Assessing Australia's Agricultural Sustainability in the Context of Global Climate Change Indicators

Shevindi Basthian Koralalge

2023-10-29

Abstract

Australia's agricultural industry is vital to the country's trade, food security, and economy. However, the effects of climate change present serious difficulties for it. The critical challenge of evaluating Australia's agricultural sustainability in light of global climate change indicators is covered in this report.This analysis reviewed Australian agricultural sustainability in relation to global climate change indicators, involving data preprocessing and Exploratory Data Analysis (EDA) to identify significant correlations between agricultural and environmental indicators, addressing missing values and eliminating unnecessary columns.This analysis reviewed Australian agricultural sustainability in relation to global climate change indicators, involving data preprocessing and Exploratory Data Analysis (EDA) to identify significant correlations between agricultural and environmental indicators, addressing missing values and eliminating unnecessary columns.The study used regression analyses and clustering analysis to evaluate the impact of factors on agricultural land. Australia was placed in Cluster 2, indicating potential sustainability concerns. Principal Component Analysis was used to reduce data dimensionality. However, the analysis has limitations as it needs further research. The findings and achievements of this study can inform strategies and aimed at promoting agricultural sustainability in the face of climate change in Australia.

Introduction

For the economy, food security, and international trade of Australia, the agricultural sector is essential. However, it faces a significant challenge in the form of climate change, which introduces uncertainty and potential disruptions to traditional farming practices. This report explores into the critical task of assessing Australia's agricultural sustainability in the context of global climate change indicators. It is essential to comprehend how climate change affects agriculture and to assess sustainable strategies in order to guarantee a resilient and adaptable farming industry. Climate change has emerged as a major global concern, affecting many industries, including agriculture. A significant threat to agricultural productivity is the changing climate, which is characterized by rising temperatures, changed rainfall patterns, and increased climate variability. Extreme weather events such as droughts and floods, which are caused by climate change, make farming in Australia more difficult. The Australian agricultural sector is vulnerable to climate change-related uncertainties, such as temperature variations, rainfall shifts, and extreme heat, CO2 and other gas emissions which can impact crop yields, farm profitability, and disrupt farmers' livelihoods. The report explores the impact of these uncertainties and understanding their behavior on Australian agriculture. This highlights the need to tailor strategies and adaptation measures to specific areas, taking into consideration both the challenges and opportunities posed by climate change indicators. The objectives of the report are to evaluate climate change indicators, compare and contrast the climate change indicators in the context of several countries with Australia.

Data

Data Source Climate change is a defining issue of our age(Climate Change, n.d.).The dataset used in this study was obtained from the World Bank Group is a global partnership comprising five related institutions

working for sustainable solutions that reduce poverty and build shared prosperity in developing countries(Who We Are, n.d.).The data was collected with the World Bank's intention of to go further in helping countries reduce poverty and rise to the challenges of climate change(Climate Change, n.d.).

The variables (columns) correspond to country characteristics or indicators gathered by the World Bank and classified as important to climate change. Every record (row) is associated with a particular country. This data collection is cross-sectional rather than longitudinal, despite the fact that most of studies on climate change focuses on how things are changing over time. It is a snapshot of one particular recent value for each of these attributes for each country.

Data Description The dataset contains a set of variables, each representing specific climate related indicators.The dataset includes records for various countries, where each record provides information relevant to these indicators.

```
## Number of Records (Rows): 217
```

```
## Number of Variables(Columns)): 79
```

These variables contain both character and numerical data.Each record of the dataset represent a specific country.Each value of the dataset is the most recent value of that indicator available for that country between 2001 and 2020. The dataset is summarised in Table 1 below.

```
## Warning: package 'dplyr' was built under R version 4.1.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
## Warning: package 'kableExtra' was built under R version 4.1.3
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
## Warning: package 'corrplot' was built under R version 4.1.3
```

```
## corrplot 0.92 loaded
```

Variable

Description

Variable_Type

Missing_Values

iso3c

iso3c

Country code

character

0

country

country

Country

character

0

AG.LND.AGRI.K2

AG.LND.AGRI.K2

Agricultural land refers to the share of land area that is arable, under permanent crops, and under permanent pastures. Arable land includes land defined by the FAO as land under temporary crops (double-cropped areas are counted once), temporary meadows for mowing or for pasture, land under market or kitchen gardens, and land temporarily fallow. Land abandoned as a result of shifting cultivation is excluded. Land under permanent crops is land cultivated with crops that occupy the land for long periods and need not be replanted after each harvest, such as cocoa, coffee, and rubber. This category includes land under flowering shrubs, fruit trees, nut trees, and vines, but excludes land under trees grown for wood or timber. Permanent pasture is land used for five or more years for forage, including natural and cultivated crops.

numeric

6

AG.LND.AGRI.ZS

AG.LND.AGRI.ZS

Agricultural land refers to the share of land area that is arable, under permanent crops, and under permanent pastures. Arable land includes land defined by the FAO as land under temporary crops (double-cropped areas are counted once), temporary meadows for mowing or for pasture, land under market or kitchen gardens, and land temporarily fallow. Land abandoned as a result of shifting cultivation is excluded. Land under permanent crops is land cultivated with crops that occupy the land for long periods and need not be replanted after each harvest, such as cocoa, coffee, and rubber. This category includes land under flowering shrubs, fruit trees, nut trees, and vines, but excludes land under trees grown for wood or timber. Permanent pasture is land used for five or more years for forage, including natural and cultivated crops.

numeric

6

AG.LND.ARBL.ZS

AG.LND.ARBL.ZS

Arable land includes land defined by the FAO as land under temporary crops (double-cropped areas are counted once), temporary meadows for mowing or for pasture, land under market or kitchen gardens, and land temporarily fallow. Land abandoned as a result of shifting cultivation is excluded.

numeric

9

AG.LND.EL5M.RU.K2

AG.LND.EL5M.RU.K2

Rural land area below 5m is the total rural land area in square kilometers where the elevation is 5 meters or less.

numeric

38

AG.LND.EL5M.RU.ZS

AG.LND.EL5M.RU.ZS

Rural land area below 5m is the percentage of total land where the rural land elevation is 5 meters or less.

numeric

38

AG.LND.EL5M.UR.K2

AG.LND.EL5M.UR.K2

Urban land area below 5m is the total urban land area in square kilometers where the elevation is 5 meters or less.

numeric

38

AG.LND.EL5M.UR.ZS

AG.LND.EL5M.UR.ZS

Urban land area below 5m is the percentage of total land where the urban land elevation is 5 meters or less.

numeric

38

AG.LND.EL5M.ZS

AG.LND.EL5M.ZS

Land area below 5m is the percentage of total land where the elevation is 5 meters or less.

numeric

38

AG.LND.FRST.K2

AG.LND.FRST.K2

Forest area is land under natural or planted stands of trees of at least 5 meters in situ, whether productive or not, and excludes tree stands in agricultural production systems (for example, in fruit plantations and agroforestry systems) and trees in urban parks and gardens.

numeric

3

AG.LND.FRST.ZS

AG.LND.FRST.ZS

Forest area is land under natural or planted stands of trees of at least 5 meters in situ, whether productive or not, and excludes tree stands in agricultural production systems (for example, in fruit plantations and agroforestry systems) and trees in urban parks and gardens.

numeric

3

AG.LND.IRIG.AG.ZS

AG.LND.IRIG.AG.ZS

Agricultural irrigated land refers to agricultural areas purposely provided with water, including land irrigated by controlled flooding.

numeric

93

AG.LND.PRCP.MM

AG.LND.PRCP.MM

Average precipitation is the long-term average in depth (over space and time) of annual precipitation in the country. Precipitation is defined as any kind of water that falls from clouds as a liquid or a solid.

numeric

35

AG.YLD.CREL.KG

AG.YLD.CREL.KG

Cereal yield, measured as kilograms per hectare of harvested land, includes wheat, rice, maize, barley, oats, rye, millet, sorghum, buckwheat, and mixed grains. Production data on cereals relate to crops harvested for dry grain only. Cereal crops harvested for hay or harvested green for food, feed, or silage and those used for grazing are excluded. The FAO allocates production data to the calendar year in which the bulk of the harvest took place. Most of a crop harvested near the end of a year will be used in the following year.

numeric

36

BX.KLT.DINV.WD.GD.ZS

BX.KLT.DINV.WD.GD.ZS

Foreign direct investment, net inflows (% of GDP)

numeric

19

EG.ELC.ACCS.ZS

EG.ELC.ACCS.ZS

Foreign direct investment are the net inflows of investment to acquire a lasting management interest (10 percent or more of voting stock) in an enterprise operating in an economy other than that of the investor. It is the sum of equity capital, reinvestment of earnings, other long-term capital, and short-term capital as shown in the balance of payments. This series shows net inflows (new investment inflows less disinvestment) in the reporting economy from foreign investors, and is divided by GDP.

numeric

1

EG.ELC.COAL.ZS

EG.ELC.COAL.ZS

Access to electricity is the percentage of population with access to electricity. Electrification data are collected from industry, national surveys and international sources.

numeric

76

EG.ELC.HYRO.ZS

EG.ELC.HYRO.ZS

Sources of electricity refer to the inputs used to generate electricity. Coal refers to all coal and brown coal, both primary (including hard coal and lignite-brown coal) and derived fuels (including patent fuel, coke oven coke, gas coke, coke oven gas, and blast furnace gas). Peat is also included in this category.

numeric

76

EG.ELC.NGAS.ZS

EG.ELC.NGAS.ZS

Sources of electricity refer to the inputs used to generate electricity. Hydropower refers to electricity produced by hydroelectric power plants.

numeric

76

EG.ELC.NUCL.ZS

EG.ELC.NUCL.ZS

Sources of electricity refer to the inputs used to generate electricity. Gas refers to natural gas but excludes natural gas liquids.

numeric

76

EG.ELC.PETR.ZS

EG.ELC.PETR.ZS

Sources of electricity refer to the inputs used to generate electricity. Nuclear power refers to electricity produced by nuclear power plants.

numeric

76

EG.ELC.RNEW.ZS

EG.ELC.RNEW.ZS

Sources of electricity refer to the inputs used to generate electricity. Oil refers to crude oil and petroleum products.

numeric

0

EG.ELC.RNWX.KH

EG.ELC.RNWX.KH

Renewable electricity is the share of electricity generated by renewable power plants in total electricity generated by all types of plants.

numeric

76

EG.ELC.RNWX.ZS

EG.ELC.RNWX.ZS

Electricity production from renewable sources, excluding hydroelectric, includes geothermal, solar, tides, wind, biomass, and biofuels.

numeric

76

EG.FEC.RNEW.ZS

EG.FEC.RNEW.ZS

Electricity production from renewable sources, excluding hydroelectric, includes geothermal, solar, tides, wind, biomass, and biofuels.

numeric

4

EG.USE.COMM.GD.PP.KD

EG.USE.COMM.GD.PP.KD

Renewable energy consumption is the share of renewables energy in total final energy consumption.

numeric

54

EG.USE.ELEC.KH.PC

EG.USE.ELEC.KH.PC

Energy use per PPP GDP is the kilogram of oil equivalent of energy use per constant PPP GDP. Energy use refers to the use of primary energy before transformation to other end-use fuels, which is equal to indigenous production plus imports and stock changes, minus exports and fuels supplied to ships and aircraft engaged in international transport. PPP GDP is gross domestic product converted to 2017 constant international dollars using purchasing power parity rates. An international dollar has the same purchasing power over GDP as a U.S. dollar has in the United States.

numeric

75

EG.USE.PCAP.KG.OE

EG.USE.PCAP.KG.OE

Electric power consumption measures the production of power plants and combined heat and power plants less transmission, distribution, and transformation losses and own use by heat and power plants.

numeric

45

EN.ATM.CO2E.EG.ZS

EN.ATM.CO2E.EG.ZS

Energy use refers to the use of primary energy before transformation to other end-use fuels, which is equal to indigenous production plus imports and stock changes, minus exports and fuels supplied to ships and aircraft engaged in international transport.

numeric

49

EN.ATM.CO2E.GF.KT

EN.ATM.CO2E.GF.KT

Carbon dioxide emissions from solid fuel consumption refer mainly to emissions from use of coal as an energy source.

numeric

10

EN.ATM.CO2E.GF.ZS

EN.ATM.CO2E.GF.ZS

Carbon dioxide emissions from liquid fuel consumption refer mainly to emissions from use of natural gas as an energy source.

numeric

26

EN.ATM.CO2E.KD.GD

EN.ATM.CO2E.KD.GD

Carbon dioxide emissions are those stemming from the burning of fossil fuels and the manufacture of cement. They include carbon dioxide produced during consumption of solid, liquid, and gas fuels and gas flaring.

numeric

29

EN.ATM.CO2E.KT

EN.ATM.CO2E.KT

Carbon dioxide emissions are those stemming from the burning of fossil fuels and the manufacture of cement. They include carbon dioxide produced during consumption of solid, liquid, and gas fuels and gas flaring.

integer

26

EN.ATM.CO2E.LF.KT

EN.ATM.CO2E.LF.KT

Carbon dioxide emissions from liquid fuel consumption refer mainly to emissions from use of petroleum-derived fuels as an energy source.

numeric

10

EN.ATM.CO2E.LF.ZS

EN.ATM.CO2E.LF.ZS

Carbon dioxide emissions from liquid fuel consumption refer mainly to emissions from use of petroleum-derived fuels as an energy source.

numeric

26

EN.ATM.CO2E.PC

EN.ATM.CO2E.PC

Carbon dioxide emissions are those stemming from the burning of fossil fuels and the manufacture of cement. They include carbon dioxide produced during consumption of solid, liquid, and gas fuels and gas flaring.

numeric

26

EN.ATM.CO2E.PP.GD

EN.ATM.CO2E.PP.GD

Carbon dioxide emissions are those stemming from the burning of fossil fuels and the manufacture of cement. They include carbon dioxide produced during consumption of solid, liquid, and gas fuels and gas flaring.

numeric

31

EN.ATM.CO2E.PP.GD.KD

EN.ATM.CO2E.PP.GD.KD

Carbon dioxide emissions are those stemming from the burning of fossil fuels and the manufacture of cement. They include carbon dioxide produced during consumption of solid, liquid, and gas fuels and gas flaring.

numeric

35

EN.ATM.CO2E.SF.KT

EN.ATM.CO2E.SF.KT

Carbon dioxide emissions from solid fuel consumption refer mainly to emissions from use of coal as an energy source.

numeric

10

EN.ATM.CO2E.SF.ZS

EN.ATM.CO2E.SF.ZS

Carbon dioxide emissions from solid fuel consumption refer mainly to emissions from use of coal as an energy source.

numeric

26

EN.ATM.GHGO.KT.CE

EN.ATM.GHGO.KT.CE

Other greenhouse gas emissions are by-product emissions of hydrofluorocarbons, perfluorocarbons, and sulfur hexafluoride.

numeric

33

EN.ATM.GHGO.ZG

EN.ATM.GHGO.ZG

Other greenhouse gas emissions are by-product emissions of hydrofluorocarbons, perfluorocarbons, and sulfur hexafluoride. Each year of data shows the percentage change to that year from 1990.

numeric

29

EN.ATM.GHGT.KT.CE

EN.ATM.GHGT.KT.CE

Total greenhouse gas emissions in kt of CO2 equivalent are composed of CO2 totals excluding short-cycle biomass burning (such as agricultural waste burning and savanna burning) but including other biomass burning (such as forest fires, post-burn decay, peat fires and decay of drained peatlands), all anthropogenic CH4 sources, N2O sources and F-gases (HFCs, PFCs and SF6).

integer

26

EN.ATM.GHGT.ZG

EN.ATM.GHGT.ZG

Total greenhouse gas emissions are composed of CO2 totals excluding short-cycle biomass burning (such as agricultural waste burning and savanna burning) but including other biomass burning (such as forest fires, post-burn decay, peat fires and decay of drained peatlands), all anthropogenic CH4 sources, N2O sources and F-gases (HFCs, PFCs and SF6). Each year of data shows the percentage change to that year from 1990.

numeric

35

EN.ATM.HFCG.KT.CE

EN.ATM.HFCG.KT.CE

Hydrofluorocarbons, used as a replacement for chlorofluorocarbons, are used mainly in refrigeration and semiconductor manufacturing.

integer

80

EN.ATM.METH.KT.CE

EN.ATM.METH.KT.CE

Methane emissions are those stemming from human activities such as agriculture and from industrial methane production.

integer

26

EN.ATM.METH.ZG

EN.ATM.METH.ZG

Methane emissions are those stemming from human activities such as agriculture and from industrial methane production. Each year of data shows the percentage change to that year from 1990.

numeric

13

EN.ATM.NOXE.KT.CE

EN.ATM.NOXE.KT.CE

Nitrous oxide emissions are emissions from agricultural biomass burning, industrial activities, and livestock management.

integer

26

EN.ATM.NOXE.ZG

EN.ATM.NOXE.ZG

Nitrous oxide emissions are emissions from agricultural biomass burning, industrial activities, and livestock management. Each year of data shows the percentage change to that year from 1990.

numeric

12

EN.ATM.PFCG.KT.CE

EN.ATM.PFCG.KT.CE

Perfluorocarbons, used as a replacement for chlorofluorocarbons in manufacturing semiconductors, are a byproduct of aluminum smelting and uranium enrichment.

integer

80

EN.ATM.SF6G.KT.CE

EN.ATM.SF6G.KT.CE

Sulfur hexafluoride is used largely to insulate high-voltage electric power equipment.

integer

80

EN.CLC.DRSK.XQ

EN.CLC.DRSK.XQ

Disaster risk reduction progress score is an average of self-assessment scores, ranging from 1 to 5, submitted by countries under Priority 1 of the Hyogo Framework National Progress Reports. The Hyogo Framework is a global blueprint for disaster risk reduction efforts that was adopted by 168 countries in 2005. Assessments of 'Priority 1' include four indicators that reflect the degree to which countries have prioritized disaster risk reduction and the strengthening of relevant institutions.

numeric

134

EN.CLC.GHGR.MT.CE

EN.CLC.GHGR.MT.CE

GHG net emissions/removals by LUCF refers to changes in atmospheric levels of all greenhouse gases attributable to forest and land-use change activities, including but not limited to (1) emissions and removals of CO2 from decreases or increases in biomass stocks due to forest management, logging, fuelwood collection, etc.; (2) conversion of existing forests and natural grasslands to other land uses; (3) removal of CO2 from the abandonment of formerly managed lands (e.g. croplands and pastures); and (4) emissions and removals of CO2 in soil associated with land-use change and management. For Annex-I countries under the UNFCCC, these data are drawn from the annual GHG inventories submitted to the UNFCCC by each country; for non-Annex-I countries, data are drawn from the most recently submitted National Communication where

available. Because of differences in reporting years and methodologies, these data are not generally considered comparable across countries. Data are in million metric tons.

numeric

155

EN.CLC.MDAT.ZS

EN.CLC.MDAT.ZS

Droughts, floods, and extreme temperatures are the annual average percentage of the population that is affected by natural disasters classified as either droughts, floods, or extreme temperature events. A drought is an extended period of time characterized by a deficiency in a region's water supply that is the result of constantly below average precipitation. A drought can lead to losses to agriculture, affect inland navigation and hydropower plants, and cause a lack of drinking water and famine. A flood is a significant rise of water level in a stream, lake, reservoir or coastal region. Extreme temperature events are either cold waves or heat waves. A cold wave can be both a prolonged period of excessively cold weather and the sudden invasion of very cold air over a large area. Along with frost it can cause damage to agriculture, infrastructure, and property. A heat wave is a prolonged period of excessively hot and sometimes also humid weather relative to normal climate patterns of a certain region. Population affected is the number of people injured, left homeless or requiring immediate assistance during a period of emergency resulting from a natural disaster; it can also include displaced or evacuated people. Average percentage of population affected is calculated by dividing the sum of total affected for the period stated by the sum of the annual population figures for the period stated.

numeric

49

EN.POP.EL5M.RU.ZS

EN.POP.EL5M.RU.ZS

Rural population below 5m is the percentage of the total population, living in areas where the elevation is 5 meters or less.

numeric

38

EN.POP.EL5M.UR.ZS

EN.POP.EL5M.UR.ZS

Urban population below 5m is the percentage of the total population, living in areas where the elevation is 5 meters or less.

numeric

38

EN.POP.EL5M.ZS

EN.POP.EL5M.ZS

Population below 5m is the percentage of the total population living in areas where the elevation is 5 meters or less.

numeric

38

EN.URB.MCTY.TL.ZS

EN.URB.MCTY.TL.ZS

Population in urban agglomerations of more than one million is the percentage of a country's population living in metropolitan areas that in 2018 had a population of more than one million people.

numeric

96

ER.H2O.FWTL.K3

ER.H2O.FWTL.K3

Annual freshwater withdrawals refer to total water withdrawals, not counting evaporation losses from storage basins. Withdrawals also include water from desalination plants in countries where they are a significant source. Withdrawals can exceed 100 percent of total renewable resources where extraction from nonrenewable aquifers or desalination plants is considerable or where there is significant water reuse. Withdrawals for agriculture and industry are total withdrawals for irrigation and livestock production and for direct industrial use (including withdrawals for cooling thermoelectric plants). Withdrawals for domestic uses include drinking water, municipal use or supply, and use for public services, commercial establishments, and homes. Data are for the most recent year available for 1987-2002.

numeric

36

ER.H2O.FWTL.ZS

ER.H2O.FWTL.ZS

Annual freshwater withdrawals refer to total water withdrawals, not counting evaporation losses from storage basins. Withdrawals also include water from desalination plants in countries where they are a significant source. Withdrawals can exceed 100 percent of total renewable resources where extraction from nonrenewable aquifers or desalination plants is considerable or where there is significant water reuse. Withdrawals for agriculture and industry are total withdrawals for irrigation and livestock production and for direct industrial use (including withdrawals for cooling thermoelectric plants). Withdrawals for domestic uses include drinking water, municipal use or supply, and use for public services, commercial establishments, and homes. Data are for the most recent year available for 1987-2002.

numeric

42

ER.LND.PTLD.ZS

ER.LND.PTLD.ZS

Terrestrial protected areas are totally or partially protected areas of at least 1,000 hectares that are designated by national authorities as scientific reserves with limited public access, national parks, natural monuments, nature reserves or wildlife sanctuaries, protected landscapes, and areas managed mainly for sustainable use. Marine areas, unclassified areas, littoral (intertidal) areas, and sites protected under local or provincial law are excluded.

numeric

5

ER.MRN.PTMR.ZS

ER.MRN.PTMR.ZS

Marine protected areas are areas of intertidal or subtidal terrain–and overlying water and associated flora and fauna and historical and cultural features–that have been reserved by law or other effective means to protect part or all of the enclosed environment.

numeric

47

ER.PTD.TOTL.ZS

ER.PTD.TOTL.ZS

Terrestrial protected areas are totally or partially protected areas of at least 1,000 hectares that are designated by national authorities as scientific reserves with limited public access, national parks, natural monuments, nature reserves or wildlife sanctuaries, protected landscapes, and areas managed mainly for sustainable use. Marine protected areas are areas of intertidal or subtidal terrain–and overlying water and associated flora and fauna and historical and cultural features–that have been reserved by law or other effective means to protect part or all of the enclosed environment. Sites protected under local or provincial law are excluded.

numeric

6

IC.BUS.EASE.XQ

IC.BUS.EASE.XQ

Ease of doing business ranks economies from 1 to 190, with first place being the best. A high ranking (a low numerical rank) means that the regulatory environment is conducive to business operation. The index averages the country's percentile rankings on 10 topics covered in the World Bank's Doing Business. The ranking on each topic is the simple average of the percentile rankings on its component indicators.

integer

28

IQ.CPA.PUBS.XQ

IQ.CPA.PUBS.XQ

The public sector management and institutions cluster includes property rights and rule-based governance, quality of budgetary and financial management, efficiency of revenue mobilization, quality of public administration, and transparency, accountability, and corruption in the public sector.

numeric

130

IS.ROD.PAVE.ZS

IS.ROD.PAVE.ZS

Paved roads are those surfaced with crushed stone (macadam) and hydrocarbon binder or bituminized agents, with concrete, or with cobblestones, as a percentage of all the country's roads, measured in length.

numeric

166

NV.AGR.TOTL.ZS

NV.AGR.TOTL.ZS

Agriculture corresponds to ISIC divisions 1-5 and includes forestry, hunting, and fishing, as well as cultivation of crops and livestock production. Value added is the net output of a sector after adding up all outputs and subtracting intermediate inputs. It is calculated without making deductions for depreciation of fabricated assets or depletion and degradation of natural resources. The origin of value added is determined by the International Standard Industrial Classification (ISIC), revision 3. Note: For VAB countries, gross value added at factor cost is used as the denominator.

numeric

15

SE.ENR.PRSC.FM.ZS

SE.ENR.PRSC.FM.ZS

Gender parity index for gross enrollment ratio in primary and secondary education is the ratio of girls to boys enrolled at primary and secondary levels in public and private schools.

numeric

23

SE.PRM.CMPT.ZS

SE.PRM.CMPT.ZS

Primary completion rate, or gross intake ratio to the last grade of primary education, is the number of new entrants (enrollments minus repeaters) in the last grade of primary education, regardless of age, divided by the population at the entrance age for the last grade of primary education. Data limitations preclude adjusting for students who drop out during the final year of primary education.

numeric

28

SH.DYN.MORT

SH.DYN.MORT

Under-five mortality rate is the probability per 1,000 that a newborn baby will die before reaching age five, if subject to age-specific mortality rates of the specified year.

numeric

24

SH.MED.CMHW.P3

SH.MED.CMHW.P3

Community health workers include various types of community health aides, many with country-specific occupational titles such as community health officers, community health-education workers, family health workers, lady health visitors and health extension package workers.

numeric

157

SH.STA.MALN.ZS

SH.STA.MALN.ZS

Prevalence of underweight children is the percentage of children under age 5 whose weight for age is more than two standard deviations below the median for the international reference population ages 0-59 months. The data are based on the WHO's child growth standards released in 2006.

numeric

67

SI.POV.DDAY

SI.POV.DDAY

Poverty headcount ratio at $1.90 a day is the percentage of the population living on less than $1.90 a day at 2011 international prices. As a result of revisions in PPP exchange rates, poverty rates for individual countries cannot be compared with poverty rates reported in earlier editions.

numeric

55

SP.POP.GROW

SP.POP.GROW

Annual population growth rate for year t is the exponential rate of growth of midyear population from year t-1 to t, expressed as a percentage. Population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship.

numeric

0

SP.POP.TOTL

SP.POP.TOTL

Total population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship. The values shown are midyear estimates.

integer

0

SP.URB.GROW

SP.URB.GROW

Urban population refers to people living in urban areas as defined by national statistical offices. It is calculated using World Bank population estimates and urban ratios from the United Nations World Urbanization Prospects.

numeric

2

SP.URB.TOTL

SP.URB.TOTL

Urban population refers to people living in urban areas as defined by national statistical offices. It is calculated using World Bank population estimates and urban ratios from the United Nations World Urbanization Prospects. Aggregation of urban and rural population may not add up to total population because of different country coverages.

integer

2

SP.URB.TOTL.IN.ZS

SP.URB.TOTL.IN.ZS

Urban population refers to people living in urban areas as defined by national statistical offices. The data are collected and smoothed by United Nations Population Division.

numeric

2

Table 1. Dataset summary

The variables except isoc3, country and Population, total, have missing data for some observations.In total, 3238 observations have at least one missing value across variables.

Data Preprocessing

Before conducting the analysis, we performed some data pre-processing. This included handling missing values, removing duplicates and selecting specific indicators for analysis.

Here is the structure of the dataset before pre processing.

```
#structure of dataset
str(climateData)
```

```
## 'data.frame':    217 obs. of  79 variables:
##  $ iso3c               : chr  "ABW" "AFG" "AGO" "ALB" ...
##  $ country             : chr  "Aruba" "Afghanistan" "Angola" "Albania" ...
##  $ AG.LND.AGRI.K2      : num  20 379190 569525 11741 188 ...
##  $ AG.LND.AGRI.ZS      : num  11.1 58.1 45.7 42.8 40.1 ...
##  $ AG.LND.ARBL.ZS      : num  11.11 11.8 3.93 22.31 1.77 ...
##  $ AG.LND.EL5M.RU.K2   : num  4.28 NA 1516.94 1294.6 NA ...
##  $ AG.LND.EL5M.RU.ZS   : num  2.352 NA 0.121 4.562 NA ...
##  $ AG.LND.EL5M.UR.K2   : num  8.58 NA 54.13 84.84 NA ...
##  $ AG.LND.EL5M.UR.ZS   : num  4.7179 NA 0.00433 0.299 NA ...
##  $ AG.LND.EL5M.ZS      : num  7.07 NA 0.126 4.861 NA ...
##  $ AG.LND.FRST.K2      : num  4.2 12084.4 666073.8 7889 160 ...
##  $ AG.LND.FRST.ZS      : num  2.33 1.85 53.43 28.79 34.04 ...
##  $ AG.LND.IRIG.AG.ZS   : num  NA 5.13 NA 14.99 NA ...
##  $ AG.LND.PRCP.MM      : num  NA 327 1010 1485 NA ...
##  $ AG.YLD.CREL.KG      : num  NA 2165 753 4840 NA ...
##  $ BX.KLT.DINV.WD.GD.ZS: num  4.245 0.121 -4.584 7.857 NA ...
##  $ EG.ELC.ACCS.ZS      : num  100 97.7 45.7 100 100 ...
##  $ EG.ELC.COAL.ZS      : num  NA NA 0 0 NA ...
##  $ EG.ELC.HYRO.ZS      : num  NA NA 53.2 100 NA ...
##  $ EG.ELC.NGAS.ZS      : num  NA NA 0 0 NA ...
##  $ EG.ELC.NUCL.ZS      : num  NA NA 0 0 NA ...
##  $ EG.ELC.PETR.ZS      : num  NA NA 46.8 0 NA ...
##  $ EG.ELC.RNEW.ZS      : num  14.9 86.1 53.2 100 86.1 ...
##  $ EG.ELC.RNWX.KH      : num  NA NA 0 0 NA ...
##  $ EG.ELC.RNWX.ZS      : num  NA NA 0 0 NA ...
##  $ EG.FEC.RNEW.ZS      : num  8.02 21.42 56.79 38.27 18.51 ...
##  $ EG.USE.COMM.GD.PP.KD: num  NA NA 66.1 69.8 NA ...
##  $ EG.USE.ELEC.KH.PC   : num  NA NA 312 2309 NA ...
##  $ EG.USE.PCAP.KG.OE   : num  NA NA 545 808 NA ...
##  $ EN.ATM.CO2E.EG.ZS   : num  NA NA 2.36 2.23 NA ...
##  $ EN.ATM.CO2E.GF.KT   : num  0 319 1496 161 0 ...
##  $ EN.ATM.CO2E.GF.ZS   : num  NA 4.32 4.23 3.27 0 ...
##  $ EN.ATM.CO2E.KD.GD   : num  NA 0.354 0.274 0.382 0.132 ...
##  $ EN.ATM.CO2E.KT      : int  NA 7440 27340 5560 460 200300 177410 5550 NA 530 ...
##  $ EN.ATM.CO2E.LF.KT   : num  884 3476 17932 3363 469 ...
##  $ EN.ATM.CO2E.LF.ZS   : num  NA 47 50.6 68.2 99.9 ...
##  $ EN.ATM.CO2E.PC      : num  NA 0.2 0.887 1.94 5.973 ...
##  $ EN.ATM.CO2E.PP.GD   : num  NA 0.0961 0.125 0.1432 NA ...
##  $ EN.ATM.CO2E.PP.GD.KD: num  NA 0.0984 0.128 0.1456 NA ...
##  $ EN.ATM.CO2E.SF.KT   : num  0 4829.4 0 14.7 0 ...
##  $ EN.ATM.CO2E.SF.ZS   : num  NA 65.351 0 0.298 0 ...
##  $ EN.ATM.GHGO.KT.CE   : num  NA -1800 5203.1 11.7 NA ...
##  $ EN.ATM.GHGO.ZG      : num  -33.3 118.1 7308.3 -75 NA ...
##  $ EN.ATM.GHGT.KT.CE   : int  NA 98920 79730 10080 590 263240 365650 9360 NA 1210 ...
##  $ EN.ATM.GHGT.ZG      : num  NA 43.5 -18 -19.9 NA ...
```

```
##  $ EN.ATM.HFCG.KT.CE   : int  NA NA 31 105 NA 0 506 565 NA NA ...
##  $ EN.ATM.METH.KT.CE   : int  NA 81510 35520 3160 50 52960 117850 2430 NA 200 ...
##  $ EN.ATM.METH.ZG      : num  52.84 80.44 -13.97 3.82 NA ...
##  $ EN.ATM.NOXE.KT.CE   : int  NA 8960 16440 1100 0 7140 47930 1170 NA 20 ...
##  $ EN.ATM.NOXE.ZG      : num  -45.1 24.6 -79.5 -26.5 0 ...
##  $ EN.ATM.PFCG.KT.CE   : int  NA NA 0 0 NA 384 134 0 NA NA ...
##  $ EN.ATM.SF6G.KT.CE   : int  NA NA 0 0 NA 1038 444 0 NA NA ...
##  $ EN.CLC.DRSK.XQ      : num  NA NA NA NA NA NA 3.25 3 NA 2.75 ...
##  $ EN.CLC.GHGR.MT.CE   : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ EN.CLC.MDAT.ZS      : num  NA 1.06 1.01 5.27 NA ...
##  $ EN.POP.EL5M.RU.ZS   : num  2.351 NA 0.241 4.688 NA ...
##  $ EN.POP.EL5M.UR.ZS   : num  4.72 NA 1.19 2.38 NA ...
##  $ EN.POP.EL5M.ZS      : num  7.07 NA 1.44 7.07 NA ...
##  $ EN.URB.MCTY.TL.ZS   : num  NA 10.8 25.3 NA NA ...
##  $ ER.H2O.FWTL.K3      : num  NA 20.282 0.706 1.188 NA ...
##  $ ER.H2O.FWTL.ZS      : num  NA 43.016 0.477 4.416 NA ...
##  $ ER.LND.PTLD.ZS      : num  18.918 0.105 6.971 17.736 26.728 ...
##  $ ER.MRN.PTMR.ZS      : num  0.000105 NA 0.00493 2.718033 NA ...
##  $ ER.PTD.TOTL.ZS      : num  0.141 0.105 5.005 13.526 26.728 ...
##  $ IC.BUS.EASE.XQ      : int  NA 173 177 82 NA 16 126 47 NA 113 ...
##  $ IQ.CPA.PUBS.XQ      : num  NA 2.6 2.3 3.3 NA NA NA 3.7 NA NA ...
##  $ IS.ROD.PAVE.ZS      : num  NA NA 10.4 NA NA ...
##  $ NV.AGR.TOTL.ZS      : num  0.44 27.01 9.43 19.25 NA ...
##  $ SE.ENR.PRSC.FM.ZS   : num  1.022 0.636 0.63 1.019 NA ...
##  $ SE.PRM.CMPT.ZS      : num  101.2 85.6 46.2 103.3 NA ...
##  $ SH.DYN.MORT         : num  NA 60.3 74.7 9.7 3 7.5 9.3 11.8 NA 6.6 ...
##  $ SH.MED.CMHW.P3      : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ SH.STA.MALN.ZS      : num  NA 19.1 19 1.5 NA NA 1.7 2.6 NA NA ...
##  $ SI.POV.DDAY         : num  NA NA 49.9 1.3 NA 0 1.5 1.1 NA NA ...
##  $ SP.POP.GROW         : num  0.428 2.304 3.219 -0.578 0.154 ...
##  $ SP.POP.TOTL         : int  106766 38928341 32866268 2837743 77265 9890400 45376763 2963234 55197 9
##  $ SP.URB.GROW         : num  0.7746 3.3544 4.193 0.8539 0.0766 ...
##  $ SP.URB.TOTL         : int  46654 10131490 21962884 1762579 67928 8609395 41796990 1876112 48106 23
##  $ SP.URB.TOTL.IN.ZS   : num  43.7 26 66.8 62.1 87.9 ...
```

Handling Missing Values It is important to identify any missing values in the dataset before the analysis.The next step is handling these missing values. This should be done after a thorough analysis of the dataset as it might cause false interpretations of the dataset and loss of information.

