CSP-571 Data Preparation Analysis Project

Sustainable World Energy Insights and Visualization Using Power-BI

Team Members:

Soham Mankar – **A20543251**
Hari Shekar Reddy – A20551047
Shashank Kulkarni – **A20542907**

Let’s Walkthrough Outline

- Project Research Goal
- Project Planning and Timeline
- Data Gathering and Metrics
- Data Manipulation
- Exploratory Data Analysis
- Data processing pipeline

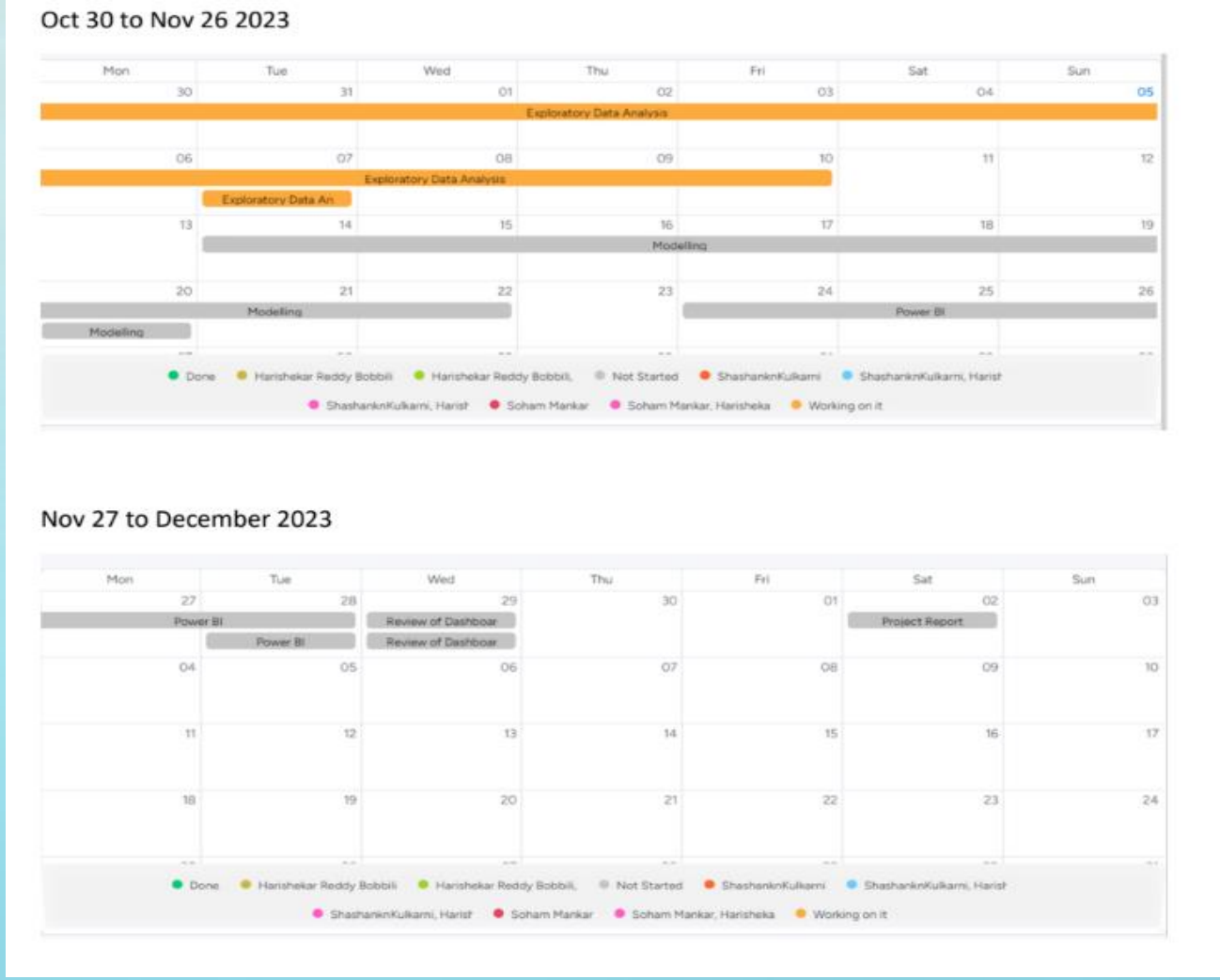
- Modeling:
  - ❑I. Primary energy consumption per capita:
    - ➢Linear Regression
    - ➢Random Forest
    - ➢Gradient Boosting
  - ❑II. Renewable energy share in the total final energy consumption (%):
    - ➢Linear Regression
  - ❑III. Co2 Emission:
    - ➢Linear Regression
    - ➢Random Forest
    - ➢Gradient Boosting

- Power BI Visualization
- Conclusion
- Future Works
- References

Project Research Goal

- The research goal of this project is to analyze the Global Data on Sustainable Energy from 2000 to 2020 to gain insights into the trends, progress, and challenges in the global sustainable energy landscape.
- The main aim is to conduct a comprehensive analysis of sustainable energy data to inform policy, investment decisions, and public awareness on this critical issue.

