# CW Group 16

# Classification and Clustering

## Exploratory Data Analysis

- We went through the complete data set of Oscar demographics before loading the data set into our environment.
- We then went ahead to carry out data cleansing and visualization.
- Initially we checked for the null values present in the data set.
- We then went ahead and created a new data set with columns such as birthplace, date of birth, race ethnicity, year of award, and the award type.
- We extracted distinct values from the column awards.
- In order to clean the date of birth column we used defined a function in such a way, so we can get the date of birth in the desired format.
- After applying the defined function to the date of birth column, we then changed the format by using datetime.
- From the cleaned date of birth column, we created a new column which is the Birth year. So, this birth Year column contains the year of Birth of each Oscar winner.
- Then we used various cleaning techniques to clean up the Na values in the data set.
- To find out the award age of the user, we created a new column award age which is the difference between the year of award and birth year of the winner.
- To clean the Birthplace column, we used string split function to divide the city, state and country of the Oscar winner.
- We also used a few more techniques to clean the Country of Birth column further and we made a few replacements as well.

## Data Exploration

We Carried out data exploration using appropriate plots to identify patterns or trends in the data

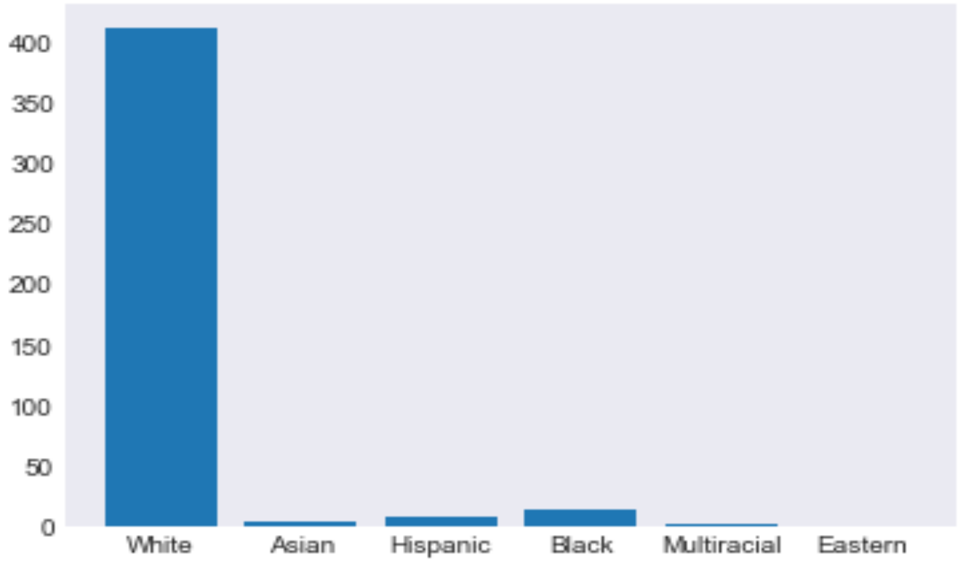
We used graphs to prove/disprove the below hypothesis.

**Most Oscar Winners are white:**

- To prove the following we first extracted the unique values in the race column.
- Using Counter library, we calculated the count value for each individual race.
- Using Matplot library, we used the values obtained using counter function and created an array and used bar plot to show our finding. See below the result

