Data Assignment 1

Name: Nakul Thureja
Roll No: 2020528

Ans 1.

Data Compilation

- I've used mysql to join the two tables using a left join query and exported them into csv files for the final dataset.
- I noticed that index (yield_index) for same year, district and crop is varying (for different seasons) which according to the definition in the assignment does not fit our needs so I again calculated the yield_index with constant values for the same district,year and crop category using the formula sum(yield_area_cc_total)/sum(area_cc_total) when year,district,crop category is same.
- After calculating the yield index I calculated the growth_rate with the formula y(t) - y(t-1)/y(t-1).

Data Cleaning

- First of all I removed all the columns from the dataset which we do not need for analysis.
- Secondly, I have removed all the null(NA) values in the data set for all the dependent and independent variables.

Further data cleaning has been done according to the needs of the question in R.
The final dataset is submitted as final.csv

Ans 2.

(A) Assumption: I have removed all duplicate values w.r.t. sdyid

	Sepsis	**LBW**	**Pneumonia**	**Diarrhea**	**Fever**	**Measles**
Mean	5.565131	17.64453	7.54585	1.888108	4.001121	0.181352
Median	2.9	16.4	4.2	0	0.9	0
Mode	0	0	0	0	0	0
S.D.	8.646767	14.12872	11.72055	7.321307	9.938436	2.845567

(B) Graphs are submitted in pdf form in the zip file
I have made the density plots for all variables after taking out the outliers using the range defined in part (C)

(C) For removing outliers I have removed all the data points where the dependent values do not lie in the range (mean - 3(standard deviation),mean + 3(standard deviation)).

(D) Correlation Coefficient (Assumption: I have removed all duplicate values w.r.t. sdyid,cropcategory)

I. Explanatory variables

	GDP	TAP	BED
Sepsis	0.1284693	-0.0731429	0.1200332
Lbw	0.2427935	0.1671283	0.05531895
Pneumonia	-0.2545924	-0.157847	-0.1786526
Diarrhea	-0.1227209	-0.08210439	-0.08023223
Fever	-0.2108153	-0.1668689	-0.1070053
Measles	-0.01502511	0.01343685	-0.03095847

II. Yield Index for each of the six crop categories

	Cash	Pulses	Cereals	Horticulture	Oil Seeds	Coarse Cereals
Sepsis	0.04243153	-0.04007136	0.03851157	-0.02484862	0.00978209	-0.01860705
Lbw	-0.1489391	-0.09653301	-0.1532917	-0.08489323	-0.0460869	-0.1627609
Pneumonia	-0.09738292	-0.00392524	-0.07424658	-0.05260997	-0.08338517	-0.09928921
Diarrhea	-0.00338107	0.0326424	0.00710508	-0.05042478	-0.02694352	0.01939885
Fever	-0.02515908	0.03288812	-0.04617684	-0.00066346	-0.07953971	0.00885643
Measles	-0.03824514	0.03744531	-0.01264069	-0.01981024	-0.01390755	-0.01004645

III. Yield Index growth rate for each of the six crop categories

	Cash	Pulses	Cereals	Horticulture	Oil Seeds	Coarse Cereals
Sepsis	-0.00761465	-0.02530641	-0.06490627	-0.00041239	-0.05256827	-0.07276482
Lbw	-0.00632138	0.00826632	0.01205284	0.02912058	0.04488454	0.02983381
Pneumonia	0.00201083	0.02601791	-0.02160198	0.01299574	0.02755668	-0.00756011
Diarrhea	0.02463909	-0.02020838	-0.00570033	-0.00796674	0.00061508	-0.02776198
Fever	0.01340119	0.01190884	-0.01359107	-0.02055875	-0.01386987	-0.01374464
Measles	-0.00448803	0.00861852	0.01972149	-0.00323752	-0.00618772	-0.01311499

Ans 3. Regression using Fever as our dependent variable

(A) Assumption: I have removed all duplicate values w.r.t. sdyid

Fever	**OLS estimates of coefficient**
Intercept	4.585e+00
GDP	-4.144e-08
Tap	-2.163e-02
Beds	8.351e-06
N = 3713	R squared = 0.06166

(B) Assumption: I have removed all duplicate values w.r.t. sdyid,cropcategory

Fever	**OLS estimates of coefficient**
Intercept	4.467e+00
GDP	-3.989e-08
Tap	-2.260e-02
Beds	9.145e-06
Yield Index (cash)	-4.565e-03
N = 3257	R squared = 0.06507

Fever	**OLS estimates of coefficient**
Intercept	4.358e+00
GDP	-4.119e-08
Tap	-2.213e-02
Beds	8.761e-06
Yield Index (pulses)	2.260e-01
N = 3593	R squared = 0.06458