Project Planning & Timeline



Data Gathering and Metric

- Data is taken from Kaggle. Below is the link :
- https://www.kaggle.com/datasets/anshtanwar/global-data-on-sustainable-energy

- Data has 3649 rows and 21 Columns

Metric

- **Entity**: The name of the country or region for which the data is reported.
- **Year**: The year for which the data is reported, ranging from 2000 to 2020.
- **Access to electricity (% of population)**: The percentage of population with access to electricity.
- **Access to clean fuels for cooking (% of population)**: The percentage of the population with primary reliance on clean fuels.
- **Renewable-electricity-generating-capacity-per-capita**: Installed Renewable energy capacity per person
- **Financial flows to developing countries (US $)**: Aid and assistance from developed countries for clean energy projects.
- **Renewable energy share in total final energy consumption (%)**: Percentage of renewable energy in final energy consumption.
- **Electricity from fossil fuels (TWh)**: Electricity generated from fossil fuels (coal, oil, gas) in terawatt-hours.
- **Electricity from nuclear (TWh)**: Electricity generated from nuclear power in terawatt-hours.
- **Electricity from renewables (TWh)**: Electricity generated from renewable sources (hydro, solar, wind, etc.) in terawatt-hours.
- **Low-carbon electricity (% electricity)**: Percentage of electricity from low-carbon sources (nuclear and renewables).
- **Primary energy consumption per capita (kWh/person)**: Energy consumption per person in kilowatt-hours.
- **Energy intensity level of primary energy (MJ/$2011 PPP GDP)**: Energy use per unit of GDP at purchasing power parity.
- **Value_co2_emissions (metric tons per capita)**: Carbon dioxide emissions per person in metric tons.
- **Renewables (% equivalent primary energy)**: Equivalent primary energy that is derived from renewable sources.
- **GDP growth (annual %)**: Annual GDP growth rate based on constant local currency.
- **GDP per capita**: Gross domestic product per person.
- **Density (P/Km2)**: Population density in persons per square kilometer.
- **Land Area (Km2)**: Total land area in square kilometers.
- **Latitude**: Latitude of the country's centroid in decimal degrees.
- **Longitude**: Longitude of the country's centroid in decimal degrees.

```r
# recontruct the dataset by replacing the old column names  to a new column names

country <- energy$Entity
year <- energy$Year
E1 <- energy$Access.to.electricity....of.population.
E2 <- energy$Access.to.clean.fuels.for.cooking
E3 <- energy$Renewable.electricity.generating.capacity.per.capita
E4 <- energy$Financial.flows.to.developing.countries..US...
E5 <- energy$Renewable.energy.share.in.the.total.final.energy.consumption....
E6 <- energy$Electricity.from.fossil.fuels..TWh.
E7 <- energy$Electricity.from.nuclear..TWh.
E8 <- energy$Electricity.from.renewables..TWh.
E9 <- energy$Low.carbon.electricity....electricity.
E10 <- energy$Primary.energy.consumption.per.capita..kWh.person.
E11 <- energy$Energy.intensity.level.of.primary.energy..MJ..2017.PPP.GDP.
E12 <- energy$Value_co2_emissions_kt_by_country
E13 <- energy$Renewables....equivalent.primary.energy.
E14 <- energy$gdp_growth
E15 <- energy$gdp_per_capita
E16 <- energy$Density.n.P.Km2.
E17 <- energy$Land.Area.Km2.
E18 <- energy$Latitude
E19 <- energy$Longitude
```

