What is Statcast?

Statcast collects data using high-resolution optical cameras along with radar equipment that has been installed in all 30 Major League ballparks since 2015. The technology tracks the location and movements of the ball and every player on the field at any given time.

In this notebook I am going to analyze and visualize statcast data from Aaron Judge and Giancarlo Stanton. There are two CSV files, judge.csv and stanton.csv, both of which contain Statcast data for 2015-2017.

In [2]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

In [3]:
```
# Load the csv files into pandas dataframes
judge = pd.read_csv('judge.csv')
stanton = pd.read_csv('stanton.csv')
```

In [4]:
```
# Display all columns so pandas does not automatically collapse them
pd.set_option('display.max_columns', None)

# Display the last five rows of the Aaron Judge data
judge.tail(5)
```

Out[4]:

	pitch_type	game_date	release_speed	release_pos_x	release_pos_z	player_name	batter	
3431	CH	2016-08-13	85.6	-1.9659	5.9113	Aaron Judge	592450	5
3432	CH	2016-08-13	87.6	-1.9318	5.9349	Aaron Judge	592450	5
3433	CH	2016-08-13	87.2	-2.0285	5.8656	Aaron Judge	592450	5
3434	CU	2016-08-13	79.7	-1.7108	6.1926	Aaron Judge	592450	5
3435	FF	2016-08-13	93.2	-1.8476	6.0063	Aaron Judge	592450	5

Batted Ball Events

Batted ball events are any batted balls that produce an outcome. This means outs, hits, and errors.

In [5]:
```
# All of Aaron Judge's batted ball events in 2017
judge_events_2017 = judge.loc[judge['game_year'] == 2017].events
print("Aaron Judge batted ball event totals, 2017:")
print(judge_events_2017.value_counts())

# All of Giancarlo Stanton's batted ball events in 2017
stanton_events_2017 = stanton.loc[stanton['game_year'] == 2017].events
print("\nGiancarlo Stanton batted ball event totals, 2017:")
print(stanton_events_2017.value_counts())
```

```
Aaron Judge batted ball event totals, 2017:
strikeout                   207
field_out                   146
walk                        116
single                       75
home_run                     52
double                       24
grounded_into_double_play    15
force_out                    11
intent_walk                  11
hit_by_pitch                  5
field_error                   4
sac_fly                       4
fielders_choice_out           4
triple                        3
strikeout_double_play         1
Name: events, dtype: int64

Giancarlo Stanton batted ball event totals, 2017:
field_out                   239
strikeout                   161
single                       77
walk                         72
home_run                     59
double                       32
grounded_into_double_play    13
intent_walk                  13
hit_by_pitch                  7
force_out                     7
field_error                   5
sac_fly                       3
strikeout_double_play         2
fielders_choice_out           2
pickoff_1b                    1
Name: events, dtype: int64
```

We can see that the stats are similar, particularly home runs. Stanton and Judge led baseball in home runs in 2017, with 59 and 52, respectively.

Now let us get into two statistics that were introduced by Statcast, launch angle and launch speed which I have defined below.

Launch angle: the vertical angle at which the ball leaves the bat
Launch speed: the speed of the baseball as it leaves the bat

