Regression for NBA Players Metrics and The Annual Salary

Background:

- **Dataset:** 2017-18 Advanced Player Metrics + Salary (https://www.kaggle.com/meicher/201718-advanced-player-metrics-salary/data)
- This dataset consists of advanced and technical metrics of NBA players in season 2017-18. Each player is an observation (sample), and his stats are the variables (features). For instance, age, team, nationality, number of the games he played, annual salary, e.t.c. For the sake of simplicity, we only take the most vital ones into consideration.
- **Variables (inputs)**:
  - Age
  - G (number of games)
  - MP (minutes played)
  - PER (player efficient rating: The rating sums up all of a player's positive accomplishments, subtracts the negative accomplishments, and returns a per-minute rating of said player's performance. The average PER across the entire league is always set to 15.)
  - WS (winshare: an estimate of the number of wins contributed by a player in a season)
  - BPM (Box Plus/Minus: A box score estimate of the points per 100 possessions that a player contributed above a league-average player, translated to an average team.)
- **Labels (outputs):** Annual salary
- **Samples:**
  - 531 players in total (after data cleaning)
  - Each player is represented by a 6-entry vector.
  - However, the content of the real input vectors might vary, and is dependent on the model we use and the experiments we conduct.

Research Question:

From a high-level point of view, we would like to explore the relationship between this professional stats of an NBA player and his salary. This is the major we would like to answer. However, we also list the technical questions to demonstrate the entire pipeline of our research project.

- **Major question:** How does the relationship between the statistics of a player and his salary look like?
- **Technical questions:**
  - How does single variable affect the salary?
  - How does the combination of certain variables affect the salary?
  - How does the order of the regression model affect the salary?
  - Is linear regression model expressive enough for this dataset?

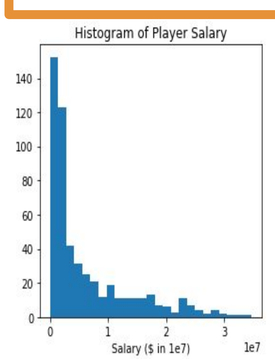
Method:

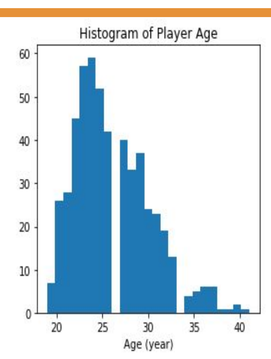
- **Main Method:**
  - **Linear Regression** a linear supervised learning approach that aims to model the relationship between inputs and continuous output. In our project, the outputs are the annual salary of a player, and it could be considered as a continuous variable, which is suitable for linear regression. Therefore, we perform regression methods to fit the distribution of the dataset throughout this work.We would examine it in lower to higher orders, and with single or multiple variables.

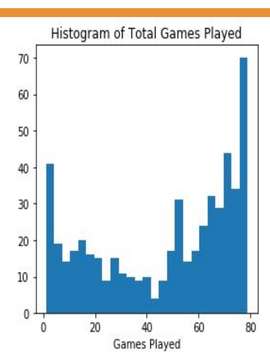
- **Implementation:**
  - First-order Linear Regression with single variable
  - First-order Linear Regression with multivariable
  - Higher-order Linear Regression with multivariable
  - The reason we are not applying more linear regression models here is based on the consideration of representativeness and time cost. First order of both single variable and multivariable Linear Regression would be enough for comparison between single variable and multivariable models. If it works better on multivariable models, we would perform higher-order Linear Regression with multivariable, vice versa.

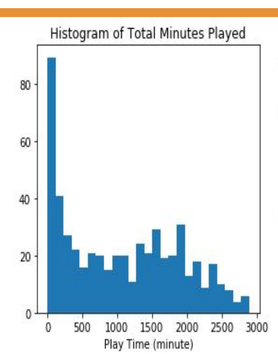
- **Supporting Analysis**:
  - Cross Validation: To avoid overfitting, we perform cross validation for all training processes. Moreover, we can visualize the statistics (MSE/SSE) of training set and test set after training, which helps us to evaluate the performance of the trained models on unseen test data.

