Andres Izquierdo Take Home Test

In [1]:
```
import csv
import os
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
```

In [2]:
```
# Loading Data
os.chdir('C:/Users/andre/OneDrive/Documents/UVA SYS ME/Job Interview Resources/pitches_
df = pd.DataFrame()
df = pd.read_csv("pitches")
df.info()
pd.set_option("display.max_columns", None)
print(df)
```

```
C:\Users\andre\anaconda3\lib\site-packages\IPython\core\interactiveshell.py:3444: DtypeW
arning: Columns (29,30) have mixed types.Specify dtype option on import or set low_memor
y=False.
  exec(code_obj, self.user_global_ns, self.user_ns)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 718961 entries, 0 to 718960
Columns: 125 entries, uid to modified_by
dtypes: float64(80), int64(25), object(20)
memory usage: 685.7+ MB
            uid  game_pk  year        date  team_id_b  team_id_p  inning  \
0      14143226   286874  2011  2011-03-31        108        118       1
1      14143227   286874  2011  2011-03-31        108        118       1
2      14143228   286874  2011  2011-03-31        108        118       1
3      14143229   286874  2011  2011-03-31        108        118       1
4      14143230   286874  2011  2011-03-31        108        118       1
...         ...      ...   ...         ...        ...        ...     ...
718956  19838192   317073  2011  2011-10-28        140        138       9
718957  19838193   317073  2011  2011-10-28        140        138       9
718958  19838194   317073  2011  2011-10-28        140        138       9
718959  19838195   317073  2011  2011-10-28        140        138       9
718960  19838196   317073  2011  2011-10-28        140        138       9

        top  at_bat_num  pcount_at_bat  pcount_pitcher  balls  strikes  fouls  \
0         1           1              1               1      0        0      0
1         1           1              2               2      1        0      0
2         1           1              3               3      2        0      0
3         1           1              4               4      2        1      0
4         1           2              1               5      0        0      0
...     ...         ...            ...             ...    ...      ...    ...
718956    1          72              3               7      1        1      0
718957    1          72              4               8      1        2      1
718958    1          72              5               9      2        2      1
718959    1          73              1              10      0        0      0
718960    1          73              2              11      0        1      0
```

	outs	is_final_pitch	final_balls	final_strikes	final_outs
0	0	0	2	1	1
1	0	0	2	1	1
2	0	0	2	1	1
3	0	1	2	1	1
4	1	0	2	2	1
...	...	...	...	...	...
718956	1	0	2	2	2
718957	1	0	2	2	2
718958	1	1	2	2	2
718959	2	0	0	1	3
718960	2	1	0	1	3

	start_tfs	start_tfs_zulu	batter_id	stand	b_height	pitcher_id
0	201226	2011-03-31 20:12:26	430895	L	5-8	460024
1	201226	2011-03-31 20:12:26	430895	L	5-8	460024
2	201226	2011-03-31 20:12:26	430895	L	5-8	460024
3	201226	2011-03-31 20:12:26	430895	L	5-8	460024
4	201354	2011-03-31 20:13:54	435062	R	5-10	460024
...	...	...	...	...	...	...
718956	31934	2011-10-29 03:19:34	435063	R	6-0	435400
718957	31934	2011-10-29 03:19:34	435063	R	6-0	435400
718958	31934	2011-10-29 03:19:34	435063	R	6-0	435400
718959	32141	2011-10-29 03:21:41	461815	L	6-3	435400
718960	32141	2011-10-29 03:21:41	461815	L	6-3	435400

	p_throws	at_bat_des	event
0	R	Maicer Izturis grounds out, second baseman Chr...	Groundout
1	R	Maicer Izturis grounds out, second baseman Chr...	Groundout
2	R	Maicer Izturis grounds out, second baseman Chr...	Groundout
3	R	Maicer Izturis grounds out, second baseman Chr...	Groundout
4	R	Howie Kendrick doubles (1) on a line drive to ...	Double
...	...	...	...
718956	R	Mike Napoli grounds out, third baseman Daniel ...	Groundout
718957	R	Mike Napoli grounds out, third baseman Daniel ...	Groundout
718958	R	Mike Napoli grounds out, third baseman Daniel ...	Groundout
718959	R	David Murphy flies out to left fielder Allen C...	Flyout
718960	R	David Murphy flies out to left fielder Allen C...	Flyout

	event2	event3	event4	away_team_runs	home_team_runs	score
0	NaN	NaN	NaN	0	0	NaN
1	NaN	NaN	NaN	0	0	NaN
2	NaN	NaN	NaN	0	0	NaN
3	NaN	NaN	NaN	0	0	NaN
4	NaN	NaN	NaN	0	0	NaN
...	...	...	...	...	...	...
718956	NaN	NaN	NaN	2	6	NaN
718957	NaN	NaN	NaN	2	6	NaN
718958	NaN	NaN	NaN	2	6	NaN
718959	NaN	NaN	NaN	2	6	NaN
718960	NaN	NaN	NaN	2	6	NaN

