FIT5147 Data Exploration Project

Chris Sale Pitch Analysis

Nicholas Pennell, 30338913

Introduction

I'm a sports fan. I like them all, football, cricket, curling, you name it I'll probably watch it. Baseball is a great sport for anyone that likes to take a deep dive into the numbers behind the sport, which I most definitely do. This is due to the shear amount of information and metrics used in the sport, you just need to know where to look.

For this data exploration project, baseball data was taken from the start of the 2010 Major League Baseball (MLB) season all the way up to the end of the regular season in 2018. This data well be used to examine one pitcher, Chris Sale. The goal being to explore,

1. what type of pitches does Sale throw and how has this changed over time,
2. and, the location of these pitches.

Data Wrangling

The data wrangling process is split into several steps

- **Get**: Sourcing the data and reading it into in R,
- **Understand**: Add any new variable, examine and alternating variables, and make sure its clear what each variable is.
- **Tidy**: Make sure the data is the right format to be explored. This is completed with several packages, `dplyr`, `tidyr`, `lubridate` and `RSQLite`.

Get

The data comes from the `pitchRx` R package. The package has a function that scrapes MLB Statcast. MLB Statcast allows people to gain access to accurate and high-level stats which are tracked by in stadium cameras in the PITCHf/x system.

The scarping function requires two dates, start and end date, and then takes all Statcast data between the dates and compiles it all into 5 tables, at bat, pitches, action, po, runner. For my use I split the data into years and save each year as an SQLite database, using the `RSQLite` package, and then extracts it from there each time it is used.

For each year the two wanted tables, pitch and at bat, with the wanted column variables are extracted and saved as data frames. The two tables are then joined based on two identifying variables. The data is then filtered to just take pitches thrown by Sale. This is repeated for each year, and then joined to make one finale set of data.

Understand

After the get step we are left with 24234 observations and 29 variables.

Two variables are added to the data set, one is based on the date (team), the other is based on two other variables to simplified it down into one (outcome).

The variable classes are then checked to make sure each is the right type and if not changed. Variable levels names are also changed to be simpler and easier to understand. leaving the following variables:

A list of variables and there meaning can be found in the raw code

Tidy

Exploring all 29 variables would be to large of a task for this project so a large number where removed. The observations that where un-wanted; pre-season games, pitches that where never meant to be hit (Pitch Out, Intentional Walk), as well as pitch types that are thrown less than 100 times.

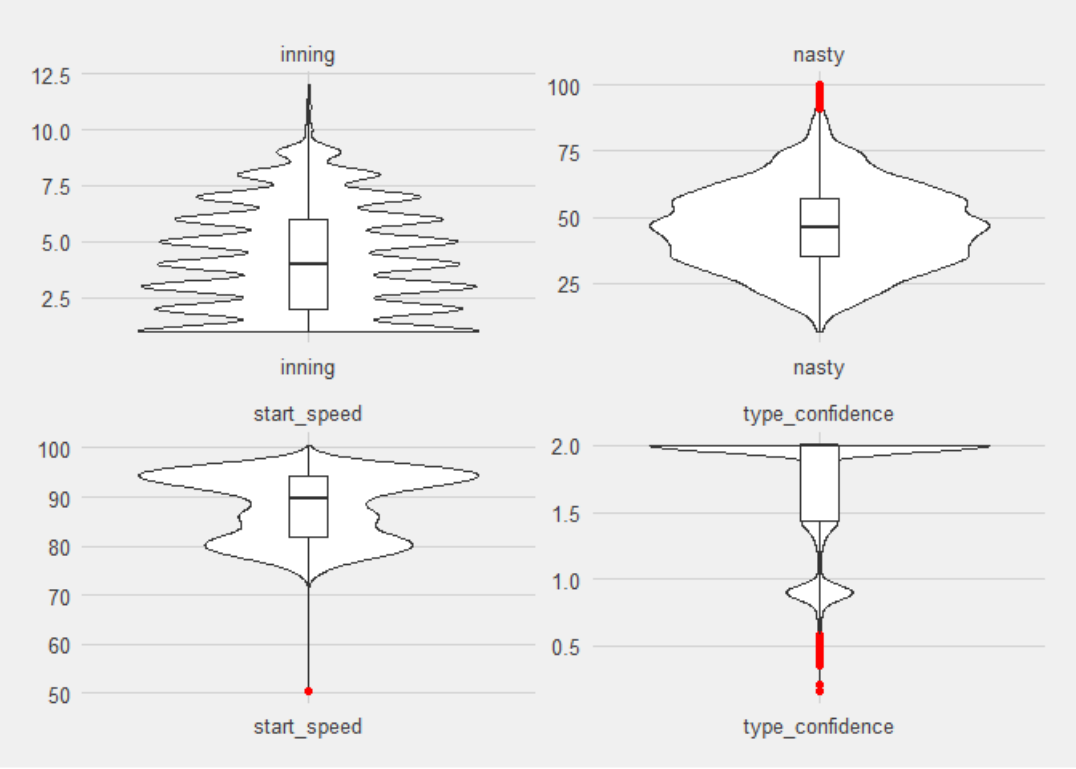
The finale check was to make sure it matched the 3 tidy data principles,

1. Each variable forms a column,
2. Each observation forms a row,
3. Each type of observational unit forms a table. The following is a sample of what is left.

Data Checking

The data checking process starts by examining the missing values that remain in the data. We find there are only missing values for nasty (62) and batters name (153), neither are a large amount of missing values, so the decision was made to remove those observations with missing values.

