Data Wrangling
And
Predictive Analysis

Purpose of The Project

The basic purpose is to perform Data Wrangling on the given dataset RegularSeasonDetailedResults and then apply clustering techniques to cluster teams and regression techniques to predict the points a team will score in the tournament. For this, we used the datasets generated from Data Wrangling part as well as TourneyCompactResults dataset.

Datasets Used

RegularSeasonDetailedResults.csv used - for Data Wrangling and Cluster Analysis
TourneyCompactResults.csv- used for Regression

Approach

1. Data Wrangling

- Taking the RegularSeasonDetailedResults dataset we first created two datasets, one for winning team and other for losing team based on their corresponding attributes.
- For each dataset, we dropped the columns Daynum, numot and Wloc which were not specific to win or loss.
- We renamed the columns in each dataset keeping common column name in both so that on combining them (Avg_score.csv) by rows to get the average of statistics per team per season. To do this, we used **rbind** and **aggregate** by mean.

2. Clustering

- For clustering, we dropped the columns Season and Team.
- We used two clustering techniques: Hierarchical clustering and K-means clustering.
- *elbow* and *silhouette* methods are used to decide number of clusters

HH-Clustering

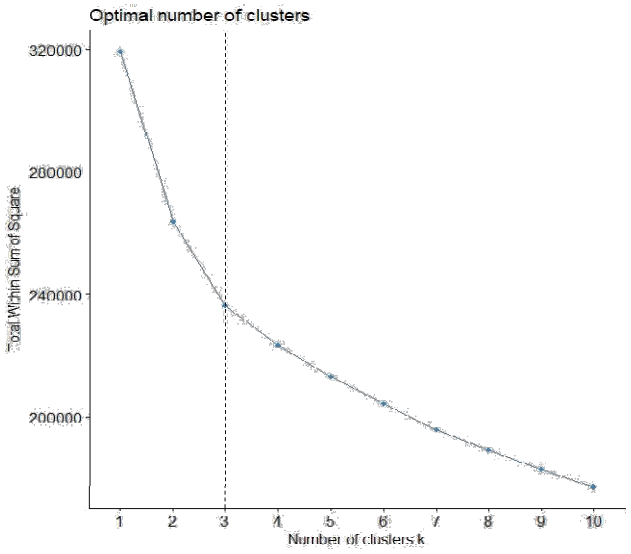

Fig: 1 Elbow Method Graph for HH-Clustering

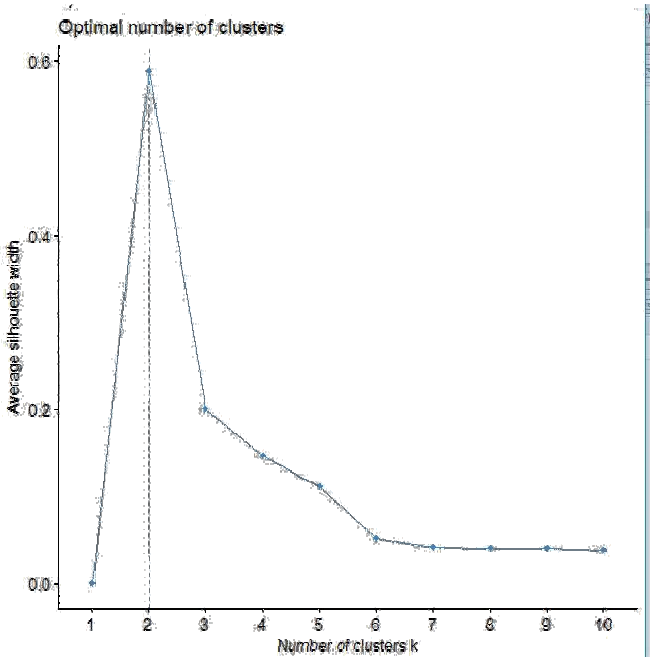

Fig: 2 Silhouette Method Graph for HH-Clustering

From the graphs above we see that according to elbow method 3 clusters are optimal whereas according to silhouette method 2 clusters are optimal.

