Name ____ Zhang Xinge,QianZhang__ Date: __2023.12.6___

NYU ID: ____N10837293, N19756113______

Net ID: XZ4513, QZ2570

Course Section: _____ CSCI-GA.2433-001_________

Project #3

Total in points (100 points total): ____________________

Professor’s Comments:

Affirmation of my Independent Effort:

_____ZHANG XINGE_Zhang Qian________

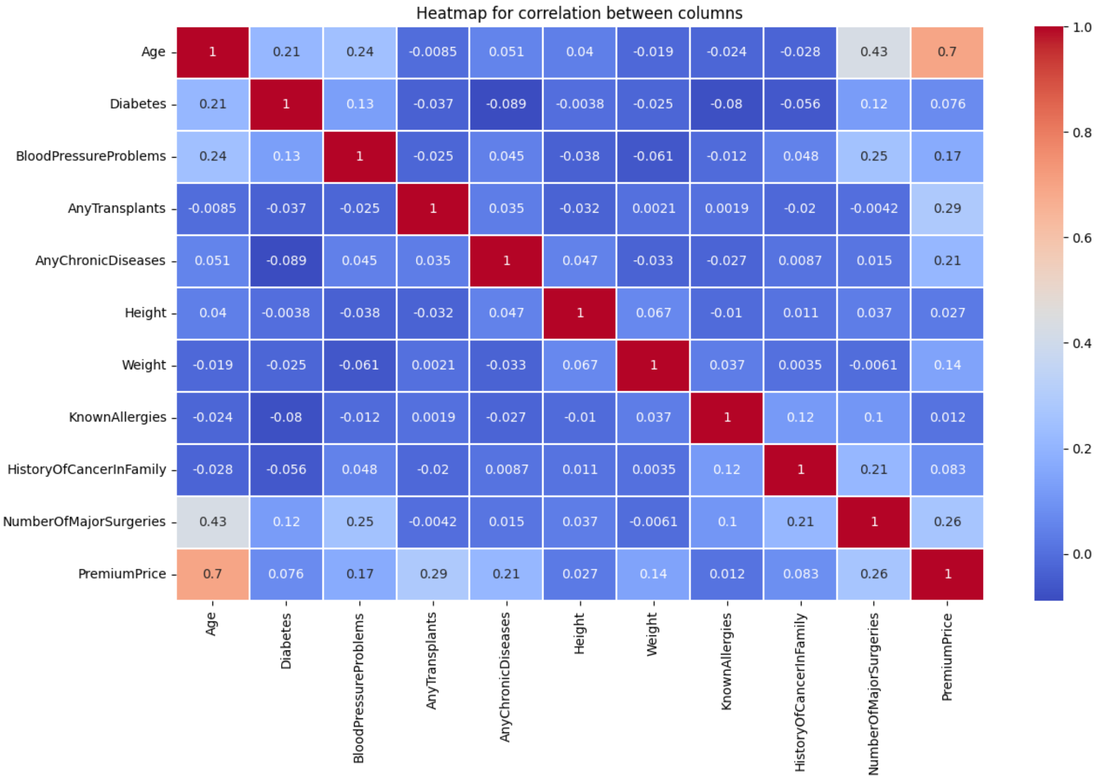
(Sign here)

README:

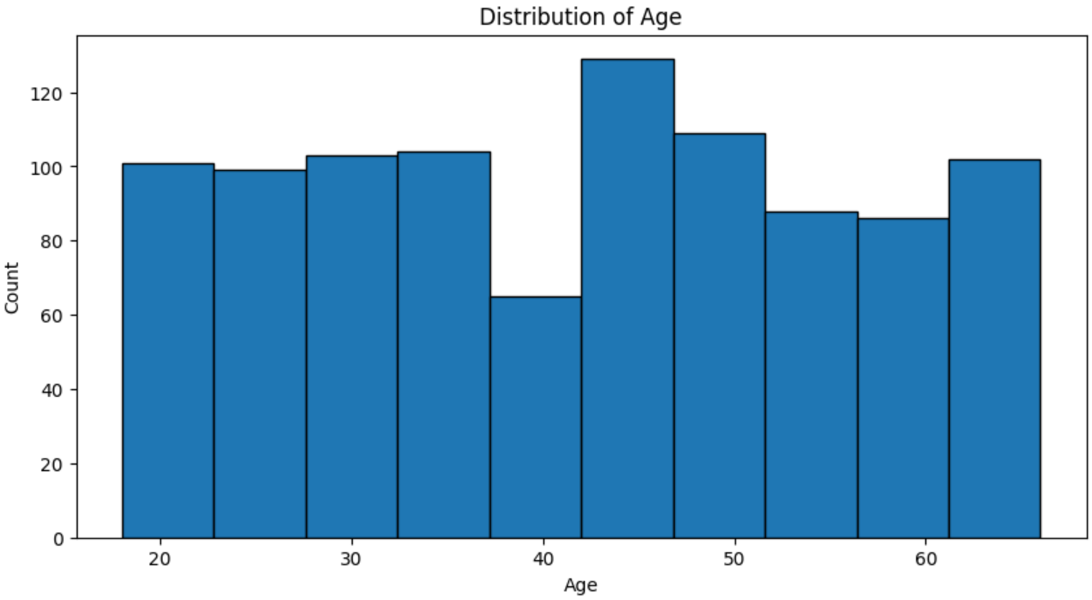
This assignment was completed by a group consisting of two students Xinge Zhang and Qian Zhang. Each member is considered to contribute equal effort to this solution.

