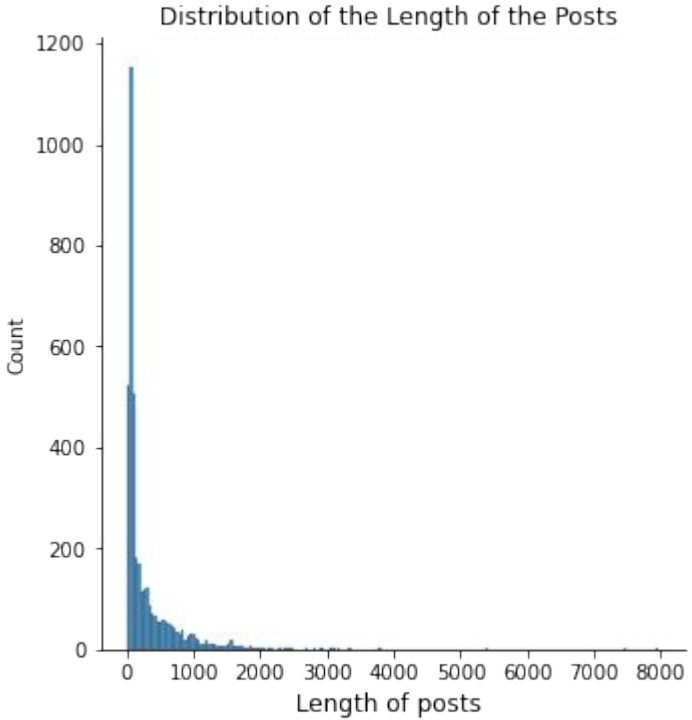
- For this project we collect data from Reddit website.Reddit is a network of online communities where people can dive into their interests , hobbies and passions. Essentially, it’s a massive collection of forums where people can share news and content or comment on other people’s posts.
- We collect using this pushshift, API on this url 'https://api.pushshift.io//reddit/search/submission'
- The data is from the subreddit of “***Urban Planning***” and “***Sustainability*** “

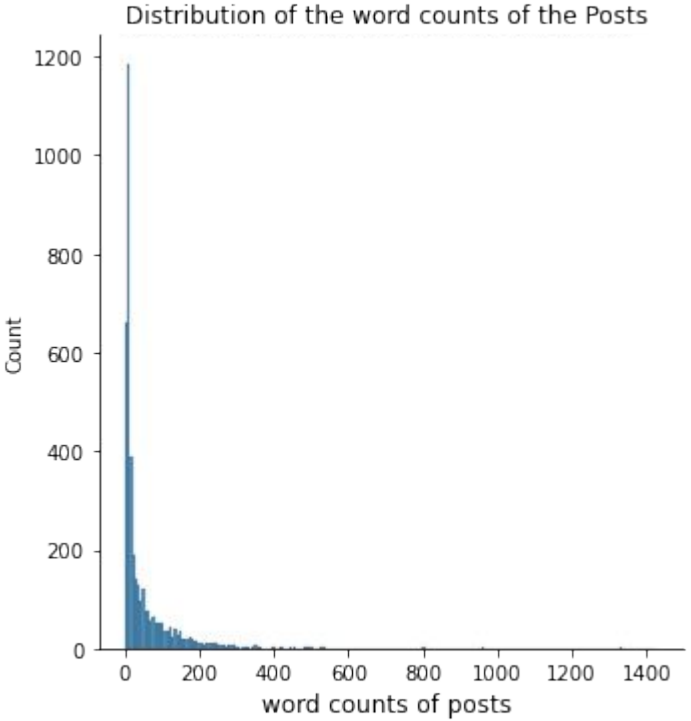
Data Analysis

- We looked at the distribution of the word counts and length of the post.

- We used different method to analyse the word from the post .The **countvectorizer** which count the occurrence of the word in our dataset and TF-IDF vectorizer means Term Frequency - Inverse Document Frequency. This is a statistic that is based on the frequency of a word in the dataset but it also provides a numerical representation of how important a word is for statistical analysis.

