BIRD EXTINCTION – POSSIBLE FACTORS

Introduction

Currently there are around 10000 species of birds to date in the world, however an alarmingly increasing number of bird species have started "disappearing" [1]. The total bird count per species worldwide has also reduced dramatically. The current bird extinction rate is 1.3% i.e. 154 species of birds that were present since we had started categorizing are now extinct [1]. As per a research conducted by Cornell Lab, the total number of birds spotted have has gone down by 3 billion in Canada and USA alone [2]. As per another study by Berkeley University, if all the critically endangered, endangered and vulnerable bird species disappears, then almost $1/5^{th}$ ~ 2220 of the total bird species will be lost. Hence the aim is to build a dataset that would help analyse – identify the causes or eliminate the causes that is likely to have resulted in the drastic reduction of the count of individual birds as well as the extinction of the bird species as a whole.

One of the most important factors that could have resulted in the loss of birds is climate change; with the global temperature increase at $1.9^{\circ}F$, the arctic ice has decreased at 12.8% per decade resulting in an increase in the sea surface level of 3.3mm/year [3]. These variations might have resulted in variations in the natural habitat of the birds, which could have costed the bird count in the affected regions. CO_2 is the major green house that has resulted in global warming, with a levels globally increased form 2 billion tonnes to 36 billion tonnes [3]. This in turn cycles back as an influencing factor to climate change and therefore is one other important factor that could have affected the bird count. Another factor that need to considered is the people population growth/density over the years. The current population growth rate is 1.1% annually with over 68% of the population living in urban regions [4]. Urbanization has resulted in increasing population density and the cities have been expanding over the years to accommodate the people, thereby destroying the natural habitat of the bird species in the process.

Objective

Based on the above conclusion, the dataset gathered should be representative of the birds, their various species and their counts spotted across countries over a period of time. The supporting datasets that could help identify/eliminate the suspicious factors for bird extinction/bird count depletion need to be collected for all countries for the same period of time.

A. Data Sources Used

The project majorly requires the following dataset:

- Bird Count by Species over the Years for different Countries
- Climate Change and Pollution related Dataset such as:
    - World Temperature
    - CO_2 Emission
- Human Population Density

Bird Dataset

The project is set out to identify the species count of various birds spotted over a period of time in different countries. The dataset gathered should have adequate information on the various species of birds including the native and migratory species, the date spotted together with the count. Also, since the loss of bird count, or the total loss of species is the research question here, it would help if the dataset has the last spotted bird information for all the world countries. To identify the trend worldwide or per country regarding the bird species counts, and to understand if it is actually stable or declining or increasing, a single reliable database that keeps updating the information about the birds spotted is required. The Cornell Lab is a significant contributor in the field of Ornithology and hence dataset gathered from their organization is a rightful choice.

Population Dataset

The population dataset is required to understand the trend of population growth over the years in different countries. This in conjunction with the bird dataset and other parameters considered might help to understand the declining reason for the birds in various countries.

Pollution Dataset

Climate change is primarily an issue due to air pollution. Pollution is measured by looking at the CO_2 emission rate of countries all around the globe. A long and historical estimate of annual carbon dioxide emissions was needed. This data set for Carbon-di-oxide emissions are obtained from Carbon Dioxide Information and Analysis Centre (CDIAC; Marland et al,2004).

Temperature Dataset

For the temperature dataset, two types of data are collected, one with the temperature anomalies between various years, and another dataset with the average surface temperature for every country. The datasets encompass data for several years from 1880 to 2015.

B. Choice of the Data Source

eBird Dataset from Cornell Lab

It is important that dataset gathered should be from a valid and reliable source. The Cornell Lab of Ornithology is one of the pioneers in this field with a mission to study, research, understand and conserve the diversity of bird species. eBird is an endeavour by Cornell Lab which involves avid bird watchers, the scientific community and the general public to come together from various countries to and to contribute in building a real-time bird sighting dataset that is freely accessible for the scientific projects. As per their website, eBirds is the world's largest database with over 100 million birds sighted and reviewed by regional experts all over the year, across geographical boundaries. The dataset is legally accessible by public for the purpose of study and research. Also, since the eBird web site has information on the same species of birds spotted across different countries, over many years and information on the first, last and highest count of bird species spotted – it is an apt choice to scrape data required for this project. The link is as follows: https://ebird.org/explore

Population Dataset from World Bank

The dataset is read as csv file from the popular and verified world bank data source. This website hosts verified and genuine data as the total population data is collected from government sources such as

1) UN Population division
2) census reports and other statistical publications from national statistical offices.
3) Eurostat: Demographic Statistics.
4) United Nations Statistical Division. Population and Vital Statistics Report (various years)
5) U.S. Census Bureau: International Database
6) Secretariat of the Pacific Community: Statistics and Demography Programme.

This dataset is also licensed by CC BY-4.0 (Creative Commons Attribution 4.0). The link for the dataset is as follows: https://data.worldbank.org/indicator/sp.pop.totl

Carbon Dataset from CDIAC

To a large extent the air pollution caused is a result of burning fossil fuels. Carbon-di-oxide is widely considered to be a good indicator of how much fossil fuel is burned and the extent of other pollutants emitted as a result. The levels of CO_2 in the atmosphere are now considerably higher than the past few decades. If the fossil fuel burning continues at this rate, then the earth would not be able to return to the pre-industry levels in the near future. The data set is used to know the countries which contribute to emitting Carbon-di-oxide in a high scale. http://cdiac.ornl.gov/ftp/ndp030/CSV-FILES/nation.1751_2014.csv

Temperature Dataset

There are two datasets for temperature used, which are as follows.

- Dataset from NOAA (National Oceanic and Atmospheric Administration) which has the merged land and oceanic surface temperature anomalies, which contains the monthly temperature data from the year 1990 till present, with a $5^{\circ} \times 5^{\circ}$ latitude-longitude spatial resolution.

  ftp://ftp.ncdc.noaa.gov/pub/data/noaaglobaltemp/operational/timeseries/

  Note : This dataset includes the temperature anomalies, which is the temperature difference from year to year.

- Global average land temperature by Country from Kaggle, which contains the average land temperature for each country from 1753 to 2013. The raw data is taken from Berkeley Earth.

  https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data

C. Intended Use of the Dataset

Using the final dataset wrangled possible research ideas include the following.

1. Does the extinction rate or loss of bird count depend on population rate, pollution rate, global warming?
   As seen by the statistics earlier, there has been a global rise in the mentioned factors. The data set would help identify if there is a dependency globally due to one or more factors. Also, this can be studied locally in a particular country, or a region. Based on the dataset put together this impact can also be studied per bird species; certain species might have fallen prey to global warming and certain others count might have gone down due to human population encroachment in their natural habitat. It is also possible that some local species native to a region might thrive as a result of human activities in that region too. Hence this dataset provides a chance analyse each bird species per country/region over a period of time.

2. Establish migratory bird patterns over the year across country
   The summer migratory birds travel from southern regions to equatorial/northern countries to escape the southern winter, while the winter migratory birds travel from northern regions to the equatorial/southern countries to escape the cold winter. Hence the same species of birds will be spotted in the same year in different countries of the world. Hence isolating the species and their count per year will provide information on the migratory patterns.

3. Variation in migratory patterns over the years due to above factors
   Once the migratory pattern is established as above per migratory bird species, possible variations in that pattern can be identified over the years. This can be studied in light of population change, pollution rate and global warming. The most likely or the least likely cause of this variation in migratory patterns if present can be deduced. The migratory pattern can be a change in the month in which the birds migrate or change in the regions (latitude/longitude) in which the birds migrate. This can be understood using the dataset put together.

4. Regional bird count variations in case of indigenous species

   The regional bird count variation of native birds in a country is an important parameter that will affect the ecological cycle and the food chain of the region. This in effect will have snowball effect on the regional ecology. Any significant increase or decrease of the native bird population over the years can be monitored and studied using this dataset.