# CW Group 16

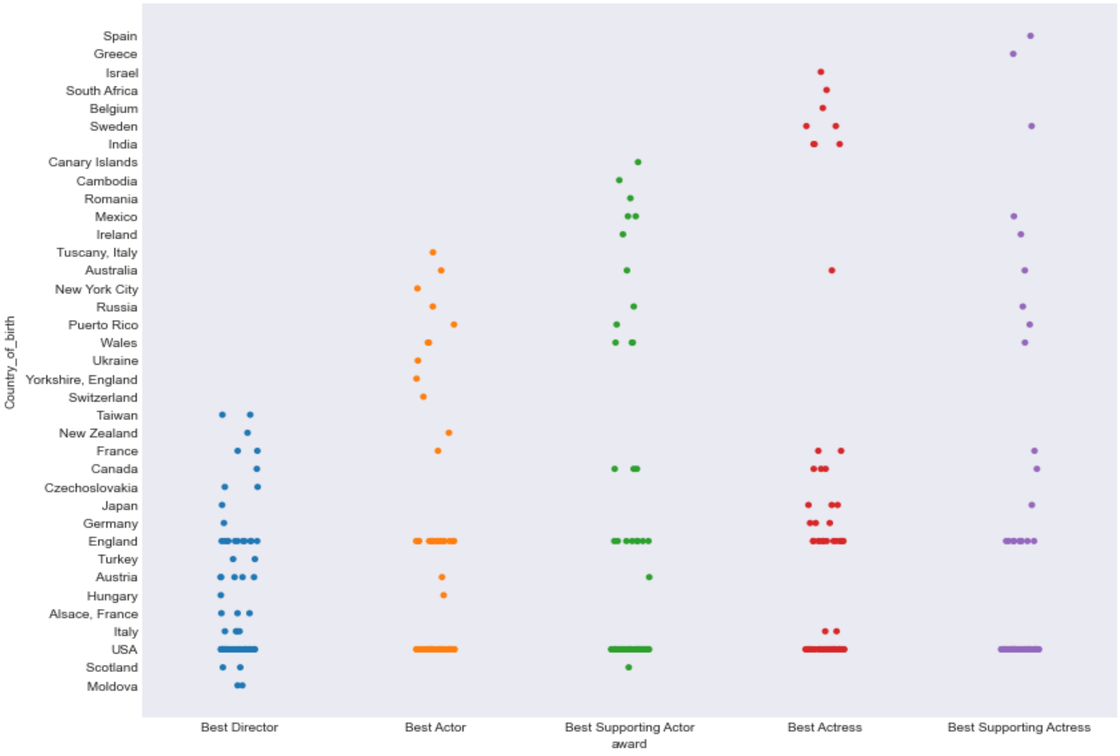
# Classification and Clustering



As you can see the Number of White people winning Oscars is more compared to any other race. So, we can Agree that the **Most Oscar Winners are white.**

## Most Oscar Winner are from USA:

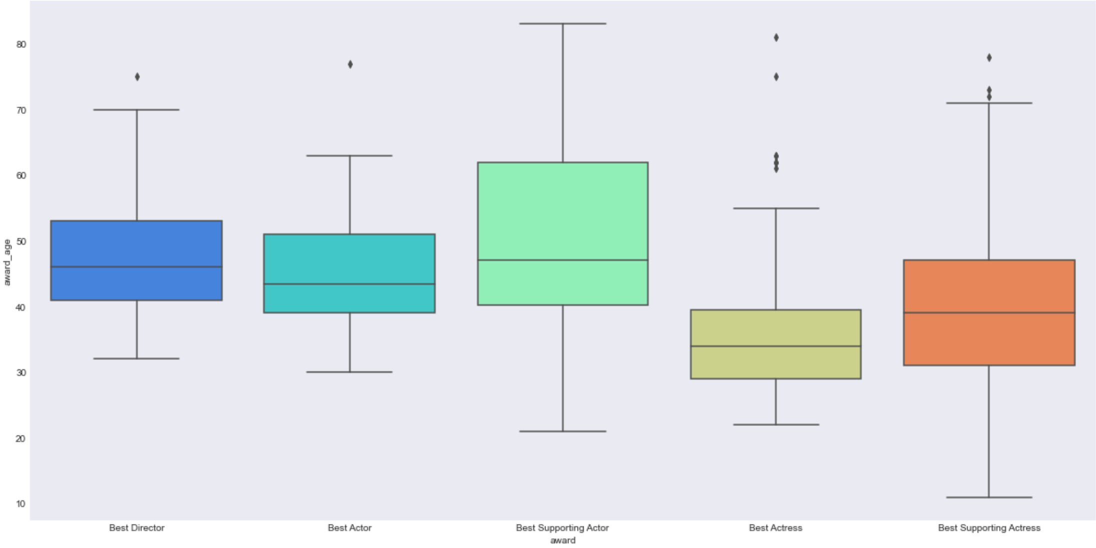
- To prove the following we first extracted the unique values in the Country of birth column.
- By using Seaborn catplot, we assigned X and Y values as award and country of birth and we obtained the below graph.