Data Manipulation

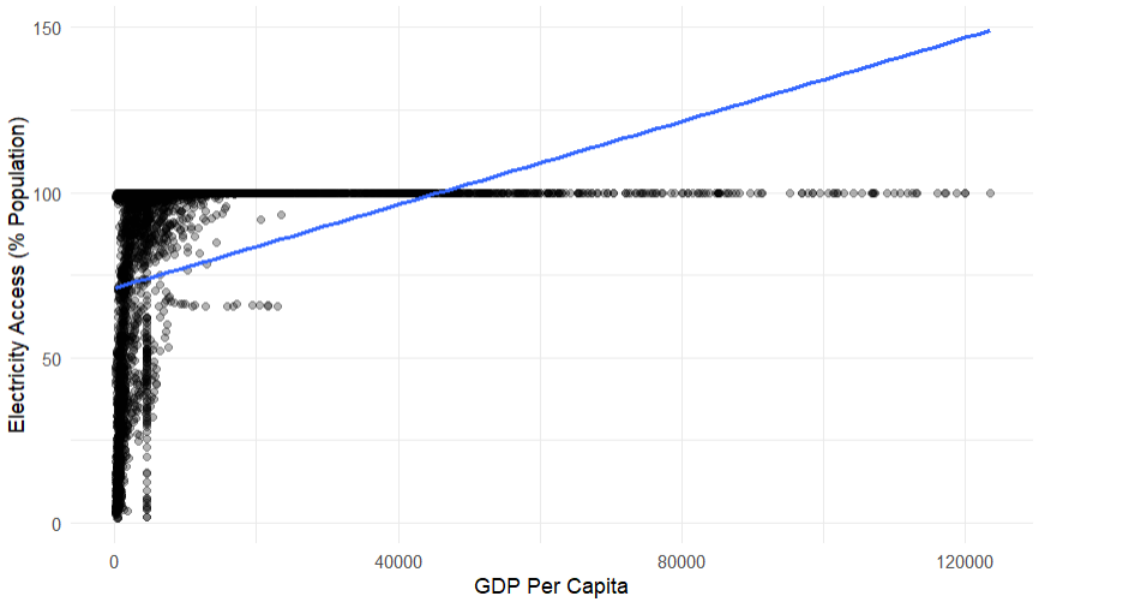
Missing values

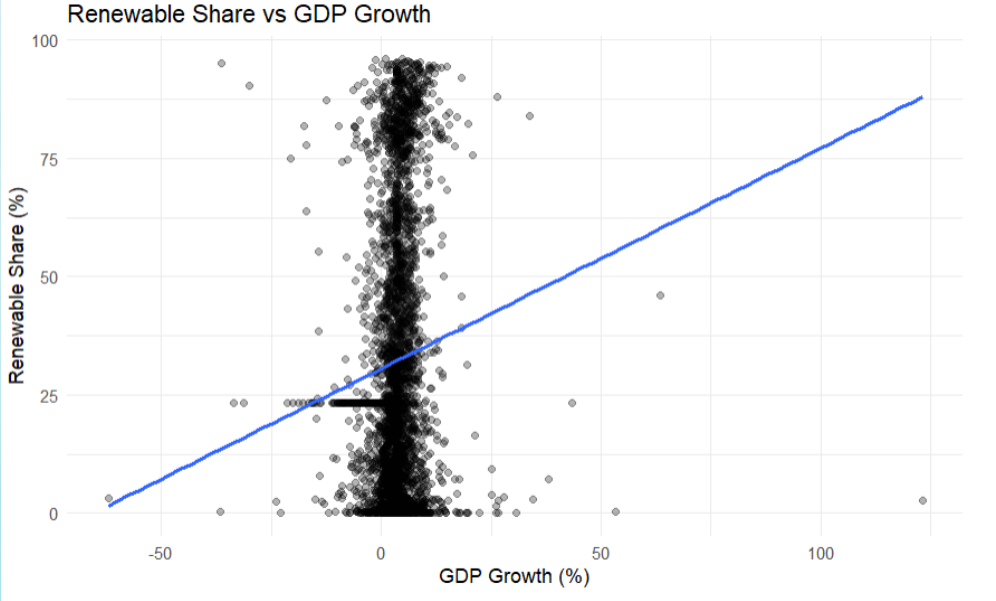
country	year	E1	E2	E3	E4
0	0	10	169	931	2089
E5	E6	E7	E8	E9	E10
194	21	126	21	42	0
E11	E12	E13	E14	E15	E16
207	428	2137	317	282	0
E17	E18	E19			
1	1	1			

Remove E17, E18 and E19 Missing values and remaining with stats and then check again full missing values then it comes out to be:

country	year	E1	E2	E3	E4
0	0	0	0	0	0
E5	E6	E7	E8	E9	E10
0	0	0	0	0	0
E11	E12	E13	E14	E15	E16
0	0	0	0	0	0
E17	E18	E19			
0	0	0			