D. Techniques Used

Web Scrapping Bird, Country and Regional Codes

The eBird web page has data but that needs to web scrapped. Web scrapping was carried out mostly using 'rvest' package in R.

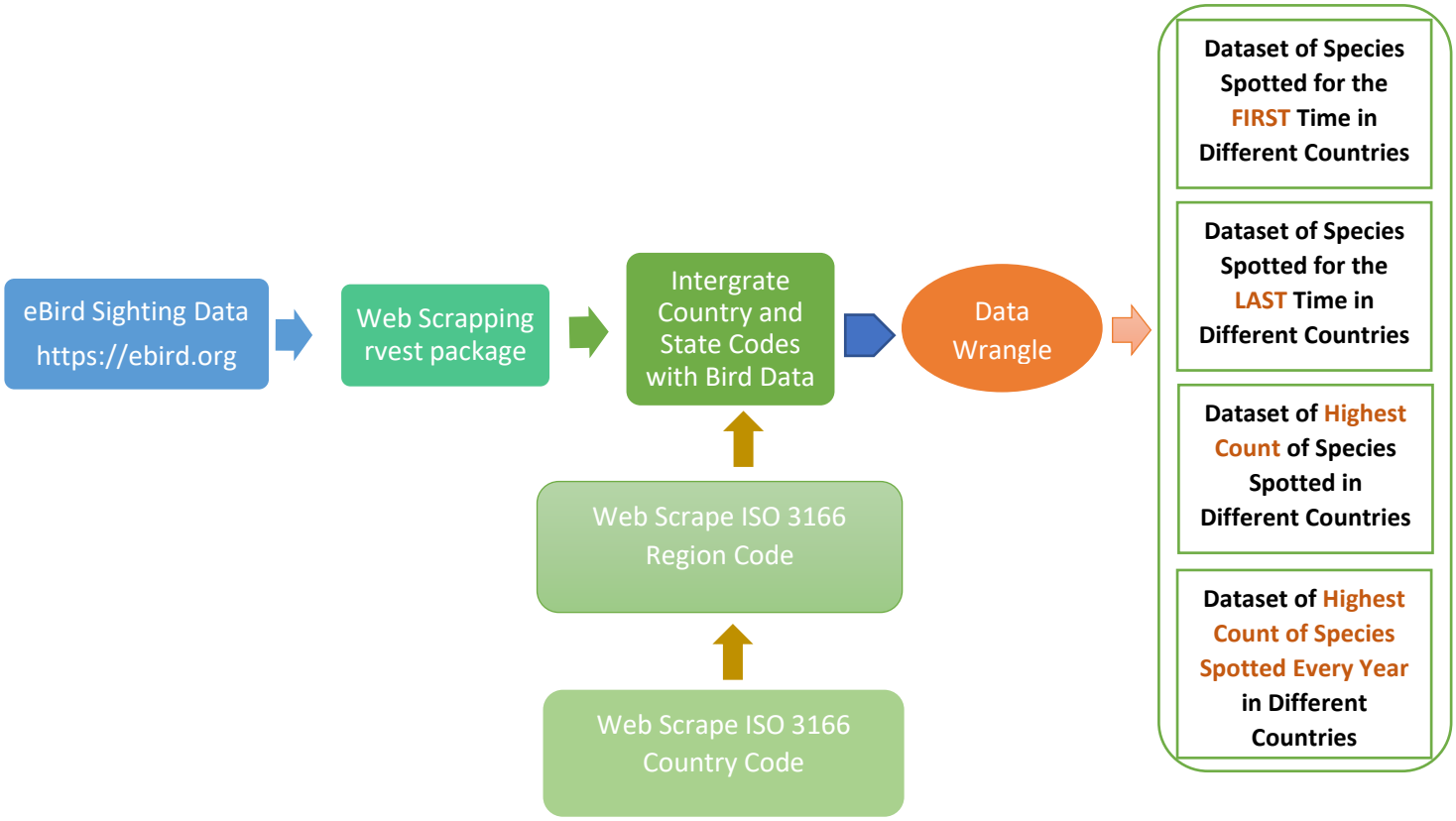

Figure 1: Data Web Scrapped for Building Bird Data Frame

The above flowchart shows the steps involved in building the final bird's dataset.

Country Code Data Frame

The input for country code is from web scrapping the Wikipedia to obtain the 2 letter country codes for all the countries of the world as represented by ISO 3166 standard, example New Zealand is represented as NZ. A function: getCountryCode <- function() that takes no input and web scrapes and returns an output data frame of the following format.

1. Country Name (Name of the Country)
2. Country Code (2-letter country code as represented by ISO 3166)
3. Subdivisions (Comment on the number of regional subdivisions)

A snapshot of the data frame is given below:

A data.frame: 249 × 3

CountryCode	CountryName	Subdivisions
<chr>	<chr>	<chr>
AD	Andorra	7 parishes
AE	United Arab Emirates	7 emirates
AF	Afghanistan	34 provinces
AG	Antigua and Barbuda	6 parishes 2 dependencies
AI	Anguilla	—
AL	Albania	12 counties
AM	Armenia	1 city 10 regions
AO	Angola	18 provinces
AQ	Antarctica	—
AR	Argentina	1 city 23 provinces
AS	American Samoa	—
AT	Austria	9 states

Figure 2: Web Scrapped Country Code Data Frame

Regional Code Data Frame

The input for regional code is also web scrapped from Wikipedia as represented by the ISO 3166 code for country region subdivisions such as county, state, cities etc. A function is written getStateCodesForCountries <- function(country) that inputs the country code and outputs the regional 2/3 lettered code for all the country subdivisions. The output data frame obtained is of the format:

1. Code (2 lettered country code - 2/3 lettered region code)
2. Subdivision (Name of the regions e.g. states or county names)
3. SubdivisionCategory (E.g. if state or county or district or parish)
4. State Code (2/3 lettered region code as represented by ISO 3166)
5. Country Code (2 lettered country code as represented by ISO 3166)

A snapshot of the data frame is given below for New Zealand and India.

A data.frame: 36 × 5

Code	Subdivision	SubdivisionCategory	StateCode	CountryCode
<chr>	<chr>	<chr>	<chr>	<chr>
IN-AP	Andhra Pradesh	State	AP	IN
IN-AR	Arunachal Pradesh	State	AR	IN
IN-AS	Assam	State	AS	IN
IN-BR	Bihar	State	BR	IN
IN-CT	Chhattisgarh[note 1]	State	CT	IN
IN-GA	Goa	State	GA	IN
IN-GJ	Gujarat	State	GJ	IN
IN-HR	Haryana	State	HR	IN
IN-HP	Himachal Pradesh	State	HP	IN
IN-JK	Jammu and Kashmir	State	JK	IN
IN-JH	Jharkhand	State	JH	IN
IN-KA	Karnataka	State	KA	IN
IN-KL	Kerala	State	KL	IN
IN-MP	Madhya Pradesh	State	MP	IN

A data.frame: 17 × 5

Code	Subdivision	SubdivisionCategory	StateCode	CountryCode
<chr>	<chr>	<chr>	<chr>	<chr>
NZ-AUK	Auckland	region	AUK	NZ
NZ-BOP	Bay of Plenty	region	BOP	NZ
NZ-CAN	Canterbury	region	CAN	NZ
NZ-GIS	Gisborne	region	GIS	NZ
NZ-HKB	Hawke's Bay	region	HKB	NZ
NZ-MBH	Marlborough	region	MBH	NZ
NZ-MWT	Manawatu-Wanganui	region	MWT	NZ
NZ-NSN	Nelson	region	NSN	NZ
NZ-NTL	Northland	region	NTL	NZ
NZ-OTA	Otago	region	OTA	NZ
NZ-STL	Southland	region	STL	NZ
NZ-TAS	Tasman	region	TAS	NZ
NZ-TKI	Taranaki	region	TKI	NZ
NZ-WKO		region	WKO	NZ

Figure 3: Web Scrapped State Code Data Frame

Bird Data Frame

Two functions are written to web scrape the bird dataset.

