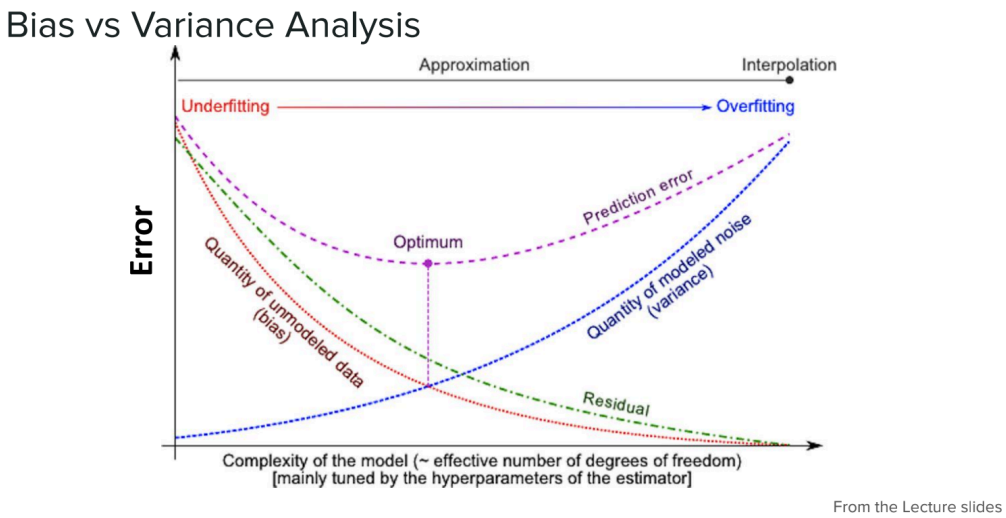
CSE 343 : Machine Learning
Assignment 1 Solutions

Section A

1. As we increase the complexity of the model , the model will be able to understand the underlying patterns easily thus reducing the bias but at the same time it will become more sensitive i.e will understand more from the noise around than the normal data thus increasing the variance of the model leading to overfitting of the data.



[Image Reference : Tutorial 2 slides (in class)]

From this image we can see that as the complexity of the model increases (plotted on x axis), the variance increases and thus leads to overfitting of the data.

2. To measure the performance of the model, We can use two commonly used metrics of evaluation : Recall and Precision.
   For spam emails :
   True Positives (TP): 200 (spam emails correctly identified as spam)
   False Positives (FP): 20 (legitimate emails incorrectly labeled as spam)
   False Negatives (FN): 50 (spam emails incorrectly labeled as legitimate)
   True Negatives (TN): 730 (legitimate emails correctly identified as legitimate)