Exploratory Data Analysis

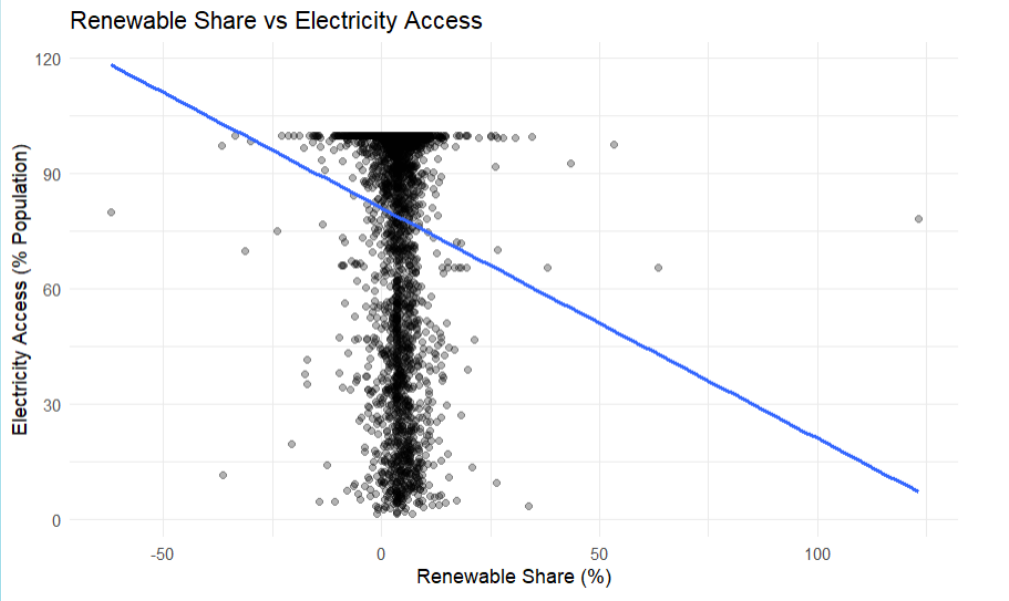


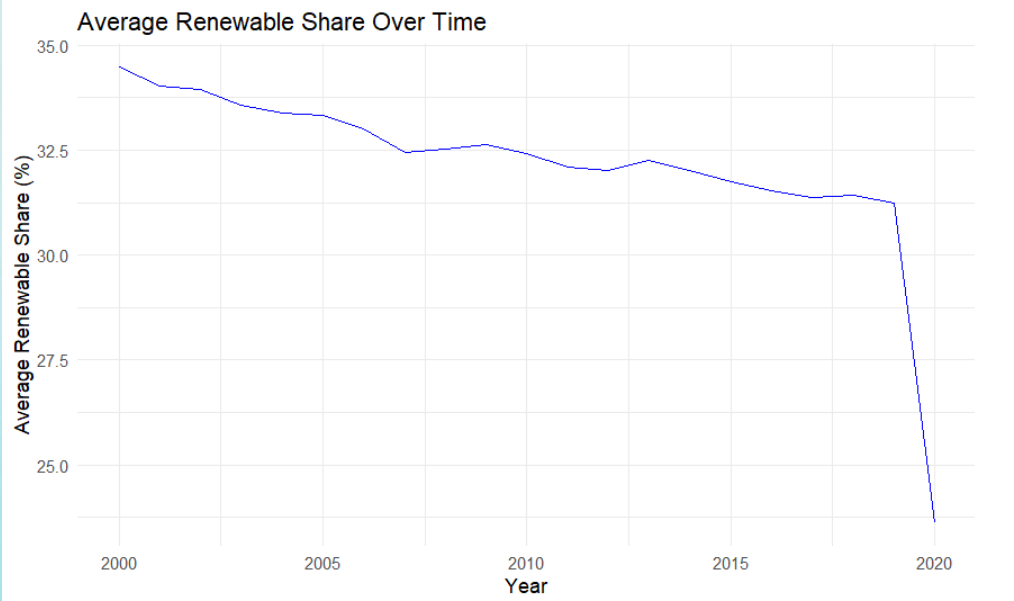


Positive correlation between GDP per capita and electricity access

Positive relationship between GDP growth and the share of renewable energy

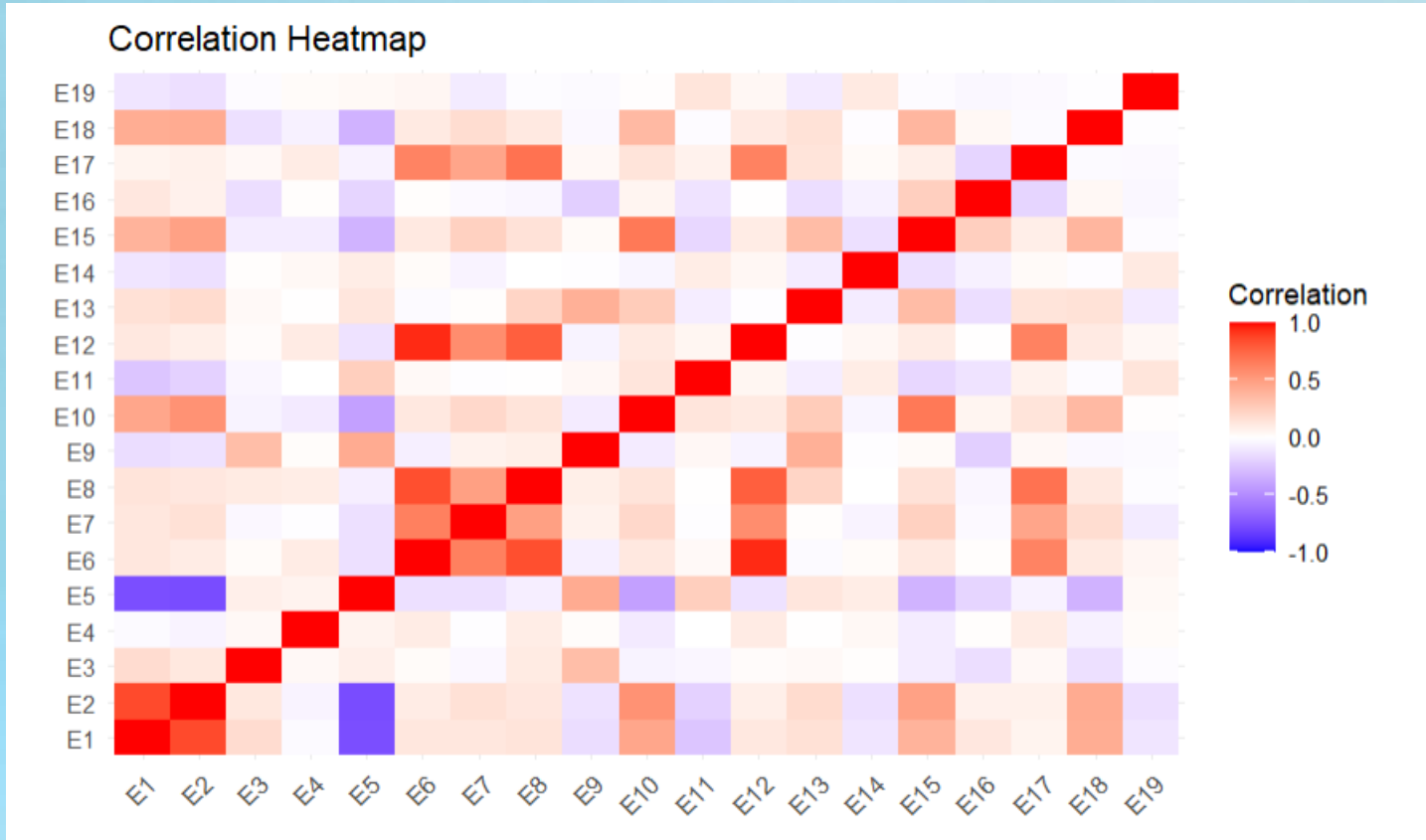
Exploratory Data Analysis





a Negative correlation between the renewable share of energy and electricity access percentage.

Overall decreasing trend in the average renewable share from the year 2000 until around 2015



The heatmap indicates a strong relationship between GDP per capita and energy access, a negative association between renewable energy share and CO2 emissions, and a positive correlation between latitude and access to electricity.

Additionally, there's a strong interconnection among various energy consumption metrics, suggesting that higher energy consumption is associated with increased energy intensity and reliance on fossil fuels.

Data processing pipeline

- First Select the features from the dataset for modelling and then split it into Train-Test Split
- We will split the data into 80% Train and 20% Test Split
- Now after these our Modeling part begins.....!

Modeling

- We are applying modeling on three different parts:

  - **Primary energy consumption per capita:**
    - **Linear Regression**
    - **Random Forest**
    - **Gradient Boosting**

  - **II. Renewable energy share in the total final energy consumption (%):**
    - **Linear Regression**

  - **III. Co2 Emission:**
    - **Linear Regression**
    - **Random Forest**
    - **Gradient Boosting**

- Now let us walk through each……..

