DPY401T Assignment

M Besho 215423085

Classification modelling problem

1. **The datasets you have chosen along with the features and response variable you would like to predict.**

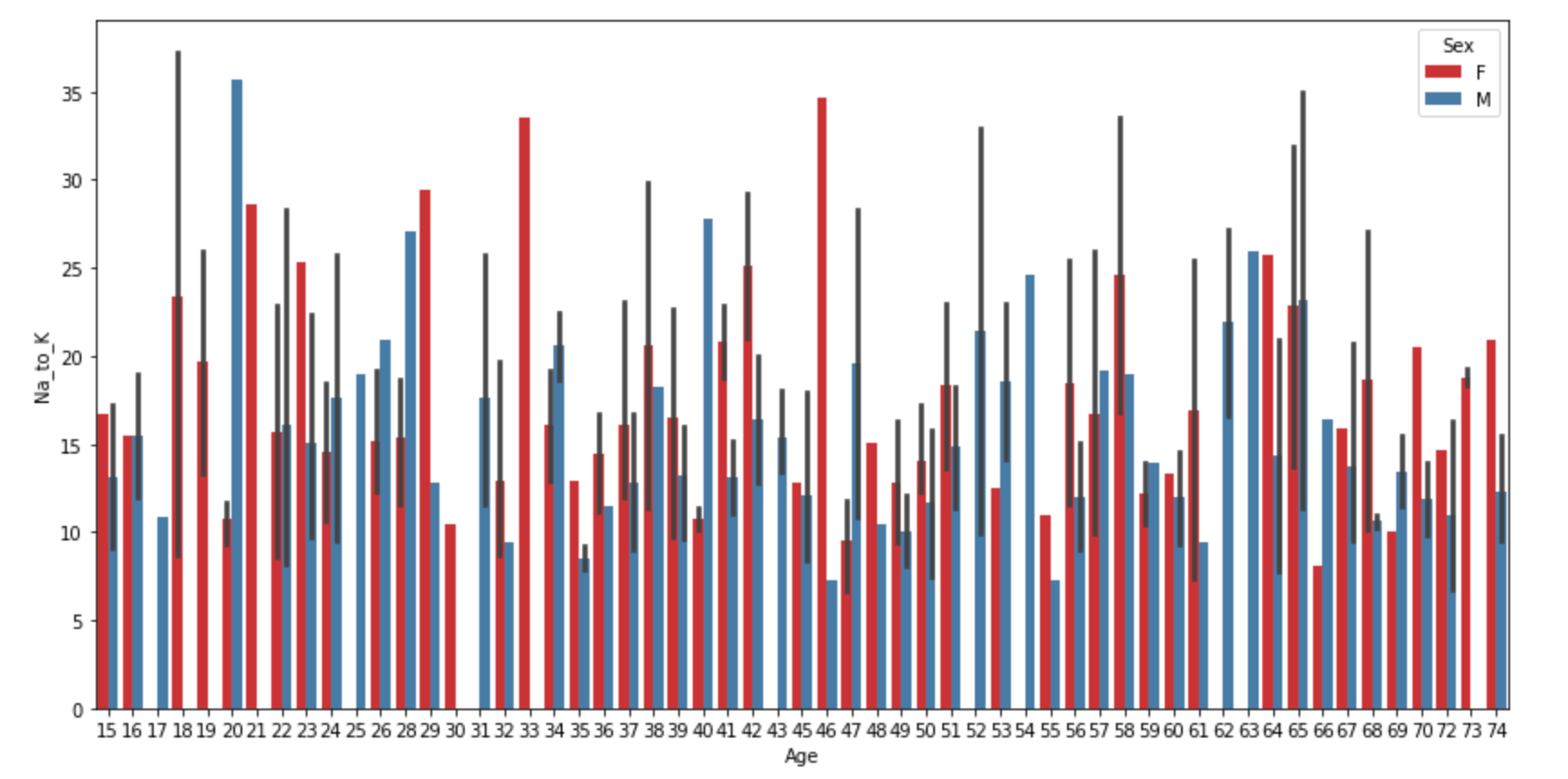
* The dataset I have chosen for classification modeling problem is Drug Classification.
* It is found here: <https://www.kaggle.com/prathamtripathi/drug-classification>
* Upon choosing this dataset I aim to predict the outcome of the drugs that will be prescribed for a patient.
* Based on the following features:
  + Age
  + Sex
  + Blood Pressure (BP)
  + Cholesterol Levels
  + Sodium to Potassium ratio in blood (Na\_to\_K)
* And my target is:
  + Drug Type
* From the dataset, we can deduce the following type of drugs:
  + Drug A
  + Drug B
  + Drug C
  + Drug X
  + Drug Y

1. **The EDA process along with visualisations and explanations that you followed to understand and clean your data.**