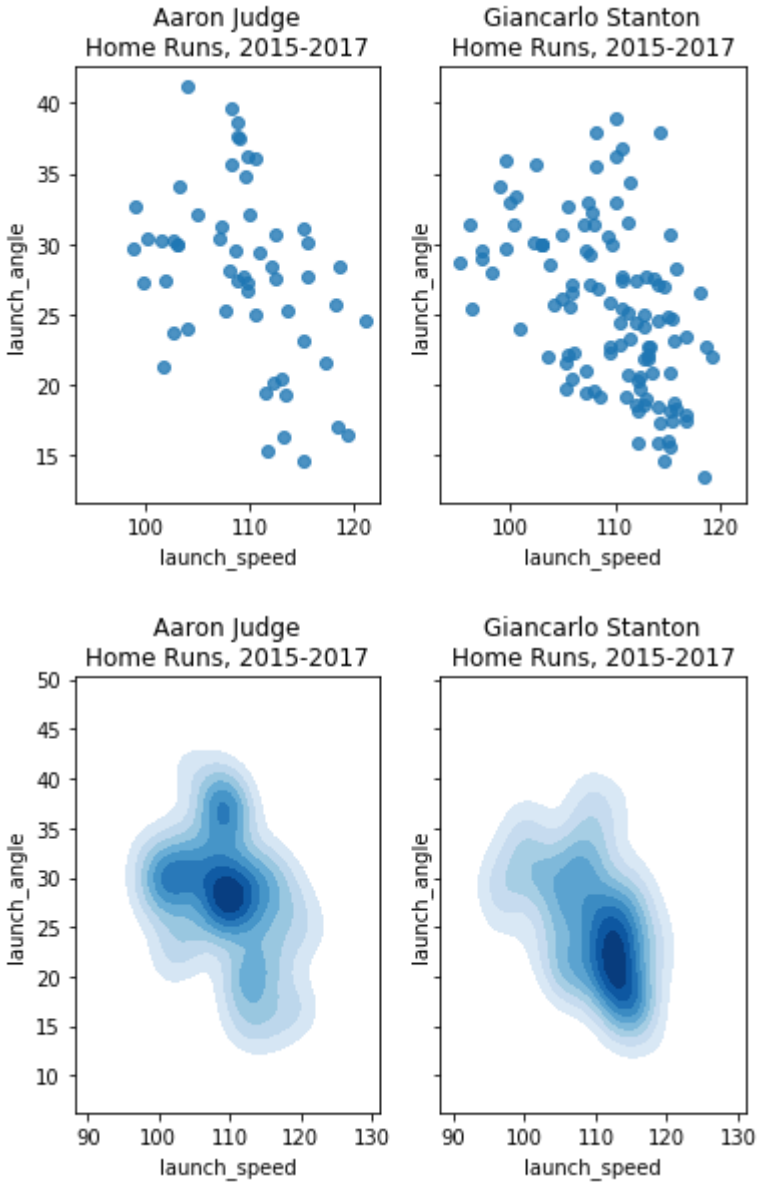
In general, balls with a greater launch angle are more likely to result in a hit.

In [6]:
```
# Filter to include home runs only
judge_hr = judge.loc[judge['events'] == 'home_run']
stanton_hr = stanton.loc[stanton['events'] == 'home_run']

# Create a figure with two scatter plots of launch speed vs. launch angle, one fo
fig1, axs1 = plt.subplots(ncols=2, sharex=True, sharey=True)
sns.regplot(x="launch_speed", y="launch_angle", fit_reg=False, color='tab:blue',
sns.regplot(x="launch_speed", y="launch_angle", fit_reg=False, color='tab:blue',

# Create a figure with two KDE plots of launch speed vs. launch angle, one for ea
fig2, axs2 = plt.subplots(ncols=2, sharex=True, sharey=True)
sns.kdeplot(judge_hr.launch_speed, judge_hr.launch_angle, cmap="Blues", shade=Tru
sns.kdeplot(stanton_hr.launch_speed, stanton_hr.launch_angle, cmap="Blues", shade
```

Out[6]: Text(0.5, 1.0, 'Giancarlo Stanton\nHome Runs, 2015-2017')



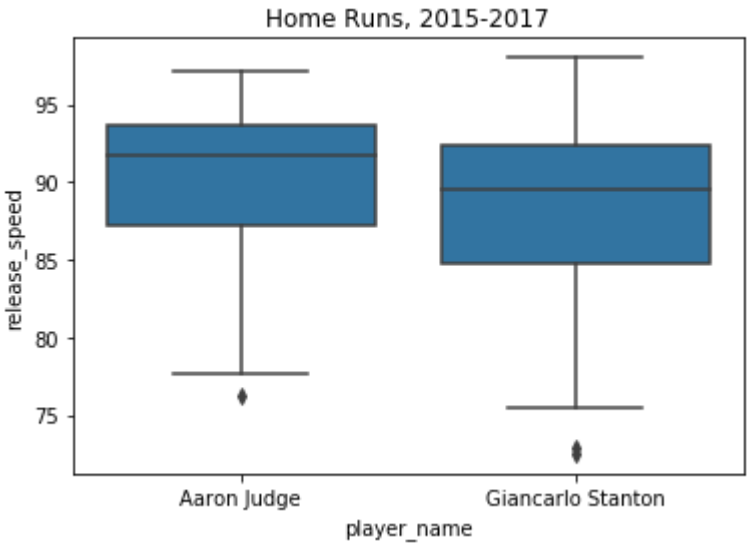
We can see from the charts that Stanton's homeruns generally have a lower launch speed and higher launch angle than Judge's homeruns.

