P2: Investigate a dataset

Baseball(MLB) data analysis and visualization

Datasets analyzed & Variables used in this project:

	Salaries table	**Batting Table**	**Pitching Table**	**Teams table**
Variables used	yearID	yearID	yearID	yearID
	playerID	playerID	playerID	G(Games played)
	salary	G(Games)	G(Games)	Rank(Position in final standings)
		AB(At Bats)	ERA(Earned Run Average)	W(Wins)
		H(Hits)		BPF(3y park factor for batters)
		BB(Base on Balls)		HR(Homeruns by batters)
		HBP(Hit by pitch)		ERA(Earned run average)
		SF(Sacrifice flies)		
		HR(Homeruns)		
		RBI(Runs Batted In)		

All the datasets used are downloaded from http://www.seanlahman.com/baseball-archive/statistics/

Questions:

1. Is there a strong positive correlation between a player's salary and batting performance?

1.1. Different parameters for batting performance were analyzed:

- Calculating batting average: BA (batting average) = H(Hits)/AB(At Bats)
- Calculate On-base percentage: On Base Percentage is calculated by adding hits, walks, and hit-by-pitches and dividing by the sum of at bats, walks, hit by pitches, and sacrifice flies: OBP=(H+BB+HBP)/(AB+BB+HBP+SF)

1.2. Spearman's rank correlation between a player's different batting metrics and salary were analyzed:

- Due to the amount of data, a function called ***<Which_year>*** is created to determine which year's data to explore. I am exploring the year of 2015 in this project.
- Only the batters whose AB (at bats) is no less than 130 times in 2015 were considered in this analysis. (AB>=130) is used to rule out the data from rookies.

1.3. Data Wrangling Steps:

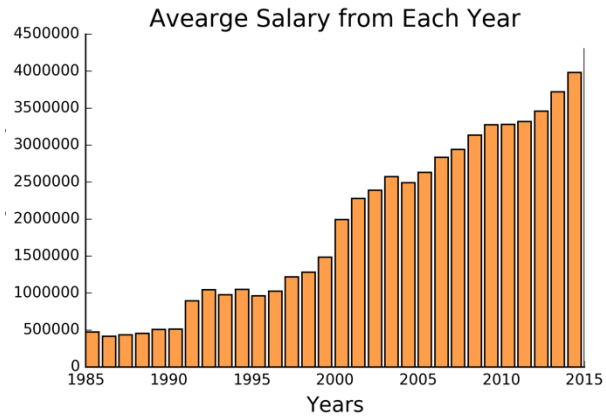
- Merge salaries table with pitching table based on the playerID, yearID, teamID and lgID.
- Discard the rows with NA values and with AB less than 130.

- Create new batting metrics BA and OBP. They were saved as new columns in the original dataframe.
- Perform correlation analysis. Since the salary variable is not normally distributed, Spearman's rank correlation was used to explore the relationship between player's salary and different performance metrics, and Pearson's correlation was used to examine the relationship between different performance metrics.

