Contribution of the Team members:

We all worked together on each question, however Karim focused mainly on the design and implementation of the database and statistical graphs, Philipp Merz worked on the data preparation and import of the dataset, and Catarina focused on the sql questions and the import of the data as well.

Description of our work.

It follows the main points of the project description starting with point 2.

2) Made an Entity Relation Diagram(ERD) with two tables (country, happiness) with a 1 to many relation between them in which happiness containing a foreign key to country where it is also NOT NULL, also made the relational schema and created the tables using sql in the database.

3) To prepare the data some countries have to be renamed or removed from the happiness datasets as they do not appear in the countries_of_the_world file. To find all unique countries we imported all wh_happiness files and concatenated them to one dataframe in a first step. To extract the columns of interest later, we also merge some columns that mean the same but are named differently in the different years. For example in some years the wh_happiness files contained the column 'Country', while it was named 'Country or region' in other years. We use the pandas function combine_first that adds the missing values to the first named column, such that column 'Country' finally contains all countries of all years. To find the unique values we copy the dataframe and drop all duplicates. As the countries_of_the_world file lists all countries with a space at the end, we further remove this space to enable an easier comparison. We continue by comparing the unique countries of the wh_happiness files with the countries_of_the_world file by using a simple for loop. For each country that is not found in the countries_of_the_world file the user has two options. First, the country can be renamed or second, it can be removed from the dataset. The script also prints proposals that were found in the countries_of_the_world file to ease the decision for the user. However, for some countries (e.g. Ivory Coast) we cannot find proposals even though they exist in the countries_of_the_world file (as Cote d'Ivoire). So, be careful when deciding on the removal of a country. Both actions are then changed in the wh_happiness dataframe. Since, this comparison needs to be completed only once we finally store the data in separated csv files. We stored this files in the folder new_files in case they are needed.

4)

a) To load the data into our database we use the approach presented in class. First, we create the connection to the database. Second, we delete all dependent tables from the database using the DELETE command. Further, we import the relevant file into a dataframe to extract the columns of interest. We transform the dataframe to a nested list using the tolist() command. This is a crucial step for the insertation of the data in the database. We update the database tables using the INSERT INTO command. We finally execute the sql query. However, Western Sahara always produced an out of range error, which is why we eclude this country from the database.

b) The import of the happiness data in the database follows the same logic as 4a). Note that we always import only one year that may be given by the user when running the script.

5) The queries are saved in question_x.sql, x being each question (including 2 extra questions). We also added a pdf file with all the queries and showing the output table for easier reading called sql.pdf

6) Made some sql queries with Python, in which we used them to produce the statistical graphs to infer more information from the database where firstly found the average of happiness score for each country over the years, then produced the graphs which are by order are the following:

The number of happiness score in which we found the most frequent happiness score which is a little under 6, then a scatter plot to show linear regression between GDP and happiness score and concluded that there is a strong positive correlation between GDP and Happiness Score, which means as DGP go higher so does the happiness score, then did a linear regression between GDP and happiness score to further prove this correlation, after that a histogram that shows the frequency of GDP andmost of the countries have low GDP and very few countries or one have GDP higher than 50000, tried to find the correlation between GDP and literacy and found that this is a nonlinear relationship between GDP and Literacy as well as correlation between GDP and infant mortality, after that a heat map between GDP, literacy and infant mortality and lastly a histogram to show the frequency of regions in the database.