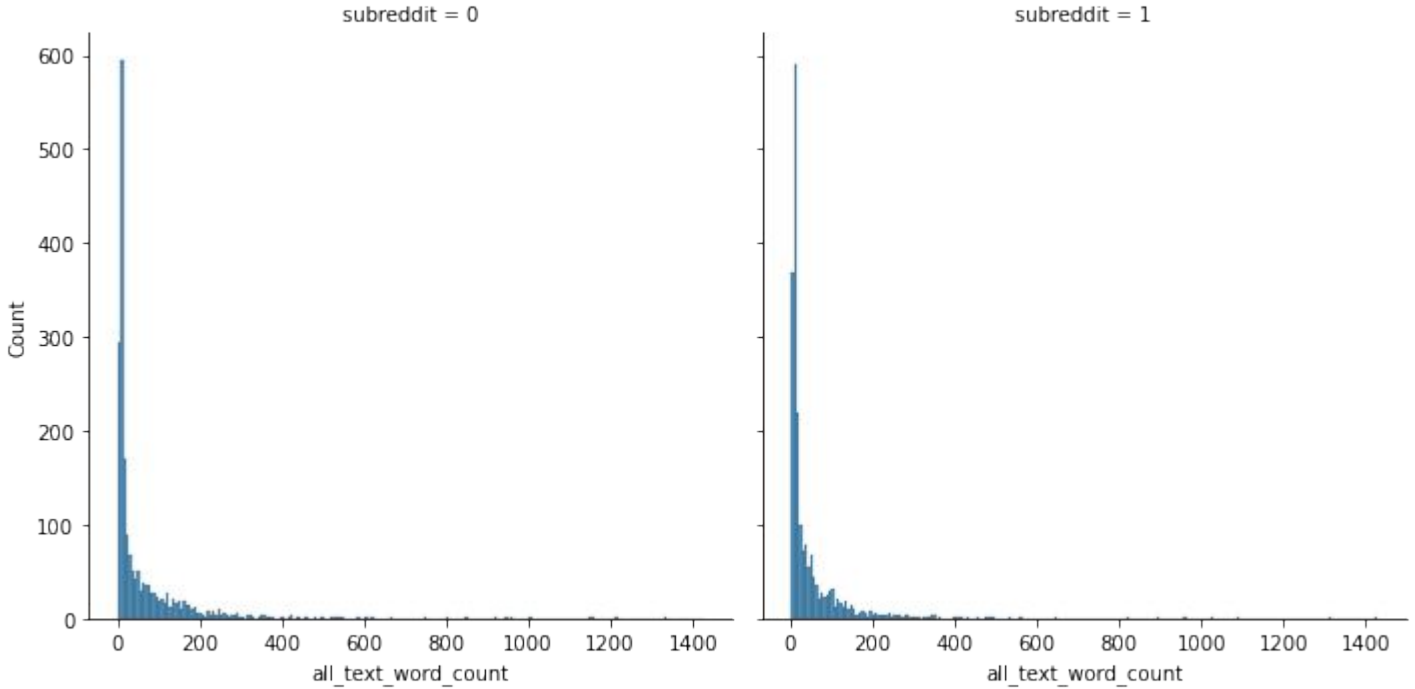
Data Analysis





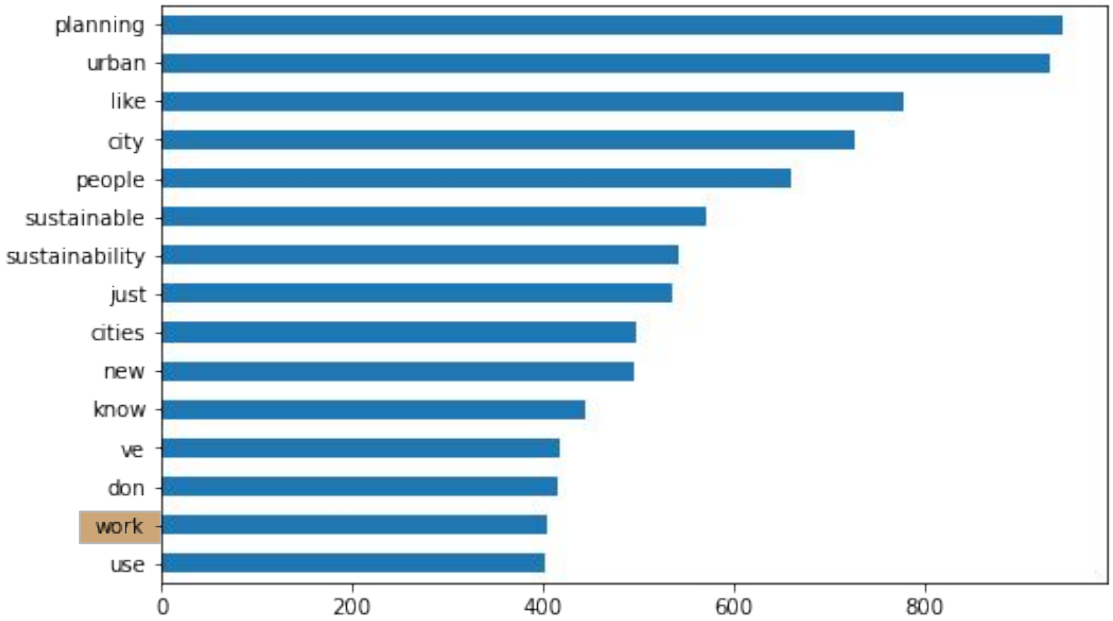
We look into the distribution of the overall length of the posts and word counts and as shown in this graph , the distribution right skewed which means that the most post are between **0-200 words** or between **0-1000 characters** .