As you can see from the above graph that most of the Oscar winners are from USA.so we can agree the hypothesis that **Most Oscar Winner are from USA**

### Best Directors tend to be older than best Actors:

- To prove the following hypothesis, we used seaborn box plot to perform the graph.
- We obtained the following graph by giving the X and Y values as award and award age.



As you can see in the above boxplot that the age of Best Director is mostly between 45-55. whereas for best actor it varies from 40-50 and for best actress the age goes from 40-65.

So, we can conclude that the **best Director are not older than the best actor and best actress.**

### Age Buckets.

We have also discrete the age by using4 buckets for each such as <35, 35<age<45, 45<age55 and age>55.

## Modeling

- To perform Model fitting we initially imported the required libraries.
- We performed label encoding to normalize the data points
- We Used Logistic Regression for Model fitting.
- We created a variable called independent and assigned all the predictor to that variable.
- We split the data into train and test sets
- Then we fit the Model for our training set which is the X and Y training set.

# CW Group 16

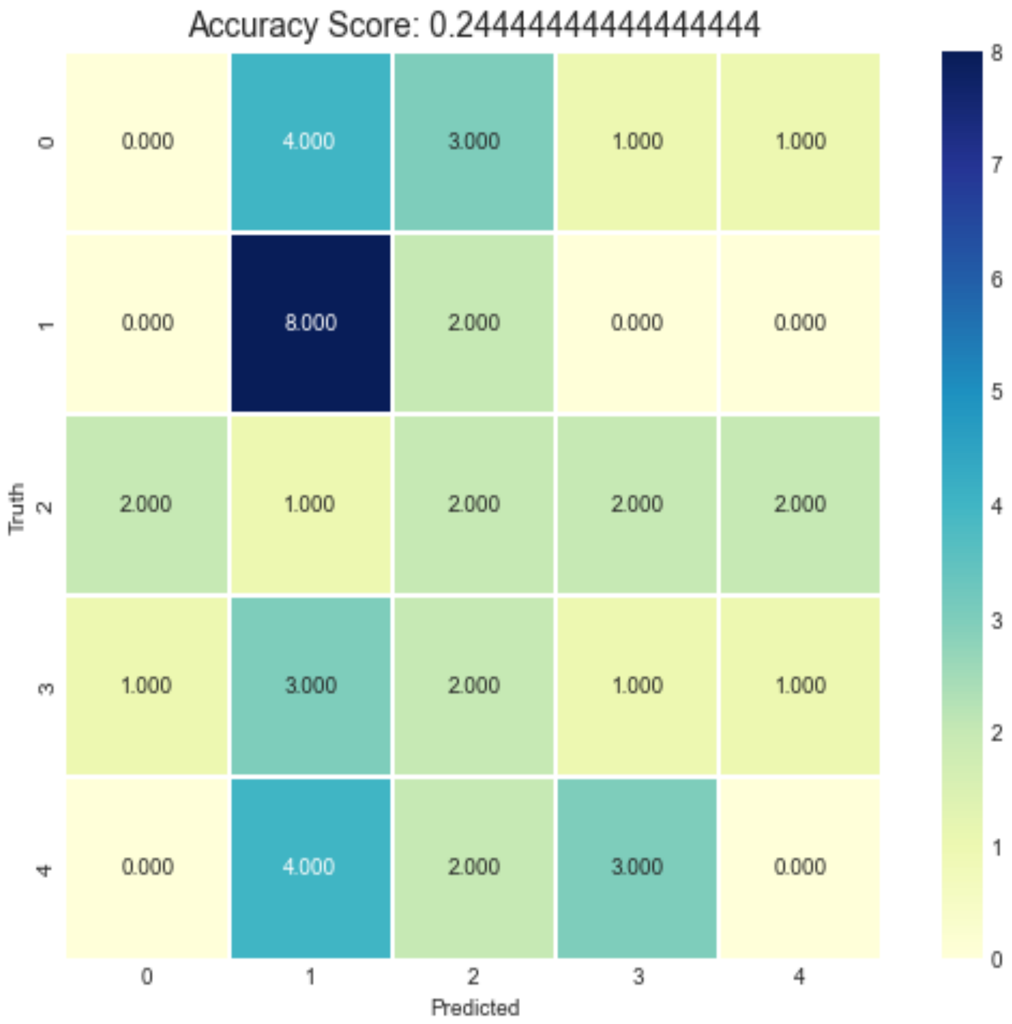