Replacing missing values (NA) with zeros (0) can be suitable in some cases, but it really depends on the context and the nature of your data,In this analysis, missing values were handled by imputing the mean value for numerical variables, as on the context of my study and since the dataset represent climate change indicator values which cannot be zero.Replacing NA with zeros can mask the fact that climate data is missing. Since missing data is significant, I have imputed them with mean value. However,for three of the indicators which I have used in my study the imputing will interpret information differently and may results in anomalies. So that I have removed the records with missing values only for those indicator columns, "AG.LND.AGRI.ZS", "AG.LND.ARBL.ZS", "AG.LND.IRIG.AG.ZS" which are key response variables in my study.

Note: In order to impute the missing values with mean, I have converted the non numerical data(integer data) into numeric using a function convertToNum.

```
#load relevant libraries
library(ggplot2)
library( naniar )
```

```
## Warning: package 'naniar' was built under R version 4.1.3
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.1.3
## Warning: package 'tibble' was built under R version 4.1.3
## Warning: package 'tidyr' was built under R version 4.1.3
## Warning: package 'readr' was built under R version 4.1.3
## Warning: package 'purrr' was built under R version 4.1.3
## Warning: package 'stringr' was built under R version 4.1.3
## Warning: package 'forcats' was built under R version 4.1.3
## Warning: package 'lubridate' was built under R version 4.1.3
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v stringr   1.5.0
## v lubridate 1.9.2     v tibble    3.2.1
## v purrr     1.0.1     v tidyr     1.3.0
## v readr     2.1.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter()          masks stats::filter()
## x kableExtra::group_rows() masks dplyr::group_rows()
## x dplyr::lag()             masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(mice)
```

```
## Warning: package 'mice' was built under R version 4.1.3
##
## Attaching package: 'mice'
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```
library(dplyr)

#check for missing values
sum(is.na(climateData))
```

```
## [1] 3238
```

```
colSums(is.na(climateData))
```

```
##                iso3c              country       AG.LND.AGRI.K2
##                    0                    0                    6
##       AG.LND.AGRI.ZS       AG.LND.ARBL.ZS    AG.LND.EL5M.RU.K2
##                    6                    9                   38
##    AG.LND.EL5M.RU.ZS    AG.LND.EL5M.UR.K2    AG.LND.EL5M.UR.ZS
##                   38                   38                   38
##       AG.LND.EL5M.ZS       AG.LND.FRST.K2       AG.LND.FRST.ZS
##                   38                    3                    3
```

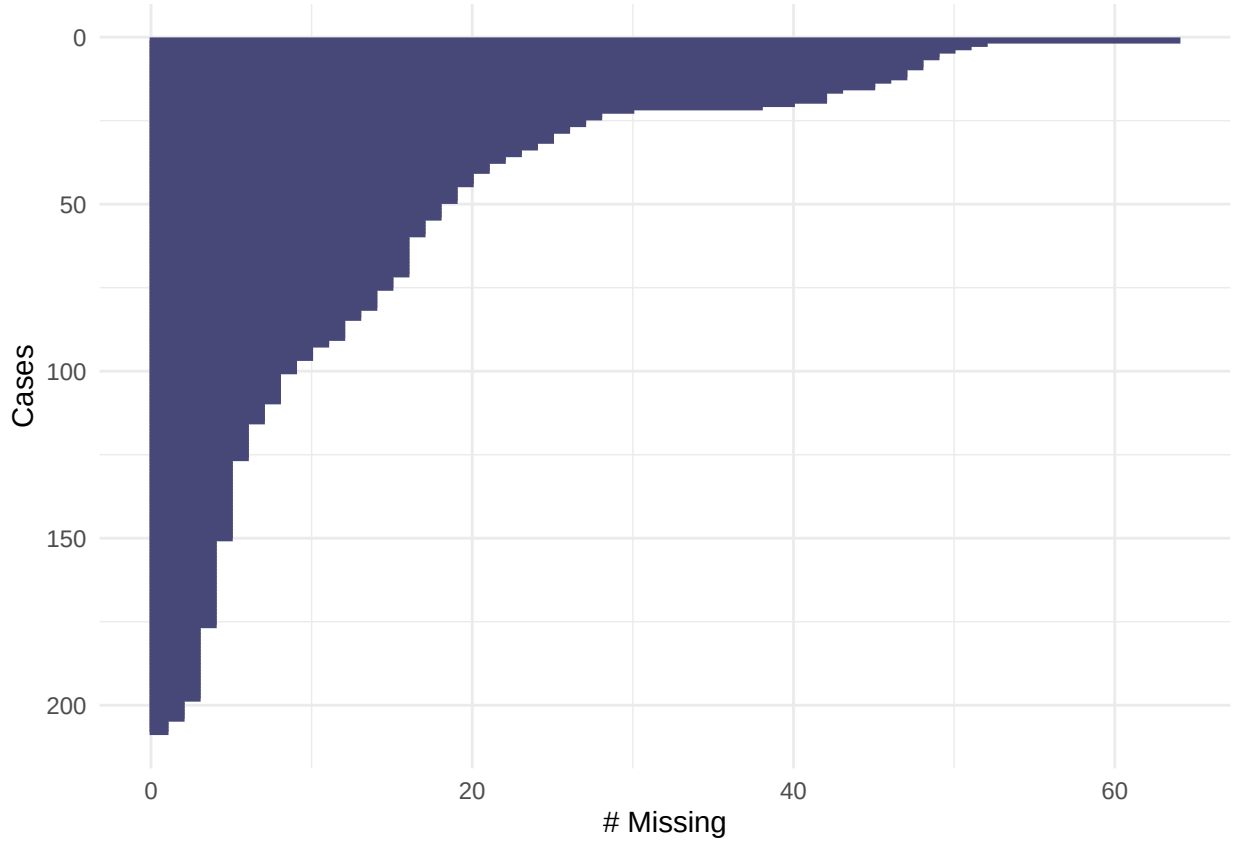
```
##    AG.LND.IRIG.AG.ZS       AG.LND.PRCP.MM       AG.YLD.CREL.KG
##                   93                   35                   36
## BX.KLT.DINV.WD.GD.ZS       EG.ELC.ACCS.ZS       EG.ELC.COAL.ZS
##                   19                    1                   76
##       EG.ELC.HYRO.ZS       EG.ELC.NGAS.ZS       EG.ELC.NUCL.ZS
##                   76                   76                   76
##       EG.ELC.PETR.ZS       EG.ELC.RNEW.ZS       EG.ELC.RNWX.KH
##                   76                    0                   76
##       EG.ELC.RNWX.ZS       EG.FEC.RNEW.ZS EG.USE.COMM.GD.PP.KD
##                   76                    4                   54
##    EG.USE.ELEC.KH.PC    EG.USE.PCAP.KG.OE    EN.ATM.CO2E.EG.ZS
##                   75                   45                   49
##    EN.ATM.CO2E.GF.KT    EN.ATM.CO2E.GF.ZS    EN.ATM.CO2E.KD.GD
##                   10                   26                   29
##       EN.ATM.CO2E.KT    EN.ATM.CO2E.LF.KT    EN.ATM.CO2E.LF.ZS
##                   26                   10                   26
##       EN.ATM.CO2E.PC    EN.ATM.CO2E.PP.GD EN.ATM.CO2E.PP.GD.KD
##                   26                   31                   35
##    EN.ATM.CO2E.SF.KT    EN.ATM.CO2E.SF.ZS    EN.ATM.GHGO.KT.CE
##                   10                   26                   33
##       EN.ATM.GHGO.ZG    EN.ATM.GHGT.KT.CE       EN.ATM.GHGT.ZG
##                   29                   26                   35
##    EN.ATM.HFCG.KT.CE    EN.ATM.METH.KT.CE       EN.ATM.METH.ZG
##                   80                   26                   13
##    EN.ATM.NOXE.KT.CE       EN.ATM.NOXE.ZG    EN.ATM.PFCG.KT.CE
##                   26                   12                   80
##    EN.ATM.SF6G.KT.CE       EN.CLC.DRSK.XQ    EN.CLC.GHGR.MT.CE
##                   80                  134                  155
##       EN.CLC.MDAT.ZS    EN.POP.EL5M.RU.ZS    EN.POP.EL5M.UR.ZS
##                   49                   38                   38
##       EN.POP.EL5M.ZS    EN.URB.MCTY.TL.ZS       ER.H2O.FWTL.K3
##                   38                   96                   36
##       ER.H2O.FWTL.ZS       ER.LND.PTLD.ZS       ER.MRN.PTMR.ZS
##                   42                    5                   47
##       ER.PTD.TOTL.ZS       IC.BUS.EASE.XQ       IQ.CPA.PUBS.XQ
##                    6                   28                  130
##       IS.ROD.PAVE.ZS       NV.AGR.TOTL.ZS    SE.ENR.PRSC.FM.ZS
##                  166                   15                   23
##       SE.PRM.CMPT.ZS          SH.DYN.MORT       SH.MED.CMHW.P3
##                   28                   24                  157
##       SH.STA.MALN.ZS          SI.POV.DDAY          SP.POP.GROW
##                   67                   55                    0
##          SP.POP.TOTL          SP.URB.GROW          SP.URB.TOTL
##                    0                    2                    2
##    SP.URB.TOTL.IN.ZS
##                    2
```

```
# Check for records with non NA values for columns "AG.LND.AGRI.ZS", "AG.LND.ARBL.ZS", "AG.LND.IRIG.AG.
complete_cases <- complete.cases(climateData[, c( "AG.LND.AGRI.ZS", "AG.LND.ARBL.ZS")])

# Subset the data frame to omit NA records
climateData <- climateData[complete_cases, ]
```

Here shows the visualization of the missing values.

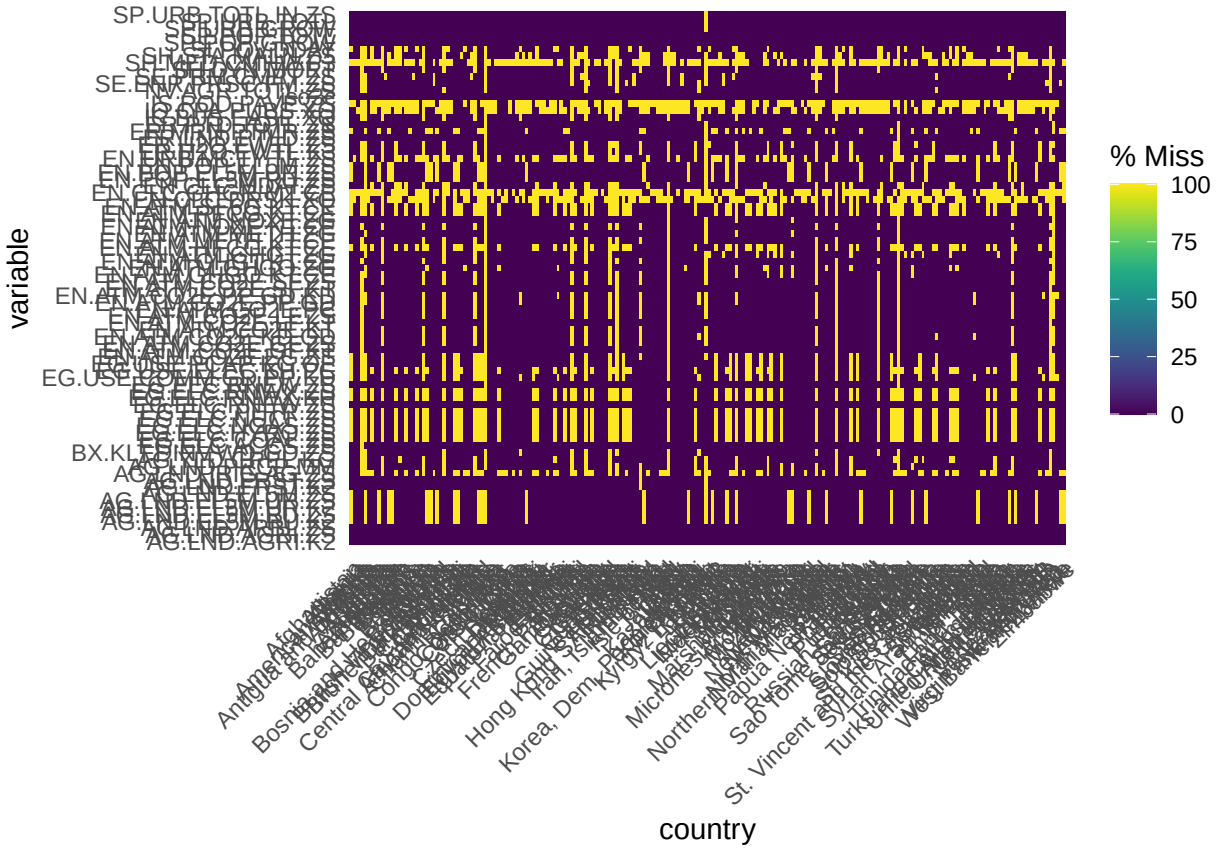
Below is a plot of cases (rows) with missing data. Each row is represented as a horizontal line, and the length of the line corresponds to the number of missing values in that case.

```
#visualization of missing values
gg_miss_case(climateData)
```



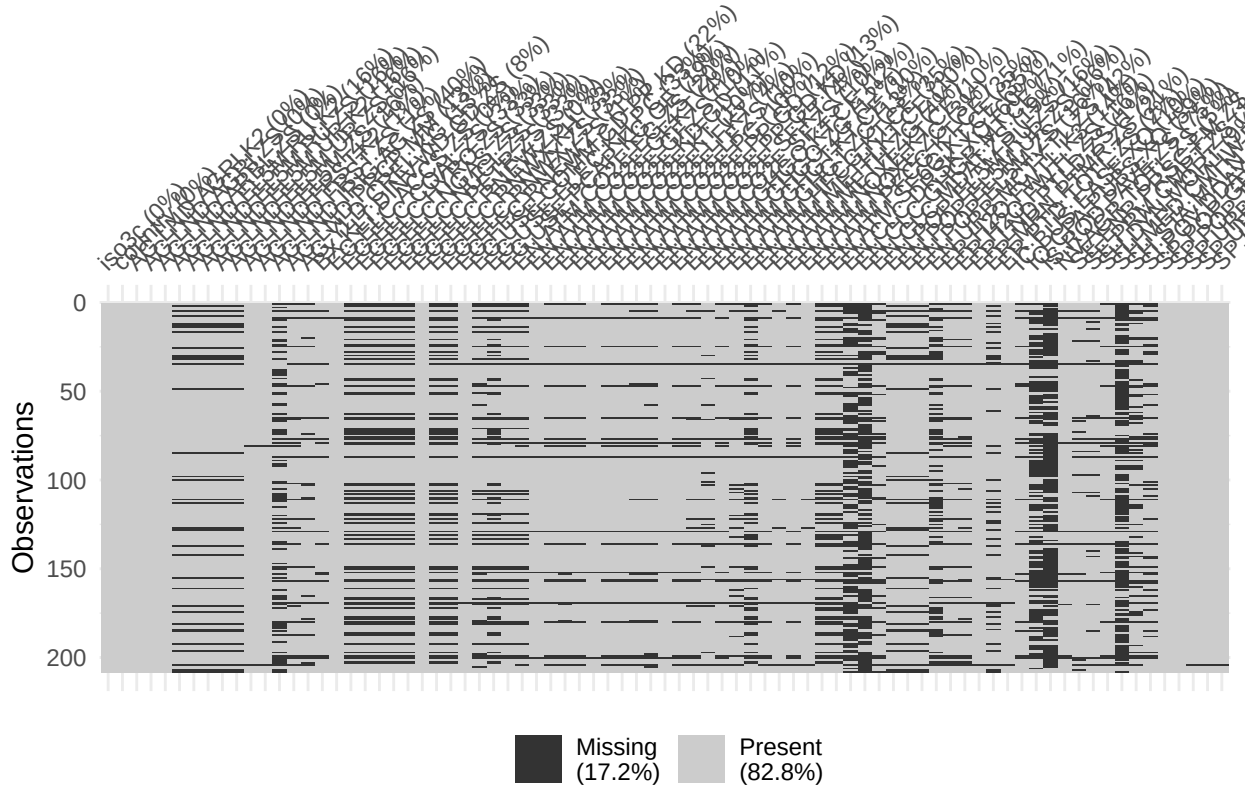
Below plot visualize missing data in a specific factor variable (country) in the dataset

```
gg_miss_fct(x = climateData, fct = country)
```



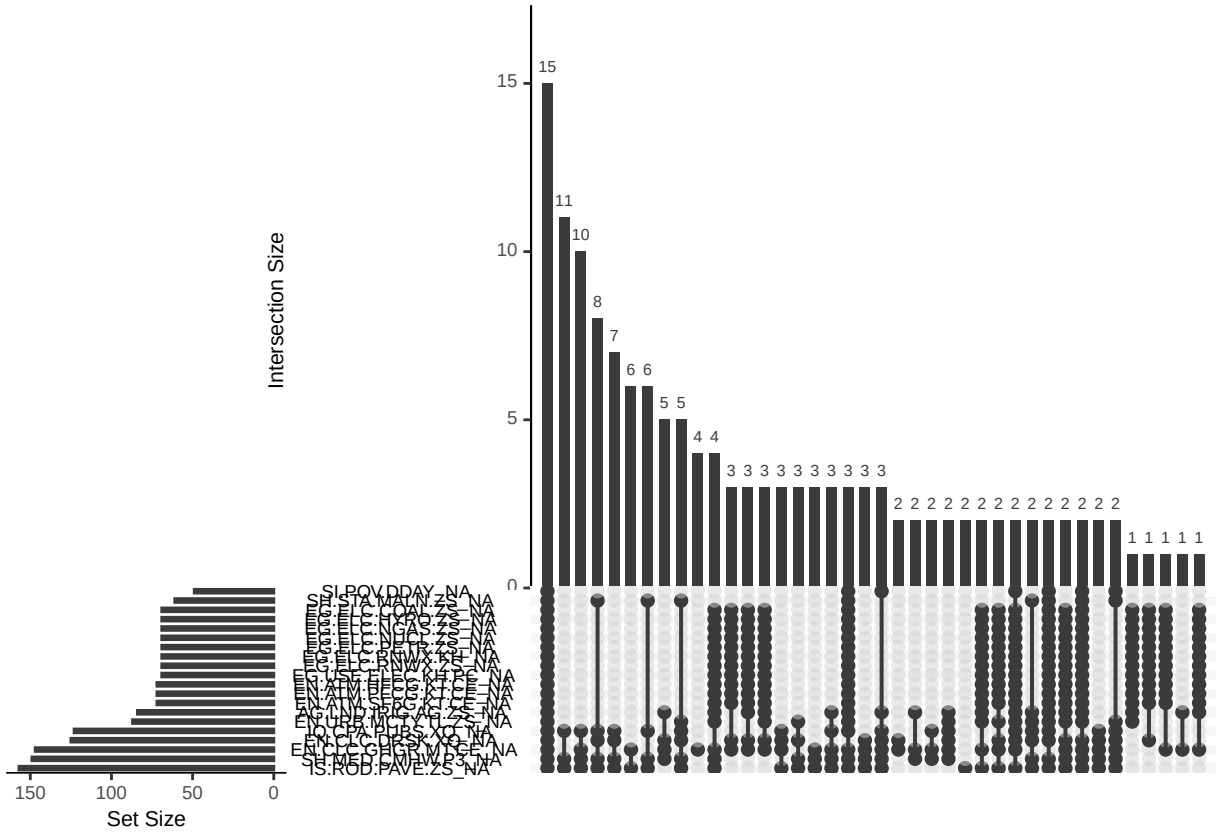
Below shows a mosaic plot that represents the presence or absence of missing data across all variables in the dataset. Each row in the plot represents a variable, and each column represents a case (row). Filled cells indicate the presence of data, while empty cells represent missing data.

```
vis_miss(climateData)
```



Below plot visualizes intersections of missing data across multiple variables (columns).

```
gg_miss_upset(climateData, nsets = 20)
```



Converted the non numerical data(integer data) into numeric using a function convertToNum.

```
#converted the non numerical data(integer data) into numeric
convertToNum <-function(column){
  if(class(column)=="integer"){
    return(as.numeric(column))
  } else {
    return(column)
  }
}

#NA replace with mean value
c.d<- lapply(climateData, convertToNum)

mean_values <- sapply(c.d[-1],mean, na.rm=TRUE)
```

```
## Warning in mean.default(X[[i]], ...): argument is not numeric or logical:
## returning NA
```

```
climateData[-1] <- lapply(c.d[-1], function(x) ifelse(is.na(x), mean_values, x))
head(climateData)
```

```
##   iso3c              country AG.LND.AGRI.K2 AG.LND.AGRI.ZS AG.LND.ARBL.ZS
## 1   ABW                Aruba          20.00      11.111111     11.1111111
## 2   AFG          Afghanistan      379190.00      58.081365     11.7988543
## 3   AGO               Angola      569524.90      45.682594      3.9303762
## 4   ALB              Albania       11740.81      42.849672     22.3118978
## 5   AND              Andorra         188.30      40.063830      1.7659574
```

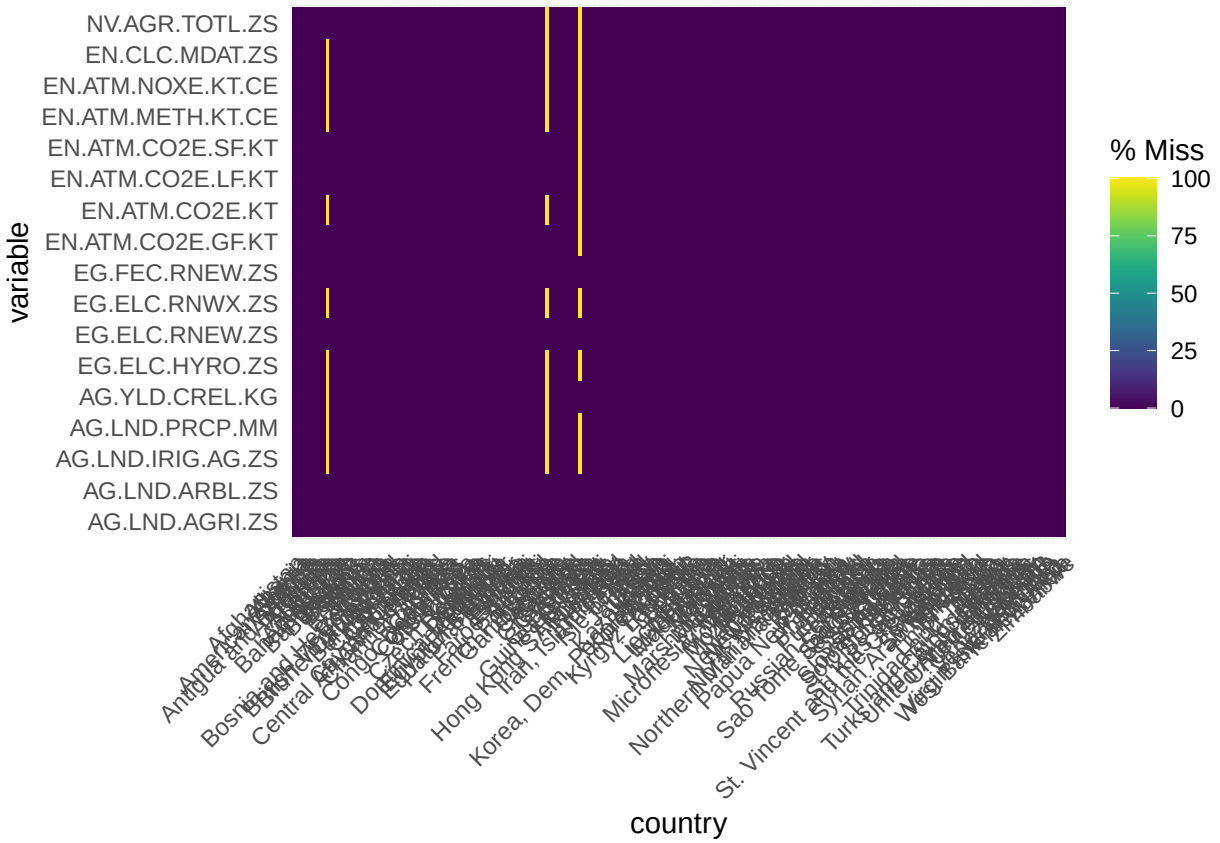
```
## 6   ARE United Arab Emirates        3817.50       5.375246      0.5956069
##   AG.LND.EL5M.RU.K2 AG.LND.EL5M.RU.ZS AG.LND.EL5M.UR.K2 AG.LND.EL5M.UR.ZS
## 1      4.278099e+00      2.352464e+00      8.579794e+00      4.717904e+00
## 2      2.291771e+05      2.291771e+05      2.291771e+05      2.291771e+05
## 3      1.516944e+03      1.213919e-01      5.413068e+01      4.331751e-03
## 4      1.294595e+03      4.562272e+00      8.484498e+01      2.990014e-01
## 5      6.626235e+03      6.626235e+03      6.626235e+03      6.626235e+03
## 6      2.611439e+03      3.299636e+00      1.198021e+03      1.513737e+00
##   AG.LND.EL5M.ZS AG.LND.FRST.K2 AG.LND.FRST.ZS AG.LND.IRIG.AG.ZS AG.LND.PRCP.MM
## 1   7.070368e+00            4.2       2.333333                NA             NA
## 2   2.291771e+05        12084.4       1.850994          5.134629        327.000
## 3   1.257237e-01       666073.8      53.426951         37.375560       1010.000
## 4   4.861273e+00         7889.0      28.791971         14.990448       1485.000
## 5   6.626235e+03          160.0      34.042553       6626.235431       6626.235
## 6   4.813373e+00         3173.0       4.467756         21.401441         78.000
##   AG.YLD.CREL.KG BX.KLT.DINV.WD.GD.ZS EG.ELC.ACCS.ZS EG.ELC.COAL.ZS
## 1             NA             4.244634      100.00000             NA
## 2       2164.900             0.121323       97.70000     229177.095
## 3        753.300            -4.583547       45.67031          0.000
## 4       4840.500             7.856693      100.00000          0.000
## 5       6626.235          6626.235431      100.00000       6626.235
## 6      27582.100             3.273826      100.00000          0.000
##   EG.ELC.HYRO.ZS EG.ELC.NGAS.ZS EG.ELC.NUCL.ZS EG.ELC.PETR.ZS EG.ELC.RNEW.ZS
## 1             NA             NA             NA             NA     14.8561614
## 2   229177.09460   229177.09460     229177.095   229177.09460     86.0501113
## 3       53.17493        0.00000          0.000       46.82507     53.1749283
## 4      100.00000        0.00000          0.000        0.00000    100.0000000
## 5     6626.23543     6626.23543       6626.235     6626.23543     86.1167002
## 6        0.00000       98.52551          0.000        1.24209      0.2324011
##   EG.ELC.RNWX.KH EG.ELC.RNWX.ZS EG.FEC.RNEW.ZS EG.USE.COMM.GD.PP.KD
## 1             NA             NA         8.0241                   NA
## 2   2.291771e+05   2.291771e+05        21.4227         229177.09460
## 3   0.000000e+00   0.000000e+00        56.7855             66.09474
## 4   0.000000e+00   0.000000e+00        38.2664             69.77348
## 5   6.626235e+03   6.626235e+03        18.5060           6626.23543
## 6   2.960000e+08   2.324011e-01         0.2592            122.61245
##   EG.USE.ELEC.KH.PC EG.USE.PCAP.KG.OE EN.ATM.CO2E.EG.ZS EN.ATM.CO2E.GF.KT
## 1                NA                NA                NA             0.000
## 2       229177.0946       229177.0946      2.291771e+05           319.029
## 3          312.2289          544.6094      2.360158e+00          1496.136
## 4         2309.3665          808.4558      2.234864e+00           161.348
## 5         6626.2354         6626.2354      6.626235e+03             0.000
## 6        11088.3419         7648.3941      2.608633e+00        135022.607
##   EN.ATM.CO2E.GF.ZS EN.ATM.CO2E.KD.GD EN.ATM.CO2E.KT EN.ATM.CO2E.LF.KT
## 1                NA                NA             NA           883.747
## 2          4.317037         0.3544900           7440          3476.316
## 3          4.225179         0.2743991          27340         17931.630
## 4          3.272779         0.3822239           5560          3362.639
## 5          0.000000         0.1323525            460           469.376
## 6         67.694078         0.5032381         200300         52335.424
##   EN.ATM.CO2E.LF.ZS EN.ATM.CO2E.PC EN.ATM.CO2E.PP.GD EN.ATM.CO2E.PP.GD.KD
## 1                NA             NA                NA                   NA
## 2          47.04081      0.2001511         0.0961047         9.841215e-02
## 3          50.64002      0.8873804         0.1249836         1.279845e-01
```

```
## 4          68.20769      1.9397316         0.1431908         1.455845e-01
## 5          99.86723      5.9734054      6626.2354306         6.626235e+03
## 6          26.23856     20.7974984         0.3032761         3.105577e-01
##   EN.ATM.CO2E.SF.KT EN.ATM.CO2E.SF.ZS EN.ATM.GHGO.KT.CE EN.ATM.GHGO.ZG
## 1             0.000                NA               NA      -33.33333
## 2          4829.439        65.3510014       -1800.00358      118.12500
## 3             0.000         0.0000000        5203.06714     7308.31228
## 4            14.668         0.2975254          11.68136      -75.01428
## 5             0.000         0.0000000        6626.23543     6626.23543
## 6          6985.635         3.5022736      -16708.68762       68.43972
##   EN.ATM.GHGT.KT.CE EN.ATM.GHGT.ZG EN.ATM.HFCG.KT.CE EN.ATM.METH.KT.CE
## 1                NA             NA                NA               NA
## 2             98920       43.52876        229177.095             81510
## 3             79730      -18.00827            31.000             35520
## 4             10080      -19.92390           105.000              3160
## 5               590     6626.23543          6626.235                50
## 6            263240      182.61068             0.000             52960
##   EN.ATM.METH.ZG EN.ATM.NOXE.KT.CE EN.ATM.NOXE.ZG EN.ATM.PFCG.KT.CE
## 1      52.835341                NA      -45.11953                NA
## 2      80.439484              8960       24.57941        229177.095
## 3     -13.973051             16440      -79.45488             0.000
## 4       3.818568              1100      -26.54102             0.000
## 5    6626.235431                 0        0.00000          6626.235
## 6      91.970807              7140      129.22388           384.000
##   EN.ATM.SF6G.KT.CE EN.CLC.DRSK.XQ EN.CLC.GHGR.MT.CE EN.CLC.MDAT.ZS
## 1                NA             NA                NA             NA
## 2        229177.095   2.291771e+05      2.291771e+05       1.059290
## 3             0.000   3.737556e+01      3.737556e+01       1.011765
## 4             0.000   1.409996e+01      1.409996e+01       5.269577
## 5          6626.235   6.626235e+03      6.626235e+03    6626.235431
## 6          1038.000   3.854346e+00      3.854346e+00       3.854346
##   EN.POP.EL5M.RU.ZS EN.POP.EL5M.UR.ZS EN.POP.EL5M.ZS EN.URB.MCTY.TL.ZS
## 1      2.351123e+00      4.718044e+00   7.069167e+00                NA
## 2      2.291771e+05      2.291771e+05   2.291771e+05          10.84437
## 3      2.411559e-01      1.194939e+00   1.436094e+00          25.34452
## 4      4.688154e+00      2.382807e+00   7.070961e+00          14.09996
## 5      6.626235e+03      6.626235e+03   6.626235e+03        6626.23543
## 6      1.618829e+00      1.316581e+01   1.478464e+01          61.12805
##   ER.H2O.FWTL.K3 ER.H2O.FWTL.ZS ER.LND.PTLD.ZS ER.MRN.PTMR.ZS ER.PTD.TOTL.ZS
## 1             NA             NA      18.917576   1.051567e-04      0.1411635
## 2        20.2820     43.0159067       0.104707   2.291771e+05      0.1047070
## 3         0.7057      0.4768243       6.971427   4.929827e-03      5.0047108
## 4         1.1880      4.4163569      17.736095   2.718033e+00     13.5256825
## 5      6626.2354   6626.2354306      26.727725   6.626235e+03     26.7277248
## 6         2.5620   1708.0000000      17.954921   1.127035e+01     15.0438948
##   IC.BUS.EASE.XQ IQ.CPA.PUBS.XQ IS.ROD.PAVE.ZS NV.AGR.TOTL.ZS SE.ENR.PRSC.FM.ZS
## 1             NA             NA             NA      0.4403979           1.02156
## 2        173.000       2.600000   2.291771e+05     27.0135608           0.63602
## 3        177.000       2.300000   1.040000e+01      9.4297730           0.63032
## 4         82.000       3.300000   1.409996e+01     19.2520985           1.01863
## 5       6626.235    6626.235431   6.626235e+03   6626.2354306        6626.23543
## 6         16.000       3.854346   3.854346e+00      0.7308402           0.94828
##   SE.PRM.CMPT.ZS SH.DYN.MORT SH.MED.CMHW.P3 SH.STA.MALN.ZS SI.POV.DDAY
## 1      101.18056          NA             NA             NA          NA
```

```
## 2       85.62533        60.3   2.291771e+05      19.100000  229177.095
## 3       46.18715        74.7   3.737556e+01      19.000000      49.900
## 4      103.32275         9.7   1.409996e+01       1.500000       1.300
## 5     6626.23543         3.0   6.626235e+03    6626.235431    6626.235
## 6      103.52020         7.5   3.854346e+00       3.854346       0.000
##   SP.POP.GROW SP.POP.TOTL SP.URB.GROW SP.URB.TOTL SP.URB.TOTL.IN.ZS
## 1   0.4280169      106766  0.77463055       46654            43.697
## 2   2.3038121    38928341  3.35442116    10131490            26.026
## 3   3.2185304    32866268  4.19296193    21962884            66.825
## 4  -0.5779423     2837743  0.85386648     1762579            62.112
## 5   0.1541341       77265  0.07658096       67928            87.916
## 6   1.2194287     9890400  1.51740240     8609395            87.048
```