Fever	**OLS estimates of coefficient**
Intercept	4.579e+00
GDP	-4.018e-08
Tap	-2.069e-02
Beds	8.551e-06
Yield Index (cereals)	-5.712e-02
N = 3700	R squared = 0.06033

Fever	**OLS estimates of coefficient**
Intercept	4.530e+00
GDP	-4.005e-08
Tap	-2.317e-02
Beds	9.103e-06
Yield Index (horticulture)	-2.611e-02
N = 3508	R squared = 0.06329

Fever	**OLS estimates of coefficient**
Intercept	4.682e+00
GDP	-4.143e-08
Tap	-1.988e-02
Beds	8.960e-06
Yield Index (oil seeds)	-1.444e-01
N = 3532	R squared = 0.06604

Fever	**OLS estimates of coefficient**
Intercept	4.094e+00
GDP	-3.840e-08
Tap	-2.159e-02
Beds	8.499e-06
Yield Index (coarse cereals)	2.543e-01
N = 3041	R squared = 0.06206

(C) Assumption: I have removed all duplicate values w.r.t. sdyid,cropcategory

Dependent Variables	**OLS estimates of coefficient**
Intercept	4.473e+00
GDP	-4.039e-08
Tap	-2.157e-02
Beds	8.866e-06
Yield Index(cash)	-3.973e-03
Yield Index(pulses)	1.065e-01
Yield Index(cereals)	-2.090e-02
Yield Index(coarse cereals)	7.777e-02
Yield Index(horticulture)	-2.164e-02
Yield Index(oil seeds)	-6.172e-02
N = 20631	R squared = 0.06316

(D) Assumption: I have removed all duplicate values w.r.t. sdyid,cropcategory

Fever	**OLS estimates of coefficient**
Intercept	4.105e+00
GDP	-3.886e-08
Tap	-2.081e-02
Beds	9.958e-06
Yield Index Growth Rate(cash)	1.186e-02
N = 3025	R squared = 0.06357

Fever	**OLS estimates of coefficient**
Intercept	4.400e+00
GDP	-4.086e-08
Tap	-2.081e-02
Beds	9.615e-06
Yield Index Growth Rate(pulses)	6.628e-02
N = 3371	R squared = 0.06504

Fever	**OLS estimates of coefficient**
Intercept	4.316e+00
GDP	-3.903e-08
Tap	-2.094e-02
Beds	9.178e-06
Yield Index Growth Rate(cereals)	-8.075e-02
N = 3457	R squared = 0.06033

Fever	**OLS estimates of coefficient**
Intercept	4.277e+00
GDP	-3.952e-08
Tap	-2.343e-02
Beds	9.441e-06
Yield Index Growth Rate(horticulture)	-1.616e-04
N = 3563	R squared = 0.06038

Fever	**OLS estimates of coefficient**
Intercept	4.579e+00
GDP	-4.018e-08
Tap	-2.069e-02
Beds	8.551e-06
Yield Index Growth Rate(oil seeds)	-5.712e-02
N = 3593	R squared = 0.06033

Fever	**OLS estimates of coefficient**
Intercept	4.202e+00
GDP	-3.858e-08
Tap	-1.914e-02
Beds	1.000e-05
Yield Index Growth Rate (coarse cereals)	-1.722e-02
N = 3088	R squared = 0.06135

(E) Assumption: I have removed all duplicate values w.r.t. sdyid,cropcategory

Dependent Variables	**OLS estimates of coefficient**
Intercept	4.416e+00
GDP	-4.006e-08
Tap	-2.186e-02
Beds	8.912e-06
Yield Index Growth Rate (cash)	-3.348e-03
Yield Index Growth Rate (pulses)	1.374e-01
Yield Index Growth Rate (cereals)	-6.618e-03
Yield Index Growth Rate (coarse cereals)	1.009e-01
Yield Index Growth Rate (horticulture)	-1.890e-02
Yield Index Growth Rate (oil seeds)	2.093e-02
N = 20631	R squared = 0.06313

(F) Assumption: I have removed all duplicate values w.r.t. sdyid,cropcategory and removed all the 0s

Fever	**OLS estimates of coefficient**
Intercept	32.16245
log(GDP)	-2.62288
log(Tap)	-0.45354
log(Beds)	1.57108
log(Yield Index (cash))	-0.07893
N = 3025	R squared = 0.1181

Fever	**OLS estimates of coefficient**
Intercept	34.25918
log(GDP)	-2.82040
log(Tap)	-0.43883
log(Beds)	1.70731
log(Yield Index (pulses))	0.60767
N = 3371	R squared = 0.1231

