

Lab 16: Moneyball


Some of you may have read the book (or seen the movie) “Moneyball”, which told the story of how two savvy baseball managers used math to help pick a team that was wildly successful – even though the players they picked were passed over by other teams.

For years, baseball managers picked their teams based on physical appearance or favorite benchmarks (for example, tall players were selected more often than short ones, and faster ones more often than slower ones).



Now, in But the A’s managers had a problem – they didn’t have a whole lot of money to buy the fastest, tallest players (and, in fact they had just lost three of their best players to free agency). So, they took a different tack: using historical data to help them pick players that otherwise might “look” undesirable.

And, after they did, they started to notice some things. Take a look at that scatterplot at right¹:

Each of those little team logos is a point (one for each team in the league), and each of the points has two coordinates. One of the coordinates has a unit of “dollars” (as it’s representing the amount of money the team spends to keep the team running over a 15 years, from 2002 to 2016), and the other is the number of wins the team has accrued over time. So, for example, the Pittsburgh Pirates (logo is ) logged just over 1,100 wins during the 15-year time period this graph shows and have spent about \$750 million to get those wins.

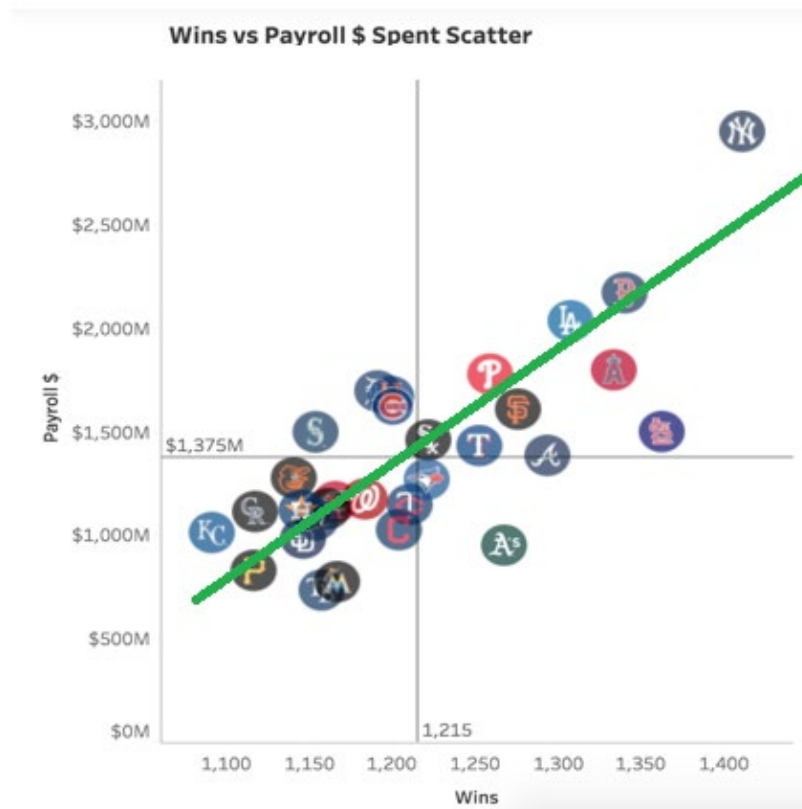
1. (2 points) What do you notice about the behavior of the points as the wins increase from the left side of the graph to the right side?
2. (2 points) What does this imply about what it takes, on *average*, to get more wins?



Do you see how that scatterplot, while not following a perfectly straight line, sure seems to be **suggesting** a straight line? What **that** tells us is that there sure seems to be a positive relationship between the number of wins a team can expect to get and the amount of money they’re willing to spend (which is basically what you just said in the last question. 😊).

In fact, like in the last lab, we can draw a trendline!

¹ Source: <https://public.tableau.com/app/profile/walter.allen/viz/15YearsofMoneyball/Story1>



See how that trendline slopes upward? That's called a **positive correlation**. Since the slope is positive, it means that, as one of the axis values goes up, so does the other axis value (in this case, as wins go up, so does the amount that the teams spent on the payroll).

And this makes sense! If you can buy the best players for your team, it stands to reason that they might do better than a "cheaper team", right?

Maybe. 😊

Take a look at the vertical line that has the "1215" next to it and the horizontal line with the "\$1,375 M" next to it. Those are (respectively) the average wins for **all** the teams over this time period, and the average costs for all the teams over the same time period. See how the A's logo is **right** of the 1215 and **below** the \$1,375 line?