The distribution, of the numerical values as well as their outliers, are then plotted.



Plot 1: Outliers

Start at the top left with inning we can see the distribution matches what we would think, the values fall on integers, and decrease the larger the value becomes. This matches what we would expect as Sale is a starting pitcher meaning he starts the game and really see's the end of a game. The interesting note from the distribution is the decrease in the seconds innings.

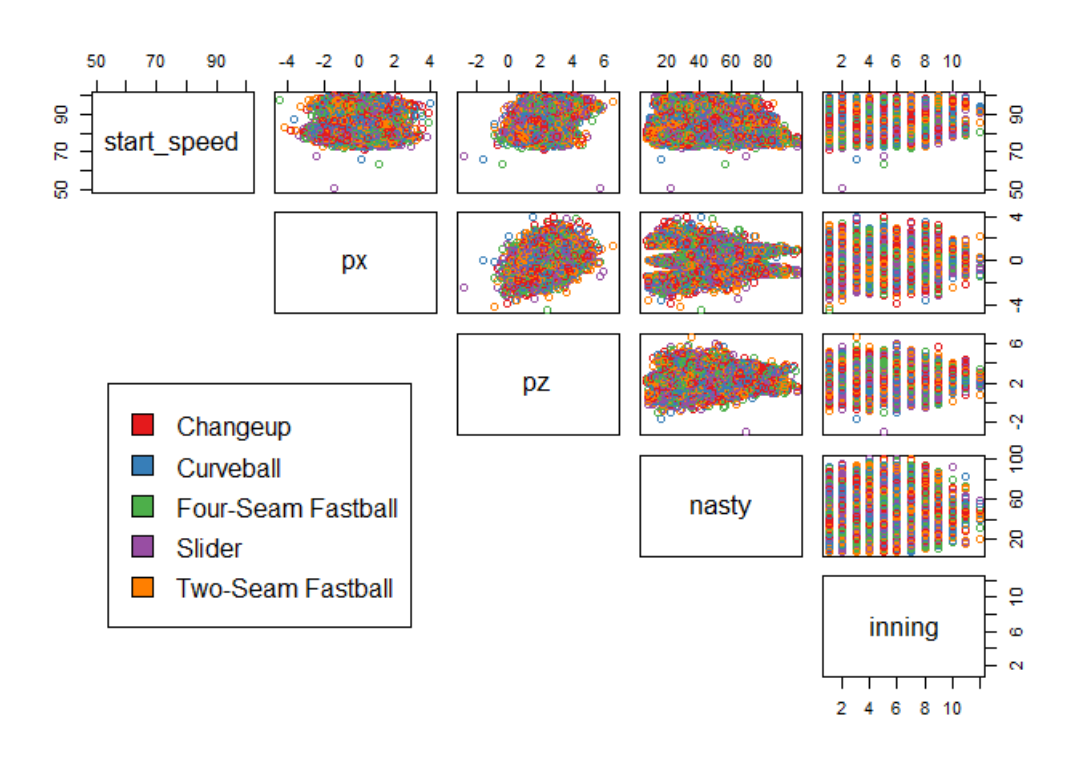
The nasty value has, some outliers, on the top end, none of these values are over one-hundred, as nasty is a

score from zero to one-hundred, none of these outliers need to be removed, as they are just Sale's best pitches for his career so far.

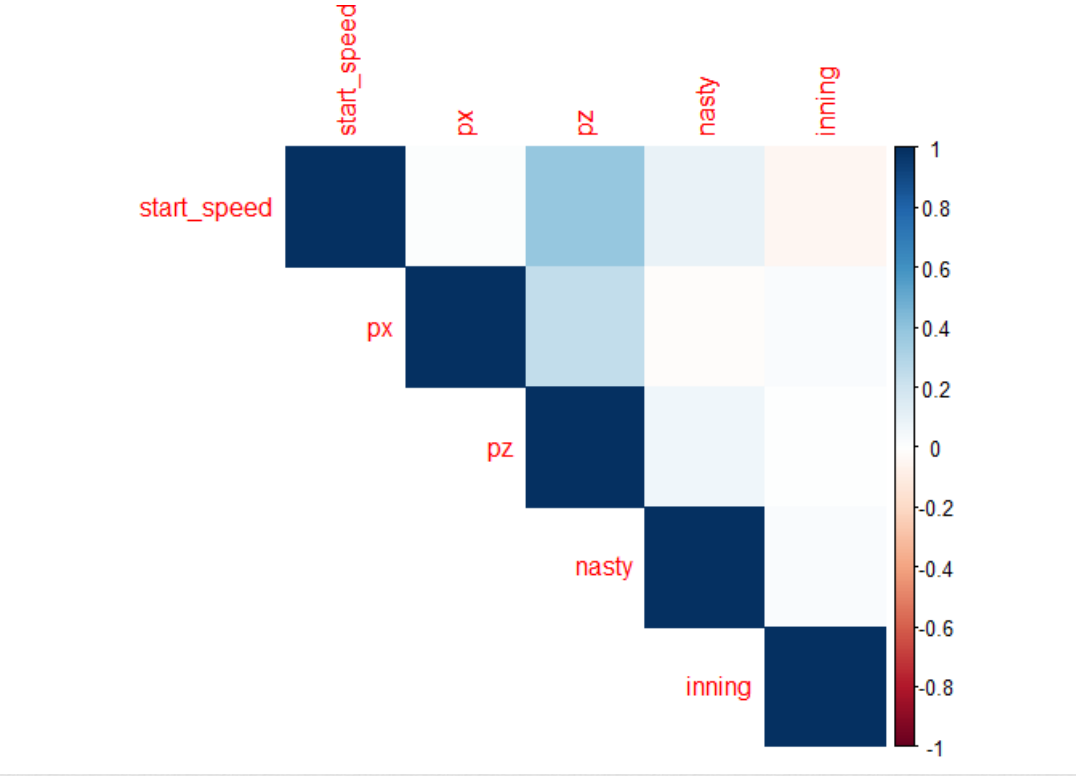
The start speed, variable also has an outlier, this one is will be removed, due to the fact it falls so far from the rest of the distribution, as Sale's slower pitches seem to fall around 80mph. Thus, the outliers most likely belongs to a pitch we don't want to have in the analysis or was a miss recording.

