SPARK ASSIGNMENT

Ashish Singh (as9tq)

MOTIVATION:

I developed a binary classification model on a dataset that contained various features of a person related to his professional and personal life that could be used to gauge how much does he earn every year. The applications of the model can be various, for example, it could be used by credit and banking agencies to confirm the credit score and establish the credibility of the documents submitted by an individual.

The dataset had a lot of different variable but I selected few that made sense to me and try to develop a model on top of them.

EXPLANATION:

```
from pyspark.ml.feature import VectorAssembler
from pyspark.ml.linalg import Vectors
```

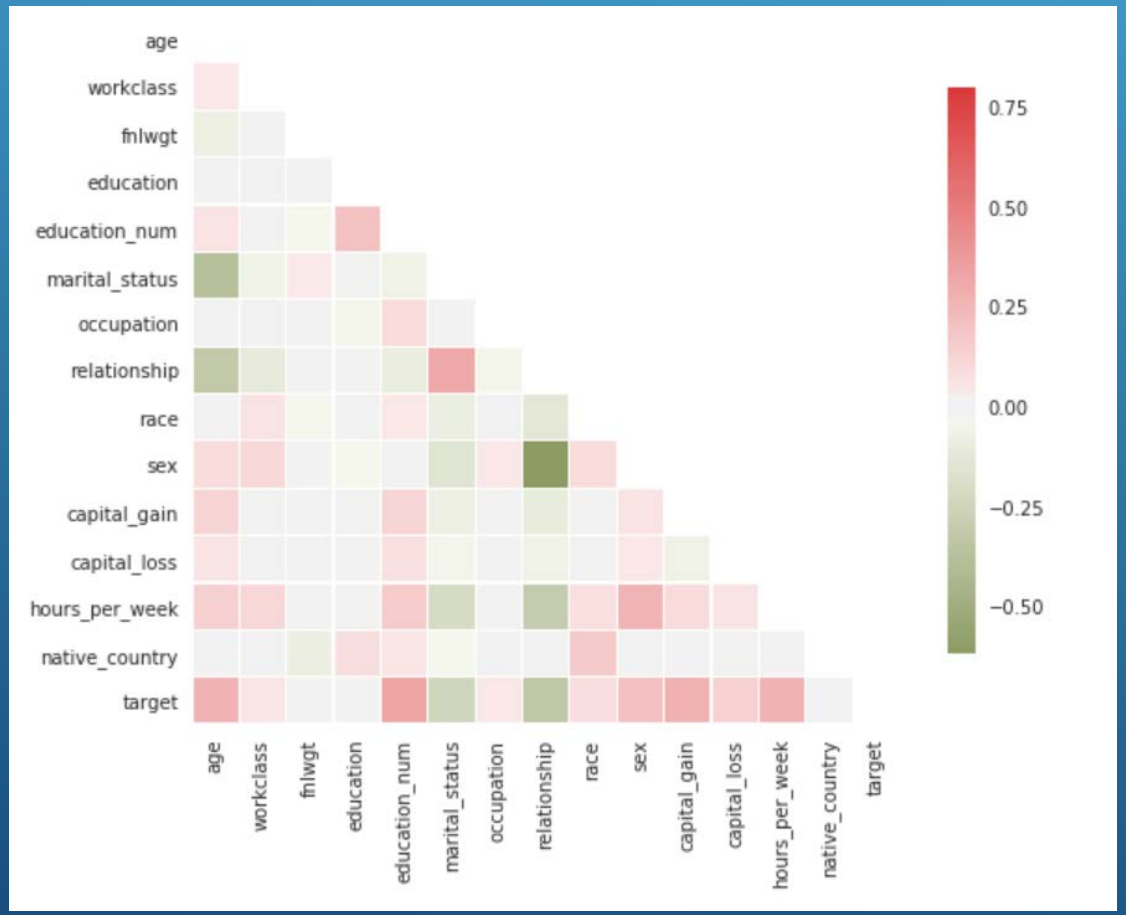
```
assembler = VectorAssembler(
    inputCols=["age","workclass","fnlwgt","education","education_num","marital_status","occu
    outputCol="features")
```

```
output = assembler.transform(df)
output.select("features", "target").show(truncate=False)
```

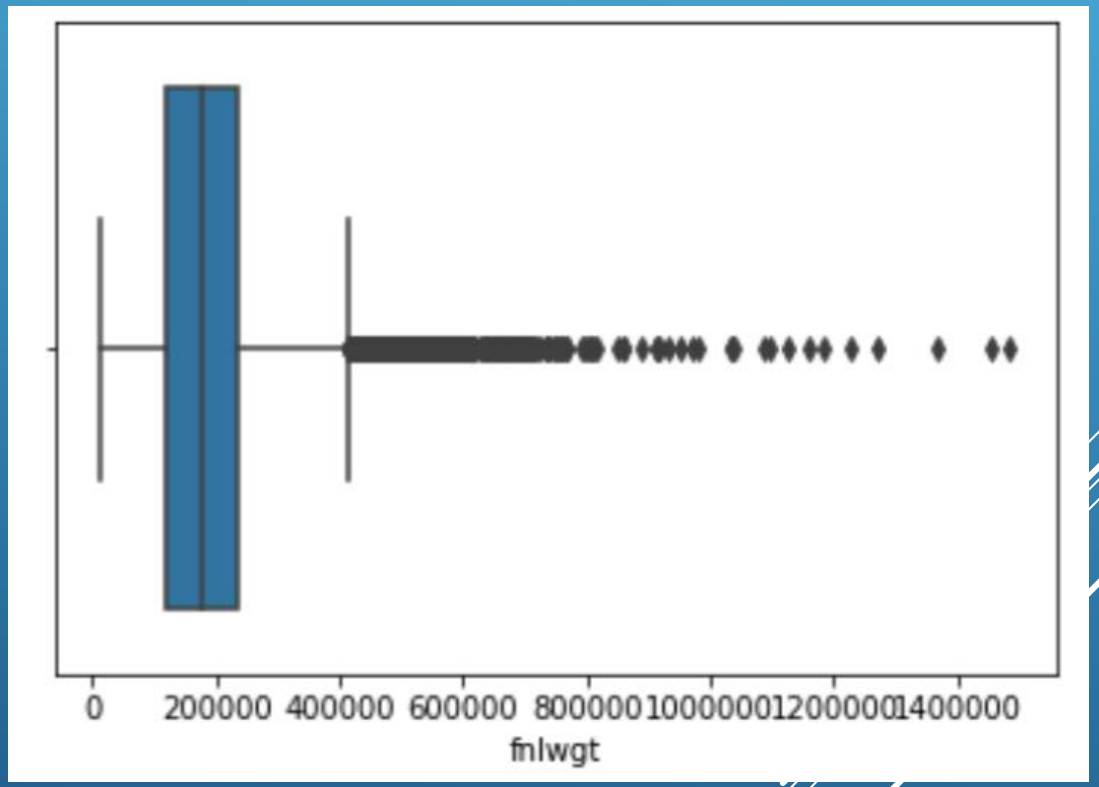
I followed almost the same sequence of code as taught in the class except a few changes:

1. I did some visualization of the features to see if there are any outliers
2. I used Vector Assembler to combine all the feature and transform them
3. I developed a binary classification model instead of linear regression model

VISUALIZATION:



Correlation Matrix



Outlier Detection

Area Under Curve: 0.5