## Classification and Clustering

- We calculated the accuracy score for the training and test set.
- We Predicted the labels for the data using the information the model learned during the model training process by making predictions on the entire test.
- We calculated the accuracy on the test set.
- We computed the confusion matrix for the test and predicted value.

**Confusion Matrix**:

```
array([[0, 4, 3, 1, 1],
       [0, 8, 2, 0, 0],
       [2, 1, 2, 2, 2],
       [1, 3, 2, 1, 1],
       [0, 4, 2, 3, 0]])
```

**We generated the F1 score test and predicted values.**
**Then we used seaborn Heat Map to visualize the confusion Matrix.**



# CW Group 16

## Classification and Clustering

**Using Logistic Regression, we calculated the Mean hits, Accuracy Score, Test Score and cross validation.**

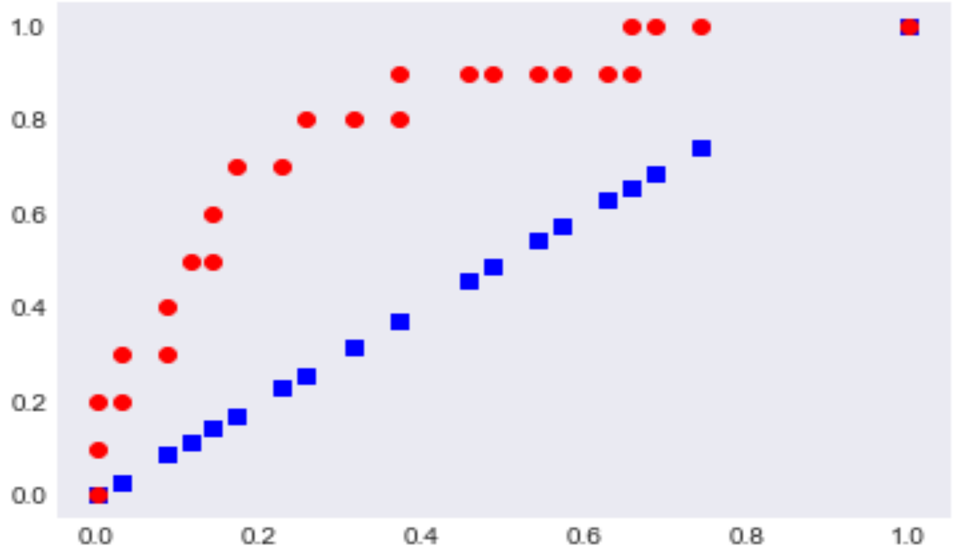
**RESULTS**

1. Mean hits: 0.4
2. Accuracy score: 0.4
3. Test score: 0.4

Cross validation mean scores: 0.3469967532467533

**We then Calculated the AUC (Area Under the curve) Score.**

AUC = 0.8214285714285714



We then used **Random forest classifier** to improve the model my classifying the data. We calculated the accuracy score for training and test data.

**RESULTS**

Accuracy score for training data is: 0.520
Accuracy score for test data: 0.022
The Oob score is: 0.326