The type confidence has the most outliers. As type confidence is a measure of how accurate an observation is, all the outliers will be removed. The mean is also very close to the maximum value, which is very positive as we want as many values close to or at the maximum.

The next exploration step is a quick correlation check, we don't expect to find any correlation, but will take a quick look to be safe.



Plot 2: Corelation Plot 1



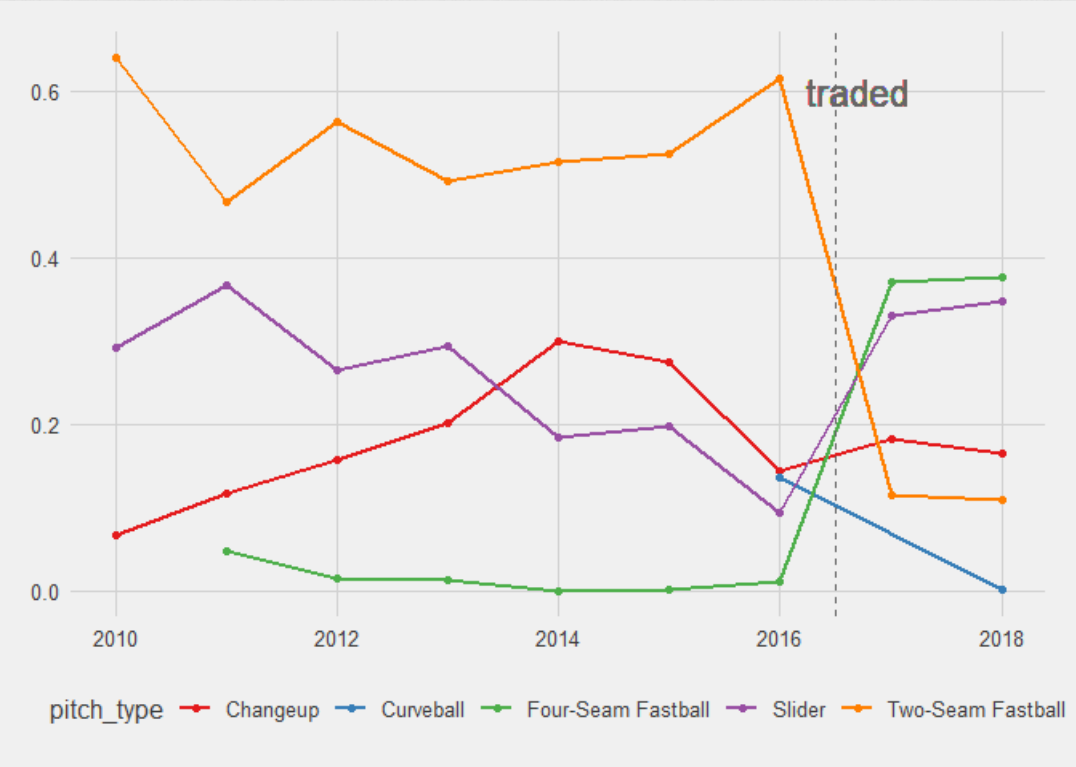
Plot 3: Corelation Plot 2

The two plots make it clear there is no strong correlation (positive or negative) between variables. So it's safe to move on.

Data Exploration

The first correlation plot gives the first look at the 5 types of pitches that Sale throws, a Changeup, Curveball, Four-Seam Fastball, Slider and the Two-Seam Fastball (this site explains a little about each one).

From here we want to find out how often he uses each and which one he uses the most. A good way to show this is to start with a simple time series.



Plot 4: Frequncey By Year

year	pitch_type	type_freq	year_freq	perc
2010	Changeup	25	372	0.0672043
2011	Changeup	127	1090	0.1165138
2012	Changeup	472	3009	0.1568627
2013	Changeup	666	3304	0.2015738
2014	Changeup	829	2764	0.2999276
2015	Changeup	907	3294	0.2753491
2016	Changeup	478	3308	0.1444982
2017	Changeup	656	3596	0.1824249
2018	Changeup	384	2322	0.1653747
2016	Curveball	448	3308	0.1354293
2018	Curveball	3	2322	0.0012920
2011	Four-Seam Fastball	53	1090	0.0486239
2012	Four-Seam Fastball	44	3009	0.0146228
2013	Four-Seam Fastball	45	3304	0.0136199
2014	Four-Seam Fastball	1	2764	0.0003618
2015	Four-Seam Fastball	5	3294	0.0015179
2016	Four-Seam Fastball	38	3308	0.0114873
2017	Four-Seam Fastball	1334	3596	0.3709677
2018	Four-Seam Fastball	875	2322	0.3768303
2010	Slider	109	372	0.2930108
2011	Slider	400	1090	0.3669725
2012	Slider	795	3009	0.2642074
2013	Slider	969	3304	0.2932809
2014	Slider	509	2764	0.1841534
2015	Slider	650	3294	0.1973285
2016	Slider	309	3308	0.0934099
2017	Slider	1190	3596	0.3309232
2018	Slider	807	2322	0.3475452

