Task 1:

RMSE Scores (for 2014 and 2015 data):

	Linear	Quadratic	Cubic	Quartic
Training	1.067259118407	0.9987073	1.02295269	1.06304244
Testing	4.2690530	4.152741	4.2368845	4.374525657

R2 Scores (for 2014 and 2015 data):

	Linear	Quadratic	Cubic	Quartic
Training	0.4788209985	0.5436231	0.521195458	0.482931156
Testing	-1.8027395	-1.65209695	-1.7606597987	-1.94294113

Please explain which model can be the best to predict this small dataset? Why?:

We want R2 to be as close to 1 as possible and for RMSE we want to be as close to 0 as possible because we want the error to be as close to 0 as possible and we want to fit to be as close to 1 as possible. In this case, quadratic seems to have the lowest RMSE and the closest R2 score to 1. Since we are only using one feature, GDP, to predict life expectancy, it would make sense that our model would use a simpler function for predicting it.

Task 2:

RMSE

Developing	**Linear**	**Quadratic**	**Cubic**	**Quartic**
Training Data	1.72948685183	1.64553374372	1.618137832	1.61873601971
Testing Data	2.58741365457	2.52952122	2.5993915775	2.766231888
Developed	**Linear**	**Quadratic**	**Cubic**	**Quartic**
Training Data	1.980129369	1.95568669	1.9456897845	1.9445265106
Testing Data	2.40421912699	2.417707779	2.422x8846976	2.42722553888

R2

Developing	**Linear**	**Quadratic**	**Cubic**	**Quartic**
Training Data	0.2399223528	0.3058872754	0.33042716641	0.3329767169
Testing Data	-390.57301108	-356.1508296	-372.418316887	-429.438537755

Developed	**Linear**	**Quadratic**	**Cubic**	**Quartic**
Training Data	0.13755907547	0.1551672217	0.16043733947	0.1585641804
Testing Data	-208.6146107	-243.106471846	-259.330875619	-274.85061179

Please explain which model(s) can be the best to predict developing and developed countries; why?:

On average, quadratic is still the best model that can predict developing and developed countries because we want R2 to be as close to 1 as possible and for RMSE we want to be as close to 0 as possible because we want the error to be as close to 0 as possible and we want to fit to be as close to 1 as possible. In this case, quadratic seems to have the lowest RMSE and the closest R2 score to 1. Like task 1, we are still using the features, so keeping the function simpler seems to still be better in this case.

Task 3:

	RMSE	R2
Developed Training	2.989008282146381	0.4306814377702445
Developed Testing	4.101801853208987	-0.03622364892985641
Developing Training	5.224095831079608	0.679894792454102
Developing Testing	3.78443486148735	0.7693129604973308

	Adult Morality	Alcohol	BMI	GDP	Schooling
Developed	-2.57913886e-02	-3.80884115e-01	-5.31020372e-03	4.42521225e-05	5.85475321e-01
Developing	-2.85063506e-02	-1.69782025e-01	8.80479377e-02	7.92917522e-05	1.30176233e+00

Comparing developing and developed countries (two models that you build), can you find some interesting results?:

It is interesting that the coefficients of all the features between both developed and developing countries are pretty low. They also both seem to have a really high RMSE, which is concerning and tells me that this model is not that accurate. However, for developing countries, both testing and training, they do have a pretty high R2 score (close to 1), indicating that this model may be

a good fit. This likely means that we can explain a lot of the variance in the response variable, but it's still not able to make that accurate of predictions. This could be due to overfitting.

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?

One limitation of the linear regression model is that it is very sensitive to outliers which leads to very unpredictable results and can possibly explain the high RMSE values in all of the data. Another limitation is that it assumes that the relationship between the independent variable and the dependent variable is linear. If this is not the case, then the model will have very poor accuracy. In order to improve the model, we can use regularization, such as ridge or lasso regression, to combat any overfitting problems, like with the good fitting model for developed countries, but with poor accuracy. If we can determine that our independent and dependent variables form a nonlinear relationship, we can transform the data into a linear relationship through a mathematical transformation to improve its accuracy.