3. (4 points) What's the implication of the A's position? Be sure to comment on both the "right" and "down" locations!

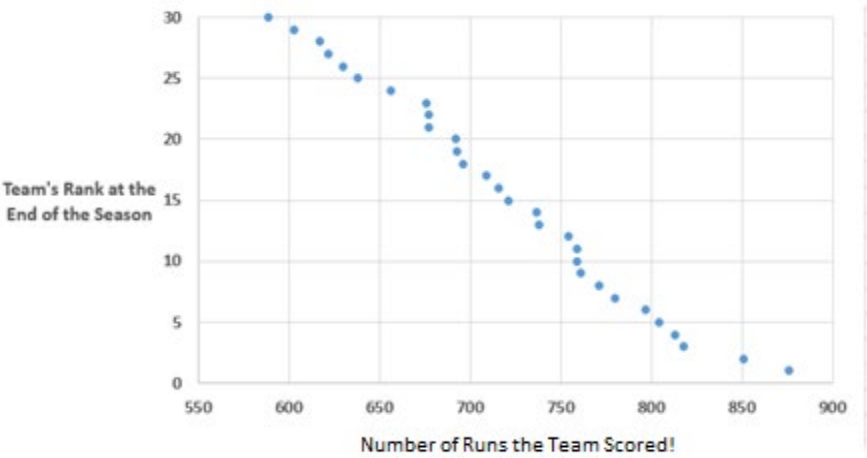
But **why**?

Well, what the A's managers (in particular, Billy Beane) did was look at what makes a "better" player differently. They had hunches that there might be data that other folks were missing that might give a clue to player performance. And one of the ways they did this was to look at how many games players won versus a lot of other, seemingly non-related data. They checked to see if players who had more at-bats had more home runs, or whether players who hit more triples tended to strike out more. Effectively, they **data mined** the available baseball data – pairing it up in various ways to see if one variable was correlated with another in a surprising – and cost-effective – way!

Now, we don't have the data from the actual A's season that was used, but we can use 2018's MLB data² to help us explore the main idea of data mining in good detail. Let's take a look at some of it, and do some data mining!

² Or any year you want, actually! IN fact, maybe I oughta update this. 😊

Now, the idea of “data mining” is to pair up two different variables to see if there’s a relationship between them. I’ve started by collecting the data at right – it shows the teams, ranked 1st to last (at the end of the season), as well as how many runs they scored. Now, there is a cool relationship that exists with this data, but it might not jump out at you from the table, so check out the scatterplot of it below.



Remember earlier how we had a scatterplot that described data that had a positive correlation? This data has a **negative correlation**. That is, as the team’s rank number gets larger (i.e. the rank gets worse), the number of runs goes **down**.

Let’s talk more on the next page!

(sorry about all this blank space – these datasets are pretty huge³. ☹)

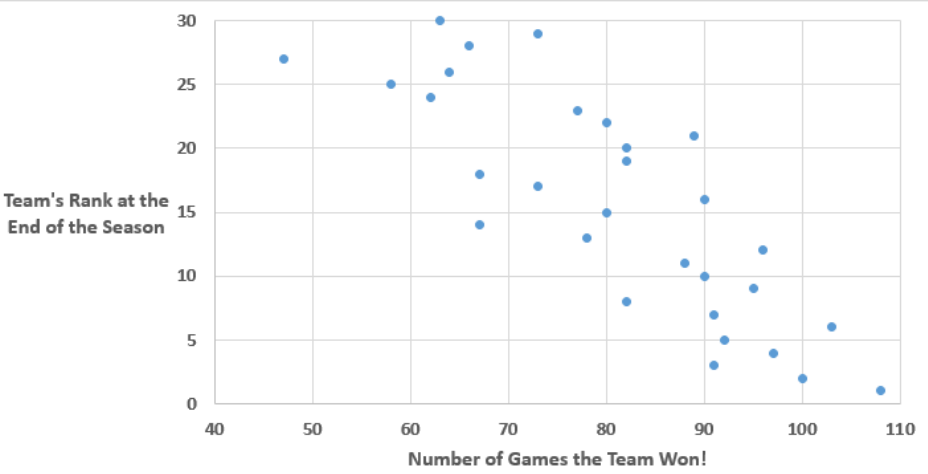
Team	Rank	Runs
Boston	1	876
NY Yankees	2	851
Cleveland	3	818
Oakland	4	813
LA Dodgers	5	804
Houston	6	797
Colorado	7	780
Washington	8	771
Chicago Cubs	9	761
Atlanta	10	759
St. Louis	11	759
Milwaukee	12	754
Minnesota	13	738
Texas	14	737
LA Angels	15	721
Tampa Bay	16	716
Toronto	17	709
Cincinnati	18	696
Arizona	19	693
Pittsburgh	20	692
Seattle	21	677
Philadelphia	22	677
NY Mets	23	676
Chicago Sox	24	656
Kansas City	25	638
Detroit	26	630
Baltimore	27	622
San Diego	28	617
San Francisco	29	603
Miami	30	589