1.4. Results & Data visualization

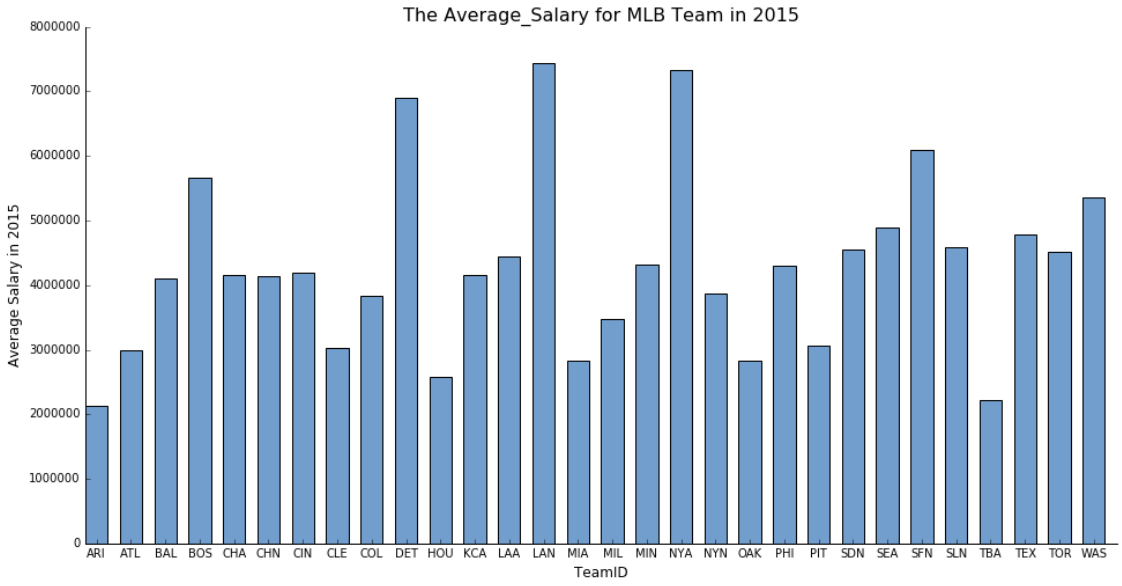
1.4.1. Single-variable (1d) exploration

- Variable: player's average salary from 1985 – 2015



This figure indicates that the average salary of MLB players increased steadily each year. Especially for the last 15 years, the growth rate is nearly linear. It implies the sustained increasing popularity of baseball.

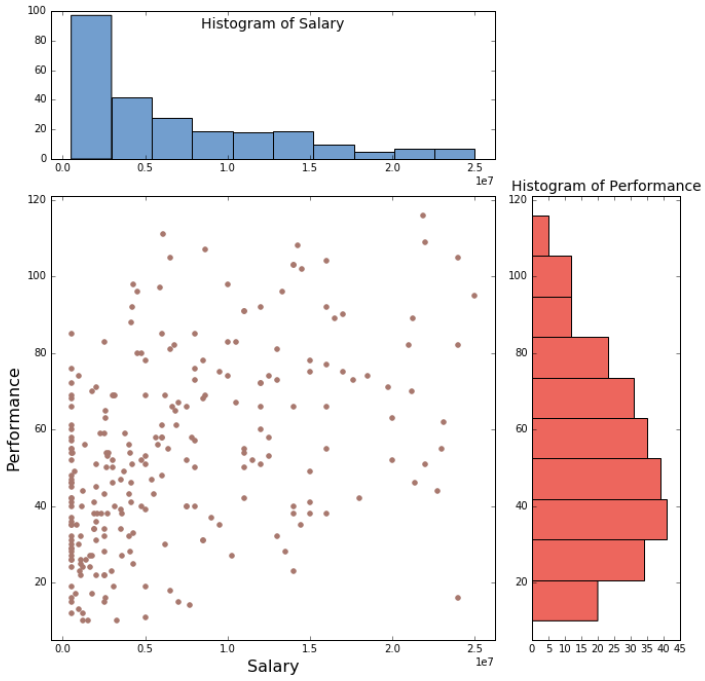
- Variable: Player's average salary for each team in 2015



This figure shows that the variances of average salary between different teams are large. For example, in 2015, the average salary of NYA is almost two times higher than TBA.

1.4.2. Multiple-variable (2d) explorations

- A function called **<*Multiplots*>** is used to visualize the 2D explorations between any two variables.
- Variables: Player's salary in 2015; player's RBI in 2014



This figure is composed of the 1d exploration of batter's salary in 2015, 1d exploration of batter's RBI in 2014 and 2d exploration of the two variables. The histogram on the top (blue) shows the distribution of batter's salary in 2015. It is highly skewed to the left (lower salary), which might be caused by the big variance between different team's average salary. The histogram on the right (red) shows the distribution of batter's RBI in 2014. It could be considered as normal distribution, although skewed a little bit to the right (higher RBI). The scatter plot in the middle shows the relationship between player's RBI in 2014 and salary in 2015. They are positively correlated (The Spearman correlation coefficient is 0.456645).

- Correlation Analysis

Spearman Correlation Coefficient	Players Salary in 2015
Player's BA in 2015	*0.185445*
Player's BA in 2014	*0.217967*
Player's RBI in 2015	*0.375105*

Player's RBI in 2014	***0.456645***
Player's OBP in 2015	*0.241051*
Player's OBP in 2014	*0.340708*
Player's HR in 2015	*0.351503*
Player's HR in 2014	*0.331338*

Pearson Correlation Coefficient	Player's RBI in 201	Player's OBP in 2015
Player's RBI in 2015		0.498064
Player's HR in 2015	***0.885034***	0.441864

1.5. Conclusion

1.5.1. First, I investigated the average salary (mean) of all the players from each year. The bar plot indicates a linear yearly growth trend of salary. The economic growth and the increasing popularity of baseball might be the key reasons for the sustained salary growth.