1. Highest Count, First and Last Spotted Counts for all Countries from 1800 to 2019

The first function inputs the attribute of interest (First Spotted, Last Spotted, Highest Count) and the country of interest and outputs a dataset for that country data available from 1800 till date. This function signature is getBirdDSPerCountry <- function(country, attribute). This function is iterated for all world countries using the information web scraped country code information and a dataset is generated that has the first, last, highest count spotted from 1800 to 2019, for all countries of the world.

The output data frame is as follows:

1. Species (Name of the Bird Species)
2. Count (Count of the Bird Species Spotted)
3. Location (Comments on the whereabouts of the birds spotted)
4. Date (includes Year/Month/Date)
5. Country (Country in which the Bird was Spotted)
6. Attribute (High-count: the highest count spotted from 1800 till 2019, First: the first spotted bird count from 1800 till 2019, Last: the last spotted bird count from 1800 till 2019)

This snapshot shows a snippet of the data frame for Australia. The X represent the count unknown.

A data.frame: 1515 × 6

Species	Count	Location	Date	Country	Attribute
<chr>	<chr>	<chr>	<date>	<chr>	<chr>
Common Ostrich	13	AU-NSW-Perricoota Road - -35.6237x144.3397 - 25 Apr 2017, 14:13	2017-04-25	AU	high_count
Emu	286	Murray-Sunset, Victoria, AU (-34.176, 141.196)	2019-03-01	AU	high_count
Spotted Whistling-Duck	84	Keatings Lagoon Conservation Park (Cooktown)	2017-12-05	AU	high_count
Wandering Whistling-Duck	10000	Kakadu National Park--Mamukala hunting track	2018-09-08	AU	high_count
Graylag Goose (Domestic type)	100	King Island	2014-07-22	AU	high_count
Graylag x Swan Goose (Domestic type) (hybrid)	1	Alcoa Wellard Wetlands	2018-07-07	AU	high_count
Canada Goose	4	Killalea State Park--Killalea Lagoon	2008-03-05	AU	high_count
Freckled Duck	1500	Lake Bael Bael	2018-03-09	AU	high_count
Black Swan	10000	Moulting Lagoon Game Reserve	2007-12-06	AU	high_count

A data.frame: 1515 × 6

Species	Count	Location	Date	Country	Attribute
<chr>	<chr>	<chr>	<date>	<chr>	<chr>
Common Ostrich	5	Emeroo Station - Turnoff at Charlton Parade	1983-07-23	AU	first
Emu	X	Big Desert Wilderness	1941-11-28	AU	first
Spotted Whistling-Duck	5	Napranum Sewage Ponds	1996-05-26	AU	first
Wandering Whistling-Duck	X	Townsville 10' Cell	1952-08-24	AU	first
Graylag Goose (Domestic type)	X	La Trobe University Bundoora Campus	1983-03-07	AU	first
Graylag x Swan Goose (Domestic type) (hybrid)	1	Alcoa Wellard Wetlands	2018-07-07	AU	first

Figure 4: Web Scrapped for Bird Data Frame -First/Last/High Count from 1800-till date for All Countries

2. Species Count of Birds in a Specific Country and Region every Year

The second function inputs the country of interest e.g. New Zealand, the year of interest e.g. 1900 and on providing the appropriate input it web scrapes the eBird Cornell Lab data site. getBirdDSPerYearStateAndCountry <- function(country, year, state) and outputs a data frame that has:

1. ID (a reference ID column)
2. Species (Name of the Bird Species)
3. Count (Count of the Bird Species Spotted)
4. Date (includes Year/Month/Date)
5. Country Code (Country in which the Bird was Spotted)
6. State Code (State in which the Bird was Spotted)

This dataset for our case is web scraped from 1970 till date. But the function can be used to web scrape data from 1800 till date, for any country of the world. A snapshot of the data frame for top 10 countries with the highest number of bird population is as below:

A data.frame: 4251 × 7

Species	Count	Location	Date	State	Country	Attribute
<chr>	<chr>	<chr>	<date>	<chr>	<chr>	<chr>
North Island Brown Kiwi	6	Tawharanui Regional Park	2018-12-29	AUK	NZ	high_count
Graylag Goose	250	Western Springs Park	2018-12-04	AUK	NZ	high_count
Canada Goose	500	Kaiaua Quarry Lakes	2018-12-27	AUK	NZ	high_count
Black Swan	600	puketutu is.	2018-03-21	AUK	NZ	high_count
Muscovy Duck (Domestic type)	7	Kaiaua Quarry Lakes	2018-01-09	AUK	NZ	high_count
Pacific Black Duck	14	Lake Kereta - southern half	2018-04-25	AUK	NZ	high_count

Figure 5: Web Scrapped for Bird Data Frame – for each Region in New Zealand

A data.frame: 26948 × 7

Species	Count	Location	Date	State	Country	Attribute
<chr>	<chr>	<chr>	<date>	<chr>	<chr>	<chr>
Fulvous Whistling-Duck	1	Kondakarla Ava	2018-09-21	AP	IN	high_count
whistling-duck sp.	5	Pulicat Lake	2018-03-03	AP	IN	high_count
Domestic goose sp. (Domestic type)	3	Vuda Park	2018-12-25	AP	IN	high_count
Ruddy Shelduck	26	Visakhapatnam Airport Backside	2018-12-31	AP	IN	high_count
Garganey	200	Pulicat Bird Sanctuary--SHAR	2018-12-23	AP	IN	high_count
Gadwall	30	Visakhapatnam Airport Backside	2018-01-03	AP	IN	high_count
Indian Spot-billed Duck	74	Gangampalli Forest	2018-08-15	AP	IN	high_count

Figure 6: Web Scrapped for Bird Data Frame – for each Region in India

Wrangling Population CSV Dataset

The CSV file had in total 63 columns in the data frame resulting in a wide data format, so we decided to reshape the data as a long format to help merge with the bird dataset.

The data was in the wide format, so we used the stack function in Julia to make it as a long data frame as we are displaying the end result based on Countries and Years. After reshaping the data, the final output would like the below snapshot.

: 264 rows × 63 columns (omitted printing of 57 columns)

	Country	Code	Indicator	Indicator_Code	1960	1961
	String	String	String	String	Int64	Int64
1	Aruba	ABW	Population, total	SP.POP.TOTL	54211	55438
2	Afghanistan	AFG	Population, total	SP.POP.TOTL	8996973	9169410
3	Angola	AGO	Population, total	SP.POP.TOTL	5454933	5531472
4	Albania	ALB	Population, total	SP.POP.TOTL	1608800	1659800
5	Andorra	AND	Population, total	SP.POP.TOTL	13411	14375
6	Arab World	ARB	Population, total	SP.POP.TOTL	92197753	94724510
7	United Arab Emirates	ARE	Population, total	SP.POP.TOTL	92418	100796
8	Argentina	ARG	Population, total	SP.POP.TOTL	20481779	20817266
9	Armenia	ARM	Population, total	SP.POP.TOTL	1874121	1941492
10	American Samoa	ASM	Population, total	SP.POP.TOTL	20123	20602

15,576 rows × 4 columns

	Year	Population_Count	Country	Code
	Symbol	Int64	String	String
1	1960	54211	Aruba	ABW
2	1960	8996973	Afghanistan	AFG
3	1960	5454933	Angola	AGO
4	1960	1608800	Albania	ALB
5	1960	13411	Andorra	AND
6	1960	92197753	Arab World	ARB
7	1960	92418	United Arab Emirates	ARE
8	1960	20481779	Argentina	ARG
9	1960	1874121	Armenia	ARM
10	1960	20123	American Samoa	ASM

