Introduction to Data Science: Project II

Instructions to students:

1. This is an **individual based** assignment.

2. Submission: The student must prepare a Jupyter Notebook file that includes for each step: (a) code to carry out the step, (b) output showing the output of the code, and (c) a short description of how the code works. In case of providing plots, the student has to provide a short description (2-3 sentences) of what the intent of the plot is (the student has to think in terms of variation, co-variation, central trend, spread, skew, etc.), a short text description of the plot, and a sentence or two of interpretation of the plot (again the student has to think concerning variation, co-variation, etc.). Finally, each student must provide max. 10 slides PowerPoint presentation to present her/his work in a Storytelling way, emphasizing the main findings of each part. The project package (Jupyter, Presentation, etc.) submission is allowed though the LMS (e-learning) only.

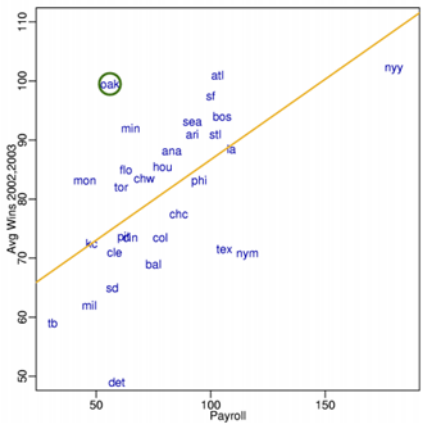
3. Assessment: Assessment will be on the Jupyter Notebook submitted, in addition to recorded PowerPoint presentation. Note that the recorded presentation is a mandatory requirement. In the recorded presentation, each student should emphasize his views, regarding the approaches and techniques she/he adopted, and the results obtained in a Story Telling way.

4. Feedback: Personalized feedback will be given through discussions. However, final feedback will be provided through the LMS (eLearning).

INTRODUCTION

In this project, the student will apply data wrangling and exploratory data analysis skills to baseball data. In particular, we want to know how well did Moneyball work for the Oakland A's. Was it worthy of a movie?

In the early 2000's the Oakland A's were winning as much as teams with much bigger payrolls by evaluating players using data differently than other teams.



Therefore, the student will be provided with background about Major League Baseball, USA, along with the business perspective concerning the return on investment in terms of the number of winning games to be acquainted with the context of this project. The very useful database on baseball teams, players and seasons curated by Sean Lahman will be used throughout this project. The student can read more about the dataset here: http://seanlahman.com/files/database/readme2014.txt.

By accomplishing this project, the student will be able to give insight on how efficient teams have been historically at spending money and getting wins in return. In the case of Moneyball, one would expect that Oakland was not much more efficient than other teams in their spending before 2000, were much more efficient between 2000 and 2005 (they made a movie about it after all), and by then other teams may have caught up. Therefore, the student should provide an insight to see how this is reflected in the provided data.

PART ONE: Data Wrangling (6 Marks)

The student will be asked to use SQL to compute a relation containing the total payroll and winning percentage (number of wins/number of games * 100) for each team. The student should describe how she/he dealt with any missing data in these two relations.

Besides, the student should include other columns that will help when performing Exploratory Data Analysis later.

PART TWO: Exploratory Data Analysis (12 Marks)

Payroll distribution (6 Marks)

The student will be asked to illustrate the distribution of payrolls across teams conditioned on time (from 1990-2014). Then, she/he should comment on this distribution of payrolls conditioned on time based on the plots in terms of central tendency, spread, etc.? The student should provide evidence for the provided statements.

Correlation between payroll and winning percentage (6 Marks)

The student will be asked to write code to discretize years into five time-periods and then make a scatterplot showing the mean winning percentage (y-axis) versus the mean payroll (x-axis) for each of the five time-periods. (4 Marks)

The student should comment on the team payrolls across these periods. Then, she/he has to answer the questions “Are there any teams that stand out as being particularly good at paying for wins across these periods? What can you say about the Oakland A's spending efficiency across these periods?” (2 Marks)

PART THREE: Data Transformations (12 Marks)

Standardizing across years (4 Marks)

Since comparing payrolls across years is misleading, so the student will be asked to do data transformation (*Eqn 1.*) that will help with these comparisons. (2 Marks)

$$standardized\_payroll_{ij} = \frac{payroll_{ij} - avg\_payroll_j}{s_j} \qquad \text{(Eqn 1.)}$$

for the team *i* in year *j*, where $avg\_payroll_j$ is the average payroll for year *j*, and s_j is the standard deviation of payroll for year *j*.

Then, the student will be asked to plot the winning percentage (y-axis) versus the transformed (standardized) payroll and discuss the relationship with respect to the graph before the data transformation. (2 Marks)

Expected wins (4 Marks)

It's hard to see global trends across time periods using these multiple plots, but now that we have standardized payrolls across time, we can look at a single plot showing the correlation between winning percentage and payroll across time.

In this way, the student will be asked to make a single scatter plot of the winning percentage (y-axis) vs. the standardized payroll (x-axis) and add a regression line to highlight the relationship. (2 Marks)

The regression line gives the expected winning percentage as a function of standardized payroll. Then, the student will be asked to analyze the results with respect to the expected winning percentage (Eqn. 2). (2 Marks)

$$expected\_win\_pct_{ij} = 50 + 2.5 \times standardized\_payroll_{ij} \quad \text{(Eqn. 2)}$$

for the team *i* in year *j*.

Spending efficiency (4 Marks)

Using the result of the previous requirements, the student can now create a single plot that makes it easier to compare teams' efficiency (Eqn. 3). The idea is to create a new measurement unit for each team based on their winning percentage and their expected winning percentage that we can plot across time summarizing how efficient each team is in their spending. (2 Marks)

$$efficiency_{ij} = win\_pct_{ij} - expected\_win\_pct_{ij} \quad \text{(Eqn. 3)}$$

for the team *i* in year *j*.

Concerning the student's experience throughout the performed analyses, the student will be asked to make use of such experience and answer the question "How good was Oakland's efficiency during the Moneyball period?" (2 Marks)