year	pitch_type	type_freq	year_freq	perc
2010	Two-Seam Fastball	238	372	0.6397849
2011	Two-Seam Fastball	510	1090	0.4678899
2012	Two-Seam Fastball	1698	3009	0.5643071
2013	Two-Seam Fastball	1624	3304	0.4915254
2014	Two-Seam Fastball	1425	2764	0.5155572
2015	Two-Seam Fastball	1732	3294	0.5258045
2016	Two-Seam Fastball	2035	3308	0.6151753
2017	Two-Seam Fastball	416	3596	0.1156841
2018	Two-Seam Fastball	253	2322	0.1089578

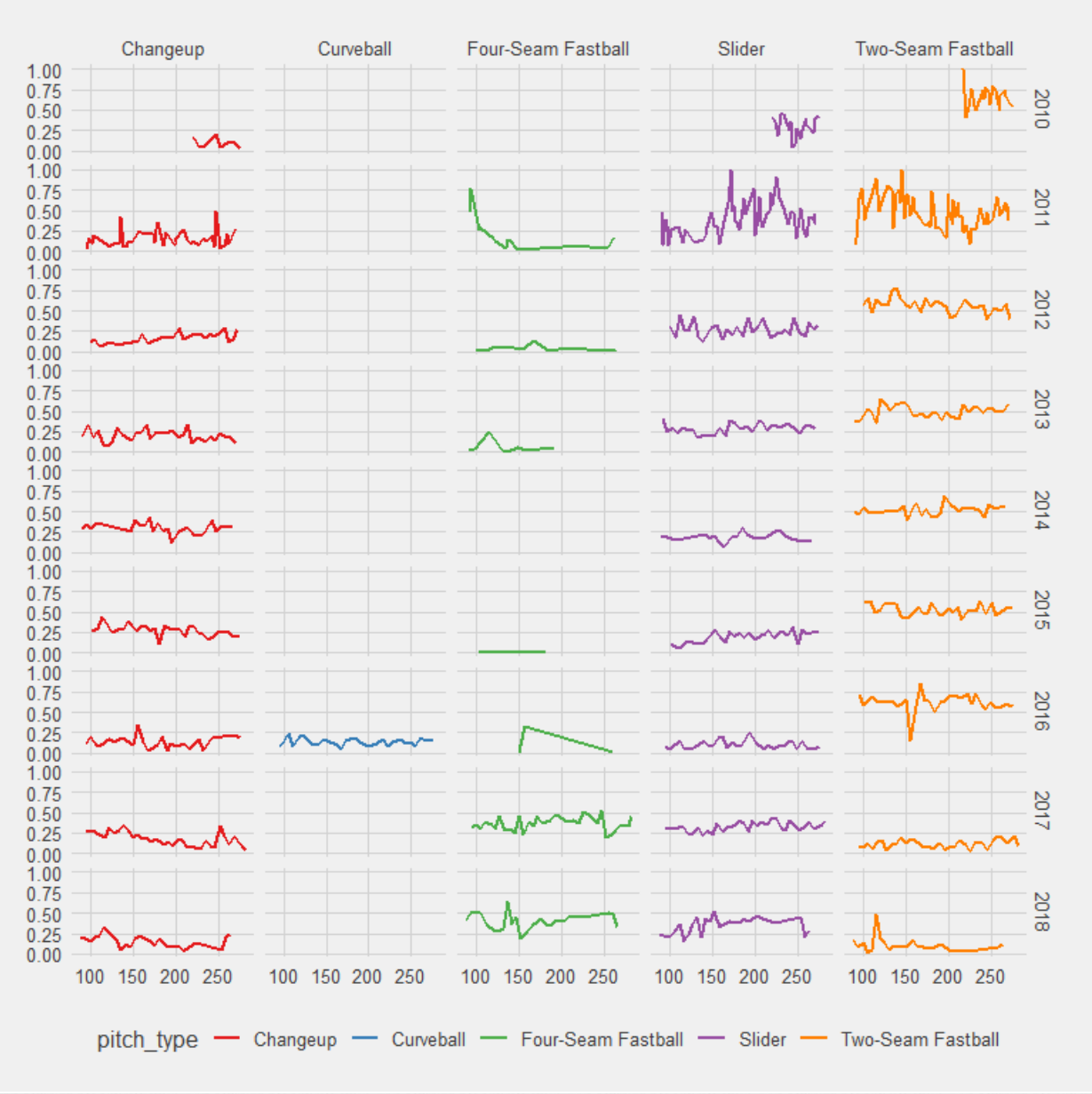
The y-axis shows the frequency a pitch was throw as a percentage of the total pitches thrown in that year, while the x-axis shows the year.

This plot shows what pitches where thrown in which year, note the Curveball was only used in 2016, while the Four-Seam Fastball wasn't used in 2010. The plot also labels the date Sale got traded, form the Chicago White Sox's to the Boston Red Sox's (1 of 236 players to play for both Sox teams. This is an event which clearly changes the tactic behind what pitch is thrown.