Figure 7: Wrangling Wide to Long Dataset for Population Data

Web Scraping Geographical Positions of Countries in Julia

The dataset that would be useful to plot the geographic points is the country wise latitude and longitude dataset. This data is gathered from the google developer site (https://developers.google.com/public-data/docs/canonical/countries_csv). This was web scrapped and the output of the data frame is:

1) Code
2) Country
3) Latitude
4) Longitude

244 rows × 4 columns

	Code	Latitude	Longitude	Country
	String	Float64	Float64	String
1	AD	42.5462	1.60155	Andorra
2	AE	23.4241	53.8478	United Arab Emirates
3	AF	33.9391	67.71	Afghanistan
4	AG	17.0608	-61.7964	Antigua and Barbuda
5	AI	18.2206	-63.0686	Anguilla
6	AL	41.1533	20.1683	Albania
7	AM	40.0691	45.0382	Armenia
8	AN	12.2261	-69.0601	Netherlands Antilles
9	AO	-11.2027	17.8739	Angola
10	AQ	-75.251	-0.071389	Antarctica

Figure 8: Wrangled Population Data over the Years

The population dataset and latitude and longitude dataset are merged based on the country names, using the join function in the Julia. This is written as a CSV file to be merged to the final bird dataset.

From this dataset we can filter the population count for any country using both country name and country code. We can also plot the information in a graphical representation.

Wrangling Pollution CSV Dataset

To plot the data set on world map, a new data frame was created. The new data frame comprises of a measurement index for pollution which was obtaining by taking the difference of CO_2 emission from 1970 to till 2014, which was the latest available. This index when sorted in descending order gives the top 10 countries emitting CO_2. The top 10 countries listed in the new data frame are plotted in a separate graph to look into detailed variations in pollution between them. The regional codes are mutated into the data frame by using library country codes.

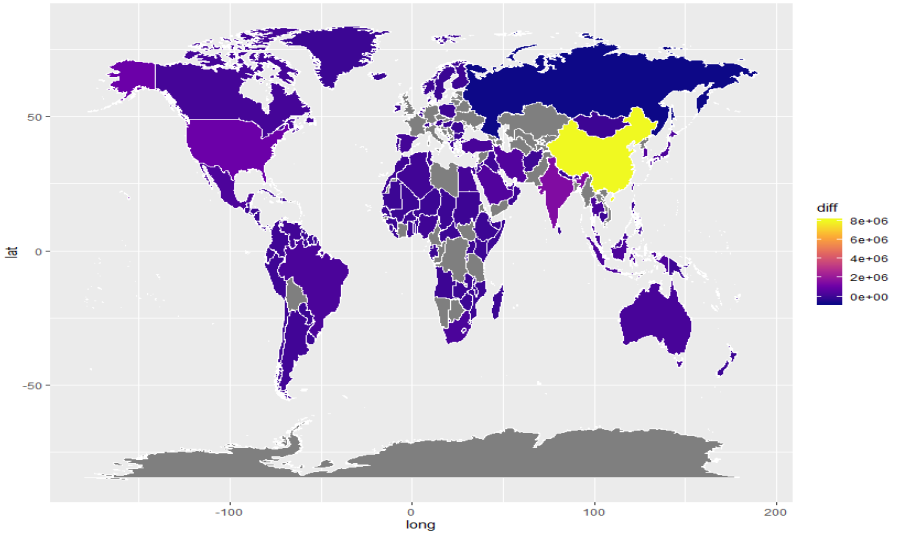

Figure 9: Plotting CO_2 emission in global scale

	region	diff
1	China	8004422.942
2	India	1524547.916
3	USA	1066627.624
4	South Korea	512932.626
5	Iran	481158.071
6	Saudi Arabia	473200.681
7	Japan	402801.615
8	Indonesia	392937.385
9	Mexico	350235.170
10	Brazil	325992.633

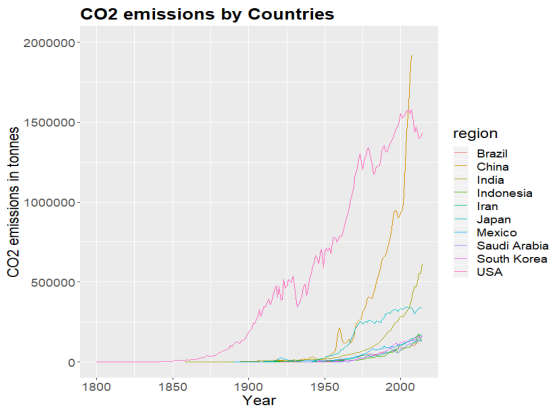

Figure 10: Table and Graph showing Top 10 countries emitting CO_2

	region	year	co2_total_emissions	iso2c
1	Afghanistan	1970	1672.152	AF
2	Afghanistan	1971	1895.839	AF
3	Afghanistan	1972	1532.806	AF
4	Afghanistan	1973	1639.149	AF
5	Afghanistan	1974	1917.841	AF
6	Afghanistan	1975	2126.860	AF
7	Afghanistan	1976	1987.514	AF
8	Afghanistan	1977	2390.884	AF
9	Afghanistan	1978	2159.863	AF
10	Afghanistan	1979	2240.537	AF

Figure 11: Wrangled Dataset for CO2 Emission

Wrangling Temperature CSV Datasets

The dataset is in 'ASCII' format, which is then converted to a csv file, after wrangling it to extract the date in the necessary format and generating a 'Space-Time-Coordinate' matrix, and finally merging the matrix with the temperature dataset before plotting the contour map showing the temperature anomalies using R.

Flow-chart for code implementation to plot global temperature anomalies using NOAA Dataset

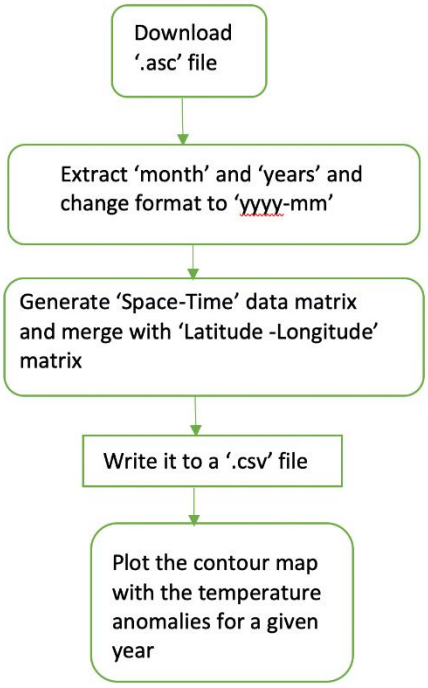

Figure 12: Flowchart for Wrangled Average Temperature Dataset

NOAA Dataset for Global Surface Temperature Data

The dataset obtained from NOAA for the merged land and oceanic surface temperature dataset is not tidy and has a lot of missing values which are filled with '-999.9'.

	1	2	3	4	5	6	7	8	9	10	11
1		LAT	LON	1880-1	1880-2	1880-3	1880-4	1880-5	1880-6	1880-7	1880-8
2	1	-87.5	2.5	-999.9	-999.9	-999.9	-999.9	-999.9	-999.9	-999.9	-999.9
3	2	-87.5	7.5	-999.9	-999.9	-999.9	-999.9	-999.9	-999.9	-999.9	-999.9
4	3	-87.5	12.5	-999.9	-999.9	-999.9	-999.9	-999.9	-999.9	-999.9	-999.9
5	4	-87.5	17.5	-999.9	-999.9	-999.9	-999.9	-999.9	-999.9	-999.9	-999.9
6	5	-87.5	22.5	-999.9	-999.9	-999.9	-999.9	-999.9	-999.9	-999.9	-999.9
7	6	-87.5	27.5	-999.9	-999.9	-999.9	-999.9	-999.9	-999.9	-999.9	-999.9
8	7	-87.5	32.5	-999.9	-999.9	-999.9	-999.9	-999.9	-999.9	-999.9	-999.9
9	8	-87.5	37.5	-999.9	-999.9	-999.9	-999.9	-999.9	-999.9	-999.9	-999.9
10	9	-87.5	42.5	-999.9	-999.9	-999.9	-999.9	-999.9	-999.9	-999.9	-999.9

Figure 13: NOAA Average Temperature Dataset

Plot of NOAA Global Temperature Anomalies

The temperature anomalies for any given year between January, 1880 to January, 2017 can be plotted. The figure below depicts the temperature anomalies for May, 2016.

