**Relation between global crop production, yield and temperature (using excel, sql, tableau)**

The aim of this project is to showcase the skills learned by asking SMART questions, structured thinking, data management, data cleaning, data analysis & data visualization.

**About the Project**

**Here we are going to see the relation between global crop production, crop yield, crop production in hectares of land and global temperature from 1900 – 2017**

# [Twentieth Century Crop Statistics, v1 (1900 – 2017)](https://sedac.ciesin.columbia.edu/data/set/food-twentieth-century-crop-statistics-1900-2017)

Data collected from [**Socioeconomic Data and Applications Center (sedac)**](https://sedac.ciesin.columbia.edu/)*A Data Center in NASA's Earth Observing System Data and Information System (*[*EOSDIS*](https://earthdata.nasa.gov/)*) — Hosted by*[*CIESIN*](http://www.ciesin.columbia.edu/)*at*[*Columbia University*](http://www.columbia.edu/)

**https://sedac.ciesin.columbia.edu/data/set/food-twentieth-century-crop-statistics-1900-2017**

This document outlines the methodology and data sets used to construct the Twentieth Century Crop Statistics, 1900-2017, and consists of national or subnational maize and wheat production, yield, and harvested area statistics for all available years from 1900- 2017.

# Climate Change: Earth Surface Temperature Data

# Global land temperature data collected from kaggle website <https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data>

# Here it contains

Global Land and Ocean-and-Land Temperatures (GlobalTemperatures.csv):

* Date: starts in 1750 for average land temperature and 1850 for max and min land temperatures and global ocean and land temperatures
* LandAverageTemperature: global average land temperature in celsius
* LandAverageTemperatureUncertainty: the 95% confidence interval around the average

### ****Process****

In the process phase is where we create and transform data but still maintaning its integrity. It is also where we test & validate data that can help us determine which data fields to clean if necessary

**In this phase is where I**:

* Data is collected in excel and done possible cleaning here itself
* **used Microsoft SQL server management studio as an analysis tool**, since it processes data faster compared to spreadsheets for big datasets.
* **checked for data integrity** by running different queries across field names that would help me in data cleaning.
* **transformed data** by adding and updating those columns that will add other categories to group user types even more to uncover patterns or give more context.
* verified and **documented data cleaning results**.

#### ****Data Cleaning and Manipulation Documentation****

#### ****Data manipulation Using Spreadsheet****

As there are many sheets and I decided to take only important sheets and renamed the sheets that I am using.

#### ****Data validation using SQL (BigQuery)****

After collecting necessary sheets I have imported data to SQL using Import data function.

Here I

1. Checked for **NULL values** for each column / data field.
2. Checked for **duplicate** entries.
3. Checked for **misspelled words** for columns with STRING as its data type.
4. Checked for **trailing & leading spaces** on STRING values
5. Checked for data types that are used

**Commands used**: SELECT, FROM, WHERE, GROUP BY, ORDER BY, COUNT, IS NULL, OR & LIKE

#### ****Queries for data validation, cleaning & manipulation**** ****of Crop statistics****

select \* from [Portfoio\_project].[dbo].[Sheet1$]

--checking the columns present in data

select admin0 from [Portfoio\_project].[dbo].[Sheet1$]

select admin0, count(Harvest\_year) as year from [Portfoio\_project].[dbo].[Sheet1$]

WHERE admin0 IS NOT NULL

group by admin0

order by year

select admin0, count(Harvest\_year) as year, count(crop) as crop\_count from [Portfoio\_project].[dbo].[Sheet1$]

WHERE admin0 IS NOT NULL

group by admin0

order by year

-- checking and selecting the necessary columns by removing nulls

select admin0, Harvest\_year, crop, [hectares (ha)], [production (tonnes)], [yield(tonnes/ha)] from [Portfoio\_project].[dbo].[Sheet1$]

where [hectares (ha)] is not null

and [yield(tonnes/ha)] is not null

and [production (tonnes)] is not null

order by admin0, Harvest\_year

--checking the data for particular country

select admin0, Harvest\_year, crop, [hectares (ha)], [production (tonnes)], [yield(tonnes/ha)] from [Portfoio\_project].[dbo].[Sheet1$]

where [hectares (ha)] is not null

and [yield(tonnes/ha)] is not null

and [production (tonnes)] is not null

and admin0 = 'India'