I.Primary energy consumption per capita:

- **First, we have selected feature modeling is done...**

```r
# Correct target variable name
target <- 'E10'

# Correct features variable names based on the provided list
features <- c(
  'E1',
  'E15',
  'E4',
  "E5",
  'E6'
)
```

```r
# Train Gradient Boosting model with best parameters
set.seed(42) # For reproducibility when training
gradient_boosting_model <- gbm(
  formula = y_train ~ .,
  data = data.frame(y_train, x_train),
  distribution = "gaussian",
  n.trees = best_gb_params$n.trees,
  interaction.depth = best_gb_params$interaction.depth,
  n.minobsinnode = best_gb_params$n.minobsinnode,
  shrinkage = best_gb_params$shrinkage,
  verbose = FALSE # to avoid verbose output
)

# Train Linear Regression model
linear_regression_model <- lm(y_train ~ ., data = data.frame(y_train, x_train))

# Make predictions on the test set
rf_predictions <- predict(random_forest_model, newdata = x_test)
lr_predictions <- predict(linear_regression_model, newdata = x_test)
gb_predictions <- predict(gradient_boosting_model, newdata = x_test, n.trees = best_gb_params$n.trees)
```

I. Primary energy consumption per capita:

- Calculating MSE and R2 score for 3 model and compare it and pickout the best model

Model <chr>	MSE <dbl>	R_squared <dbl>
Random Forest	226876536	0.8217806
Linear Regression	677905233	0.4674820
Gradient Boosting	102878440	0.9191854

- And, we have predicted next 5 years of renewable energy share for 2021-2025

Description: df [10 × 2]

Year <int>	E10_Predictions <dbl>
2021	2076.144
2022	2078.811
2023	1451.279
2024	1562.081
2025	2017.392
2021	1966.803
2022	2078.811
2023	1966.803
2024	1969.084
2025	1966.803

1-10 of 10 rows

The above is next 5 year E10 Predictions using gradient Boosting

II. Renewable energy share in the total final energy consumption (%)

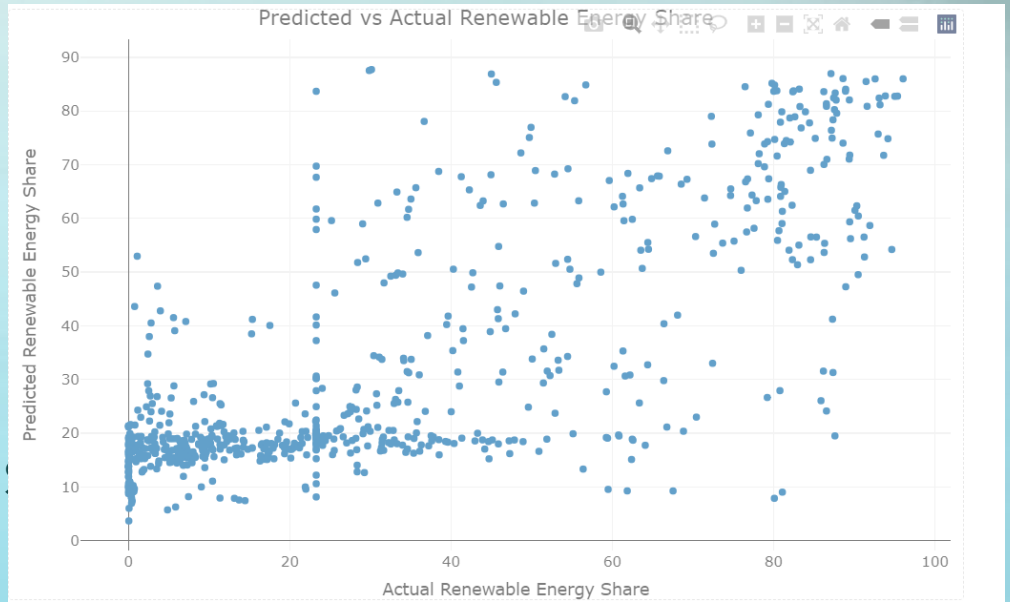
- **First, we have selected feature on which linear Regression model is done…**

```
# Select features
selected_features <- c('year', 'E15', 'E10',
                       'E1', 'E12')
```

- model<- lm(y_train ~ ., data = X_train)

- Then we will calculate mse value

- Plot a scatter plot between Predicted Renewable Energy Share



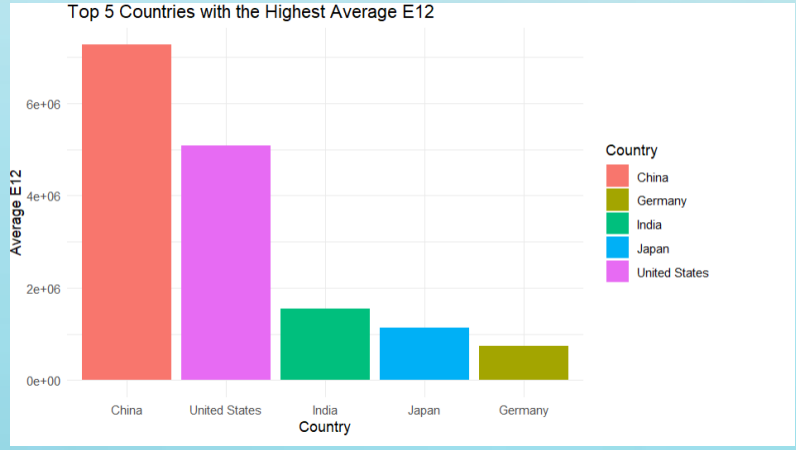
- And, we have predicted next 5 years of renewable energy share for 2021

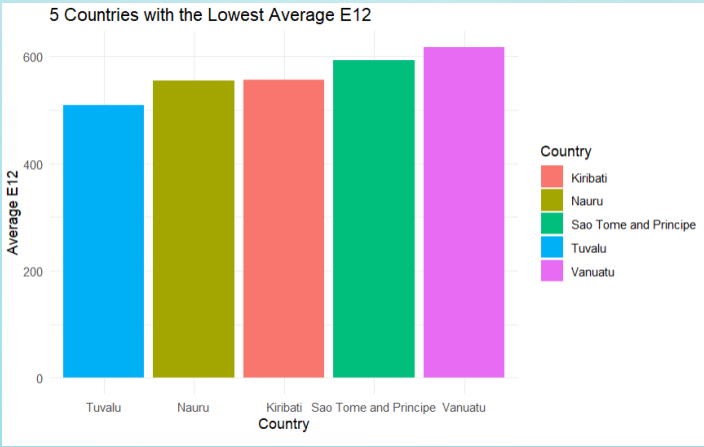
-2025

```
       1        2        3        4        5
54.25737 54.43583 54.61428 54.79274 54.97120
```