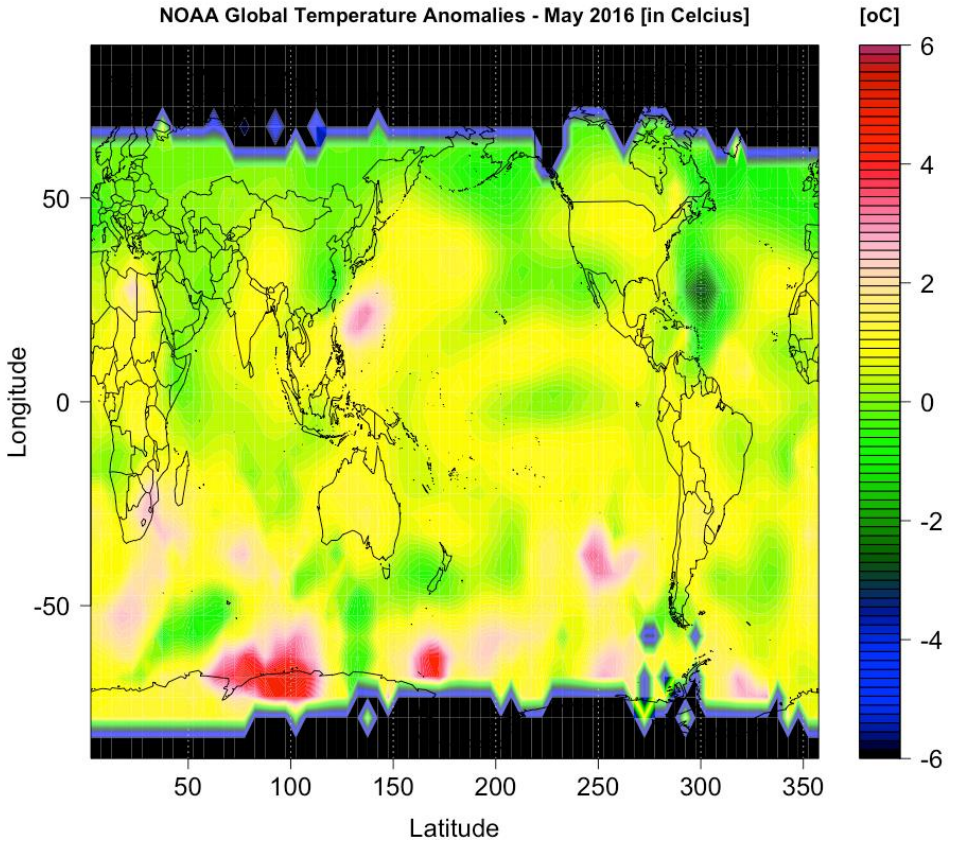

Figure 14: Temperature Anomalies

Flow-chart for code implementation to plot global average surface temperature using Berkeley Earth Dataset

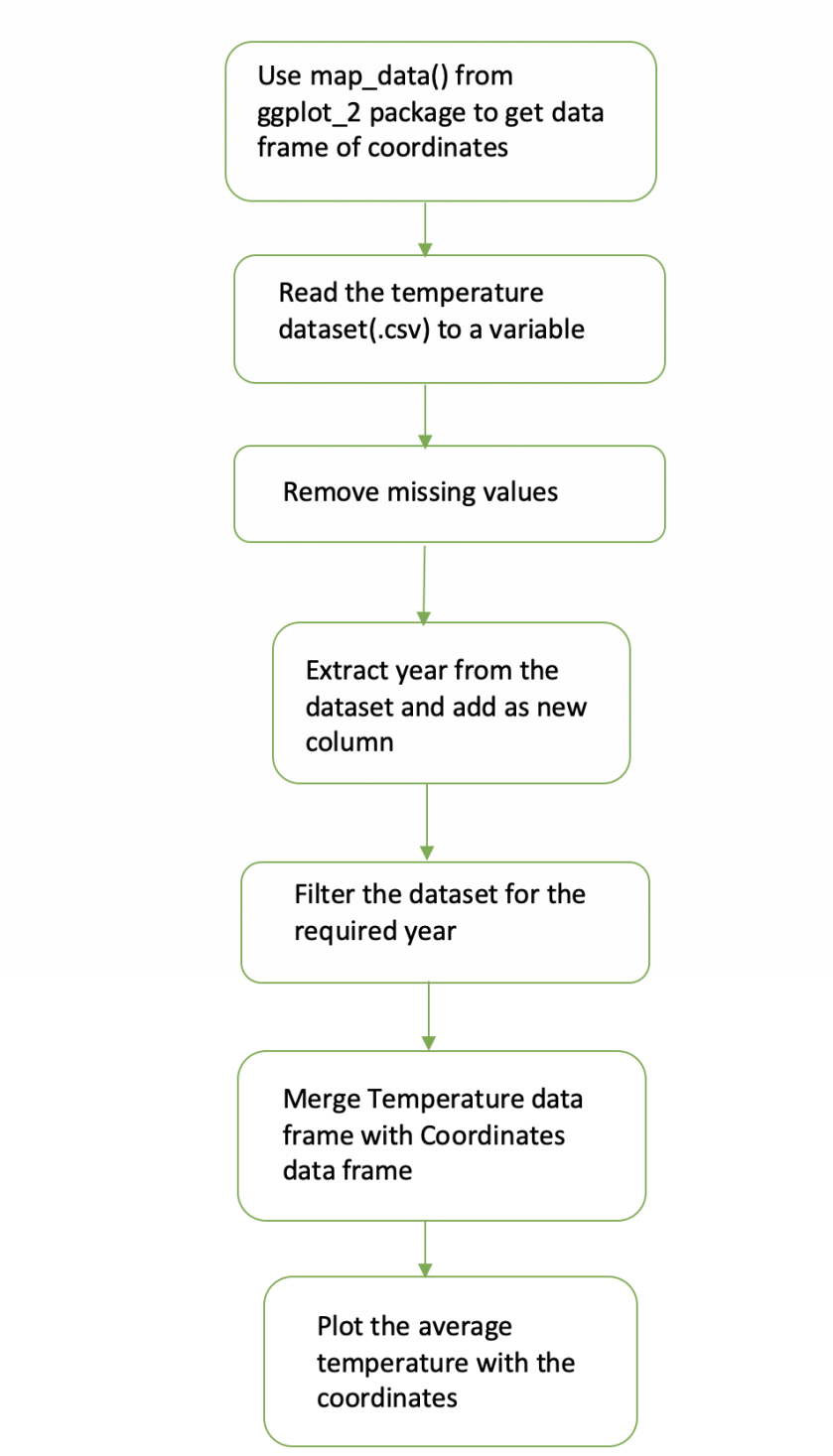

Figure 15: Flowchart for Code Implementation

Berkeley Earth Dataset for Global Surface Temperature Data

The figure below depicts Berkeley Earth dataset obtained from Kaggle. This dataset has the average surface temperature in comparison to the temperature anomalies listed in the NOAA dataset along with Country and Date.