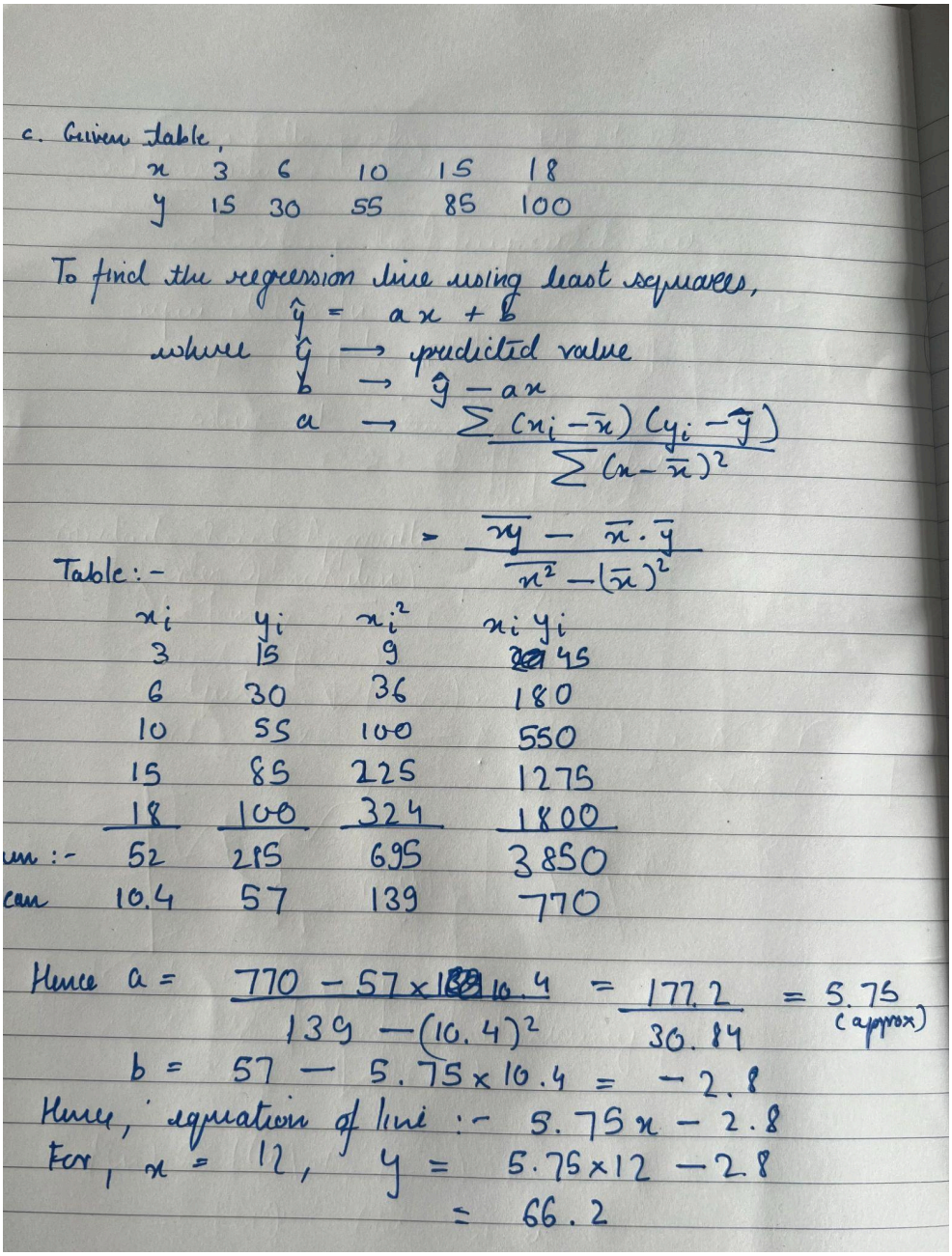
   Hence , Recall = TP / (TP + FN) = 200 / (200 + 50) = 0.80
   And Precision = TP / (TP + FP) = 200 / (200 + 20) = 0.9091

For legitimate emails :

True Positives (TP): 730 (legitimate emails correctly identified)
False Positives (FP): 50 (spam emails incorrectly labeled as legitimate)
False Negatives (FN): 20 (legitimate emails incorrectly labeled as spam)
True Negatives (TN): 200 (spam emails correctly identified as spam)
Hence , Recall = TP / (TP + FN) = 730 / (730+ 20) = 0.973
And Precision = TP / (TP + FP) = 730 / (730 + 50) = 0.9358

Avg. Precision = 0.9224
Avg. Recall = 0.8865
Accuracy = TP+TN / TP + TN + FN + FP = 930 / 1000 = 93%

