Rural Urban dynamics (DTE Datathon)

Group 7

2023-09-26

introduction

This analysis is intended for the DTE datathon, which aims to cultivate a data-driven, innovative, and collaborative environment for creatively addressing societal issues and producing actionable solutions.

```r
library(tidyverse)
library(gridExtra)
library(readxl)
library(reshape2)
library(janitor)
library(ggthemes)

library(knitr)
# library(kableExtra)

df <- data.frame(
  Details = c("Time Frame","Deadline","Software/Tools"),
  values  = c("1 Week","28th September","R - studio")
)
df %>% kable(caption = "The rules of the datathon")
```

Table 1: The rules of the datathon

Details	values
Time Frame	1 Week
Deadline	28th September
Software/Tools	R - studio

> `The objective:` The objective of this challenge is to analyze open data sets in Kenya using real-world data. You are therefore to explore, visualize, and draw insights from the provided dataset to provide better insights to different stakeholders on how urbanization is impacting different aspects of Kenyan society. Participants should work in teams and are encouraged to utilize technologies such as big data, machine learning, and artificial Intelligence that train test, and evaluate multiple data sets to uncover innovative solutions.

```r
colors <- c('#BC5308', '#FFECD1', '#C5CAB8', '#FF7D00', '#8AA79F', '#FFB569', '#15616D', '#001524')

df <- read_excel("Dataset/Dataset.xlsx")
df %>% select(1:7) %>% head() %>% knitr::kable(caption = "A sample of the data")
```

Table 2: A sample of the data

Country	County	Rural_ppn	Urban_ppn	Total Population	Urban/Total	Status
Kenya	Baringo	591474	75289	666763	0.1129172	Rural
Kenya	Bomet	847718	27971	875689	0.0319417	Rural
Kenya	Bungoma	1480458	190112	1670570	0.1138007	Rural
Kenya	Busia	779928	113753	893681	0.1272859	Rural
Kenya	Elgeiyo-Marakwet	433901	20579	454480	0.0452803	Rural
Kenya	Embu	532675	75924	608599	0.1247521	Rural

data

As per the rules, we sourced the data from KNBS on the following website: KNBS Data.

Additionally, we obtained supplementary data from the following website: Kenya County Fact Sheets Report (PDF).

The collected data encompasses a wide range of aspects of Kenyan life, including population, GDP, infrastructure, education, healthcare, and employment.

data soourcing and cleaning

The data collection process was primarily manual, and it followed the following general steps:

1. **Data Compilation:** We collaborated in Google Sheets to collect various data points simultaneously, gradually building up the dataset we needed.
2. **Data Cleaning:** We employed a manual query system to identify and address any anomalies or outliers in the dataset. This process involved thorough cross-checking to ensure data accuracy.
3. **Feature Engineering:** We created additional columns in the dataset, such as:
   - Total Population: $totalPopulation = femalePopulation + malePopulation$
   - Urban-Rural Classification: $urban_rural = \frac{urbanPopulation}{totalPopulation}$

   Based on the urban-rural classification, we categorized certain counties as either rural or urban. Counties were classified as fully urban if the urban population exceeded 40% of the total population.