By looking at the years just before the trader we can see the Two-Seamer is the go-to pitch, while being predominantly backed up by the Slider and Changeup, with the Slider being preferred from 2010 to 2013, then the Changeup becomes the preferred secondary pitch. In 2016, all secondary pitches fall below 20%, this coincides with adding the Curveball to the arsenal, and a spike in the Two-Seamer, reaching above 60% for the first time since 2010 (this was due to a change in game play Sale and the White Sox implanted for the 2016 season).

Once the trade happens Sale returned to the old game plane with, according to the data, a new twist. In 2017 the Curveball which had just been introduced was no longer in use, the Splitter was back to being the secondary pitch, back over 30% first time scene 2011. The change up back down to below 20%. The most drastic change and new twist to the game plane, was the Two-Seamer was being used ~10% and the rarely use Four-Seamer was now the primary pitch at ~37%.

Another way to look at the changing frequency is to beak the time series into years and pitch-type.



Plot 5: Frequncey By Year Days

In this plot the y-axis is the same as before, the x-axis is the Year day, this allows the time series to be faceted easily into year and pitch type, to show the change of frequency thorough out a year.

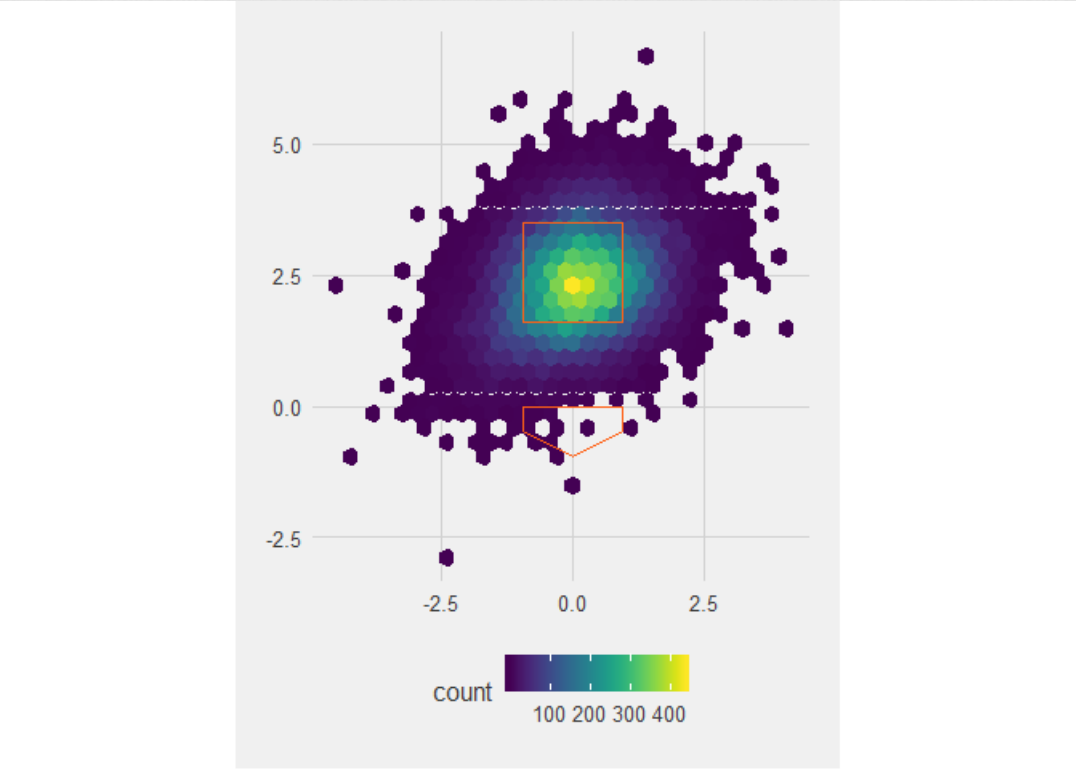
This plot helps show the frequency variance of a pitch type for each year. 2010 was the first year Sale played in the MLB, and he came in later in the year, the 6th of August. For the most part each year and pitch type keep a consistent frequency, All but 2011. This could be due to the reduced number of pitches thrown in 2011,

year	Total Pitches
2010	372
2011	1090
2012	3009
2013	3304
2014	2764
2015	3294

year	Total Pitches
2016	3308
2017	3596
2018	2322

Or could be a result of growing pains and trying to establish him self in his first full year in the MLB.

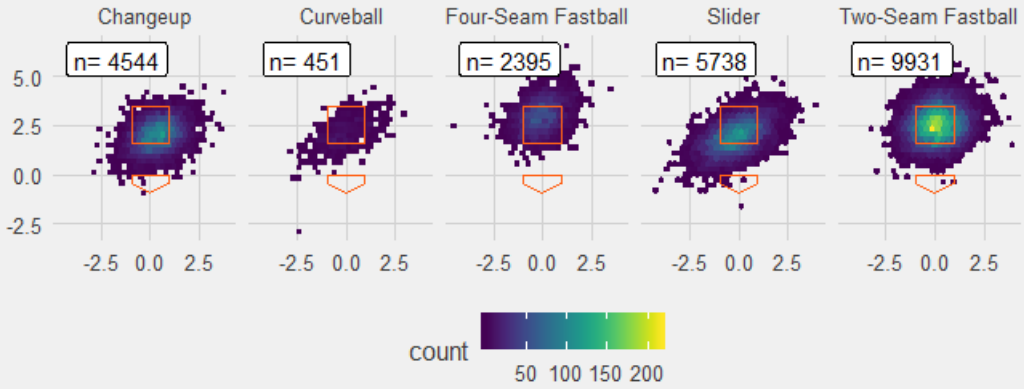
Another aspect of the data set is the location of the pitches,



Plot 6: Location of All Pitches

From this data we can see that most pitches end up in the strike zone.