Removing Unnecessary columns The whole set of indicators given in the dataset does not relevant to my study on assessing the agricultural sustainability of Australia in the context of climate change indicators. Thus, I have removed the unnecessary columns from the dataset while retaining the most important indicators for my study. These indicators are essential for assessing the climate-related characteristics of different countries, and the analysis of this data can potentially give some useful insights for my study.

```
#remove unnecessary columns
climateData_new <- climateData[,c("country", "AG.LND.AGRI.ZS", "AG.LND.ARBL.ZS",
                                  "AG.LND.IRIG.AG.ZS", "AG.LND.PRCP.MM","AG.YLD.CREL.KG",
                                  "EG.ELC.HYRO.ZS", "EG.ELC.RNEW.ZS","EG.ELC.RNWX.ZS",
                                  "EG.FEC.RNEW.ZS", "EN.ATM.CO2E.GF.KT", "EN.ATM.CO2E.KT", "EN.ATM
                                  "EN.ATM.CO2E.SF.KT", "EN.ATM.METH.KT.CE", "EN.ATM.NOXE.KT.CE", "

#checked new dataset for any missing data
gg_miss_fct(x = climateData_new, fct = country)
```



Removing duplicates

Checked for any duplicates and if any, they were removed.

```
#duplicates
duplicated(climateData)
```

```
##   [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [25] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [37] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [49] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [61] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [73] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [85] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [97] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [109] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [121] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [133] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [145] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [157] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [169] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [181] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [193] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [205] FALSE FALSE FALSE FALSE
```

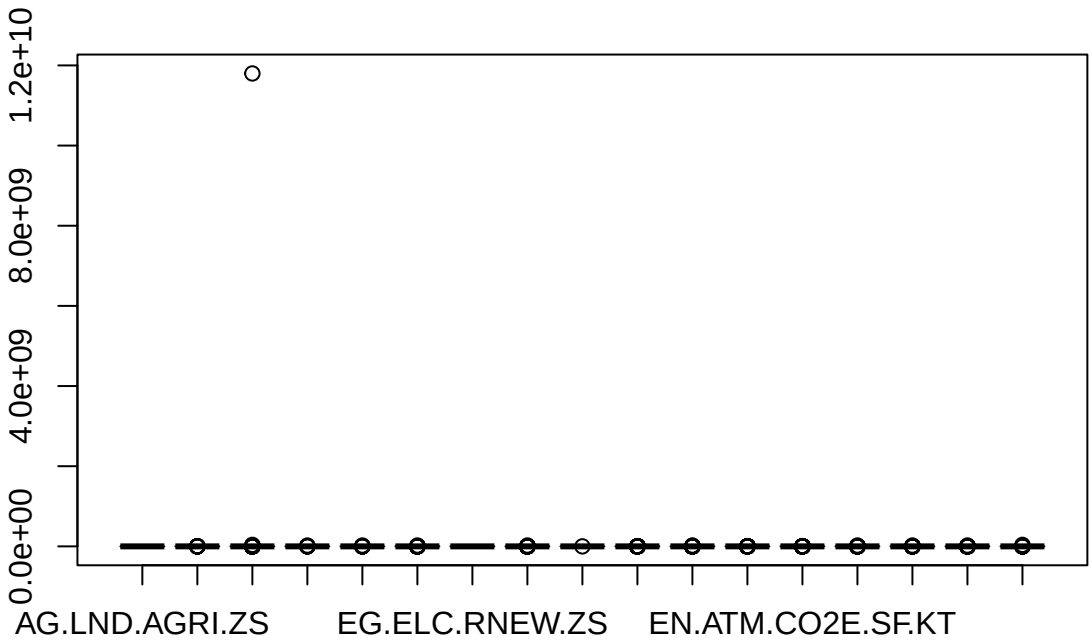
```
#find dxuplicates
dim(climateData[duplicated(climateData)])
```

```
## [1] 208   0
```

There were no duplicates in the dataset.

Outliers Outliers, if present, were identified.

```
boxplot(climateData_new[-1])
```



Below shows a snapshot of the new dataset after pre processing.

```
head(climateData_new)
```

```
##                country AG.LND.AGRI.ZS AG.LND.ARBL.ZS AG.LND.IRIG.AG.ZS
## 1                Aruba      11.111111     11.1111111                NA
## 2          Afghanistan      58.081365     11.7988543          5.134629
## 3               Angola      45.682594      3.9303762         37.375560
## 4              Albania      42.849672     22.3118978         14.990448
## 5              Andorra      40.063830      1.7659574       6626.235431
## 6 United Arab Emirates       5.375246      0.5956069         21.401441
##   AG.LND.PRCP.MM AG.YLD.CREL.KG EG.ELC.HYRO.ZS EG.ELC.RNEW.ZS EG.ELC.RNWX.ZS
## 1             NA             NA             NA     14.8561614             NA
## 2        327.000       2164.900   229177.09460     86.0501113   2.291771e+05
## 3       1010.000        753.300       53.17493     53.1749283   0.000000e+00
## 4       1485.000       4840.500      100.00000    100.0000000   0.000000e+00
## 5       6626.235       6626.235     6626.23543     86.1167002   6.626235e+03
## 6         78.000      27582.100        0.00000      0.2324011   2.324011e-01
##   EG.FEC.RNEW.ZS EN.ATM.CO2E.GF.KT EN.ATM.CO2E.KT EN.ATM.CO2E.LF.KT
## 1         8.0241             0.000             NA           883.747
```

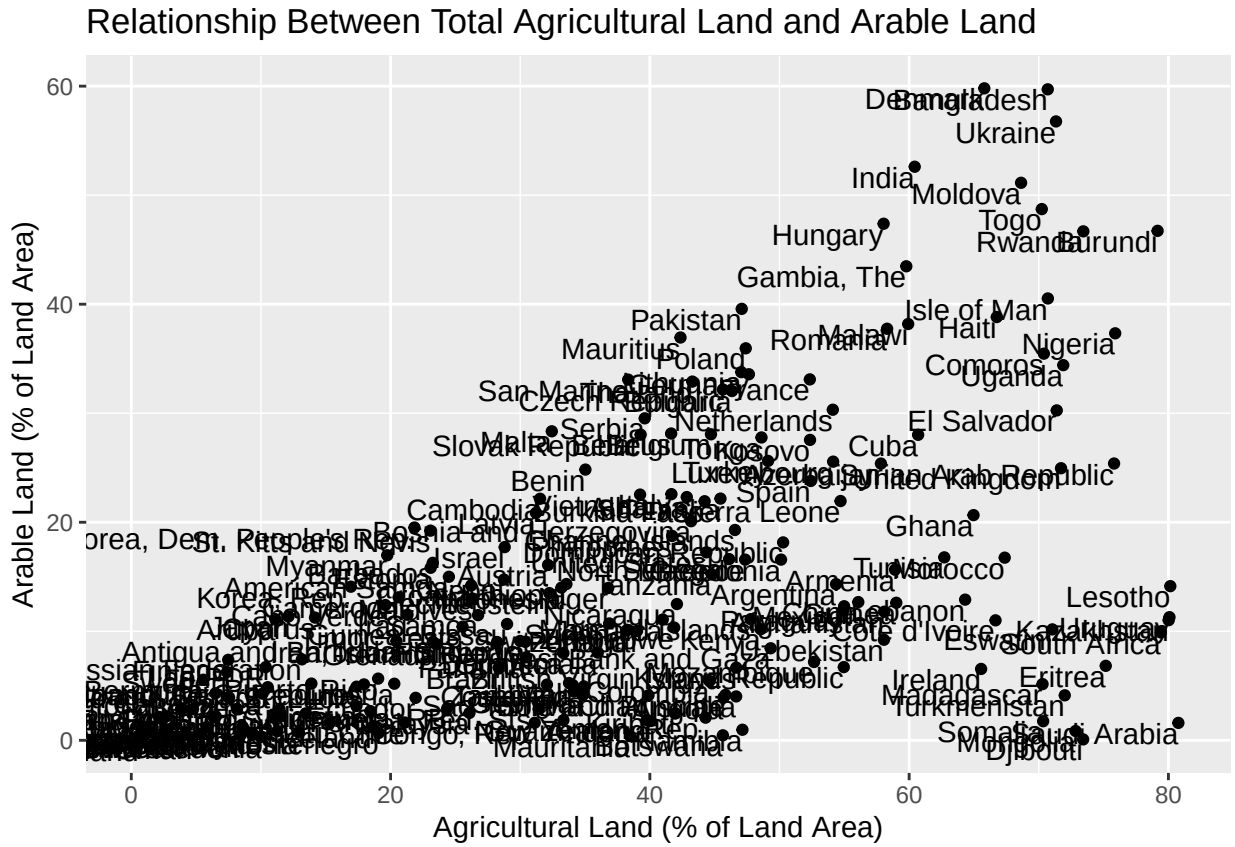
```
## 2        21.4227           319.029           7440          3476.316
## 3        56.7855          1496.136          27340         17931.630
## 4        38.2664           161.348           5560          3362.639
## 5        18.5060             0.000            460           469.376
## 6         0.2592        135022.607         200300         52335.424
##   EN.ATM.CO2E.SF.KT EN.ATM.METH.KT.CE EN.ATM.NOXE.KT.CE EN.CLC.MDAT.ZS
## 1             0.000                NA                NA             NA
## 2          4829.439             81510              8960       1.059290
## 3             0.000             35520             16440       1.011765
## 4            14.668              3160              1100       5.269577
## 5             0.000                50                 0    6626.235431
## 6          6985.635             52960              7140       3.854346
##   NV.AGR.TOTL.ZS
## 1      0.4403979
## 2     27.0135608
## 3      9.4297730
## 4     19.2520985
## 5   6626.2354306
## 6      0.7308402
```

```
#converted to a dataframe for the further analysis
climateData_df <- data.frame(climateData_new)
```

Methodology

In this section, I have outlined the statistical learning methods employed to process and analyze the data in the context of my study. Initially the following illustrate the relationship between "Agricultural land (% of land area)" with "Arable land (% of land area)". This comparison allows us to understand the relationship between total agricultural land and land specifically used for crops.

```
#compare compare the "Agricultural land (% of land area)" with "Arable land (% of land area)" to unders
#land and land specifically used for crops in differebt countries.
ggplot(climateData_df, aes(x = AG.LND.AGRI.ZS, y = AG.LND.ARBL.ZS)) +
  geom_point() +
  geom_text(aes(label = country), hjust = 1, vjust = 1) +
  labs(
    x = "Agricultural Land (% of Land Area)",
    y = "Arable Land (% of Land Area)",
    title = "Relationship Between Total Agricultural Land and Arable Land",
  )
```



Interpretation and Discussion Plot from the comparison between "Agricultural land (% of land area)" with "Arable land (% of land area)" indicates that there is a significant relationship between these two indicators and every country used to have a land specifically assigned for crop farming. Arable land grows with Agricultural land which has been assigned for a particular country from its total land area.

Below analysis shows a comparison of the global climate change indicator values of Australia with selected countries which shows a reasonable interest towards agricultural field in recent ages. Since the study focus on Australia, following visualization will give a brief overview on the position of Australia in the world regarding agricultural sustainability.

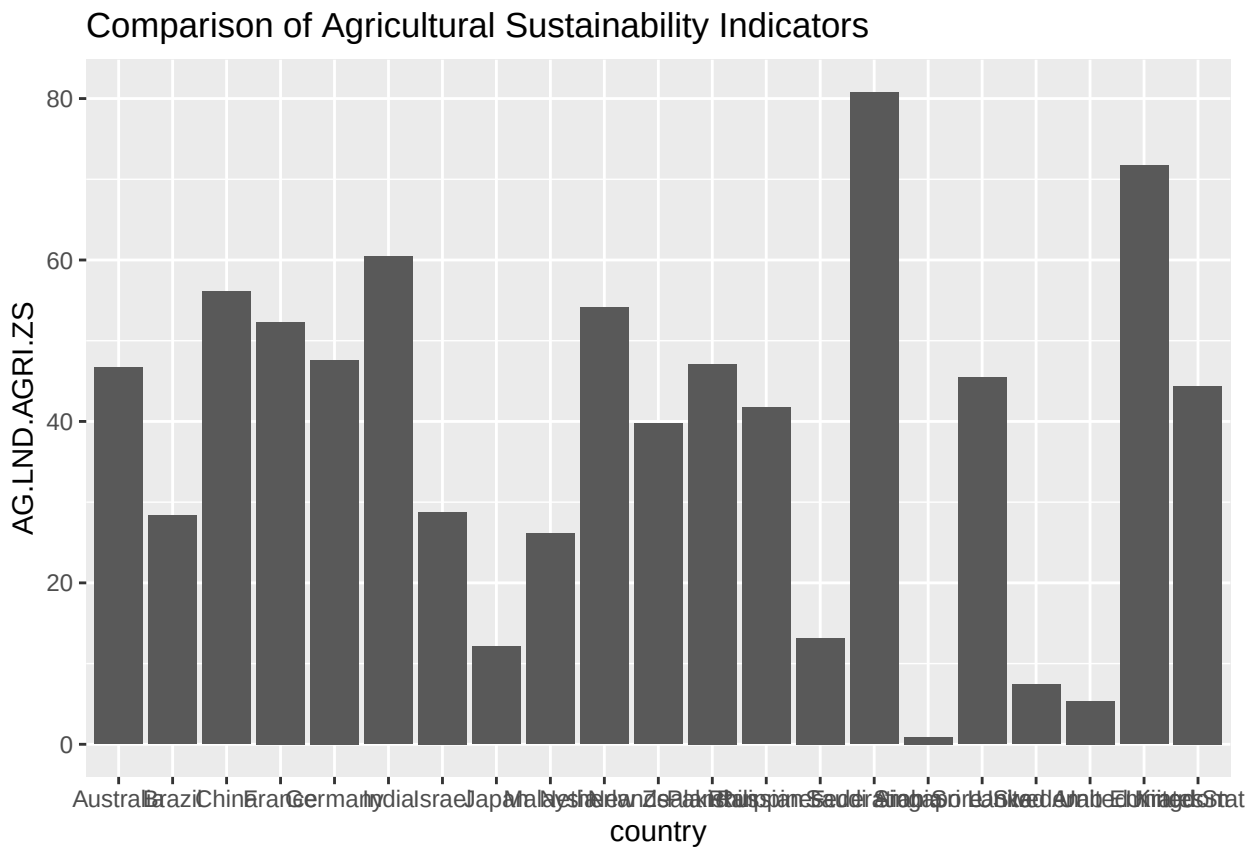
```
summary(climateData_df[climateData_df$country == "Australia", ])
```

```
##    country          AG.LND.AGRI.ZS  AG.LND.ARBL.ZS  AG.LND.IRIG.AG.ZS
##  Length:1           Min.   :46.66   Min.   :4.027   Min.   :0.6403
##  Class :character   1st Qu.:46.66   1st Qu.:4.027   1st Qu.:0.6403
##  Mode  :character   Median :46.66   Median :4.027   Median :0.6403
##                     Mean   :46.66   Mean   :4.027   Mean   :0.6403
##                     3rd Qu.:46.66   3rd Qu.:4.027   3rd Qu.:0.6403
##                     Max.   :46.66   Max.   :4.027   Max.   :0.6403
##  AG.LND.PRCP.MM AG.YLD.CREL.KG EG.ELC.HYRO.ZS  EG.ELC.RNEW.ZS  EG.ELC.RNWX.ZS
##  Min.   :534    Min.   :2036   Min.   :5.296   Min.   :13.64   Min.   :8.342
##  1st Qu.:534    1st Qu.:2036   1st Qu.:5.296   1st Qu.:13.64   1st Qu.:8.342
##  Median :534    Median :2036   Median :5.296   Median :13.64   Median :8.342
##  Mean   :534    Mean   :2036   Mean   :5.296   Mean   :13.64   Mean   :8.342
##  3rd Qu.:534    3rd Qu.:2036   3rd Qu.:5.296   3rd Qu.:13.64   3rd Qu.:8.342
##  Max.   :534    Max.   :2036   Max.   :5.296   Max.   :13.64   Max.   :8.342
##  EG.FEC.RNEW.ZS  EN.ATM.CO2E.GF.KT EN.ATM.CO2E.KT   EN.ATM.CO2E.LF.KT
##  Min.   :9.645   Min.   :79123     Min.   :386620   Min.   :118910
```

```
##  1st Qu.:9.645   1st Qu.:79123     1st Qu.:386620   1st Qu.:118910
##  Median :9.645   Median :79123     Median :386620   Median :118910
##  Mean   :9.645   Mean   :79123     Mean   :386620   Mean   :118910
##  3rd Qu.:9.645   3rd Qu.:79123     3rd Qu.:386620   3rd Qu.:118910
##  Max.   :9.645   Max.   :79123     Max.   :386620   Max.   :118910
##  EN.ATM.CO2E.SF.KT EN.ATM.METH.KT.CE EN.ATM.NOXE.KT.CE EN.CLC.MDAT.ZS
##  Min.   :172598    Min.   :139070    Min.   :76760     Min.   :3.047
##  1st Qu.:172598    1st Qu.:139070    1st Qu.:76760     1st Qu.:3.047
##  Median :172598    Median :139070    Median :76760     Median :3.047
##  Mean   :172598    Mean   :139070    Mean   :76760     Mean   :3.047
##  3rd Qu.:172598    3rd Qu.:139070    3rd Qu.:76760     3rd Qu.:3.047
##  Max.   :172598    Max.   :139070    Max.   :76760     Max.   :3.047
##  NV.AGR.TOTL.ZS
##  Min.   :1.898
##  1st Qu.:1.898
##  Median :1.898
##  Mean   :1.898
##  3rd Qu.:1.898
##  Max.   :1.898
```

Comparison of Australia with other set of countries

```
#bar chart comparing Australia to a few countries
countries_to_compare <- c("Australia","United Arab Emirates", "Brazil", "China", "Germany", "France", "

comparison_data <- climateData_df[climateData_df$country %in% countries_to_compare, ]
ggplot(comparison_data, aes(x = country, y = AG.LND.AGRI.ZS)) +
  geom_bar(stat = "identity") +
  labs(title = "Comparison of Agricultural Sustainability Indicators")
```



######Interpretation and Discussion In the context of the above study, Australia which the pioneer focus of the study which assess its agricultural sustainability, indicates that country has about 50% of land area for agriculture. This indicates a good sign for its agricultural sustainability.

Explotary Data Analysis

Correlation Analysis Following code shows the calculation for correlations between different indicators to identify relationships between indicators.

```
# Calculate correlation matrix
(correlation_matrix <- cor(climateData_df[-1], use = "complete.obs"))
```

```
##                   AG.LND.AGRI.ZS AG.LND.ARBL.ZS AG.LND.IRIG.AG.ZS
## AG.LND.AGRI.ZS       1.0000000000     0.56416847      -0.019792069
## AG.LND.ARBL.ZS       0.5641684694     1.00000000       0.041141411
## AG.LND.IRIG.AG.ZS   -0.0197920688     0.04114141       1.000000000
## AG.LND.PRCP.MM      -0.1182356520    -0.07460141      -0.003217479
## AG.YLD.CREL.KG      -0.1183677933    -0.07451871      -0.003237244
## EG.ELC.HYRO.ZS      -0.1192687760    -0.07544733      -0.003562721
## EG.ELC.RNEW.ZS       0.0265478530    -0.04387623       0.032010228
## EG.ELC.RNWX.ZS      -0.1192666608    -0.07544345      -0.003564226
## EG.FEC.RNEW.ZS       0.0650159051     0.12744027       0.021814899
## EN.ATM.CO2E.GF.KT   -0.0001263166     0.01319974      -0.018376705
## EN.ATM.CO2E.KT      -0.0680781830    -0.04139307      -0.009276841
## EN.ATM.CO2E.LF.KT    0.0606079231     0.06195136      -0.016503680
## EN.ATM.CO2E.SF.KT    0.0786361826     0.04630478      -0.009063822
## EN.ATM.METH.KT.CE   -0.1113842332    -0.06783778      -0.004147276
```

```
## EN.ATM.NOXE.KT.CE  -0.1146341174    -0.07139254      -0.003665546
## EN.CLC.MDAT.ZS     -0.1197765314    -0.07557246      -0.003360324
## NV.AGR.TOTL.ZS     -0.0502756769     0.02718972      -0.001809057
##                   AG.LND.PRCP.MM AG.YLD.CREL.KG EG.ELC.HYRO.ZS EG.ELC.RNEW.ZS
## AG.LND.AGRI.ZS      -0.118235652   -0.118367793   -0.119268776     0.02654785
## AG.LND.ARBL.ZS      -0.074601408   -0.074518707   -0.075447329    -0.04387623
## AG.LND.IRIG.AG.ZS   -0.003217479   -0.003237244   -0.003562721     0.03201023
## AG.LND.PRCP.MM       1.000000000    0.999997206    0.999724435     0.10678454
## AG.YLD.CREL.KG       0.999997206    1.000000000    0.999718091     0.10659711
## EG.ELC.HYRO.ZS       0.999724435    0.999718091    1.000000000     0.10519869
## EG.ELC.RNEW.ZS       0.106784545    0.106597113    0.105198685     1.00000000
## EG.ELC.RNWX.ZS       0.999724393    0.999718053    1.000000000     0.10518657
## EG.FEC.RNEW.ZS      -0.020191734   -0.020507301   -0.020284875     0.11316982
## EN.ATM.CO2E.GF.KT   -0.018694633   -0.018175212   -0.019771886    -0.11301407
## EN.ATM.CO2E.KT       0.868262594    0.868440887    0.867616154     0.05895482
## EN.ATM.CO2E.LF.KT   -0.019006185   -0.018536571   -0.020111587    -0.08077017
## EN.ATM.CO2E.SF.KT   -0.009602535   -0.009359841   -0.010155462    -0.04626832
## EN.ATM.METH.KT.CE    0.996173956    0.996196837    0.995831225     0.10251484
## EN.ATM.NOXE.KT.CE    0.999425132    0.999432665    0.999114853     0.10559140
## EN.CLC.MDAT.ZS       0.999895821    0.999895991    0.999701813     0.10526845
## NV.AGR.TOTL.ZS      -0.004963869   -0.004946467   -0.005267247     0.08755977
##                   EG.ELC.RNWX.ZS EG.FEC.RNEW.ZS EN.ATM.CO2E.GF.KT
## AG.LND.AGRI.ZS      -0.119266661    0.065015905     -0.0001263166
## AG.LND.ARBL.ZS      -0.075443446    0.127440272      0.0131997396
## AG.LND.IRIG.AG.ZS   -0.003564226    0.021814899     -0.0183767048
## AG.LND.PRCP.MM       0.999724393   -0.020191734     -0.0186946332
## AG.YLD.CREL.KG       0.999718053   -0.020507301     -0.0181752121
## EG.ELC.HYRO.ZS       1.000000000   -0.020284875     -0.0197718855
## EG.ELC.RNEW.ZS       0.105186568    0.113169816     -0.1130140741
## EG.ELC.RNWX.ZS       1.000000000   -0.020286551     -0.0197705991
## EG.FEC.RNEW.ZS      -0.020286551    1.000000000     -0.0767770420
## EN.ATM.CO2E.GF.KT   -0.019770599   -0.076777042      1.0000000000
## EN.ATM.CO2E.KT       0.867616464   -0.042511785      0.2914266815
## EN.ATM.CO2E.LF.KT   -0.020110529   -0.063047454      0.8736191388
## EN.ATM.CO2E.SF.KT   -0.010155014   -0.032950008      0.3838717465
## EN.ATM.METH.KT.CE    0.995831135   -0.024031210      0.0395997362
## EN.ATM.NOXE.KT.CE    0.999114783   -0.021116141     -0.0010113123
## EN.CLC.MDAT.ZS       0.999701786   -0.020778935     -0.0184494945
## NV.AGR.TOTL.ZS      -0.005270088   -0.001788174     -0.0183783381
##                   EN.ATM.CO2E.KT EN.ATM.CO2E.LF.KT EN.ATM.CO2E.SF.KT
## AG.LND.AGRI.ZS      -0.068078183       0.060607923       0.078636183
## AG.LND.ARBL.ZS      -0.041393071       0.061951360       0.046304777
## AG.LND.IRIG.AG.ZS   -0.009276841      -0.016503680      -0.009063822
## AG.LND.PRCP.MM       0.868262594      -0.019006185      -0.009602535
## AG.YLD.CREL.KG       0.868440887      -0.018536571      -0.009359841
## EG.ELC.HYRO.ZS       0.867616154      -0.020111587      -0.010155462
## EG.ELC.RNEW.ZS       0.058954818      -0.080770166      -0.046268321
## EG.ELC.RNWX.ZS       0.867616464      -0.020110529      -0.010155014
## EG.FEC.RNEW.ZS      -0.042511785      -0.063047454      -0.032950008
## EN.ATM.CO2E.GF.KT    0.291426681       0.873619139       0.383871746
## EN.ATM.CO2E.KT       1.000000000       0.403568529       0.467443926
## EN.ATM.CO2E.LF.KT    0.403568529       1.000000000       0.669816705
## EN.ATM.CO2E.SF.KT    0.467443926       0.669816705       1.000000000
## EN.ATM.METH.KT.CE    0.902463302       0.048604398       0.060025364
```

```
## EN.ATM.NOXE.KT.CE    0.882897793       0.007368761       0.019596488
## EN.CLC.MDAT.ZS       0.868087171      -0.019353151      -0.009774043
## NV.AGR.TOTL.ZS      -0.010480323      -0.017711083      -0.006441388
##                    EN.ATM.METH.KT.CE EN.ATM.NOXE.KT.CE EN.CLC.MDAT.ZS
## AG.LND.AGRI.ZS         -0.111384233      -0.114634117   -0.119776531
## AG.LND.ARBL.ZS         -0.067837784      -0.071392537   -0.075572455
## AG.LND.IRIG.AG.ZS      -0.004147276      -0.003665546   -0.003360324
## AG.LND.PRCP.MM          0.996173956       0.999425132    0.999895821
## AG.YLD.CREL.KG          0.996196837       0.999432665    0.999895991
## EG.ELC.HYRO.ZS          0.995831225       0.999114853    0.999701813
## EG.ELC.RNEW.ZS          0.102514844       0.105591400    0.105268449
## EG.ELC.RNWX.ZS          0.995831135       0.999114783    0.999701786
## EG.FEC.RNEW.ZS         -0.024031210      -0.021116141   -0.020778935
## EN.ATM.CO2E.GF.KT       0.039599736      -0.001011312   -0.018449494
## EN.ATM.CO2E.KT          0.902463302       0.882897793    0.868087171
## EN.ATM.CO2E.LF.KT       0.048604398       0.007368761   -0.019353151
## EN.ATM.CO2E.SF.KT       0.060025364       0.019596488   -0.009774043
## EN.ATM.METH.KT.CE       1.000000000       0.998273578    0.996037941
## EN.ATM.NOXE.KT.CE       0.998273578       1.000000000    0.999308704
## EN.CLC.MDAT.ZS          0.996037941       0.999308704    1.000000000
## NV.AGR.TOTL.ZS         -0.005931312      -0.005504337   -0.005065895
##                    NV.AGR.TOTL.ZS
## AG.LND.AGRI.ZS      -0.050275677
## AG.LND.ARBL.ZS       0.027189722
## AG.LND.IRIG.AG.ZS   -0.001809057
## AG.LND.PRCP.MM      -0.004963869
## AG.YLD.CREL.KG      -0.004946467
## EG.ELC.HYRO.ZS      -0.005267247
## EG.ELC.RNEW.ZS       0.087559770
## EG.ELC.RNWX.ZS      -0.005270088
## EG.FEC.RNEW.ZS      -0.001788174
## EN.ATM.CO2E.GF.KT   -0.018378338
## EN.ATM.CO2E.KT      -0.010480323
## EN.ATM.CO2E.LF.KT   -0.017711083
## EN.ATM.CO2E.SF.KT   -0.006441388
## EN.ATM.METH.KT.CE   -0.005931312
## EN.ATM.NOXE.KT.CE   -0.005504337
## EN.CLC.MDAT.ZS      -0.005065895
## NV.AGR.TOTL.ZS       1.000000000
```

Interpretation and Discussion

From the correlation analysis, the study was able to agree on the following insights.

