ECO ASSIGNMENT-01

Zip file includes 6 different R scripts for all 6 questions, and one csv file for question1.
NOTE: Some of the details are mentioned here in the pdf.Also the R scripts are commented and can be referred.

Don't delete/ remove data variables which are created with each R file.
Some of the scripts use data structures which were formed in the previous R scripts.

Question 1:

To fill the columns for gdp, tap and beds. Three separate csv files were made containing data about the variable for a state/district and the year for which the data is.
Csv files: beds, gdp and tap.
NOTE : GDP csv file data unit is in lakhs.
To fill the data in the columns a double loop is used to fill where the particular index matches the variable.
The original csv file is loaded to variable main where it is converted into a dataframe.
Main[i, j] is used in a double loop where i = row, j = column.
3 double for loops are used for given three variables.

Question 2:

a)
A function summarizeDV is used to print mean, median, mode and sd of the dataset.
As there is no base function to calculate mode of the dataset, so to calculate the mode calcMode function is implemented.
na.rm = TRUE as a parameter is used to ignore NA values in the dataset.
After calculating the values are then printed.

b)
In plotting the histogram,
abbreviations are used for the 6 variables.These are used interchangeably.
v40:sepsis
v42:lbw
v43:measles
v44:diarrhoea
v45:fever
v46:measles

Here for a particular variable for all the years and the seasons the histograms are combined. So for each variable there are two plots one for year based and the other one season based. The graphs are plotted using ggplot2

Ex:

```
ggplot(v40_year, aes(x=v40_year$data.v40, fill = year))+
  geom_histogram( color='#e9ecef', alpha=0.6, position='identity') +
xlab("sepsis")+ggtitle("Histogram for v40 (year)")

ggplot(v40_season, aes(x=data.v40, fill = season))+
  geom_histogram( color='#e9ecef', alpha=0.6, position='identity') +
xlab("sepsis")+ ggtitle("Histogram for v40 (season)")
```

C)
The function boxplot.stats(data)$out can be used to find any outliers in the dataset.
So this function is used to find the outliers for each of the 6 variables.

D)

cor function when supplied with a df makes a correlation matrix with the columns of df provided.

For part 2)
6 different cor matrix are made for 6 different types of crop_categories.
generally here cropcategory_df.index represents the yield index for the given crop category. pulse_df.index represents the yield index.

For part 3)
data_2 is a separate df which will be used to add a new column yield_index_growth. It is formed from some columns of data df.
A double loop is used to calculate yield_index_growth for a particular year for a given crop.

NOTE: for 2011 there is no data of yield_index for previous years so yield_index_growth cannot be calculated or it will be a NaN value. Instead of this here yield_index_growth for 2011 is taken to be the value of its yield_index.

Question- 3

- Institutional deliveries as a percentage variable: v15 to be taken as a health indicator.
- for C,E,G parts six separate models had to made for different crop categories.
- Lm function is used t fit linear models to data frames in the R Language and the function summary(linear_model) is used to represent regression results in a table.
- Refer R script ECO_Q3.R(commented).

lm(formula, data)
the formula is y ~ x1 + x2 + x3 …..
Where y is a dependent variable and x1,x2,x3…. are independent variables.

Question 4:

```
# NOTE : Dont delete/ remove data variables which are created with each R file.
# Some of the scripts use data structures which were formed in the previous R scripts.
```

We know goodness of fit or R^2 is given by
and correlation coefficient is root(R^2) or R.

