1. Read the McCracken article:

2. (Pitching Statistics) In this question we will become familiar with the data by analyzing some standard pitching statistics.

   (a) Estimate each of the following probabilities. Each is a single value computed using all of the seasons in the dataset.

      i. Probability of a plate appearance ending in a walk.

      $$P(\texttt{walk}) = 0.0749$$

      ii. Probability of a plate appearance ending with a strikeout.

      $$P(\texttt{strike out}) = 0.2064$$

      iii. Conditional probability of a plate appearance ending with a strikeout given a walk did not occur.

      $$P(\texttt{strikeout | no walk}) = 0.2231$$

      iv. Average number of home runs per plate appearance.

      $$P(\texttt{home runs}) = 0.0276$$

      v. Conditional probability of a plate appearance ending with a home run given neither a walk nor a strikeout occurred.

      $$P(\texttt{home run | no walk or strikeout}) = 0.0384$$

      vi. Conditional probability of a plate appearance ending with a non-HR hit given that the plate appearance didn't end with a walk, strikeout, or homerun.

      $$P(\texttt{hit | no walk, strikeout, or homerun}) = 0.2915$$

(b) Compute for each season and pitcher, the average number of strike-outs per plate appearance that did not end in a walk (the second McCracken component). Dis- play the top 10 players in seasons 2016 and 2017 according to this statistic (two tables, one for each season). Include pitcher, krate (your statistic), and the num- ber of batters faced. Restrict your analysis to pitchers with at least 500 batters faced, and sort your table in decreasing order by krate.

	index	y	pitcher	bf	krate
0	1190	2016	Fernandez_Jose_605228	731	0.370968
1	3524	2016	Scherzer_Max_453286	900	0.335697
2	3721	2016	Strasburg_Stephen_544931	597	0.330325
3	2021	2016	Kershaw_Clayton_477132	543	0.322702
4	3775	2016	Syndergaard_Noah_592789	742	0.310984
5	3183	2016	Ray_Robbie_592662	772	0.309220
6	3445	2016	Salazar_Danny_517593	581	0.309021
7	3942	2016	Velasquez_Vincent_592826	550	0.300395
8	3964	2016	Verlander_Justin_434378	902	0.300236
9	120	2016	Archer_Chris_502042	850	0.297573

	index	y	pitcher	bf	krate
0	3452	2017	Sale_Chris_519242	851	0.381188
1	3525	2017	Scherzer_Max_453286	778	0.369655
2	3184	2017	Ray_Robbie_592662	662	0.367003
3	2058	2017	Kluber_Corey_446372	775	0.357625
4	1754	2017	Hill_Rich_448179	551	0.330020
5	2948	2017	Peacock_Brad_502748	546	0.329243
6	121	2017	Archer_Chris_502042	852	0.314394
7	3556	2017	Severino_Luis_622663	783	0.314208
8	3722	2017	Strasburg_Stephen_544931	696	0.311927
9	2022	2017	Kershaw_Clayton_477132	679	0.311248

(c) Repeat the previous problem with hrate (in place of krate), the fourth Mc- Cracken component. Recall that hrate is defined as the average number of non- HR hits that did not end in a walk, strikeout, or home run.

	index	y	pitcher	bf	hrate
0	3183	2016	Ray_Robbie_592662	772	0.347732
1	2939	2016	Paxton_James_572020	508	0.346260
2	2958	2016	Pelfrey_Mike_460059	541	0.341981
3	759	2016	Cole_Gerrit_543037	503	0.339726
4	2980	2016	Perdomo_Luis_606131	655	0.336066
5	2511	2016	McHugh_Collin_543521	795	0.335185
6	412	2016	Bradley_Archie_605151	630	0.334951
7	1047	2016	Duffey_Tyler_608648	593	0.334118
8	4009	2016	Wacha_Michael_608379	600	0.333333
9	3775	2016	Syndergaard_Noah_592789	742	0.332627

	index	y	pitcher	bf	hrate
0	2619	2017	Montero_Rafael_606160	545	0.361345
1	3778	2017	Taillon_Jameson_592791	584	0.348148
2	3223	2017	Richard_Clayton_453385	852	0.346154
3	248	2017	Bauer_Trevor_545333	749	0.333333
4	2722	2017	Nelson_Jimmy_519076	727	0.333333
5	1361	2017	Gausman_Kevin_592332	816	0.331471
6	2564	2017	Miley_Wade_489119	727	0.329060
7	3055	2017	Pivetta_Nick_601713	584	0.328729
8	788	2017	Colon_Bartolo_112526	648	0.328629
9	396	2017	Boyd_Matt_571510	602	0.327830

3. (Predicting McCracken Components) In this problem we forecast the McCracken com- ponents for a given season using data from the preceding season. Restrict to rows with at least 200 batters faced.

