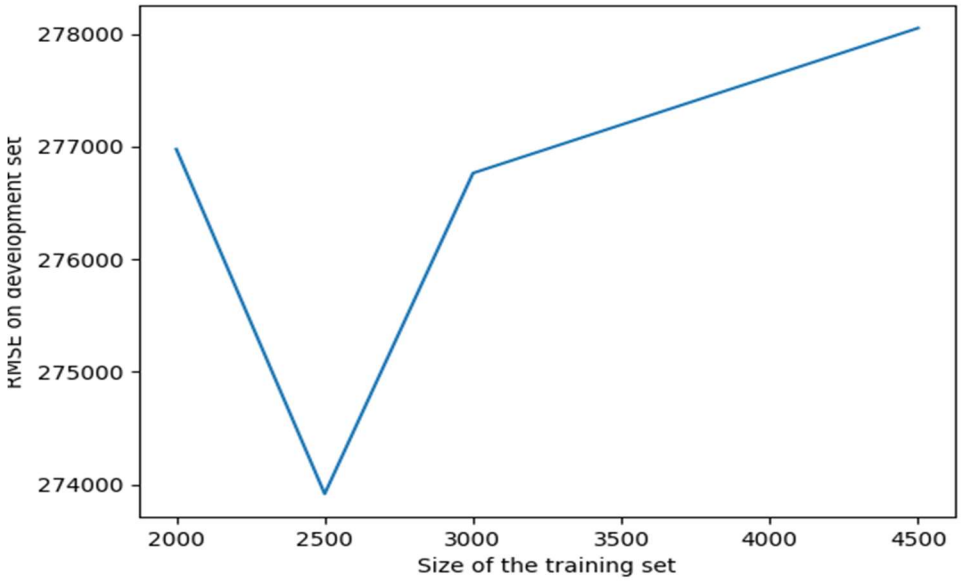
190040082

(May get little different (not much) answer than reported as weights are randomly initialized)

1. Using gradient descent method, RMSE for df_train.csv = 253454.81043638688, RMSE for dev.csv = 246664.26598675892
    a. Absolute difference between the following two calls, compute RMSE (phi, w1, y) and compute RMSE (phi, w2, y) on dev.csv, where w1 and w2 were obtained using the closed-form solution and gradient descent, respectively = 3.9447594998055138
    b. Stopping criterion used or gradient descent converged when RMSE on development set becomes minimum (as can be seen in code).
    c. Absolute difference between the two calls compute RMSE (phi, w1, y) and compute RMSE (phi, w2, y) on dev.csv = 0.033429565373808146.
2. RMSE on all the development set samples when p=2 is 246664.26133395056and when p=4 is 246663.05907888332
3. Basis function developed using known sets of basis function to me which are applied on features which are not one hot encoded and on all categorical columns of one hot encoded features to minimize RMSE on development basis set. Tried same for better test set prediction. RMSE basis using L2 norm on dev set = 246499.14583434275
4. plot of RMSE on development set using gradient descent vs Size of the training set



5. Mileage, seats are the two least features found using 2 minimum weights of features using closed-form solution. For features which are one hot encoded maximum weight out of weight of category of those features are considered.
6. Best linear regression model which best predicted on 50% test data on Kaggle is SGD model. Some data of categorical data is not given in dev set which is given in dev set and vice versa leading to dimension mismatch of phi_dev with weights, so in hope that union of train + dev set covers all data, concatenation of set given is done with train+dev set followed by split to avoid dimension mismatch.

7. Link to final results of competition - https://www.kaggle.com/c/cs419-assgmt1-20-21/overview