DATA ASSIGNMENT:

- **Please read the data description before attempting the questions.**
- **Please submit an ordered document with appropriate axes labels and figure titles.**
- **Please present regression results in a table as indicated in the last page of this document.**

Q1) Construct a dataset for the following three variables and append it to your existing dataset (main.csv) as three columns. Submit the complete file in .csv or .txt format.

Explanatory Variables:	**Links**	**Frequency/ Comments**	**Variable Name (to be assigned)**
State Wise GDP	https://esopb.gov.in/static/PDF/GSDP/Statewise-Data/statewisedata.pdf	Yearly	gdp
State Wise Number of Hospital Beds (as of 2020)	https://cddep.org/wp-content/uploads/2020/04/State-wise-estimates-of-current-beds-and-ventilators_24Apr2020.pdf	In 2020	beds
District Wise Tap Water Access (Percentage of Households) as of 2019	https://ejalshakti.gov.in/jjmreport/JJMIndia.aspx	As of 2019	tap

Table 1: Explanatory variables

Q2) Recall from class that we are interested in understanding the impact of agricultural growth on health outcomes. In this context, consider the following dependent variables (DVs):

Variable Description	**Variable Label**
Percentage of infant deaths due to Sepsis (to total reported infant deaths)	sepsis
Percentage of infant deaths due to Low Birth Weight (LBW) (to total reported infant deaths)	lbw
Percentage of infant deaths due to Pneumonia (to total reported infant deaths)	pneumonia
Percentage of infant deaths due to Diarrhoea (to total reported infant deaths)	diarrhoea
Percentage of infant deaths due to Fever (to total reported infant deaths)	fever
Percentage of infant deaths due to Measles (to total reported infant deaths)	measles

(A) Summarize each of the 6 dependent variables (DVs) by reporting the following statistics:

- Mean

- Median
- Mode
- Standard Deviation

(B) Additionally, please provide year-wise and season-wise histograms for each variable.

(C) Please report any outliers.

(D) Compute the pair-wise correlation coefficients between each of the 6 dependent variables and:

1. Explanatory variables described in table 1 above.
2. Yield Index for each of the six crop categories (i.e., pulses, cereals, coarse cereals,
3. Yield Index growth rate for each of the six crop categories.

Q3) Estimate the following regression models and interpret the results for any one dependent variable:

(A)
$$H_{i,t} = \beta_0 + \beta_1(gdp_{i,t}) + \beta_2(beds_{i,t}) + \beta_3(tap_{i,t}) + u_{i,t}$$
i : districts | t : year

(B)
$$H_{i,t} = \beta_0 + \beta_1(gdp_{i,t}) + \beta_2(beds_{i,t}) + \beta_3(tap_{i,t}) + \beta_4(yield\_index_{c,i,t}) + u_{i,t}$$
i : districts | t : year
c : crop category $\in$ {cereals,coarse cereals,cash,oilseeds, horticulture}

(C)
$$H_{i,t} = \beta_0 + \beta_1(gdp_{i,t}) + \beta_2(beds_{i,t}) + \beta_3(tap_{i,t}) + \sum_c \beta_c(yield\_index_{c,i,t}) + u_{i,t}$$
i : districts | t : year
c : crop category $\in$ {cereals,coarse cereals,cash,oilseeds, horticulture}

(D)
$$H_{i,t} = \beta_0 + \beta_1(gdp_{i,t}) + \beta_2(beds_{i,t}) + \beta_3(tap_{i,t}) + \beta_4(yield\_index\_gr_{c,i,t}) + u_{i,t}$$
i : districts | t : year
c : crop category $\in$ {cereals,coarse cereals,cash,oilseeds, horticulture}
gr : represents yearly, district-wise growth rate of the yield index for each crop category

(E)
$$H_{i,t} = \beta_0 + \beta_1(gdp_{i,t}) + \beta_2(beds_{i,t}) + \beta_3(tap_{i,t}) + \sum_c \beta_c(yield\_index\_gr_{c,i,t}) + u_{i,t}$$
i : districts | t : year
c : crop category $\in$ {cereals,coarse cereals,cash,oilseeds, horticulture}
gr : represents yearly, district-wise growth rate of the yield index for each crop category