1.There is a very strong positive correlation (approximately 0.999995) between the percentage of agricultural land (AG.LND.AGRI.ZS) and the percentage of electricity generated from hydroelectric power (EG.ELC.HYRO.ZS). This indicates that countries with more agricultural land tend to produce a higher proportion of their electricity from hydroelectric sources.

2.There is a moderately strong negative correlation (approximately -0.0427) between agricultural land and the percentage of electricity generated from renewable sources (EG.ELC.RNEW.ZS). This implies that a countries with more agricultural land may have a slightly lower percentage of electricity from renewable sources.

3.The variables related to renewable energy (EG.ELC.RNEW.ZS) have negative correlations with various carbon emissions indicators (EN.ATM.CO2E.GF.KT, EN.ATM.CO2E.KT, EN.ATM.CO2E.LF.KT, and

EN.ATM.CO2E.SF.KT). This means that countries with higher renewable energy generation tend to have lower carbon emissions from different sources.

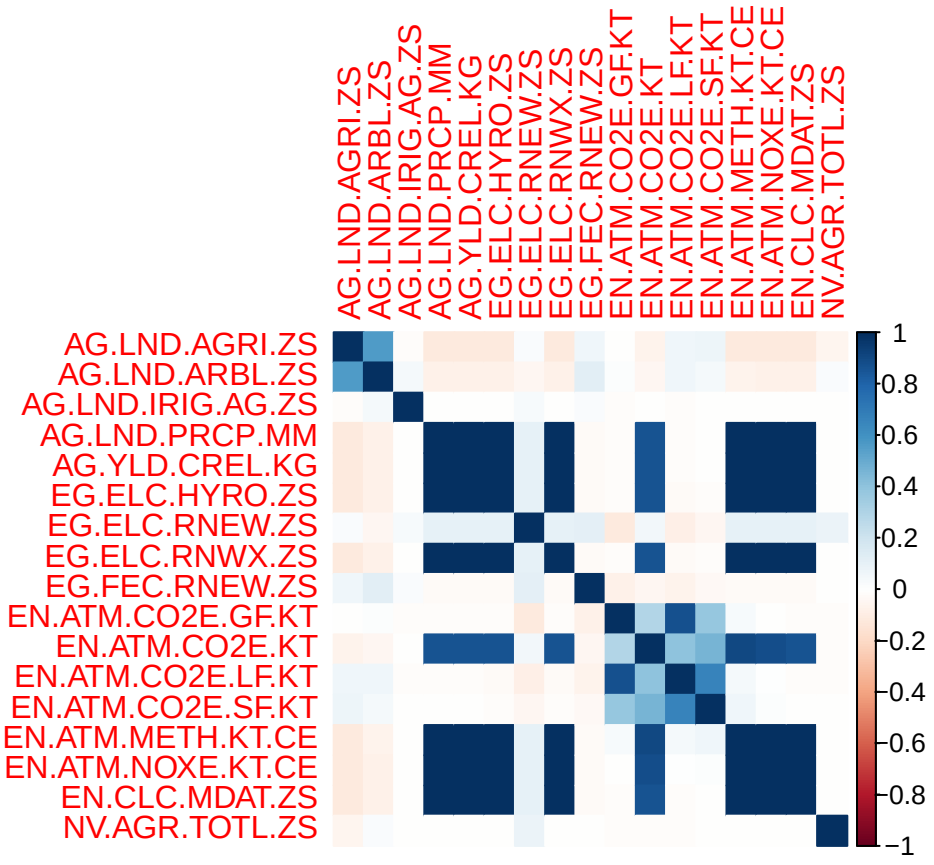
4.There is a very strong positive correlation (approximately 0.999995) between the percentage of irrigated agricultural land (AG.LND.IRIG.AG.ZS) and the percentage of electricity generated from hydroelectric power (EG.ELC.HYRO.ZS). This implies a close relationship between the extent of irrigated agricultural land and hydroelectric power generation.

5.There is a negative correlation between the percentage of agricultural land and methane emissions, although it's relatively weak (approximately -0.0058). This indicates that regions with larger areas of agricultural land may have slightly lower methane emissions.

6.There is a weak positive correlation (approximately 0.0061) between agricultural land and nitrogen oxide emissions. This suggests that regions with more agricultural land may have slightly higher nitrogen oxide emissions.

```
# Visualize correlation matrix
corrplot(correlation_matrix, method = "color")
```

Heatmap for correlation



Multiple Linear Regression with Regularisation Following regression analysis was conducted in order to assess the relationship between percentage of arable land, percentage of agricultural land and irrigation land with other indicators and how they impact on these key response variables.

Fitting the linear model for response variables: "AG.LND.AGRI.ZS": Agricultural land (% of land area)

```r
climate.lm <-lm(formula = climateData_df$AG.LND.AGRI.ZS ~ climateData_df$AG.LND.PRCP.MM+climateData_df$A
                  climateData_df$EG.ELC.RNEW.ZS+climateData_df$EG.ELC.RNWX.ZS+climateData_df$EG.FEC.RNEW
                  climateData_df$EN.ATM.CO2E.KT+climateData_df$EN.ATM.CO2E.LF.KT+climateData_df$EN.ATM.C
                  climateData_df$EN.ATM.NOXE.KT.CE+climateData_df$EN.CLC.MDAT.ZS+climateData_df$NV.AGR.T

summary(climate.lm)
```

```
##
## Call:
## lm(formula = climateData_df$AG.LND.AGRI.ZS ~ climateData_df$AG.LND.PRCP.MM +
##     climateData_df$AG.YLD.CREL.KG + climateData_df$EG.ELC.HYRO.ZS +
##     climateData_df$EG.ELC.RNEW.ZS + climateData_df$EG.ELC.RNWX.ZS +
##     climateData_df$EG.FEC.RNEW.ZS + climateData_df$EN.ATM.CO2E.GF.KT +
##     climateData_df$EN.ATM.CO2E.KT + climateData_df$EN.ATM.CO2E.LF.KT +
##     climateData_df$EN.ATM.CO2E.SF.KT + climateData_df$EN.ATM.METH.KT.CE +
##     climateData_df$EN.ATM.NOXE.KT.CE + climateData_df$EN.CLC.MDAT.ZS +
##     climateData_df$NV.AGR.TOTL.ZS, data = climateData_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -44.36 -15.83  -0.07  13.51  48.95
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      3.665e+01  2.555e+00  14.342   <2e-16 ***
## climateData_df$AG.LND.PRCP.MM    3.384e-04  4.681e-04   0.723   0.4706
## climateData_df$AG.YLD.CREL.KG   -3.114e-04  4.625e-04  -0.673   0.5015
## climateData_df$EG.ELC.HYRO.ZS   -1.717e-01  6.935e-02  -2.476   0.0142 *
## climateData_df$EG.ELC.RNEW.ZS    1.037e-01  6.090e-02   1.703   0.0901 .
## climateData_df$EG.ELC.RNWX.ZS    1.717e-01  6.935e-02   2.476   0.0142 *
## climateData_df$EG.FEC.RNEW.ZS    1.287e-02  1.837e-02   0.701   0.4844
## climateData_df$EN.ATM.CO2E.GF.KT 3.545e-05  9.617e-05   0.369   0.7128
## climateData_df$EN.ATM.CO2E.KT   -5.487e-05  7.344e-05  -0.747   0.4560
## climateData_df$EN.ATM.CO2E.LF.KT 5.331e-05  6.239e-05   0.854   0.3940
## climateData_df$EN.ATM.CO2E.SF.KT 6.276e-05  9.186e-05   0.683   0.4953
## climateData_df$EN.ATM.METH.KT.CE -9.691e-06  4.705e-05  -0.206   0.8370
## climateData_df$EN.ATM.NOXE.KT.CE 1.424e-04  1.390e-04   1.025   0.3067
## climateData_df$EN.CLC.MDAT.ZS   -8.554e-05  7.615e-05  -1.123   0.2627
## climateData_df$NV.AGR.TOTL.ZS   -2.741e-07  5.956e-07  -0.460   0.6459
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.77 on 190 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.08566,    Adjusted R-squared:  0.01828
## F-statistic: 1.271 on 14 and 190 DF,  p-value: 0.2284
```

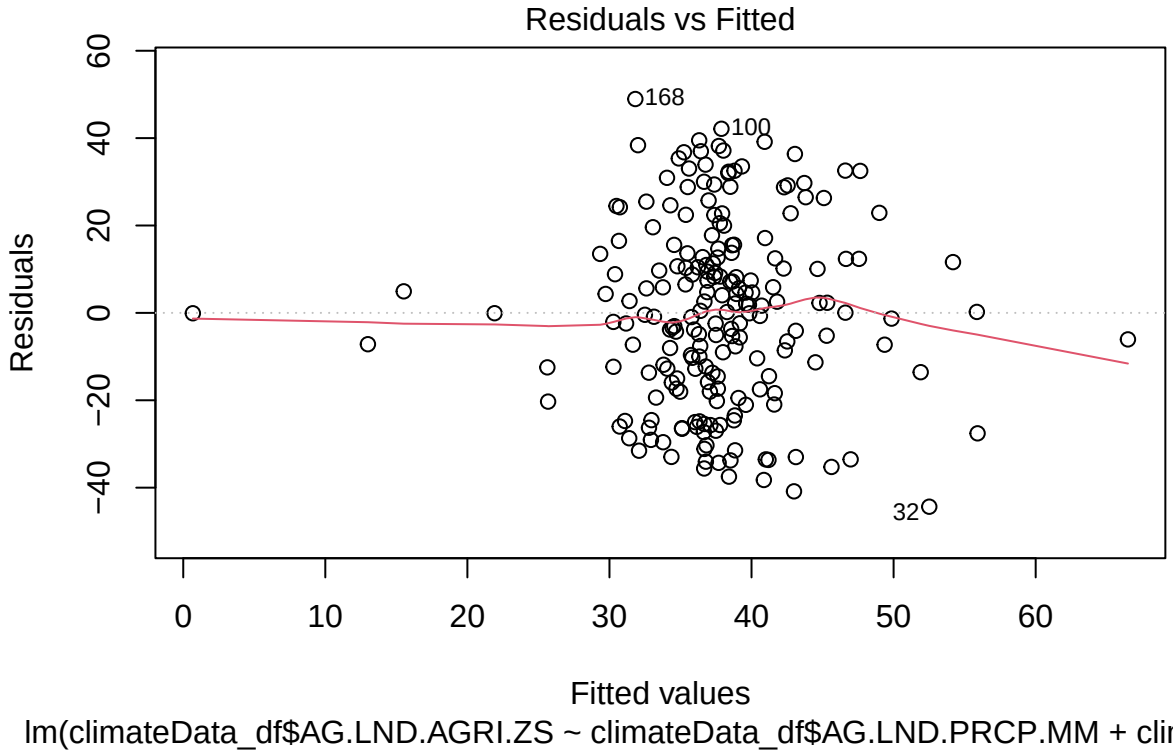
Interpretation and Discussion

The F-statistic is 2.583, with a p-value of 0.00304, indicating that the model as a whole is statistically significant. But the R-squared value relatively low which is 0.1548, which means model perform relatively low with the given predictors. However, Average precipitation in depth (mm per year),Electricity production from hydroelectric sources (% of total), Renewable energy consumption (% of total final energy consumption) and Agriculture, forestry, and fishing, value added (% of GDP) have significant impact on agricultural land. This means annual precipitation of a country, renewable energy consumption and hydroelectric generation of

a country has a significant affect for the growth of agricultural land.This imply that increased precipitation is linked to reduced agricultural land use, areas relying more on hydropower have less agricultural land use and that a greater reliance on renewable energy sources is associated with more agricultural land use.

Residuals Plot for the model A diagnostic plot of the residuals is:

```
plot(climate.lm, which = 1)
```



The residual plot shows a few outliers, more noticeably observations 32, 100 and 168, which may impose some distortion but are not critical because the linear relationship is very clear/strong.

Apply Ridge Regression by calling function glmnet().This function requires that we provide the response variable Y separated from a data matrix X with the predictors. To build this matrix X more easily, just using the linear regression formula and the dataset, we can call function model.matrix():

```
library (glmnet)
```

```
## Warning: package 'glmnet' was built under R version 4.1.3
```

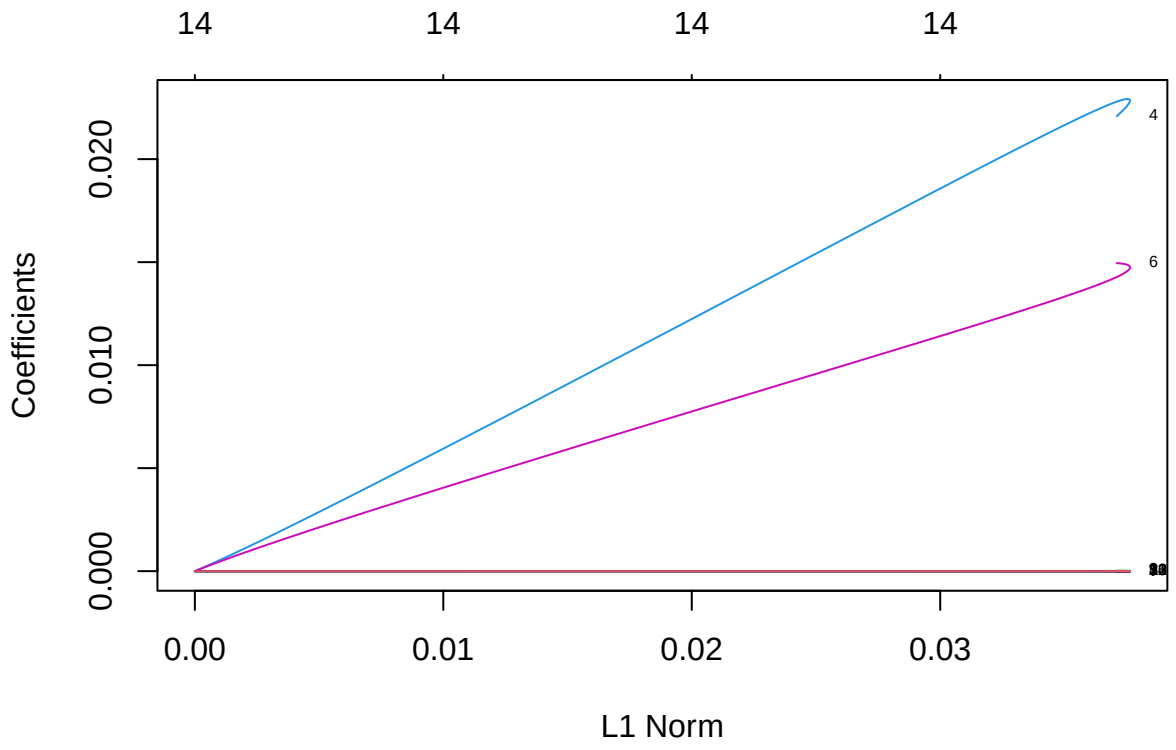
```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
## Loaded glmnet 4.1-7
```

```
predictors <- model.matrix(climateData_df$AG.LND.AGRI.ZS ~ climateData_df$AG.LND.PRCP.MM+climateData_df
                           climateData_df$EG.ELC.RNEW.ZS+climateData_df$EG.ELC.RNWX.ZS+climateData_df
                           climateData_df$EN.ATM.CO2E.KT+climateData_df$EN.ATM.CO2E.LF.KT+climateData
                           climateData_df$EN.ATM.NOXE.KT.CE+climateData_df$EN.CLC.MDAT.ZS+climateData
                  data = climateData_df)
#contain Na values
rows_to_omit <- c(1, 79, 157)

response <- climateData_df$AG.LND.AGRI.ZS
response <- response[-rows_to_omit]
predictors <- predictors[,-1] # Remove the first column

#predictors

# apply ridge with alpha 0
ridge = glmnet(predictors,response,alpha=0)
plot(ridge,label=TRUE)
```



In this plot as lambda increases, the coefficients are shrunk towards zero and, accordingly, their L1 norm decreases. So, the way we read this plot is from right to left: L1 norm decreasing, lambda increasing, coefficients shrinking. The top bar indicates the number of non-zero coefficients (excluded the intercept).

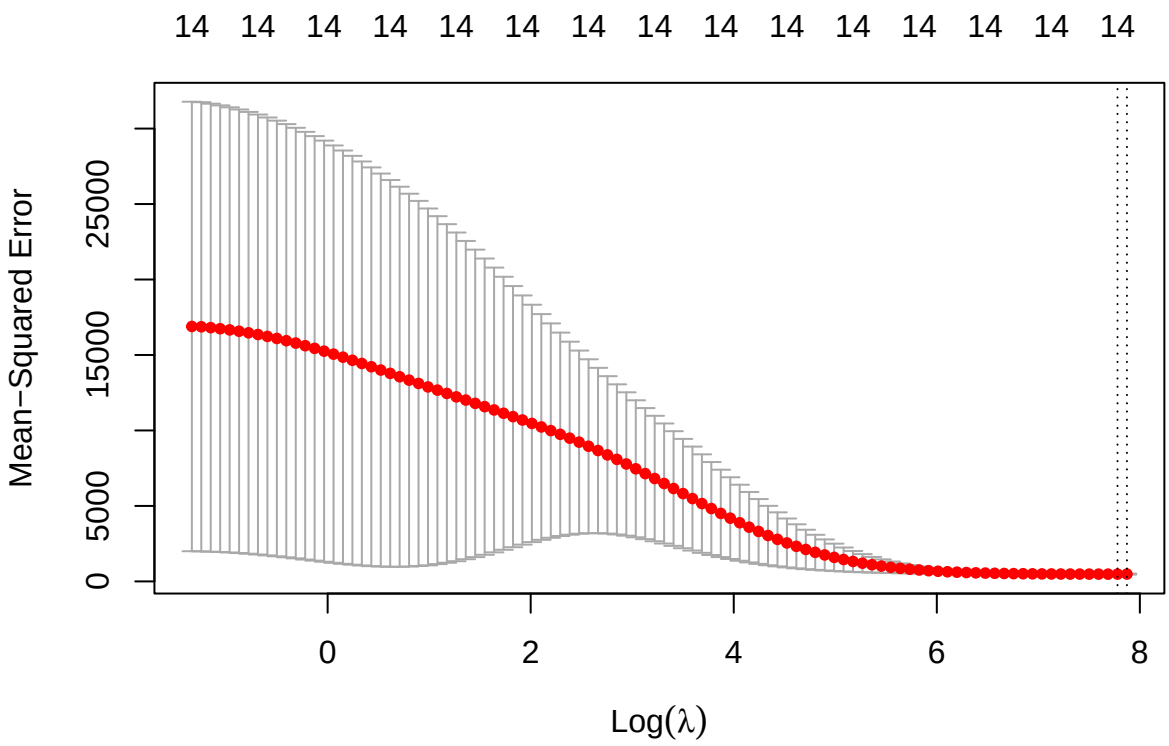
We need to choose one model. We can archive this using cross-validation (CV). We can run CV for Ridge Regression by calling function cv.glmnet().

```
cv_ridge = cv.glmnet(predictors,response,alpha=0)
cv_ridge
```

```
##
## Call:  cv.glmnet(x = predictors, y = response, alpha = 0)
##
## Measure: Mean-Squared Error
##
##     Lambda Index Measure    SE Nonzero
## min   2392     2   479.7 30.98      14
## 1se   2625     1   484.4 33.22      14
```

We can visualize it as,

```
plot(cv_ridge)
```



Above plot, the red dots are the test Mean Squared Error (MSE) for the corresponding model — i.e., the model with the given lambda value in the x-axis — as estimated by CV. The model corresponding to the minimum MSE is indicated by the leftmost dotted vertical line. The rightmost dotted vertical line indicates the model with the largest value of lambda for which the MSE is not above one standard error from the minimum.

Check the values of lambda corresponding to these two models (i.e., the ones indicated by the vertical dotted lines) as follows:

```
cv_ridge$lambda.min
```

```
## [1] 2392.023
```

```
cv_ridge$lambda.1se
```

```
## [1] 2625.243
```

Corresponding coefficients for the each of the above model:

```
#The coefficients for each model can be checked as follows. S is for lambda
coef(cv_ridge, s = cv_ridge$lambda.min)
```

```
## 15 x 1 sparse Matrix of class "dgCMatrix"
##                                          s1
## (Intercept)                        3.764138e+01
## climateData_df$AG.LND.PRCP.MM     -1.525621e-08
## climateData_df$AG.YLD.CREL.KG     -1.527585e-08
## climateData_df$EG.ELC.HYRO.ZS     -1.539810e-08
## climateData_df$EG.ELC.RNEW.ZS      1.672120e-04
## climateData_df$EG.ELC.RNWX.ZS     -1.539775e-08
## climateData_df$EG.FEC.RNEW.ZS      1.532862e-04
## climateData_df$EN.ATM.CO2E.GF.KT -1.088708e-09
## climateData_df$EN.ATM.CO2E.KT     -7.466286e-09
## climateData_df$EN.ATM.CO2E.LF.KT  6.354259e-08
## climateData_df$EN.ATM.CO2E.SF.KT  3.094170e-08
## climateData_df$EN.ATM.METH.KT.CE -1.429574e-08
## climateData_df$EN.ATM.NOXE.KT.CE -1.476498e-08
## climateData_df$EN.CLC.MDAT.ZS     -1.546743e-08
## climateData_df$NV.AGR.TOTL.ZS     -3.874608e-09
```

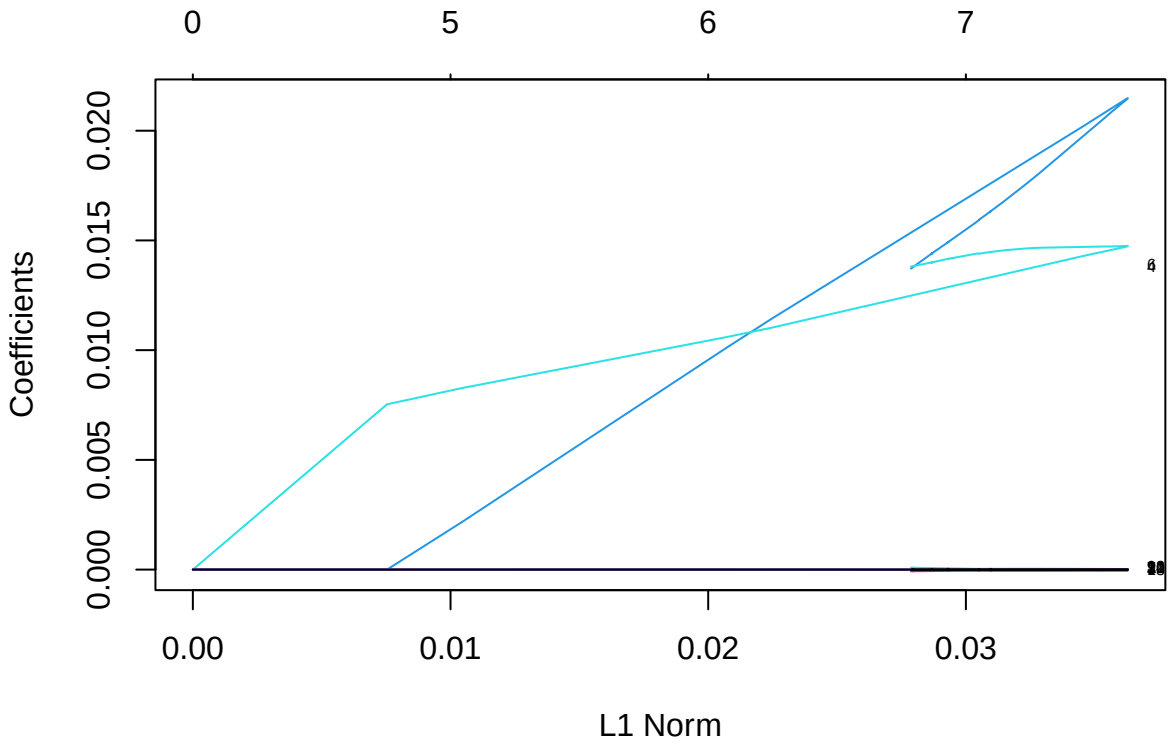
```
coef(cv_ridge, s = cv_ridge$lambda.1se)
```

```
## 15 x 1 sparse Matrix of class "dgCMatrix"
##                                          s1
## (Intercept)                        3.764266e+01
## climateData_df$AG.LND.PRCP.MM     -1.796151e-42
## climateData_df$AG.YLD.CREL.KG     -1.798341e-42
## climateData_df$EG.ELC.HYRO.ZS     -1.811816e-42
## climateData_df$EG.ELC.RNEW.ZS      1.804150e-38
## climateData_df$EG.ELC.RNWX.ZS     -1.811783e-42
## climateData_df$EG.FEC.RNEW.ZS      1.709051e-38
## climateData_df$EN.ATM.CO2E.GF.KT -2.157522e-44
## climateData_df$EN.ATM.CO2E.KT     -9.052233e-43
## climateData_df$EN.ATM.CO2E.LF.KT  7.097116e-42
## climateData_df$EN.ATM.CO2E.SF.KT  3.444197e-42
## climateData_df$EN.ATM.METH.KT.CE -1.688851e-42
## climateData_df$EN.ATM.NOXE.KT.CE -1.741635e-42
## climateData_df$EN.CLC.MDAT.ZS     -1.819466e-42
## climateData_df$NV.AGR.TOTL.ZS     -4.306193e-43
```

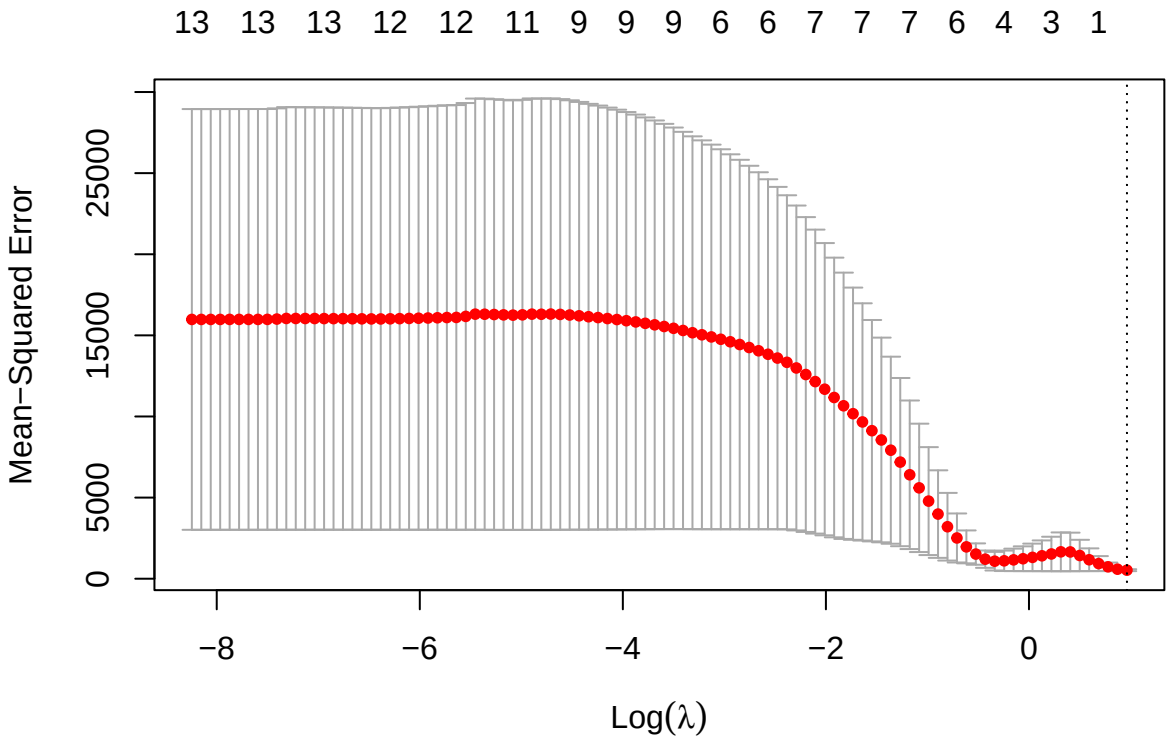
In this case we choose the model with lambda.1se which is the model with the largest value of lambda for which the MSE is not above one standard error from the minimum.The reason is within this analysis my aim is to interpret the results rather than predicting. So I have chosen the model with lambda.1se for my analysis.

Next, similarly we have perform lasso regression with alpha =1.

```
#lasso
lasso <- glmnet(predictors,response,alpha=1)
plot(lasso,label=TRUE)
```




```
cv_lasso = cv.glmnet(predictors,response,alpha=1)
plot(cv_lasso)
```




```
cv_lasso$lambda.min
```

```
## [1] 2.625243
```

```
cv_lasso$lambda.1se
```

```
## [1] 2.625243
```

Similar to ridge with also lasso I have chosen lamda.1se,

```
coef(cv_lasso)
```

```
## 15 x 1 sparse Matrix of class "dgCMatrix"
##                                       s1
## (Intercept)                     37.64266
## climateData_df$AG.LND.PRCP.MM     .
## climateData_df$AG.YLD.CREL.KG     .
## climateData_df$EG.ELC.HYRO.ZS     .
## climateData_df$EG.ELC.RNEW.ZS     .
## climateData_df$EG.ELC.RNWX.ZS     .
## climateData_df$EG.FEC.RNEW.ZS     .
## climateData_df$EN.ATM.CO2E.GF.KT  .
## climateData_df$EN.ATM.CO2E.KT     .
## climateData_df$EN.ATM.CO2E.LF.KT  .
## climateData_df$EN.ATM.CO2E.SF.KT  .
## climateData_df$EN.ATM.METH.KT.CE  .
## climateData_df$EN.ATM.NOXE.KT.CE  .
## climateData_df$EN.CLC.MDAT.ZS     .
```

```
## climateData_df$NV.AGR.TOTL.ZS     .
```

Compare the best cross-validated MSE achieved for Ridge Regression and the Lasso.

```
min(cv_ridge$cvm)
```

```
## [1] 479.7282
```

```
min(cv_lasso$cvm)
```

```
## [1] 525.6017
```

Comparing the best cross-validated MSE achieved for Ridge Regression and the Lasso, we can see that the Lasso achieves a smaller estimated error.