III. Co2 Emission:

- **First, we have selected feature on 3 models modelling is done…**
- **We have done a few plots for co2 emission:**





- Then calculated the Growth Rate and Average Growth Rate of Co2 Emission

- After then now modeling is done:

```r
# Load required libraries
library(caret)
library(ranger) # ranger is used instead of randomForest
library(xgboost)

# Assuming you have a data frame named 'global_co2' with columns 'country' and 'E12'

# Convert 'country' to a numeric variable
global_co2$country_code <- as.numeric(as.factor(global_co2$country))

# Prepare the data
X <- global_co2[, !(colnames(global_co2) %in% c("E12", "country"))] # Remove 'country' and use 'country_code' instead
y <- global_co2$E12

# Split the data into training and testing sets
set.seed(42)  # Set random seed for reproducibility
train_indices <- createDataPartition(y, p = 0.8, list = FALSE)
X_train <- X[train_indices, ]
y_train <- y[train_indices]
X_test <- X[-train_indices, ]
y_test <- y[-train_indices]

# Create a list of regression models
set.seed(42) # Set the seed before training the models
models <- list(
  'Linear Regression' = lm(E12 ~ ., data = data.frame(E12 = y_train, X_train)),
  'Random Forest' = ranger(E12 ~ ., data = data.frame(E12 = y_train, X_train), num.trees = 500),
  'Gradient Boosting' = train(E12 ~ ., data = data.frame(E12 = y_train, X_train), method = 'gbm', verbose = FALSE)
)

best_model <- NULL
best_r2 <- -Inf # Start with negative infinity for comparison
```

- Finding of R-square and RMSE values to find out the best model for fit..

```
    precision, recall

Linear Regression :
R2 Score: 0.97
Mean Absolute Error (MAE): 53,083.37
Mean Squared Error (MSE): 1.756714e+10
Root Mean Squared Error (RMSE): 132,541.1
-----------------------------------------------
Random Forest :
R2 Score: 0.99
Mean Absolute Error (MAE): 12,963.49
Mean Squared Error (MSE): 6.818058e+09
Root Mean Squared Error (RMSE): 82,571.53
-----------------------------------------------
Gradient Boosting :
R2 Score: 0.99
Mean Absolute Error (MAE): 33,424.19
Mean Squared Error (MSE): 1.015382e+10
Root Mean Squared Error (RMSE): 100,766.2
-----------------------------------------------
The best performing model is: Random Forest with R2 score: 0.99
```

The best model performance is Random Forest with R2 Score: 0.99 which approximately equal to 1.

Global Sustainable Energy Trend Dashboard (2000-2020)

Global Sustainable Energy Trend Visualization (2000-2020)

year

2000 2020

Continent

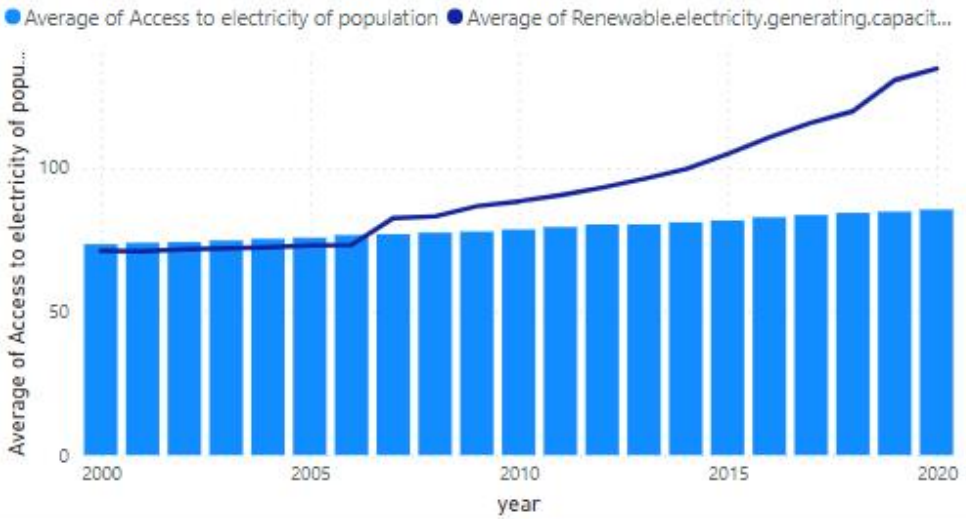
Africa | Americas | Asia | Europe | Oceania

Average of Access to electricity of population by Country

78.98	142.38K	32.14
Average of Access to electric...	Average of Value_co2_emiss...	Average of Renewable....