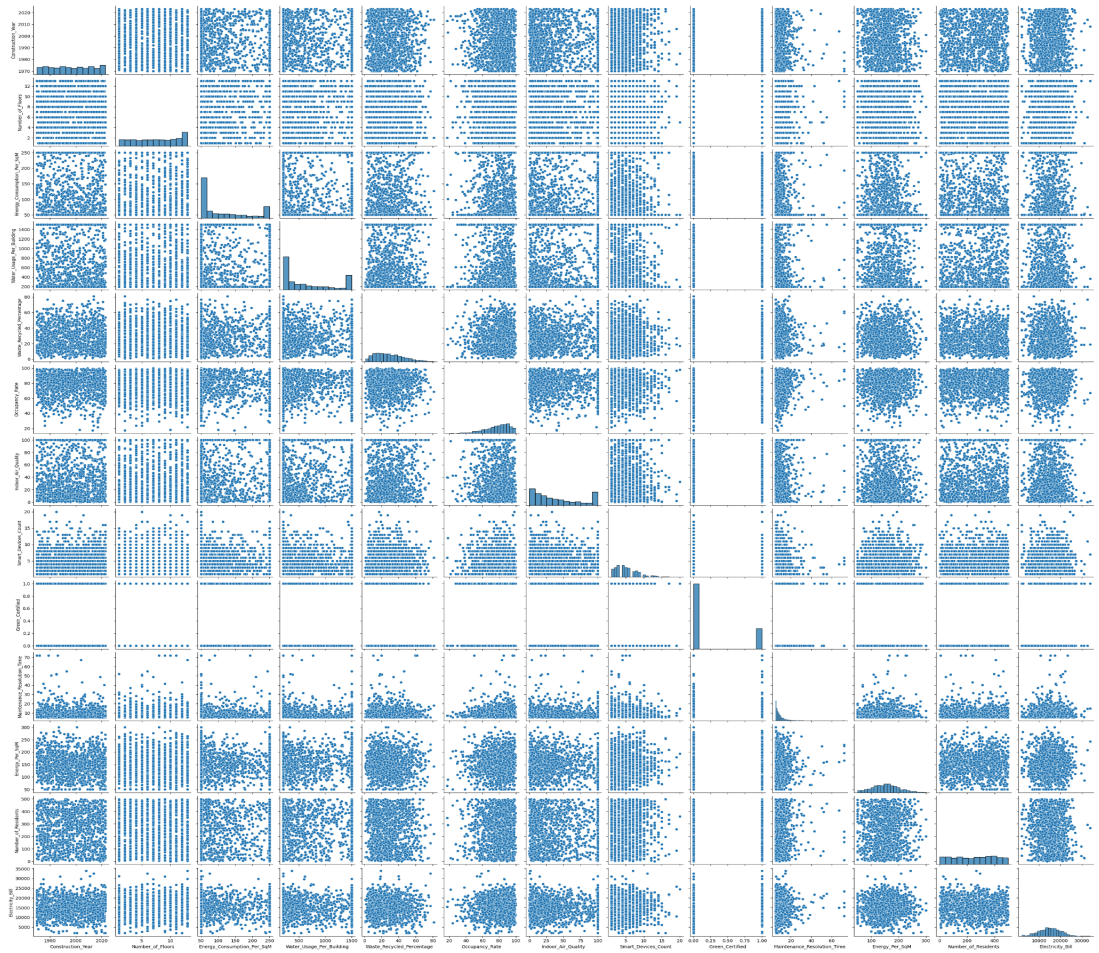
3. The expected value for y is 66.2 when x = 12. I have attached all the working in the picture below (did the implementation and working using pen and paper)



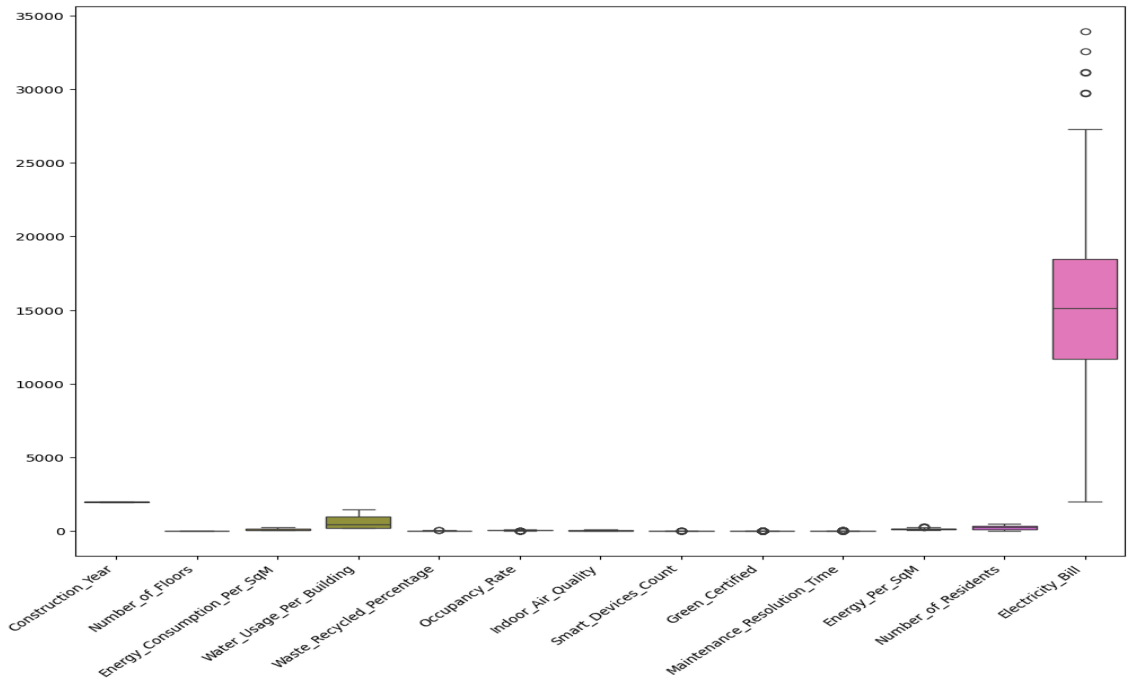
4. The above situation where the empirical loss on f1 might be lower but it may not generalize better describes the situation of overfitting the model. For example, Let us take a simple classification model to predict the tumors as malignant and benign. There might be a possibility that f1 is a complex model with many parameters, such as a high-degree polynomial regression or a deep neural network with multiple layers. Due to its complexity, f1 can fit the training data very well, potentially even memorizing it leading to a low empirical loss. On the other hand, f2 might be a simpler model. While f1 might excel on the training data, it's more susceptible to overfitting.In contrast, f2 might have a slightly higher empirical loss on the training set, but its simplicity can help it avoid overfitting and generalize better to new data.