	1	2	3	4	5	6	7
1	dt	AverageTem	AverageTem	City	Country	Latitude	Longitude
2	1743-11-01	6.068	1.737	√Örhus	Denmark	57.05N	10.33E
3	1743-12-01			√Örhus	Denmark	57.05N	10.33E
4	1744-01-01			√Örhus	Denmark	57.05N	10.33E
5	1744-02-01			√Örhus	Denmark	57.05N	10.33E
6	1744-03-01			√Örhus	Denmark	57.05N	10.33E
7	1744-04-01	5.788	3.624	√Örhus	Denmark	57.05N	10.33E
8	1744-05-01	10.644	1.283	√Örhus	Denmark	57.05N	10.33E
9	1744-06-01	14.051	1.347	√Örhus	Denmark	57.05N	10.33E
10	1744-07-01	16.082	1.396	√Örhus	Denmark	57.05N	10.33E

Figure 16: Berkeley Earth Temperature Dataset

Plot of Average Global Surface Temperature - (Berkeley Earth Dataset)

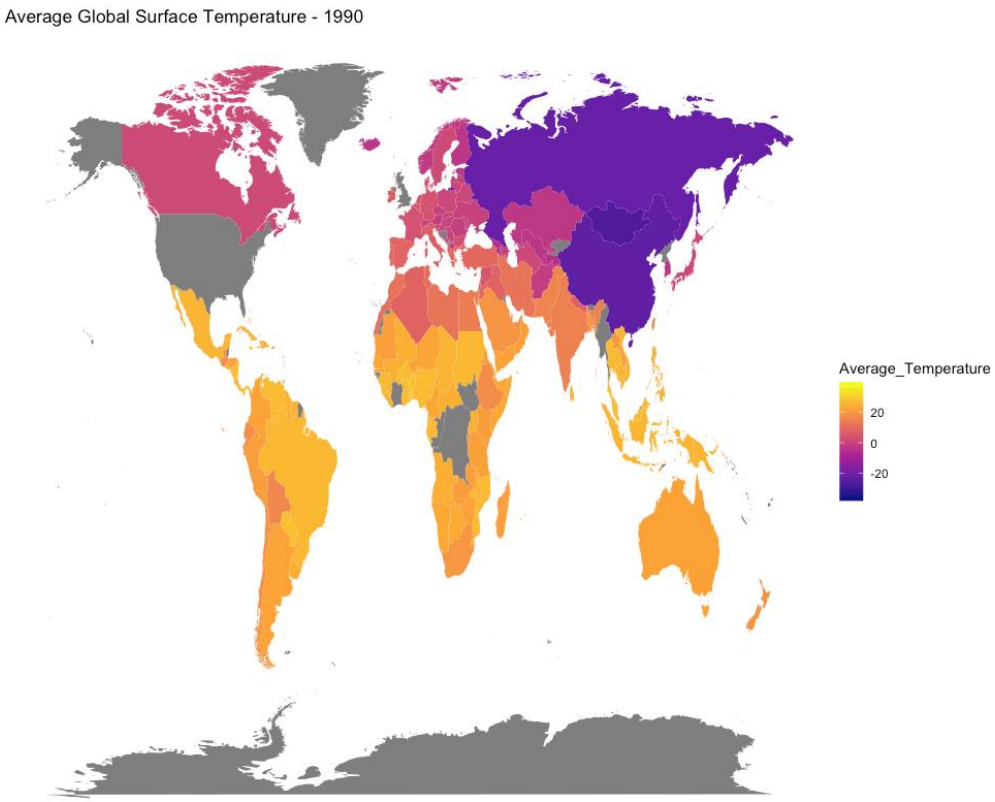

Figure 17: Plot of Average Global Temperature

Final Dataset

There are two final data set generated:

1. Bird Species and its count over 1800 to 2019 for countries of the world

A data.frame: 5 × 8

ID	Species	Count	Location	Date	Country	Attribute	Year
<int>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<int>
1	Mallard	X	Andorra la Vella	2010-05-17	AD	first	2010
2	Gray Partridge	7	Carretera de Prats Sobirans, La Massana, La Massana, AD (42,579, 1,477)	2007-06-01	AD	first	2007
3	Rock Ptarmigan	1	Pas de la Casa	1979-06-19	AD	first	1979
4	Common Wood-Pigeon	X	Andorra: Vall D' Ordino	1969-04-22	AD	first	1969
5	Common Cuckoo	2	Andorra: Vall D' Ordino	1969-04-22	AD	first	1969

2. Bird Species and its count together with temperature, population and pollution.

A data.frame: 5 × 13

Country	Country_Code	Species	Count	Location	Date	Year	Attribute	Population_Count	CO2_Emission	Average_Temperature	Latitude	Longitude
<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<int>	<chr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>
United Arab Emirates	AE	Graylag Goose	1	Abu Dhabi--general area	1976-12-03	1976	first	637922	39651.27	26.42933	23.42408	53.84782
United Arab Emirates	AE	Lesser White-fronted Goose	1	Umm al-Qaiwain Rubbish Tip	1996-11-20	1996	first	2539126	41059.40	27.20233	23.42408	53.84782
United Arab Emirates	AE	Mute Swan	3	Eastern Lagoon (Eastern Mangroves NP)	1984-12-13	1984	first	1293971	46394.88	26.29108	23.42408	53.84782
United Arab Emirates	AE	Tundra Swan	2	Nad al-Sheba	1994-01-25	1994	first	2294385	73130.98	27.25042	23.42408	53.84782
United Arab Emirates	AE	Egyptian Goose	200	Abu al-Abyadh Island	1978-01-01	1978	first	835508	44814.41	26.89042	23.42408	53.84782

E. Difficulties Overcome to Wrangle Data Source

Understanding Country and Regional Codes:

Bird Dataset required information in form of country and regional code as one of its input. To understand that there was a universal ISO standard of representation for each country and its subdivisions took a lot of time and effort. The web scraped data for bird dataset now inputs this ISO format of the data to obtain the output data frame.

Web scrapping Bird Dataset:

The information on the eBird web page is available in 2 different sources, one just on the web page, each page dedicated to a particular species and one other as a form that you fill in to get all species in a country. A significant amount of time was spent trying to web scrape from the first mentioned page and consolidate the information for all species which was less productive and clumsy. The code provided now is from the second approach, better optimised. Also, for a few of the countries, the web scraping could not be done globally, hence a different method that involves country together with regional information was added to web scrape those data.

Wrangling Bird Dataset:

The final bird data set gathered is really vast, with around 249 countries and different species of birds over the year 1800 to 2019. Any data wrangling done for the case study derived or for the joining the dataset with other parameters proved difficult as every time, the count of rows and columns were constantly checked and rechecked if the data wrangled were right.

Wrangling Wide Population Dataset to Long Population Dataset:

The obtained population dataset was wide and merging was difficult hence we changed the dataset to long to suit our needs. This took considerable amount of time to check the rows if they were consistent.

Web Scraping Geographical Position Dataset:

Web scrapping of the dataset resulted in an empty element, and converting it to a data frame was challenging. This was done by eliminating that empty HTML element.

Plotting Pollution data in world map and regional code:

Plotting the data in a world map is cumbersome considering that the individual region names and column name have to be matched. The same is needed to obtain country codes using library “countrycode”. The list of regions unmapped in the world map during the initial trial were further analysed. The region name from “world_map” data and our data set of interest is listed by using unique function and later replaced with its corresponding name as seen in world map data. Also, to plot a set measure is needed. Hence an index which is measured as a difference of $C0_2$ emitted from 1970 till 2014 the latest data available was used to plot the data in the world map.

Missing data in the Temperature Datasets:

Both the datasets from NOAA and Berkeley Earth have a lot of missing values. This result is obvious in the plots, as the NOAA dataset does not have any data for the region near the north and south pole. Similarly, the Berkeley dataset also has a lot of missing values for various countries.

