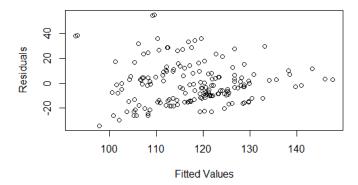**Problem 1**

Data preparation

- Latitude and longitude were shifted 90 so the boxcox would not get affected, it doesn't matter if we do so from the start for all the problems.
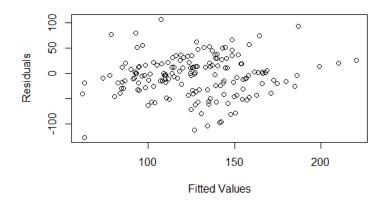- Data was split into 80/20 for the whole problem

Part A

The latitude model had an R-Squared in the training data equal to 0.32, and in the testing data equal to 0.29, while mean square error over the testing data was 269.79.

While the longitude had an R-Squared in the training data equal to 0.37, and in the testing data equal to 0.36. Mean square error over the testing data was 1619.16.
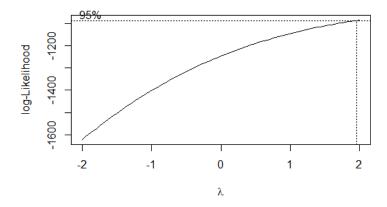

Residuals against values for latitude regression

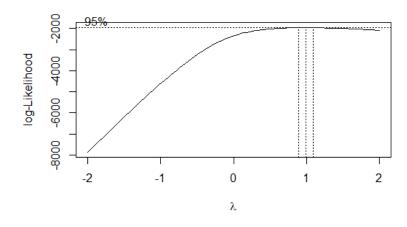*Residuals against values for longtitude regression*

Part B

For the BoxCox Lambda selection over the latitude, I applied the boxcox function and it gave me best lambda equal to 2 as follows:

I then applied the transformation over the data and built a regression model over the transformed data, then converted the data back to calculate the R-Squared for training which was 0.29. As for testing data, R-Squared was 0.25 after reconstructing the data. MSE over the testing data was 273.
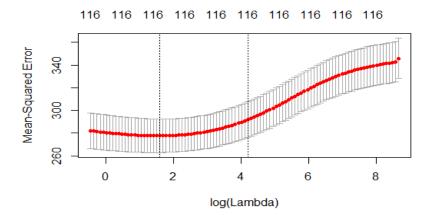
As for longitude regression, best lambda was 1 as shown in the figure below, so it doesn't make since to try it as there's no actual transformation for the data, just shifting.

Decision is not to use BoxCox furthermore since it hasn't added any value in neither the R-Squared nor the MSE.
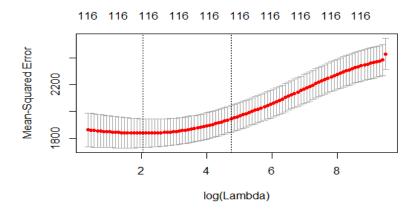
Part C.1 (Ridge regularization)

For the case of latitude regression with a ridge regularization, min lambda was 4.94 (deducted from the below figure), and it resulted in R-Squared over the training equal to 0.21, and over the testing equal to 0.08, and MSE equal to 313.97

Lambda selection and MSE over training data for Ridge Latitude regression

For the case of longitude regression with a ridge regularization, min lambda was 7.69 (deducted from the below figure), and it resulted in R-Squared over the training equal to 0.28, and over the testing equal to 0.12, and MSE equal to 1875.2.
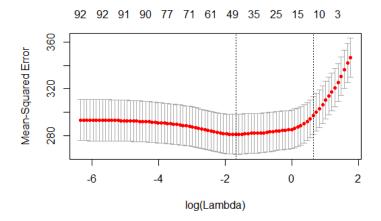