<u>Section - C</u>

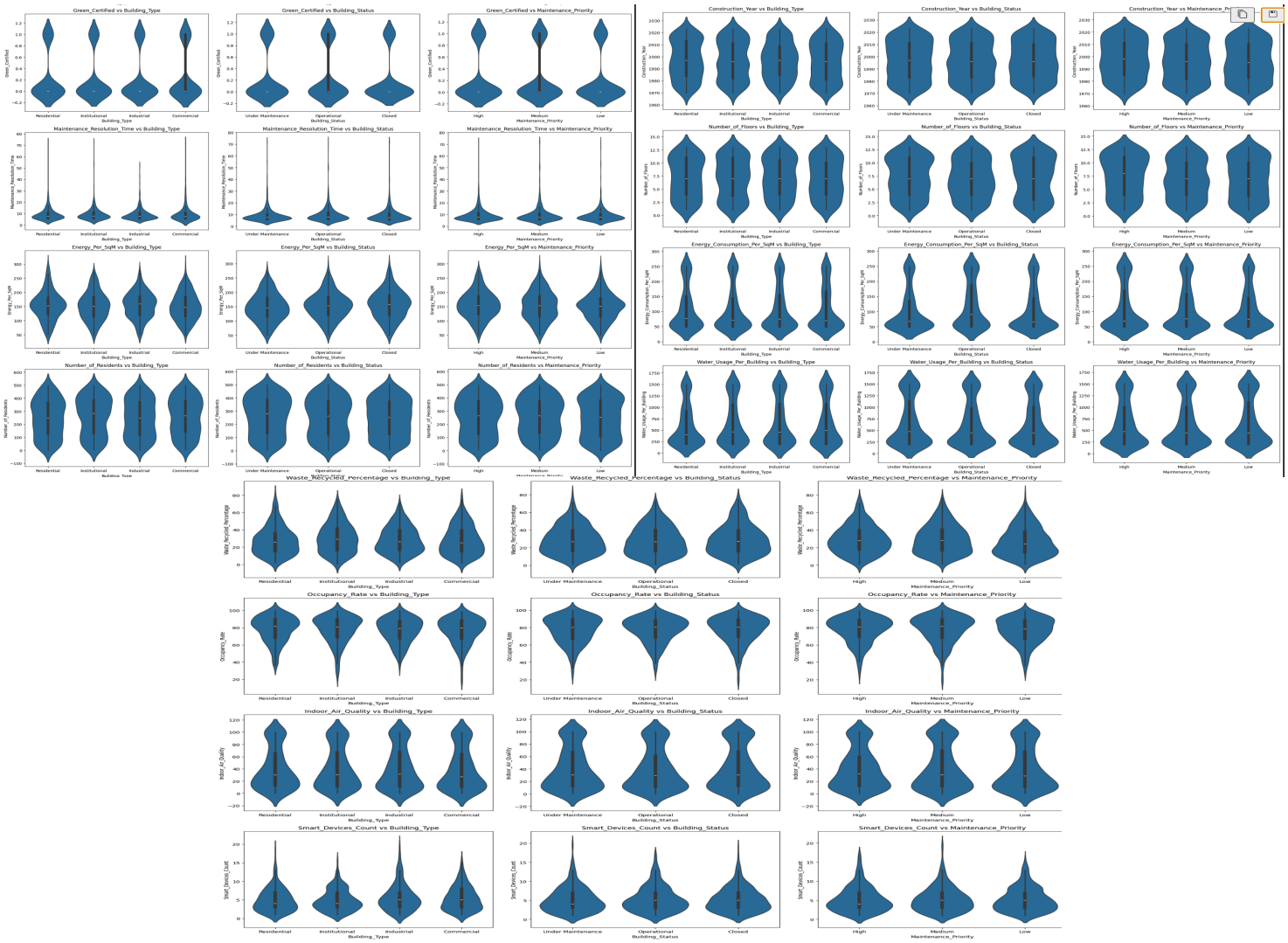
a. Pairplot for the data is as follows :