   (a) Fit the following linear regression model:

   $$\texttt{bbrate} \sim \texttt{bbrate prev}.$$

   This is a single regression fit on the entire dataset. The response bbrate (average walks per plate appearance, the first McCracken component) varies over pitchers and seasons (except 2012), and the covariate bbrate prev is average walks per plate appearance for the same pitcher over the preceding season. A row should be excluded if the pitcher had strictly fewer than 200 batters faced in either the current season, or the preceding season.

   i. Report the two coefficients from your model.

   ```
   Intercept:  0.0335
   bbrate_prev:  0.5474
   ```

   ii. By using additional features from the preceding season, try to improve your fit of `bbrate`.

   A. List some of the features you tried, whether they had significant coefficients, and what the value of the corresponding coefficient was.

   I decided to create `prev` versions of the four other types of pitches (`lo, po, fo, go`). Since we want to see what combination of `prev` features factor into `bbrate`, I decided to incorporate forward selection.

   The best features and corresponding coefficients can be found below:

Features	Coefficient
bbrate_prev	0.5272
krate_prev	0.0357
forate_prev	-0.0277
hrrate_prev	0.0879
porate_prev	-0.0262

   The coefficients that were statistically significant were `bbrate, krate, hrrate`.

   B. For each feature listed above, give a brief explanation of your findings.

In the end we chose bbrate, krate, and hrrate which makes sense because strikeouts and homeruns represent dead ball situations just like walks.

Your features can be other McCracken components, as well as other features you create using the columns in the given data.

(b) Repeat the previous part using krate, the second McCracken component (recall krate measures the average number of strikeouts per plate appearance that didn't end in a walk).

i. Report the two coefficients from your model.

```
Intercept:  0.0583
bbrate_prev:  0.7474
```

ii. By using additional features from the preceding season, try to improve your fit of `krate`.

A. List some of the features you tried, whether they had significant coefficients, and what the value of the corresponding coefficient was. The best features and corresponding coefficients can be found below:

Features	Coefficient
krate_prev	0.7344
lorate_prev	0.1037
bbrate_prev	0.0781
porate_prev	0.1234
forate_prev	-0.0484
hrate_prev	0.0528

The only coefficient that was statistically significant was `porate`.

B. For each feature listed above, give a brief explanation of your findings.

It's interesting that porate was the only other related feature because it refers to popouts which in general don't have anything to do with strikeouts because they are a different way of getting a player out. But if porate were to be correlated, then it is strange that forate (flyout rate) was not because both are different ways

(c) Repeat the previous part using `hrrate`, the third McCracken component (recall hrrate measures the average number of home runs per plate appearance that didn't end in a walk or strikeout).

i. Report the two coefficients from your model.

```
Intercept:  0.0273
bbrate_prev:  0.2954
```

ii. By using additional features from the preceding season, try to improve your fit of bbrate.

A. List some of the features you tried, whether they had significant coefficients, and what the value of the corresponding coefficient was. The best features and corresponding coefficients can be found below:

Features	Coefficient
gorate_prev	-0.0464
hrrate_prev	0.1850
lorate_prev	0.1003
krate_prev	-0.0155
porate_prev	0.0272

The coefficients that were statistically significant include `gorate, hrrate, lorate, krate`.

B. For each feature listed above, give a brief explanation of your findings.

With home run rate, it's not surprising that krate would rank as a statistically significant feature because we have shown that they are related features. gorate (ground out rate) and lorate (line out rate) are particularly surprising because both represent live ball situations in which a player can be taken out via fielders touching one of the bases. But if we look at the coefficients given in the chart above, the highest coefficient is for hrrate at .185. This signifies that homerun rate from one year is not particularly telling for the subsequent year and thus the correlations found may not be very strong ones.

(d) Repeat the previous part using hrate, the fourth McCracken component (recall hrate measures the average number of non-HR hits per plate appearance that didn't end in a walk, strikeout, or home run).

i. Report the two coefficients from your model.

```
Intercept:  0.2404
bbrate_prev:  0.1720
```

ii. By using additional features from the preceding season, try to improve your fit of bbrate.

A. List some of the features you tried, whether they had significant coefficients, and what the value of the corresponding coefficient was. The best features and corresponding coefficients can be found below:

Features	Coefficient
porate_prev	-0.2068
krate_prev	-0.0675
hrate_prev	0.0901
gorate_prev	-0.0361
bbrate_prev	0.0534

The coefficients that were statistically significant include `porate, krate, hrate`

B. For each feature listed above, give a brief explanation of your findings.

As with hrrate, we've shown that hrate is also related to krate. Porate is the only one of the four non-strikeout methods of getting a batter out that had statistical significance with hrate. Since hrate typically depicts live ball situations I would expect that lorate, forate, and gorate would correlate better than they did. However what is really interesting about the coefficients is that porate and krate were better indicators of current hrate than previous hrate was, which shows that the best indicator for hrate might change every year.