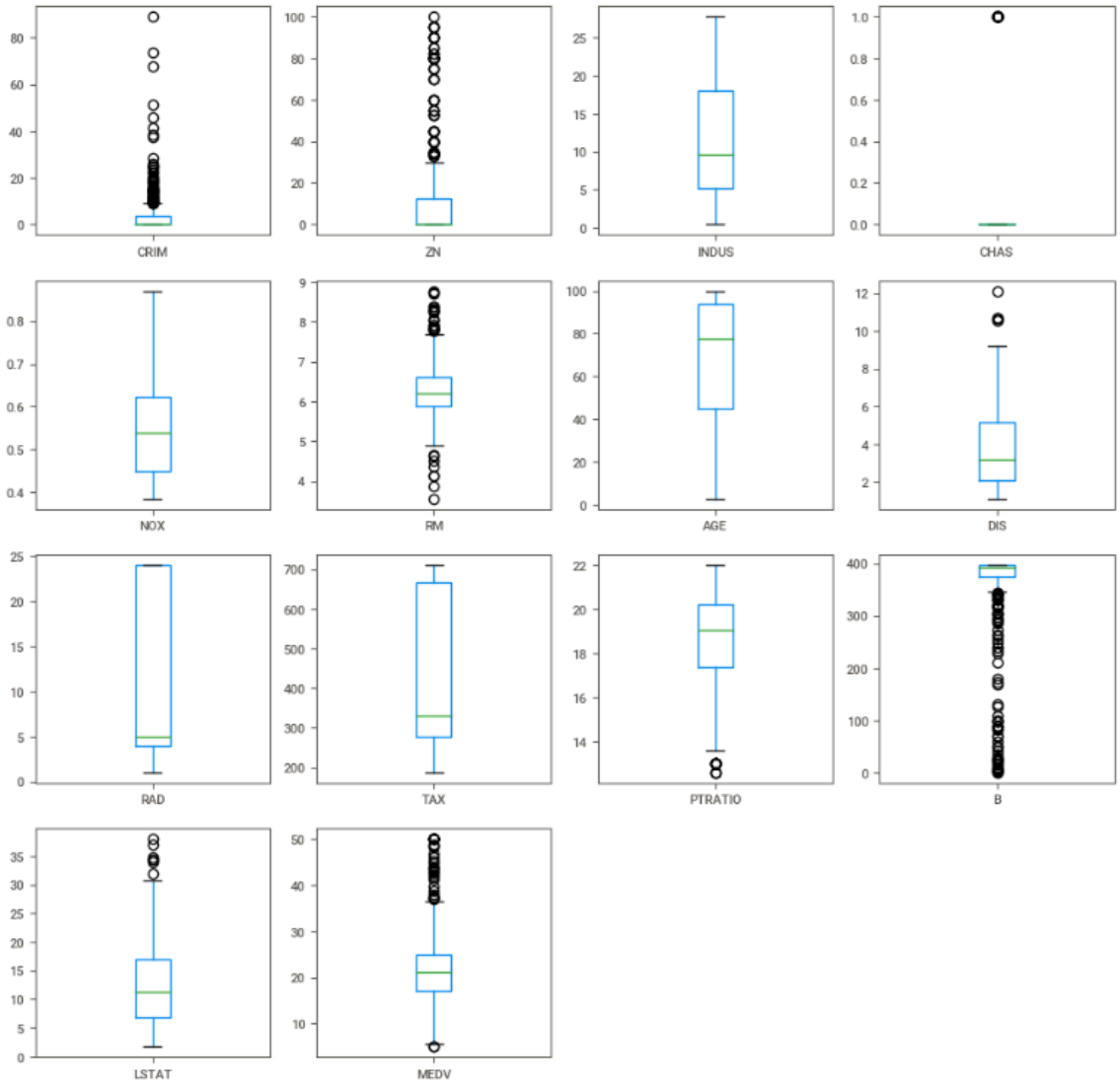
Assignment 2

Rashmi S

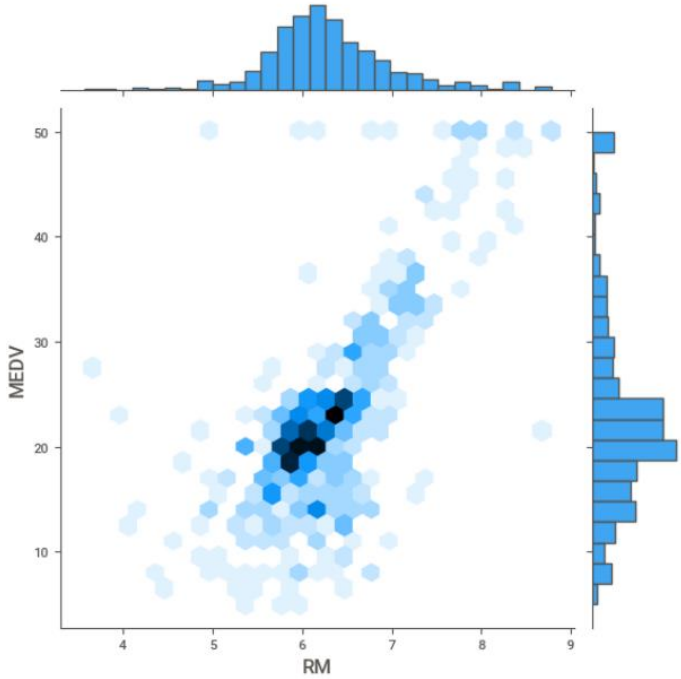
21BDA02

1. Document 5-6 key insights from EDA and support each point with a visualization.

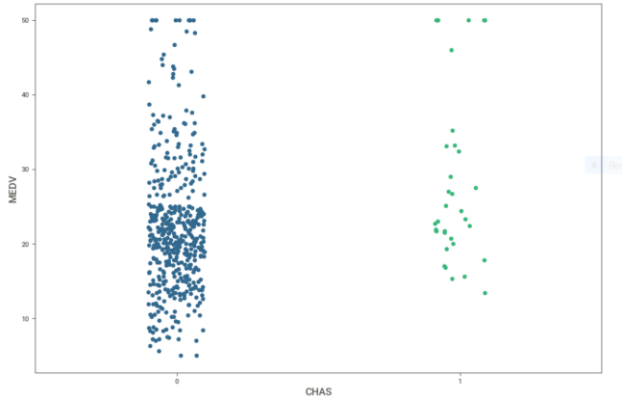
- We observe outliers in the following columns
  CRIM, ZN, NOX, RM,PTRATIO,B AND LSTAT



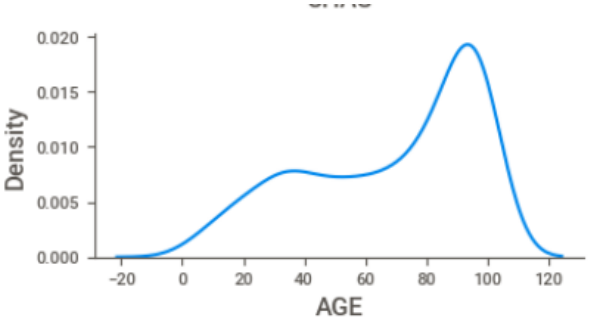
- In the joint plot below, we can observe a good correlation between RM and target i.e. MEDV, also it displays histogram



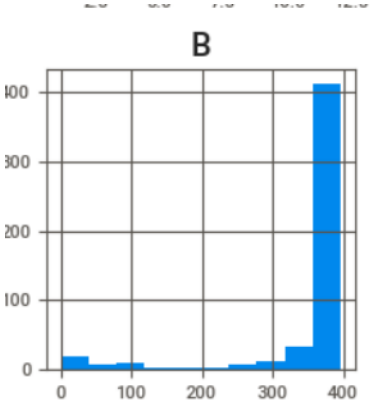
- We observe that the cost of the house increases with increase in the river arenas nearby



- So most of them in Boston own house in their late 90's



- Proportion of blacks in the town falls in the range of 300 - 400



2. Answer the following questions:

i. What are the assumptions of linear regression?

There are four assumptions associated with a linear regression model:

1. **Linearity**: The relationship between X and the mean of Y is linear.
2. **Homoscedasticity**: The variance of residual is the same for any value of X.
3. **Independence**: Observations are independent of each other.
4. **Normality**: For any fixed value of X, Y is normally distributed.

ii. How can we evaluate a Regression model? Define each metric and its interpretation.

We can evaluate using Mean square error, Root Mean square error

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n}}$$

Where yi is the actual value

Y cap is the predicted value

n = no of observations

iii. Can R squared be negative?

R2 score can be negative. R2 is not always the square of anything, so it can have a negative value without violating any rules of math. R2 is negative only when the chosen model does not follow the trend of the data

iv. What is dummy variable trap?