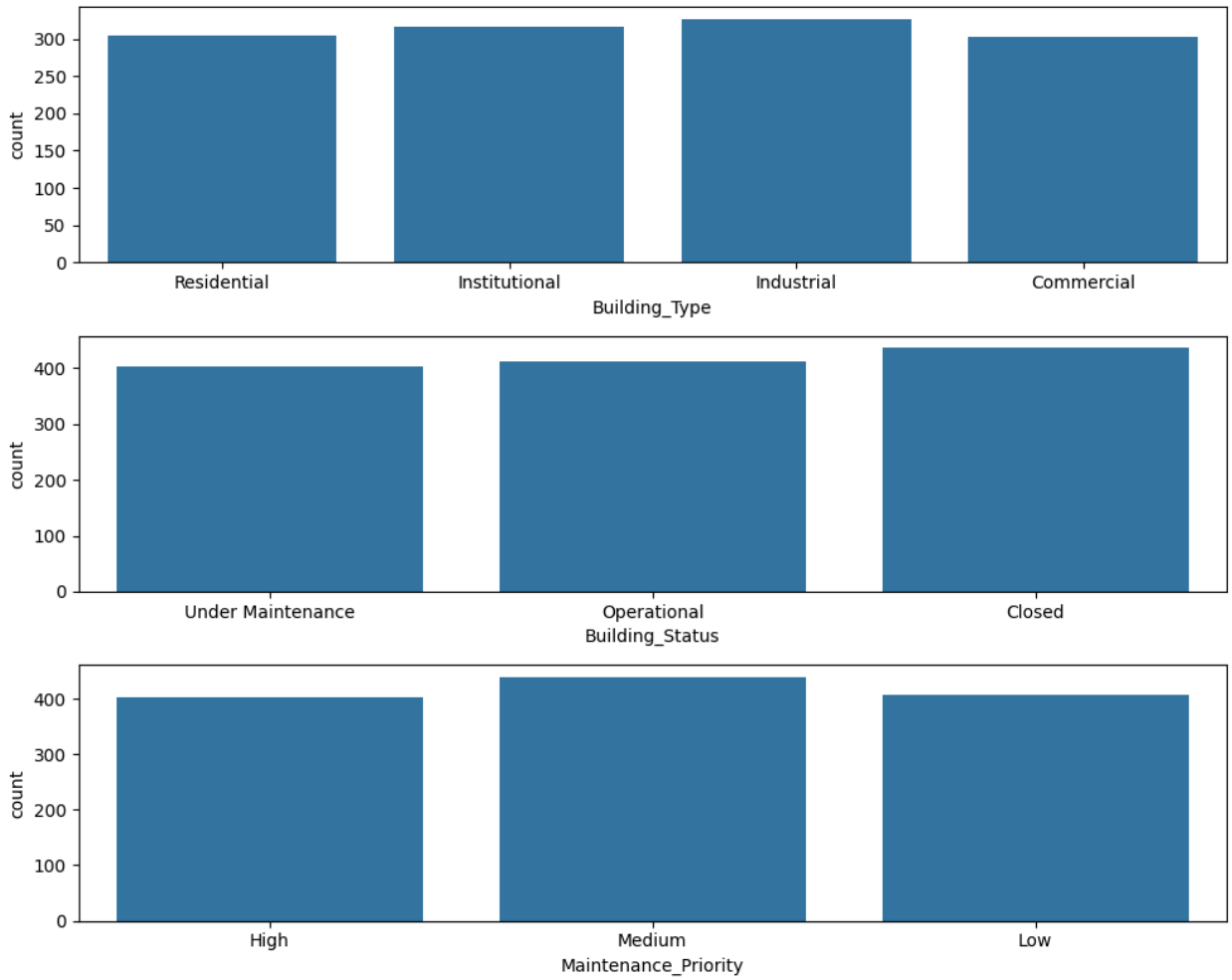
The box plot is as follows (I have calculated this only for numerical columns) :