In this section, we plan to use the datasets Medical Insurance Premium Prediction as our data lake. ossible to handle.

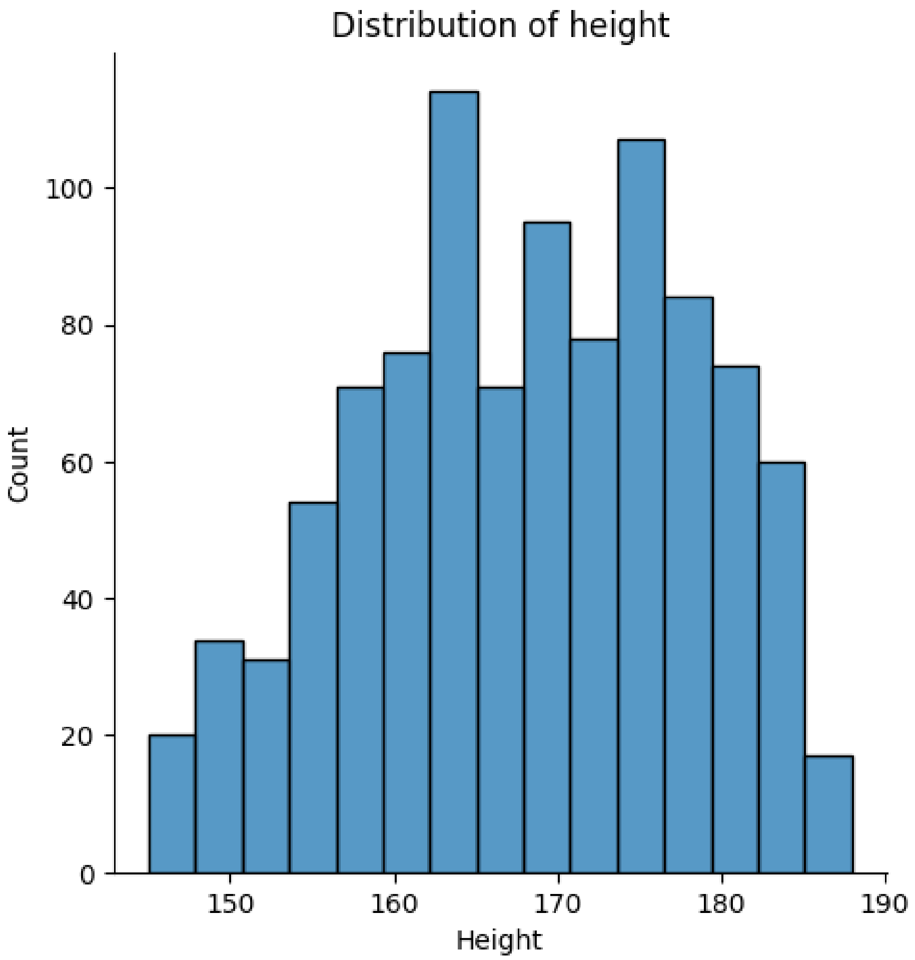
Heatmap for correlation between columns:



Distribution of Age:



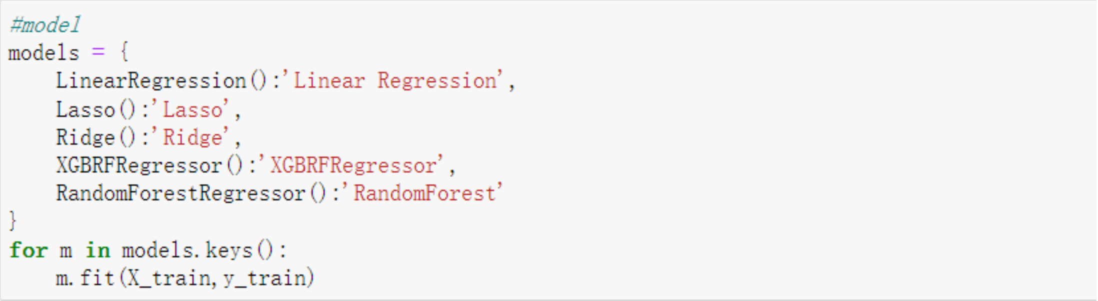
Distribution of height:



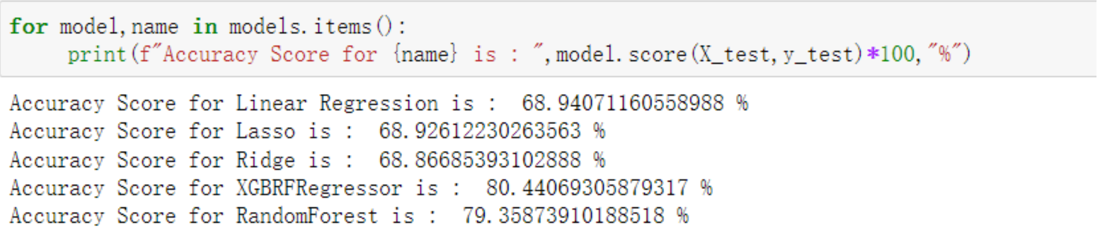
Dependent and independent feature split

Daata normalization

Train test split



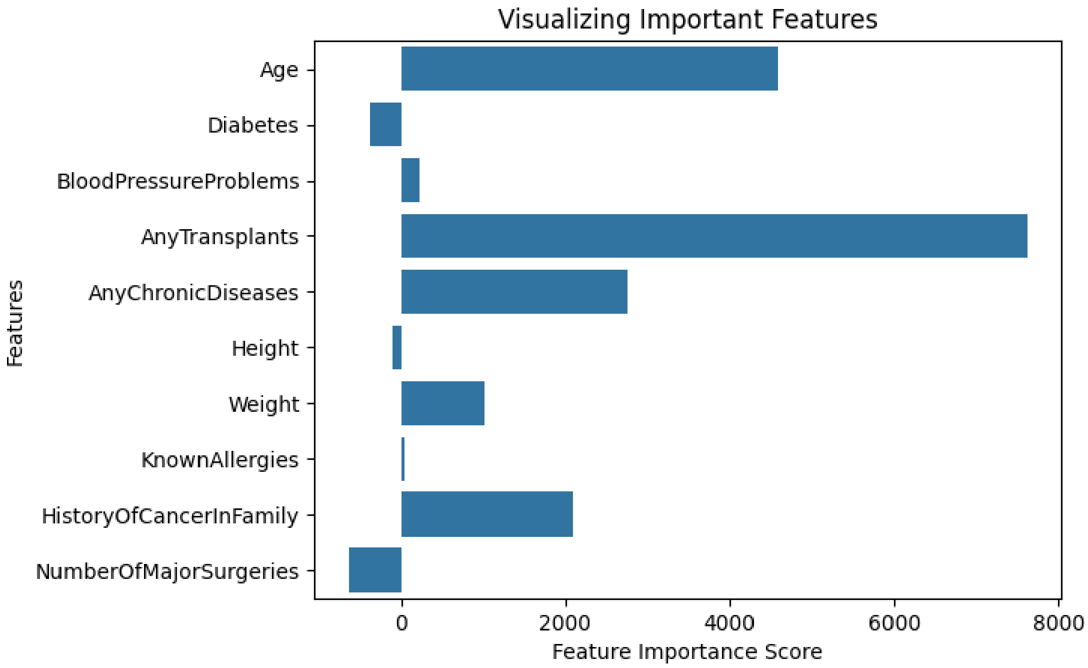
```
#model
models = {
    LinearRegression():’Linear Regression’,
    Lasso():’Lasso’,
    Ridge():’Ridge’,
    XGBRFRegressor():’XGBRFRegressor’,
    RandomForestRegressor():’RandomForest’
}
for m in models.keys():
    m.fit(X_train,y_train)
```