Data Analysis...



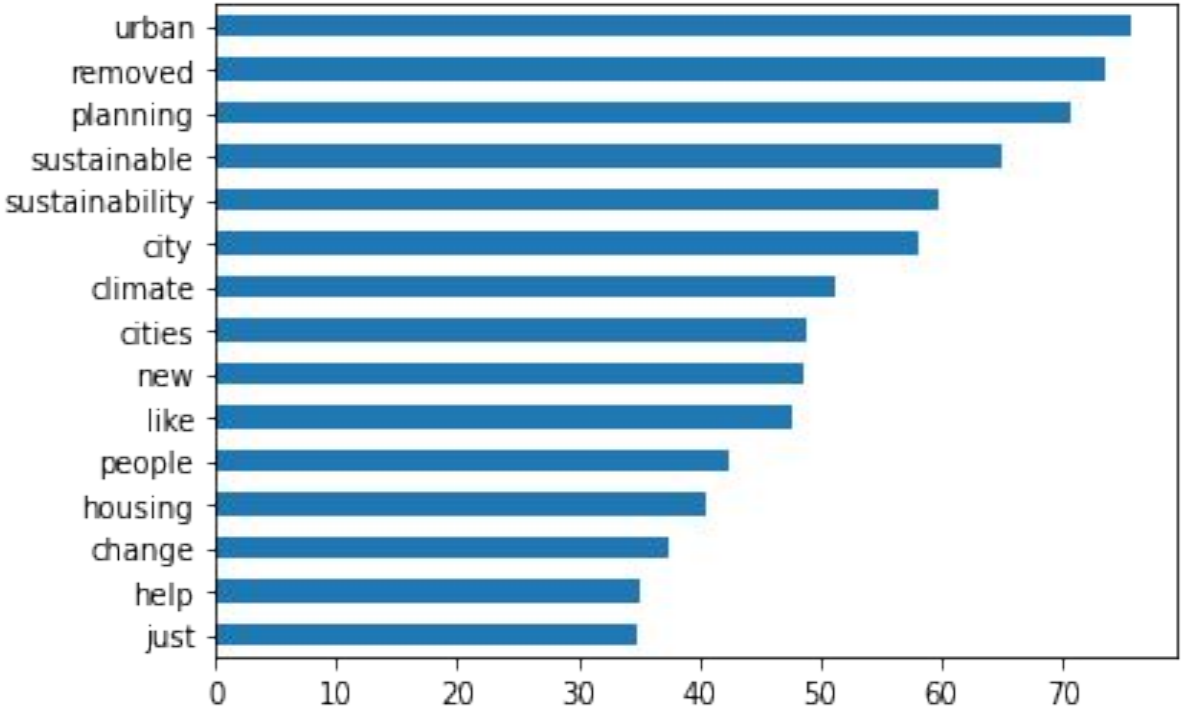
Comparison of distribution of the word counts on posts by category. *Subreddit :0* is **Urban Planning** and *subreddit :1* is **Sustainability.** We can see that the distribution is the same .

