PROGRESS REPORT ON DATA ANALYSIS PROJECT

**Student name: Shweta Dutta**

**Registration No: 18MSMS19**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sr. No.** | **Step** | **Due date** | **Progress Report** |
| 1 | Collect multiple datasets | 18/10/19 | Datasets collected:   |  |  | | --- | --- | | **Sl. No.** | **Dataset Name / Source Link / Brief Description** | | 1 | **Water Quality (India)**  <https://data.gov.in/resources/water-quality-creek-canal-drains-2014>26 Data on characteristics of water in creek, canal and drains in few states in India for 2014. | | 2 | **Physical Performance of SRTU (India)**  <https://data.gov.in/resources/physical-performance-srtus-glance-2011-12-2015-16>  Data on parameters of ‘State Road Transport Undertaking’ for all states of India over 2011-2016 | | 3 | **Fuel Economy Data (United States)**  <https://www.epa.gov/compliance-and-fuel-economy-data/data-cars-used-testing-fuel-economy>  Data on cars used for testing fuel economy | | 4 | **Emission of GHG by all sectors (World)**  <http://www.fao.org/faostat/en/#data/EM>  Amount of emission of greenhouse gases from various sources over 1990-2010 | | 5 | **Crops Production (World)**  <http://www.fao.org/faostat/en/#data/QC>  Data on production of crops from 1961-2017 | | 6 | **Oil Consumption (World)**  <https://data.gov.in/resources/country-wise-world-oil-consumption-2011-2017>  Oil consumption data based on countries from 2011-2017 | |
| 2 | Study the background of the data | 22/10/19 | |  |  | | --- | --- | | **Dataset** | **Brief Background/ Description** | | **Water Quality** | * Data onwater quality parameters from stations located on all important rivers, lakes, wells, from **13** states of India, in the year 2014. * Purpose: Groundwater assessment for National water Monitoring Program (NWMP). * Total number of observations: **91** * Variables:**7 Numerical** variables (Temperature, Dissolved Oxygen, pH Conductivity, Biochemical oxygen Demand (B.O.D), Nitrate-N+Nitrite-N, Fecal Coliform, Total Coliform) * Columns contain minimum, maximum and mean values of the 7 variables. * Missing Values: Yes | | **Physical Performance of SRTU** | * Data on physical parameters of **47** SRTUs across states in India from 2011-12 to 2015-16. * Purpose: Fleet utilization, Staff productivity etc. * Total number of observations: **47** * Variables: **7 Numerical** variables (Average Fleet Held, Average Age of Fleet, Revenue Earning km, Staff Strength, Fuel Efficiency, Passenger kms performed, Passenger carried in Lakhs.) * Columns contain data on 7 variables for 5 years. * Missing Values: Yes | | **Fuel Economy Data** | * Test data to determine fuel economy estimates derived from vehicle testing done at EPA’s National Vehicle and Fuel Emissions Laboratory, for the year 2019. * Purpose: Department of Energy, Transport uses this data to administer their fuel economy related programs. * Total number of observations: **4,610** * Variables: **65 (Mixed)** variables (Test Vehicle ID, Rated Horsepower, Equivalent Test weights, Axle Ratio, etc.) of data on**25** vehicle manufacturers. * Missing Values: Yes | | **Emission of GHG by all sectors** | * FAOSTAT Agri-Environmental Indicators section provides data on emissions of greenhouse gases (GHG) by gas, economic sector, country and year, from 1990-2010. * Purpose: Provides global database in exploring amount of GHG emissions by countries, worldwide. * Number of observations: **40,586** * Data on **14 factors of elements** (shares of each sector in total emissions of each gas, shares of each gas in emissions from each sector viz., emissions of CO2eq, emissions of CO2eq from CH4, etc.) from **10 items/sources** (energy, transport etc.) for **231 countries**. * Missing Values: Yes | | **Crop Production** | * FAOSTAT provides data on crop statistics for products covering categories like primary crops, primary fiber crops, cereals, fruits etc. from 1961-2017. * Purpose: To cover production of all primary crops for all countries in the world. * Number of observations: **49,370** * Data on **3** elements (Area harvested, yield, production quantity) for **180** items (apples, cereals, barley etc.), for **258** countries. * Missing Values: Yes | | **Oil Consumption** | * Data on consumption of non-petroleum additive, substitute fuels etc. * Purpose: Gives an idea of global petroleum scenario. * Total number of observations: 96 on **96** countries across the world. * Columns contain consumption data for **7** years from 2011-2017 and % share in 2017. * Missing Values: Yes, only for % share in 2017. | |
| 3 | Identify the data to be used | 22/10/19 | **Emission of GHG by all sectors, worldwide.** |