F. Managed to Achieve

The following are the achievements as a result of this project:

1. Managed to web scrape and create a bird dataset that will provide information of the highest count of birds sighted per species, the first date when a bird species was spotted and the last date when a bird species was spotted. All this information was web scraped for all countries of the world globally and for a period from 1800 up until 2019 (for which the information was available).
2. Managed to web scrape the bird count locally per species region per country (highest count). This information is vital to look at a specific region (county / district / state) locally to understand the bird sighting count. This information can be web scrapped from 1800 up until 2019. However, for this project the years chosen were 1970 until 2019, based on the information available form rest of the datasets needed to merge.
3. Managed to write a method that would help web scrape the country code and regional codes of the world.
4. Managed to wrangle population using Julia, for the years starting 1960 to 2018.
5. Managed to web scrape geographical position dataset for various countries using Julia.
6. Managed to obtain a pollution index which identifies the top most polluting country and identify them with their corresponding country codes.
7. Conducted a case study of India from the dataset gathered and identified the trends of local, migrating summer and winter birds. Also identified the species of birds that are in decline and critically endangered and birds that have not been spotted / extinct in India from 1800, and over the decades.
8. Managed to wrangle both datasets from NOAA and Berkeley Earth and plot the respective temperature values for the coordinates, thereby, depicting the change in average temperature and temperature anomalies between various years.

Case Study from the Dataset - India

The following is one of the many case studies that can be studied with the dataset that we have built. India was chosen for the following reasons: it currently ranks 8th in the number of diverse species spotted as per the eBirds web site. It ranks 2nd as the world's most populous country in the world and it is in the top 10 list of countries that emit higher proportion of greenhouse gases especially CO_2. Although India being in the equatorial region, the temperature hasn't changed as much compared to other countries in the temperate world, it has been a go-to destination for many summer and winter migratory bird species from the temperate zones of the world. Hence studying India might provide interesting insights.

Birds of India

The birds of India can be studied as two categories: the native Indian birds and the seasonal migratory birds of India.

Native Species

A few of the most common native species are chosen and their highest count spotted over a couple of decades is plotted. The highest count spotted is on the raise for these species.

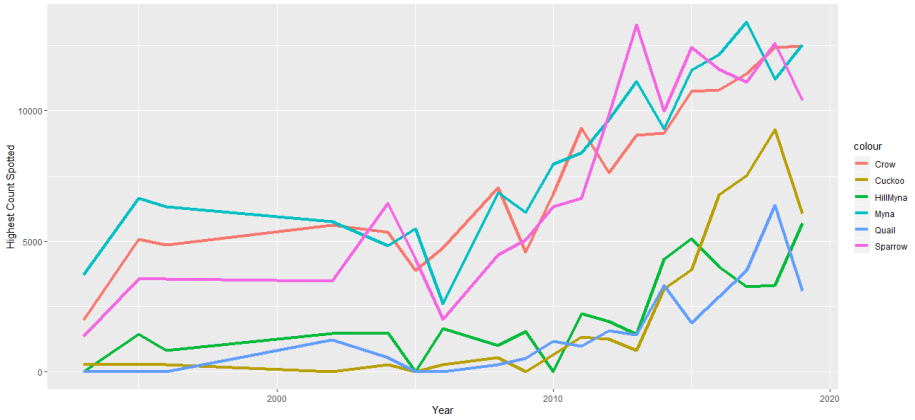

Figure 18: Plot showing the Highest Count Spotted VS Year for India Native Birds

Migratory Species

Winter migratory birds migrate to India from the cold temperate northern regions mostly during the months of November to February, while the summer migratory birds migrate to India from southern regions of the world mostly during June to August. The below plot shows the count of few species of summer and winter migratory birds over the years.

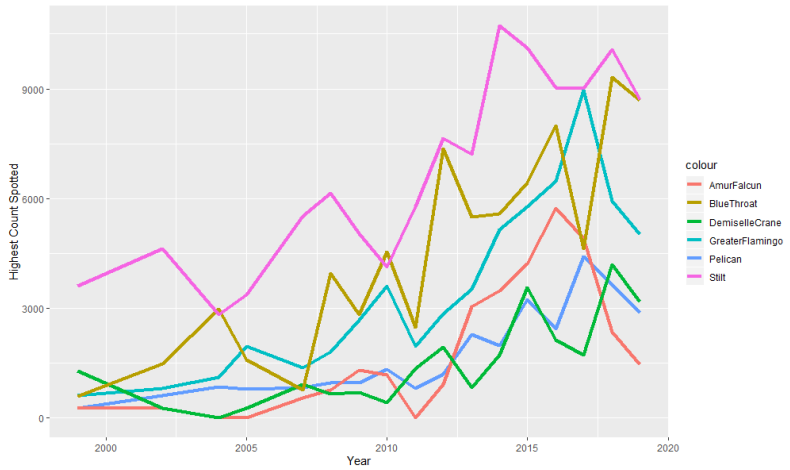

Figure 19: Plot showing the Highest Count Spotted VS Year for Winter Migratory Birds

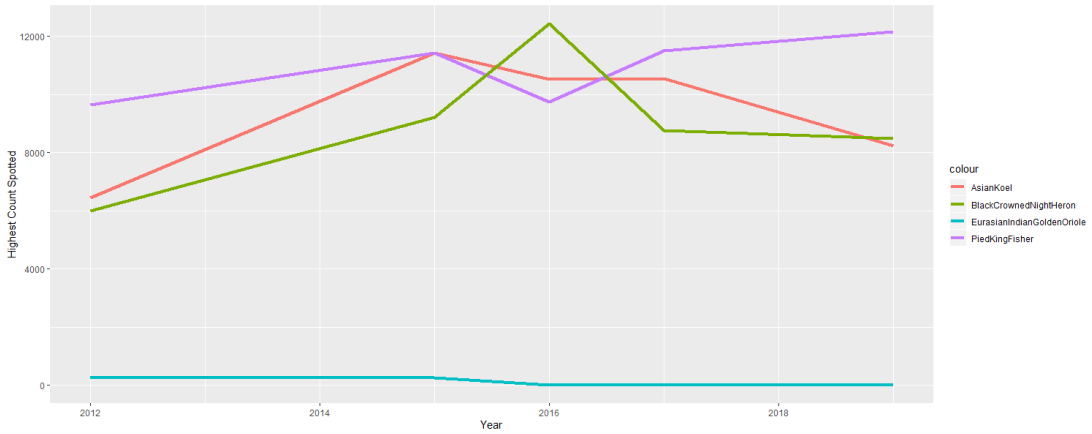

Figure 20: Plot showing the Highest Count Spotted VS Year for Summer Migratory Birds

Species of Bird Extinct/Not Spotted in India

One other important question that the dataset can address is the count of species that were lost over a period of time. This can be obtained from the last seen/spotted dataset generated. As can be seen from the below graph, a considerable number of species were not spotted after a specific decade and the count of all these species in different decades gives the total count of bird “not spotted”/possibly extinct in India.