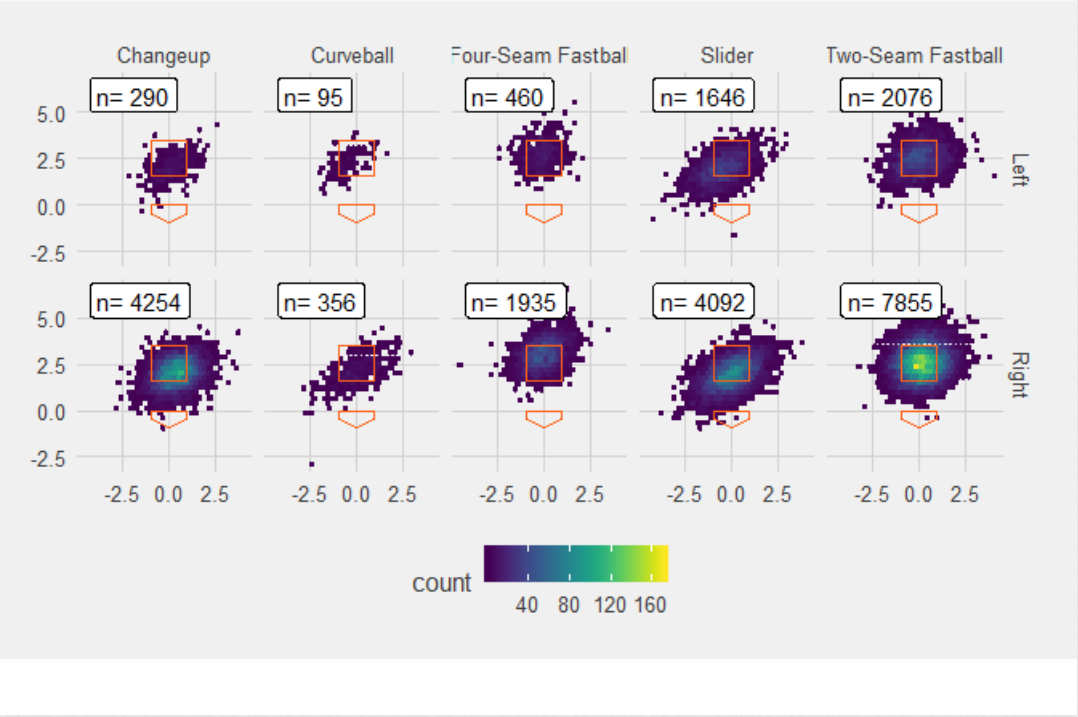
This data is use for to see where certain pitch types are thrown, and where, if not thrown correctly they miss the strike zone (the strike zone is show by the orange box).



Plot 7: Location by Pitch Type

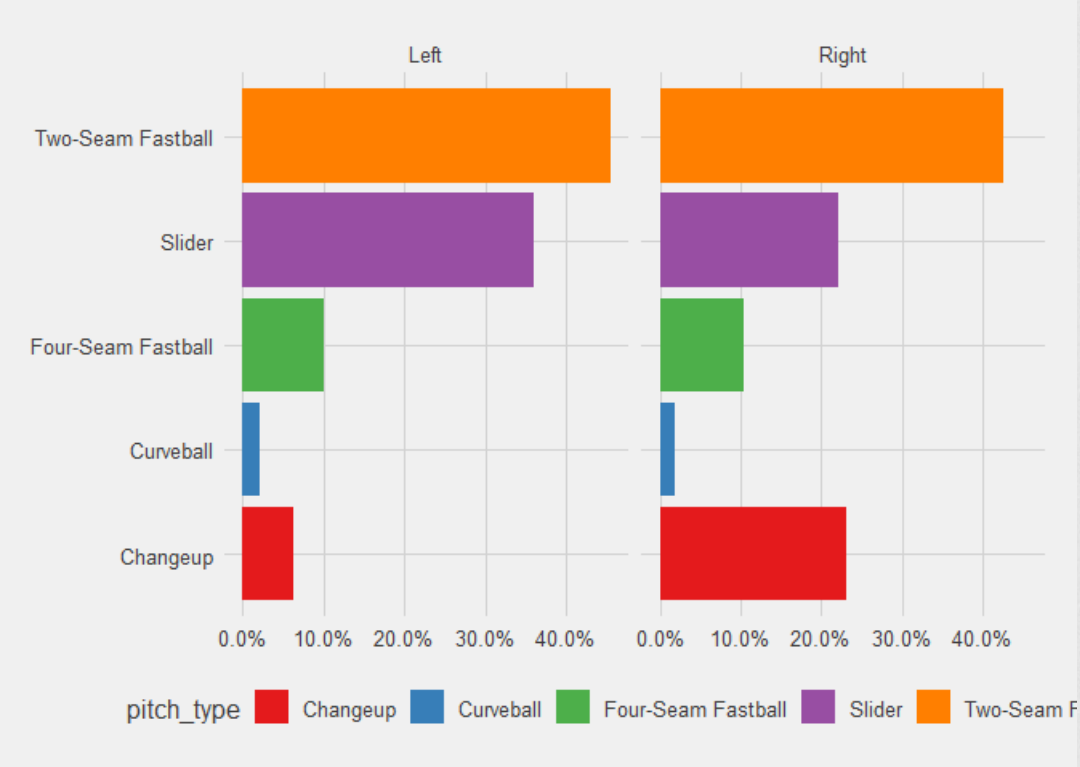
Starting with the Changeup, the target is, shown by the higher frequency in the bottom right corner of the strike zone, the misses are also in the same area with a lot of pitches missing to the right hand side, indicating this pitch is probably thrown to right hander more than left handers, as it would hit a left hander. The Curveball with its much smaller sample size, is much harder to so see the area that is being aimed for, its clear that the pitch wasn't meant for the top left corner. The Four-Seamer is a pitch that's fast and used to challenge the batter to see if he can hit it, same as the Two-Seamer. The Four-Seamer is thrown in the top half of the strike zone and thus the misses for the Four-seamer are above the strike zone. If this is compared to the pitch is has replaced the Two-seamer, the location of which is more central and has more misses all around the strike zone. Finally, the Slider a pitch Sale is trying to throw down and left in the strike zone, has a similar shape to the Curveball, with misses above the top right and below the bottom left corner, this could be the pitch get away from Sale, or him trying to use it to get swings and misses.