Fitting the linear model for response variables: "AG.LND.ARBL.ZS": Arable land (% of land area)

```
climate_arable.lm <-lm(formula = climateData_df$AG.LND.ARBL.ZS ~ climateData_df$AG.LND.PRCP.MM+climateDa
                  climateData_df$EG.ELC.RNEW.ZS+climateData_df$EG.ELC.RNWX.ZS+climateData_df$EG.FEC.RNEW
                  climateData_df$EN.ATM.CO2E.KT+climateData_df$EN.ATM.CO2E.LF.KT+climateData_df$EN.ATM.C
                  climateData_df$EN.ATM.NOXE.KT.CE+climateData_df$EN.CLC.MDAT.ZS+climateData_df$NV.AGR.T

summary(climate_arable.lm)
```

```
##
## Call:
## lm(formula = climateData_df$AG.LND.ARBL.ZS ~ climateData_df$AG.LND.PRCP.MM +
##     climateData_df$AG.YLD.CREL.KG + climateData_df$EG.ELC.HYRO.ZS +
##     climateData_df$EG.ELC.RNEW.ZS + climateData_df$EG.ELC.RNWX.ZS +
##     climateData_df$EG.FEC.RNEW.ZS + climateData_df$EN.ATM.CO2E.GF.KT +
##     climateData_df$EN.ATM.CO2E.KT + climateData_df$EN.ATM.CO2E.LF.KT +
##     climateData_df$EN.ATM.CO2E.SF.KT + climateData_df$EN.ATM.METH.KT.CE +
##     climateData_df$EN.ATM.NOXE.KT.CE + climateData_df$EN.CLC.MDAT.ZS +
##     climateData_df$NV.AGR.TOTL.ZS, data = climateData_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.928  -9.569  -2.471   6.212  44.094
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      1.304e+01  1.539e+00   8.471 6.57e-15 ***
## climateData_df$AG.LND.PRCP.MM   -1.991e-04  2.819e-04  -0.706 0.480929
## climateData_df$AG.YLD.CREL.KG    1.962e-04  2.785e-04   0.704 0.481987
## climateData_df$EG.ELC.HYRO.ZS   -1.462e-01  4.177e-02  -3.501 0.000577 ***
## climateData_df$EG.ELC.RNEW.ZS    4.858e-02  3.668e-02   1.324 0.186971
## climateData_df$EG.ELC.RNWX.ZS    1.462e-01  4.177e-02   3.501 0.000577 ***
## climateData_df$EG.FEC.RNEW.ZS    2.183e-02  1.106e-02   1.973 0.049899 *
## climateData_df$EN.ATM.CO2E.GF.KT  1.365e-05  5.792e-05   0.236 0.813872
## climateData_df$EN.ATM.CO2E.KT    -3.984e-05  4.423e-05  -0.901 0.368911
## climateData_df$EN.ATM.CO2E.LF.KT  4.650e-05  3.757e-05   1.238 0.217407
## climateData_df$EN.ATM.CO2E.SF.KT  4.200e-05  5.532e-05   0.759 0.448716
## climateData_df$EN.ATM.METH.KT.CE  1.061e-05  2.833e-05   0.375 0.708352
## climateData_df$EN.ATM.NOXE.KT.CE  7.646e-05  8.370e-05   0.914 0.362099
## climateData_df$EN.CLC.MDAT.ZS    -3.082e-05  4.586e-05  -0.672 0.502369
## climateData_df$NV.AGR.TOTL.ZS     3.126e-07  3.587e-07   0.872 0.384527
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 13.11 on 190 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.1195, Adjusted R-squared:  0.05458
## F-statistic: 1.841 on 14 and 190 DF,  p-value: 0.03534
```

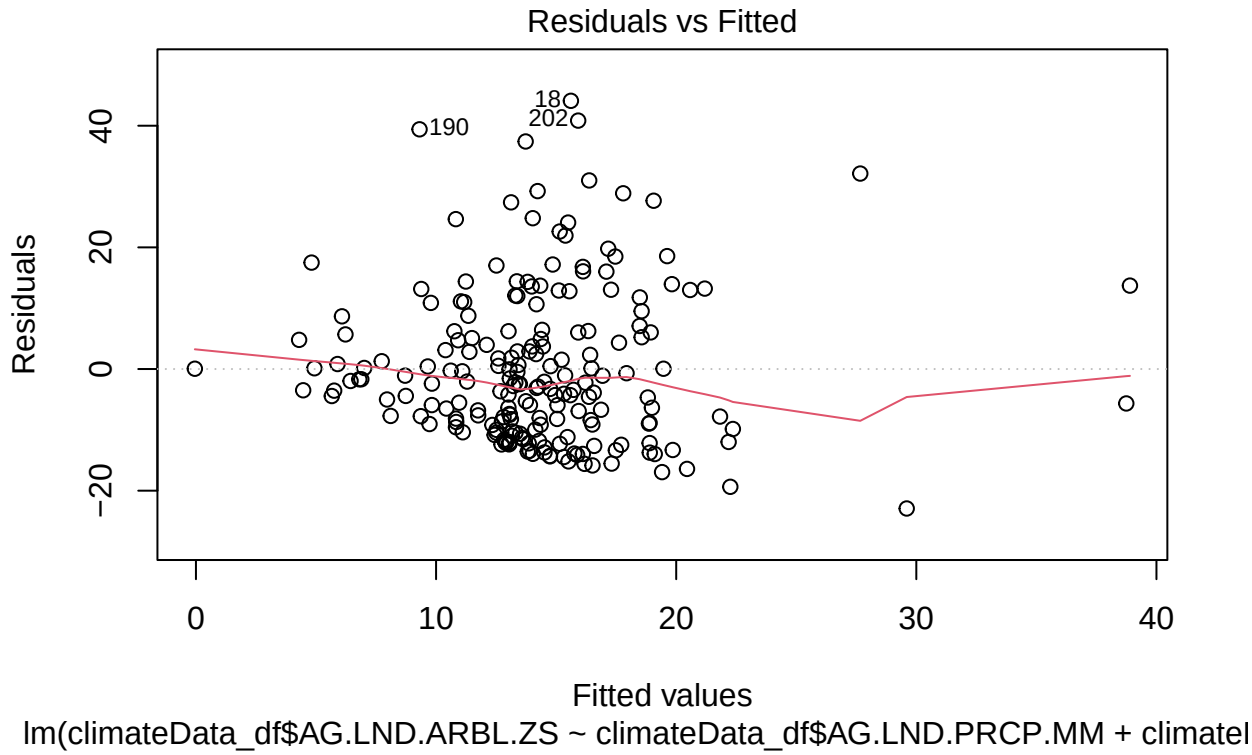
Interpretation and Discussion

The F-statistic is 2.228, with a p-value of 0.01079, indicating that the model as a whole is statistically significant. But the R-squared value relatively low which is 0.1227, which means model perform relatively low with the given predictors. However, Electricity production from hydroelectric sources (% of total), Renewable electricity output have significant impact on arable land. This means that factors like electricity generation from hydropower, renewable energy, and renewable fertilizer consumption are associated with arable land.Negative correlation indicates that areas relying more on hydropower have less arable land.A positive coefficient (1.822e-01) suggests that a higher percentage of electricity generation from renewable sources is associated with an increase in "climateData_df$AG.LND.ARBL.ZS," indicating that a greater reliance on renewable energy sources is associated with more arable land.A positive coefficient (1.554e-01) indicates that a higher percentage of fertilizer consumption from renewable sources is associated with more arable land. This may suggest that sustainable agricultural practices are associated with greater arable land.

Residuals Plot for the model

A diagnostic plot of the residuals is:

```
plot(climate_arable.lm, which = 1)
```

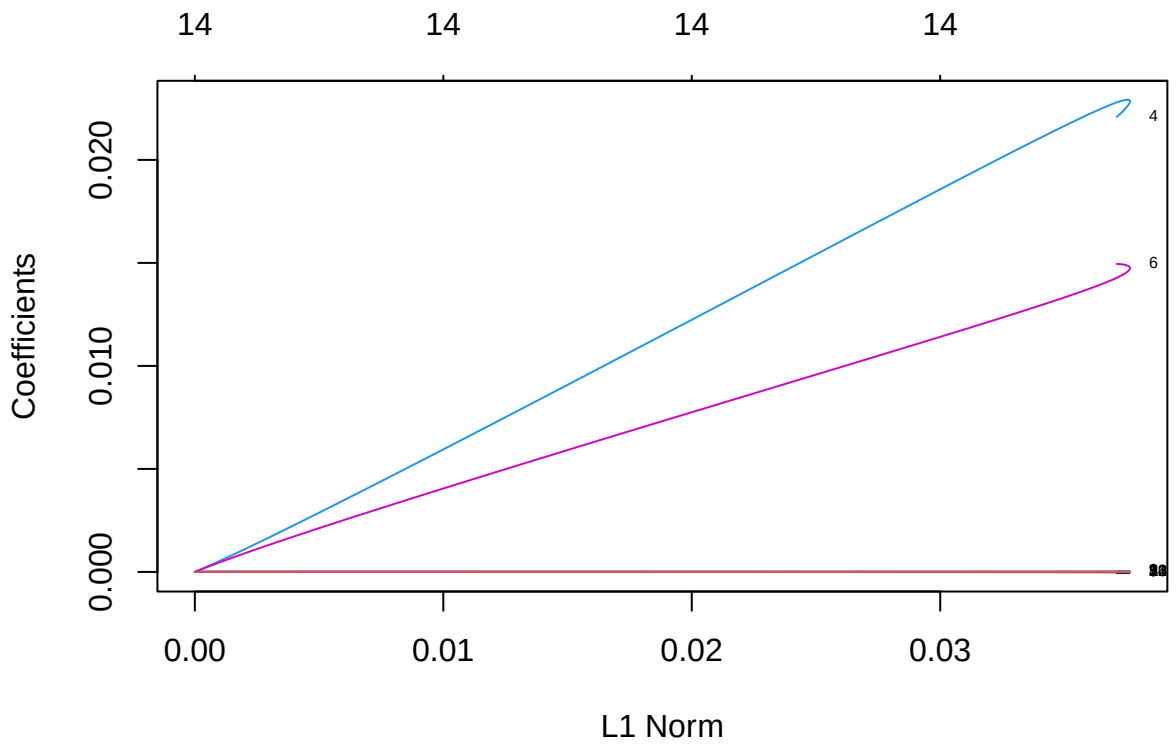


The residual plot shows a few outliers, more noticeably observations 18, 190 and 202, which may impose some distortion but are not critical because the linear relationship is very clear/strong.

Apply Ridge Regression by calling function glmnet().This function requires that we provide the response

variable Y separated from a data matrix X with the predictors. To build this matrix X more easily, just using the linear regression formula and the dataset, we can call function model.matrix():

```
predictors <- model.matrix(climateData_df$AG.LND.ARBL.ZS ~ climateData_df$AG.LND.PRCP.MM+climateData_df$
                             climateData_df$EG.ELC.RNEW.ZS+climateData_df$EG.ELC.RNWX.ZS+climateData_df$
                             climateData_df$EN.ATM.CO2E.KT+climateData_df$EN.ATM.CO2E.LF.KT+climateData_
                             climateData_df$EN.ATM.NOXE.KT.CE+climateData_df$EN.CLC.MDAT.ZS+climateData_
                           data = climateData_df)

predictors <- predictors[,-1] # Remove the first column

#contain Na values
rows_to_omit <- c(1, 79, 157)

response <- climateData_df$AG.LND.AGRI.ZS
response <- response[-rows_to_omit]

# apply ridge with alpha 0
ridge = glmnet(predictors,response,alpha=0)
plot(ridge,label=TRUE)
```



We need to choose one model. We can archive this using cross-validation (CV). We can run CV for Ridge Regression by calling function cv.glmnet().

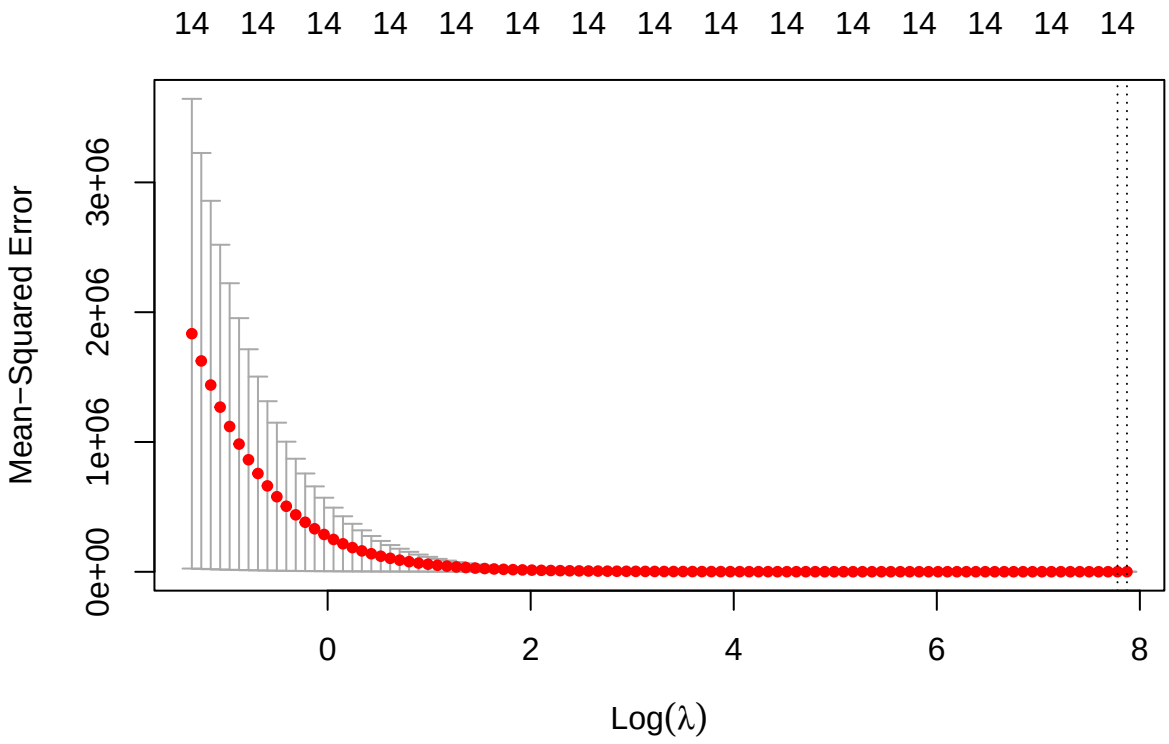
```
cv_ridge = cv.glmnet(predictors,response,alpha=0)
cv_ridge
```

```
##
```

```
## Call:  cv.glmnet(x = predictors, y = response, alpha = 0)
##
## Measure: Mean-Squared Error
##
##     Lambda Index Measure    SE Nonzero
## min   2392     2   489.1 48.70      14
## 1se   2625     1   489.1 48.71      14
```

We can visualize it as,

```
plot(cv_ridge)
```



Above plot, the red dots are the test Mean Squared Error (MSE) for the corresponding model — i.e., the model with the given lambda value in the x-axis — as estimated by CV. The model corresponding to the minimum MSE is indicated by the leftmost dotted vertical line. The rightmost dotted vertical line indicates the model with the largest value of lambda for which the MSE is not above one standard error from the minimum.

Check the values of lambda corresponding to these two models (i.e., the ones indicated by the vertical dotted lines) as follows:

```
cv_ridge$lambda.min
```

```
## [1] 2392.023
```

```
cv_ridge$lambda.1se
```

```
## [1] 2625.243
```

Corresponding coefficients for the each of the above model:

```r
#The coefficients for each model can be checked as follows. S is for lambda
coef(cv_ridge, s = cv_ridge$lambda.min)
```

```
## 15 x 1 sparse Matrix of class "dgCMatrix"
##                                          s1
## (Intercept)                     3.764138e+01
## climateData_df$AG.LND.PRCP.MM  -1.525621e-08
## climateData_df$AG.YLD.CREL.KG  -1.527585e-08
## climateData_df$EG.ELC.HYRO.ZS  -1.539810e-08
## climateData_df$EG.ELC.RNEW.ZS   1.672120e-04
## climateData_df$EG.ELC.RNWX.ZS  -1.539775e-08
## climateData_df$EG.FEC.RNEW.ZS   1.532862e-04
## climateData_df$EN.ATM.CO2E.GF.KT -1.088708e-09
## climateData_df$EN.ATM.CO2E.KT  -7.466286e-09
## climateData_df$EN.ATM.CO2E.LF.KT  6.354259e-08
## climateData_df$EN.ATM.CO2E.SF.KT  3.094170e-08
## climateData_df$EN.ATM.METH.KT.CE -1.429574e-08
## climateData_df$EN.ATM.NOXE.KT.CE -1.476498e-08
## climateData_df$EN.CLC.MDAT.ZS  -1.546743e-08
## climateData_df$NV.AGR.TOTL.ZS  -3.874608e-09
```

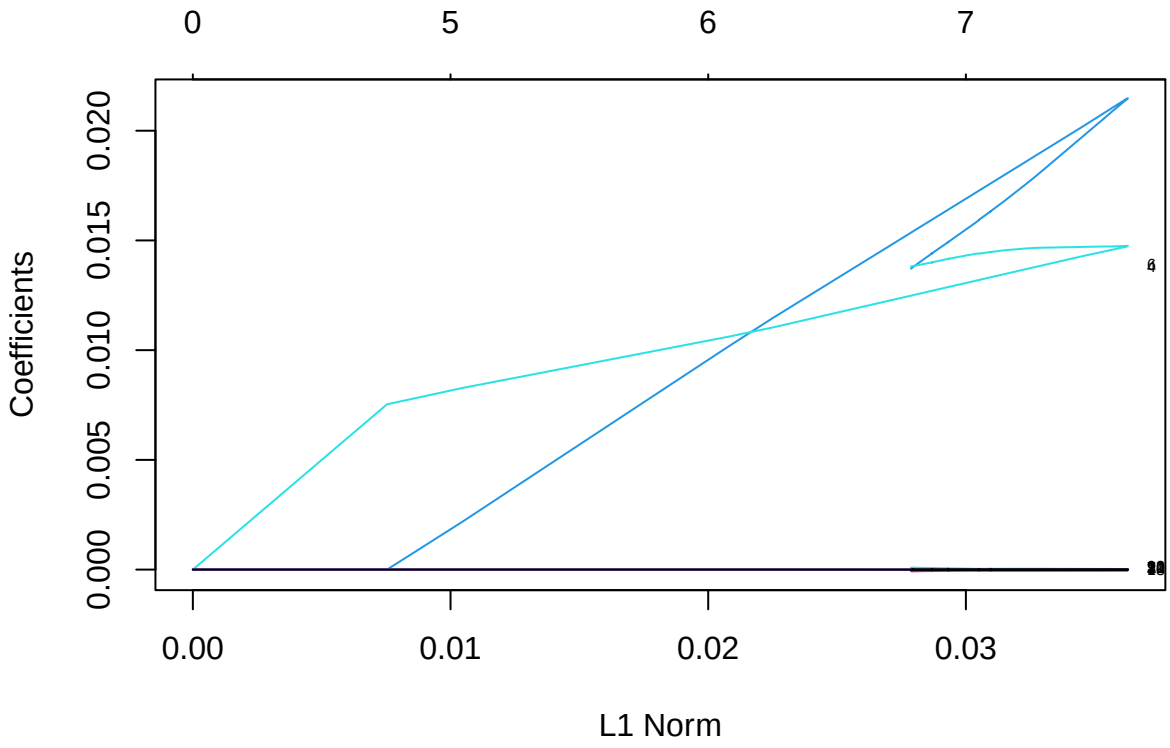
```r
coef(cv_ridge, s = cv_ridge$lambda.1se)
```

```
## 15 x 1 sparse Matrix of class "dgCMatrix"
##                                          s1
## (Intercept)                     3.764266e+01
## climateData_df$AG.LND.PRCP.MM  -1.796151e-42
## climateData_df$AG.YLD.CREL.KG  -1.798341e-42
## climateData_df$EG.ELC.HYRO.ZS  -1.811816e-42
## climateData_df$EG.ELC.RNEW.ZS   1.804150e-38
## climateData_df$EG.ELC.RNWX.ZS  -1.811783e-42
## climateData_df$EG.FEC.RNEW.ZS   1.709051e-38
## climateData_df$EN.ATM.CO2E.GF.KT -2.157522e-44
## climateData_df$EN.ATM.CO2E.KT  -9.052233e-43
## climateData_df$EN.ATM.CO2E.LF.KT  7.097116e-42
## climateData_df$EN.ATM.CO2E.SF.KT  3.444197e-42
## climateData_df$EN.ATM.METH.KT.CE -1.688851e-42
## climateData_df$EN.ATM.NOXE.KT.CE -1.741635e-42
## climateData_df$EN.CLC.MDAT.ZS  -1.819466e-42
## climateData_df$NV.AGR.TOTL.ZS  -4.306193e-43
```

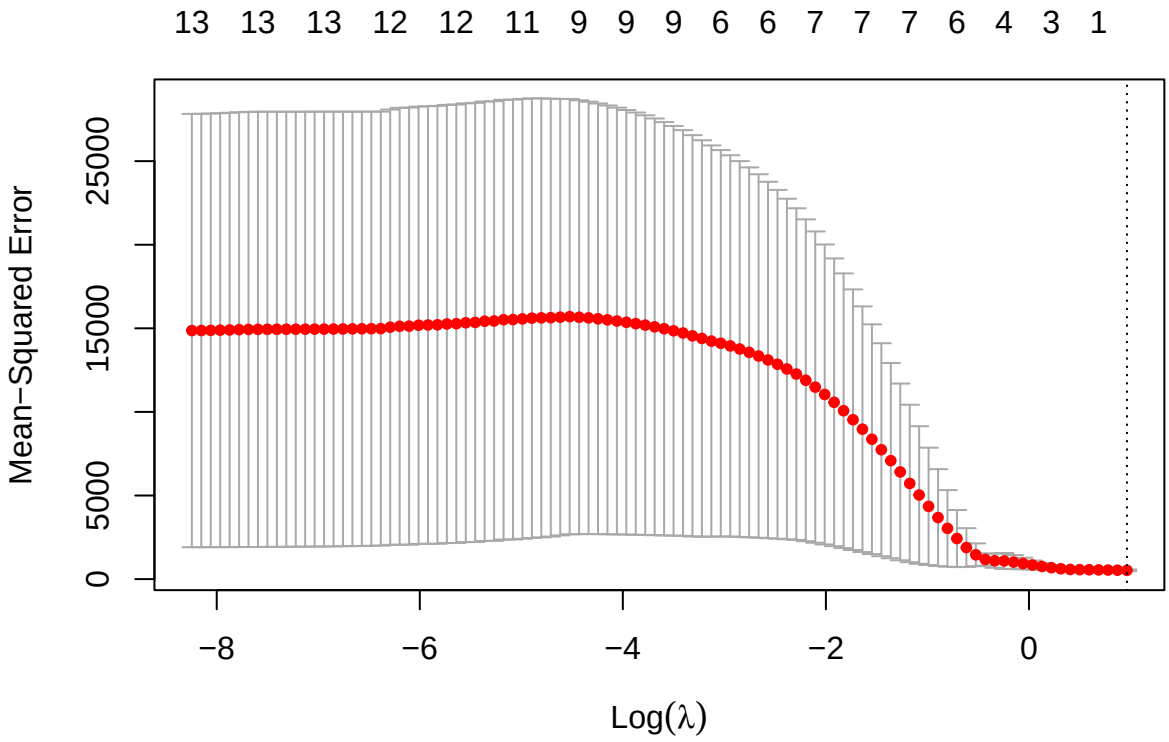
In this case we choose the model with lambda.1se which is the model with the largest value of lambda for which the MSE is not above one standard error from the minimum.The reason is within this analysis my aim is to interpret the results rather than predicting. So I have chosen the model with lambda.1se for my analysis.

Next, similarly we have perform lasso regression with alpha =1.

```r
#lasso
lasso <- glmnet(predictors,response,alpha=1)
plot(lasso,label=TRUE)
```




```
cv_lasso = cv.glmnet(predictors,response,alpha=1)
plot(cv_lasso)
```




```
cv_lasso$lambda.min
```

```
## [1] 2.625243
```

```
cv_lasso$lambda.1se
```

```
## [1] 2.625243
```

Similar to ridge with also lasso I have chosen lamda.1se,

```
coef(cv_lasso)
```

```
## 15 x 1 sparse Matrix of class "dgCMatrix"
##                                       s1
## (Intercept)                     37.64266
## climateData_df$AG.LND.PRCP.MM     .
## climateData_df$AG.YLD.CREL.KG     .
## climateData_df$EG.ELC.HYRO.ZS     .
## climateData_df$EG.ELC.RNEW.ZS     .
## climateData_df$EG.ELC.RNWX.ZS     .
## climateData_df$EG.FEC.RNEW.ZS     .
## climateData_df$EN.ATM.CO2E.GF.KT  .
## climateData_df$EN.ATM.CO2E.KT     .
## climateData_df$EN.ATM.CO2E.LF.KT  .
## climateData_df$EN.ATM.CO2E.SF.KT  .
## climateData_df$EN.ATM.METH.KT.CE  .
## climateData_df$EN.ATM.NOXE.KT.CE  .
## climateData_df$EN.CLC.MDAT.ZS     .
```

```
## climateData_df$NV.AGR.TOTL.ZS     .
```

Compare the best cross-validated MSE achieved for Ridge Regression and the Lasso.

```
min(cv_ridge$cvm)
```

```
## [1] 489.0503
```

```
min(cv_lasso$cvm)
```

```
## [1] 524.7098
```

Comparing the best cross-validated MSE achieved for Ridge Regression and the Lasso, we can see that the Lasso achieves a smaller estimated error.

Fitting the linear model for response variables: "AG.LND.IRIG.AG.ZS: Agricultural irrigated land (% of total agricultural land)

```
climate.lm <-lm(formula = climateData_df$AG.LND.IRIG.AG.ZS ~ climateData_df$AG.LND.PRCP.MM+climateData_
                  climateData_df$EG.ELC.RNEW.ZS+climateData_df$EG.ELC.RNWX.ZS+climateData_df$EG.FEC.RNE
                  climateData_df$EN.ATM.CO2E.KT+climateData_df$EN.ATM.CO2E.LF.KT+climateData_df$EN.ATM.
                  climateData_df$EN.ATM.NOXE.KT.CE+climateData_df$EN.CLC.MDAT.ZS+climateData_df$NV.AGR.
summary(climate.lm)
```

```
##
## Call:
## lm(formula = climateData_df$AG.LND.IRIG.AG.ZS ~ climateData_df$AG.LND.PRCP.MM +
##     climateData_df$AG.YLD.CREL.KG + climateData_df$EG.ELC.HYRO.ZS +
##     climateData_df$EG.ELC.RNEW.ZS + climateData_df$EG.ELC.RNWX.ZS +
##     climateData_df$EG.FEC.RNEW.ZS + climateData_df$EN.ATM.CO2E.GF.KT +
##     climateData_df$EN.ATM.CO2E.KT + climateData_df$EN.ATM.CO2E.LF.KT +
##     climateData_df$EN.ATM.CO2E.SF.KT + climateData_df$EN.ATM.METH.KT.CE +
##     climateData_df$EN.ATM.NOXE.KT.CE + climateData_df$EN.CLC.MDAT.ZS +
##     climateData_df$NV.AGR.TOTL.ZS, data = climateData_df)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2.584e+08 -7.566e+07 -4.537e+07 -1.959e+07  1.165e+10
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       4.477e+07  9.984e+07   0.448    0.654
## climateData_df$AG.LND.PRCP.MM     8.672e+02  1.829e+04   0.047    0.962
## climateData_df$AG.YLD.CREL.KG     7.078e+02  1.807e+04   0.039    0.969
## climateData_df$EG.ELC.HYRO.ZS     2.774e+06  2.710e+06   1.024    0.307
## climateData_df$EG.ELC.RNEW.ZS    -7.882e+05  2.380e+06  -0.331    0.741
## climateData_df$EG.ELC.RNWX.ZS    -2.774e+06  2.710e+06  -1.024    0.307
## climateData_df$EG.FEC.RNEW.ZS     1.655e+05  7.177e+05   0.231    0.818
## climateData_df$EN.ATM.CO2E.GF.KT -7.214e+02  3.757e+03  -0.192    0.848
## climateData_df$EN.ATM.CO2E.KT     2.409e+02  2.869e+03   0.084    0.933
## climateData_df$EN.ATM.CO2E.LF.KT  1.504e+02  2.438e+03   0.062    0.951
## climateData_df$EN.ATM.CO2E.SF.KT -2.847e+02  3.589e+03  -0.079    0.937
## climateData_df$EN.ATM.METH.KT.CE  4.286e+02  1.838e+03   0.233    0.816
## climateData_df$EN.ATM.NOXE.KT.CE -1.894e+03  5.430e+03  -0.349    0.728
## climateData_df$EN.CLC.MDAT.ZS    -1.257e+02  2.975e+03  -0.042    0.966
## climateData_df$NV.AGR.TOTL.ZS    -4.410e+00  2.327e+01  -0.190    0.850
##
## Residual standard error: 850600000 on 190 degrees of freedom
```

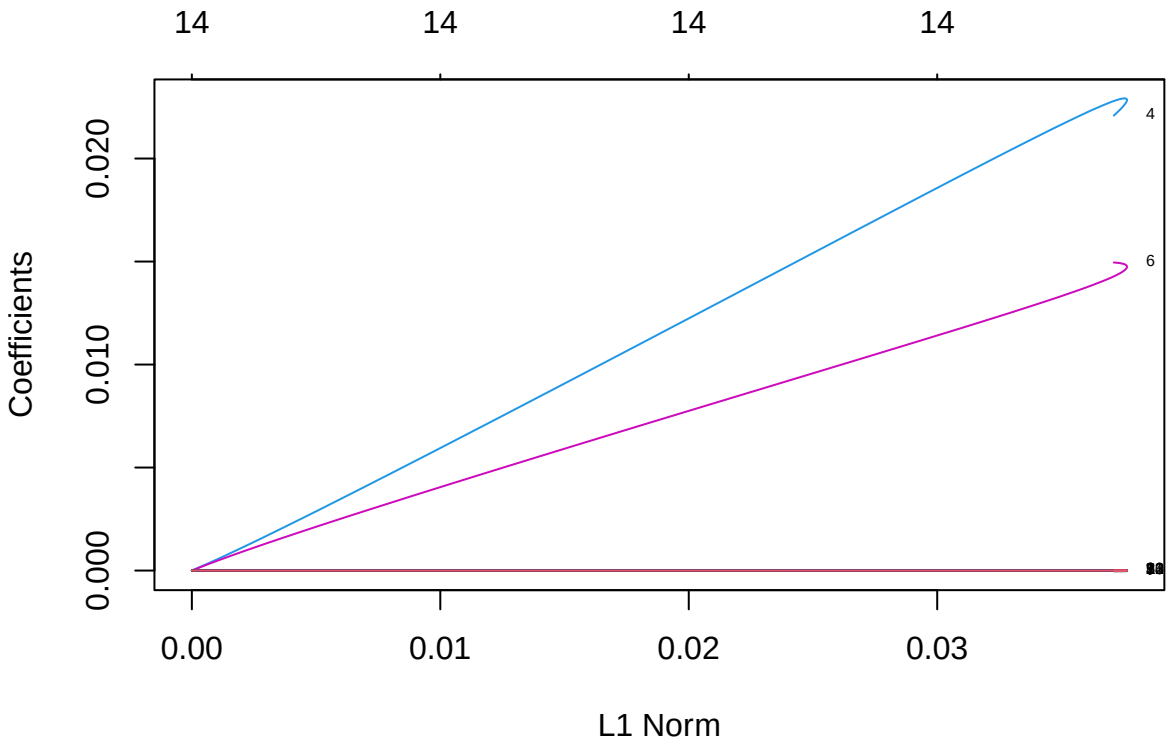
```
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.00769,    Adjusted R-squared:  -0.06543
## F-statistic: 0.1052 on 14 and 190 DF,  p-value: 1
```

Interpretation and Discussion

The the F-statistic is 1.467, with a p-value of 0.1356, indicating that the model is not statistically significant with siginificance level 0.05.The regression model suggests that cereal yield and metadata completeness for environmental data are associated with the percentage of irrigated agricultural land for some extent.