Lambda selection and MSE over training data for Ridge Longtitude regression

Conclusion is that regularization is not better than the original regression in neither latitude nor longitude for the different tried metrics.
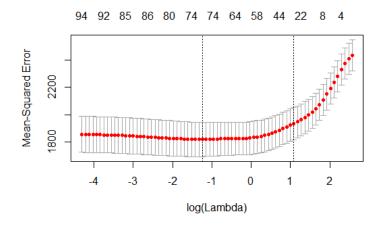
Part C.2(Lasso regularization)

For the case of latitude regression with a lasso regularization, min lambda was 0.18 (deducted from the below figure), and it resulted in R-Squared over the training equal to 0.26, and over the testing equal to 0.06, and MSE equal to 314.5 and number of non-zero coefficients for minimum lambda is 46

Lambda selection and MSE over training data for Lasso Latitude regression

For the case of longitude regression with a lasso regularization, min lambda was 0.28 (deducted from the below figure), and it resulted in R-Squared over the training equal to 0.33, and over the testing equal to 0.12, and MSE equal to 1835.1 and number of non-zero coefficients for minimum lambda is 80

Conclusion is that regularization is not better than the original regression in neither latitude nor longitude for the different tried metrics.

Parc C.3

Here are the table that represents different numbers for Elastic net regularization against different alpha values.

Regression Task	R-Squared Training	R-Squared Testing	MSE Over Testing	# of Coefficients
Latitude	0.24	0.06	313	74
Longitude	0.31	0.15	1765	102

Metrics for Elastic net regularization with alpha equal to 0.25

Regression Task	R-Squared Training	R-Squared Testing	MSE Over Testing	# of Coefficients
Latitude	0.26	0.05	316	70
Longitude	0.31	0.12	1828	94

Metrics for Elastic net regularization with alpha equal to 0.5

Regression Task	R-Squared Training	R-Squared Testing	MSE Over Testing	# of Coefficients
Latitude	0.24	0.06	312	56
Longitude	0.26	0.11	1853	73

Metrics for Elastic net regularization with alpha equal to 0.75

Conclusion is that regularization is not better than the original regression in neither latitude nor longitude for the different tried metrics.

Problem 2

Data Preparation

I am using the bigger file with more features, I removed the first row and the IDs column. The data was split into 80/20 so I can test with unseen data.

Regression

Regression training was always executed with training set and then error rate was calculated on the testing set.
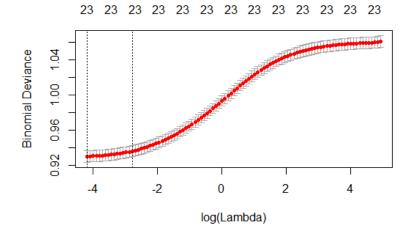
I tried different methods for regression, no regularization, where glmnet was used (CV is not used as there are no parameters to tune), and then the three regularization techniques we learned.

For regularization techniques, I am using lambda min as it gives me better results, I think it's not overfitting the data as the data is big enough, so no need for lambda.1se
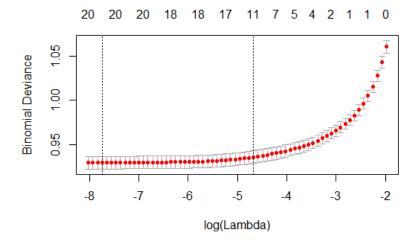
Technique	Error Rate (over testing set)	R-Squared
No regularization	19.13	N/A
Ridge	19.5	0.12
Lasso	19.16	0.13
Elastic net (0.5)	19.18	0.13

Conclusion, it doesn't seem that regularization made a big difference in the error rate for this testing set, difference is small though and thus I'd prefer applying it as it's less susceptible for overfitting.
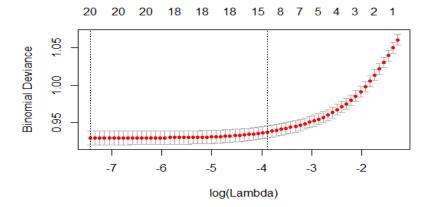
Finally, below are different deviance plots for the different regression techniques.

Deviance against Lambda for Ridge regression

Deviance against Lambda for Lasso regression

Deviance against Lambda for elastic net regression (alpha=0.5)