1.5.2. Second, I explored different team's salary in 2015. The variance is pretty large, and the average salary from the highest teams could be two times more than the lowest teams.

1.5.3. Third, I analyzed the correlations between batter's salary and batting performance. Before performing the analysis, I hypothesized that a player's batting performance determines his salary in the next year. BA (batting average), OBP (On Base Percentage), RBI (Runs Batted In) and HR (Homeruns) were used to represent a player's batting performance. The larger those values are, the better the player performs. Because the salary level is not normally distributed, Spearman's Rank Correlation was used to calculate the correlation between different batting metrics and salary. Pearson's Correlation was used to calculate the correlation between different batting metrics. However, the results disapprove my hypothesis. There is only a moderate positive correlation between baseball players' batting performance and salary level. What's more, the players' salary in current year is slightly more correlated with the batting performance from previous year. Finally, there is a strong correlation between a player's HR and RBI.

1.5.4. Correlation does not imply causation. The correlation analysis I did here does not account for all the variables that define the relationship between a batter's salary and performance. Other factors, such as different teams, years played in MLB and value of a contract would all affect the salary. However, it reveals a trend that a batter's RBI is likely to affect his salary in the next year. It also disproved my hypothesis that a player's batting performance determines his salary in the next year.

What's more, the association between HR and RBI is a necessary step along the path to establishing causality. It makes sense that higher HR would lead to better RBI scores.

2. Is there a strong positive correlation between a player's salary and his pitching performance?

2.1. ERA (Earned Run Average) was used to represent a player's pitching performance.

- The smaller the ERA value is, the better the pitcher performs.

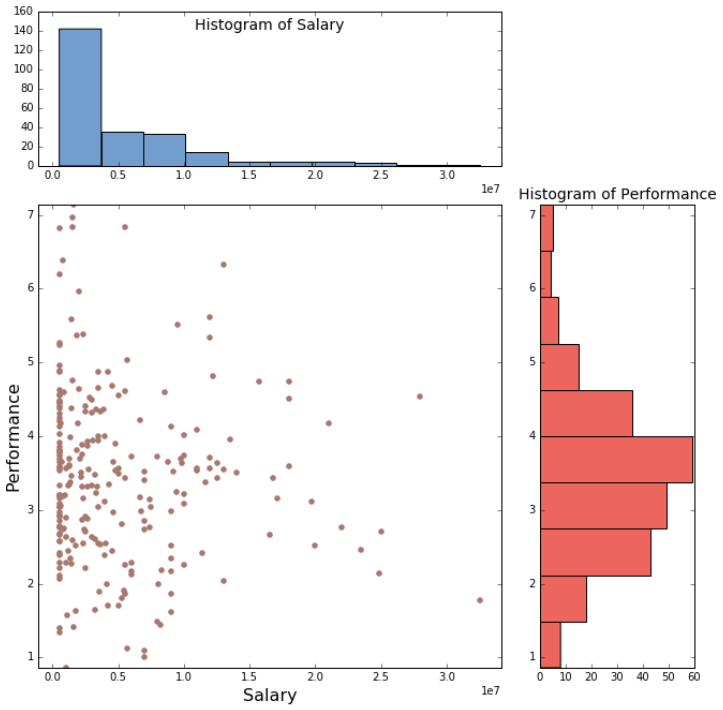
2.2. Spearman's rank correlation between a player's ERA and salary were analyzed:

- Only the pitchers whose G (Games) is no less than 10 were considered in this analysis to rule out the data from rookies.

2.3. Data Wrangling Steps: same as Question 1.

2.4. Results & Data Visualization

- Variables: Player's salary in 2015; player's ERA in 2014



This figure is composed of the 1d exploration of pitcher's salary in 2015, 1d exploration of pitcher's ERA in 2014 and 2d exploration of the two variables. The histogram on the top (blue) shows the distribution of pitcher's salary in 2015. It is highly skewed to the left (lower salary). The histogram on the right (red) shows the distribution of pitcher's ERA in 2014. It is normal distributed. The scatter plot in the middle shows the relationship between player's ERA in 2014 and salary in 2015. They are not correlated (The Spearman correlation coefficient is -0.098451).