Data Analysis: The most common words



Using the countvectorizer method we see that apart from the word in the title of the forums , we see the *word* ***city,people,like,work and use***.

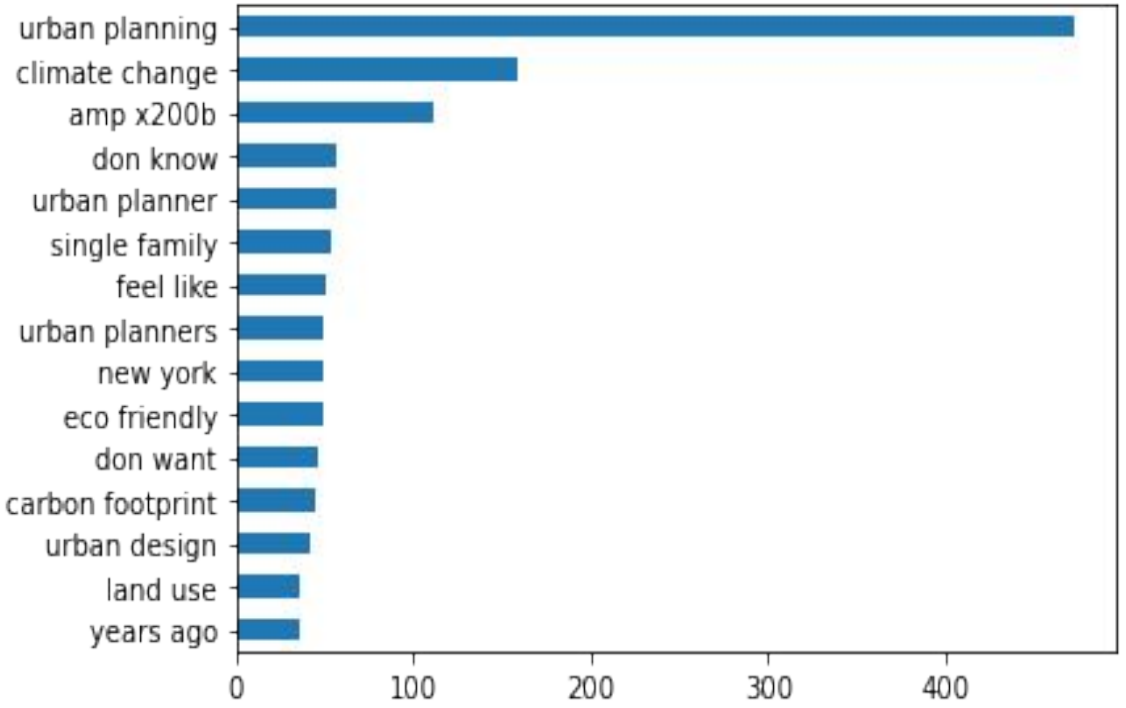
Data Analysis: The most common words



Using the TF-IDF Vectorizer method we see that the words that we see most are : ***removed,climate,city,like,people,housing,change,help and just***.