-- selecting unique countries

SELECT distinct admin0 FROM [Portfoio\_project].[dbo].[Sheet1$]

select admin0, Avg([hectares (ha)]), AVG([production (tonnes)]), Avg([yield(tonnes/ha)]) from [Portfoio\_project].[dbo].[Sheet1$]

where [hectares (ha)] is not null

and [yield(tonnes/ha)] is not null

and [production (tonnes)] is not null

--group by admin0

order by admin0, Harvest\_year

select Avg([yield(tonnes/ha)]) from [Portfoio\_project].dbo.Sheet1$

-- changing hectares column data type from varchar to bigint

alter table [Portfoio\_project].[dbo].[Sheet1$]

alter column [hectares (ha)] money

SELECT CONVERT(int, '[hectares (ha)]') from [Portfoio\_project].[dbo].[Sheet1$]

go

select admin0, Harvest\_year, [hectares (ha)] from [Portfoio\_project].[dbo].[Sheet1$]

where [hectares (ha)] = '2101756.9'

SELECT CONVERT(money, '2101756.9')

alter table [Portfoio\_project].[dbo].[Sheet1$]

alter column [hectares (ha)] bigint

alter table [Portfoio\_project].[dbo].[Sheet1$]

alter column [production (tonnes)] money

-- changing productin column data type from varchar to bigint

alter table [Portfoio\_project].[dbo].[Sheet1$]

alter column [production (tonnes)] bigint

select Avg([hectares (ha)]) from [Portfoio\_project].dbo.Sheet1$

where [hectares (ha)] is not null

select Avg([production (tonnes)]) from [Portfoio\_project].dbo.Sheet1$

where [production (tonnes)] is not null

-- selecting necessary columnns by using aggregate and group by functions

select country, crop, sum(Harvest\_year) as Sum\_year, Avg([hectares (ha)]) as Avg\_hectares, AVG([production (tonnes)]) as Avg\_production, Avg([yield(tonnes/ha)]) as Avg\_yield from [Portfoio\_project].[dbo].[Sheet1$]

where [hectares (ha)] is not null

and [yield(tonnes/ha)] is not null

and [production (tonnes)] is not null

group by country, crop

order by country

-- creating new table by grouping necessary columns

select country, crop, cast(sum(Harvest\_year) as bigint) as Sum\_year, cast(Avg([hectares (ha)]) as bigint) as Avg\_hectares,

cast(AVG([production (tonnes)]) as bigint) as Avg\_production, Avg([yield(tonnes/ha)]) as Avg\_yield

into Table\_1 from [Portfoio\_project].[dbo].[Sheet1$]

where [hectares (ha)] is not null

and [yield(tonnes/ha)] is not null

and [production (tonnes)] is not null

group by country, crop

order by country

select \* from Table\_1

-- changing the column name by using sp\_rename changed admin0 to country

alter table

sp\_rename '[Portfoio\_project].[dbo].[Sheet1$].admin0', 'country', 'column'

select \* from Table\_1

order by country

-- dropping table\_1

drop table Table\_1

-- creating new table as Crops\_Table by grouping and aggregating necessary columns

select country, crop, count(Harvest\_year) as count\_Year, cast(Avg([hectares (ha)]) as bigint) as Avg\_hectares,

cast(AVG([production (tonnes)]) as bigint) as Avg\_production, Avg([yield(tonnes/ha)]) as Avg\_yield

into Crops\_Table from [Portfoio\_project].[dbo].[Sheet1$]

where [hectares (ha)] is not null

and [yield(tonnes/ha)] is not null

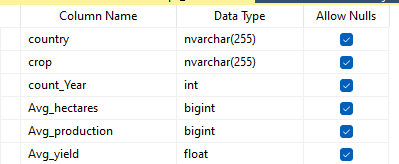
and [production (tonnes)] is not null

group by country, crop

order by country

select \* from Crops\_Table

order by country



Country crop count\_Year Avg\_hectares Avg\_production Avg\_yield

Argentina wheat 236 5451790 8005423 1.49582203389831

Australia maize 938 27090 58856 3.02568990345629

Australia wheat 1786 971862 1275707 1.26890184877268

Austria wheat 10 303002 1706993 5.62336

Belgium wheat 10 203851 1784058 8.7535

Brazil wheat 588 419490 594350 1.17919469117007

Brazil maize 522 1147837 3178929 2.51577011494253

Canada wheat 220 9774674 16654355 1.66893682428182

Canada maize 220 531516 3634964 4.92160724141818

Chile wheat 176 588659 1167560 2.43120001802273

Chile maize 176 85894 516488 5.05535832886364

#### ****Queries for data validation, cleaning & manipulation**** ****of climate change****

-- selecting temp table

select \* from [Portfoio\_project].[dbo].[Temp$]