(F)
$$H_{i,t} = \beta_0 + \beta_1(\log(gdp_{i,t})) + \beta_2(\log(beds_{i,t})) + \beta_3(\log(tap_{i,t})) + \beta_4(\log(yield\_index_{c,i,t})) + u_{i,t}$$
i : districts | t : year
c : crop category $\in$ {cereals,coarse cereals,cash,oilseeds, horticulture}

$$H_{i,t} = \beta_0 + \beta_1(\log(gdp_{i,t})) + \beta_2(\log(beds_{i,t})) + \beta_3(\log(tap_{i,t})) + \sum_c \beta_c(\log(\_yield\_index_{c,i,t})) + u_{i,t}$$

(G) i : districts | t : year

c : crop category $\in$ {cereals,coarse cereals,cash,oilseeds, horticulture}

Q4) Comment on the results in question 3 in line with the theoretical relationship between correlation coefficient and goodness of fit.

Q5) What could be a potential issue in including yield indices for all six crop categories together?

Q6) Is the relation between yield growth and health indicators similar across crop categories?

Data Description:

main.csv provides information on health indicators and agricultural yield indices for the following different crop categories (by consumption) as given by the Ministry of Statistics and Planning Implementation:

- The data source for all health indicators is the Health Management Information System, Ministry of Health and Family Welfare, Government of India.
- Data source for all the agricultural indicators is the Ministry of Agriculture and Farmers' Welfare, Government of India.

Horticulture (Vegetables, Fruits, Spices)	Pulses	Oilseeds	Cash Crops	Coarse Cereals	Cereals
Arecanut	Arhar/Tur	Groundnut	Castor seed	Bajra	Maize
Banana	Cowpea(Lobia)	Linseed	Cotton(lint)	Barley	Rice
Black pepper	Gram	Niger seed	Jute	Jowar	Wheat
Cardamom	Guar seed	Oilseeds total	Khesari	Other Cereals	
Cashewnut	Horse-gram	Rapeseed &Mustard	Mesta	Ragi	
Coconut	Masoor	Safflower	Sugarcane	Small Millets	
Coriander	Moong(Green Gram)	Soyabean	Tobacco		
Dry Ginger	Moth	Sunflower			
Dry chillies	Other Kharif pulses	Other Oilseeds			
Garlic	Other Rabi pulses				
Ginger	Other Summer Pulses				
Onion	Peas & beans (Pulses)				
Potato	Sannhamp				
Sesamum	Urad				
Sweet potato					
Tapioca					
Turmeric					

Note that the crop categories as followed by the Reserve Bank of India are:

(Please ignore this variable for now).

RBI Categories	Commercial	Food Grain
Crop Consumption Categories	Cash, Horticulture, Oilseed	Cereals, Coarse Cereals, Pulses

Calculating the Yield Index:

For a crop category, in a district, the way we calculate the yield index for a year is as follows:

Crop Type	Crop Name	Yield	Area	Yield*Area
Cereals	**A**	12	100	1200
	B	25	12	300
	C	33	15	495
	D	45	33	1485
	Sum	115	160	3480
Yield Index = (3480/160) = 21.75				

Data Decisions:

Note that we remove the data for banana and coconut from the analysis since the yields for these crops is not comparable with other crops in the same category.