	pitch_des	pitch_id	type	pitch_tfs	pitch_tfs_zulu
0	Ball	3	B	201301.0	2011-03-31 20:13:01
1	Ball	4	B	201319.0	2011-03-31 20:13:19
2	Called Strike	5	S	201327.0	2011-03-31 20:13:27
3	In play, out(s)	6	X	180441.0	2011-03-31 18:04:41
4	Called Strike	10	S	201404.0	2011-03-31 20:14:04
...	...	...	...	...	...
718956	Foul	614	S	32000.0	2011-10-29 03:20:00

```
718957            Ball       615    B    32030.0  2011-10-29 03:20:30
718958  In play, out(s)      616    X   223106.0  2011-10-28 22:31:06
718959    Called Strike      620    S    32140.0  2011-10-29 03:21:40
718960  In play, out(s)      621    X    32222.0  2011-10-29 03:22:22

             x       y          sv_id  start_speed  end_speed  sz_top  sz_bot  \
0       105.58  180.46            NaN          NaN        NaN     NaN     NaN
1        99.57  170.96            NaN          NaN        NaN     NaN     NaN
2        95.28  152.83            NaN          NaN        NaN     NaN     NaN
3        93.56  168.37            NaN          NaN        NaN     NaN     NaN
4        99.57  170.96            NaN          NaN        NaN     NaN     NaN
...        ...     ...            ...          ...        ...     ...     ...
718956  102.15  140.74  111028_221959         97.9       89.1    3.14    1.50
718957   90.99  170.10  111028_222029         91.4       84.3    3.14    1.43
718958  109.01  155.42  111028_222049         97.3       88.0    3.14    1.50
718959  105.58  158.01  111028_222139         93.1       85.3    3.74    1.76
718960  102.15  129.52  111028_222158         97.4       88.3    3.74    1.76

        pfx_x  pfx_z     px     pz     x0     z0    y0    vx0    vz0      vy0  \
0         NaN    NaN    NaN    NaN    NaN    NaN   NaN    NaN    NaN      NaN
1         NaN    NaN    NaN    NaN    NaN    NaN   NaN    NaN    NaN      NaN
2         NaN    NaN    NaN    NaN    NaN    NaN   NaN    NaN    NaN      NaN
3         NaN    NaN    NaN    NaN    NaN    NaN   NaN    NaN    NaN      NaN
4         NaN    NaN    NaN    NaN    NaN    NaN   NaN    NaN    NaN      NaN
...       ...    ...    ...    ...    ...    ...   ...    ...    ...      ...
718956  -3.35  10.56 -0.069  2.689 -0.899  5.513  50.0  3.560 -6.112 -143.374
718957   0.94   3.39  0.275  1.446 -0.863  5.415  50.0  2.680 -5.564 -133.947
718958  -4.43   9.53 -0.261  2.122 -0.774  5.554  50.0  3.031 -7.265 -142.543
718959 -10.53   3.05 -0.136  2.053 -0.989  5.394  50.0  5.952 -4.042 -136.340
718960  -4.45   8.56 -0.561  3.704 -0.851  5.730  50.0  2.425 -3.044 -142.847

            ax      az      ay  break_length  break_y  break_angle pitch_type  \
0          NaN     NaN     NaN           NaN      NaN          NaN        NaN
1          NaN     NaN     NaN           NaN      NaN          NaN        NaN
2          NaN     NaN     NaN           NaN      NaN          NaN        NaN
3          NaN     NaN     NaN           NaN      NaN          NaN        NaN
4          NaN     NaN     NaN           NaN      NaN          NaN        NaN
...        ...     ...     ...           ...      ...          ...        ...
718956  -6.886 -10.386  36.837           2.6     23.7         26.6         FF
718957   1.707 -25.962  29.860           6.1     23.8         -4.7         FC
718958  -8.925 -12.891  38.935           3.4     23.7         28.7         FF
718959 -19.664 -26.404  32.209           7.2     23.8         33.8         FT
718960  -9.040 -14.714  38.276           3.5     23.7         28.4         FF

        type_confidence  zone  nasty  spin_dir  spin_rate   cc  on_1b  on_2b  \
0                   NaN   NaN    NaN       NaN        NaN  NaN    NaN    NaN
1                   NaN   NaN    NaN       NaN        NaN  NaN    NaN    NaN
2                   NaN   NaN    NaN       NaN        NaN  NaN    NaN    NaN
3                   NaN   NaN    NaN       NaN        NaN  NaN    NaN    NaN
4                   NaN   NaN    NaN       NaN        NaN  NaN    NaN    NaN
...                 ...   ...    ...       ...        ...  ...    ...    ...
718956              2.0   2.0   29.0   197.537   2312.186  NaN    NaN    NaN
718957              2.0   9.0   47.0   164.636    697.763  NaN    NaN    NaN
718958              2.0   4.0   29.0   204.835   2162.620  NaN    NaN    NaN
718959              2.0   8.0   44.0   253.646   2180.650  NaN    NaN    NaN
718960              2.0   1.0   57.0   207.372   1996.857  NaN    NaN    NaN

        on_3b  runner1_id  runner1_start  runner1_end  runner1_event  \
0         NaN         NaN            NaN          NaN            NaN
1         NaN         NaN            NaN          NaN            NaN
```

```
2         NaN         NaN            NaN          NaN            NaN
3         NaN         NaN            NaN          NaN            NaN
4         NaN         NaN            NaN          NaN            NaN
...       ...         ...            ...          ...            ...
718956    NaN         NaN            NaN          NaN            NaN
718957    NaN         NaN            NaN          NaN            NaN
718958    NaN         NaN            NaN          NaN            NaN
718959    NaN         NaN            NaN          NaN            NaN
718960    NaN         NaN            NaN          NaN            NaN

        runner1_score  runner1_rbi  runner1_earned  runner2_id  runner2_start  \
0                 NaN          NaN             NaN         NaN            NaN
1                 NaN          NaN             NaN         NaN            NaN
2                 NaN          NaN             NaN         NaN            NaN
3                 NaN          NaN             NaN         NaN            NaN
4                 NaN          NaN             NaN         NaN            NaN
...               ...          ...             ...         ...            ...
718956            NaN          NaN             NaN         NaN            NaN
718957            NaN          NaN             NaN         NaN            NaN
718958            NaN          NaN             NaN         NaN            NaN
718959            NaN          NaN             NaN         NaN            NaN
718960            NaN          NaN             NaN         NaN            NaN

        runner2_end  runner2_event  runner2_score  runner2_rbi  \
0               NaN            NaN            NaN          NaN
1               NaN            NaN            NaN          NaN
2               NaN            NaN            NaN          NaN
3               NaN            NaN            NaN          NaN
4               NaN            NaN            NaN          NaN
...             ...            ...            ...          ...
718956          NaN            NaN            NaN          NaN
718957          NaN            NaN            NaN          NaN
718958          NaN            NaN            NaN          NaN
718959          NaN            NaN            NaN          NaN
718960          NaN            NaN            NaN          NaN

        runner2_earned  runner3_id  runner3_start  runner3_end  runner3_event  \
0                  NaN         NaN            NaN          NaN            NaN
1                  NaN         NaN            NaN          NaN            NaN
2                  NaN         NaN            NaN          NaN            NaN
3                  NaN         NaN            NaN          NaN            NaN
4                  NaN         NaN            NaN          NaN            NaN