Apply Ridge Regression by calling function glmnet().This function requires that we provide the response variable Y separated from a data matrix X with the predictors. To build this matrix X more easily, just using the linear regression formula and the dataset, we can call function model.matrix():

```
predictors <- model.matrix(climateData_df$AG.LND.IRIG.AG.ZS ~ climateData_df$AG.LND.PRCP.MM+climateData
                           climateData_df$EG.ELC.RNEW.ZS+climateData_df$EG.ELC.RNWX.ZS+climateData_df
                           climateData_df$EN.ATM.CO2E.KT+climateData_df$EN.ATM.CO2E.LF.KT+climateData
                           climateData_df$EN.ATM.NOXE.KT.CE+climateData_df$EN.CLC.MDAT.ZS+climateData
                         data = climateData_df)

predictors <- predictors[,-1] # Remove the first column

#contain Na values
rows_to_omit <- c(1, 79, 157)

response <- climateData_df$AG.LND.AGRI.ZS
response <- response[-rows_to_omit]

# apply ridge with alpha 0
ridge = glmnet(predictors,response,alpha=0)
plot(ridge,label=TRUE)
```



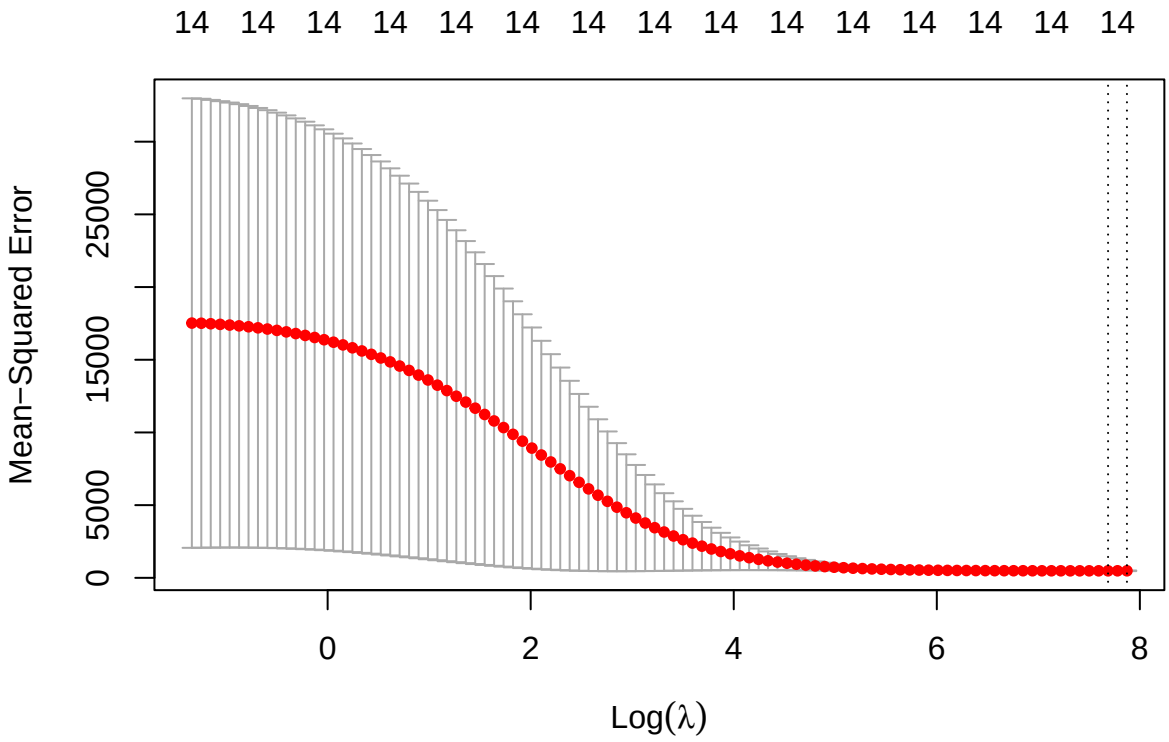
We need to choose one model. We can archive this using cross-validation (CV). We can run CV for Ridge Regression by calling function cv.glmnet().

```
cv_ridge = cv.glmnet(predictors,response,alpha=0)
cv_ridge
```

```
##
## Call:  cv.glmnet(x = predictors, y = response, alpha = 0)
##
## Measure: Mean-Squared Error
##
##     Lambda Index Measure    SE Nonzero
## min   2180     3   482.3 23.37      14
## 1se   2625     1   484.4 21.21      14
```

We can visualize it as,

```
plot(cv_ridge)
```



Above plot, the red dots are the test Mean Squared Error (MSE) for the corresponding model — i.e., the model with the given lambda value in the x-axis — as estimated by CV. The model corresponding to the minimum MSE is indicated by the leftmost dotted vertical line. The rightmost dotted vertical line indicates the model with the largest value of lambda for which the MSE is not above one standard error from the minimum.

Check the values of lambda corresponding to these two models (i.e., the ones indicated by the vertical dotted lines) as follows:

```
cv_ridge$lambda.min
```

```
## [1] 2179.523
```

```
cv_ridge$lambda.1se
```

```
## [1] 2625.243
```

Corresponding coefficients for the each of the above model:

```
#The coefficients for each model can be checked as follows. S is for lambda
coef(cv_ridge, s = cv_ridge$lambda.min)
```

```
## 15 x 1 sparse Matrix of class "dgCMatrix"
##                                         s1
## (Intercept)                   3.764116e+01
## climateData_df$AG.LND.PRCP.MM -1.663805e-08
## climateData_df$AG.YLD.CREL.KG -1.666049e-08
## climateData_df$EG.ELC.HYRO.ZS -1.679537e-08
## climateData_df$EG.ELC.RNEW.ZS  1.838594e-04
```

```
## climateData_df$EG.ELC.RNWX.ZS    -1.679550e-08
## climateData_df$EG.FEC.RNEW.ZS     1.680490e-04
## climateData_df$EN.ATM.CO2E.GF.KT -1.289451e-09
## climateData_df$EN.ATM.CO2E.KT    -8.120496e-09
## climateData_df$EN.ATM.CO2E.LF.KT  6.964826e-08
## climateData_df$EN.ATM.CO2E.SF.KT  3.392568e-08
## climateData_df$EN.ATM.METH.KT.CE -1.558791e-08
## climateData_df$EN.ATM.NOXE.KT.CE -1.610200e-08
## climateData_df$EN.CLC.MDAT.ZS    -1.687251e-08
## climateData_df$NV.AGR.TOTL.ZS    -4.249004e-09
```

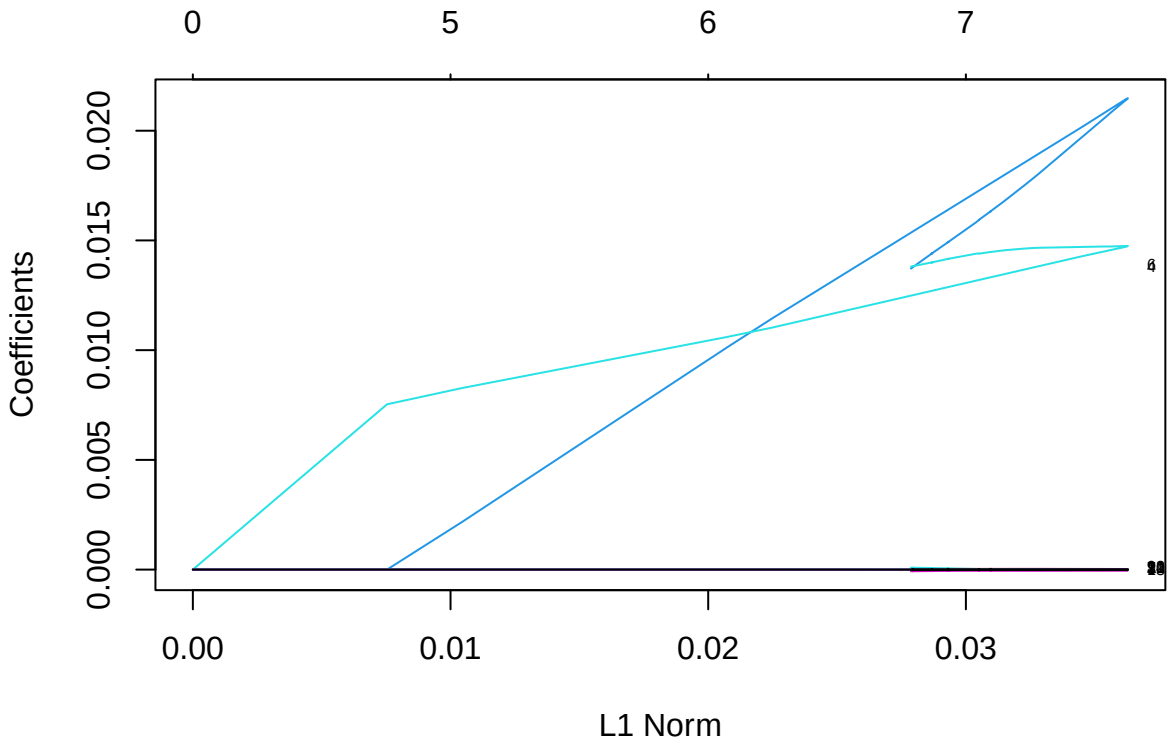
```
coef(cv_ridge, s = cv_ridge$lambda.1se)
```

```
## 15 x 1 sparse Matrix of class "dgCMatrix"
##                                          s1
## (Intercept)                    3.764266e+01
## climateData_df$AG.LND.PRCP.MM -1.796151e-42
## climateData_df$AG.YLD.CREL.KG -1.798341e-42
## climateData_df$EG.ELC.HYRO.ZS -1.811816e-42
## climateData_df$EG.ELC.RNEW.ZS  1.804150e-38
## climateData_df$EG.ELC.RNWX.ZS -1.811783e-42
## climateData_df$EG.FEC.RNEW.ZS  1.709051e-38
## climateData_df$EN.ATM.CO2E.GF.KT -2.157522e-44
## climateData_df$EN.ATM.CO2E.KT    -9.052233e-43
## climateData_df$EN.ATM.CO2E.LF.KT  7.097116e-42
## climateData_df$EN.ATM.CO2E.SF.KT  3.444197e-42
## climateData_df$EN.ATM.METH.KT.CE -1.688851e-42
## climateData_df$EN.ATM.NOXE.KT.CE -1.741635e-42
## climateData_df$EN.CLC.MDAT.ZS    -1.819466e-42
## climateData_df$NV.AGR.TOTL.ZS    -4.306193e-43
```

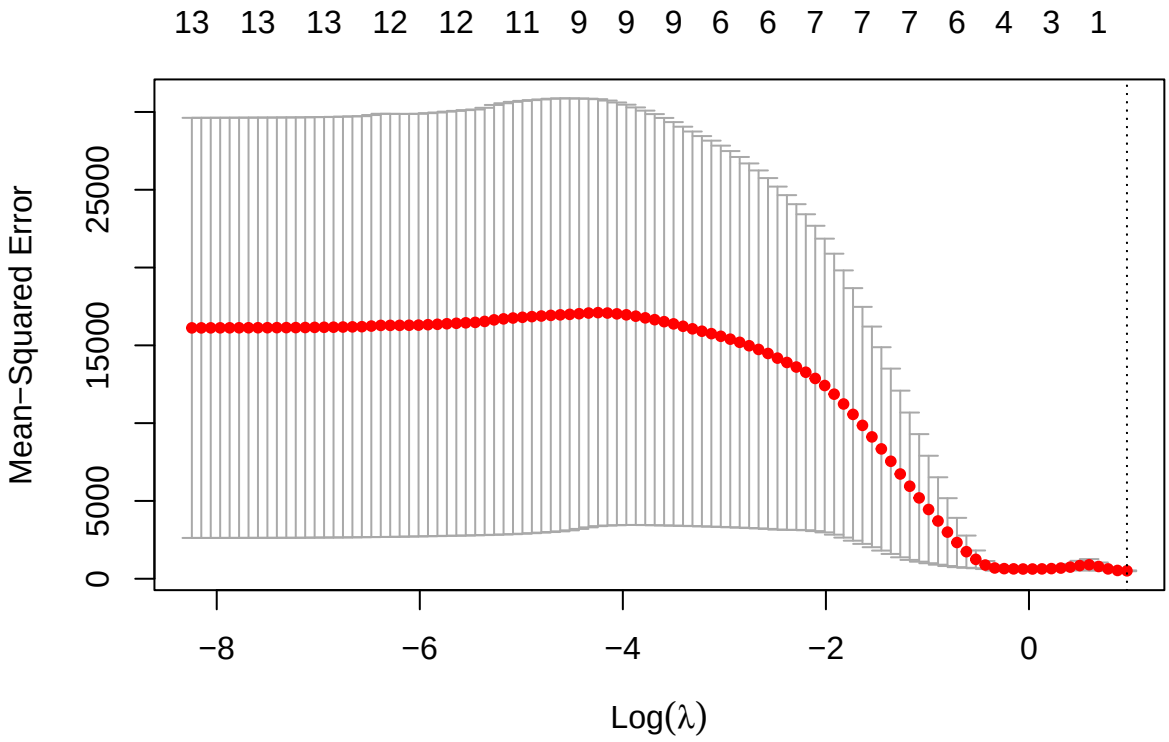
In this case we choose the model with lambda.1se which is the model with the largest value of lambda for which the MSE is not above one standard error from the minimum.The reason is within this analysis my aim is to interpret the results rather than predicting. So I have chosen the model with lambda.1se for my analysis.

Next, similarly we have perform lasso regression with alpha =1.

```
#lasso
lasso <- glmnet(predictors,response,alpha=1)
plot(lasso,label=TRUE)
```




```
cv_lasso = cv.glmnet(predictors,response,alpha=1)
plot(cv_lasso)
```




```
cv_lasso$lambda.min
```

```
## [1] 2.625243
```

```
cv_lasso$lambda.1se
```

```
## [1] 2.625243
```

Similar to ridge with also lasso I have chosen lamda.1se,

```
coef(cv_lasso)
```

```
## 15 x 1 sparse Matrix of class "dgCMatrix"
##                                       s1
## (Intercept)                     37.64266
## climateData_df$AG.LND.PRCP.MM     .
## climateData_df$AG.YLD.CREL.KG     .
## climateData_df$EG.ELC.HYRO.ZS     .
## climateData_df$EG.ELC.RNEW.ZS     .
## climateData_df$EG.ELC.RNWX.ZS     .
## climateData_df$EG.FEC.RNEW.ZS     .
## climateData_df$EN.ATM.CO2E.GF.KT  .
## climateData_df$EN.ATM.CO2E.KT     .
## climateData_df$EN.ATM.CO2E.LF.KT  .
## climateData_df$EN.ATM.CO2E.SF.KT  .
## climateData_df$EN.ATM.METH.KT.CE  .
## climateData_df$EN.ATM.NOXE.KT.CE  .
## climateData_df$EN.CLC.MDAT.ZS     .
```

```
## climateData_df$NV.AGR.TOTL.ZS      .
```

Compare the best cross-validated MSE achieved for Ridge Regression and the Lasso.

```
min(cv_ridge$cvm)
```

```
## [1] 482.3318
```

```
min(cv_lasso$cvm)
```

```
## [1] 507.7596
```

Comparing the best cross-validated MSE achieved for Ridge Regression and the Lasso, we can see that the Lasso achieves a smaller estimated error.

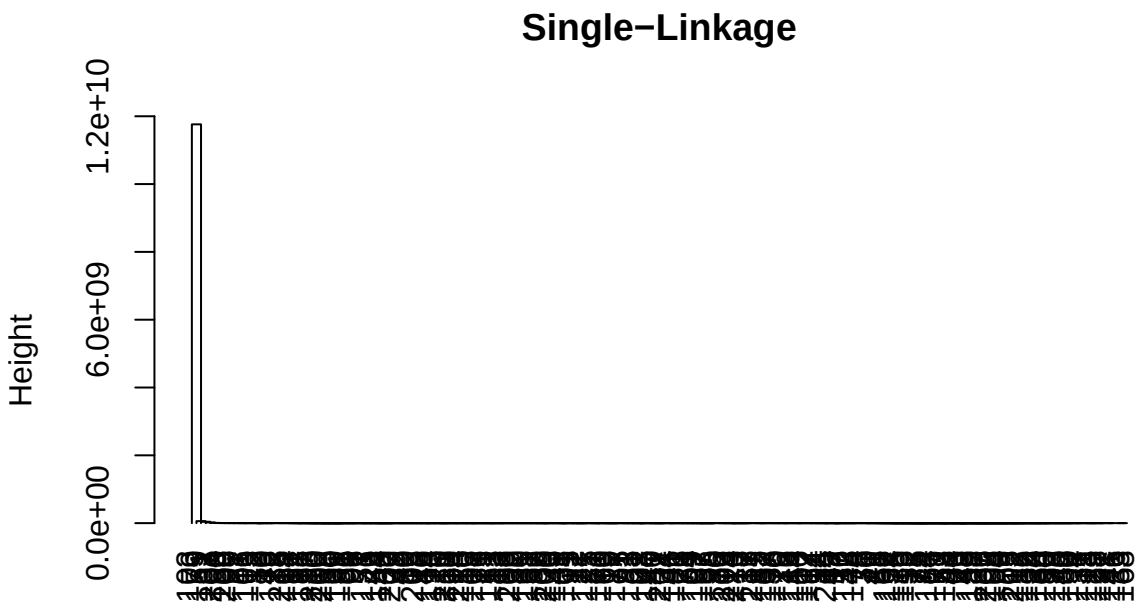
Cluster Analysis Following analysis explores whether there are clusters or groups of countries with similar climate and sustainability patterns.

```
climateData_df<- na.omit(climateData_df)
# Make a dissimiliarity matrix:
climateData.d <- dist(climateData_df[-1], method = "euclidean")
# Cluster the data using hclust():
c.h.s <- hclust(climateData.d,method="single")
c.h.c <- hclust(climateData.d,method="complete")
c.h.a <- hclust(climateData.d,method="average")
c.h.w <- hclust(climateData.d,method="ward.D2")
```

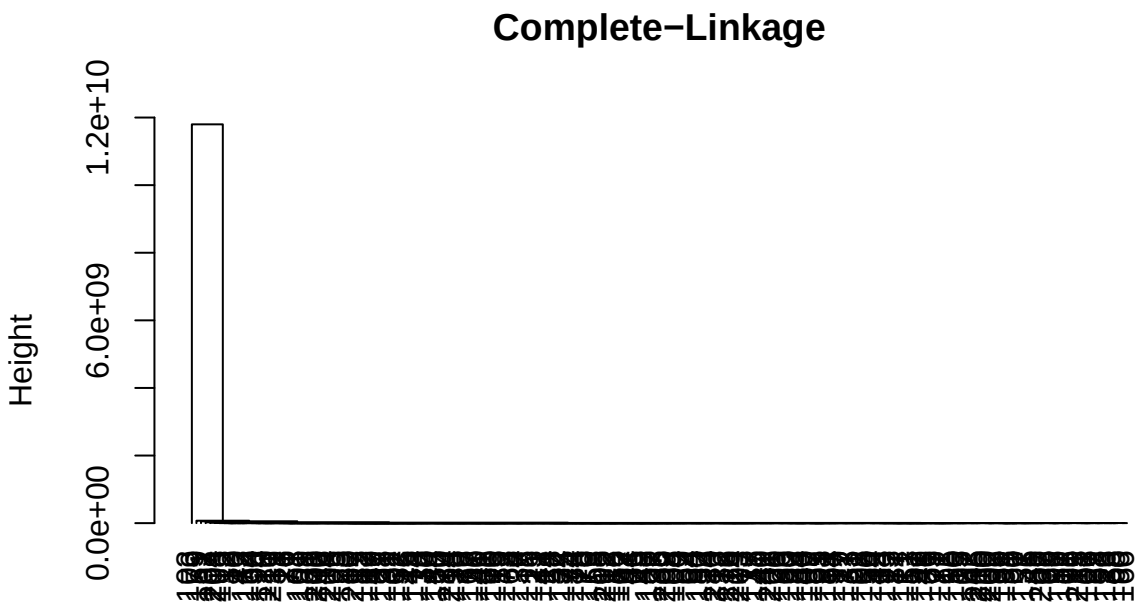
Carry out hierarchical clustering with SL, CL, AL, and Ward's

```
plot(c.h.s, main = "Single-Linkage", xlab = "", sub = "", hang = -1)
```

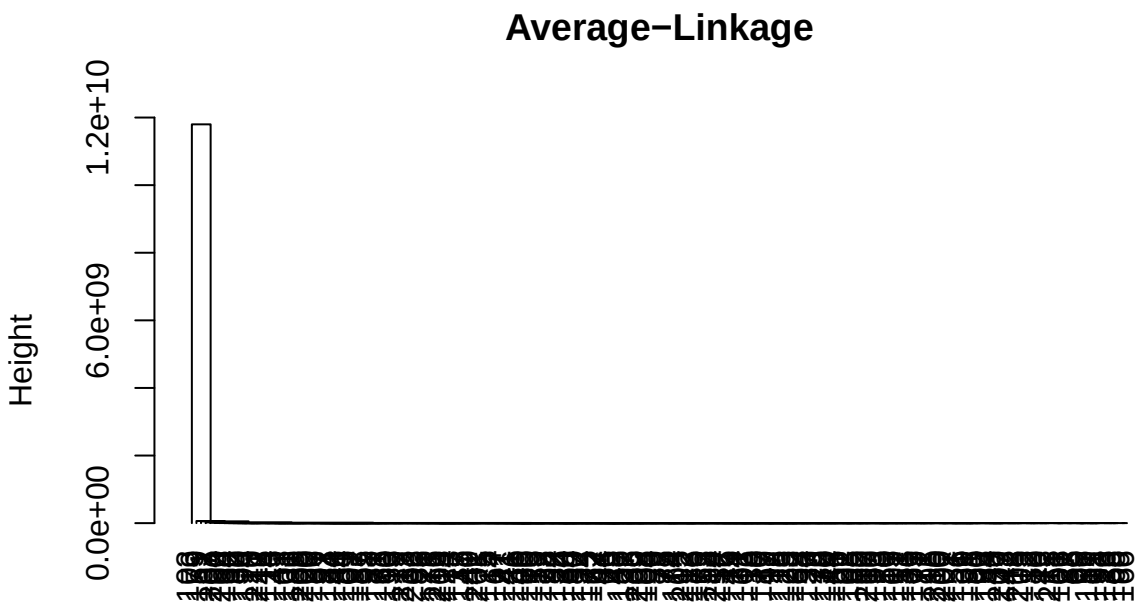
Plot dendograms for SL, CL, AL and Wards



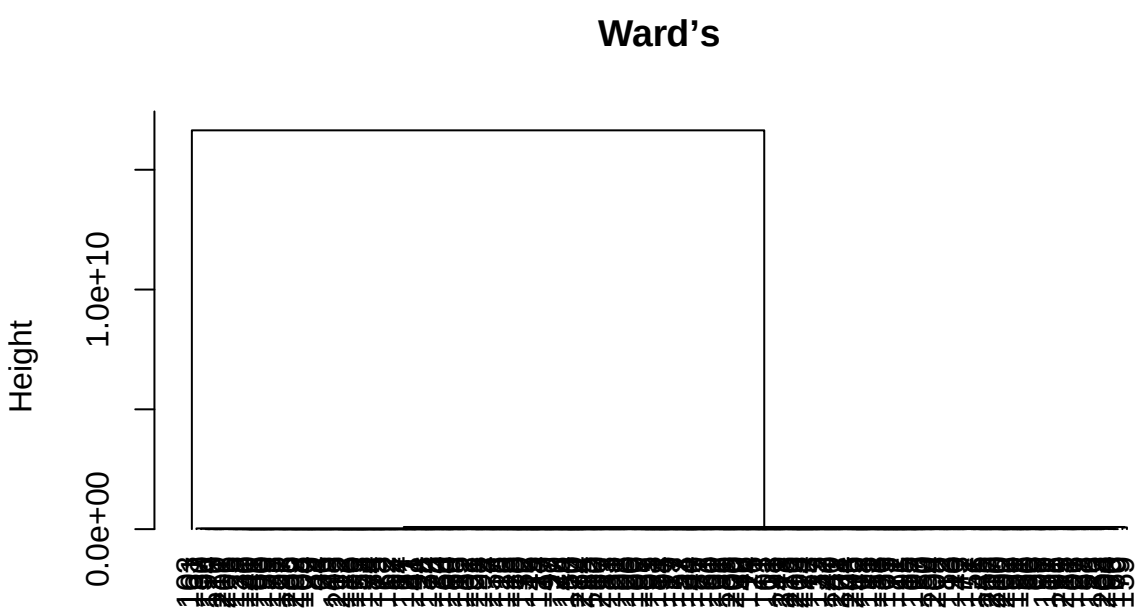

```
plot(c.h.c, main = "Complete-Linkage", xlab = "", sub = "", hang = -1)
```




```
plot(c.h.a, main = "Average-Linkage", xlab = "", sub = "", hang = -1)
```




```
plot(c.h.w, main = "Ward's", xlab = "", sub = "", hang = -1)
```



Choose a number of clusters to cut the tree Above algorithms produce dendrogram structures with more noticeable major clusters composed of subclusters.At a first glance, these dendrograms seem to suggest that a reasonable cut could be at about 4 clusters.

Using Silhouette Width Criterion (SWC) choose a number of cluster to cut the tree. In the following we perform a cut with k = 2, 3, 4, 5 and 6 and use the silhouette to assess Ward's results:

```
require(dplyr)
Wards_2clusters = cutree(tree = c.h.w, k = 2)
Wards_3clusters = cutree(tree = c.h.w, k = 3)
Wards_4clusters = cutree(tree = c.h.w, k = 4)
Wards_5clusters = cutree(tree = c.h.w, k = 5)
Wards_6clusters = cutree(tree = c.h.w, k = 6)

SWC <- function(clusterLabels, distanceMatrix){
  require(cluster)
  sil <- silhouette(x = clusterLabels, dist = distanceMatrix)
  return(mean(sil[,3]))
}
#k=2
(SWC_Wards2 <- SWC(clusterLabels = Wards_2clusters, dist = climateData.d))
```

```
## Loading required package: cluster
```

```
## [1] 0.9949768
```

```
#k=3
( SWC_Wards3 <- SWC(clusterLabels = Wards_3clusters, dist = climateData.d) )
```

```
## [1] 0.9727799
```

```
#k=4
( SWC_Wards4 <- SWC(clusterLabels = Wards_4clusters, dist = climateData.d) )
```

```
## [1] 0.9734515
```

```
#k=5
( SWC_Wards5 <- SWC(clusterLabels = Wards_5clusters, dist = climateData.d) )
```

```
## [1] 0.9621015
```

```
#k=6
( SWC_Wards6 <- SWC(clusterLabels = Wards_6clusters, dist = climateData.d) )
```

```
## [1] 0.9518022
```

Using the above results, it can be seen that k = 2, 3 and 4 clusters exhibit better silhouette values than k = 5 and 6. Thus I have used three solutions k=2, 3 and 4 for further analysis.

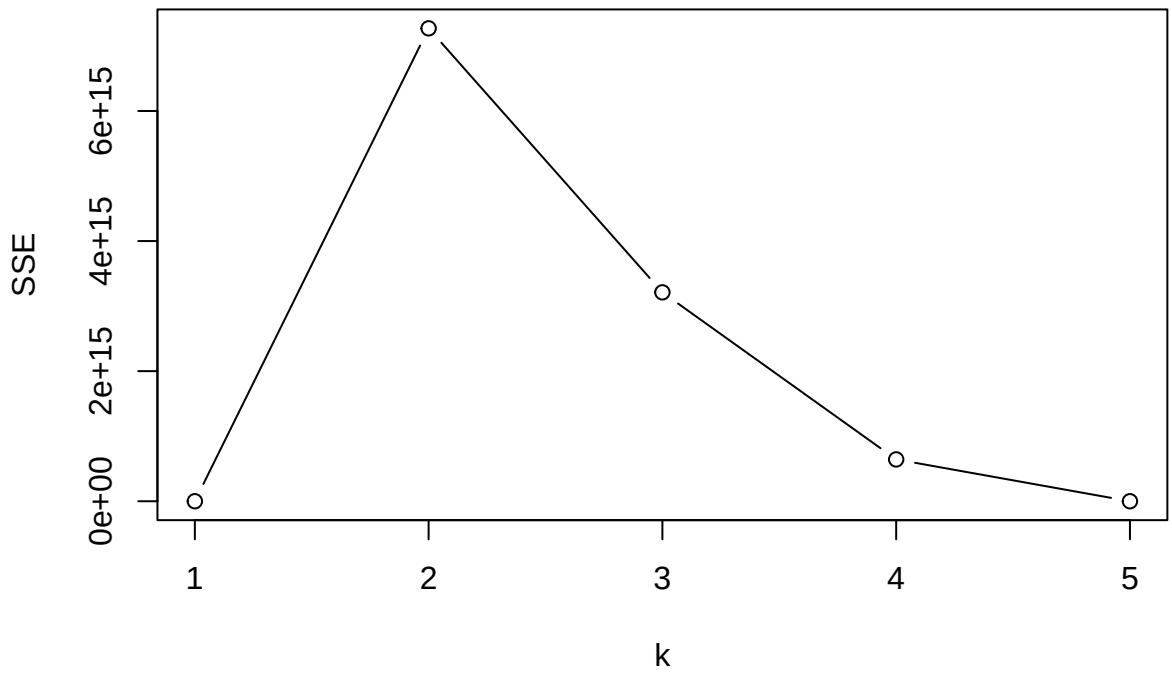
Run k-means Created a confusion matrix using the cluster labels and run k-means using the same number of clusters as the one selected to cut the trees in the previous analysis. Then, compared the k-means findings against the cut(s) of one hierarchical technique. Here I have calculated sum of squared distances(SSE) for each cluster to correctly determine the number of clusters(k).

```
set.seed(0)
SSE <- rep(0, 5)
#climateData_df<- na.omit(climateData_df)
km2 = kmeans(climateData_df[-1], centers = 2, nstart = 30)
SSE[2] <- km2$tot.withinss

km3 = kmeans(climateData_df[-1], centers = 3, nstart = 30)
SSE[3] <- km3$tot.withinss

km4 = kmeans(climateData_df[-1], centers = 4, nstart = 30)
SSE[4] <- km4$tot.withinss
```

Using the elbow method, we can determine the number of clusters (k) in K-means clustering. Plotting the sum of squared distances (SSE) for various values of k and locating a "elbow" point on the plot are the steps involved in the elbow method. The elbow point is the point on the SSE when the variance within each cluster is not appreciably reduced by adding further clusters.

To determine the appropriate number of clusters, I have ploted these SSE values and look for the elbow point.

```
plot(SSE, xlab="k", ylab="SSE", type = "b")
```



In order to compare with the results from Ward's compute the confusion matrix

```
( tab2 = table(km2$cluster,Wards_2clusters) )
```

```
##    Wards_2clusters
##       1   2
##   1   0   1
##   2 204   0
```

```
( tab3 = table(km3$cluster,Wards_3clusters) )
```

```
##    Wards_3clusters
##       1   2   3
##   1   0   2   0
##   2   0   0   1
##   3 202   0   0
```

```
( tab4 = table(km4$cluster,Wards_4clusters) )
```

```
##    Wards_4clusters
##       1   2   3   4
##   1 202   0   0   0
##   2   0   0   0   1
##   3   0   1   0   0
##   4   0   0   1   0
```

From elbow method above, it was able to select two k values which could be used for further analysis.

To interpret the cluster results in the analysis, I have examined the characteristics and patterns of the clusters

which I have selected before. This analysis will further able to give a correct value for number of clusters(k).

Calculate the mean or median values for each variable within each cluster and create visualizations to compare the clusters for each indicator.

Explore Relationships: Examine how the clusters relate to each other and what sets them apart. For example, you might find that Cluster 1 consists of countries with high agricultural yields and low CO2 emissions, while Cluster 2 has the opposite characteristics.

For the below plot I have used the indicator "EN.ATM.CO2E.GF.KT":CO2 emissions from gaseous fuel consumption (kt). These plots examine how the clusters relate to each other and what sets them apart. For example, you might find that Cluster 1 consists of countries with high agricultural yields and low CO2 emissions, while Cluster 2 has the opposite characteristics.