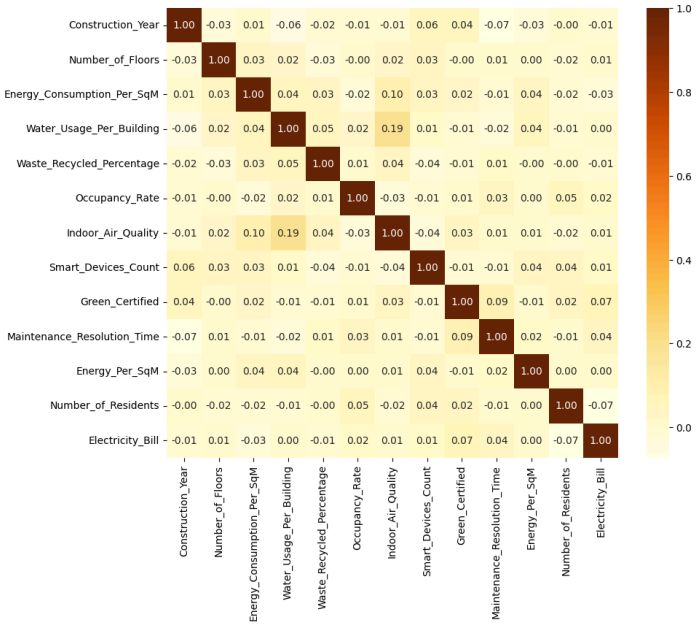
The violin plot is as follows (I have calculated for numerical vs categorical data)



The count plots for categorical data is as follows :



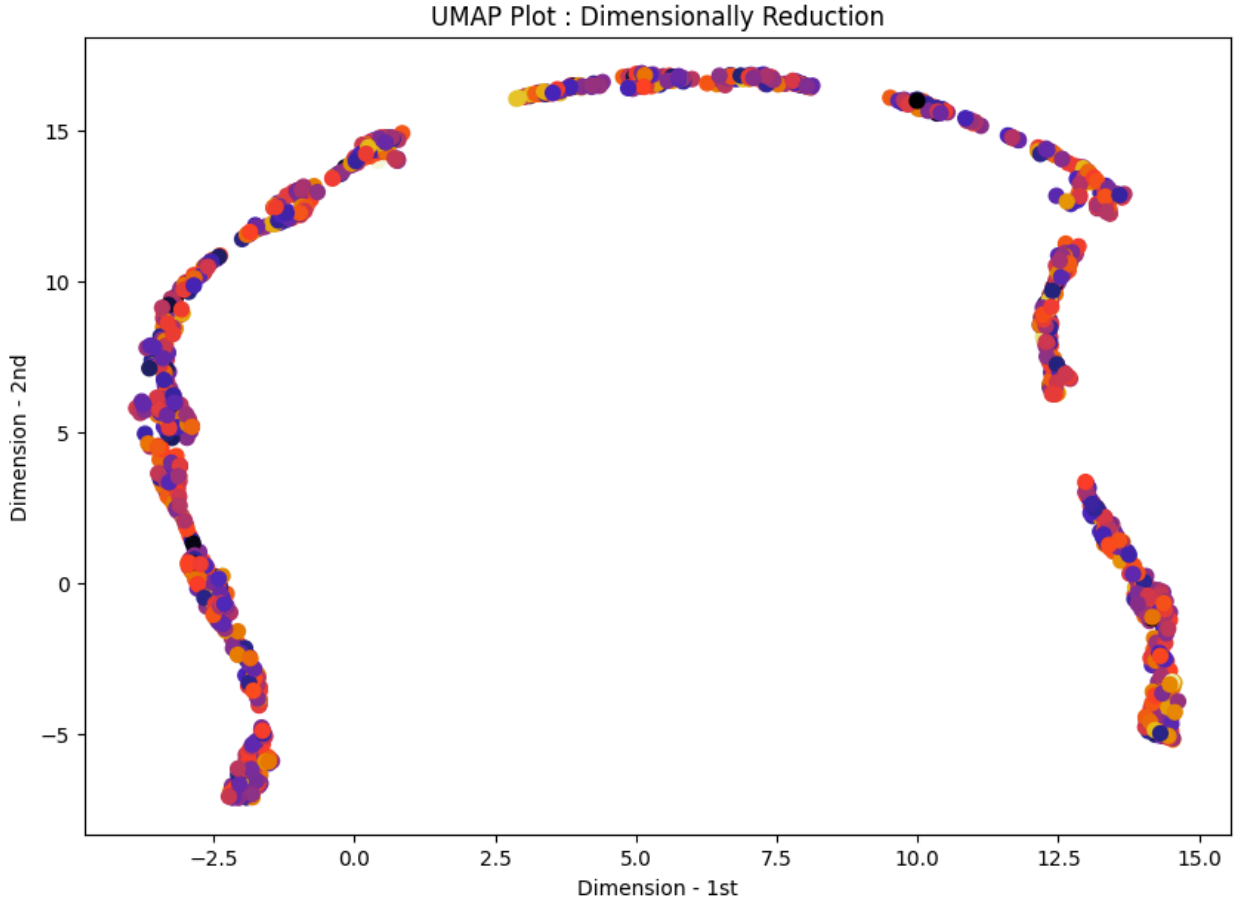
The correlation heatmap is as follows :