...                ...         ...            ...          ...            ...
718956             NaN         NaN            NaN          NaN            NaN
718957             NaN         NaN            NaN          NaN            NaN
718958             NaN         NaN            NaN          NaN            NaN
718959             NaN         NaN            NaN          NaN            NaN
718960             NaN         NaN            NaN          NaN            NaN

        runner3_score  runner3_rbi  runner3_earned  runner4_id  runner4_start  \
0                 NaN          NaN             NaN         NaN            NaN
1                 NaN          NaN             NaN         NaN            NaN
2                 NaN          NaN             NaN         NaN            NaN
3                 NaN          NaN             NaN         NaN            NaN
4                 NaN          NaN             NaN         NaN            NaN
...               ...          ...             ...         ...            ...
718956            NaN          NaN             NaN         NaN            NaN
718957            NaN          NaN             NaN         NaN            NaN
718958            NaN          NaN             NaN         NaN            NaN
718959            NaN          NaN             NaN         NaN            NaN
```

```
718960           NaN          NaN             NaN         NaN            NaN

        runner4_end  runner4_event  runner4_score  runner4_rbi  \
0               NaN            NaN            NaN          NaN
1               NaN            NaN            NaN          NaN
2               NaN            NaN            NaN          NaN
3               NaN            NaN            NaN          NaN
4               NaN            NaN            NaN          NaN
...             ...            ...            ...          ...
718956          NaN            NaN            NaN          NaN
718957          NaN            NaN            NaN          NaN
718958          NaN            NaN            NaN          NaN
718959          NaN            NaN            NaN          NaN
718960          NaN            NaN            NaN          NaN

        runner4_earned  runner5_id  runner5_start  runner5_end  runner5_event  \
0                  NaN         NaN            NaN          NaN            NaN
1                  NaN         NaN            NaN          NaN            NaN
2                  NaN         NaN            NaN          NaN            NaN
3                  NaN         NaN            NaN          NaN            NaN
4                  NaN         NaN            NaN          NaN            NaN
...                ...         ...            ...          ...            ...
718956             NaN         NaN            NaN          NaN            NaN
718957             NaN         NaN            NaN          NaN            NaN
718958             NaN         NaN            NaN          NaN            NaN
718959             NaN         NaN            NaN          NaN            NaN
718960             NaN         NaN            NaN          NaN            NaN

        runner5_score  runner5_rbi  runner5_earned  runner6_id  runner6_start  \
0                 NaN          NaN             NaN         NaN            NaN
1                 NaN          NaN             NaN         NaN            NaN
2                 NaN          NaN             NaN         NaN            NaN
3                 NaN          NaN             NaN         NaN            NaN
4                 NaN          NaN             NaN         NaN            NaN
...               ...          ...             ...         ...            ...
718956            NaN          NaN             NaN         NaN            NaN
718957            NaN          NaN             NaN         NaN            NaN
718958            NaN          NaN             NaN         NaN            NaN
718959            NaN          NaN             NaN         NaN            NaN
718960            NaN          NaN             NaN         NaN            NaN

        runner6_end  runner6_event  runner6_score  runner6_rbi  \
0               NaN            NaN            NaN          NaN
1               NaN            NaN            NaN          NaN
2               NaN            NaN            NaN          NaN
3               NaN            NaN            NaN          NaN
4               NaN            NaN            NaN          NaN
...             ...            ...            ...          ...
718956          NaN            NaN            NaN          NaN
718957          NaN            NaN            NaN          NaN
718958          NaN            NaN            NaN          NaN
718959          NaN            NaN            NaN          NaN
718960          NaN            NaN            NaN          NaN

        runner6_earned  runner7_id  runner7_start  runner7_end  runner7_event  \
0                  NaN         NaN            NaN          NaN            NaN
1                  NaN         NaN            NaN          NaN            NaN
2                  NaN         NaN            NaN          NaN            NaN
3                  NaN         NaN            NaN          NaN            NaN
4                  NaN         NaN            NaN          NaN            NaN
```

```
...              ...         ...            ...          ...            ...
718956           NaN         NaN            NaN          NaN            NaN
718957           NaN         NaN            NaN          NaN            NaN
718958           NaN         NaN            NaN          NaN            NaN
718959           NaN         NaN            NaN          NaN            NaN
718960           NaN         NaN            NaN          NaN            NaN

        runner7_score  runner7_rbi  runner7_earned           created_at  \
0                 NaN          NaN             NaN  2016-03-03 21:33:20
1                 NaN          NaN             NaN  2016-03-03 21:33:20
2                 NaN          NaN             NaN  2016-03-03 21:33:20
3                 NaN          NaN             NaN  2016-03-03 21:33:20
4                 NaN          NaN             NaN  2016-03-03 21:33:20
...               ...          ...             ...                  ...
718956            NaN          NaN             NaN  2016-03-03 22:23:19
718957            NaN          NaN             NaN  2016-03-03 22:23:19
718958            NaN          NaN             NaN  2016-03-03 22:23:19
718959            NaN          NaN             NaN  2016-03-03 22:23:19
718960            NaN          NaN             NaN  2016-03-03 22:23:19

                  added_at          modified_at  modified_by
0       2016-03-03 21:33:20  2016-03-03 21:33:20            1
1       2016-03-03 21:33:20  2016-03-03 21:33:20            1
2       2016-03-03 21:33:20  2016-03-03 21:33:20            1
3       2016-03-03 21:33:20  2016-03-03 21:33:20            1
4       2016-03-03 21:33:20  2016-03-03 21:33:20            1
...                     ...                  ...          ...
718956  2016-03-03 22:23:19  2016-03-03 22:23:19            1
718957  2016-03-03 22:23:19  2016-03-03 22:23:19            1
718958  2016-03-03 22:23:19  2016-03-03 22:23:19            1
718959  2016-03-03 22:23:19  2016-03-03 22:23:19            1
718960  2016-03-03 22:23:19  2016-03-03 22:23:19            1

[718961 rows x 125 columns]
```

