Report

a)

All lines of the code are commented but for a brief overview of the process:

1. Data is loaded from csv files
2. Missing values are handled
3. Dataframes are divided into 2 groups: those that have 'Country' as a column (i.e. df_country, df_temp, df_prec, df_land, df_density, df_quality) and those that have 'Abr' as a column (i.e. df_invest, df_main)
4. Each of the groups mentioned above are merged together first
5. The resulting dataframes from step 4 are merged once again to form the final dataset
6. Several issues with string formats are found and fixed in the final dataframe
7. The 'Country Name', 'Abr' columns are remover per instructions of the project

b)

The feature most correlated with 'Road Quality' is 'Avg Temperature' which has a negative correlation of magnitude 0.34.

Other features might not be very useful since they have a low correlation with Road Quality.

All the codes are explained by their comments, but for a brief overview:

1. Using the .corr() methos, we can calculate the correlations between all numeric columns of the dataframe.
2. The correlation dataframe is plotted as a heatmap to help with visualisation.
3. The correlations between 'Road Quality' and all other features are printed to find the maximum absolute value.
4. Remove unnecessary columns (low correlation)
5. Reset index from 0 to 579
6. Save the resulting dataframe to a csv file named data.csv

c)

The best model is model 4 because it has the lowest test (and train) errors. Model 5 is overfitting since the train error is going down but the test error is going up.

The training error is 1.16% and the testing error is 2.99%

d)

The main problem is that there are a lot of missing values in the data but, nevertheless, the Random Forest Regressor model still performs really well (testing accuracy of 97%)!

To improve the model, we can use k-fold cross-validation to choose an even better hyperparameter!