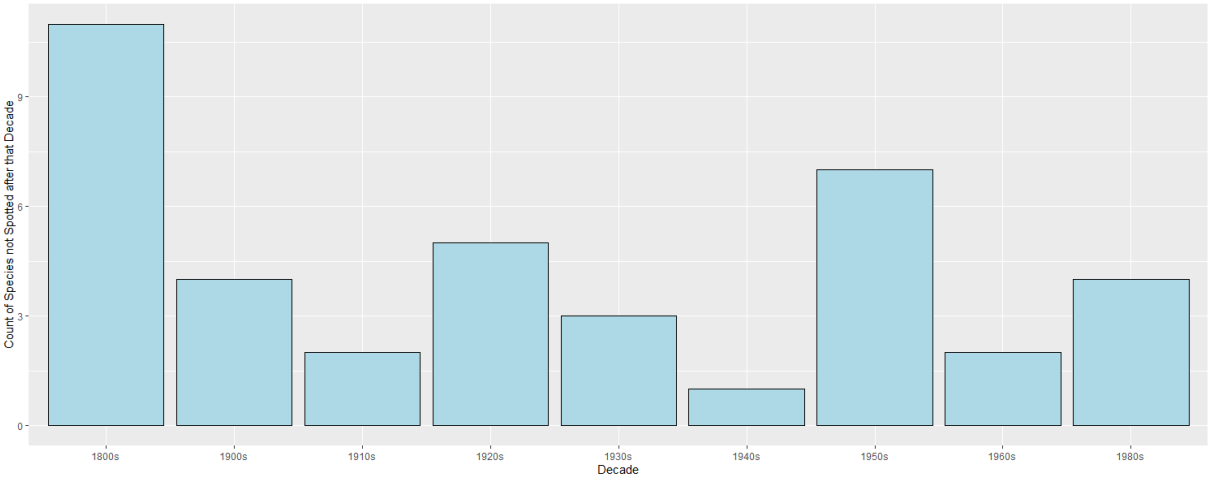

Figure 21: Plot showing the Count of Bird Species that were not spotted after a particular Decade

The below plots show the scatter plot of CO2 and temperature vs highest bird count per year for India.

As can be seen from the plots, the temperature variation is ~1 degree centigrade, over three decades for which the data was plotted. The plot shows a week positive correlation.

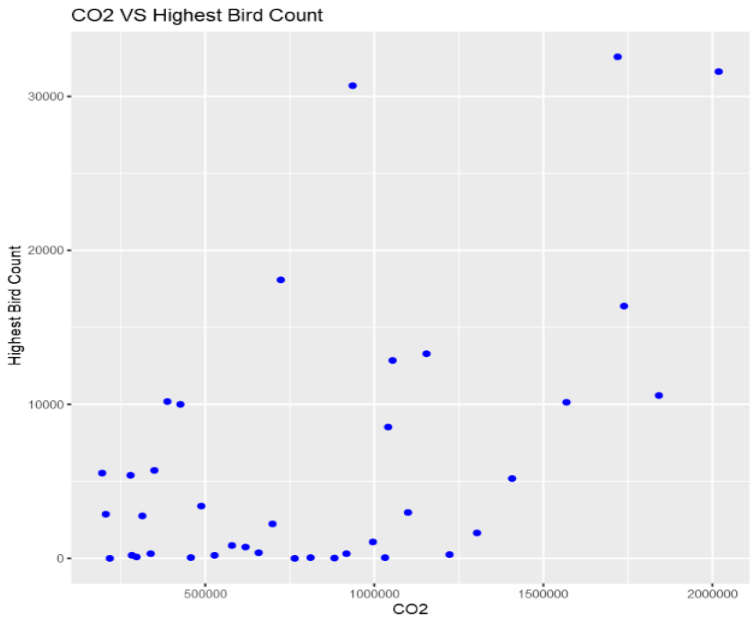

Figure 22: Scatter Plot showing the Count of Bird Vs Levels of CO2 over the Years

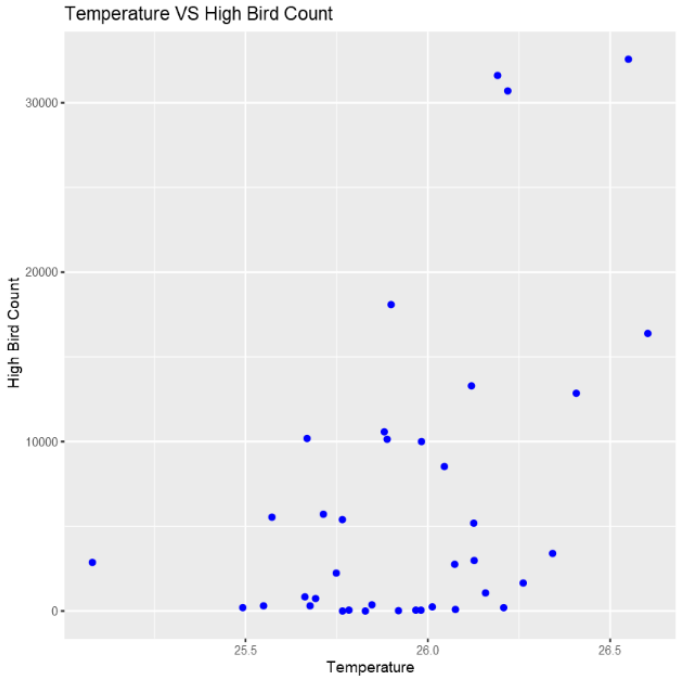

Figure 22: Scatter Plot showing the Count of Bird Vs Levels of Temperature over the Years

G. Failed to Achieve

The following are the achievements as a result of this project:

1. The dataset contained only the common species name of birds, the scientific name of the bird with genus is not available.
2. The dataset takes a long time to web scrape, although it does web scrape for all countries from 1800 to 2019, need to check if there are any way to do this optimally. This is the case for region-based dataset too.
3. There are certainly other factors that can be considered with regards to bird extinction such as deforestation, local hunting, etc. This additional information can be gathered and added to this dataset to provide better insights into the bird depletion rate.
4. The NOAA dataset could not be merged with the final dataset, as it needed the ISO Alpha 2 Country Code, as the common data column. This was tried to be implemented using the 'geoname' package. However, the error 'HTTP status was 401 Unauthorized' kept persisting. In order to use the package, it requires that a username be created and activated. Another package, namely, 'revgeo' was also explored, but a paid membership to 'Google Maps API' or 'BING API' was required, as the default 'Photon API' did not return any valid values.
5. The final dataset has information on the bird count, pollution, population, temperature for years starting from 1970 to 2014, although the bird dataset was web scrapped from 1800 to 2019. Proper datasets prior to 1970 for the population and pollution need to be collected to have a better insight for those years.

REFERENECES:

1. Biello, D. (2006, July 5). Bird Extinction Estimates May Be Too Low. Retrieved from https://www.scientificamerican.com/article/bird-extinction-estimates/.
2. Powell, H. (n.d.). Nearly 3 Billion Birds Gone. Retrieved from https://www.birds.cornell.edu/home/bring-birds-back.
3. (n.d.). Retrieved from https://evolution.berkeley.edu/evolibrary/article/0_0_0/massextinct_10.
4. Climate Change: Vital Signs of the Planet. (n.d.). Retrieved from https://climate.nasa.gov/.
5. Population. (n.d.). Retrieved from https://www.un.org/en/sections/issues-depth/population/.
6. 68% of the world population projected to live in urban areas by 2050, says UN | UN DESA Department of Economic and Social Affairs. (n.d.). Retrieved from https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html.
7. User, S. (n.d.). Global Carbon Emissions. Retrieved from https://www.co2.earth/global-co2-emissions.
8. Ritchie, H., & Roser, M. (2017, May 11). CO_2 and Greenhouse Gas Emissions. Retrieved from https://ourworldindata.org/co2-and-other-greenhouse-gas-emissions.

9. countries.csv | Dataset Publishing Language | Google Developers. (n.d.). Retrieved from https://developers.google.com/public-data/docs/canonical/countries_csv.
10. Population, total. (n.d.). Retrieved from https://data.worldbank.org/indicator/SP.POP.TOTL.
11. Data Overview. (n.d.). Retrieved from http://berkeleyearth.org/data/.
12. Earth, B. (2017, May 1). Climate Change: Earth Surface Temperature Data. Retrieved from https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data.
13. Laboratory, O. R. N. (n.d.). Carbon Dioxide Information Analysis Center. Retrieved from http://cdiac.ornl.gov/.
14. National Oceanic and Atmospheric Administration. (n.d.). Retrieved from https://www.noaa.gov/.