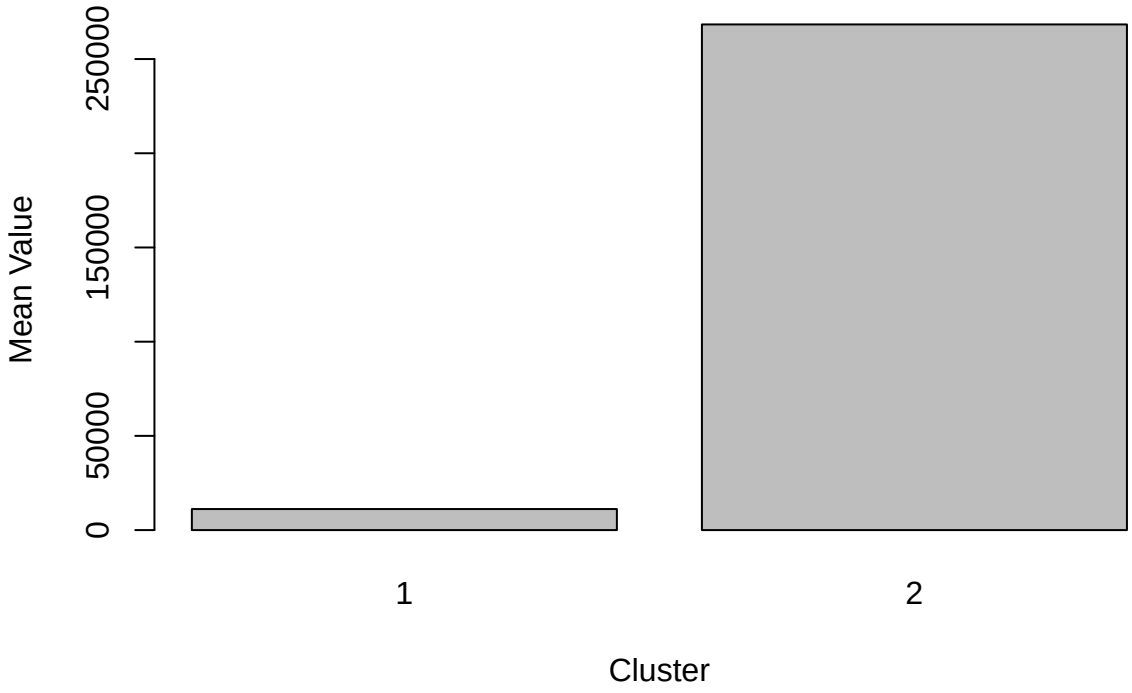
Similarly I have compared the plots for each indicator of the dataset with clusters k=2 and k=3.

Below plot cluster profiles for k=2 relevant to the indicator, percentage of CO2 emissions .First cluster contains countries with higher mean value of CO2 emissions while second cluster contains counties with low mean values.

```
#k=2
climateData_df$cluster <- km2$cluster

(cluster_profiles <- aggregate(. ~ cluster, data = climateData_df[-1], FUN = mean))
```

```
##   cluster AG.LND.AGRI.ZS AG.LND.ARBL.ZS AG.LND.IRIG.AG.ZS AG.LND.PRCP.MM
## 1       1       31.53184       22.15613      1.179859e+10         1904.0
## 2       2       37.67262       14.20080      3.938459e+05       105765.8
##   AG.YLD.CREL.KG EG.ELC.HYRO.ZS EG.ELC.RNEW.ZS EG.ELC.RNWX.ZS EG.FEC.RNEW.ZS
## 1         3623.5       45.48556       46.41801   9.324539e-01       61.83770
## 2       107890.8   111100.60717       31.78855   1.110873e+05       35.46117
##   EN.ATM.CO2E.GF.KT EN.ATM.CO2E.KT EN.ATM.CO2E.LF.KT EN.ATM.CO2E.SF.KT
## 1              0.00        11160.0          6435.585          2684.244
## 2          33956.01       268367.7         50895.846         68045.542
##   EN.ATM.METH.KT.CE EN.ATM.NOXE.KT.CE EN.CLC.MDAT.ZS NV.AGR.TOTL.ZS
## 1           20310.0            5160.0   6.643791e+00       22.83697
## 2          143630.2          118317.6   1.068504e+05   183123.22253
```

```
barplot(cluster_profiles$EN.ATM.CO2E.KT, names.arg = cluster_profiles$cluster, xlab = "Cluster", ylab =
```

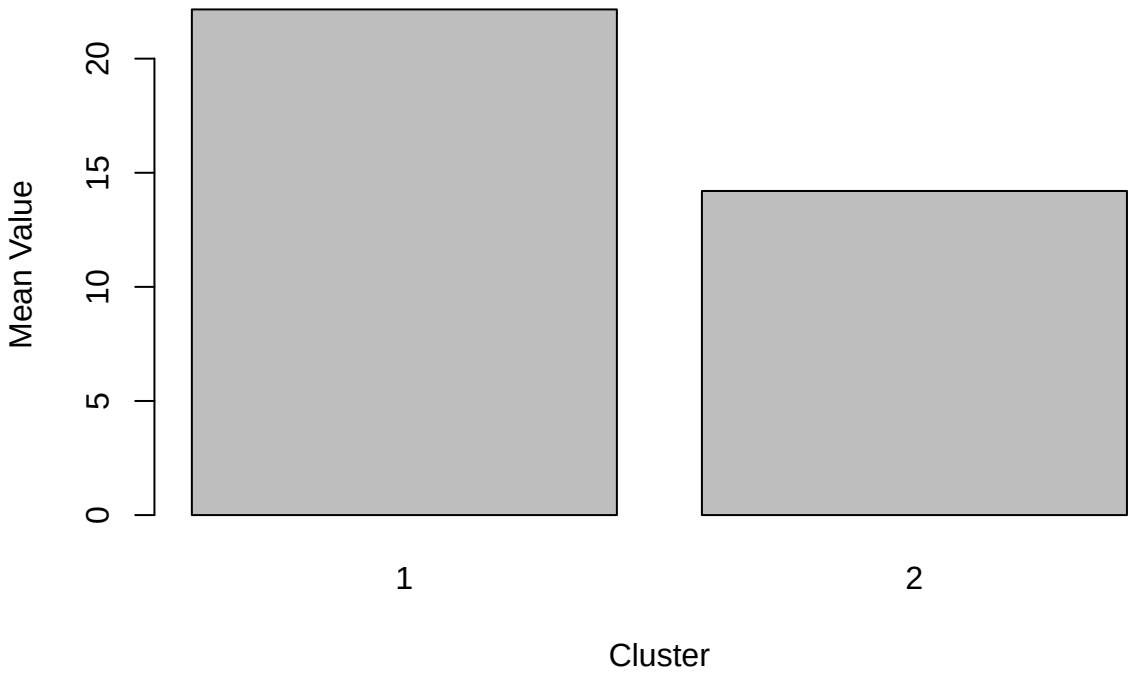


Same plot is created but with indicator which represent arable land percentage.

```
#k=2
climateData_df$cluster <- km2$cluster

(cluster_profiles <- aggregate(. ~ cluster, data = climateData_df[-1], FUN = mean))
```

```
##   cluster AG.LND.AGRI.ZS AG.LND.ARBL.ZS AG.LND.IRIG.AG.ZS AG.LND.PRCP.MM
## 1       1       31.53184       22.15613      1.179859e+10         1904.0
## 2       2       37.67262       14.20080      3.938459e+05       105765.8
##   AG.YLD.CREL.KG EG.ELC.HYRO.ZS EG.ELC.RNEW.ZS EG.ELC.RNWX.ZS EG.FEC.RNEW.ZS
## 1         3623.5       45.48556       46.41801   9.324539e-01       61.83770
## 2       107890.8   111100.60717       31.78855   1.110873e+05       35.46117
##   EN.ATM.CO2E.GF.KT EN.ATM.CO2E.KT EN.ATM.CO2E.LF.KT EN.ATM.CO2E.SF.KT
## 1              0.00        11160.0          6435.585          2684.244
## 2          33956.01       268367.7         50895.846         68045.542
##   EN.ATM.METH.KT.CE EN.ATM.NOXE.KT.CE EN.CLC.MDAT.ZS NV.AGR.TOTL.ZS
## 1           20310.0            5160.0   6.643791e+00       22.83697
## 2          143630.2          118317.6   1.068504e+05   183123.22253
```

```
barplot(cluster_profiles$AG.LND.ARBL.ZS, names.arg = cluster_profiles$cluster, xlab = "Cluster", ylab =
```



Interpretation and Discussion The interpretation of these clusters suggests that cluster 1 might represent countries with higher CO2 emissions, a more significant share of agriculture in their GDP, and greater reliance on hydropower. In contrast, cluster 2 includes regions with higher arable land proportions, more irrigated agricultural land, and higher cereal yields, despite lower CO2 emissions.This implies that CO2 emission significantly impact on agricultural sustainability.

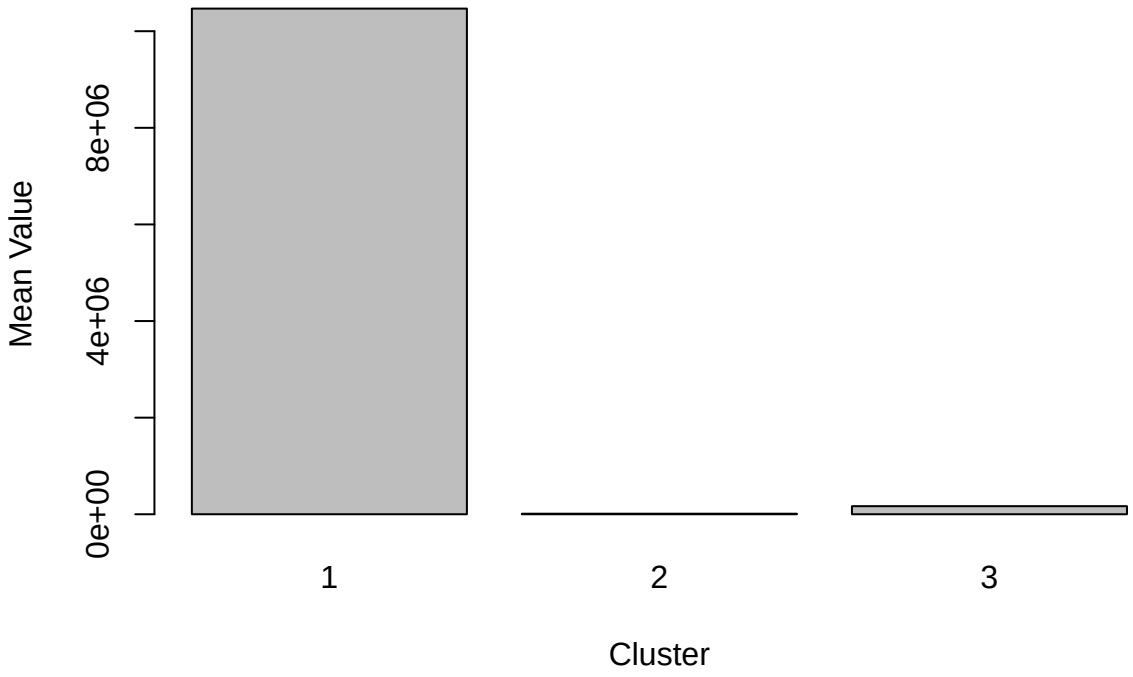
Below plot cluster profiles for k=3 relevant to the indicator, percentage of CO2 emissions. Second cluster contains countries with higher mean value of CO2 emissions while second and third clusters contains counties with low mean values.

```
#k=3
climateData_df$cluster <- km3$cluster

(cluster_profiles <- aggregate(. ~ cluster, data = climateData_df[-1], FUN = mean))
```

```
##   cluster AG.LND.AGRI.ZS AG.LND.ARBL.ZS AG.LND.IRIG.AG.ZS AG.LND.PRCP.MM
## 1       1       11.21717       9.759422      2.900910e+07   10459748.408
## 2       2       31.53184      22.156130      1.179859e+10       1904.000
## 3       3       37.93455      14.244770      1.105265e+05       3251.163
##   AG.YLD.CREL.KG EG.ELC.HYRO.ZS EG.ELC.RNEW.ZS EG.ELC.RNWX.ZS EG.FEC.RNEW.ZS
## 1   10461003.558   1.045926e+07       77.04381   1.045922e+07       22.56755
## 2       3623.500   4.548556e+01       46.41801   9.324539e-01       61.83770
## 3       5384.751   8.643605e+03       31.34048   8.630510e+03       35.58883
##   EN.ATM.CO2E.GF.KT EN.ATM.CO2E.KT EN.ATM.CO2E.LF.KT EN.ATM.CO2E.SF.KT
## 1              0.00     10468281.4          1855.502         10845.153
## 2              0.00        11160.0          6435.585          2684.244
## 3          34292.21       167378.4         51381.394         68611.883
##   EN.ATM.METH.KT.CE EN.ATM.NOXE.KT.CE EN.CLC.MDAT.ZS NV.AGR.TOTL.ZS
```

```
## 1          10468576       10460426.41   1.045922e+07   1.854989e+07
## 2             20310           5160.00   6.643791e+00   2.283697e+01
## 3             41403          15920.53   4.351690e+03   1.274031e+03
```

```
barplot(cluster_profiles$EN.ATM.CO2E.KT, names.arg = cluster_profiles$cluster, xlab = "Cluster", ylab =
```



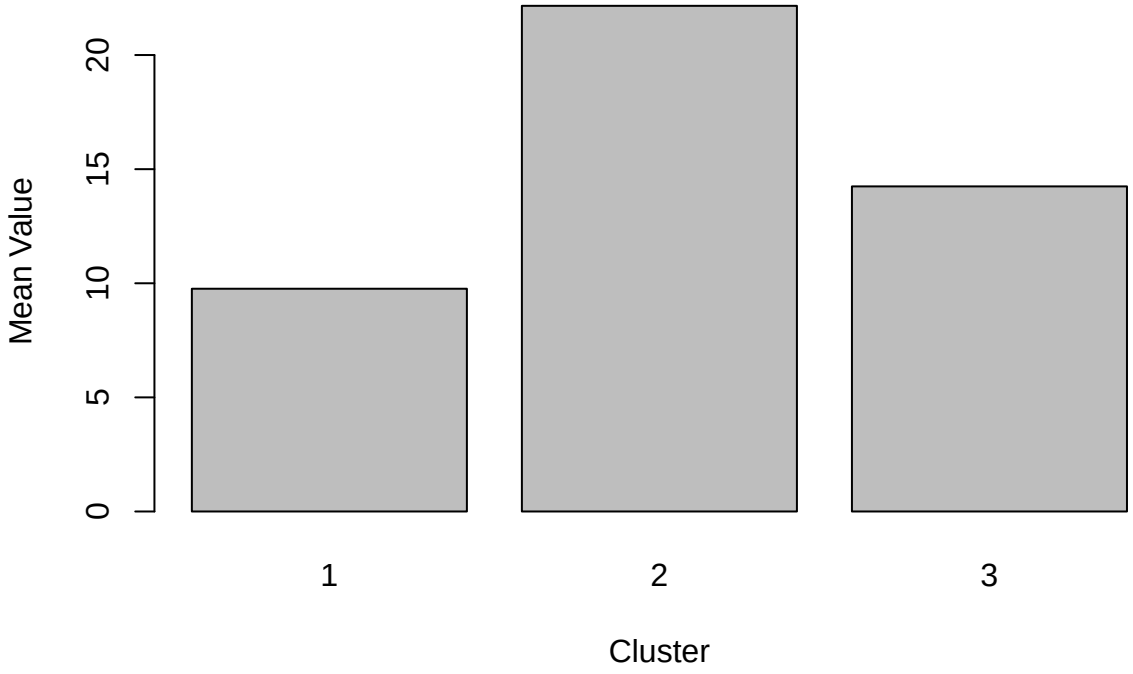
Same plot is created but with indicator which represent arable land percentage.

```
#k=3
climateData_df$cluster <- km3$cluster

(cluster_profiles <- aggregate(. ~ cluster, data = climateData_df[-1], FUN = mean))
```

```
##   cluster AG.LND.AGRI.ZS AG.LND.ARBL.ZS AG.LND.IRIG.AG.ZS AG.LND.PRCP.MM
## 1       1       11.21717       9.759422      2.900910e+07   10459748.408
## 2       2       31.53184      22.156130      1.179859e+10       1904.000
## 3       3       37.93455      14.244770      1.105265e+05       3251.163
##   AG.YLD.CREL.KG EG.ELC.HYRO.ZS EG.ELC.RNEW.ZS EG.ELC.RNWX.ZS EG.FEC.RNEW.ZS
## 1   10461003.558   1.045926e+07       77.04381   1.045922e+07       22.56755
## 2       3623.500   4.548556e+01       46.41801   9.324539e-01       61.83770
## 3       5384.751   8.643605e+03       31.34048   8.630510e+03       35.58883
##   EN.ATM.CO2E.GF.KT EN.ATM.CO2E.KT EN.ATM.CO2E.LF.KT EN.ATM.CO2E.SF.KT
## 1              0.00     10468281.4          1855.502         10845.153
## 2              0.00        11160.0          6435.585          2684.244
## 3          34292.21       167378.4         51381.394         68611.883
##   EN.ATM.METH.KT.CE EN.ATM.NOXE.KT.CE EN.CLC.MDAT.ZS NV.AGR.TOTL.ZS
## 1          10468576       10460426.41   1.045922e+07   1.854989e+07
## 2             20310           5160.00   6.643791e+00   2.283697e+01
```

```
## 3             41403          15920.53   4.351690e+03   1.274031e+03
```

```
barplot(cluster_profiles$AG.LND.ARBL.ZS, names.arg = cluster_profiles$cluster, xlab = "Cluster", ylab =
```



From the above plots its is shown that clusters with k= 3, gives a better interpretation to the data as well as to my study. With these it is concluded that the best choice for k is 3.

Interpretation and Discussion

The dataset consists of three distinct clusters, each with varying agricultural and environmental factors. Cluster 1 has a significant focus on agriculture, with 40.79% of land area used for agriculture. The cluster also has a moderate emphasis on crop cultivation, with 16.47% of land classified as arable. Approximately 8.89% of agricultural land is irrigated, suggesting some investment in irrigation. The average precipitation is 1065.99 mm, providing enough water for agricultural needs.

Cluster 2 has unique characteristics, with a higher percentage of arable land, lower irrigation of agricultural land, and higher precipitation. The electricity production and consumption profiles are distinct, with a focus on renewable energy, particularly wind energy. CO2 emissions are significantly higher in this cluster, suggesting a different economic and environmental profile compared to cluster 1.

Cluster 3 has an average of 32.44% of land area used for agriculture, with 28.34% classified as arable. The cluster does not rely on hydropower for electricity production, with renewable energy production being relatively low. Wind energy production is relatively low in this cluster, with renewable energy consumption accounting for about 7.46% of the final energy consumption.

The dataset's completeness is high, with an average of 5.48e+07 (representing a very high level of completeness). Agriculture contributes only minimally to the GDP, with an average of approximately 0.65%. These three clusters represent distinct agricultural and environmental profiles, with varying levels of agricultural focus, renewable energy utilization, emissions, and contributions of agriculture to the GDP.

However according to the data provided, Australia belongs to cluster 2 which characterized with higher percentage of arable land, lower irrigation of agricultural land, and higher precipitation.Not only, CO2 emissions are significantly higher within this cluster, which means Australia might have an impact by CO2 emission for its agricultural sustainability.

PCA Analysis Principal component analysis (PCA) is a classic multivariate analysis technique that is used for different purposes in the context of statistical learning, most noticeably for dimensionality reduction and data visualisation.

Apply PCA for the dataset Here scale attribute has set to true in order to standardized the dataset.

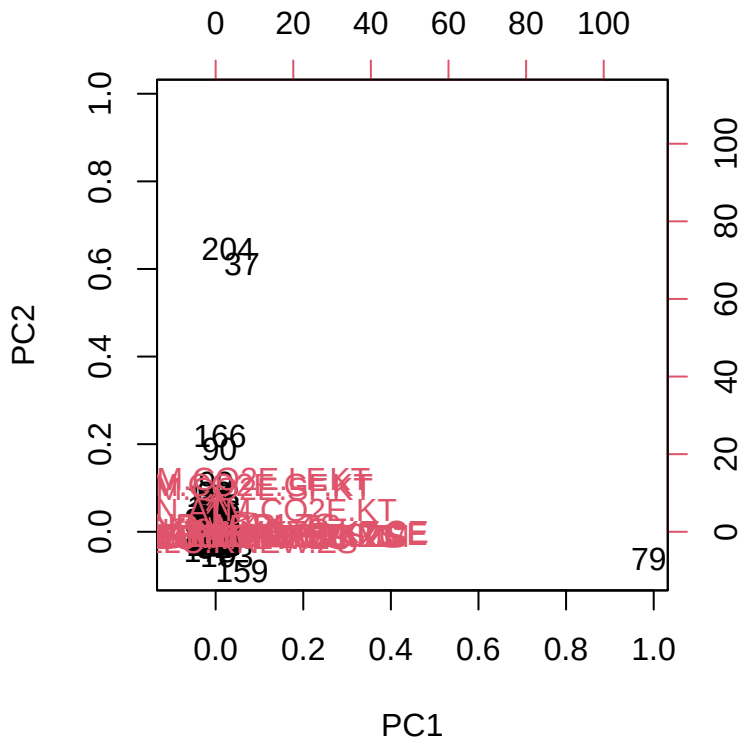
```
PCA <- prcomp(climateData_df[-1], scale = TRUE)
PCA$rotation
```

```
##                            PC1          PC2          PC3         PC4
## AG.LND.AGRI.ZS    -0.047890682  0.086473556 -0.596050130  0.30574162
## AG.LND.ARBL.ZS    -0.030403596  0.076191268 -0.643999445  0.22622103
## AG.LND.IRIG.AG.ZS  0.009714981 -0.035659793 -0.170767212 -0.30638260
## AG.LND.PRCP.MM     0.344999120 -0.039895793  0.006749026  0.06909185
## AG.YLD.CREL.KG     0.345008946 -0.039595624  0.006852991  0.06900770
## EG.ELC.HYRO.ZS     0.344919265 -0.040517113  0.007733823  0.06934066
## EG.ELC.RNEW.ZS     0.041824096 -0.107868513 -0.133423308 -0.11410821
## EG.ELC.RNWX.ZS     0.344919103 -0.040515561  0.007733756  0.06934466
## EG.FEC.RNEW.ZS    -0.009677441 -0.064329227 -0.247609602  0.07090094
## EN.ATM.CO2E.GF.KT  0.011281827  0.532949906  0.072622334 -0.09483056
## EN.ATM.CO2E.KT     0.312895405  0.251310691  0.002442041  0.02358188
## EN.ATM.CO2E.LF.KT  0.016634081  0.604671806  0.014921150 -0.08555425
## EN.ATM.CO2E.SF.KT  0.022193379  0.488973699 -0.029323518 -0.05619826
## EN.ATM.METH.KT.CE  0.345812569  0.006577882  0.005111175  0.06313372
## EN.ATM.NOXE.KT.CE  0.345630630 -0.022385623  0.005121805  0.06748828
## EN.CLC.MDAT.ZS     0.344974639 -0.039967097  0.008011644  0.06884939
## NV.AGR.TOTL.ZS     0.021027481 -0.057322847 -0.255176681 -0.66386254
## cluster           -0.248101166  0.076716981  0.223403063  0.50025460
##                            PC5           PC6           PC7           PC8
## AG.LND.AGRI.ZS    -0.131507309 -0.0755994200  0.2023881195 -0.009986241
## AG.LND.ARBL.ZS    -0.186923923 -0.0132523049 -0.0328793950  0.097400922
## AG.LND.IRIG.AG.ZS -0.057319855  0.8875731315  0.1139048574 -0.009768009
## AG.LND.PRCP.MM    -0.012093959 -0.0004209977 -0.0071179037  0.031299203
## AG.YLD.CREL.KG    -0.012335813 -0.0004391948 -0.0070626703  0.031543032
## EG.ELC.HYRO.ZS    -0.012991753 -0.0003795369 -0.0084146729  0.030403857
## EG.ELC.RNEW.ZS     0.685692914 -0.1031567936  0.6650434843  0.131585488
## EG.ELC.RNWX.ZS    -0.013000906 -0.0003787798 -0.0084219586  0.030402921
## EG.FEC.RNEW.ZS     0.656728049  0.1002258466 -0.6912946418  0.004491896
## EN.ATM.CO2E.GF.KT  0.015671061  0.0146267788 -0.0548737941  0.595321730
## EN.ATM.CO2E.KT     0.044410305  0.0004398880  0.0125870245 -0.191981323
## EN.ATM.CO2E.LF.KT  0.073632431  0.0069604701  0.0006891739  0.173819373
## EN.ATM.CO2E.SF.KT  0.122451193 -0.0033732315  0.0602634018 -0.738706142
## EN.ATM.METH.KT.CE -0.003504335 -0.0003037018 -0.0039236353  0.012950463
## EN.ATM.NOXE.KT.CE -0.008169153 -0.0005273715 -0.0053597855  0.017334863
## EN.CLC.MDAT.ZS    -0.013093488 -0.0002773504 -0.0081535997  0.031237938
## NV.AGR.TOTL.ZS    -0.101922396 -0.4304375226 -0.1233347249 -0.023360691
## cluster            0.095417213 -0.0092177628  0.0494621198 -0.002279746
##                            PC9         PC10          PC11          PC12
## AG.LND.AGRI.ZS     0.6908739799  0.009778254 -0.0010810563  0.0001997860
## AG.LND.ARBL.ZS    -0.6936715059  0.021824476 -0.0041366842 -0.0006622450
```

```
## AG.LND.IRIG.AG.ZS  0.0462404120  0.002000047  0.0047706060  0.0459481403
## AG.LND.PRCP.MM     0.0050607750 -0.013867003 -0.0828835265 -0.2596537319
## AG.YLD.CREL.KG     0.0047118370 -0.013599003 -0.0861466288 -0.2714790713
## EG.ELC.HYRO.ZS     0.0050078059 -0.014428634 -0.3099467399  0.5186065160
## EG.ELC.RNEW.ZS    -0.1221986464  0.020666283 -0.0043018937  0.0004035812
## EG.ELC.RNWX.ZS     0.0050072715 -0.014429664 -0.3100175694  0.5186302108
## EG.FEC.RNEW.ZS     0.1009129267  0.001238731 -0.0000255339 -0.0002363732
## EN.ATM.CO2E.GF.KT  0.0597193291  0.578237483 -0.0368692978  0.0089387665
## EN.ATM.CO2E.KT    -0.0166602692  0.053899310 -0.0933974521 -0.3282140846
## EN.ATM.CO2E.LF.KT  0.0040625380 -0.762365891  0.0229247030  0.0433495337
## EN.ATM.CO2E.SF.KT -0.0695393291  0.276406253 -0.0216944085  0.1099693465
## EN.ATM.METH.KT.CE  0.0005937244  0.049565649  0.8514026143  0.2886527633
## EN.ATM.NOXE.KT.CE  0.0028837974 -0.027184854  0.1839556320 -0.1921931783
## EN.CLC.MDAT.ZS     0.0044329433 -0.009262668 -0.1528039826 -0.2233925579
## NV.AGR.TOTL.ZS     0.0518673836  0.002187011  0.0100437330  0.0912622080
## cluster           -0.0524040814  0.007041329  0.0147315829  0.1362667359
##                            PC13          PC14          PC15          PC16
## AG.LND.AGRI.ZS     0.0006770439 -0.0001372959  0.0008031312 -3.977912e-04
## AG.LND.ARBL.ZS    -0.0004937362 -0.0008176836 -0.0003638123 -5.420324e-05
## AG.LND.IRIG.AG.ZS  0.0458544699 -0.1568924860 -0.1972774609  3.934606e-02
## AG.LND.PRCP.MM     0.0828868857  0.3146076451 -0.4128955699 -1.275274e-01
## AG.YLD.CREL.KG     0.0852732606  0.3184302668 -0.4150481044 -1.321979e-01
## EG.ELC.HYRO.ZS    -0.0877379859 -0.0075914903  0.0086572314  1.534386e-02
## EG.ELC.RNEW.ZS     0.0002935308 -0.0005850585  0.0005144278  1.535109e-04
## EG.ELC.RNWX.ZS    -0.0876567087 -0.0074532686  0.0085519830  1.519243e-02
## EG.FEC.RNEW.ZS     0.0002594267 -0.0001724697 -0.0001231798 -2.637404e-04
## EN.ATM.CO2E.GF.KT  0.0337853449  0.0381045719 -0.0044160941  4.140519e-02
## EN.ATM.CO2E.KT    -0.5693540933 -0.4992309388  0.0556279945 -3.213511e-01
## EN.ATM.CO2E.LF.KT  0.0670860092  0.0508313408 -0.0054806710  1.148479e-02
## EN.ATM.CO2E.SF.KT  0.2142658687  0.1929976043 -0.0219864658  1.209482e-01
## EN.ATM.METH.KT.CE  0.0630561082 -0.0408705547 -0.0030417017 -2.446030e-01
## EN.ATM.NOXE.KT.CE -0.1905837058 -0.0507650909  0.0382685393  8.735945e-01
## EN.CLC.MDAT.ZS     0.7247909113 -0.4029482240  0.3326561375 -4.080446e-02
## NV.AGR.TOTL.ZS     0.0911896189 -0.3135089911 -0.3939099306  7.827577e-02
## cluster            0.1368266522 -0.4684833580 -0.5892864255  1.171095e-01
##                            PC17          PC18
## AG.LND.AGRI.ZS    -1.812781e-04 -5.778272e-07
## AG.LND.ARBL.ZS     1.908737e-04 -2.418918e-06
## AG.LND.IRIG.AG.ZS -5.833286e-04  2.311152e-06
## AG.LND.PRCP.MM     7.115379e-01 -3.133698e-05
## AG.YLD.CREL.KG    -7.026058e-01 -1.290136e-04
## EG.ELC.HYRO.ZS    -3.811674e-03 -7.071234e-01
## EG.ELC.RNEW.ZS    -5.466044e-05  8.283583e-06
## EG.ELC.RNWX.ZS    -3.908218e-03  7.070901e-01
## EG.FEC.RNEW.ZS    -1.975773e-04  4.880880e-07
## EN.ATM.CO2E.GF.KT  3.569836e-04 -3.196308e-06
## EN.ATM.CO2E.KT     1.654886e-03  5.213067e-05
## EN.ATM.CO2E.LF.KT -1.878850e-04 -7.973439e-06
## EN.ATM.CO2E.SF.KT -3.234796e-04 -2.239932e-05
## EN.ATM.METH.KT.CE -3.557602e-03  1.137077e-05
## EN.ATM.NOXE.KT.CE -2.227312e-03  1.243751e-04
## EN.CLC.MDAT.ZS     1.977787e-03  1.411864e-05
## NV.AGR.TOTL.ZS    -1.128370e-03  4.235267e-06
## cluster           -1.697940e-03  4.292642e-06
```

Visualize the relationships between variables and principal components

```
biplot(PCA , scale = 1)
```



Variable sdev, which stores the standard deviation along each of the principal components derived from prcomp(), quantifies the idea of the order of principal components according to the variance of the data along the corresponding directions. The Proportion of the Variance Explained (PVE) by each component is obtained by dividing the respective variances, which are returned by the square sdev^2, by the total sum of variances.