Another important stat is release speed, how fast the ball leaves the pitchers hand before Stanton or Judge hit it. We can use this as another way to compare their homeruns.

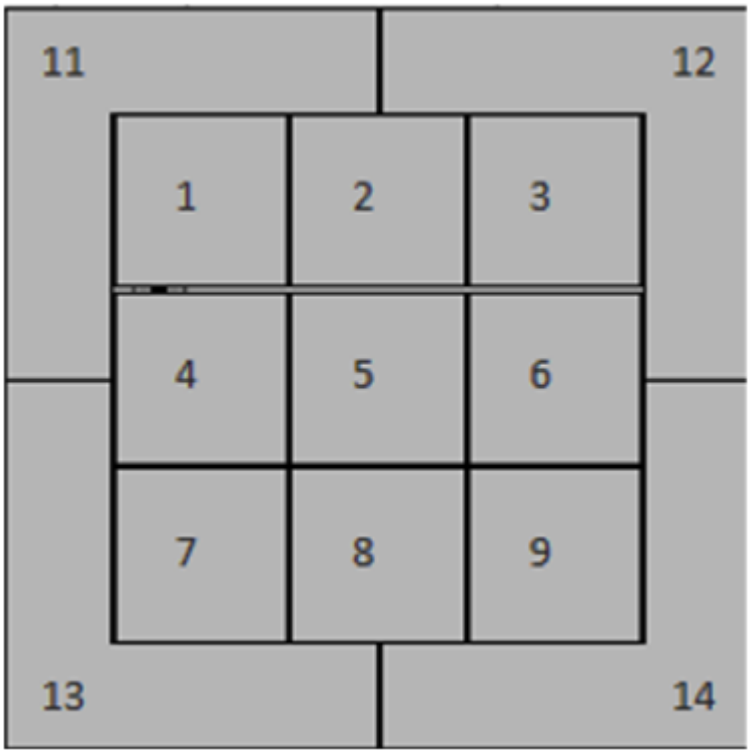
In [7]:
```
# Combine the Judge and Stanton home run DataFrames for easy boxplot plotting
judge_stanton_hr = pd.concat([judge_hr, stanton_hr])

# Create a boxplot that describes the pitch velocity of each player's home runs
sns.boxplot(x='player_name', y='release_speed', data=judge_stanton_hr, color='tab
```

Out[7]:
```
Text(0.5, 1.0, 'Home Runs, 2015-2017')
```



Judge hits his homeruns on faster pitches than Stanton making him more of a fastball hitter. Statcast also tracks pitch location. The zone the pitch is in when it crosses the plate. The zone numbering goes from 1-14 with 11-14 being outside of the strikezone. (Pictured below)



We can plot this with a 2D histogram on a 9x9 grid. We can view each zone as coordinates on a 2D plot, the bottom left corner being (1,1) and the top right corner being (3,3). Below I will setup a function to assign x coordinates to the strikezone.

In [8]:
```
def assign_x_coord(row):
    """
    Assigns an x-coordinate to Statcast's strike zone numbers. Zones 11, 12, 13,
    and 14 are ignored since they are not in the strikezone.
    """
    # Left third of strike zone
    if row.zone in [1, 4, 7]:
        return 1
    # Middle third of strike zone
    if row.zone in [2, 5, 8]:
        return 2
    # Right third of strike zone
    if row.zone in [3, 6, 9]:
        return 3
```

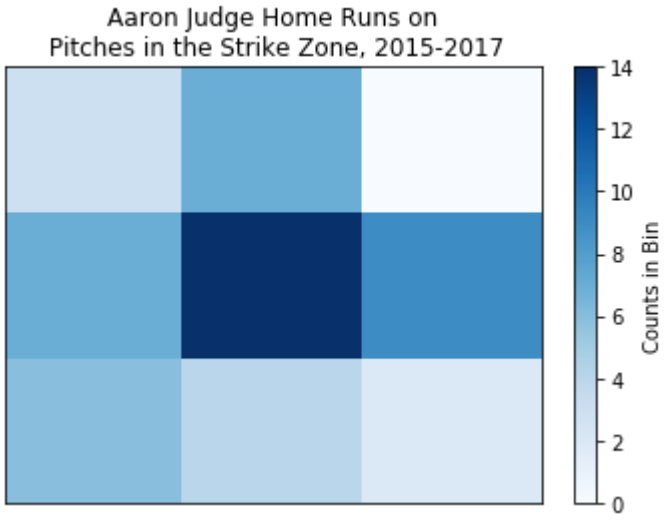
In [9]:
```
def assign_y_coord(row):
    """
    Assigns a y-coordinate to Statcast's strike zone numbers. Zones 11, 12, 13,
    and 14 are ignored since they are not in the strikezone.
    """
    # Upper third of strike zone
    if row.zone in [1, 2, 3]:
        return 3
    # Middle third of strike zone
    if row.zone in [4, 5, 6]:
        return 2
    # Lower third of strike zone
    if row.zone in [7, 8, 9]:
        return 1
```