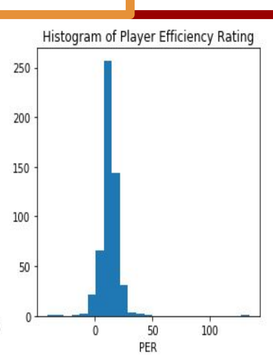
- **WHY NOT?**
  - Why not using PCA?
    PCA is used to compress the data so that we would be able to represent the target with less information. A good example is PCA on images. However, it is obvious that in our case compress the data would be meaningless. One player only have limited data with him, and compressing the data would make prediction unreal, and lose its representativeness.
  - Why not using Clustering?
    There is no identical relationship of clustering here. NBA Players cannot be separated into groups. Each position is equally important, so we cannot cluster them by position. Maybe we can cluster them by team, but there is too much teams and too less players, making the clustering way too unstable. Moreover, we lack the actual-world meaning of predict which team does a NBA player belongs to.

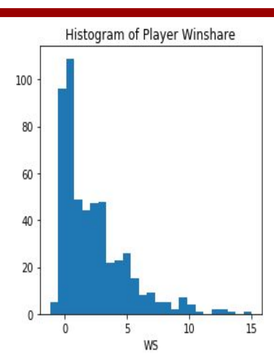


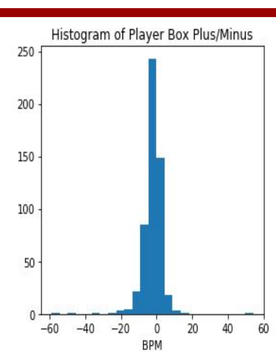




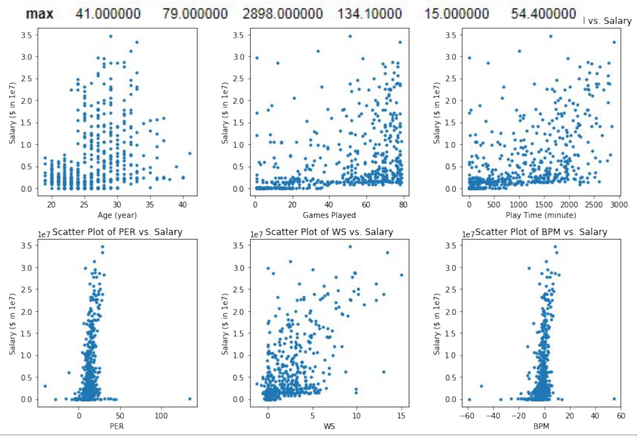








	Age	G	MP	PER	WS	BPM
count	531.000000	531.000000	531.000000	531.00000	531.000000	531.000000
mean	26.071563	46.892655	1067.429379	12.89774	2.253672	-2.080038
std	4.166189	26.176645	826.639558	8.96386	2.638646	6.078032
min	19.000000	1.000000	1.000000	-41.10000	-1.200000	-59.000000
25%	23.000000	20.000000	240.000000	9.30000	0.200000	-3.950000
50%	25.000000	54.000000	1014.000000	12.90000	1.400000	-1.500000
75%	29.000000	70.000000	1753.000000	16.20000	3.350000	0.400000
max	41.000000	79.000000	2898.000000	134.10000	15.000000	54.400000



Result:

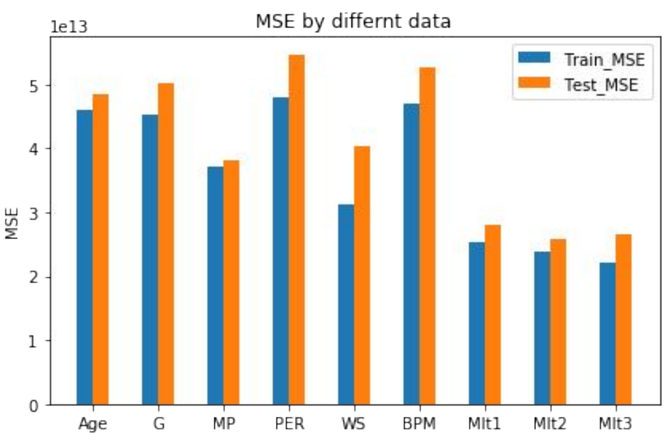
- **Visualization**
  - Histogram of Each Variable (Shown on the top)
    We created Histogram for each variable, including the target. As we could see, no variable is evenly spreaded. Most People in NBA is receiving relatively low salary, this tendency is the same for playing time. Players' age is like a normal distribution. Their player efficient rating and Box plus/minus are centered. Their win share skews left, and their games played skews to two polars.
  - Data Frame Description (Shown on the top right)
    We then took a look at some basic description of the input variables, there is nothing too special here because we cannot find out the relationship simply by looking at the datas.
  - Scatter Plot Between All Input Variables and the Output Variable (Shown as the second figure on the right)
    Scatter Plot shows something more than Histogram, that the dataset is sparse, very sparse. By raw prediction it might be hard to find out tendency with the tools we learned in the class. Moreover, we cannot find any division of multiple cluster with in one graph.