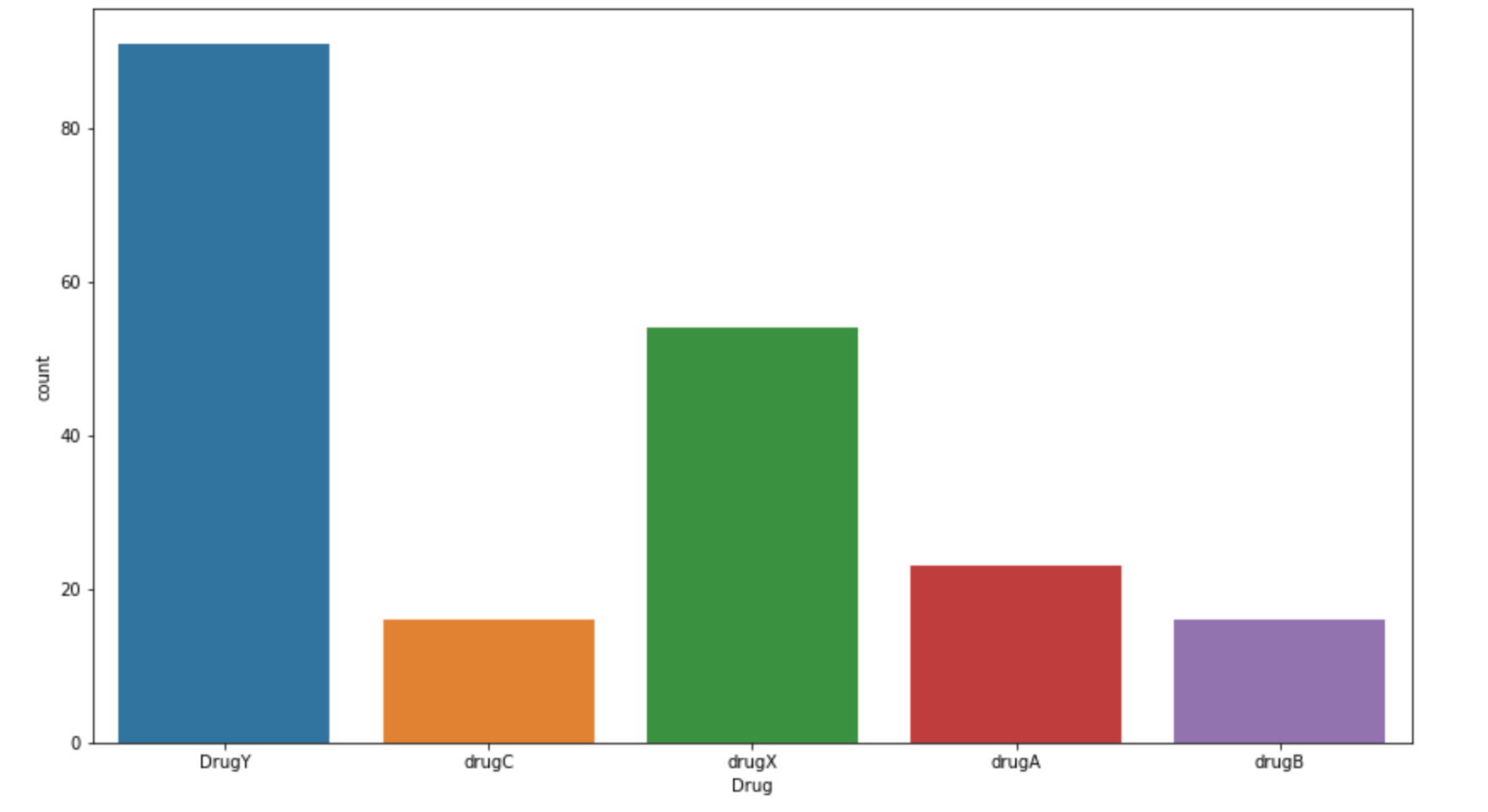
* The structure of the dataset is 200 rows and 6 columns
* 200 rows being data values from 5 data inputs: Age, Sex, BP, Cholesterol and Na\_to\_K
* From the structure of the data there are no abnormalities and null values
* I then described numeric columns to get the **mean, medium, min, max, std** and  **interquartile range**.
* The following table shows the above topics:

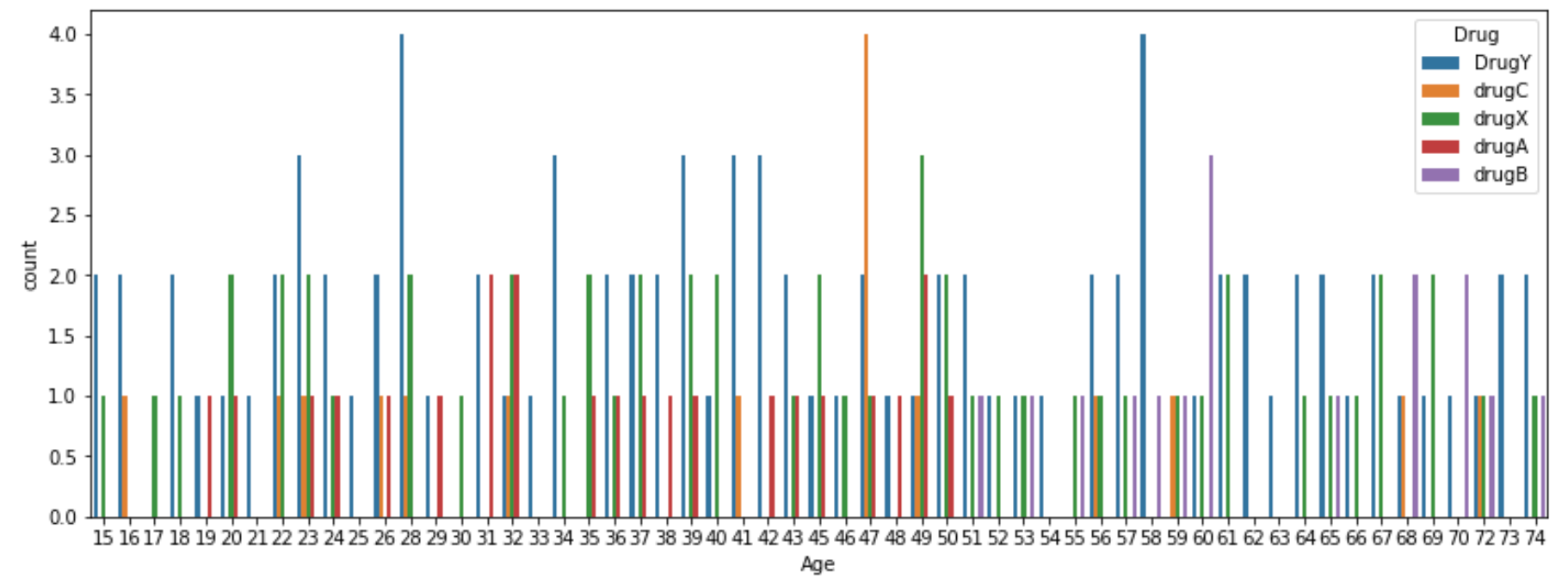
| **Age** | **Na\_to\_K** |  |
| --- | --- | --- |
| **count** | 200.000000 | 200.000000 |
| **mean** | 44.315000 | 16.084485 |
| **std** | 16.544315 | 7.223956 |
| **min** | 15.000000 | 6.269000 |
| **25%** | 31.000000 | 10.445500 |
| **50%** | 45.000000 | 13.936500 |
| **75%** | 58.000000 | 19.380000 |
| **max** | 74.000000 | 38.247000 |