- Correlation Analysis

Spearman Correlation Coefficient	Players Salary in 2015
Pitcher's ERA in 2015	*-0.091880*
Pitcher's ERA in 2014	*-0.098451*

2.5. Conclusion

2.5.1. There is no correlation between a baseball players' pitching performance and salary level (correlation coefficient: $|r| < 0.1$). As stated before, correlation does not imply causation, but it could be used to disapprove a causal theory. Thus, there is no direct causation between a pitcher's ERA and salary.

2.5.2. There are at least several reasons which might lead to this result: First, a pitcher's salary is dependent on the team's salary level or other factors. Second, a pitcher's performance could not be correctly measured by ERA; Third, even for a good pitcher, ERA is a highly unstable variable. Some pitchers might be really good in a season or in a game, however, it might be hard to maintain once the other teams start to analyze his pitching speed and trajectory. When a person becomes good in a certain motor skill, the trajectories of movements are usually highly stereotyped. With the video analyzing techniques, it would become harder and harder for a pitcher to maintain their record in the long run.

Python Code for Q1 & Q2:
https://github.com/super-penguin/Udacity_Data_Analyst/blob/master/P2/Q1_2.py

3. What is the most important factor for a team's success, batting performance, pitching performance or both? Does a team's average salary relate to its success?

3.1. Data Wrangling Steps:

- Calculate the average salary for each team in 2015.
- Merge the team performance table with the average salary based on teamID.
- Perform correlation analysis. BPF (Three-year park factor for batters) was used to represent the batting performance (hitting) of a team and ERA (Earned run average) was used to represent the pitching performance (defense) of a team. Pearson correlation was used to examine the relationship between different team performance and averaged salary.

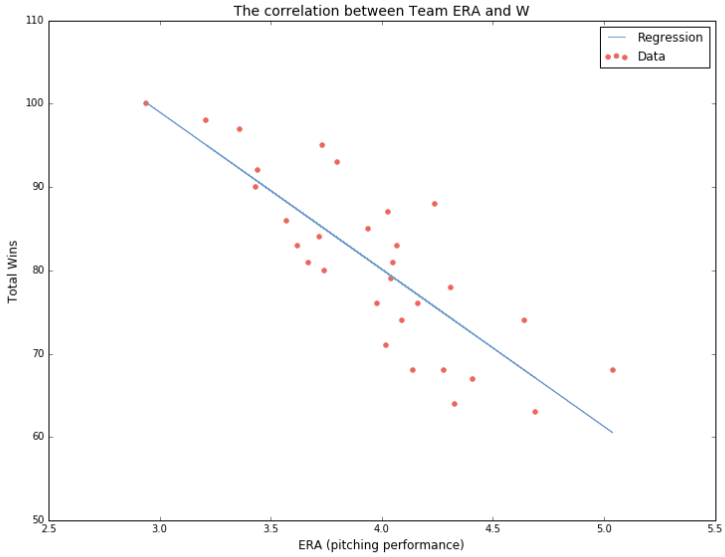
3.2. Results & Data visualization

3.2.1. Correlation Analysis:

Pearson Correlation Coefficient	Team W	Team W/G	Team Rank
Team BPF	*-0.051974*		
Team ERA	***-0.819600***	***-0.819131***	
Team average salary		*0.205825*	*-0.170484*

3.2.2. 2d Explorations

- Variables: Team ERA in 2015; Team W (winning record) in 2015



This plot is the 2d exploration of team's ERA and W. It shows the winning record of a team is highly correlated with the ERA. Smaller ERA represents better pitching performance. The correlation coefficient is -0.819131. It is suggested that a team's success is largely dependent on their pitching performance. The blue line is the linear regression line and the descriptive statistics are showed in the following table.

- OLS Regression Results

==========

Dep. Variable:	*W*	*R-squared:*	*0.672*
Model:	*OLS Adj.*	*R-squared:*	*0.660*
Method:	*Least Squares*	*F-statistic:*	*57.30*
Date:	*Wed, 06 Jul 2016*	*Prob (F-statistic):*	*3.03e-08*
Time:	*11:57:18*	*Log-Likelihood:*	*-95.758*