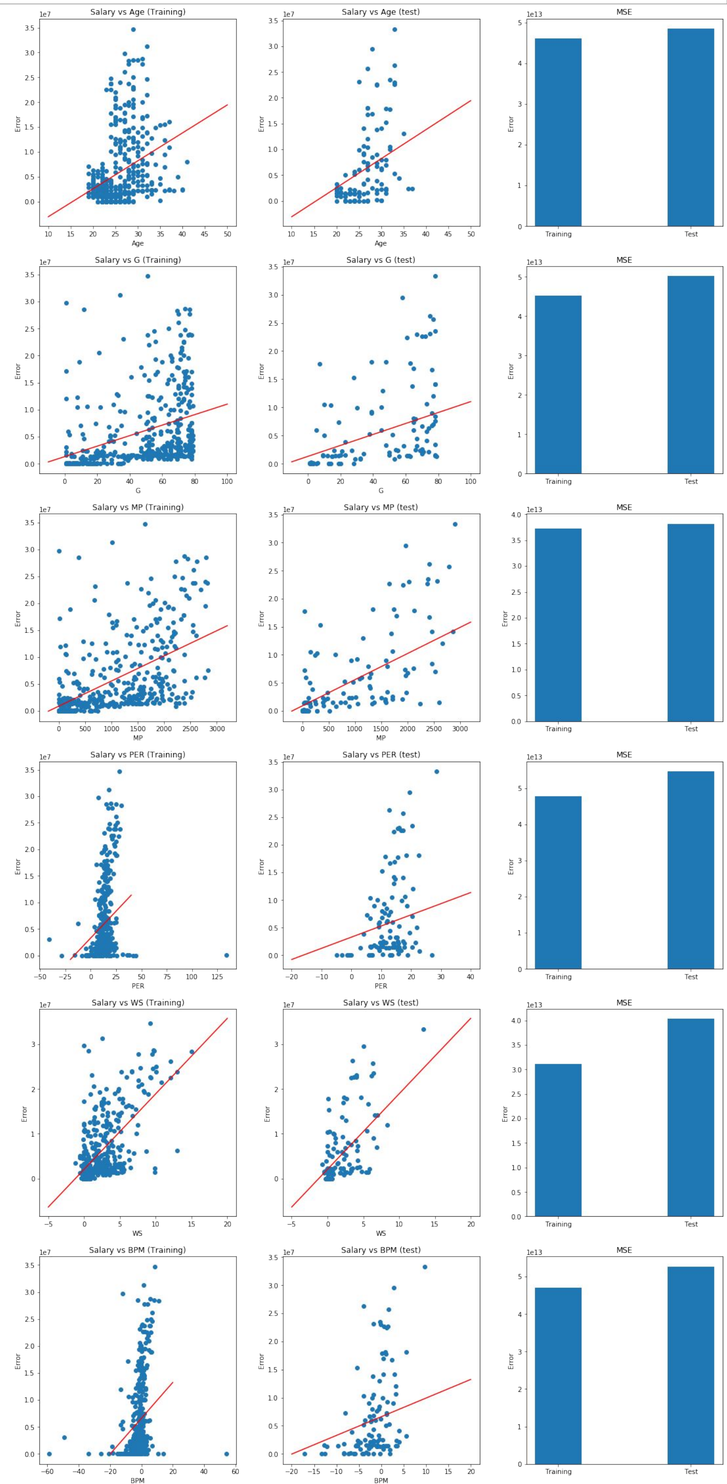
- Single Variable Linear Regression (Report of Model is posted as the table on the right, Printed Plot is the last figure on the right)
  - As we could seen, no matter which variable we choose, it would be barely hard to find a actual correlation between the variable and the salary because the data is too sparse. This has further cause the MSE of linear regression to become very large. This is true even if there is better fit and worse fit.