To verify this relationship, in a multiple regression model we need to find correlation between, the fitted values given by the model and the orignal values.
R^2 value for a model can be found out using summary(model_name).

```
# proving this for model_a
# model_a R^2 = 0.2188
cor(model_a$fitted.values, model_a$model$v15)
# comes out to be 0.4677898. CHECK: root(0.2188)  = 0.46776 (almost equal)
# so the theoretical relation between goodness of fit and correlation coefficient can be
# proven practically by finding the correlation and goodness of fits of the MLRM.

# model_b R^2 = 0.2204
cor(model_b$fitted.values, model_b$model$v15)
# CHECK: root(0.2204)  = 0.4694678 (almost equal)


# this can be done for all the 21 models and similar results are obtained.
```

```
cor(model_c_1$fitted.values, model_c_1$model$v15)
# Multiple R-squared:  0.216
# cor = .4647398  :  .4647398^2 = 0.2159831
```

```
cor(model_d$fitted.values, model_d$model$v15)
# Multiple R-squared:  0.2993
# cor = 0.5470577  :  0.5470577^2 = 0.2992721
```

Question -5

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

The issue is that these crops are different, all the crops have different needs and conditions to grow fully.
Some crop types might require more land than the others. Some crop types might be more vulnerable to pests.
Some crop types could also be highly dependent on large amount of fertilizers to grow.
Crops types might have different harvest indices.
As the crop types are not the same they require different conditions, which results in dissimilarity in the yield index. For example: yield index for Pulses would be different than that for Horticulture crops, combining the data of yield_index of the two crop types wouldn't make
sense as the range of yield_index for both the crop types could be very different.
This creates inconsistency in data, and the regression model wouldn't make much sense.
It would be better if we model six regressions, one for each crop category.

```
# making linear models with independent var as yield_index_growth and dependent variable as a health indicator.
pulse_df = subset(main, main$cropcategory == "Pulse")
model_q5_1 = lm(v15 ~ yield_index_growth, data = pulse_df)
summary(model_q5_1)

cash_df = subset(main, main$cropcategory == "Cash")
model_q5_2 = lm(v15 ~ yield_index_growth, data = cash_df)
summary(model_q5_2)

Cereal_df = subset(main, main$cropcategory == "Cereal")
model_q5_3 = lm(v15 ~ yield_index_growth, data = Cereal_df)
summary(model_q5_3)

Oilseed_df = subset(main, main$cropcategory == "Oilseed")
model_q5_4 = lm(v15 ~ yield_index_growth, data = Oilseed_df)
summary(model_q5_4)

Horticulture_df = subset(main, main$cropcategory == "Horticulture")
model_q5_5 = lm(v15 ~ yield_index_growth, data = Horticulture_df)
summary(model_q5_5)

Coarse_Cereal_df = subset(main, main$cropcategory == "Coarse Cereal")
model_q5_6 = lm(v15 ~ yield_index_growth, data = Coarse_Cereal_df)
summary(model_q5_6)
```

Q-6

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

No.The relation between yield growth index and health indicator is somewhat dissimilar across crop categories. This can be confirmed by making models with the dependent variable as v15 which is a health indicator and independent variable being yield_index_growth for a particular crop.

By analyzing these said models, 5 of them have a positive slope and one of them has a negative slope (Horticulture). Also the slope in absolute value varies for all 6 crops.
though the intercepts of the regression models are all pretty close and in the range [87, 88].

R script:

```
# finding out the models to compare v15 to yield index growth for all 6 crop types.

pulse_df = subset(main, main$cropcategory == "Pulse")
model_q6_1 = lm(v15 ~ yield_index_growth, data = pulse_df)
summary(model_q6_1)

cash_df = subset(main, main$cropcategory == "Cash")
model_q6_2 = lm(v15 ~ yield_index_growth, data = cash_df)
summary(model_q6_2)

Cereal_df = subset(main, main$cropcategory == "Cereal")
model_q6_3 = lm(v15 ~ yield_index_growth, data = Cereal_df)
summary(model_q6_3)

Oilseed_df = subset(main, main$cropcategory == "Oilseed")
model_q6_4 = lm(v15 ~ yield_index_growth, data = Oilseed_df)
summary(model_q6_4)

Horticulture_df = subset(main, main$cropcategory == "Horticulture")
model_q6_5 = lm(v15 ~ yield_index_growth, data = Horticulture_df)
summary(model_q6_5)

Coarse_Cereal_df = subset(main, main$cropcategory == "Coarse Cereal")
model_q6_6 = lm(v15 ~ yield_index_growth, data = Coarse_Cereal_df)
summary(model_q6_6)
```