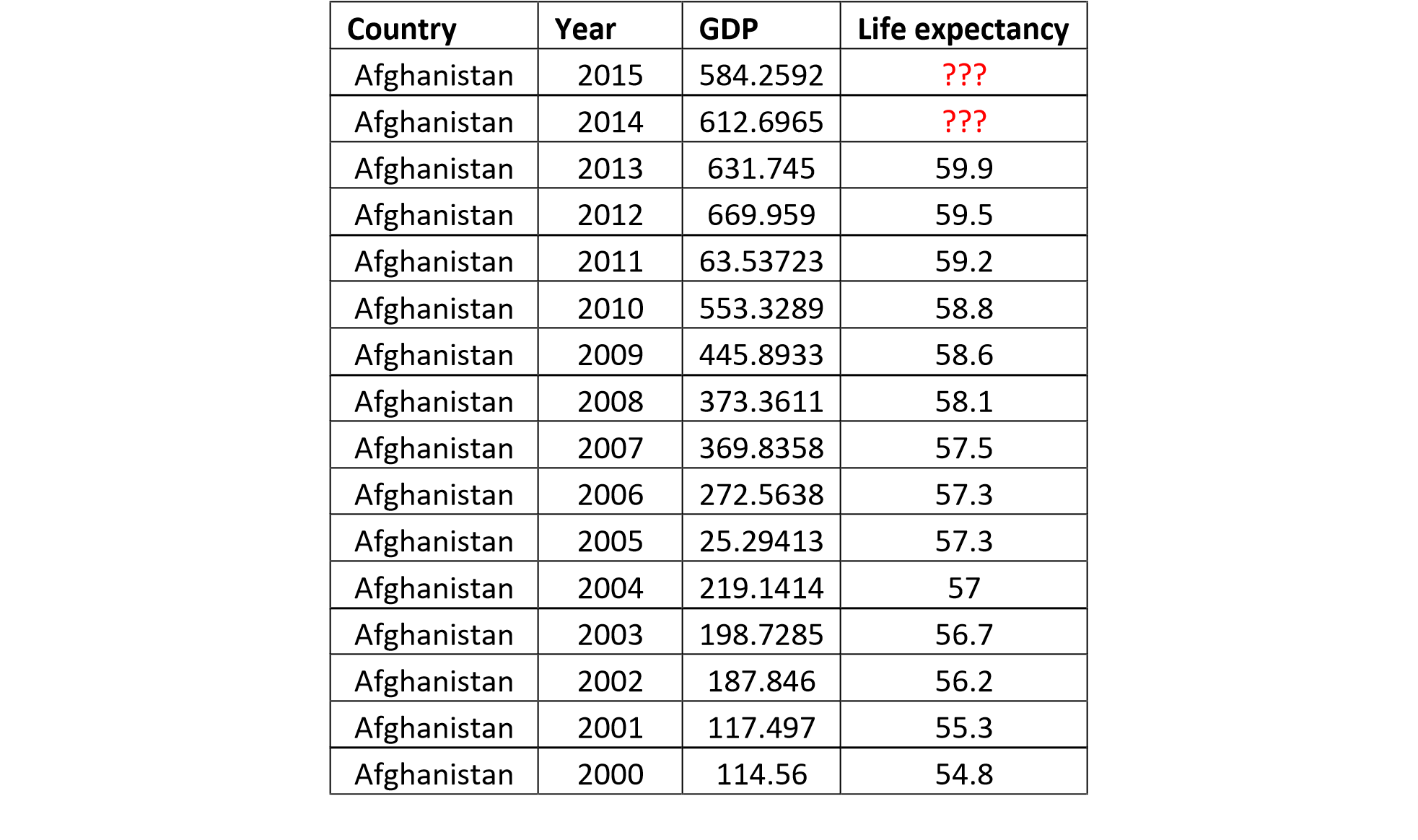
Assignment 2 - Alex Li

**Task 1**:

In this task, we will use different kinds of models to explore the relationships between economic status and life expectancy. For Afghanistan for instance, as the following table shows, we can use older data (from 2000 to 2013) to train models and use the trained models to predict life expectancy of 2014 and 2015. The model input can be GDP number and the model output will be life expectancy for that year.



Please train 4 functions, Linear Function, Quadratic Function, Cubic Function, and Quartic Function, to fit this data (only using Afghanistan data), and then calculate RMSE and R2 scores. Please fill the following table:



-1.102

0.539

0.519

0.366

0.243

-0.471

-2.198

-0.815

3.060

0.878

4.063

0.897

2.756

1.030

3.294

1.126

**Note to Grader: The reason for negative R2 values is that when you go to higher-degree polynomials, the model might fit the training data too closely (overfitting) but fail to generalize to the test data, leading to negative R2 values.**

Please submit your code (named *calculate\_Afghanistan.py*). Please explain which model can be the best to predict this small dataset? why?:

**The Quadratic Model (Degree = 2) is the best choice** for predicting this small dataset because it has the lowest test RMSE of 2.756, indicating the most accurate predictions on unseen data compared to the other models. It also has a good balance between simplicity and expressiveness, capturing the underlying curve in the data better than the linear model (Degree = 1) while avoiding the major overfitting issues seen in more complex models like the cubic (Degree = 3) and quartic (Degree = 4) ones. Although its test R2 value is still negative, indicating chance of overfitting, the quadratic model still generalizes better and demonstrates the most reliable performance among all the options, making it the best fit for this dataset.

**Task 2**:

Please repeat this process for all the countries in this dataset. Then, you can average the RMSE and R2 scores for all the developing and developed countries. Please fill the following table:

RMSE Scores:



2.587

2.404

1.980

2.432

1.793

2.670

1.647

2.603

1.603

5.283

1.192

2.874

1.312

2.428

1.444

1.729

R2 Scores:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Developing Countries** | **Linear Function** | **Quadratic Function** | **Cubic Function** | **Quartic Function** |
| Training Data  (2000 - 2013) | 0.240 | 0.473 | 0.547 | 0.622 |
| Testing Data (2014 and 2015) | -390.573 | -320.929 | -572.513 | -17391.211 |

-385.314

-414.995

-294.317

-208.615

0.410

0.370

0.279

0.138

**Developed**

**Countries**

**Linear**

**Function**

**Quadratic**

**Function**

**Cubic Function**

**Quartic**

**Function**

Training Data

(2000

-

2013)

Testing Data

(2014

and

2015)

Please submit your code (named *calculate\_all\_country.py*). Please explain which model(s) can be the best to predict developing and developed countries; why?:

ANSWER HERE

**Task 3**:

For this task, we will use 5 variables - Adult Mortality, Alcohol, BMI, GDP, Schooling – to build regression models (Multiple Linear Regression) to predict the life expectancy of the target country for a specific year, e.g., use a model to predict “*Libya’s life expectancy in year 2010*”. We can train two different models (developing country model and developed country model) to predict the data. Similarly, we can use older data (from 2000 to 2013) to train models and use the trained models to predict life expectancy of 2014 and 2015.

Please fill this table (for testing with 2014 and 2015 data):



Please submit your code (named *calculate\_regression.py*). Comparing developing and developed countries (two models that you build), can you find some interesting results?:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Task 4**:

For task 3, we used the Linear Regression model to address the prediction problem. Please tell us the limitation(s) of the model, and can you improve it?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Submission: Your code files and a PDF file (containing your solution for tasks and the results to report).