Report

For part A and B, the code was straightforward and we made use of standard linear Regression method and found the correct outputs and weights

For 1.1.b The best regularisation factor λ came out to be $\lambda = 10$ in the ridge regression

$$L(w : X, Y) = \|Y - Xw\|_2^2 + \lambda \|w\|^2$$

Part C required some more indepth analysis.

I have used 273 features and given below is more details into creating the features

(A) Dropping Identical (one-one mapped features)

Multiple features by the name <nam>_code

and <name>_description had a one on one mapping

= We dropped all <name>_code

(B) FEATURE TYPE ANALYSIS

There were mostly categorical features with 2 real valued features

with 2 real valued features

Length of Stay Birth weight

Note ⇒ A features looked categorical but still their mapping defined a total order

For example APR Severity of Illness Description
APR Risk of Mortality

C) ONE HOT ENCODING

I however didn't proceed with having an one hot encoding since there were multiple features with large number of categories. This allowed us to escape the CURSE OF DIMENTIONALITY

Additionally one-hot-encoding is made on viewed classes so if we train it on a subset, there is a chance that we might miss some labels and make a lower dimentional one hot encoding.

D) POLYNOMIAL FEATURES

From the above features, we ge polynomial features

$$\phi_{ij} = x_i x_j \quad \forall\, i,j \in \text{no of prev features}$$

This is then concatenated with previous features

E) LASSO FEATURE SELECTION

Lasso Regression using λ is done

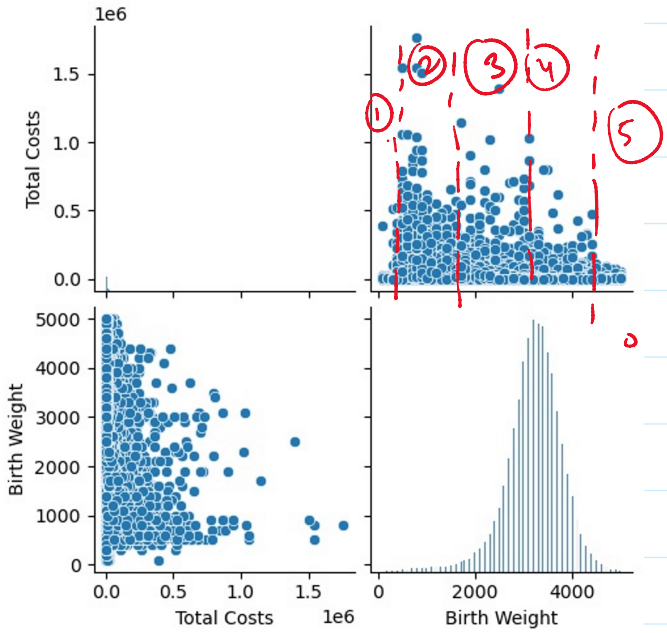
λ was found using 10-fold validation
(although for making feature-selection.py
I have directly mentioned $\lambda = 0.003$

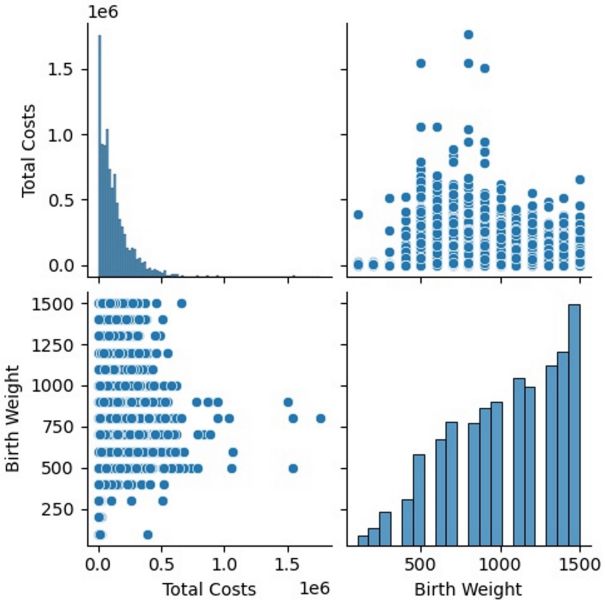
The coefficients that were 0 have been ignored

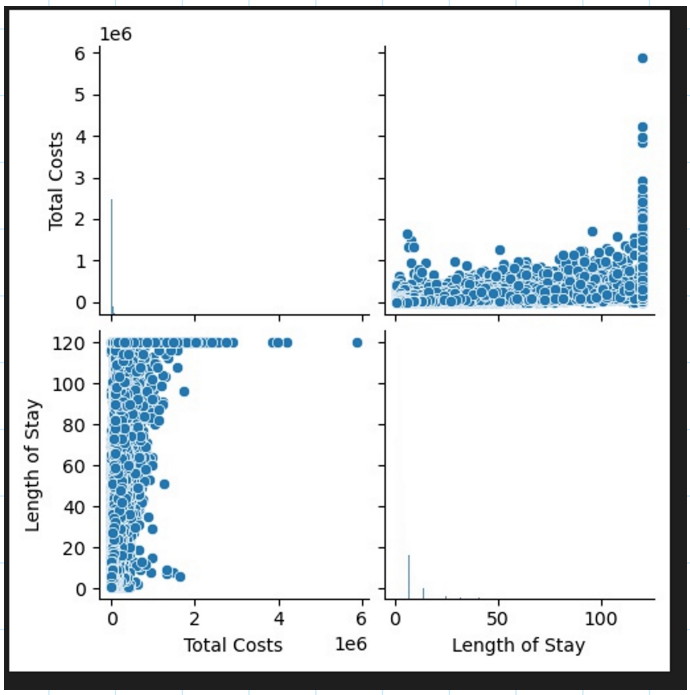
E) DOMAIN SPICIFIC FEATURES

It came to my notice that there was a short increase in cost when the weight of the child was below 1000 (typically). This could be attributed to certain diseases which would require more treatment thus increasing the cost

So I tried to create hand crafted categories by priliminary data analysis.







I tried the same for length of stay

Also we tried to find the maximum correlation b/w features and total cost

The top 10 values were

```
(Pdb) abs(data[data["Birth Weight"] ==0 ].corr()["Total Costs"]).nlargest(10)
Total Costs                            1.000000
Length of Stay                         0.667088
APR Severity of Illness Description    0.249514
APR Medical Surgical Description       0.228554
APR Risk of Mortality                  0.206154
Patient Disposition                    0.115013
Operating Certificate Number           0.088694
Age Group                              0.082672
Zip Code - 3 digits                    0.054797
Gender                                 0.052908
Name: Total Costs, dtype: float64
```

when birth weight = 0

```
(Pdb) abs(data[data["Birth Weight"] !=0 ].corr()["Total Costs"]).nlargest(10)
Total Costs                            1.000000
Length of Stay                         0.861056
Birth Weight                           0.355507
APR Medical Surgical Description       0.334788
APR Severity of Illness Description    0.247343
APR Risk of Mortality                  0.197094
CCS Procedure Description              0.087210
Type of Admission                      0.060246
Patient Disposition                    0.056354
Race                                   0.049308
Name: Total Costs, dtype: float64
```

when birth weight ≠ 0

After all the considerations I had **273** features left and I used them to do the regression.