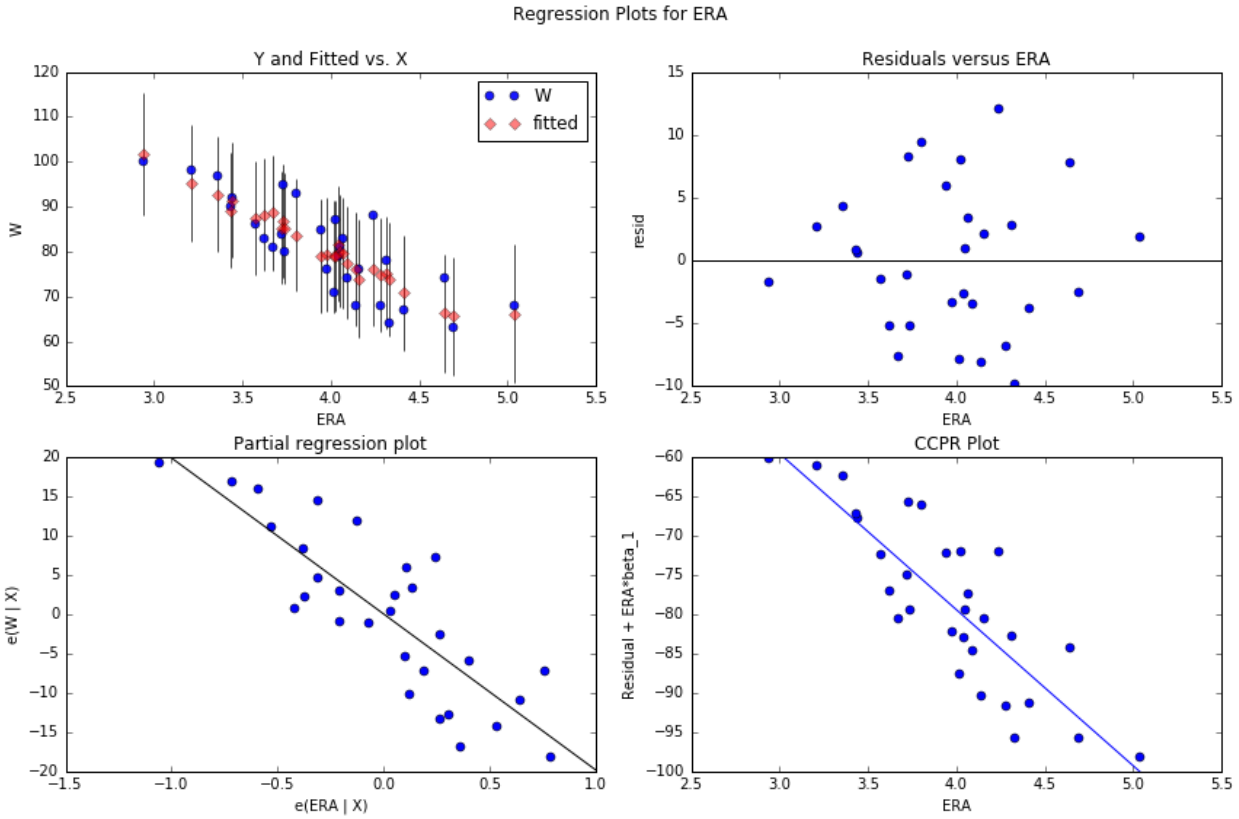
No. Observations:	*30*	*AIC:*	*195.5*
Df Residuals:	*28*	*BIC:*	*198.3*
Df Model:	*1*		
Covariance Type:	*nonrobust*		

	coef	*std err*	*t*	*P>|t|*	*[95.0% Conf. Int.]*	
const	*155.6139*	*9.924*	*15.681*	*0.000*	*135.286*	*175.942*
ERA	*-18.8678*	*2.493*	*-7.570*	*0.000*	*-23.974*	*-13.762*

Omnibus:	*1.504*	*Durbin-Watson:*	*1.863*
Prob(Omnibus):	*0.472*	*Jarque-Bera (JB):*	*1.115*
Skew:	*0.215*	*Prob(JB):*	*0.573*
Kurtosis:	*2.159*	*Cond. No.*	*37.7*

3.2.3. Multiple Explorations

- Dependent variable: Team W in 2015
- Independent variable: Team ERA in 2015; Team BPF in 2015
- Explore the relationship of dependent variable (W) and independent variable (ERA) conditional on the other independent variable (BPF)



- OLS Regression Results

Dep. Variable:	W	R-squared:	0.698
Model:	OLS	Adj. R-squared:	0.676
Method:	Least Squares	F-statistic:	31.24
Date:	Thu, 07 Jul 2016	Prob (F-statistic):	9.44e-08
Time:	14:46:39	Log-Likelihood:	-94.494
No. Observations:	30	AIC:	195.0
Df Residuals:	27	BIC:	199.2
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]	
Intercept	123.6929	22.874	5.408	0.000	76.760	170.626
ERA	-19.8587	2.517	-7.889	0.000	-25.023	-14.694
BPF	0.3572	0.232	1.541	0.135	-0.119	0.833