Counting the number of pitch types in the dataframe.

In [3]:

```
# Counting all the different types of pitches there are
df = df[df['pitch_type'].notna()] # removing NaNs from pitch_type column
df['pitch_type'].value_counts()
```

Out[3]:

```
FF    238541
SL    109756
SI     87740
FT     81056
CH     72641
CU     56379
FC     41702
FS     10503
KC      8490
KN      4450
IN      4058
PO       559
FO       329
FA       204
EP       134
SC       120
UN        17
```

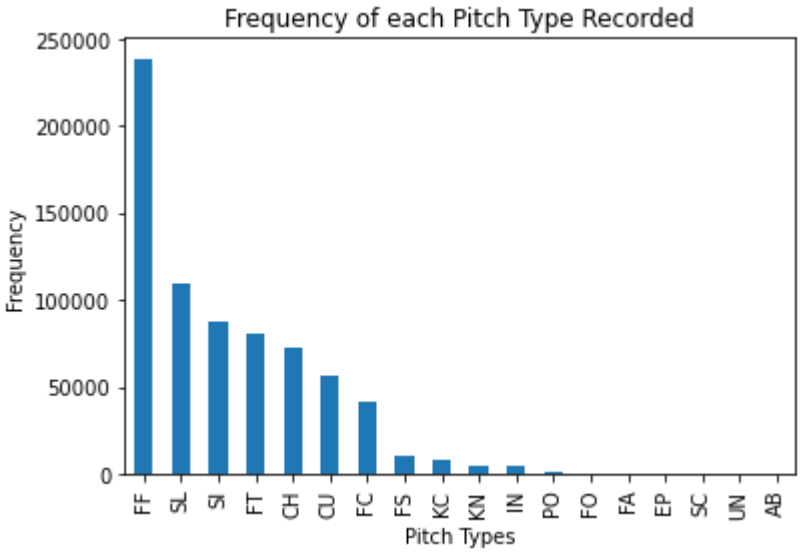
```
AB          2
Name: pitch_type, dtype: int64
```

In [4]:

```
# Visualization of the amount of each pitch type
ax = df['pitch_type'].value_counts().plot(kind='bar',
                                          title="Frequency of each Pitch Type Recorded")
ax.set_xlabel("Pitch Types")
ax.set_ylabel("Frequency")
```

Out[4]:

```
Text(0, 0.5, 'Frequency')
```



I will be creating two Logistic Regression models for predicting the probability of fastballs and sliders. The process for building these models will be the same for both and any other pitches I consider doing in the future.

In [5]:

```
# Creating two dataframes so I can do feature engineering on the dataframe in order to
FF = df.copy()
SL = df.copy()
```

In [6]:

```
# Setting Fastball indicator as 1 and all other pitch types to zero in order get model
FF["pitch_type"].replace({"FF": "1"
                          , "SL": "0"
                          , "SI": "0"
                          , "FT": "0"
                          , "CH": "0"
                          , "CU": "0"
                          , "FC": "0"
                          , "FS": "0"
                          , "KC": "0"
                          , "KN": "0"
                          , "IN": "0"
                          , "PO": "0"
                          , "FO": "0"
                          , "FA": "0"
                          , "EP": "0"
                          , "SC": "0"
                          , "UN": "0"
                          , "AB": "0"}, inplace=True)
```

```
# making the pitch type column into numeric in order to do analysis
FF["pitch_type"] = pd.to_numeric(FF["pitch_type"])
# confirming total number of fastballs in data frame
FF['pitch_type'].value_counts()[1]
```