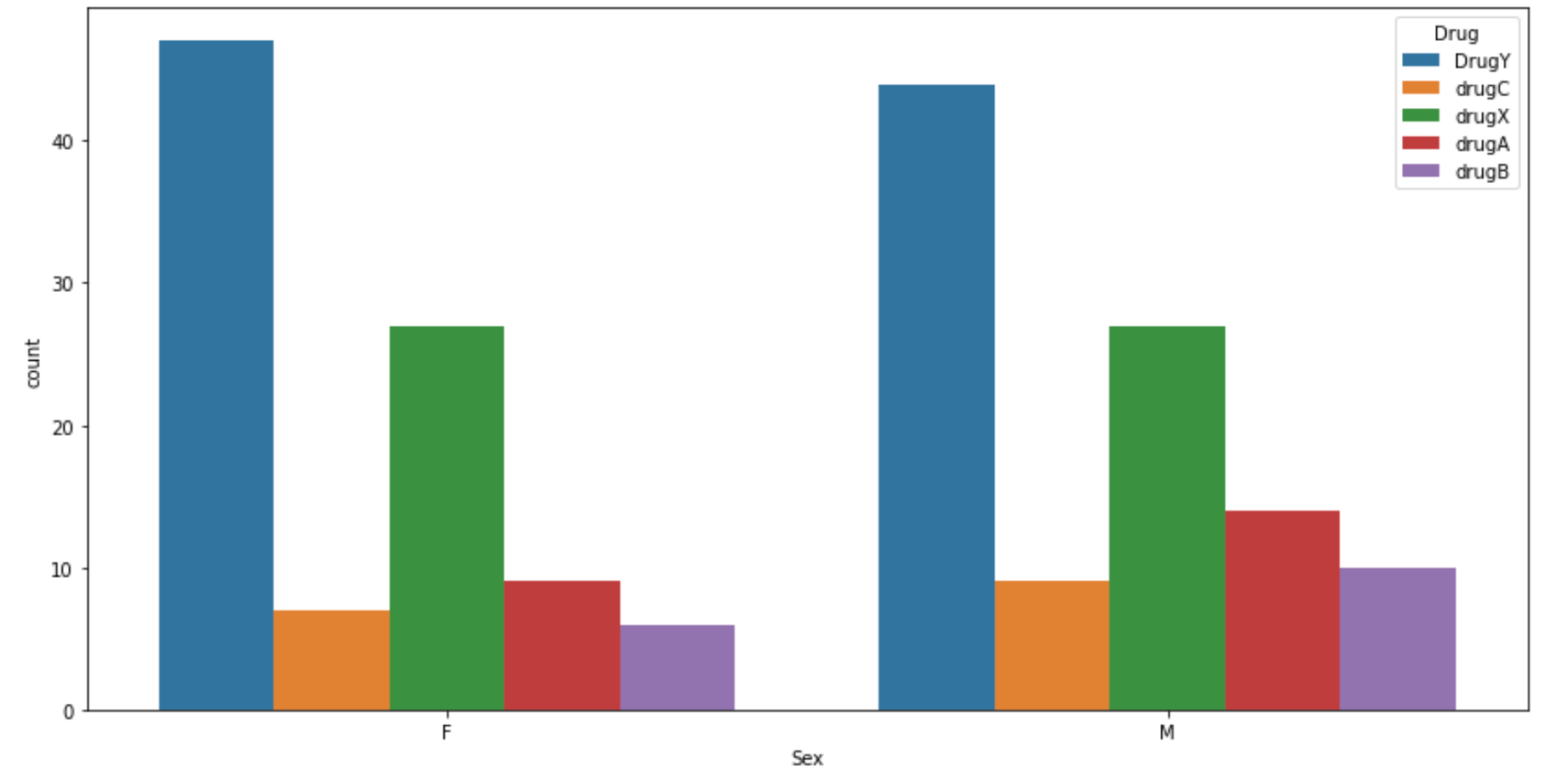
* From the table above I can take out that:
  + The highest age of a patient is 74
  + The lowest sodium to potassium ratio is 6.27
* I therefore plotted a bar graph comparing age with Na\_to\_K per Sex:



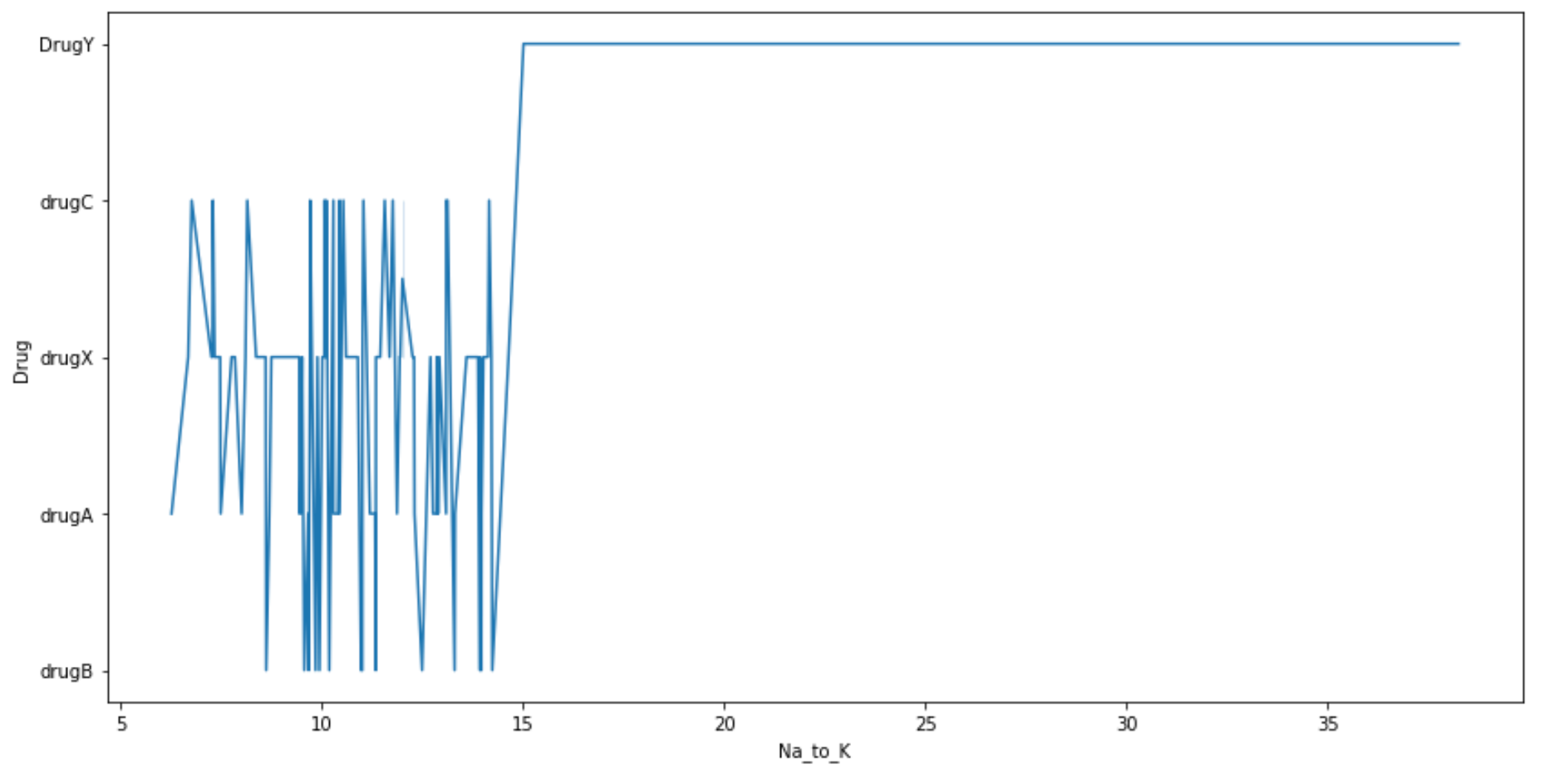
* The above graph can show us which gender and age will suffer from hypertension as high sodium to potassium ratio increases.
* From exploring age vs Na\_to\_k, I explored which drug type is used the most to better understand how features lead to which drug is prescribed to the patient



* From the bar graph above, Drug Y is the most prescribed drug to patients.
* We can then look at which drug is being prescribed the most based on ag
* 
* I noticed that between age 46 and 47 drug C is the most prescribed.
* We can look at the gender per drug prescribed as well.



* We can conclude that drug Y is prescribed to mostly females and drug A and B being prescribed for males.
* I then checked how BP, Cholesterol levels and Sodium to potassium ratio affects the type of drug being prescribed.

****

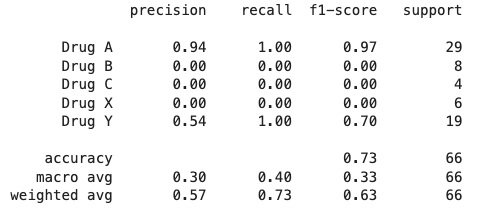
* Na\_to\_K value that is greater than 15 causes Drug Y to be prescribed to patients.
* From the other 2 graphs in the notebook, high BP causes drug X to be prescribed and high cholesterol causes drug C to be prescribed to patients.
* Partially we can conclude that all features affect our target (drug type prescribed)

1. **Three different supervised machine learning models should be used for each. Therefore, you should detail which models you have chosen and how you tuned the hyperparameters of the models using GridSearchCV.**

* The following are the classification models I have used:
  + Support Vector Machine (SVM)
  + K Nearest Neighbor (KNN)
  + Decision Tree Classifier
* **Support Vector Machine**:
  + Is a linear model for classification and regression problems. It can solve linear and non-linear problems.
* **K Nearest Neighbor:**
  + Is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems.
  + It assumes that similar things exist in close proximity. In other words, similar things are near to each other.
  + The model determines the class of a data point by majority voting principle. If k is set to 5 which is a default value for the model, the classes of 5 closest points are checked.
* **Decision Tree Classifier:**
  + is a graphical representation of all possible solutions to a decision based on certain conditions.
  + On each step or node of a decision tree, used for classification, we try to form a condition on the features to separate all the labels or classes contained in the dataset to the fullest purity.
* I choose these 3 algorithms because my dataset is linear by nature and I have different classes to compare through.
* For **SVM**:
  + The accuracy of the model is 73%
  + Although for classes Drug B, C, X the precision is always 0 means always classifying everything into 2 classes i.e. Drug A and Y
  + That is why I had to use GridSearchCV to find the best parameters.
  + After tunning we got almost 90% accuracy.
* For **KNN**:
  + The accuracy of the model is 70%
  + Although it classifies everything
  + The model is too specific and not generalized well
  + That is why I had to use GridSearchCV to find the best parameters.
  + After tunning accuracy remained the same but precision values in classes changed.
* For the Decision Tree Classifier we did not apply GridSearchCV because the accuracy is 100% on both training and test.

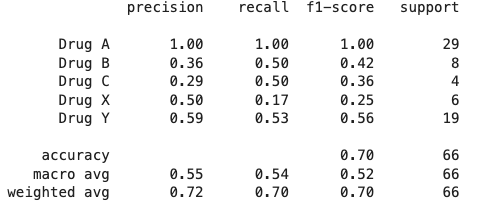
1. **Report and discuss the performances of your models by using adequate performance measures.**

* For SVM:



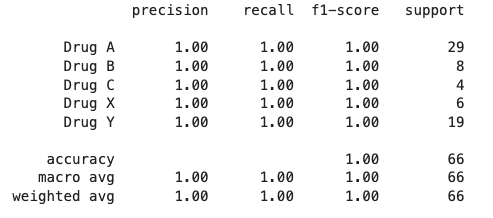
The model was 73% accurate although it had a good score on the test set compared to training.

* For KNN:

****

The model was 70% accurate although it was specific and did not know much about unseen data.

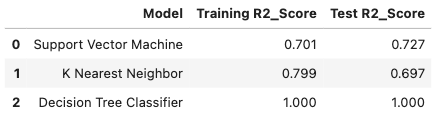
* For DTC:



The model is 100% and knew unseen data and predicted well. No overfitting.