Variable Description:

Variable Description	Variable Name (In the Data)
ROWID	rowid
Country	country
State LGD Code	statelgdcode
State	state
District LGD Code	districtlgdcode
District	district
Year	year
Name of the Crop	crop
Name of the Season	season
Area under production (in hectares)	areahectares
Production (in tonnes)	productiontonnes
Yield (Tonnes/ Hectare)	yieldtonneshectare
State - District - year ID	sdyid
State - District - Crop Name - Season - year ID	sdcsyid
Pregnant women registered for Ante Natal Care(ANC)	v1
Pregnant women registered for Ante Natal Care within first trimester	v2
Pregnant women received 3 Ante Natal Care (ANC) check-ups	v3

Pregnant women received the second dose of tetanus-toxoid vaccine (TT2) or booster	v4
Pregnant women received 100 Iron and Folic Acid (IFA) tablets	v5
Women having tested moderately anemic with hemoglobin(Hb)<11	v6
Women having tested with severe anemia with hemoglobin(Hb) and are being treated at an institution	v7
Home deliveries	v8
Home deliveries attended by doctor or nurse or ANMs trained as SBA	v9
Home deliveries attended by TBA or Dai non-trained as SBA	v10
Deliveries conducted at public institutions	v11
Percentage of women discharged in less than 48 hours of delivery (to total reported deliveries in public institutions)	v12
Institutional deliveries	v13
Total reported deliveries	v14
Percentage of institutional deliveries (to total reported deliveries)	v15
Percentage of safe deliveries (to total reported deliveries)	v16
Percentage of home deliveries (to total reported deliveries)	v17
Women received postpartum checkup within 48 hours of delivery	v18
Women received Post-Natal Care or Post Partum check-up between 48 hours and 14 days of delivery	v19
Percentage of women received post partum check up within 48 hours of delivery (to total reported deliveries)	v20
Percentage of women received a postpartum checkup or Post-Natal Care between 48 hours to 14 days of delivery	v21
Reported live births	v22
Percentage of total reported live births (to total reported deliveries)	v23
Reported still births	v24
Percentage of reported live births (to reported births)	v25
Newborns weighed at birth	v26
Newborns having weight less than 2.5 kg	v27
Percentage of newborns having weight less than 2.5 kg (to newborns weighed at birth)	v28
New borns breastfed within 1 hour	v29
Percentage of new borns breastfed within 1 hour (to total live births)	v30
Sex ratio at birth	v31
Percentage of male sterilization (to total sterilisation)	v32
Percentage of female sterilisations (to total sterilisation)	v33
Fully immunized children in the age group of 9 to 11 months	v34

Percentage of children with Polio in the age group of 0 to 5 years	v35
Percentage of children with Measles in the age group of 0 to 5 years	v36
Percentage of children with Diarrhea and Dehydration in the age group of 0 to 5 years	v37
Percentage of children with Malaria in the age group of 0 to 5 years	v38
Infant deaths reported	v39
Percentage of infant deaths due to Sepsis (to total reported infant deaths)	v40
Percentage of infant deaths due to Asphyxia (to total reported infant deaths)	v41
Percentage of infant deaths due to Low Birth Weight (LBW) (to total reported infant deaths)	v42
Percentage of infant deaths due to Pneumonia (to total reported infant deaths)	v43
Percentage of infant deaths due to Diarrhea (to total reported infant deaths)	v44
Percentage of infant deaths due to Fever (to total reported infant deaths)	v45
Percentage of infant deaths due to Measles (to total reported infant deaths)	v46
Percentage of infant deaths due to other causes (to total reported infant deaths)	v47
Mark for Deletion	v48
Merge ID	_merge
Crop Category (detailed above)	cropcategory
Crop Categories as followed by the Reserve Bank of India (RBI)	rbicat
State - District _Crop Category_Season - Year ID	sd_cc_syid
Yield*Area	yield_area
Sum (Yield*Area by SDCCSYID)	yield_area_cc_total
Sum (Area under production by SDCCSYID)	area_cc_total
Index (Yield_Area/Area)	index

Sample Regression Table:

Dependent Variable	Model 1
	Coefficient (SE)
Intercept	$\hat{\beta}_0$ (SE of $\hat{\beta}_0$)
Independent Variable 1 (e.g., Yield Index of Cereal Crops)	$\hat{\beta}_1$ (SE of $\hat{\beta}_1$)
N= ? | R squared = ?	

***, **, *** indicate the p-values.**

Note that different model specifications can be incorporated as additional columns in the same table.