K-means

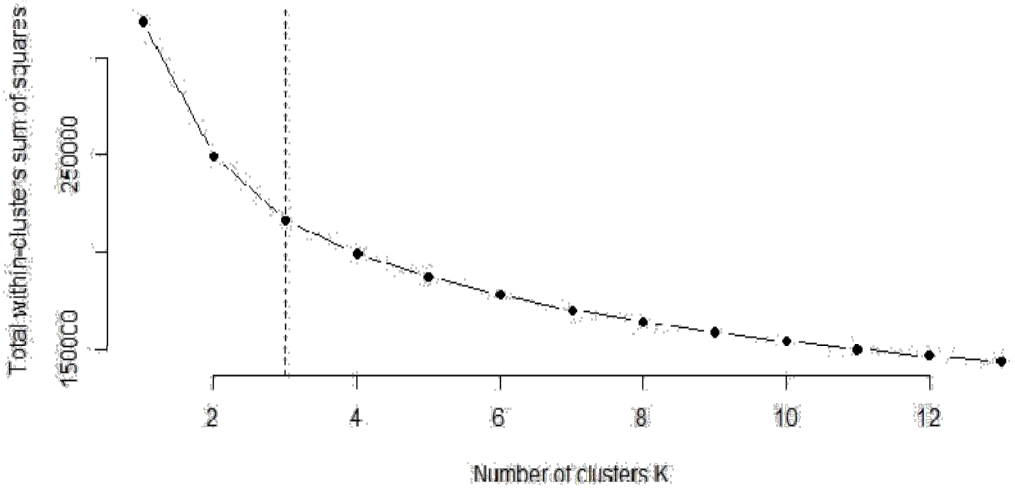

Fig: 3 Elbow-Method Graph for K-means

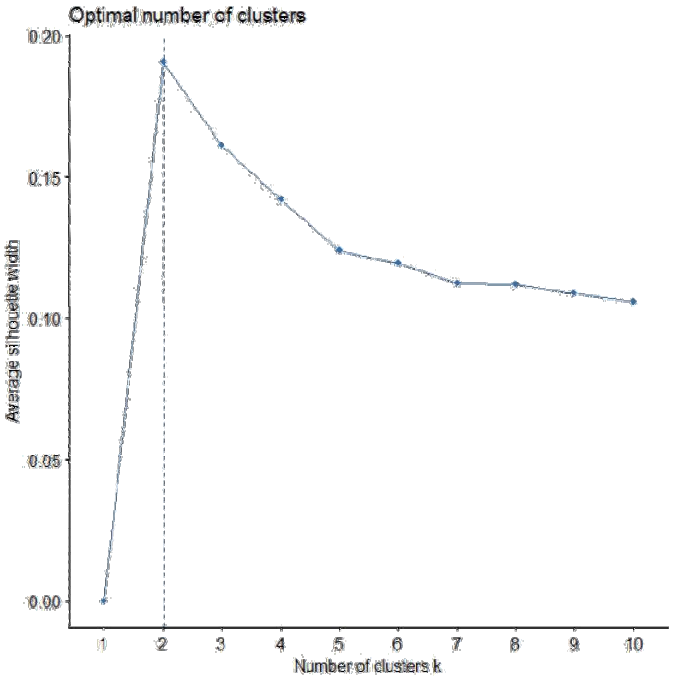

Fig: 4 Silhoutte Method Graph for K-means

From the graphs above we see that according to elbow method 3 clusters are optimal whereas according to silhouette method 2 clusters are optimal.

Using Kmeans on number of clusters = 2,3 and 7 the results are tabulated as shown below. We see that for K=7, our model is **overfitting**, hence, the optimal clusters should be 3.

K	Between_SS / Total_SS
2	22%
3	32.2%
7	46.2%

Table-1: Between_SS/Total_SS values for different K values

To get a better idea of the optimal number of clusters, we applied **PCA** with 3 Principal Components to get better value for Between_SS/Total_SS which turned out to be 50.5%.

3. Regression

- Taking the TourneyCompactResults dataset we first created two datasets, winning_tourney and losing_tourney based on their corresponding attributes.
- For each dataset, we dropped the columns Daynum, numot and Wloc.
- We filtered the rows in both the datasets ,keeping only the rows where **season>=2003.**
- We renamed the columns in each dataset keeping common column name as before so that we can combine both of them by rows per team per season. To do this, we used **rbind**.

- We combined this dataset with the myDataSet_avg dataset created in Data Wrangling part.
- We split the data into train and test data in 80:20 ratio respectively.
- First we build linear regression model using **lm** on train data and predicting on test data
- Then we built our linear regression model using **10 fold cross-validation**.
- We also built our model using **glm**.
- We compared Mean Square Errors with other techniques such as **PCR** and regularization techniques such as **Lasso** and **Ridge.**

- **Statistics using lm are as follows:**
  - lm model with data-split (test-train split) statistics
    - **Residual standard error**: 11.09 on 1448 degrees of freedom
    - **Multiple R-squared**: 0.1679,
    - **Adjusted R-squared**: 0.1605
    - **F-statistic**: 22.48 on 13 and 1448 DF, p-value: < 2.2e-16
  - lm model with cross-validation statistics
    - **Residual standard error**: 11.01 on 1814 degrees of freedom
    - **Multiple R-squared**: 0.1616,
    - **Adjusted R-squared**: 0.1556
    - **F-statistic**: 26.9 on 13 and 1814 DF, p-value: < 2.2e-16

MSEs for various models are as follows:

Model	MSE
lm	121.8237
glm	122.1304
PCR	123.3935
Lasso	121.9172
Ridge	121.832

Summary

- By undertaking this project we got an opportunity to apply techniques such as: **Kmeans clustering, hierarchical clustering, PCA, linear regression, PCR, Lasso and Ridge** along with cross validation to a real-life dataset.
- Apart from understanding the working of key techniques, we also learnt the importance of **pre-processing** and **data wrangling** while dealing with multiple related data sources.
- Given the techniques we used, we learnt how the results of K-means can be enhanced by using PCA. As for linear regression, we got better accuracies using PCR, Lasso and Ridge.
- This project also got us acquainted with robustness of R as a programming language dealing with a huge amount of data and machine learning algorithms. It also laid strong foundations of machine learning algorithms in us.