- Multivariable Linear Regression (Report of Model is posted as the table at the bottom, Printed Plot is the right figure at the bottom)
  - Mlt1 denotes first order, Mlt2 denotes second order, and Mlt3 denotes third order. As we could have seen, no matter which order we choose, the MSE is still pretty large. Although by using the multivariate model the MSE is nearly halved, but it still is very large.

Age	Salary = -8698754.32382574 + 562764.64270265 * Age
G	Salary = 1302986.19586075 + 97216.85467595 * G
MP	Salary = 900809.2719447 + 4665.70713391 * MP
PER	Salary = 3273179.93812688 + 202111.62086122 * PER
WS	Salary = 2062778.17394801 + 1684734.23106476 * WS
BPM	Salary = 6597299.27871832 + 331872.99070467 * BPM

First Order	Salary = -8.71699485e+06 + 4.79835523e+05 * Age + -1.23132427e+05 * G + 4.80011760e+03 * MP + 1.58972214e+04 * PER + 1.17321937e+06 * WS + 1.18557039e+04 * BPM
Second Order	Salary = -4.07180227e+07 + 2.66836914e+06 * Age + -3.64836212e+04 * G + 8.08017755e+03 * MP + 9.76495728e+04 * PER + 3.13757893e+05 * WS + -4.40991544e+04 * BPM + -3.92958769e+04 * Age ^ 2 + -1.27194878e+03 * G ^ 2 + -8.01954344e-01 * MP ^ 2 + -7.24415590e+02 * PER ^ 2 + 6.87362512e+04 * WS ^ 2 + 3.82052401e+02 * BPM ^ 2
Third Order	Salary = 7.67992462e+07 + -1.01311280e+07 * Age + -2.56657314e+05 * G + 2.04440714e+04 * MP + 1.39264864e+05 * PER + -8.08413573e+05 * WS + 3.29034761e+04 * BPM + 4.17155180e+05 * Age ^ 2 + 2.71544923e+03 * G ^ 2 + -1.06237375e+01 * MP ^ 2 + -2.34130668e+03 * PER ^ 2 + 3.50680870e+05 * WS ^ 2 + 9.52476292e+03 * BPM ^ 2 + -5.29866853e+03 * Age ^ 3 + -2.34290776e+01 * G ^ 3 + 2.20862571e-03 * MP ^ 3 + -1.15905347e+01 * PER ^ 3 + -1.67934427e+04 * WS ^ 3 + 1.29257978e+02 * BPM ^ 3





Discussion:

When carrying out linear regression with a single input variable on first order each time, we found the results to be less than ideal. The plot lacks homogeneity and the data points don't always fit the linear projection very well. However, the plots show a positive correlation between each of the input variables and the output variable as all first order models have a significant positive w1 value. **Minutes Played** has the least significant coefficient value while **Winshare** has the most significant coefficient value. Each of the testing set shows a slight increase in its MSE compared to that of the training set, which is within expected range, but could be slight risk of overfitting the data.

Implementing a multivariate linear regression gives us more insights on which features have the most significant correlations with player salary. The first order multivariate regression suggests that **Games Played** has the only negative correlation, and **Winshare** has the most significant positive correlation with player salary. The second and third order multivariate regression show correlations that are relatively difficult to pinpoint to a single feature being the most significant one. Furthermore, all three multivariate methods show a large MSE when switching from the training set to testing set, showing signs of overfitting the data.

In conclusion, the order of model does not seem to matter much here, as all three multivariate models produced a similarly high level of MSE. In the end, **Winshare** seems to be the metric that is most significantly correlated with player salary among all statistics of a player.

Future plans: include more features (example: 3pt attempt rate and 3pt field goal %), collect data from multiple seasons, etc.