Out[6]:
```
238541
```

In [7]:
```
# Setting Slider indicator as 1 and all other pitch types to zero in order get model re
SL["pitch_type"].replace({"FF": "0"
                          , "SL": "1"
                          , "SI": "0"
                          , "FT": "0"
                          , "CH": "0"
                          , "CU": "0"
                          , "FC": "0"
                          , "FS": "0"
                          , "KC": "0"
                          , "KN": "0"
                          , "IN": "0"
                          , "PO": "0"
                          , "FO": "0"
                          , "FA": "0"
                          , "EP": "0"
                          , "SC": "0"
                          , "UN": "0"
                          , "AB": "0"}, inplace=True)
# making the pitch type column into numeric in order to do analysis
SL["pitch_type"] = pd.to_numeric(SL["pitch_type"])
# confirming total number of sliders in data frame
SL['pitch_type'].value_counts()[1]
```

Out[7]:
```
109756
```

Setting up Fastball Model

In [8]:
```
# Setting up correlation matrix to see what variables show correlation with the pitch t
corr_FF = FF.corr()
corr_FF.style.background_gradient(cmap='coolwarm')
```

```
C:\Users\andre\anaconda3\lib\site-packages\pandas\io\formats\style.py:2813: RuntimeWarni
ng: All-NaN slice encountered
  smin = np.nanmin(gmap) if vmin is None else vmin
C:\Users\andre\anaconda3\lib\site-packages\pandas\io\formats\style.py:2814: RuntimeWarni
ng: All-NaN slice encountered
  smax = np.nanmax(gmap) if vmax is None else vmax
```

Out[8]:

	uid	game_pk	year	team_id_b	team_id_p	inning	top	at_bat_num	p
uid	1.000000	0.390780	nan	0.019080	0.019230	-0.007139	-0.000892	0.001742	
game_pk	0.390780	1.000000	nan	0.071178	0.073782	-0.004490	-0.001576	0.000299	
year	nan	nan	nan	nan	nan	nan	nan	nan	
team_id_b	0.019080	0.071178	nan	1.000000	-0.073066	-0.002456	-0.005154	-0.006554	
team_id_p	0.019230	0.073782	nan	-0.073066	1.000000	0.003914	0.003741	-0.001520	

	uid	game_pk	year	team_id_b	team_id_p	inning	top	at_bat_num	p
inning	-0.007139	-0.004490	nan	-0.002456	0.003914	1.000000	0.040169	0.976341	
top	-0.000892	-0.001576	nan	-0.005154	0.003741	0.040169	1.000000	-0.051980	
at_bat_num	0.001742	0.000299	nan	-0.006554	-0.001520	0.976341	-0.051980	1.000000	
pcount_at_bat	0.002171	0.000916	nan	-0.001440	-0.003043	0.001987	-0.002008	0.002824	
pcount_pitcher	-0.012253	-0.016561	nan	0.004429	0.010576	0.011561	-0.003248	-0.010417	
balls	-0.005117	-0.002658	nan	0.003328	-0.006731	0.000590	-0.006637	0.002917	
strikes	0.005940	0.003369	nan	-0.004021	0.005034	-0.000823	0.003657	-0.001847	
fouls	0.006703	0.003333	nan	-0.007033	-0.006135	0.012520	0.001433	0.012091	
outs	0.001811	-0.000789	nan	0.001415	0.004695	0.006771	-0.000118	0.055724	
is_final_pitch	-0.000431	0.000122	nan	0.000706	0.000362	0.000162	0.000773	-0.000296	
final_balls	-0.010316	-0.002899	nan	0.007661	-0.009953	0.002893	-0.012730	0.006964	
final_strikes	0.010756	0.005459	nan	-0.008023	0.014904	0.009145	0.010341	0.007473	
final_outs	-0.000249	-0.001402	nan	0.001317	0.009067	0.001277	0.005850	0.042191	
start_tfs	-0.041532	-0.027894	nan	-0.002767	-0.003658	-0.201844	0.011827	-0.202907	
batter_id	0.066375	-0.003258	nan	0.002937	-0.009212	0.007382	0.003876	0.006802	
pitcher_id	0.027437	-0.011546	nan	-0.003897	-0.115061	0.039526	0.005108	0.041579	
away_team_runs	0.026047	0.013042	nan	-0.016549	-0.014822	0.488752	-0.038259	0.582259	
home_team_runs	0.028451	0.022960	nan	-0.019408	-0.016573	0.478855	-0.015593	0.569420	
pitch_id	0.003006	0.001331	nan	-0.009104	-0.004432	0.964648	-0.050919	0.995346	
pitch_tfs	-0.040438	-0.020530	nan	-0.004019	-0.004009	-0.197081	0.011635	-0.197378	
x	-0.010647	-0.004231	nan	0.007697	-0.003205	0.002184	0.001073	0.001569	
y	-0.009731	0.004193	nan	0.019807	0.030795	-0.014789	-0.006033	-0.014862	
start_speed	0.032115	0.021043	nan	-0.011282	-0.003862	0.060975	0.007507	0.061263	
end_speed	0.054650	0.029973	nan	-0.013044	-0.003899	0.055342	0.007583	0.055790	
sz_top	0.033549	0.019338	nan	-0.021110	0.007204	-0.016973	0.018257	-0.016469	
sz_bot	0.021399	0.011670	nan	-0.011309	0.003799	-0.012171	-0.001194	-0.009573	
pfx_x	0.010359	0.004256	nan	-0.011228	0.059711	-0.023553	-0.002222	-0.021485	
pfx_z	-0.010501	0.008258	nan	-0.011585	-0.019493	-0.025919	0.006518	-0.025768	
px	0.011032	0.006416	nan	-0.008223	0.003028	-0.005013	-0.001935	-0.004309	
pz	0.008976	-0.000102	nan	-0.012683	-0.025179	0.012942	0.006758	0.012882	
x0	0.011112	0.002959	nan	-0.006118	0.086777	-0.052607	-0.010203	-0.051102	
z0	0.017892	0.025436	nan	-0.006460	0.024459	-0.137384	0.003359	-0.140359	
y0	nan	nan	nan	nan	nan	nan	nan	nan	