Average of Access to electricity of population by Continent

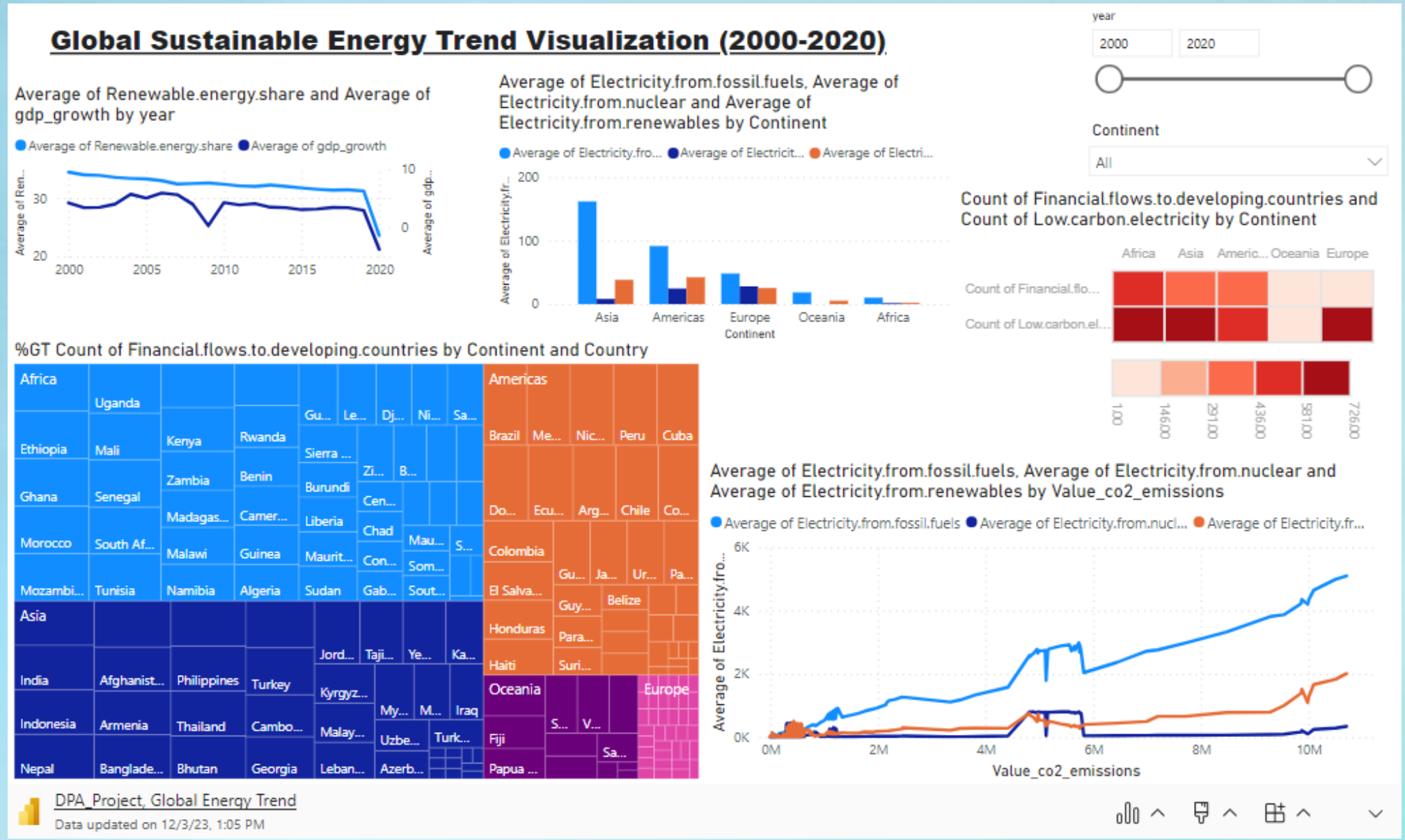
Average of Access to electricity of population and Average of Renewable.electricity.generating.capacity.per.capita by year



DPA_Project, Access to Electricity Measures

Data updated on 12/3/23, 1:05 PM

Global Sustainable Energy Trend Dashboard (2000-2020)



Global Sustainable Energy Trend Dashboard (2000-2020)

Global Sustainable Energy Trend Visualization (2000-2020)

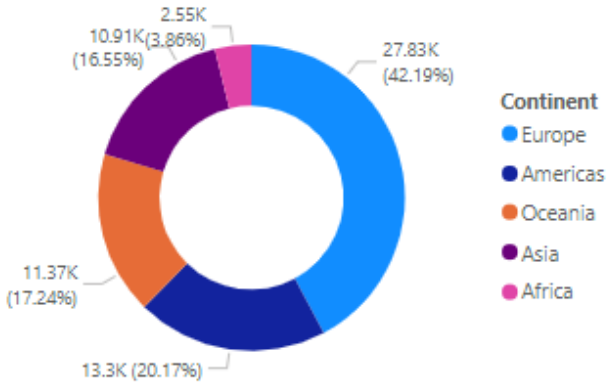
year

2000 2020

Continent

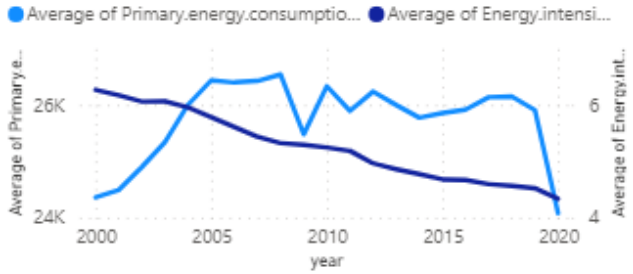
All

Average of gdp_per_capita by Continent

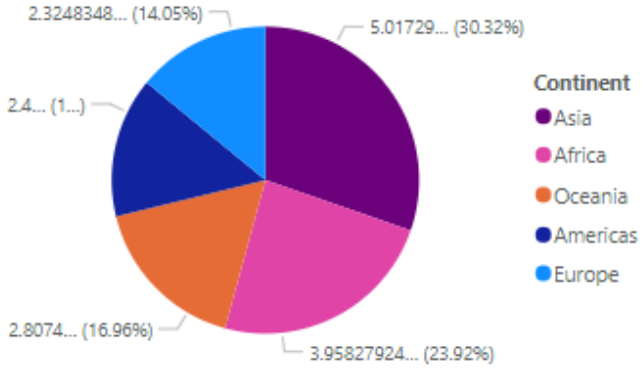


Country	Average of gdp_growth	Sum of gdp_per_capita
Myanmar	9.53	15,373.27
Ethiopia	8.80	8,840.91
China	8.70	106,339.07
Equatorial Guinea	8.26	252,090.38
Qatar	8.20	1,242,521.09
Azerbaijan	8.15	86,070.18
Turkmenistan	7.86	90,101.94
Tajikistan	7.54	13,672.63
Cambodia	7.22	18,102.90
Rwanda	7.22	11,656.07
Mongolia	6.38	53,951.54
Bhutan	6.28	43,344.88
Uzbekistan	6.25	30,203.26
Total	**3.45**	**46,013,064.16**