It was brought up earlier that the Changeup was probably thrown to right handers more, so to check this, plot 6 was split into left and right handers,



Plot 8: Location by Pitch Type and Stand

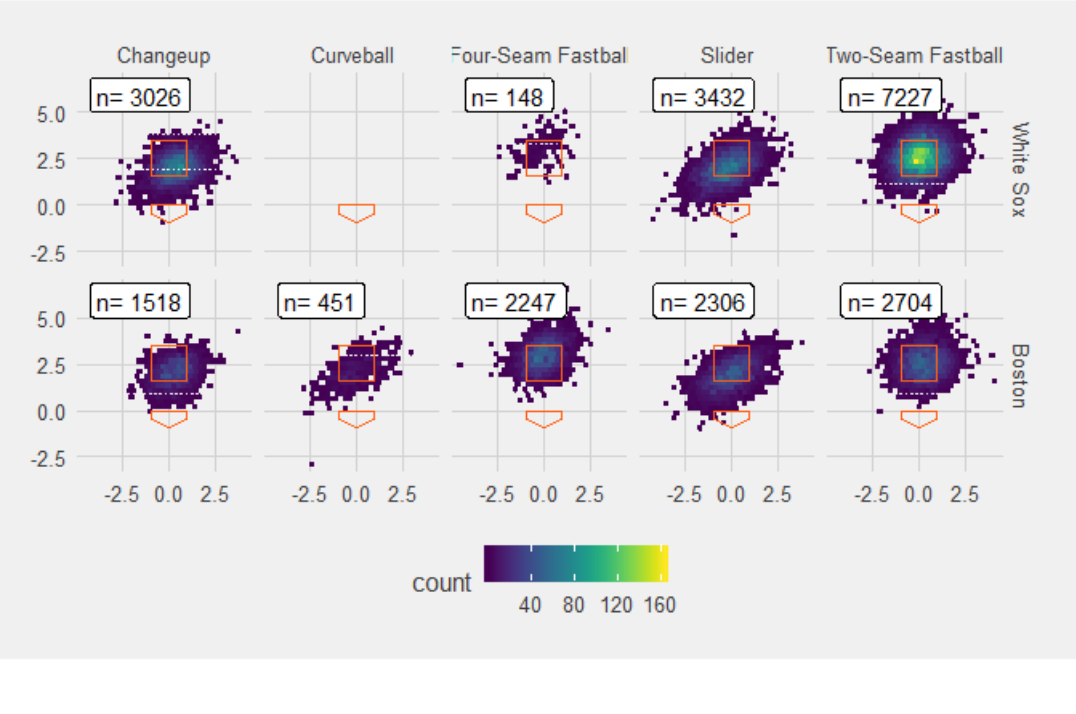
A frequency pitch type plot was also added to help add some incite.



Plot 9: Freqcency of Pitches by Stand

In analyzing these plots, we must be careful with sample sizes as its clear Sale throws to left hands much less than right handers, 4567 pitch's to left handers compared to 18492 pitch's to right handers. The distribution shape of the pitches is similar no matter what hand the batter is. But the frequency does clearly show what was hypothesized before, that the change up is use far less often to left handers compared to right handers.