Based on these plots, following insights can be made for the data :

- The violin plots and box plot suggest that different building types have distinct energy consumption patterns due to variations in usage, occupancy rates, and building characteristics
- Occupancy rates are generally associated with increased energy consumption, as more people are using the building. This is reflected in the positive correlation between occupancy rate and electricity bill. There's a strong positive correlation between energy consumption and electricity bills, indicating that higher energy usage directly translates to higher costs
- The violin plots, box plot, and pair plot reveal the presence of outliers in several numerical variables. These outliers might be indicative of unusual data points or errors
- The count plots suggest that buildings with different maintenance priorities and statuses have varying distributions
- Indoor air quality doesn't appear to be strongly related to the other variables in the dataset

b.



As per the Umap data, We can observe that there are various different clusters in the map. The distance between the clusters is large and significant, each of them being clearly separated. These observations suggest the following things :

1. Observations is the original data have various categories and is well diverse
2. For each of the category , the data points share similar characteristics with strong underlying pattern

These observations can be confirmed by the original dataset given to us.

c. For the linear regression model implemented, the results obtained were as follows :

	Training Metrics	Testing Metrics:
MSE:	0.9898407279477682	0.9818735957078486
RMSE:	0.9949073966695434	0.9908953505329655
R2 Score:	0.013922520844610098	3.7344733075372893e-05
Adjusted R2 Score:	-0.0011091480449538782	-0.0640628254763429
MAE:	0.8056895778274504	0.7727247430098398

d. After using RFE, the 3 most important features came out to be Building type, Green certified and Building Status. The results obtained are as follows :

	Training Metrics	Testing Metrics:
MSE:	0.997869743108829	0.9779576904865817
RMSE:	0.9989343037001127	0.9889174336043336
R2 Score:	0.005924030979948647	0.004025392685427898
Adjusted R2 Score:	-0.00922956610877157	-0.05981913342448064
MAE:	0.80578119407792	0.7693546911355292

On comparing these results from part -c , following observations can be made :

1. For the training results , MSE and RMSE is slightly higher indicating that the model is making larger errors on the training data. R2 score has also decreased meaning a weaker fit on the training data

2. For the testing results, MSE and RMSE are slightly lower indicating that the model is making smaller errors on the testing data. R2 score has also increased indicating a better fit on the testing data.

