Raunak Sarbajna and Christoph F. Eick

COSC 3337 *"Data Science I"* Fall 2024
Problem Set1
Team Tasks[1]
Second Draft

Last Updated: September 5, 10am

Task1: Exploratory Data Analysis for a Baseball Databank

Remark: This is a first, somewhat preliminary specification of Task1; there might be still ninor changes, corrections and additions in the next 5 days. More details submission instructions should also be available by Sept. 14 or earlier.

Task1 Due: Saturday, Sept. 21, 11:59p (electronic Submission)

Tentative weight: about 23-30% of the points allocated with the course's ProblemSet tasks.

Responsible TA: Raunak

Dataset Link:
https://github.com/RaunakDune/3337_Fall2024_EDA/blob/bf7b9ca17ed5e08ff25677c329ea1473cb935b49/Baseball_Databank_Teams_1871_2023_Modded.csv

Learning Objectives:

1. Learn how to manage and preprocess datasets and how to compute basic statistics and to create basic data visualizations (using R/Python or other tools)
2. Learn how to interpret popular displays, such as histograms, scatter plots, box plots, density plots,… and to interpret basic statistics.
3. Get some practical experience in exploratory data analysis
4. Learn how to create background knowledge for a dataset
5. Learn to distinguish expected from unexpected results in data analysis and data mining—in general, this task is quite challenging, as it requires background knowledge with respect to the employed data mining technique, and also practical experience.

[1] Collaboration with other teams in the course is not allowed!

Baseball Databank is a compilation of historical baseball data in a convenient, tidy format, distributed under Open Data terms. The Sean Lahman Baseball Databank is a collection of baseball statistics for every team and player in Major League history. Starting in 1995, he made this database freely available for download from the Internet, helping to launch a new era of baseball research and analytics by making the raw data available to everyone.

The goal of this project is to perform exploratory data analysis for the Processed Baseball Databank, which is a modification of the original Baseball Databank by Sean Lahman. The *Processed Baseball Databank* is a (34+1)D dataset with a nominal `Target` attribute added; the attributes of this dataset, their meanings are listed below, and the range of values:

1. **yearID**: Year [1871, 2023]
2. **name**: Team's full name
3. **Rank**: Position in final standings [1, 13]
4. **G:** Total Number of Games **[6,165]**
5. **W**: Wins [0,116]
6. **L**: Losses [4,134]
7. WP: Winning Percentage [0,0.87]
8. **WSWin**: World Series Winner (Y or N)
9. **R**: Runs scored [24, 1220]
10. **AB**: At bats [211, 5781]
11. **H**: Hits by batters [33, 1783]
12. **2B**: Doubles [1, 376]
13. **3B**: Triples [0, 150]
14. **HR**: Homeruns by batters [0, 307]
15. **BB**: Walks by batters [0, 835]
16. **SO**: Strikeouts by batters [3, 1654]
17. **SB**: Stolen bases [0, 581]
18. **CS**: Caught stealing [0, 191]
19. HBP: Batters hit by pitch [7, 160]
20. SF: Sacrifice Flies [7, 77]
21. **RA**: Opponents runs scored [34, 1252]
22. **ER**: Earned runs allowed [23, 1023]
23. **ERA**: Earned run average [1.22, 8]
24. **CG**: Complete games [0, 148]
25. **SHO**: Shutouts [0, 32]
26. **SV**: Saves [0, 68]
27. **IPOuts**: Outs Pitched (innings pitched x 3) [162, 4518]
28. **HA**: Hits allowed [49, 1993]
29. **HRA**: Homeruns allowed [0, 305]
30. **BBA**: Walks allowed [1, 827]
31. **SOA**: Strikeouts by pitchers [0,1687]
32. **E**: Errors [20, 639]
33. **DP**: Double Plays [0, 460]
34. **FP**: Fielding percentage [0.761,0.991]
35. **TARGET**: Winning Percentage in Categorical Order.

The first 3 examples of the dataset are listed below:

yearID	***name***	***Rank***	***G***	***W***	***L***	***WP***	***R***	***AB***	***H***
1871	Boston Red Stockings	3	31	20	10	0.65	401	1372	426
1871	Chicago White Stockings	2	28	19	9	0.68	302	1196	323
1871	Cleveland Forest Citys	8	29	10	19	0.34	249	1186	328

2B	***3B***	***HR***	***BB***	***SO***	***SB***	***CS***	***HBP***	***SF***	***RA***	***ER***	***ERA***
70	37	3	60	19	73	16			303	109	3.55
52	21	10	60	22	69	21			241	77	2.76
35	40	7	26	25	18	8			341	116	4.11

CG	***SHO***	***SV***	***IPouts***	***HA***	***HRA***	***BBA***	***SOA***	***E***	***DP***	***FP***	***Target***
22	1	3	828	367	2	42	23	243	24	0.834	HIGH
25	0	1	753	308	6	28	22	229	16	0.829	HIGH
23	0	0	762	346	13	53	34	234	15	0.818	LOW

The values of the class attribute `Target` have been computed from the `WP` attribute as follows: (0.63,1]→HIGH, (0.4,0.63] → AVERAGE, [0…0.4]→LOW; In general, we are interested in predicting Attributes 7 (WINNING PERCENTAGE) and 35 (TARGET) using the other attributes; that is, we like to predict which of the collected baseball statistics contribute most to the games won by the team. Another subject we are interested in is finding relationships between the attributes in the dataset, and to understand what factors influences successful teams the most.