**We then computed the Validation score for logistic regression and Random Forest.**

**RESULTS**

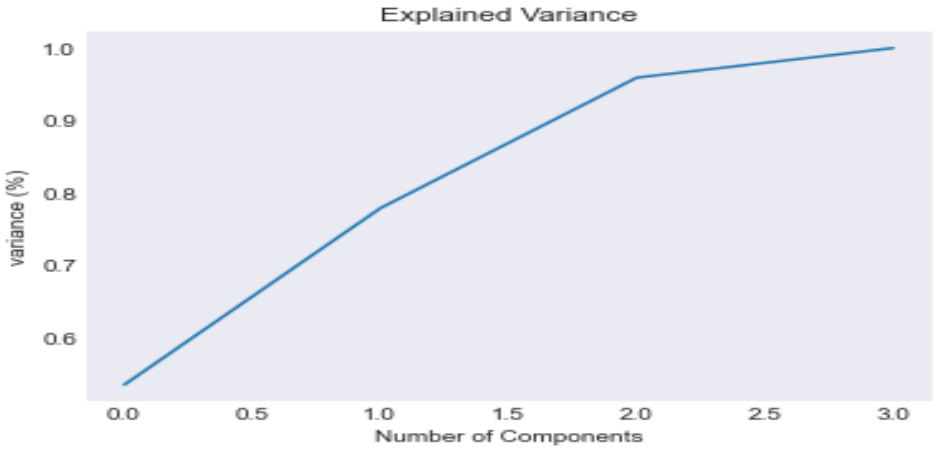
X validation score for Logistic regression is: 0.3377
X validation score for Random Forest: 0.3719

## K Means Clustering

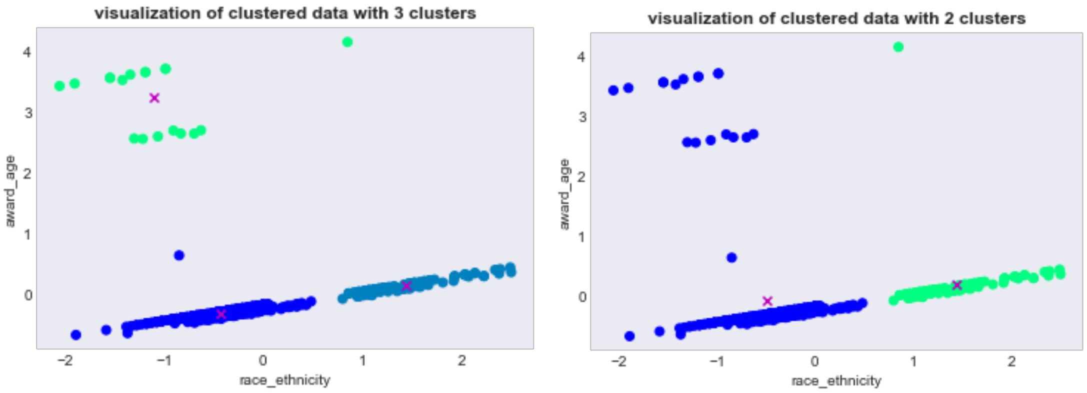
- Using K means clusters we assumed cluster value as 4.
- We defined our indicator list using the predictors.
- Removed values such as zero or missing values.
- We transformed award column into indexes value for each row in the data set.
- We then got the cluster centers.
- We got the labels for clustering.

**We Used principle component analysis within the 2 dimensional clustering.**

Based on the below graph which has the value for Variance, Explained Variance and Number of components. We created an independent data set using the values obtained for each component.



