1.Introduction

1.1 Background

Tanzania is a country in East Africa within the African Great Lakes region. It borders Uganda to the north; Kenya to the northeast; Comoro Islands and the Indian Ocean to the east; Mozambique and Malawi to the south; Zambia to the southwest; and Rwanda, Burundi, and the Democratic Republic of the Congo to the west. Mount Kilimanjaro, Africa's highest mountain, is in northeastern Tanzania.

The population distribution in Tanzania is extremely uneven. Most people live on the northern border or the eastern coast, with much of the remainder of the country being sparsely populated. Density varies from 12 per square kilometer (31/sq mi) in the Katavi Region to 3,133 per square kilometer (8,110/sq mi) in Dar es Salaam. Approximately 70 percent of the population is rural, although this percentage has been declining since at least 1967

According to the 2012 census, the total population was 44,928,923 compared to 12,313,469 in 1967, resulting in an annual growth rate of 2.9 percent. The under 15 age group represented 44.1 percent of the population, with 35.5 percent being in the 15–35 age group, 52.2 percent being in the 15–64 age group, and 3.8 percent being older than 64. [1]

1.2 Problem

According to the Wikipedia [2] and [3], most of the regions in Tanzania, though not all, with high number of people (according to census done in 2012), contribute much to the national GDP. This project aims to find out the relationship between the population and national GDP and also explore some of the available markets in one of the high GDP contributing region so as to enable different stakeholders who are looking for where they can sell their products.

2. Data acquisition and cleaning

2.1 Data sources

Several sources were used for data collection. List of regions of Tanzania by GDP and population in each regions of Tanzania data was scraped from Wikipedia [2] [3], while the geographical

coordinates of each region were obtained from [4]. The coordinates were disorganized, so after obtaining the coordinated of each region, the CSV file was prepared which contained the list of all regions and thein coordinates. The geographical coordinates of the highest contributing region to the national GDP were utilized to collect venues data within 5000 m radius using Foursquare API.

2.2 Data cleaning

Data downloaded or scraped from multiple sources were combined into one table. The regions such as Songwe, Geita, Simiyu, Katavi and others, which came into existence after the census done in 2012 were removed since there were no population data about them.

The remaining data were used to examine the relationship between the population trend in each region according to the census done in 2012 and the contribution of that region to the national GDP.

2.3 Feature selection

After data cleaning, there were 22 samples and 10 features in the data. Some features were kept for further usages but other features were discarded. The unrequired features were dropped because they don't provide any necessary information required for task completion. Table 1, shows the dropped features.

Table 1 Shows the Features dropped

S/No.	Feature dropped
1.	Rank
2.	GDP in mil USD (PPP)
3.	Equivalent country[3]

References

[1] "Demographics of Tanzania".

[2] "List of regions of Tanzania by GDP".

[3] "Regions of Tanzania".

[4] Latitude.to, "regions," 2020. [Online]. Available: https://latitude.to/map/tz/tanzania/regions.