EDA

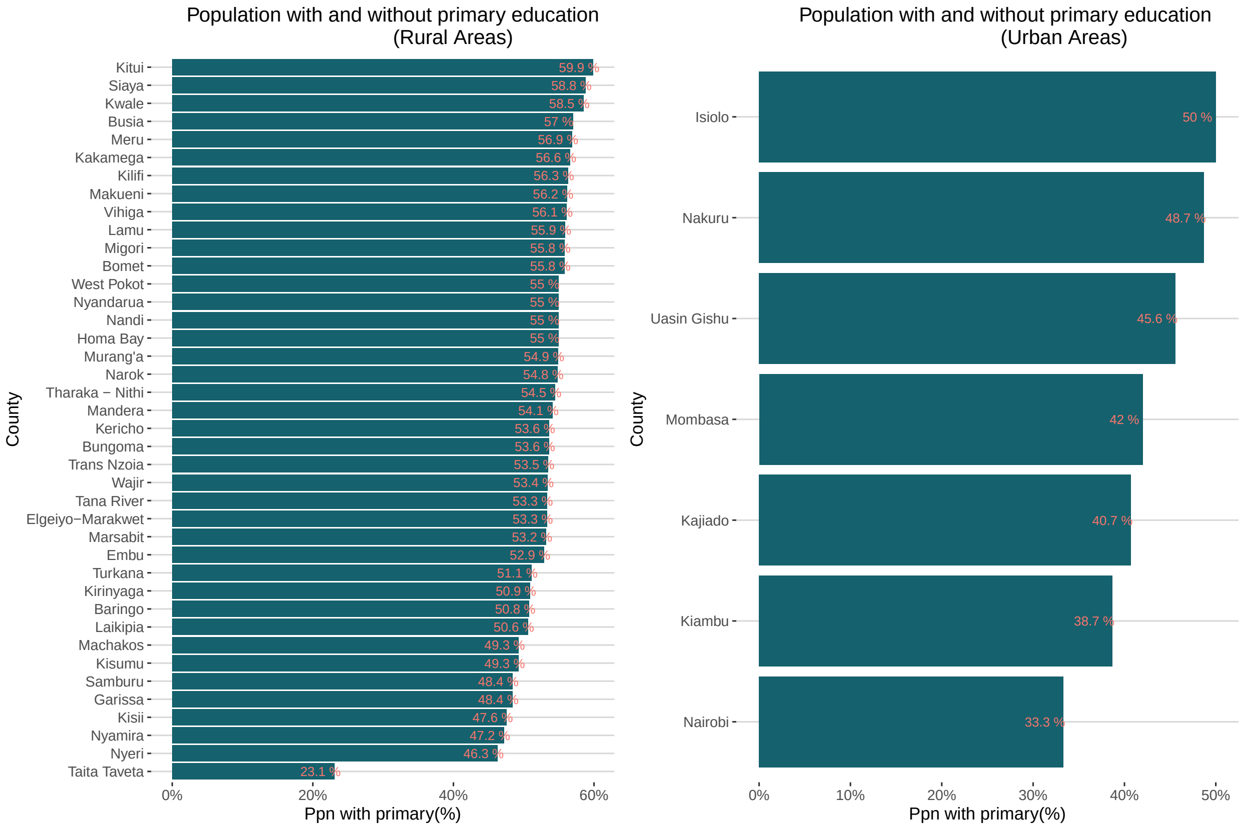
Rural-Urban population dynamics

Rural-Urban health dynamics

Rural-Urban Education dynamics

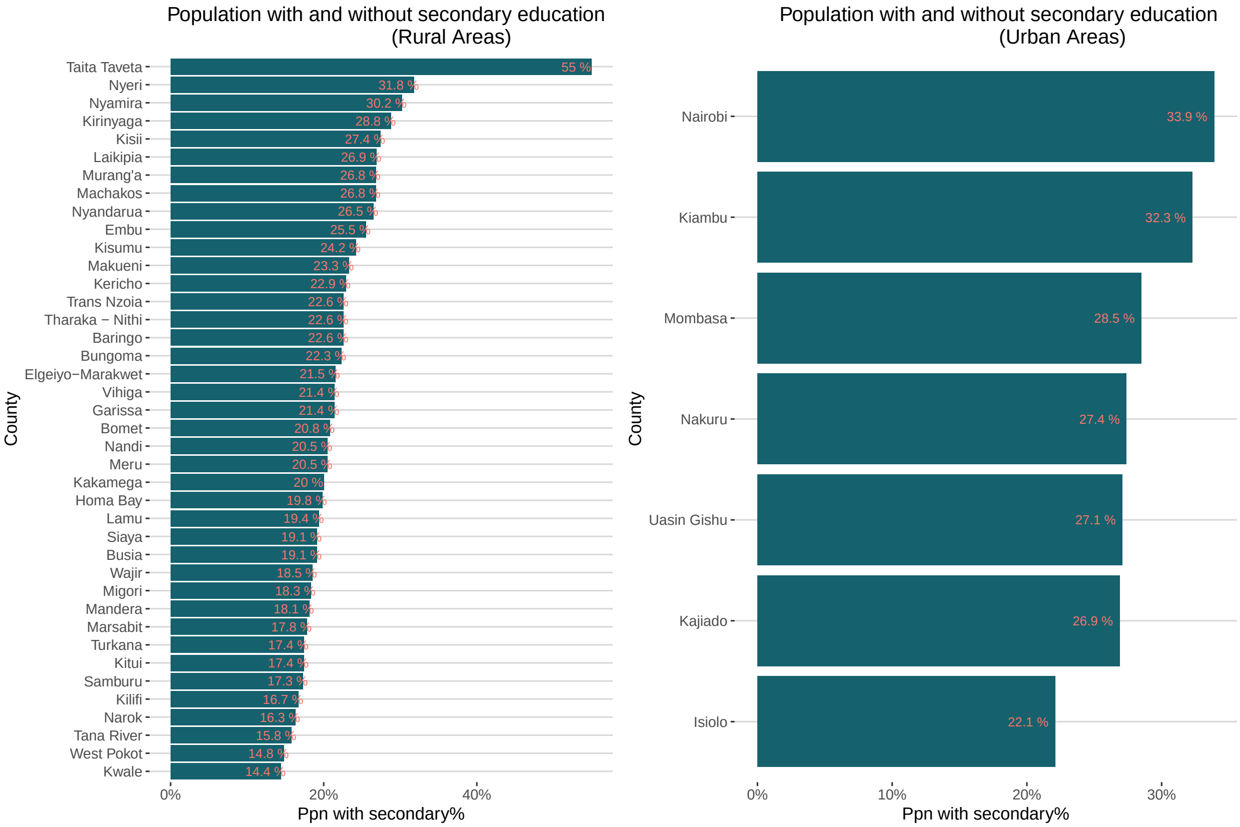
Population with and without education

```
# find an arrangement index
y <- df %>%  select(County,`Ppn with primary(%)`,Status) %>% arrange(`Ppn with primary(%)`)
y$County <- factor(y$County,levels = y$County)



# create a graph for urban
rural <- y %>% filter(Status == 'Rural') %>% mutate(perc = paste(`Ppn with primary(%)`,'%')) %>%
  ggplot(aes(x = County,y  =`Ppn with primary(%)`,label = perc))+geom_bar(stat = 'identity',fill = colo
  #scale_fill_manual(values = c(colors[1],colors[7]))+
  labs(title = "Population with and without primary education\n (Rural Areas)")+scale_y_continuous(label
  geom_text(aes(y = `Ppn with primary(%)` - 2, color = colors[2]), size = 3, position = position_dodge(
    plot.title = element_text(hjust = 0.7)  # Center the title
  )+
  theme_hc()

# create a graph for urban
urban <- y %>% filter(Status == 'Urban') %>% mutate(perc = paste(`Ppn with primary(%)`,'%')) %>%
  ggplot(aes(x = County,y  =`Ppn with primary(%)`,label = perc))+geom_bar(stat = 'identity',fill = colo
  #scale_fill_manual(values = c(colors[1],colors[7]))+
  labs(title = "Population with and without primary education\n (Urban Areas)")+scale_y_continuous(label
  geom_text(aes(y = `Ppn with primary(%)` - 2, color = colors[2]), size = 3, position = position_dodge(
    plot.title = element_text(hjust = 0.7)  # Center the title
  )+
  theme_hc()
grid.arrange(rural,urban,ncol =2)
```