Data Analysis: The most 2 common words

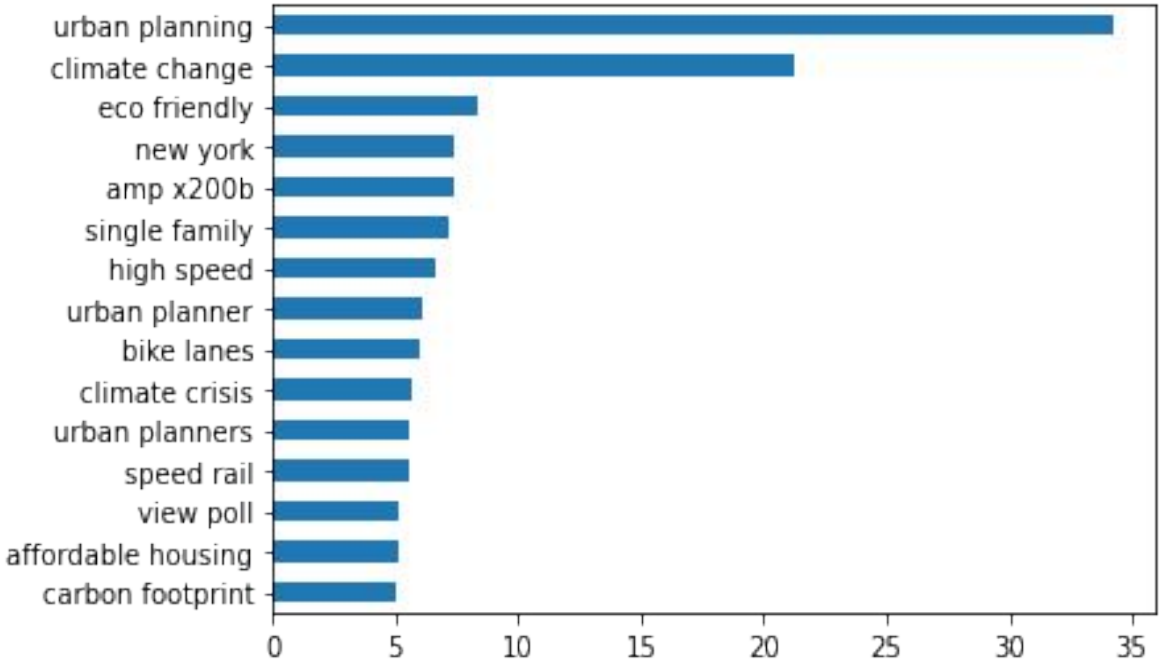
- *Climate change*
- *Single family*
- *Eco friendly*
- *Carbon footprint*
- *Urban design*
- *Land use*
- *Years ago*



This is using the CountVectorizer method.

Data Analysis: The most 2 common words

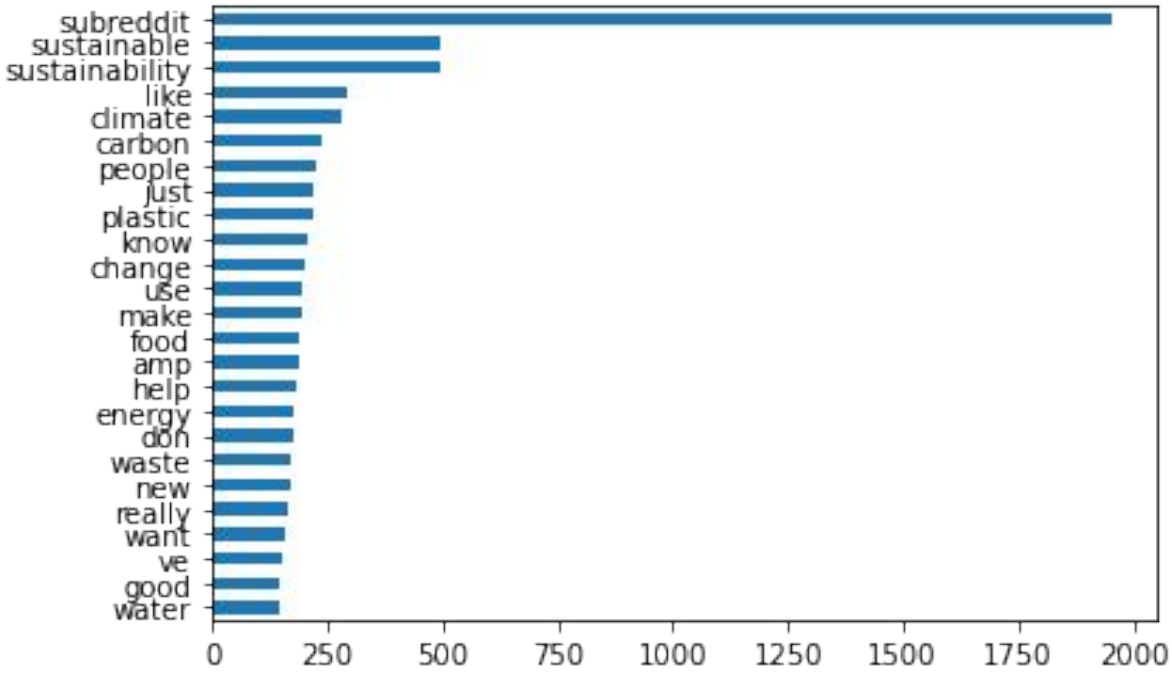
- *Climate change*
- *Eco friendly*
- *Single family*
- *High speed*
- *Bike lanes*
- *Climate crisis*
- *Speed rail*
- *Affordable housing*
- *Carbon footprint*