where country is not null

order by country, dt

--trying to alter dt to int but there was a error

alter table [Portfoio\_project].[dbo].[Temp$]

alter column dt int

-- using cast to to convert datatype to date

select cast([dt] as date) from [Portfoio\_project].[dbo].[Temp$] as Date

-- creating new tabe to make it simple as Table\_1

select CAST(dt as date) as Date, AverageTemperature, AverageTemperatureUncertainty, Country into Table\_1

from [Portfoio\_project].[dbo].[Temp$]

select \* from Table\_1

--using format to select only year from the date into temp\_table

select FORMAT(Date, 'yyyy') as Year, AverageTemperature, AverageTemperatureUncertainty,

Country into temp\_table from Table\_1

where country is not null

group by country

order by country

alter table Table\_1

alter column Date date

select CAST(Date as date)

format(Date, 'yyyy') from Table\_1

select FORMAT(Date, 'yyyy') as Year from Table\_1

-- creating table as temperature by using aggregate and group by functions

select count(Year) as Count\_Year, AVG(AverageTemperature) as Avg\_Temp, Avg(AverageTemperatureUncertainty) as Avg\_Temp\_Uncertainty,

Country into temperature from temp\_table

group by Country

order by Country

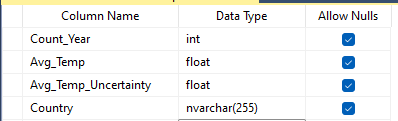
-- checking the data of united states

select \* from temperature

where Country = 'United States'

select \* from temperature

order by Country



1365 5.66115395894428 0.463844574780059 Ã…land

1365 14.225853372434 0.494984604105572 Afghanistan

1365 24.2422653958944 0.235238269794721 Africa

1365 12.8859809384164 0.431643695014663 Albania

1365 23.2662118768328 0.517530058651026 Algeria

1365 26.7195095307918 0.483642228739003 American Samoa

1365 11.462202346041 0.369406158357771 Andorra

1365 21.962348973607 0.5805784457478 Angola

1365 26.8114164222874 0.37850219941349 Anguilla

Now two tables are prepared as Crops\_Table and temperature

Joining the two table where country is common to find the relation between temperature and crops

--Joining the two tables using join on country column from 2 tables

--and creating crops\_vs\_Temp table

SELECT temperature.Country, Crops\_Table.crop, temperature.Avg\_Temp, temperature.Count\_Year, Crops\_Table.Avg\_production, Crops\_Table.Avg\_yield

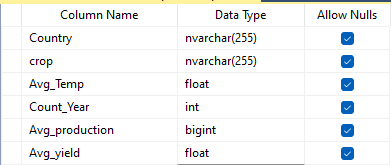
into Crop\_vs\_Temp FROM Crops\_Table

JOIN temperature

ON Crops\_Table.country = temperature.Country

select \* from Crop\_vs\_Temp

order by Country



Argentina wheat 14.814848973607 1365 8005423 1.49582203389831

Australia maize 21.7401341642229 1365 58856 3.02568990345629

Australia wheat 21.7401341642229 1365 1275707 1.26890184877268

Austria wheat 6.49602052785924 1365 1706993 5.62336

Belgium wheat 9.78588782991203 1365 1784058 8.7535

Brazil maize 24.9535234604105 1365 3178929 2.51577011494253

Brazil wheat 24.9535234604105 1365 594350 1.17919469117007

Canada maize -4.74358388278389 1365 3634964 4.92160724141818

Canada wheat -4.74358388278389 1365 16654355 1.66893682428182

Chile wheat 9.57297800586512 1365 1167560 2.43120001802273

Chile maize 9.57297800586512 1365 516488 5.05535832886364

### ****Analysis****

Here we created new table as Crop\_vs\_Temp from joining two tables Crops\_Table and temperature which contains country, crop, Avg\_Temp, Count\_Year, Avg\_production, Avg\_yield as columns.

By using these we can find relation between temperature, crop production and crop yield of major countries from 1900 – 2017

And also we can separately analyze the statistics of crop production, land used, and crop yield of different countries by using visualizations

And also we can analyze average temperature and temperature uncertainty globally by using viz.

#### ****Data visualization****

**In this phase is where I:**

* used **Tableau as a tool that will create effective & compelling data viz**