| 4 | Examining the data  Deal with missing observations  Understanding Data Characteristics: Number of observations, number of variables  Variable type: numerical/categorical | 24/10/19 | * Data contains **40,586** observations of **49** variables. * Three factor variables:   + **Area** with **268** levels (231 countries and other cumulative levels)   + **Item** with **14** levels   + **Element** with **14** levels * **Area code**, **Item Code** and **Element Code** columns contain the unique codes of the levels of area, item and element, respectively. * Elements are expressed in Gigagram and % units, denoted in **Unit** columns. * This is a **time series dataset** containing **numerical values** (in Gigagram or %) for each level of elements, under each items, for each country, from the year **1990-2010**. * Contains multiple missing values for each year, viz., 4,904 rows have missing values for the year 1990, whereas 1,191 rows have missing values for 2010. |
| 5 | Purpose of the Analysis: Modeling, Predicting | 25/10/19 | The purposes of the analysis are to:   * Identify the main source of greenhouse gases for different continents of the world. * Find which greenhouse gas is most prevalent. * Find how the emission values have changed over years * Forecast the amount of emission in the upcoming years   The above analysis will help the developing countries to take proper steps to reduce the emission of greenhouse gases and plan the policies of GHG emissions accordingly, thus improving the environment. |
| 6 | Data exploration using graphical and numerical methods | 25/10/19 | * Considering the level ‘**Emissions**’ of ‘**Elements**’, since it is the total of the other levels: ‘Emissions from CO2, CH4, etc.’, and applying descriptive statistics on the emission values (ignoring the missing values) under following categories:  1. Each levels of ‘Item (Energy, Transport etc.)’ (for all Area, Year) 2. Each ‘Area (country)’ levels (for all Item, Year) 3. Each year from 1990-2010 (for all Area, Item) 4. Each ‘Item’ under each ‘Year’ (for all Area) 5. Each ‘Item’ under each ‘Area’ (for all Year) 6. Each ‘Area’ under each ‘Year’ (for all Item) 7. **Graphical Methods:**   Following graphical methods are used   1. **Multiple** and **Sub-Divided bar** diagram to note the emission values of each Item, Country, Year. 2. **Histograms** and describing the shape of the distribution as symmetric, left/right skewed, uni/bi/multi-modal. 3. Interpreting through **box plot**. 4. **Numerical Methods:**   Obtaining summary statistics:   1. Measures of central tendency (**mean**, **quantiles**) 2. Measures of spread (**standard deviation**) 3. **Outliers**  * It was observed that maximum average emissions are  1. By ‘Energy’ (considering all the Years and Areas). 2. In the year 2010 (considering all Items and Areas). 3. By USA, followed by China (considering all Items and Years).  * Similarly, we can obtain the results, for other levels of ‘Elements’ for a particular ‘Item’ or ‘Area’. |
| 7 | Basic analysis | 30/10/19 | 1. **Reshaped Data:** 2. **7** levels of Area: Africa, Asia, Europe, Northern America, Central America, Southern America and Oceania. 3. **15** Variables: Energy, Transport, Residential, Industrial, Waste, Other, Agricultural, Land, Forest, Energy Total, Land use total, Emission of CO2, CH4, NO2, F gases from all resources. 4. **Principal Component Analysis:** The purpose of PCA is to find the best low dimensional representation of variation in multivariate data. In this data set, the amounts of emission values are recorded from 9 main resources from 7 levels. By carrying out a principal component analysis, we obtained 3 principal components which capture most of the variation of emissions from the resources. Each of these **3** principal components is a particular linear combination of the 9 resources. 5. **Linear Discriminant Analysis:** The purpose of LDA is to find the linear combinations of the original variables (the 9 resources) that gives the best possible separation between the groups in the dataset. At most min(7,9)= 7 useful discriminant functions can be found to separate the emission values by cultivar, using the 9 independent variables. 6. **Multiple Regression:** For each level of Area, a model is fitted by regressing the 9 variables on the dependent variable “Emission share of CO2 from all resources” and thus can predict the value of emission of CO2 based on the value of other variables. Similar results can be obtained for “Emission share of other gases”. |
| 8 | Advanced analysis | 01/11/19 | **Time Series Analysis and Forecasting:** It is observed that for each levels, autocorrelation of the variables are large positive values (~>0.9). Thus considering that this is a multivariate time series data, we can forecast how much GHG will be emitted by each of the resources in the future. |
| 9 | Report Writing | 05/11/19 | - |
| 10 | Presentation | 08/11/19 | - |