**Below is our Obtained 2D Clusters.**

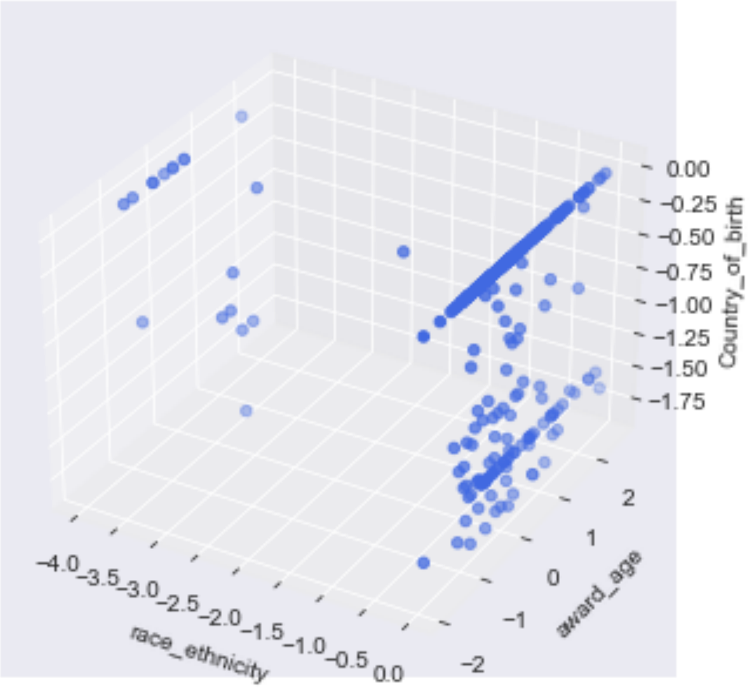


In the obtained 2D cluster the data points where not appearing as expected and we were not able to visualize clustered data with 4 cluster. So, we went for 3D Cluster using Axes3D

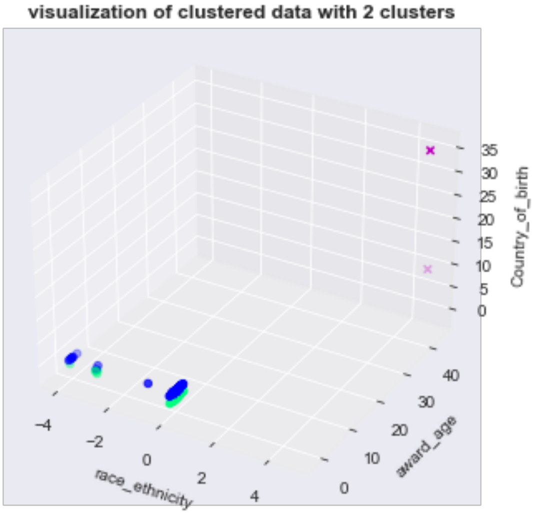
# CW Group 16

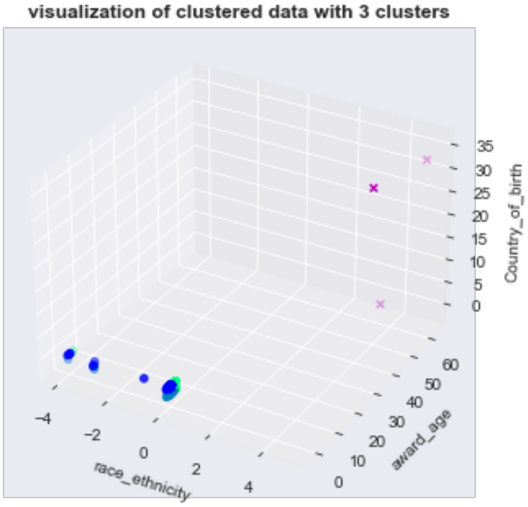
## Classification and Clustering

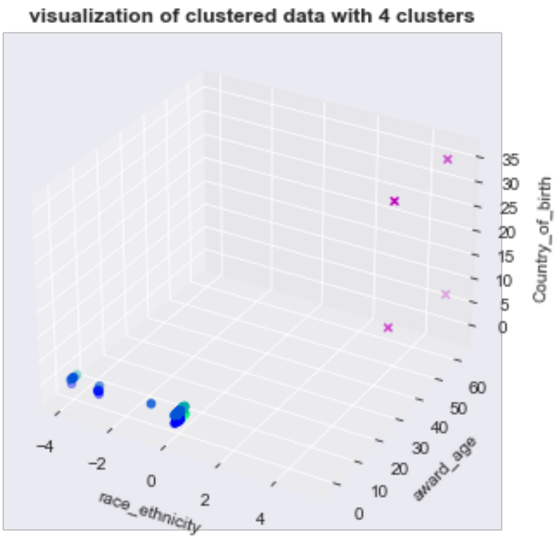
**3D which includes all the predictors.**



Below is our Obtained 3D cluster for 2,3 and 4 Cluster.


visualization of clustered data with 2 clusters


visualization of clustered data with 3 clusters



## Conclusion

- Modeling helped us to predict the outcome of award time in a effective way.
- Using Modeling we were able to get the Accuracy score for our test and training set.
- Using Python plotting libraries we were able to prove the given hypothesis.
- We used Clustering to define how well the data is distributed.
- Clustering also helped us to find the similarity with the data.
- 3D clustering helped us to visualize the data in a more clear and conclusive way.