	uid	game_pk	year	team_id_b	team_id_p	inning	top	at_bat_num	p
vx0	-0.005844	-0.001067	nan	0.004899	-0.087047	0.048075	0.007537	0.046556	
vz0	-0.009883	-0.023671	nan	0.004008	-0.016471	0.052252	-0.003004	0.053018	
vy0	-0.032322	-0.021063	nan	0.011329	0.003127	-0.061138	-0.007689	-0.061357	
ax	0.008789	0.004565	nan	-0.011317	0.067020	-0.028223	-0.002782	-0.026102	
az	-0.004819	0.013524	nan	-0.011578	-0.017629	-0.016758	0.007082	-0.016473	
ay	-0.067526	-0.019262	nan	0.000625	0.006032	0.077492	0.004551	0.076771	
break_length	-0.014105	-0.011245	nan	0.016611	0.014910	-0.004461	-0.007634	-0.004803	
break_y	0.124168	0.048114	nan	-0.008871	0.002568	-0.041883	-0.000762	-0.041053	
break_angle	-0.007035	0.000098	nan	0.010295	-0.067054	0.034019	0.005419	0.031633	
pitch_type	0.014347	0.004832	nan	-0.013564	-0.015156	0.015466	0.009027	0.015863	
type_confidence	0.020910	0.020328	nan	0.005219	-0.081324	-0.008850	0.011183	-0.015331	
zone	0.000313	0.003654	nan	0.003802	0.011479	0.000009	-0.004409	0.000777	
nasty	-0.002411	-0.001969	nan	-0.000778	0.002520	-0.017771	0.000639	-0.018668	
spin_dir	-0.006149	-0.004168	nan	0.007052	-0.032804	-0.001996	-0.003261	-0.002730	
spin_rate	-0.029124	0.013427	nan	0.004885	0.003599	-0.009530	0.002892	-0.010834	
on_1b	0.065175	-0.005285	nan	-0.007411	-0.013273	0.022826	0.015284	0.018019	
on_2b	0.061079	-0.013664	nan	-0.017993	-0.005855	0.016400	0.016719	0.014921	
on_3b	0.077527	-0.012314	nan	-0.000785	-0.017172	0.023869	0.013940	0.022160	
runner1_id	nan	nan	nan	nan	nan	nan	nan	nan	
runner1_start	nan	nan	nan	nan	nan	nan	nan	nan	
runner1_end	nan	nan	nan	nan	nan	nan	nan	nan	
runner1_event	nan	nan	nan	nan	nan	nan	nan	nan	
runner1_score	nan	nan	nan	nan	nan	nan	nan	nan	
runner1_rbi	nan	nan	nan	nan	nan	nan	nan	nan	
runner1_earned	nan	nan	nan	nan	nan	nan	nan	nan	
runner2_id	nan	nan	nan	nan	nan	nan	nan	nan	
runner2_start	nan	nan	nan	nan	nan	nan	nan	nan	
runner2_end	nan	nan	nan	nan	nan	nan	nan	nan	
runner2_event	nan	nan	nan	nan	nan	nan	nan	nan	
runner2_score	nan	nan	nan	nan	nan	nan	nan	nan	
runner2_rbi	nan	nan	nan	nan	nan	nan	nan	nan	
runner2_earned	nan	nan	nan	nan	nan	nan	nan	nan	
runner3_id	nan	nan	nan	nan	nan	nan	nan	nan	

	uid	game_pk	year	team_id_b	team_id_p	inning	top	at_bat_num	p
runner3_start	nan	nan	nan	nan	nan	nan	nan	nan	
runner3_end	nan	nan	nan	nan	nan	nan	nan	nan	
runner3_event	nan	nan	nan	nan	nan	nan	nan	nan	
runner3_score	nan	nan	nan	nan	nan	nan	nan	nan	
runner3_rbi	nan	nan	nan	nan	nan	nan	nan	nan	
runner3_earned	nan	nan	nan	nan	nan	nan	nan	nan	
runner4_id	nan	nan	nan	nan	nan	nan	nan	nan	
runner4_start	nan	nan	nan	nan	nan	nan	nan	nan	
runner4_end	nan	nan	nan	nan	nan	nan	nan	nan	
runner4_event	nan	nan	nan	nan	nan	nan	nan	nan	
runner4_score	nan	nan	nan	nan	nan	nan	nan	nan	
runner4_rbi	nan	nan	nan	nan	nan	nan	nan	nan	
runner4_earned	nan	nan	nan	nan	nan	nan	nan	nan	
runner5_id	nan	nan	nan	nan	nan	nan	nan	nan	
runner5_start	nan	nan	nan	nan	nan	nan	nan	nan	
runner5_end	nan	nan	nan	nan	nan	nan	nan	nan	
runner5_event	nan	nan	nan	nan	nan	nan	nan	nan	
runner5_score	nan	nan	nan	nan	nan	nan	nan	nan	
runner5_rbi	nan	nan	nan	nan	nan	nan	nan	nan	
runner5_earned	nan	nan	nan	nan	nan	nan	nan	nan	
runner6_id	nan	nan	nan	nan	nan	nan	nan	nan	
runner6_start	nan	nan	nan	nan	nan	nan	nan	nan	
runner6_end	nan	nan	nan	nan	nan	nan	nan	nan	
runner6_event	nan	nan	nan	nan	nan	nan	nan	nan	
runner6_score	nan	nan	nan	nan	nan	nan	nan	nan	
runner6_rbi	nan	nan	nan	nan	nan	nan	nan	nan	
runner6_earned	nan	nan	nan	nan	nan	nan	nan	nan	
runner7_id	nan	nan	nan	nan	nan	nan	nan	nan	
runner7_start	nan	nan	nan	nan	nan	nan	nan	nan	
runner7_end	nan	nan	nan	nan	nan	nan	nan	nan	
runner7_event	nan	nan	nan	nan	nan	nan	nan	nan	
runner7_score	nan	nan	nan	nan	nan	nan	nan	nan	
runner7_rbi	nan	nan	nan	nan	nan	nan	nan	nan	