<u>Task1 Subtasks</u>:

Apply the following exploratory data analysis techniques **<u>using R [Preferred] or Python</u>** to your dataset:

0. Use the *Processed Baseball Databank* dataset created by the TA or clean up the raw dataset yourself!

1. Compute the covariance matrix for each pair of the following attributes:

   `R (Runs Scored)`

E (Errors Per Game)
HR (Homeruns by Batters)
RA (Opponents Runs Scored)
SOA (Strikeouts by pitchers)

Next, compute the correlations for each of the pairs of attributes. Interpret the statistical findings! **4 points**

2. Create scatter plot for the attribute pairs `AB/H (At bats vs Hits by batters)`and `HA/BBA (Hits allowed vs Walks allowed).` Interpret the two scatter plots**! 4 points**

3. Pick any two teams at random, in addition to the Houston Astros. Create two sets histograms/bar plot for each team for the 10-year periods `yearID = [2004…2013] and [2014…2023]` for the High, Average and Low `Target` classes; interpret the obtained histograms. **6 points**

4. Create box plots for the `BB (Walks by Batters) and SB (Stolen Bases)` attributes for the instances of the 3 Target class— low/average /high — and a third box plot for all instances in the dataset. Interpret and compare the box plots for each attribute! **4 points**

5. Create supervised scatter plots for the following 3 pairs of attributes using the Target as a class variable: `HB/SO`, `CG/SHO` and `IPOuts/DP`. Use different colors for labelling the class variable. Interpret the obtained plots; in particular, address what can be said about the difficulty in predicting the `Target` attribute and the distribution of the instances of the two classes. **6 points**

6. Create 2 density plots for each instance of the 3 Target classes in the `W(Wins)Percentage vs E (Errors) Per Game` space. Compare the density plots! **6 points**

7. Create a table of all the teams who won the World Series (**WSWIN** = Y). Add three columns counting how many times each class of the Target attribute each Team obtained. Create histograms for the `W (Wins)` and `L (Losses)` attributes for the instances of each of the teams. Interpret the table and the histograms you created. **8 points**

8. Create a new dataset *<u>Z-Processed Baseball Databank</u>* from the *Processed Baseball Databank* dataset by transforming the `H (Hits), SO (Strikeouts by Batters), SOA (Strikeouts by Pitchers), SHO (Shutouts) and FP (Fielding Percentage)` attributes into z-scores. Fit a linear model that predicts the values of the `WP (Win Percentage)` attribute using the 5 z-scored, continuous attributes as the independent variables. Report the R^2 of the linear model and the coefficients of each attribute in the obtained regression function. What do the obtained coefficients tell you about the importance of each attribute for predicting a successful team? **6points**

9. Create 3 decision tree models with 25 or less nodes for the dataset (both intermediate and leaf nodes count; do not submit models with more than 25 nodes!); use the `Target` attribute as the class variable, and use ONLY Attributes 9 upto 34 of the dataset excluding all other attributes of the dataset when building the decision tree model. Explain how the 3 decision tree models were obtained! Report the training accuracy and the testing accuracy of the submitted decision trees. Interpret the learnt decision tree. What does it tell you about the importance of the chosen attributes for the classification problem? **11 points**

10. Write a conclusion (at most 13 sentences!) summarizing the most important findings of this task; in particular, address the findings obtained related to predicting a successful team (both by Rank and Target) using attributes 7-30. If possible, write about which attributes seem useful for predicting good teams and what you as an individual can learn from this dataset! **6 points (and up to 4 extra points)**

Remark: About 30-40% of the Task1 points will be allocated to interpreting statistical findings and visualizations!

Submission Guidelines Task1[2]: *Submit on MS Teams.* When you submit your task 1 for problem set 1, you should submit a word file/pdf that displays your graphs and your interpretations. **DO NOT ZIP.** Each interpretation should use complete sentences to describe your findings. If you're using ChatGPT/LLM model, please mention what you're using it for! Also in the folder, you should include all files used to complete your tasks, such as your R or python files. If you have doubts about what to submit send Raunak an e-mail.

[2] More detailed submission instructions for Task1 will be added to this specification by Sept. 18, 2024 the latest.