³ Well, that’s relative, I guess. When you data mine data outside of math classrooms, the sets are often so large it’s prohibitive to actually scroll through them. Data miners would pat our little sets on the head and say things like, “Oh, how *sweet*!”

Now, upon seeing that last relationship, you might have thought, “Well, of **course** runs would be negatively correlated with the team’s rank! The more runs they get, the better they do...and the closer to rank 1 they’d be!”⁴

You might **also** say “Why did you do **runs** instead of **wins**?! I mean, clearly, the team who wins the **most** will have the lowest **rank** number, right?”

Well, there’s the data at right. Let’s take a look!



4. (1 point) Are these data points (rank versus games won) **more** or **less** tightly clustered around a suggested line than previous data (rank versus runs scored)?

So it appears that the team’s rank is, indeed, correlated with games won (and indeed, the 1st-place team did win the most games) – but not as **closely** correlated as it is with runs scored. What this would tell a manager is that it isn’t always the teams that win the most total games that do the best – it’s the ones that score the most **runs**!

Congratulations, you just mined your first data!

Team	Rank	Wins
Boston	1	108
NY Yankees	2	100
Cleveland	3	91
Oakland	4	97
LA Dodgers	5	92
Houston	6	103
Colorado	7	91
Washington	8	82
Chicago Cubs	9	95
Atlanta	10	90
St. Louis	11	88
Milwaukee	12	96
Minnesota	13	78
Texas	14	67
LA Angels	15	80
Tampa Bay	16	90
Toronto	17	73
Cincinnati	18	67
Arizona	19	82
Pittsburgh	20	82
Seattle	21	89
Philadelphia	22	80
NY Mets	23	77
Chicago Sox	24	62
Kansas City	25	58
Detroit	26	64
Baltimore	27	47
San Diego	28	66
San Francisco	29	73
Miami	30	63

What you can see, as you begin to cross-reference one data column with another, is if any particular independent variable data (on the horizontal axis) results in some kind of predictable behavior on the dependent variable. For example, we just saw that runs scored is very highly correlated with a team’s rank, while games won is not-so-highly-correlated. That just means that “runs scored” is a better **predictor variable** for rank.

So now, data miners will ask themselves, “OK – is any **other** data in the chart is positively correlated with **runs scored**? Because if we can find a relationship between some other variable and runs, we’ve (maybe) found a predictor variable that can get us more runs, and therefore a better team rank.”