	uid	game_pk	year	team_id_b	team_id_p	inning	top	at_bat_num	p
runner7_earned	nan	nan	nan	nan	nan	nan	nan	nan	
modified_by	nan	nan	nan	nan	nan	nan	nan	nan	

In [9]:
```
# Dropping nan columns
FF = FF.drop(['year', 'y0'], axis = 1)
FF = FF.drop(FF.loc[:,'runner1_id':'modified_by'].columns, axis = 1)
```

In [10]:
```
# Setting up correlation matrix with remaining variables
corr_FF = FF.corr()
corr_FF.style.background_gradient(cmap='coolwarm')
```

Out[10]:

	uid	game_pk	team_id_b	team_id_p	inning	top	at_bat_num	pcount_
uid	1.000000	0.390780	0.019080	0.019230	-0.007139	-0.000892	0.001742	0.(
game_pk	0.390780	1.000000	0.071178	0.073782	-0.004490	-0.001576	0.000299	0.(
team_id_b	0.019080	0.071178	1.000000	-0.073066	-0.002456	-0.005154	-0.006554	-0.(
team_id_p	0.019230	0.073782	-0.073066	1.000000	0.003914	0.003741	-0.001520	-0.(
inning	-0.007139	-0.004490	-0.002456	0.003914	1.000000	0.040169	0.976341	0.(
top	-0.000892	-0.001576	-0.005154	0.003741	0.040169	1.000000	-0.051980	-0.(
at_bat_num	0.001742	0.000299	-0.006554	-0.001520	0.976341	-0.051980	1.000000	0.(
pcount_at_bat	0.002171	0.000916	-0.001440	-0.003043	0.001987	-0.002008	0.002824	1.(
pcount_pitcher	-0.012253	-0.016561	0.004429	0.010576	0.011561	-0.003248	-0.010417	0.(
balls	-0.005117	-0.002658	0.003328	-0.006731	0.000590	-0.006637	0.002917	0.(
strikes	0.005940	0.003369	-0.004021	0.005034	-0.000823	0.003657	-0.001847	0.(
fouls	0.006703	0.003333	-0.007033	-0.006135	0.012520	0.001433	0.012091	0.(
outs	0.001811	-0.000789	0.001415	0.004695	0.006771	-0.000118	0.055724	0.(
is_final_pitch	-0.000431	0.000122	0.000706	0.000362	0.000162	0.000773	-0.000296	0.(
final_balls	-0.010316	-0.002899	0.007661	-0.009953	0.002893	-0.012730	0.006964	0.(
final_strikes	0.010756	0.005459	-0.008023	0.014904	0.009145	0.010341	0.007473	0.(
final_outs	-0.000249	-0.001402	0.001317	0.009067	0.001277	0.005850	0.042191	-0.(
start_tfs	-0.041532	-0.027894	-0.002767	-0.003658	-0.201844	0.011827	-0.202907	0.(
batter_id	0.066375	-0.003258	0.002937	-0.009212	0.007382	0.003876	0.006802	0.(
pitcher_id	0.027437	-0.011546	-0.003897	-0.115061	0.039526	0.005108	0.041579	0.(
away_team_runs	0.026047	0.013042	-0.016549	-0.014822	0.488752	-0.038259	0.582259	0.(
home_team_runs	0.028451	0.022960	-0.019408	-0.016573	0.478855	-0.015593	0.569420	0.(
pitch_id	0.003006	0.001331	-0.009104	-0.004432	0.964648	-0.050919	0.995346	0.(