The Dummy variable trap is a scenario where there are attributes that are highly correlated (Multicollinear) and one variable predicts the value of others. So , ideally they should be removed before the regression is done

v. Is One Hot Encoding different from Dummy Variables?

One-hot encoding ensures that machine learning does not assume that higher numbers are more important.

A dummy (binary) variable just takes the value 0 or 1 to indicate the exclusion or inclusion of a category. In one-hot encoding, “Red” color is encoded as [1 0 0] vector of size 3. “Green” color is encoded as [0 1 0] vector of size 3. . “Blue” color is encoded as [0 0 1] vector of size 3

vi. How is polynomial regression different from linear regression?

The model is no more a line but curve of some sort, it might give us much accurate results than linear regression

vii. Interpret the screenshot below from the notebook we discussed in class today:

```
In [50]: model.score(X_test, model.predict(X_test))
Out[50]: 1.0

In [51]: model.score(X_train, model.predict(X_train))
Out[51]: 1.0

In [52]: model.score(X_test, y_test)
Out[52]: 0.9085774752313169

In [54]: model.score(X_train, y_train)
Out[54]: 0.8911672911176578
```

- The R^2 value btw x_test and predicted value of x_test is 1, which implies that the predictions of the train set are 100% accurate, which is obvious as we are comparing with the same variables

- The R^2 value btw x_train and predicted value of x_train is 1, which implies that the predictions of the test set are 100% accurate, which is obvious as we are comparing with the same variables
- The accuracy of the model predicting the y variables wrt to x is 0.91
- The accuracy of the trained values of x and y are 0.89

viii. Bonus: We saw Sweetviz as an Automated EDA option. What are the other options? Try a few of them and share which one did you find the best.

We can use Autoviz, which provides more diverse visualization tools, so as to have greater sense to visualization