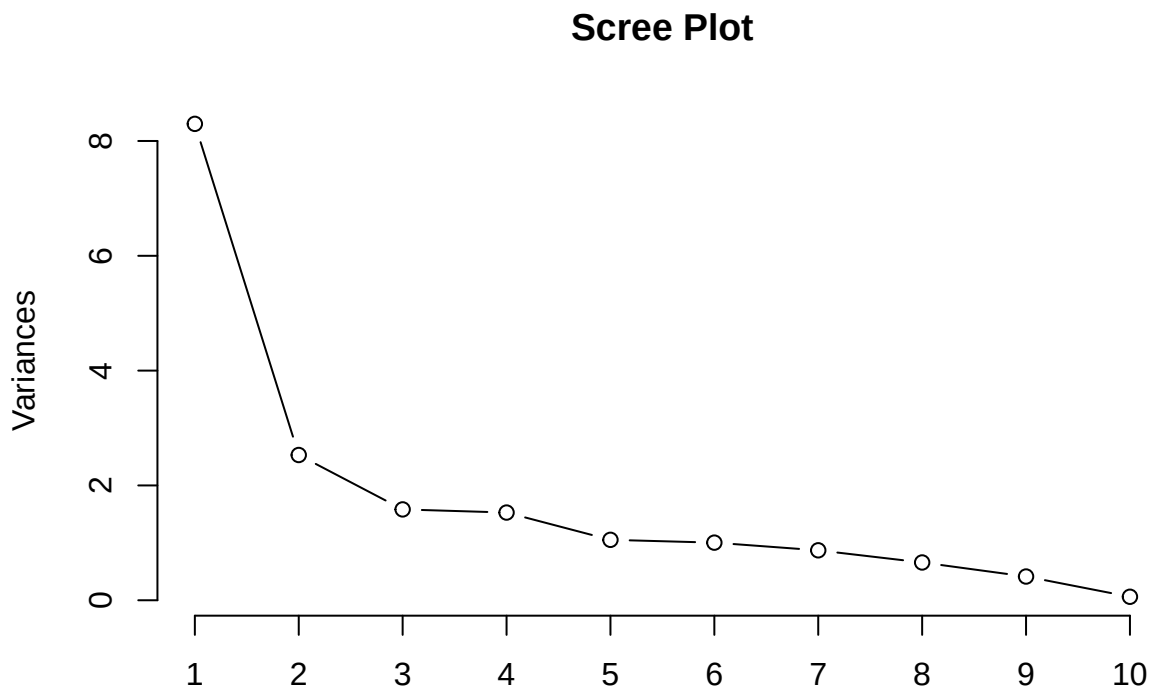
```
PVE <- (PCA$sdev^2)/sum(PCA$sdev^2)
```

Create a scree plot to visualize the proportion of variance explained by each principal component

```
summary(PCA)
```

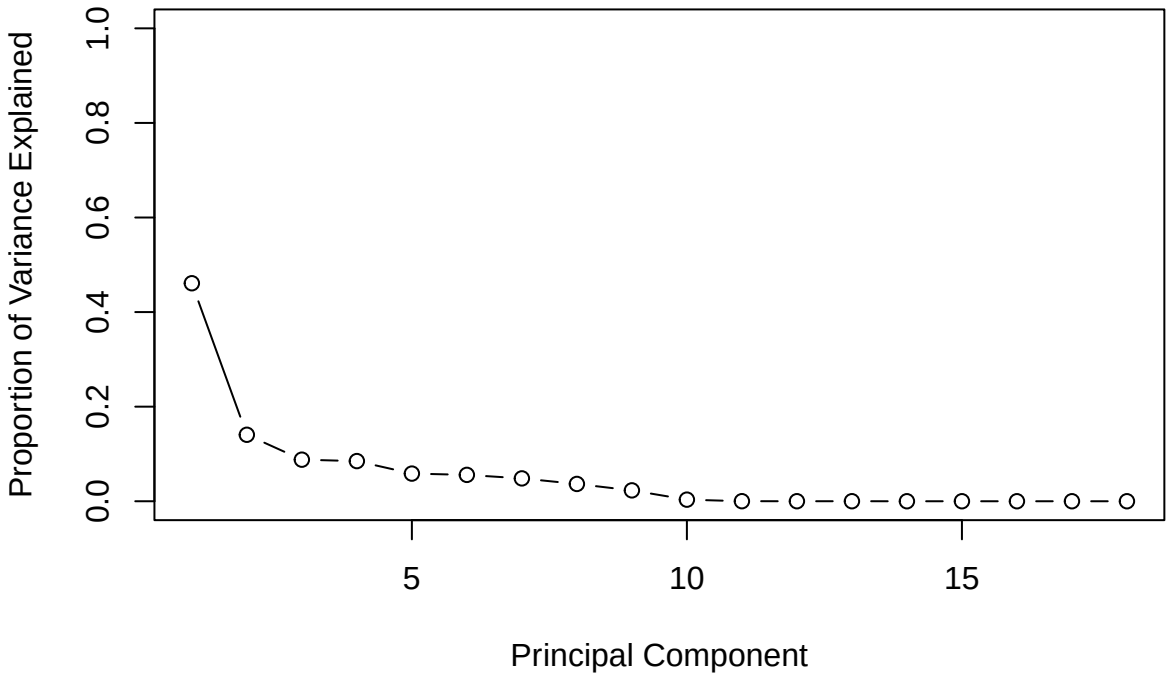
```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.881 1.5906 1.25797 1.23605 1.02603 1.00187 0.93268
## Proportion of Variance 0.461 0.1406 0.08792 0.08488 0.05849 0.05576 0.04833
## Cumulative Proportion  0.461 0.6016 0.68947 0.77435 0.83284 0.88860 0.93693
##                           PC8     PC9     PC10    PC11    PC12    PC13    PC14
## Standard deviation     0.81128 0.64304 0.24695 0.03883 0.02671 0.01435 0.01049
## Proportion of Variance 0.03657 0.02297 0.00339 0.00008 0.00004 0.00001 0.00001
## Cumulative Proportion  0.97349 0.99646 0.99985 0.99994 0.99997 0.99999 0.99999
##                           PC15     PC16     PC17      PC18
## Standard deviation     0.009573 0.006466 0.001589 1.026e-05
## Proportion of Variance 0.000010 0.000000 0.000000 0.000e+00
## Cumulative Proportion  1.000000 1.000000 1.000000 1.000e+00
```

```r
plot(PCA, type = "l", main = "Scree Plot")
```



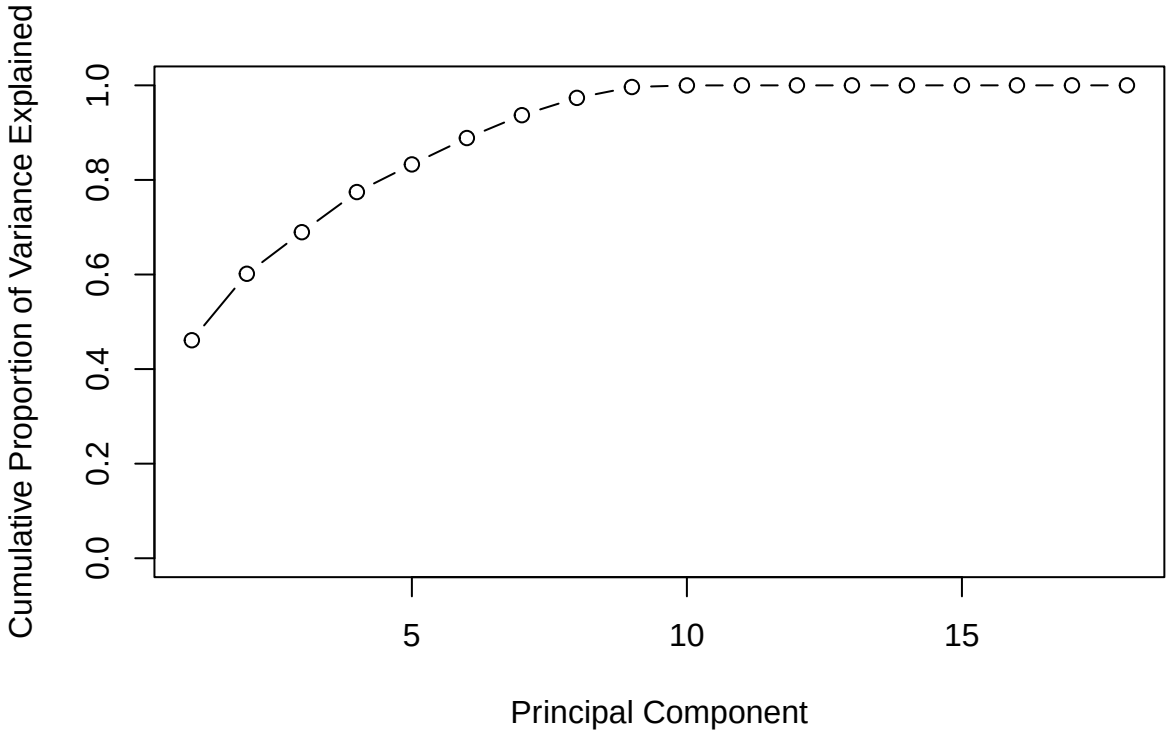
Here shows another visualization for proportion of variance Given below a plot of the proportion of variance explained by each principal component.

```r
plot(PVE , xlab = "Principal Component",
     ylab = "Proportion of Variance Explained", ylim = c(0, 1),
     type = "b")
```



Given below a plot of the cumulative proportion of variance explained.

```
plot(cumsum(PVE), xlab = "Principal Component",
     ylab = "Cumulative Proportion of Variance Explained",
     ylim = c(0, 1), type = "b")
```



Using the above plots and data we can conclude that, PC1 explains about 29.69% of the total variance, PC2 explains about 28.37%, and PC3 explains about 9.31%. The cumulative proportion of PC2 (PC1 + PC2) is 58.06%, which means that the first two components together explain 58.06% of the total variance.Thus, majority of the variance explained by both PC1 and PC2. However, as you go down the list, the explained variance by each component decreases, with PC17 explaining only a very small amount of variance.

Now let us pre process the dataset with PCA data with PC1 and PC2.

```
#install.packages("caret")
#install.packages("recipes")

library(caret)
```

```
## Warning: package 'caret' was built under R version 4.1.3
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
pca_model <- preProcess(climateData_df[-1], method = 'pca', pcaComp = 2)
```

Then transform the original data (“climateData_df[-1]”) based on the PCA model.

```
pca_data_update <- predict(pca_model, climateData_df[-1])
pca_data_update
```

```
##             PC1          PC2
## 2   -0.11667591 -0.470280966
## 3   -0.23064378 -0.420205874
## 4   -0.21948712 -0.523300145
## 5   -0.17846400 -0.601001123
## 6   -0.14122406  0.309996399
## 7   -0.22879763  0.455324000
## 8   -0.32894425 -0.231811675
## 9   -0.28974958 -0.294501942
## 10  -0.12870004 -0.353915243
## 11  -0.14878588  0.611421720
## 12  -0.19490926 -0.344006620
## 13  -0.35677152  0.007843563
## 14  -0.38162701 -0.232181505
## 15  -0.30830716  0.106419082
## 16  -0.32906721 -0.226108272
## 17  -0.33733055 -0.243635433
## 18  -0.44373962  0.404856528
## 19  -0.34144808 -0.089756124
## 20  -0.21998656 -0.287271911
## 21  -0.20527881 -0.455622987
## 22  -0.29335540 -0.295403833
## 23  -0.33389929  0.057417956
## 24  -0.17361903 -0.596884452
## 25  -0.22791182 -0.415701936
## 26  -0.23444693 -0.346452692
## 27   0.06643043  0.745457755
## 28  -0.28793594 -0.287822441
## 29  -0.20617639 -0.426019725
## 30  -0.10119980 -0.795863030
## 31  -0.30307688 -0.296598959
## 32  -0.09587960 -0.811707020
## 33   0.03597372  1.199834057
## 34  -0.21916421 -0.373778645
## 35  -0.35513932 -0.223863953
## 36  -0.17100615 -0.314989328
## 37   2.48687869 13.908044713
## 38  -0.35365828 -0.219265886
## 39  -0.16824394 -0.604648838
## 40  -0.11455958 -0.783373739
## 41  -0.21173584 -0.558655287
## 42  -0.18184977 -0.235807729
## 43  -0.32729366 -0.047999408
## 44  -0.25046667 -0.405828017
## 45  -0.16238937 -0.636067324
## 46  -0.38420292 -0.008776895
## 48  -0.22960363 -0.419901071
## 49  -0.24702481 -0.369081137
## 50  -0.33170698  0.049415034
## 51  -0.14300668  1.658302826
## 52  -0.36528844 -0.200724061
## 53  -0.27330943 -0.353381386
## 54  -0.38897690 -0.024477827
## 55  -0.33020538 -0.149483641
```

```
## 56  -0.19292867  0.130855464
## 57  -0.17605324 -0.421724005
## 58  -0.12334098  0.324030794
## 59  -0.38832702 -0.191551938
## 60  -0.24564914  0.500226593
## 61  -0.26906786 -0.336466162
## 62  -0.14966966 -0.622582708
## 63  -0.16990205 -0.464327742
## 64  -0.21789902 -0.491653120
## 65  -0.26650466  0.836263791
## 66  -0.13654720 -0.650892597
## 67  -0.27592563 -0.336976653
## 68  -0.17682316 -0.630032534
## 69  -0.26693548  1.146641764
## 70  -0.18596567 -0.557614455
## 71  -0.31970430 -0.247831814
## 73  -0.26029565 -0.471320417
## 74  -0.43663713 -0.027841588
## 75  -0.29920221 -0.361486978
## 76  -0.15653907 -0.584059303
## 77  -0.28814356 -0.129016025
## 78  -0.27435746 -0.335259724
## 79  40.78716168 -1.427955268
## 80  -0.22041031 -0.490555353
## 82  -0.11256727 -0.451956206
## 83  -0.21669132 -0.351456306
## 84  -0.23467302 -0.449848474
## 85  -0.20303408 -0.479465182
## 86  -0.43192389 -0.063372316
## 87  -0.40779472  0.126093749
## 88  -0.05120343  1.013394829
## 89  -0.44799263  0.032316104
## 90   0.36896884  4.282238543
## 91  -0.30831172 -0.163661234
## 92  -0.06380320  2.021095982
## 93  -0.21609677  0.207638270
## 94  -0.12469120 -0.774754729
## 95  -0.28375984 -0.090137887
## 96  -0.19570006  0.712449537
## 97  -0.30222823 -0.261613763
## 98  -0.22437029 -0.340593861
## 99   0.06071070  2.535158236
## 100 -0.32204362  0.444554565
## 101 -0.20922460 -0.519924497
## 102 -0.22658315 -0.491968488
## 103  1.06882709 -1.275430754
## 104 -0.29576329 -0.346600703
## 105 -0.29087892 -0.286735932
## 106 -0.10295089  1.049676362
## 107 -0.19340083 -0.076572138
## 108 -0.12592358 -0.683845162
## 109 -0.36088667 -0.083595102
## 110 -0.24994098 -0.425673878
## 111 -0.20079682 -0.289281468
```

```
## 112 -0.25275297 -0.380910214
## 113  0.02436476 -0.654900905
## 114 -0.28625584 -0.304703243
## 115 -0.28449194 -0.430305034
## 116 -0.32969195 -0.206684760
## 117 -0.33506068 -0.203111318
## 118 -0.25510658 -0.403503653
## 121 -0.34740182 -0.020364908
## 123 -0.46030493  0.067505897
## 124 -0.30093496 -0.397088908
## 125 -0.28073830 -0.294552004
## 126 -0.16113250  1.301666016
## 127 -0.33344098 -0.229023380
## 128 -0.30275307 -0.299459816
## 129 -0.21082218 -0.502732196
## 130 -0.32720390 -0.209470739
## 131 -0.18608969 -0.451889537
## 132 -0.18592990 -0.582464265
## 133 -0.34754563 -0.157757998
## 134 -0.21998718 -0.422830363
## 135 -0.22239304 -0.534148273
## 136 -0.27018805 -0.372424586
## 137 -0.34761780 -0.166701388
## 138 -0.30671021 -0.373067832
## 139 -0.17628045  0.296286096
## 140 -0.18344183 -0.613206323
## 141 -0.20631163 -0.467861279
## 142 -0.31229631 -0.302945889
## 143 -0.36638746  0.265375764
## 144 -0.25929522 -0.420541563
## 145 -0.32501303  0.408843869
## 146 -0.08291128 -0.681790678
## 147 -0.16662757 -0.640208980
## 149 -0.17282833 -0.481238134
## 150 -0.19073323 -0.208688876
## 151 -0.25484083  0.429965333
## 152 -0.20033996 -0.511894850
## 153 -0.15951508 -0.395201217
## 154 -0.25175195 -0.025791944
## 155 -0.22511263 -0.429430084
## 156 -0.16450303 -0.582255320
## 157 -0.27834022  0.445337299
## 158 -0.25455574 -0.370897441
## 159  2.48521691 -2.064527684
## 160 -0.23561976 -0.297275626
## 161 -0.21567181 -0.540683816
## 162 -0.33045208 -0.228240407
## 164 -0.13501441 -0.031377269
## 165 -0.33920911  0.015122486
## 166  0.41537521  4.975969618
## 167 -0.40045472 -0.168316132
## 168 -0.22659736  1.836151198
## 169 -0.20574370 -0.464645980
## 170 -0.32479376 -0.232528615
```

```
## 171 -0.18895784 -0.328372383
## 172 -0.12274651 -0.504991440
## 173 -0.30052991 -0.390347390
## 174 -0.35169309 -0.205162977
## 175 -0.49343384 -1.016240880
## 176 -0.36315072 -0.254340409
## 177 -0.30781486 -0.170008038
## 179 -0.30221528 -0.329864983
## 180 -0.12807215 -0.673833321
## 181 -0.31330303 -0.160495014
## 182 -0.24880072 -0.379073923
## 183 -0.14523618 -0.532106099
## 184 -0.32784797 -0.334592602
## 186 -0.20774766 -0.462357686
## 187 -0.41128492  0.058602383
## 188 -0.20803430 -0.460230538
## 189 -0.29015141 -0.352316497
## 190 -0.37600276 -0.217393073
## 191 -0.26908975  0.665356010
## 192 -0.16295265 -0.652515409
## 193 -0.33734344  0.111323311
## 194 -0.26587394 -0.325490260
## 195 -0.36445784 -0.142721945
## 196 -0.22882865 -0.226185644
## 197 -0.36509563 -0.051881180
## 198 -0.20815402  0.634072288
## 200 -0.27545954 -0.350498051
## 201 -0.31511885 -0.354196541
## 202 -0.41245521  0.587217977
## 203 -0.28683558 -0.409156877
## 204  1.19393182 14.708585371
## 205 -0.26616101  0.092128282
## 206 -0.22924472 -0.426339314
## 207 -0.12470438 -0.112569283
## 208  0.13203847 -0.258684219
## 209 -0.22993704 -0.424178230
## 210 -0.20886158  0.040248876
## 211 -0.19550381 -0.496629865
## 212 -0.25198550 -0.427854603
## 213 -0.35772461 -0.123453149
## 214 -0.29945521 -0.256210636
## 215 -0.27574220  0.567465822
## 216 -0.16012646 -0.689620993
## 217 -0.25221172 -0.464840680
```

Now we can use reduced PCA dataset for outlier detection.

Outlier Detection

Apply KNN for Outlier Detection Cluster analysis we confirmed that k=3 perform better than k=2.Thus, here I have use k=3 for outlier detection as well.

```
#install.packages("dbscan")
library(dbscan)
```

```
## Warning: package 'dbscan' was built under R version 4.1.3
```

```
##
## Attaching package: 'dbscan'
```

```
## The following object is masked from 'package:stats':
##
##     as.dendrogram
```

```
# KNN parameter
k <- 3
```

```
# KNN distance (outlier score) computation
KNN_Outlier <- kNNdist(x=climateData_df[-1], k = k, all = TRUE)[,k]
```

```
 # KNN distance (outlier score) computation for PCA dataset
KNN_Outlier_pca <- kNNdist(x=pca_data_update, k = k, all = TRUE)[,k]
```

Apply knn for reduced PCA dataset The following code sorts the observations according to their KNN Outlier scores and displays the top 50 outliers along with their scores.

```
top_n <- 50 # No. of top outliers to be displayed for the original dataset
rank_KNN_Outlier <- order(x=KNN_Outlier, decreasing = TRUE) # Sorting (descending)
KNN_Result <- data.frame(ID = rank_KNN_Outlier, score = KNN_Outlier[rank_KNN_Outlier])
head(KNN_Result, top_n)
```

```
##      ID        score
## 103  99 1.177782e+10
## 79   76 6.104678e+07
## 159 151 5.229608e+07
## 161 153 2.072407e+07
## 37   36 1.093835e+07
## 204 192 4.545659e+06
## 90   86 2.246209e+06
## 166 157 1.398561e+06
## 99   95 6.785676e+05
## 208 196 6.436738e+05
## 27   26 3.633216e+05
## 113 109 3.280266e+05
## 51   49 2.973976e+05
## 106 102 2.875398e+05
## 92   88 2.841541e+05
## 88   84 2.776143e+05
## 215 203 2.672828e+05
## 2     1 2.580184e+05
## 10    9 2.532095e+05
## 43   42 2.446151e+05
## 82   78 2.446151e+05
## 164 155 2.398678e+05
## 168 159 2.317365e+05
## 172 163 2.017561e+05
## 33   32 1.957911e+05
## 11   10 1.830669e+05
## 126 119 1.703605e+05
## 198 187 1.433689e+05
## 69   67 1.392129e+05
```

```
## 157 149 1.310766e+05
## 93   89 1.102788e+05
## 151 143 1.061916e+05
## 6     5 9.665089e+04
## 100  96 9.348546e+04
## 210 198 9.348546e+04
## 65   63 9.346977e+04
## 202 190 9.145270e+04
## 7     6 9.004318e+04
## 96   92 8.836446e+04
## 58   56 8.026631e+04
## 154 146 7.994097e+04
## 139 132 7.992201e+04
## 56   54 7.928314e+04
## 143 136 7.885364e+04
## 205 193 7.885364e+04
## 39   38 7.645109e+04
## 145 138 7.496930e+04
## 60   58 6.770054e+04
## 191 180 6.628385e+04
## 207 195 6.510694e+04
```

We can see that the most outlying observation is observation 237, with outlier score = 4.42, followed by observation 208 with outlier score = 4.17, so on and so forth.

```
top_n <- 50 # No. of top outliers to be displayed for the PCA dataset
rank_KNN_Outlier <- order(x=KNN_Outlier_pca, decreasing = TRUE) # Sorting (descending)
KNN_Result <- data.frame(ID = rank_KNN_Outlier, score = KNN_Outlier_pca[rank_KNN_Outlier])
head(KNN_Result, top_n)
```

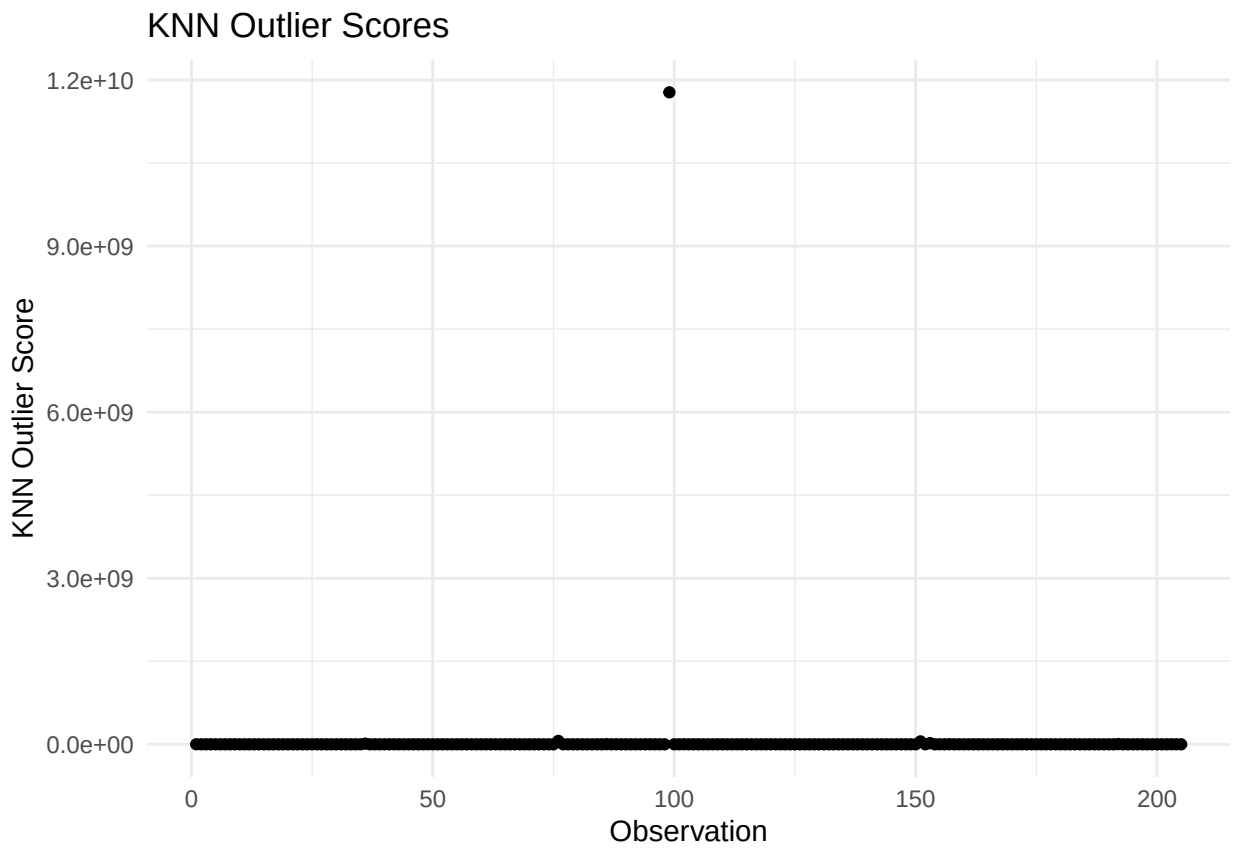
```
##      ID       score
## 79   76 40.67193428
## 204 192 10.45893265
## 37   36  9.85604822
## 166 157  2.99347460
## 159 151  2.86907978
## 90   86  2.30218534
## 103  99  1.26449521
## 99   95  0.90020896
## 168 159  0.53847939
## 92   88  0.52892694
## 175 166  0.44708330
## 51   49  0.37133821
## 208 196  0.30823094
## 27   26  0.29262262
## 126 119  0.25861919
## 69   67  0.25356472
## 88   84  0.25356472
## 33   32  0.22185719
## 65   63  0.21044287
## 106 102  0.20456633
## 202 190  0.16889969
## 18   17  0.15949370
## 143 136  0.15676643
## 113 109  0.15360810
```

```
## 58   56  0.14883164
## 11   10  0.13184055
## 6     5  0.12681920
## 93   89  0.12589275
## 207 195  0.11661383
## 96   92  0.11138923
## 164 155  0.10287715
## 191 180  0.09811597
## 215 203  0.09811597
## 139 132  0.09717910
## 56   54  0.09199680
## 210 198  0.09199680
## 150 142  0.08121848
## 198 187  0.07936053
## 154 146  0.07873831
## 87   83  0.07867565
## 123 116  0.07867565
## 107 103  0.07755641
## 205 193  0.07611361
## 145 138  0.07328201
## 60   58  0.07085994
## 100  96  0.06876816
## 151 143  0.06876816
## 86   82  0.06780902
## 187 176  0.06758155
## 10    9  0.06544800
```

We can see that the most outlying observation is observation 237, with outlier score = 4.42, followed by observation 208 with outlier score = 4.17, so on and so forth.

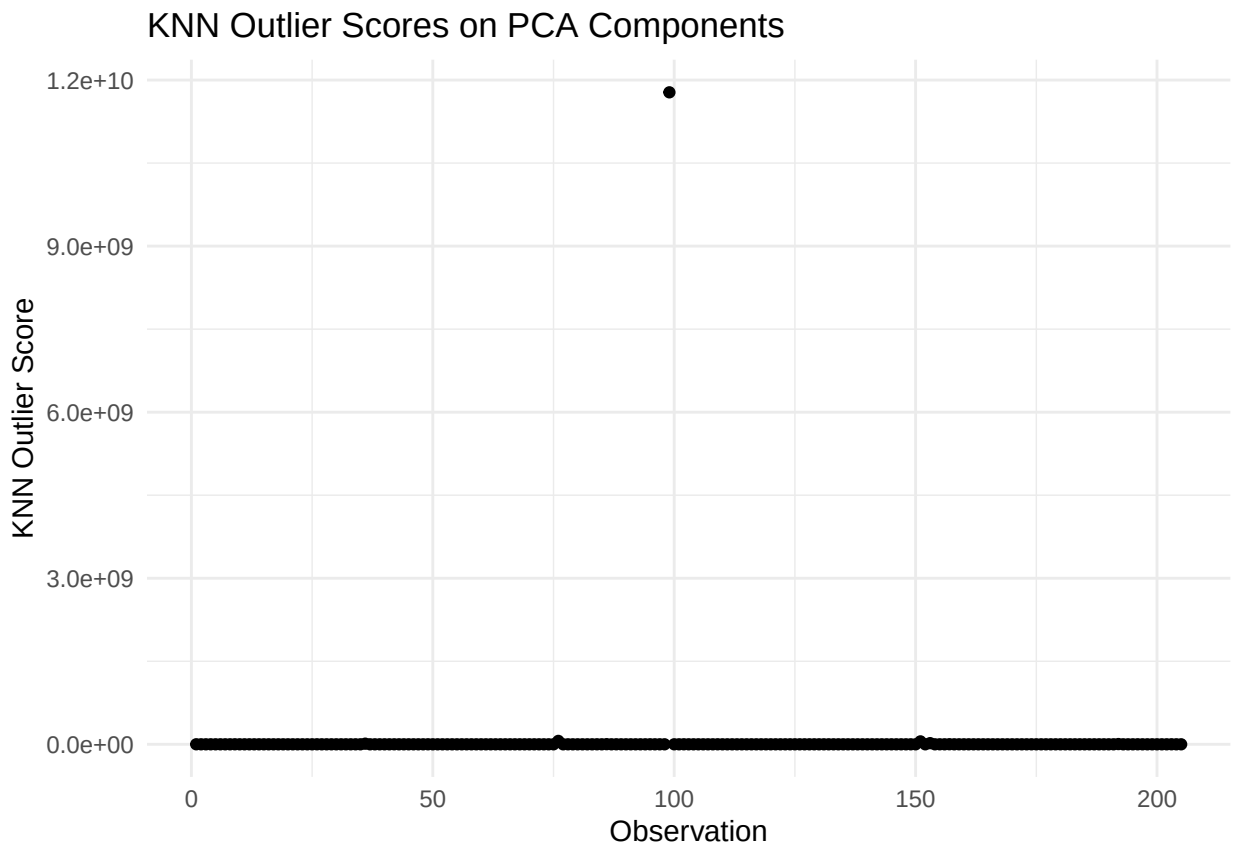
```
# Create a scatterplot of KNN outlier scores
ggplot(data = climateData_df, aes(x = 1:nrow(climateData_df), y = KNN_Outlier)) +
  geom_point() +
  labs(x = "Observation", y = "KNN Outlier Score", title = "KNN Outlier Scores") +
  theme_minimal()
```

Plot the KNN outlier scores for both original dataset and PCA dataset.




```
ggplot(data = pca_data_update, aes(x = 1:nrow(pca_data_update), y = KNN_Outlier)) +
  geom_point() +
  labs(x = "Observation", y = "KNN Outlier Score", title = "KNN Outlier Scores on PCA Components") +
  theme_minimal()
```

Visualization for KNN outlier score for PCA dataset



Conclusion

The study examines Australia's agricultural sustainability in relation to global climate change indicators. It found a strong correlation between agricultural land and hydropower generation, suggesting increased reliance on hydropower. Renewable energy consumption and hydropower electricity production are crucial for agricultural land expansion. However, Australia's high percentage of arable land and CO2 emissions highlight areas needing attention for sustainability. The findings suggest investing in sustainable practices and renewable energy to mitigate CO2 emissions, while further research and policies may be needed to balance environmental impact. The study has limitations as it requires more in-depth research for precise recommendations.

References

Agriculture and mining - Climate Adaptation. (2016, May 5). Climate Adaptation. https://research.csiro.au/climate/themes/agriculture/

Climate Change. (n.d.). World Bank. https://www.worldbank.org/en/topic/climatechange

Glossary | DataBank, https://databank.worldbank.org/metadataglossary/world-development-indicators/series/

Home - DAFF. (2023, October 27). https://www.agriculture.gov.au/

Log in to canvas. (n.d.). https://canvas.newcastle.edu.au/courses/24134/assignments/211595

Overview. (n.d.). World Bank. https://www.worldbank.org/en/topic/climatechange/overview#2

Principal Component Analysis (PCA) Explained Visually with Zero Math, https://towardsdatascience.com/principal-component-analysis-pca-explained-visually-with-zero-math-1cbf392b9e7d.

Who We Are. (n.d.). World Bank. https://www.worldbank.org/en/who-we-are