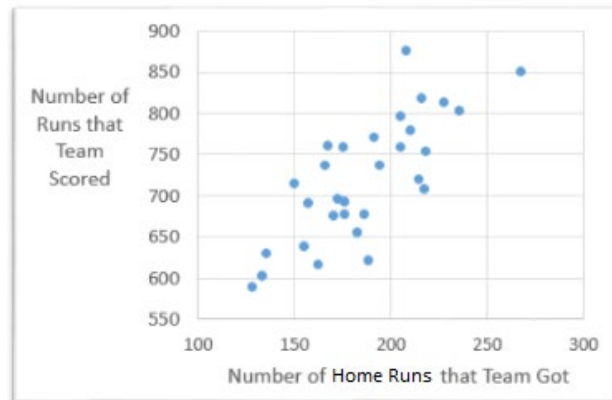
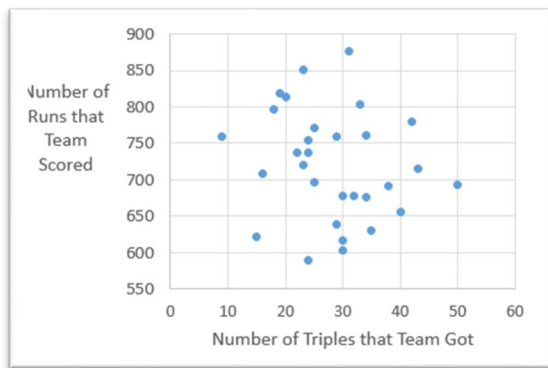
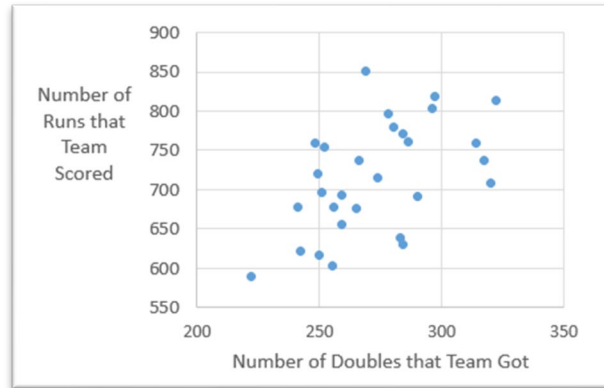
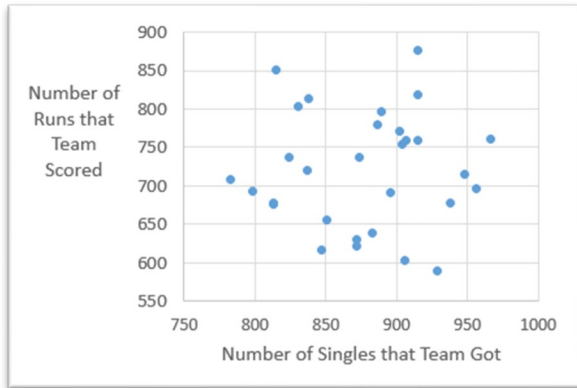
So let’s try!

⁴ Just in case you needed help with that last question. ☺

I noticed that four of the categories for “hits” in this sheet are

- “singles” (that is, a hit where the batter gets to first base only)
- “doubles” (they get to second base)
- “triples” (they get to third base) and
- “home runs” (they get all the way around and score a run).

Here are the four scatterplots that correlate each of those types of hits with total runs scored:



(Remember that each of those points in each of those plots represents one of our 30 teams, just without the logos this time.)

Please notice that I’ve moved the “runs scored” to the vertical axis – it was an independent variable in the last couple of exercises, but now, since I know it’s a predictor for rank, I want to see if either singles, doubles, triples, or home runs predict *it*.

See how different they all are? And much, much *noisier* than the last scatterplots. By “noisier”, I mean that the data doesn’t appear to have a perfect shape headed from left to right – but in a couple of them, the points tend to cluster a little more toward a line.

5. **(1 point)** In which of the data sets do the data cluster the *most* towards a line?
6. **(1 point)** In which of the data sets do the data cluster the *least* towards a line? I would vote that there are probably two answers that look the least “clustered”.
7. **(1 point)** So, which of the 4 types of hits (single, double, triples, or home runs) is *most* positively correlated with runs scored?

And that makes perfect sense! By the definition of one of those, you get at *least* one run each time one is hit (and, sometimes, more)!

8. **(1 point)** Which of the 4 types of hits is 2nd most positively correlated with runs scored?
9. **(2 points)** Did this surprise you at all? I know it surprised **me**! If it did, tell me why – and if not, tell me why **not**!

Now, what you'll do is to data mine through a variety of data sets to try to find the ones that are most highly correlated with "runs scored". We'll use [a Google sheet created for this very task](#); go ahead and open it now! Both the dependent and independent variable have pull-down menus that allow you to pick any two data sets to compare—you just need to click the arrow on the right-hand edge of each label to show the menu.

10. **(3 points)** Find a data set that appears to be very highly positively correlated with "Runs Scored" (besides the ones you've already seen or used). **Also** give a reason why that makes sense (you might have to Google what certain baseball phrases mean; I know *I* had to. 😊).
11. **(1 point)** Now, take your answer from #11, and make it be your dependent variable, and find another predictor variable highly positively correlated with **it**! (besides "Runs Scored" and "Wins").
12. **(3 points)** Find a data set that appears to be slightly negatively correlated with "Runs Scored", **and** give a reason why **that** makes sense.
13. **(3 points)** Find a data set that **doesn't** appear to be correlated at all with "Runs Scored". Give a reason why **that** makes sense, too!