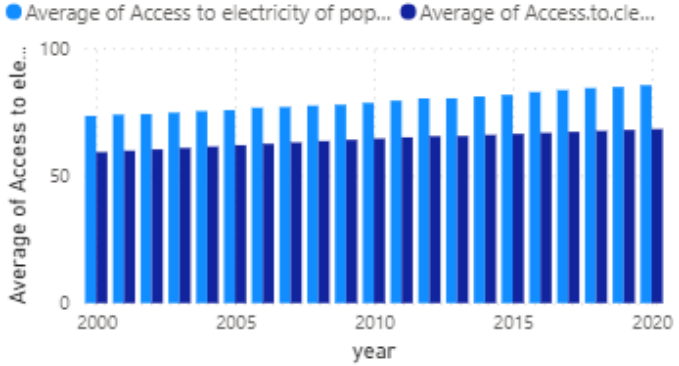
Average of Primary.energy.consumption.per.capita and Average of Energy.intensity.level by year



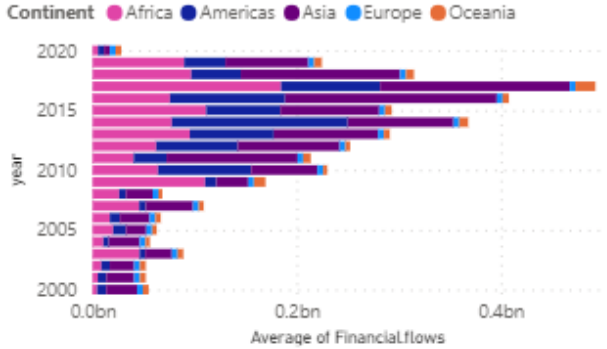
Average of gdp_growth by Continent



Average of Access to electricity of population and Average of Access.to.clean.fuels.for.cooking by year



Average of Financial.flows by year and Continent



DPA_Project, Global Energy Trend 2

Data updated on 12/3/23, 1:05 PM

Conclusion and Future Scope

- This project demonstrates the harmonious integration of data analysis, visualization, and machine learning to untangle the complexities surrounding global energy consumption.
- By combining insights derived from historical data with predictive capabilities, it strives to contribute to a sustainable energy future and a world that is better informed.
- Key findings showcase progress in improving access to electricity and cleaner cooking fuels, alongside the ongoing challenge of escalating CO2 emissions.
- The predictive models developed in this project serve as valuable tools for decision-makers, facilitating resource allocation and targeted energy initiatives.
- Opportunities for future refinement exist through the incorporation of additional variables, the application of advanced modeling techniques, and more comprehensive regional analyses.
- In essence, this project underscores the critical role of data-driven approaches in addressing global energy challenges and advancing sustainability efforts.

Future Scope

1. Data Expansion: Suggest expanding the dataset beyond 2020 to include the most recent trends and developments in sustainable energy. This could include the impact of recent global events such as the COVID-19 pandemic on energy usage patterns.

2. Advanced Analytical Techniques: Propose the use of more advanced data analysis techniques, like machine learning and predictive modeling, to forecast future energy trends and model the impact of different energy policies.

3. Interactive Dashboard Enhancements: For the dashboard, consider incorporating more interactive elements or real-time data updates. This would make it a more dynamic tool for tracking ongoing changes in global energy trends.

4. Collaborative Research Opportunities: Identify opportunities for collaboration with other researchers or institutions. This could lead to more comprehensive studies or the development of a global sustainable energy database.

5. Policy Impact Assessment: Suggest conducting an analysis of how different countries' energy policies have influenced sustainable energy trends and what lessons can be learned from these policies.

6. Public Engagement and Education: Highlight the importance of using the project's findings to educate the public and engage them in discussions about sustainable energy.

References

1. Towards Sustainable Energy: A Systematic Review of Renewable Energy Sources, Technologies, and Public Opinions Atika Qazi, Fayaz Hussian, Nasrudin Abd, KhaledShaban, and Khalid, Volume 7, 2019 IEEE.

2. Towards data-driven energy communities: A review of open-source datasets, modelsand tools, Hussain Kazmi, Ingrid Munné-Collado, Fahad Mehmood, Tahir AbbasSyed, Johan Driesen, Published by Elsevier Ltd.

3. Emerging renewable and sustainable energy technologies: State of the art AkhtarHussaina, Syed Muhammad Arifb, Muhammad Aslam, 2016 Published by ElsevierLtd.

4. An Introduction to Power BI for Data Analysis: Dr. Kalpana V. Metre, Dr. AshutoshMathur, Dr. Ranjana Prakash Dahake, Dr. Yogita Bhapkar, Mrs.Jayashri Ghadge,Prashant Jain, Santosh Gore, Published by IJISAE

5. "R for Data Science" by Hadley Wickham and Garrett Grolemund: This book provides a comprehensive introduction to data manipulation, visualization, and analysis using R.

6. "The Art of R Programming" by Norman Matloff: This book is a practical guide to R-programming and covers various aspects of R, from data structures to debugging.

7. "The Art of Data Science" by Roger D. Peng: Provides insights into the data analysis process and the art of interpreting results.

THANK YOU !!!