Overall , the model performance in part d has downgraded from part c

e. Implementing One hot encoding and Ridge Regression, the following results were obtained :

	Training Metrics	Testing Metrics
MSE:	0.9898408237206868	0.9818997517961477
RMSE:	9949074448011166	0.9909085486542881
R2 Score:	0.013922425435808017	1.0706771100044143e-05
Adjusted R2 Score:	-0.0011092449081582245	-0.0640911709999834
MAE:	0.8056772546593934	0.7727209962946449

On comparing these results from part -c , following observations can be made :

1. For the training results , MSE and RMSE have remained almost identical, indicating that the model's errors on the test data have not changed significantly. R2 score has also decreased meaning a weaker fit on the training data
2. For the testing results, MSE and RMSE are almost the same. R2 score has also decreased slightly indicating a worse fit on the testing data but it is minimal

Overall , the model performance is part e is almost similar to part c

f. After performing ICA on the one-hot encoded dataset, the results obtained were as follows :

	Training Metrics	Testing Metrics
MSE:	0.9898407418955132	0.9818634831721499
RMSE:	0.9949074036791128	0.9908902477934425
R2 Score:	0.013922506949892721	4.764357204289915e-05
Adjusted R2 Score:	-0.0011091621514809358	-0.06405186645539018
MAE:	0.8056858707072951	0.7727173764958635

MSE and RMSE are slightly lower on the testing set compared to the training set, indicating that the model is making slightly smaller errors on unseen data. The R2 score is significantly lower on the testing set, suggesting that the model's ability to generalize to new data is poor. The adjusted R2 score is also significantly lower on the testing set, further confirming that the

model's performance is not improving when considering the more number of features. The MAE is slightly lower on the testing set, indicating that the average absolute error is slightly smaller on unseen data

g. For various values of alpha (0.1 , 0.5 , 1 , 5 , 10 , 100) :

```
ElasticNet with alpha=0.1:
MSE: 0.9802244188169418
RMSE: 0.9900628357922248
R2 Score: 0.0017169044137154144
Adjusted R2 Score: -0.06227560171361057
MAE: 0.7691863999506915

ElasticNet with alpha=0.5:
MSE: 0.9784754660478332
RMSE: 0.9891791880381599
R2 Score: 0.0034980781437957686
Adjusted R2 Score: -0.060380250180319894
MAE: 0.7684754551434521

ElasticNet with alpha=1:
MSE: 0.9785667456305137
RMSE: 0.9892253260155209
R2 Score: 0.0034051169171490647
Adjusted R2 Score: -0.060479170459956766
MAE: 0.7685683497940659
```

```
ElasticNet with alpha=5:
MSE: 0.9793339785497569
RMSE: 0.9896130448562999
R2 Score: 0.0026237492640333038
Adjusted R2 Score: -0.06131062578314417
MAE: 0.769269313544698

ElasticNet with alpha=10:
MSE: 0.98087786809866
RMSE: 0.990392784756967
R2 Score: 0.0010514166343459364
Adjusted R2 Score: -0.06298374896601655
MAE: 0.7703256727953284

ElasticNet with alpha=100:
MSE: 0.9854405214954387
RMSE: 0.9926935687791266
R2 Score: -0.0035952945366743982
Adjusted R2 Score: -0.06792832623774325
MAE: 0.7727149477663728
```

From this data, we can observe following things :

1. MSE and RMSE generally decrease as alpha increases, indicating that the model's errors on the test dataset are getting smaller with higher alpha values
2. The R2 score increases as alpha increases, suggesting a slightly better fit to the test data. The adjusted R2 score also increases with higher alpha values.
3. The MAE decreases slightly as alpha increases, suggesting that the average absolute error is getting smaller

h. After applyingGradient Boosting Regressor, the results obtained were as follows :

	Training Metrics	Testing Metrics
MSE:	0.6036689062148138	0.9886374134441157
RMSE:	0.7769613286482242	0.9943024475831231
R2 Score:	0.39862616633389736	0.006851082833227773
Adjusted R2 Score:	0.3894588822841092	-0.07139281891228078

MAE:	0.6219647244467083	0.7685546378565014

On comparing these results with part -c and part -g , following observations can be made :

1. GBR significantly outperforms ElasticNet and the Linear regression on the training set, with much lower errors and higher R2 scores
2. GBR's performance on the testing set is comparable to ElasticNet, with slightly lower errors and a slightly higher R2 score
3. The Linear Regression performs the worst on both training and testing sets out of the three approaches