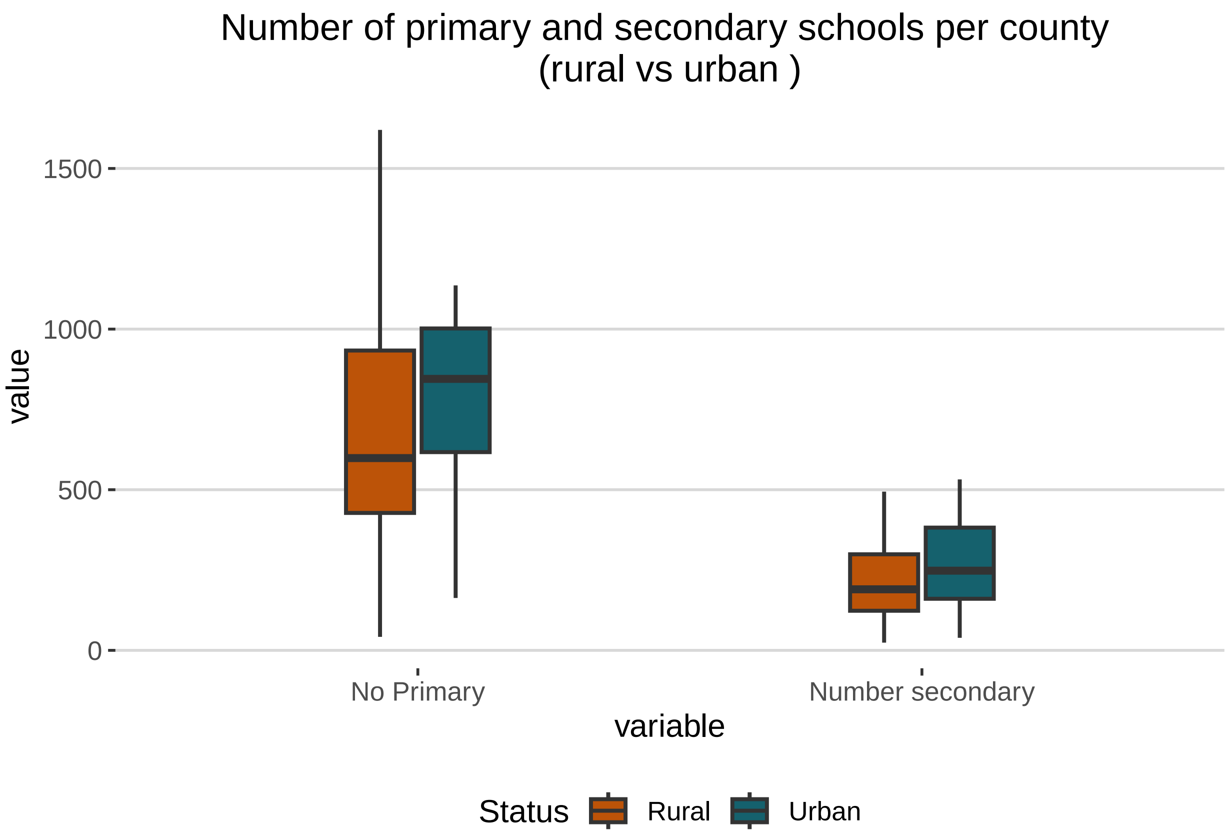
Primary

```
# find an arrangement index
y <- df %>%  select(County,`Ppn with secondary%`,Status) %>% arrange(`Ppn with secondary%`)
y$County <- factor(y$County,levels = y$County)



# create a graph for urban
rural <- y %>% filter(Status == 'Rural') %>% mutate(perc = paste(`Ppn with secondary%`,'%')) %>%
  ggplot(aes(x = County,y  =`Ppn with secondary%`,label = perc))+geom_bar(stat = 'identity',fill = colo
  #scale_fill_manual(values = c(colors[1],colors[7]))+
  labs(title = "Population with and without secondary education\n (Rural Areas)")+scale_y_continuous(la
  geom_text(aes(y = `Ppn with secondary%` - 2, color = colors[2]), size = 3, position = position_dodge(
    plot.title = element_text(hjust = 0.7)  # Center the title
  )+
  theme_hc()

# create a graph for urban
urban <- y %>% filter(Status == 'Urban') %>% mutate(perc = paste(`Ppn with secondary%`,'%')) %>%
  ggplot(aes(x = County,y  =`Ppn with secondary%`,label = perc))+geom_bar(stat = 'identity',fill = colo
  #scale_fill_manual(values = c(colors[1],colors[7]))+
  labs(title = "Population with and without secondary education\n (Urban Areas)")+scale_y_continuous(la
  geom_text(aes(y = `Ppn with secondary%` - 2, color = colors[2]), size = 3, position = position_dodge(
    plot.title = element_text(hjust = 0.7)  # Center the title
  )+
  theme_hc()
grid.arrange(rural,urban,ncol =2)
```



Secondary

```
df %>% select(Status,`No Primary`,`Number secondary`) %>%
  melt(id.vars = 'Status') %>%
  ggplot(aes(x = variable,y =value,fill = Status))+geom_boxplot(width = 0.3, size = 0.7)+
  theme_hc()+
  labs(title = "Number of primary and secondary schools per county \n(rural vs urban )")+
  theme(
    plot.title = element_text(hjust = 0.5)  # Center the title
  )+
  scale_fill_manual(values = c(colors[1],colors[7]))
```



Rural-Urban Misc

findings and conclusion