Fever	**OLS estimates of coefficient**
Intercept	31.81825
log(GDP)	-2.57073
log(Tap)	-0.45860
log(Beds)	1.48932
log(Yield Index (cereals))	0.28890
N = 3508	R squared = 0.1083

Fever	**OLS estimates of coefficient**
Intercept	31.70500
log(GDP)	-2.57133
log(Tap)	-0.46952
log(Beds)	1.52059
log(Yield Index (horticulture))	-0.05496
N = 3252	R squared = 0.1145

Fever	**OLS estimates of coefficient**
Intercept	33.030761
log(GDP)	-2.725162
log(Tap)	-0.432696
log(Beds)	1.643333
log(Yield Index (oil seeds))	0.002511
N = 3363	R squared = 0.1113

Fever	**OLS estimates of coefficient**
Intercept	31.33207
log(GDP)	-2.41767
log(Tap)	-0.51195
log(Beds)	1.32943
log(Yield Index (coarse cereals))	0.44067
N = 2892	R squared = 0.1131

(G) Assumption: I have removed all duplicate values w.r.t. sdyid,cropcategory and removed all the 0s

Dependent Variables	**OLS estimates of coefficient**
Intercept	32.40916
log(GDP)	-2.62671
log(Tap)	-0.45773
log(Beds)	1.55111
log(Yield Index(cash))	-0.06490
log(Yield Index(pulses))	0.45711
log(Yield Index(cereals))	0.03631
log(Yield Index(coarse cereals))	0.40132
log(Yield Index(horticulture))	-0.09002
log(Yield Index(oil seeds))	0.01535
N = 19310	R squared = 0.1141

Ans 4.

- Correlation Coefficient is the measure of how two variables are strongly correlated
- R^2 is the measure of how the dependent variable is being explained by the independent variable

R^2 = cor(y,y-hat)^2 = cor(y,x)^2

Part A:
R^2 = 0.061
cor(y,x)^2 = (0.21)^2 + (0.16)^2 + (0.10)^2 = 0.079

Part B:
R^2 = 0.06507
cor(y,x)^2 = (0.21)^2 + (0.16)^2 + (0.10)^2 + (0.02)^2 = 0.080
Only for cash crops similar analysis can be proved for every other crop category.

Part C:
R^2 = 0.06316
cor(y,x)^2 = (0.21)^2 + (0.16)^2 + (0.10)^2 + (0.02)^2 + (0.03)^2 + (0.04)^2 + (0.07)^2 = 0.087

Part D:
R^2 = 0.06357
cor(y,x)^2 = (0.21)^2 + (0.16)^2 + (0.10)^2 + (0.01)^2 = 0.079
Only for cash crops similar analysis can be proved for every other crop category.

Part E:
R^2 = 0.06316
cor(y,x)^2 = (0.21)^2 + (0.16)^2 + (0.10)^2 + (0.01)^2 + (0.01)^2 + (0.01)^2 + (0.02)^2 + (0.01)^2 + (0.01)^2 = 0.080

Note -> values are not exactly equal since independent variables are also correlated themselves.
cor(gdp,tap) = 0.1631174
cor(gdp,beds) = 0.7976767

Ans 5.

A potential issue in including yield indices for all six crop categories together could be the fact that each crop category has a different purpose and therefore affect the health indicators very differently and also all these crop categories are correlated as to grow crops of the crop category we need to reduce the growth of other categories.

For ex- cash crops are grown majorly for profits and export which should lead to increase in GDP therefore increase in Health Facilities, whereas pulses is not a cash crop and therefore affect the indicators differently.

We can see that it is better to take different models for different crop categories as shown in part B as the R^2 of the model in part (B) is more than the R^2 of the model in part C.

Ans 6.

The relation between yield growth and health indicators is somewhat similar across crop categories for dependent variable Fever,as for 4 out of the 6 crop category the value of Beta ols is negative in part D (i.e. the health indicators decrease with increase in growth rate), the Beta ols is positive in part D for cash and pulse crop category.

<u>Notation of similarity:</u>

Positive: the value of health indicators increase with increase in growth rate for particular crop category.

Negative: the value of health indicators increase with decrease in growth rate for particular crop category.

For more information regarding this:

We can look at the table in part D with correlation coefficient for dependent variable and growth rate for various crop categories.

We can see that major values in the table are negative which implies the same correlation as said above but it is very hard to say this with certainty as each crop category is grown for different purposes and has different selling prices and profit margins and therefore could affect the health parameters quite differently.