Using the TF-IDF Vectorizer method

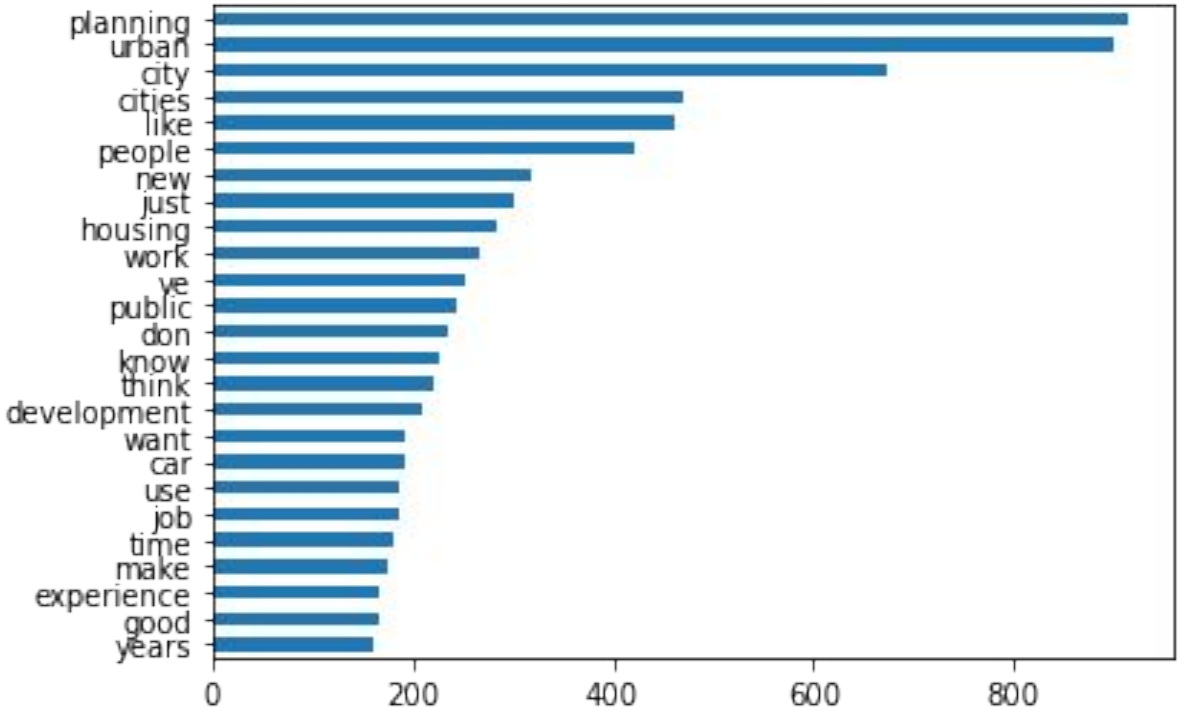
Data Analysis: The most common words in sustainability

- *Climate*
- *Carbon*
- *Plastic*
- *Change*
- *Energy*
- *Waste*
- *good*
- *Water*



Data Analysis: The most common words in urban planning

- *City*
- *People*
- *Housing*
- *Work*
- *Public*
- *Development*
- *Car*
- *Time*
- *experience*



Modelling

- For this task , we looked for different machine learning model that can help in differentiate the post from the urban planning forum from the sustainability forum.
- We compare the Random Forest Classifiers ,the Multinomial Naives Bayes and the K-Nearest Neighbors

Using the Multinomial Naives Bayes model which look into probabilities
The model has A score of **94%** on training data and **90%** on unseen data which is an indication of high variance .

Precision is 90% .
The recall is 89%
The f-1 score is 90%

Using the Random Forest Classifier model
The model has a score of **99%** on training data and **87%** on unseen data which is an indication of high variance .

Precision is 85% .

The recall is 91%
The f-1 score is 88%

Conclusion

- My recommendation for the urban planners are to incorporate more elements regarding climate change , waste management and energy and and also non-motorized mode of transportation such as bike lanes.

- The best model to use also in this case will be Multinomial Naives Bayes model because even though it doesn’t have a highest accuracy score the f-1 score is high which is the mean between precision and recall.