1. **Based on your reported results, identify the best performing models for each modelling problem.**

* Based on the dataframe table that compares the predicted values from the model to true value
* And the below table shows R2 score on both training and testing data inputs



* The best model for this classification modelling is Decision Tree Classifier.

1. **List recommendations on what can be done in future to achieve better model performances.**

* Clean data and encode your categorical data inputs
* Split the dataset into a certain percentage.
* Tune parameters.

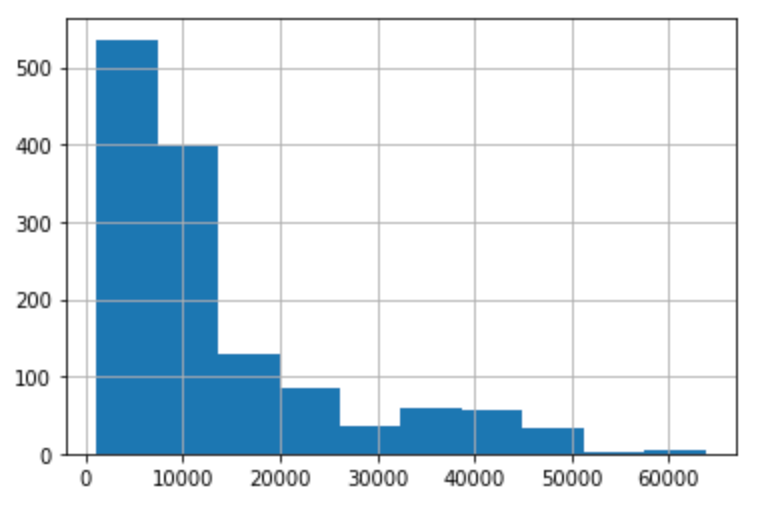
Regression modelling problem

1. **The datasets you have chosen along with the features and response variable you would like to predict.**

* The dataset I have chosen for the regression modeling problem is Medical cost prediction.
* It is found here: <https://www.kaggle.com/mirichoi0218/insurance>
* Choosing this dataset I aim to predict the medical costs that a customer needs to pay.
* Based on the following features:
  + Age
  + Sex
  + Body Mass Index (BMI)
  + Number of children
  + Whether you smoke or not
  + Region
* And my target is:
  + Charges

1. **The EDA process along with visualisations and explanations that you followed to understand and clean your data.**

* The structure of the dataset is 200 rows and 6 columns
* 200 rows being data values from 5 data inputs: Age, Sex, BP, Cholesterol and Na\_to\_K
* From the structure of the data there are no abnormalities and null values
* I then described numeric columns to get the **mean, medium, min, max, std** and  **interquartile range**.
* We can be more interested in charges as our target:

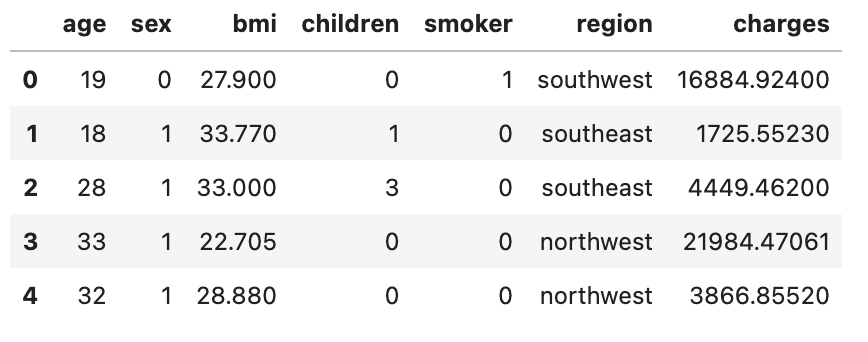


* From the above histogram, we can tell that the graphs is skewed towards the left
* What influences the charges to be either high or low? I am using EDA to theoretically and graphically find out.
* First, I explore my dataset by simply looking at the count between sex that I have:

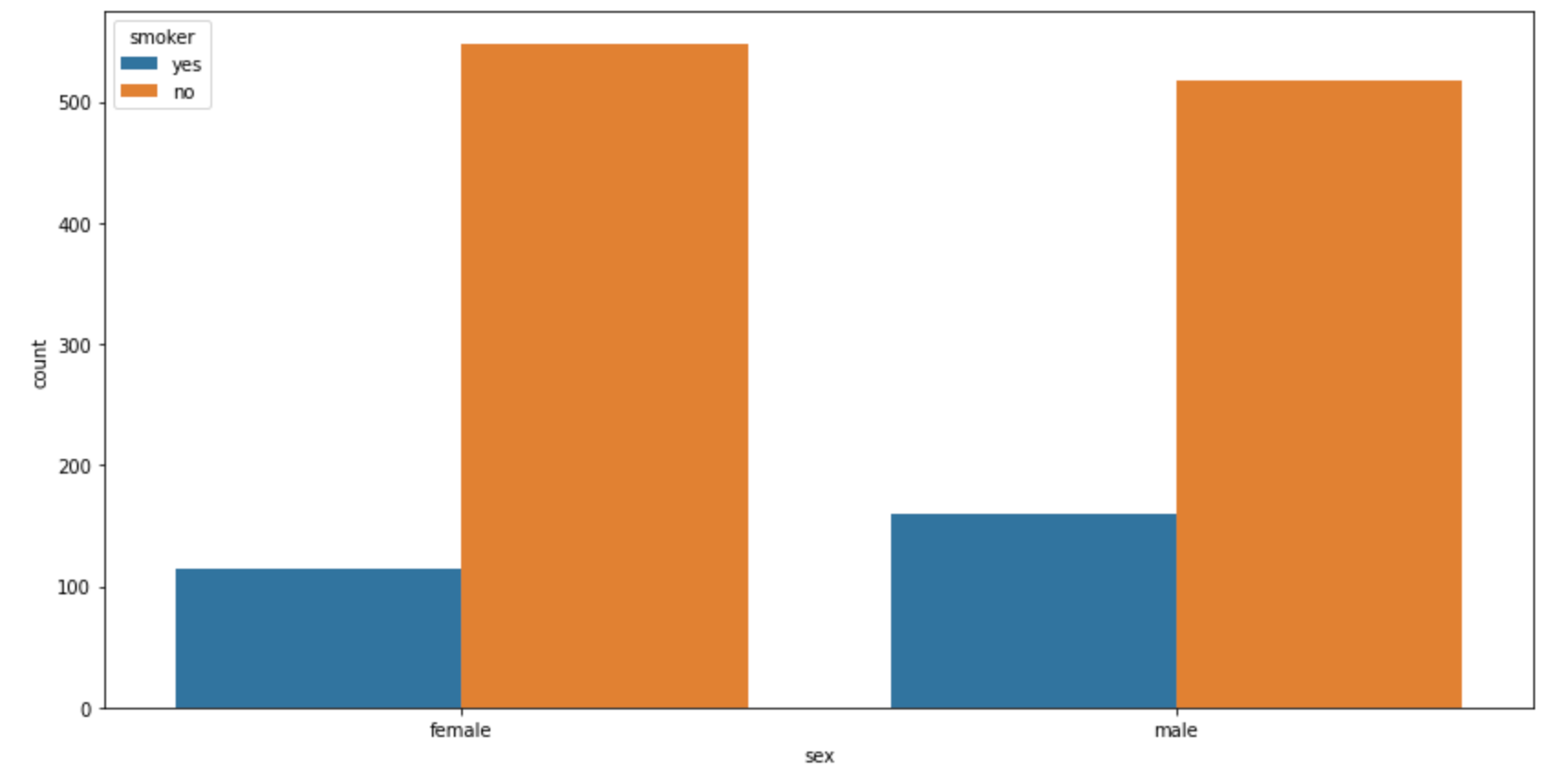
**female 662**

**male 676**

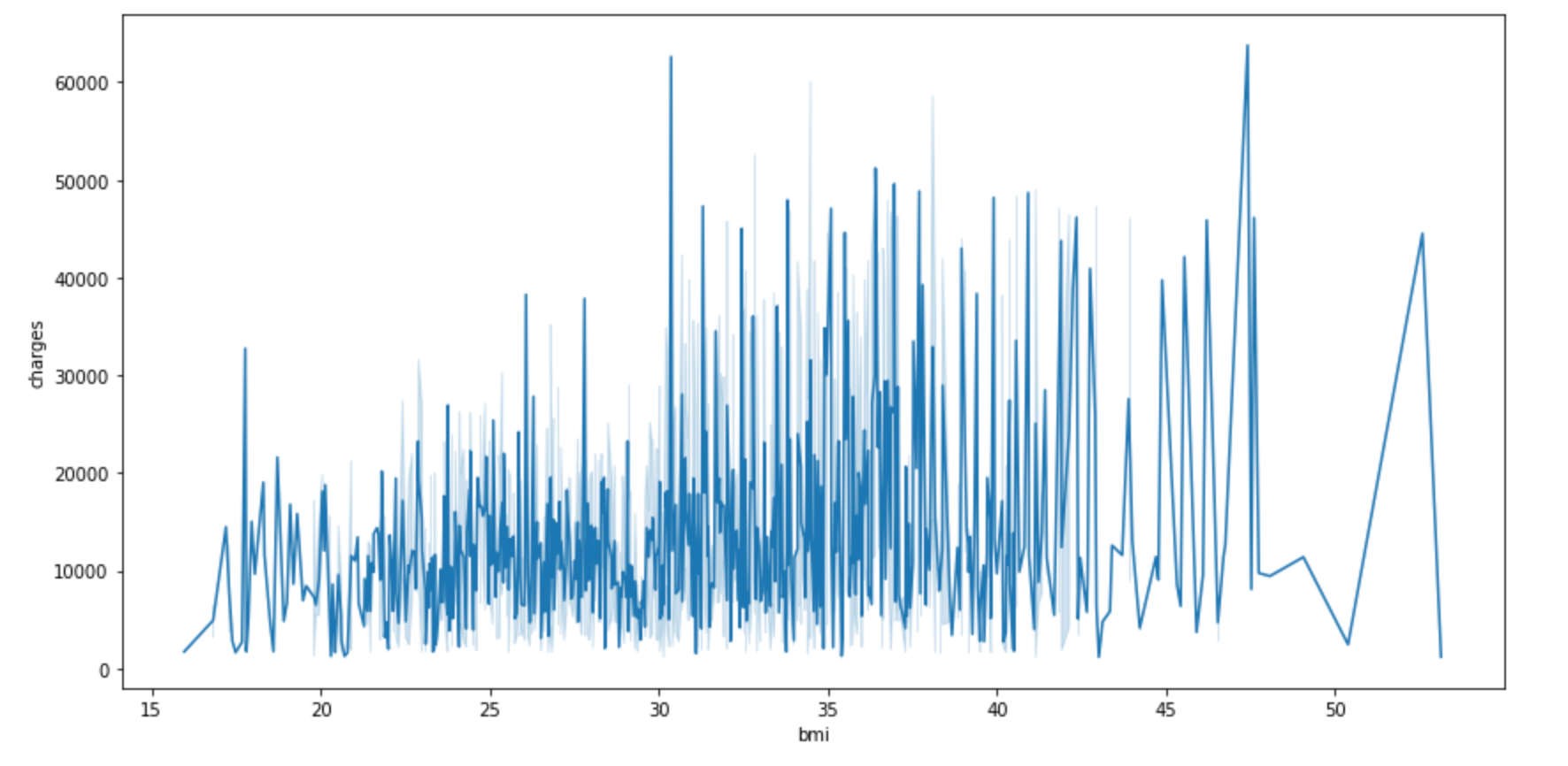
* The above shows that males are more than females.
* Exploring more on dataset, I found the following columns are categorical:
  + Sex, Smoker and Region
* I can apply OneHot, Standard or Binary Encoder and choose not to complicate the flow. I applied OneHot by simply having a dictionary of key-value data.
* After encoding I got the following results:



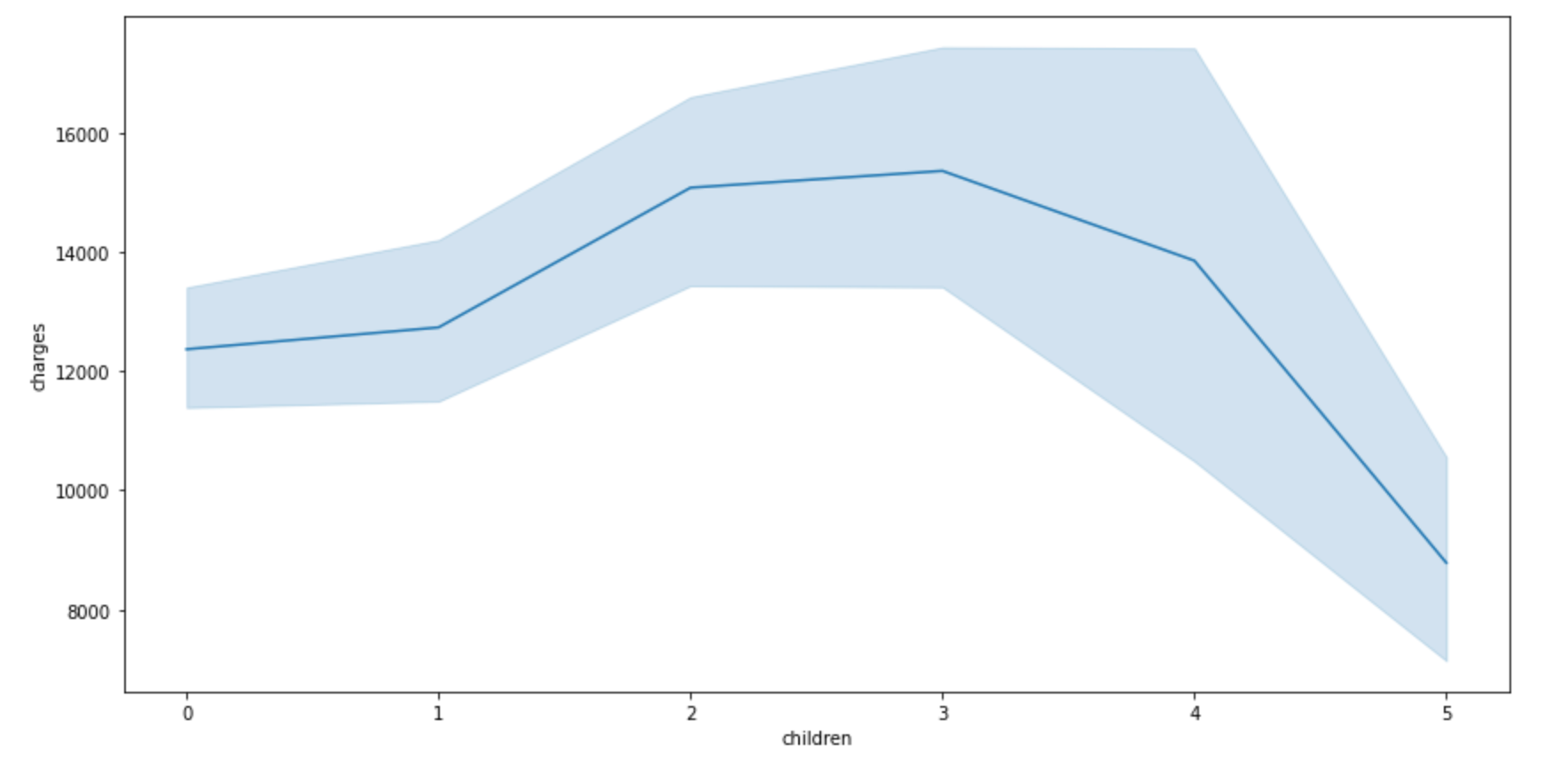
* Although region is a least affecting feature towards our prediction, I drop it off completely.
* Furthermore, I explored which sex has highest smoke count



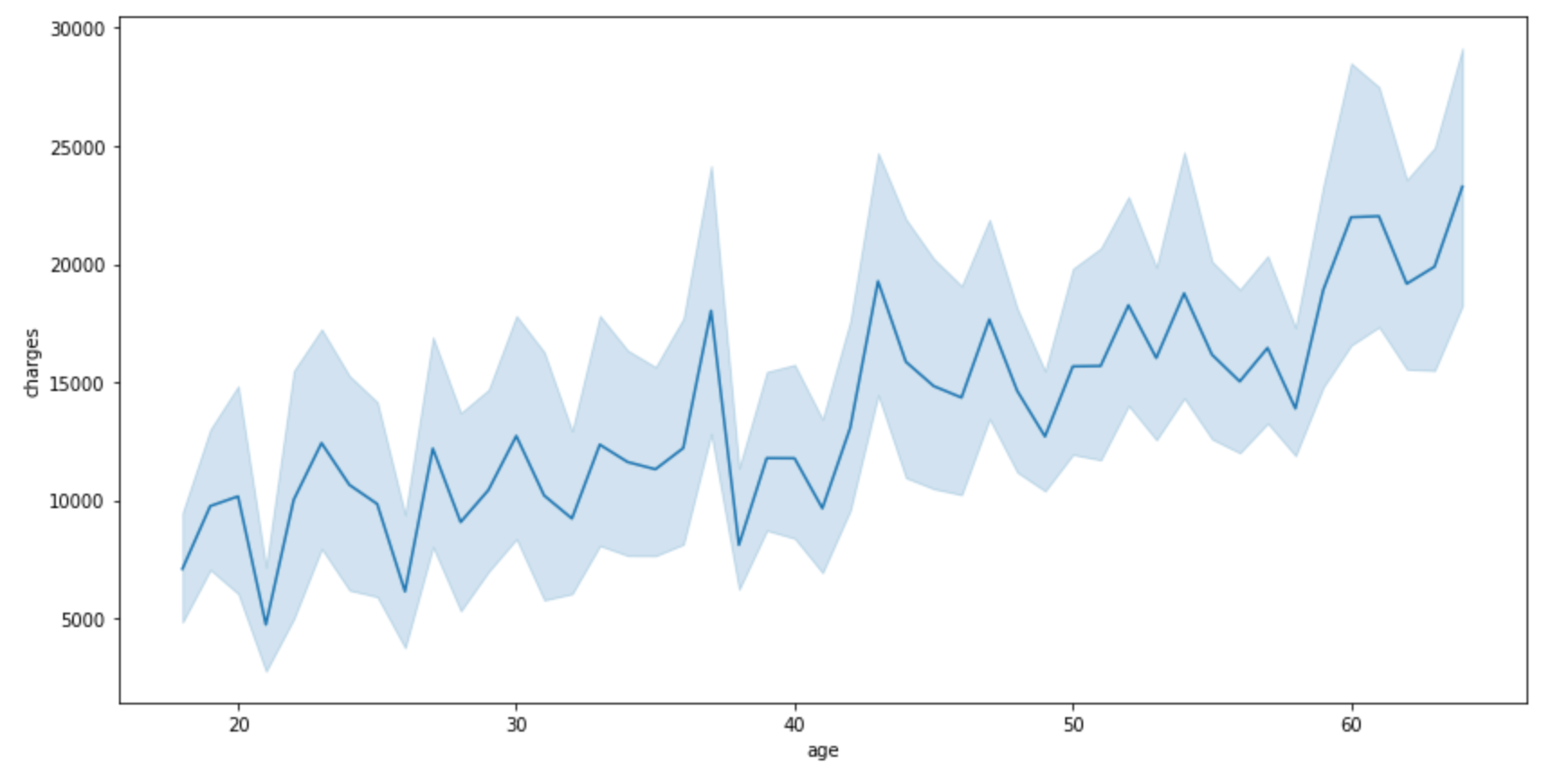
* It appears males smoke more than females although both sex has highest count of non-smokers
* I looked at how body mass index affects insurance costs.