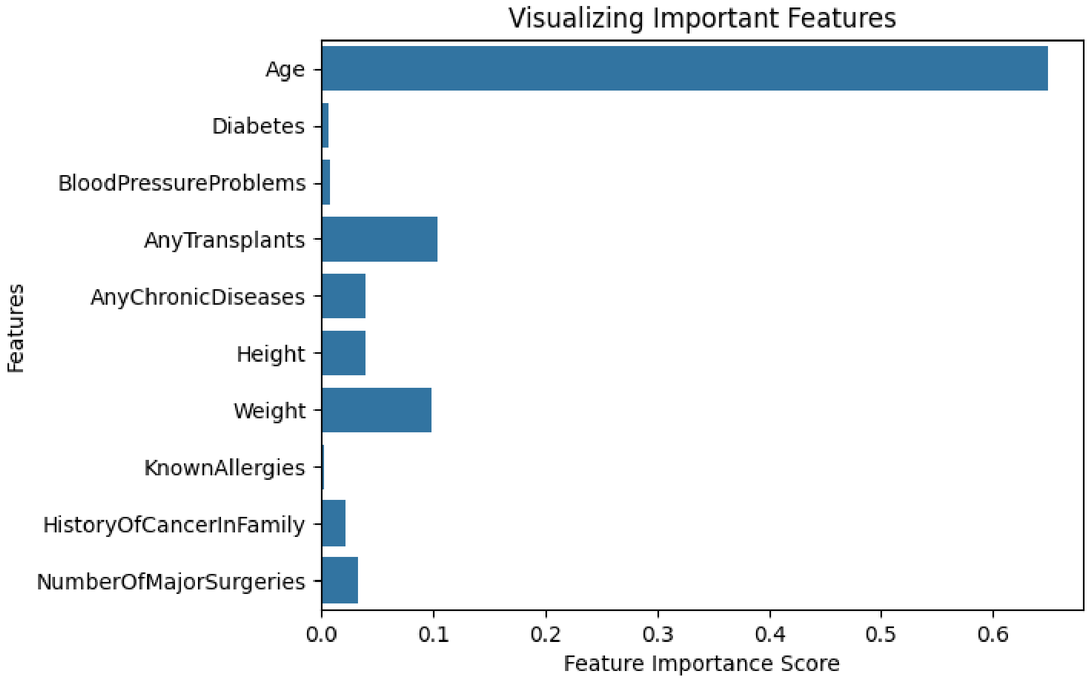
```
for model,name in models.items():
     print(f”Accuracy Score for {name} is : ”,model.score(X_test,y_test)*100,”%”)
```

```
Accuracy Score for Linear Regression is :  68.94071160558988 %
Accuracy Score for Lasso is :  68.92612230263563 %
Accuracy Score for Ridge is :  68.86685393102888 %
Accuracy Score for XGBRFRegressor is :  80.44069305879317 %
Accuracy Score for RandomForest is :  79.35873910188518 %
```

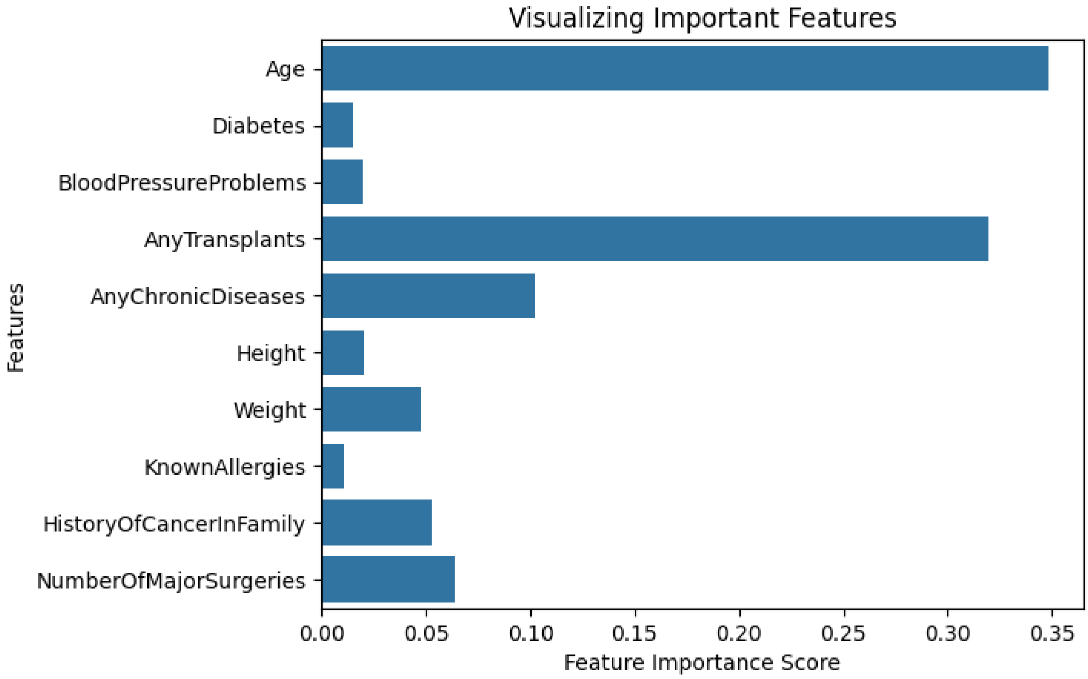
Find important feature through linear regression:



Visualize important features through random forest regressor:



Visualize important features through xgboost:


Visualizing Important Features
Features
Age
Diabetes
BloodPressureProblems
AnyTransplants
AnyChronicDiseases
Height
Weight
KnownAllergies
HistoryOfCancerInFamily
NumberOfMajorSurgeries
Feature Importance Score