First we'll create our 2D histogram for Aaron Judge. Remember this is for pitches in the strike zone that resulted in a homerun.

In [10]:
```
# Zones 11, 12, 13, and 14 are to be ignored for plotting simplicity
judge_strike_hr = judge_hr.copy().loc[judge_hr.zone <= 9]

# Assign Cartesian coordinates to pitches in the strike zone for Judge home runs
judge_strike_hr['zone_x'] = judge_strike_hr.apply(assign_x_coord, axis=1)
judge_strike_hr['zone_y'] = judge_strike_hr.apply(assign_y_coord, axis=1)

# Plot Judge's home run zone as a 2D histogram with a colorbar
plt.hist2d(judge_strike_hr['zone_x'], judge_strike_hr['zone_y'], bins = 3, cmap='
plt.title('Aaron Judge Home Runs on\n Pitches in the Strike Zone, 2015-2017')
plt.gca().get_xaxis().set_visible(False)
plt.gca().get_yaxis().set_visible(False)
cb = plt.colorbar()
cb.set_label('Counts in Bin')
```

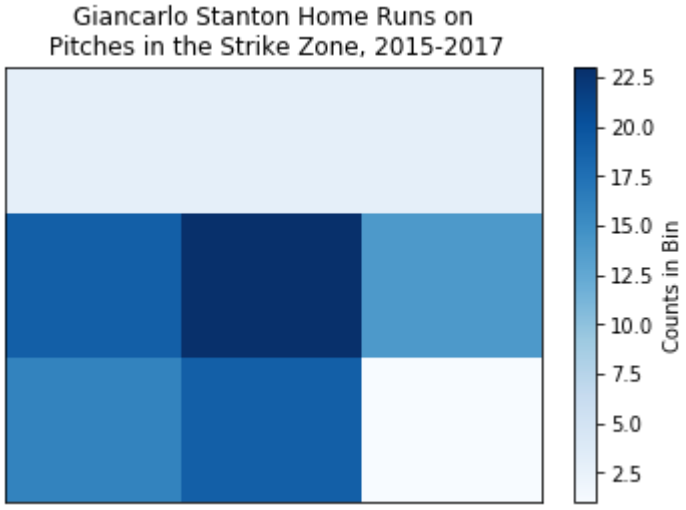


Next we can do the same for Giancarlo Stanton.

In [11]:
```
# Zones 11, 12, 13, and 14 are to be ignored for plotting simplicity
stanton_strike_hr = stanton_hr.copy().loc[stanton_hr.zone <= 9]

# Assign Cartesian coordinates to pitches in the strike zone for Stanton home run
stanton_strike_hr['zone_x'] = stanton_strike_hr.apply(assign_x_coord, axis=1)
stanton_strike_hr['zone_y'] = stanton_strike_hr.apply(assign_y_coord, axis=1)


# Plot Stanton's home run zone as a 2D histogram with a colorbar
plt.hist2d(stanton_strike_hr['zone_x'], stanton_strike_hr['zone_y'], bins = 3, cm
plt.title('Giancarlo Stanton Home Runs on\n Pitches in the Strike Zone, 2015-2017
plt.gca().get_xaxis().set_visible(False)
plt.gca().get_yaxis().set_visible(False)
cb = plt.colorbar()
cb.set_label('Counts in Bin')
```



Takeways:

Stanton does not hit many home runs on pitches in the upper third of the strike zone.

Both hitters favor the center of the strikezone.

Judge's least favorite home run pitch appears to be high-away while Stanton's appears to be low-away.

Stanton's preference for homeruns is the middle inside.

Judge's homeruns are a lot more spread out.

Stanton hits homeruns on slower pitches than Judge.

Even though both hitters had very similar homerun numbers their hitting profiles are completely different.
Regardless, I would not want to pitch against either of these hitters.