* It shows that the graph is having a high steep slope as the bmi increases. Based on the BMI rules, BMI that is over 31 is obese and BMI less than 18.5 are likely to get an insurance completely.
* Does the number of children affect the charges you will pay towards the insurance?



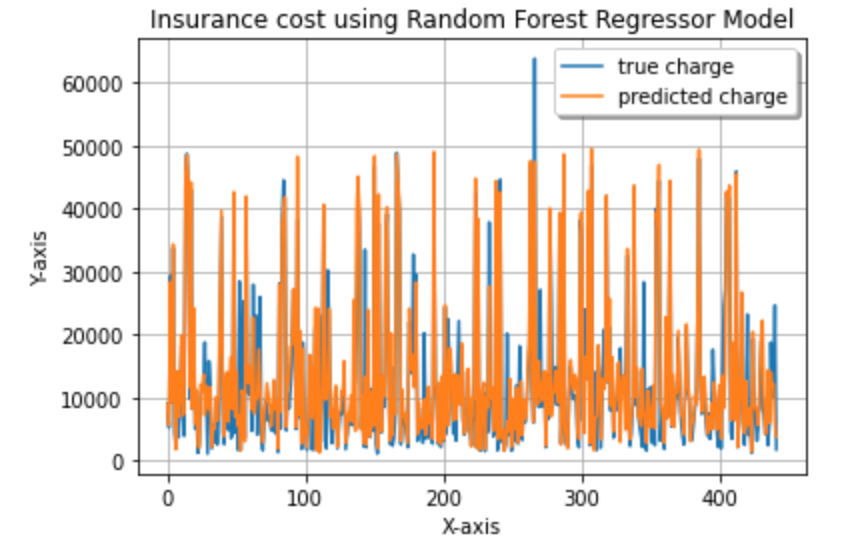
* From my observations, it doesn’t entirely affect charges
* Lastly let us explore if age feature affects charges



* Yet again we see a steep slope as the age increases.
* From EDA, we can conclude that as Age and BMI increase so as the insurance charges.

1. **Three different supervised machine learning models should be used for each. Therefore, you should detail which models you have chosen and how you tuned the hyperparameters of the models using GridSearchCV.**

* The following are the regression models I have used:
  + Linear Regression
  + Random Forest Regressor
  + Decision Tree Regressor
* **Linear Regression:**
  + is a technique for determining the relationship between two variables. It assumes that the two variables have a direct link and that this relationship may be represented by a straight line.
  + These two variables are called the independent variable and the dependent variable, and they are given these names for fairly intuitive reasons.
  + The independent variables in this case would be Sex, BMI, Children, Smoker and Region.
  + Then the dependent variable would be the insurance charges.
* **Random Forest Regressor:**
  + Partially inherits what a decision tree does although
  + It consists of a large number of individual decision trees that operate as one instead of individual points.
  + It randomly splits data inputs into prediction data inputs and runs them as models.
  + Target with the highest score becomes the model prediction
* **Decision Tree Regressor:**
  + is a graphical representation of all possible solutions to a decision based on certain conditions.
  + On each step or node of a decision tree, used for regression, we try to form a condition on the features to separate all the labels or classes contained in the dataset to the fullest purity.
* I chose these models because they work well with generalized linear dataset.
* **For Linear Regression:**
  + After running the model, the testing score is higher than the training score.
  + We can tune the parameters to see the performance of the model.
* **For Random Forest Regressor:**
  + After running the model, the model score on training is 97 and test is 87
  + The model ran well as per the graph:



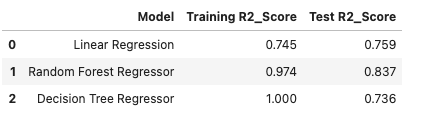
* We can tune the parameters to get better accuracy on test set
* After applying GridSearchCV, the test score of the model increased:

Test score:0.8629835247721431

* **For Decision Tree Regressor:**
  + Decision Tree Regression performance well on training set and 73% on testing.
  + We can tune the following parameters to see if we can’t boost the accuracy:
    - max\_leaf\_nodes
    - max\_features and more
  + After tuning the performance went down. The more we increase the leaf nodes on this dataset it loses to generalize.

1. **Report and discuss the performances of your models by using adequate performance measures.**

****

****

* Using sklearn metrics API and R2 score we can tell the performance of the models before tuning and from the above explanation some models after tuning they perform bad as a result of parameters or dataset at hand.

1. **Based on your reported results, identify the best performing models for each modelling problem.**

* The best model is Random Forest Regressor

1. **List recommendations on what can be done in future to achieve better model performances.**

* Clean data and encode your categorical data inputs
* Split the dataset into a certain percentage.
* Tune parameters.