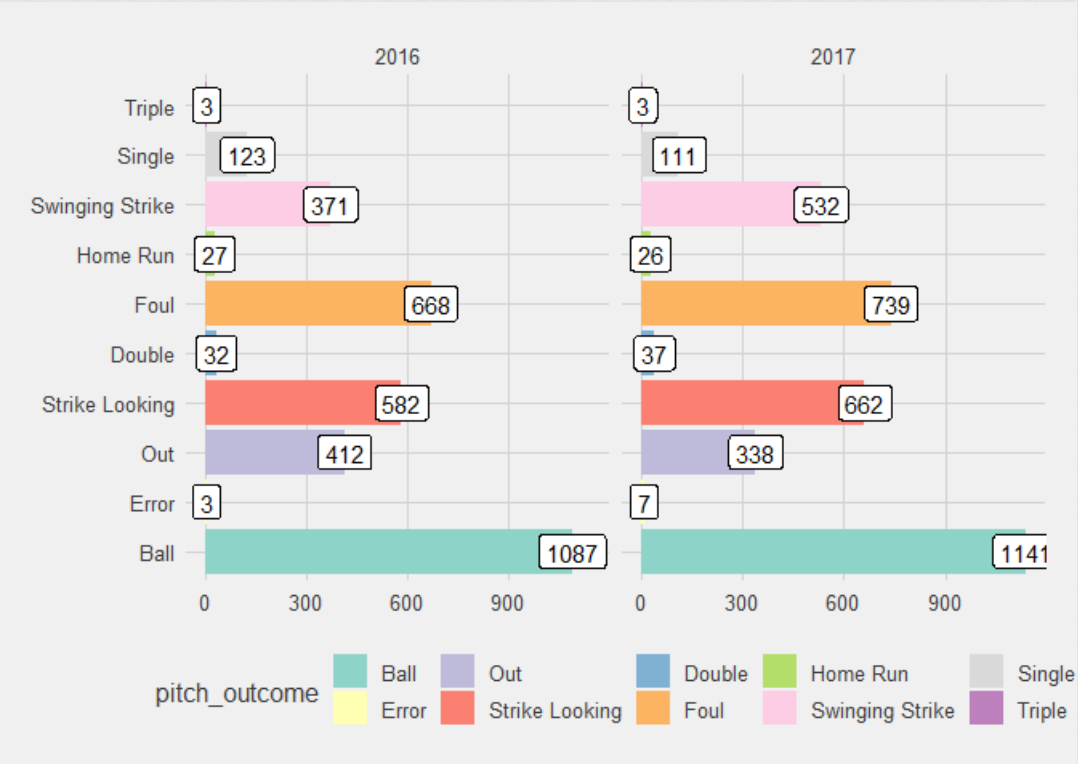
The other way to examine the location is to compare Sale's time at before and after the trade.



Plot 10: Location by Pitch Type and Team

This plot gives us no real insight, the shapes and hot spots are very similar if not the exact same for both teams. Showing, all thought Sale might have changed what pitches he throws and how often, after the trade he didn't change the location.

The last thing to look at will be the outcome of Sale's pitches, to keep things simple, just the years 2016 and 2017, Sale's last year in Chicago and his first year in Boston, the years that where mentioned regarding the different game plans earlier.



Plot 11: Outcome by Year

We know the 2016 game plan was to get more balls hit into play and get outs that way, compared to the 2017 plan of going back to getting swings and misses and strike outs. The data back ups that is what happened. 2016

had a total of 953 strikes (swing strikes plus Strike looking) and 412 outs (note an out is a ball hit in to play that results in an out), compared to 2017 which had 1,194 strikes and just 338 outs. The other points of notes are that balls and fouls went up in. The balls are expected due to trying to get batters to swing at balls outside the strike zone is a key part of getting strike outs. The fouls are more surprising, I would have expected the 2016 plan to get more balls in play would yield more bats on balls and thus more fouls, this is not the case. Equally surprising is the hits (hits = Single + Double + Triple + Home Run) are about equally for each year, with 2016 at 185 and 2017 at 177 (note in 2017 sale throw 288 more pitches than in 2016).

Conclusion

From looking at the data, it's become clear the biggest factor on the change on the pitch type frequency has been Sale's move from Chicago to Boston, and the change in game plane that came with it. In the move the Two-Seamer was replaced by the Four-Seamer while the Curveball was removed. As well as no pitch was thrown more than 40%.

The Location of the pitches was able to be shown, clear target areas for most pitches where easy to identify as well as the spots the pitches where missed.

On top of answering the two question that where originally set, the difference of Sale throwing to a left or right handed batter was explored with, both hands facing the same locations of pitches, but left handers face less Changeup's and more Sliders. The outcome of pitches from before and after Sale's trade where also looked at. Finding that the change of plan did have a major effect on the outcome of the thrown pitch, in the way he got out but not result of balls hit into play against him.

Reflection

There is plenty more that could be explored. Sale made the playoffs for the first time in 2017 and then won the World Series in 2018, it would be interesting to see if anything changed in playoff environment. A look at the speed and nasty score of the balls Sale throws would also be useful, to judge his form. A deep dive in to specific situations, such as 1 out bottom of the 9th winning by one run, and a 3-2 count, and may other things could be done with the data given the time.

Steps that could be taken to improve what has been done in this report are a look at a different way to scrape the data, as pitchRx has stop being updated making it hard to get the playoff games for 2018 and any games past that. I would recommend looking in to the baseballr package (even though it uses the pitchRx package) or scrape that data from another program. Also add some more in-depth stats, such as WAR and ERA, to judge performance from year to year.

References/Bibliography

- raw code for this project
- pitchRx
- Baseball pitches illustrated
- Baseball Reference
- The Ringer

all data wrangling was done in R, as where all the plots, with the following packages:

- pitchRx, package to scrape data
- RSQLite, loading in data of hard drive into r
- readr, read in files
- dplyr, data wrangling
- tidyr, data wrangling
- forcats, data wrangling
- lubridate, data wrangling
- outliers, dealing with outliers

- corrplot, correlation plot
- RColorBrewer, colour package
- viridis, colour package
- ggplot2, plots
- kableExtra, tables
- ggthemes, plot themes