Omnibus:	1.189	Durbin-Watson:	1.735
Prob(Omnibus):	0.552	Jarque-Bera (JB):	1.029
Skew:	0.247	Prob(JB):	0.598
Kurtosis:	2.239	Cond. No.	2.12e+03

3.3. Conclusions

3.3.1. A team's winning record has a strong correlation with its pitching performance (r = *-0.819600*). However, it is not correlated with its batting performance *($|r| < 0.1$)*. The scatter plot of W and ERA shows a likely linear relationship ($R^2 = 0.672$) and the P value for the F-test of overall significance is less than the significance level, thus I could conclude that this linear regression model provides a better fit than the intercept-only model.

3.3.2. However, the R^2 cannot determine whether the coefficient estimates and predictions are biased, and ERA is not the only independent variable which might determine the winning outcomes. Thus the relationship of dependent variable (W) and independent variable (ERA) conditional on the other independent variable (BPF) is explored as well.

The 'Y and Fitted vs. X' plot shows a very good fitting for the regression model between W and ERA. The 'Residuals versus ERA' indicates that the coefficient estimates and predictions are valid without bias in this model. Furthermore, the partial regression model shows a strong linear relationship of W and ERA conditional on BPF. Finally, the CCPR plot displays an even better linear effect of ERA on W by taking into account of the other independent variable BPF. The final regression model summary with F-test confirms the significance and credibility of this model.

3.3.3. Based on all of those analysis, I conclude that a team's ERA has a strong effect on its success. Linear model (W = -18.8678*ERA + 155.6139) fits the data in 2015 well and it has the potential to be used for unbiased prediction. It indicates that a baseball team which has good defense strategies and good pitchers is more likely to win.

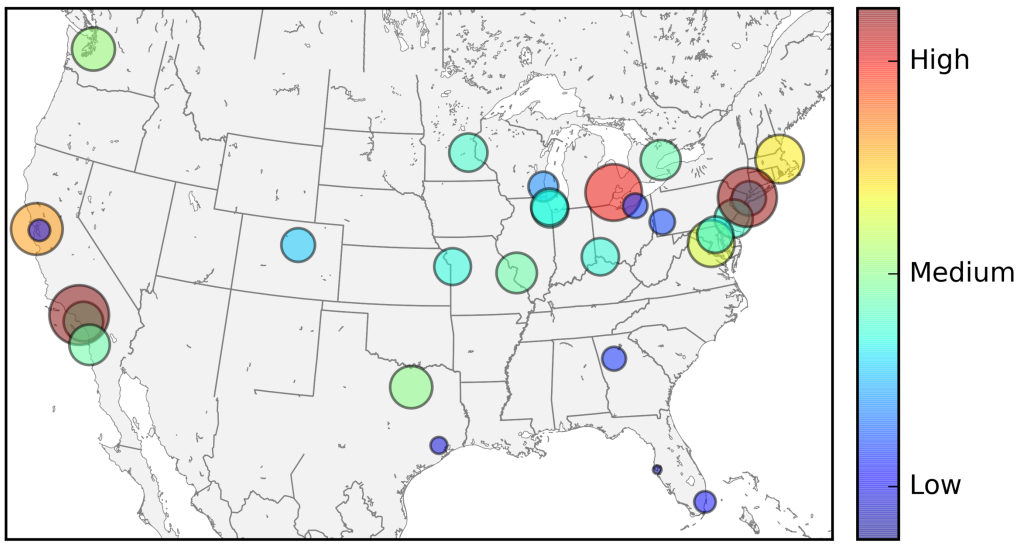
3.3.4. A team's average salary is not strongly correlated with the its success.

Python code for Q3:
https://github.com/super-penguin/Udacity_Data_Analyst/blob/master/P2/Q3.py

4. The distribution of MLB team location and average salary level.

4.1. Map the team's average salary in 2015 on the USA maps

- The team's locations are marked by the location of the MBL Stadiums.
- The size and color of each circle represents the salary level, the color is matched with the color bar on the right.



This plot shows clearly that the baseball teams on the east and west coast have higher salary compared with teams in the middle or south part of the USA

Python code for Q4
https://github.com/super-penguin/Udacity_Data_Analyst/blob/master/P2/map_team_salary.py