	uid	game_pk	team_id_b	team_id_p	inning	top	at_bat_num	pcount_
pitch_tfs	-0.040438	-0.020530	-0.004019	-0.004009	-0.197081	0.011635	-0.197378	0.(
x	-0.010647	-0.004231	0.007697	-0.003205	0.002184	0.001073	0.001569	-0.(
y	-0.009731	0.004193	0.019807	0.030795	-0.014789	-0.006033	-0.014862	0.(
start_speed	0.032115	0.021043	-0.011282	-0.003862	0.060975	0.007507	0.061263	0.(
end_speed	0.054650	0.029973	-0.013044	-0.003899	0.055342	0.007583	0.055790	0.(
sz_top	0.033549	0.019338	-0.021110	0.007204	-0.016973	0.018257	-0.016469	-0.(
sz_bot	0.021399	0.011670	-0.011309	0.003799	-0.012171	-0.001194	-0.009573	0.(
pfx_x	0.010359	0.004256	-0.011228	0.059711	-0.023553	-0.002222	-0.021485	0.(
pfx_z	-0.010501	0.008258	-0.011585	-0.019493	-0.025919	0.006518	-0.025768	-0.(
px	0.011032	0.006416	-0.008223	0.003028	-0.005013	-0.001935	-0.004309	0.(
pz	0.008976	-0.000102	-0.012683	-0.025179	0.012942	0.006758	0.012882	-0.(
x0	0.011112	0.002959	-0.006118	0.086777	-0.052607	-0.010203	-0.051102	0.(
z0	0.017892	0.025436	-0.006460	0.024459	-0.137384	0.003359	-0.140359	-0.(
vx0	-0.005844	-0.001067	0.004899	-0.087047	0.048075	0.007537	0.046556	-0.(
vz0	-0.009883	-0.023671	0.004008	-0.016471	0.052252	-0.003004	0.053018	0.(
vy0	-0.032322	-0.021063	0.011329	0.003127	-0.061138	-0.007689	-0.061357	-0.(
ax	0.008789	0.004565	-0.011317	0.067020	-0.028223	-0.002782	-0.026102	0.(
az	-0.004819	0.013524	-0.011578	-0.017629	-0.016758	0.007082	-0.016473	-0.(
ay	-0.067526	-0.019262	0.000625	0.006032	0.077492	0.004551	0.076771	0.(
break_length	-0.014105	-0.011245	0.016611	0.014910	-0.004461	-0.007634	-0.004803	0.(
break_y	0.124168	0.048114	-0.008871	0.002568	-0.041883	-0.000762	-0.041053	0.(
break_angle	-0.007035	0.000098	0.010295	-0.067054	0.034019	0.005419	0.031633	-0.(
pitch_type	0.014347	0.004832	-0.013564	-0.015156	0.015466	0.009027	0.015863	-0.(
type_confidence	0.020910	0.020328	0.005219	-0.081324	-0.008850	0.011183	-0.015331	0.(
zone	0.000313	0.003654	0.003802	0.011479	0.000009	-0.004409	0.000777	0.(
nasty	-0.002411	-0.001969	-0.000778	0.002520	-0.017771	0.000639	-0.018668	-0.:
spin_dir	-0.006149	-0.004168	0.007052	-0.032804	-0.001996	-0.003261	-0.002730	-0.(
spin_rate	-0.029124	0.013427	0.004885	0.003599	-0.009530	0.002892	-0.010834	-0.(
on_1b	0.065175	-0.005285	-0.007411	-0.013273	0.022826	0.015284	0.018019	-0.(
on_2b	0.061079	-0.013664	-0.017993	-0.005855	0.016400	0.016719	0.014921	0.(
on_3b	0.077527	-0.012314	-0.000785	-0.017172	0.023869	0.013940	0.022160	0.(

In [11]:

```
# Sorting correlation values to pitch_type in order to determine which are the best ind
corr_FF['pitch_type'].sort_values()
```

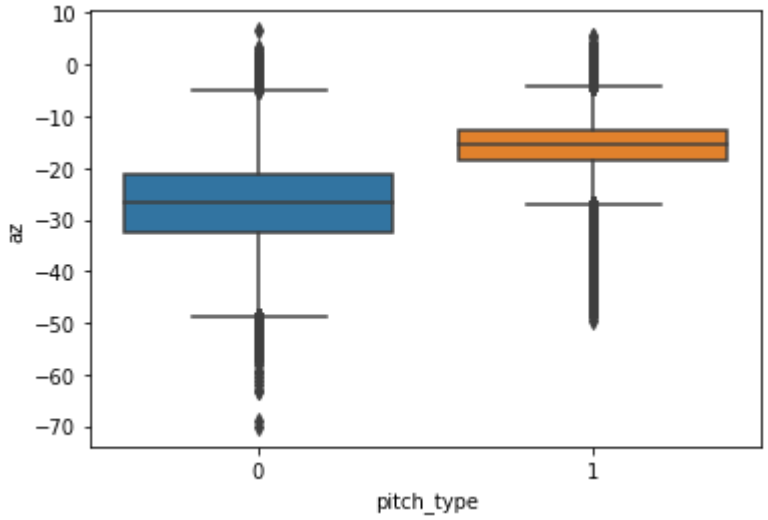
Out[11]:

```
break_length      -0.601570
vy0               -0.507116
vz0               -0.332165
y                 -0.213460
break_y           -0.195955
zone              -0.107303
pfx_x             -0.089909
ax                -0.088784
pcount_pitcher    -0.063531
strikes           -0.047665
final_outs        -0.025064
outs              -0.024273
is_final_pitch    -0.021503
px                -0.019798
fouls             -0.015306
team_id_p         -0.015156
team_id_b         -0.013564
z0                -0.003835
pcount_at_bat     -0.003167
x0                -0.002517
final_strikes     -0.002037
sz_bot             0.000016
type_confidence    0.000174
sz_top             0.000989
pitch_tfs          0.002499
batter_id          0.003063
start_tfs          0.003388
game_pk            0.004832
top                0.009027
on_2b              0.009955
on_3b              0.010135
on_1b              0.012133
home_team_runs     0.012540
uid                0.014347
inning             0.015466
at_bat_num         0.015863
pitch_id           0.020162
x                  0.022118
away_team_runs     0.022430
final_balls        0.026117
vx0                0.036516
balls              0.037757
nasty              0.043473
pitcher_id         0.090184
spin_dir           0.108575
break_angle        0.115282
pz                 0.224894
spin_rate          0.353668
ay                 0.423412
end_speed          0.486442
start_speed        0.506907
pfx_z              0.547131
az                 0.601343
pitch_type         1.000000
Name: pitch_type, dtype: float64
```

Using the boxplot charts below we see that variable az and pfx_z does contribute information towards indicating if a pitch is a fastball (FF) or not as it helps identify the types of pitches that are

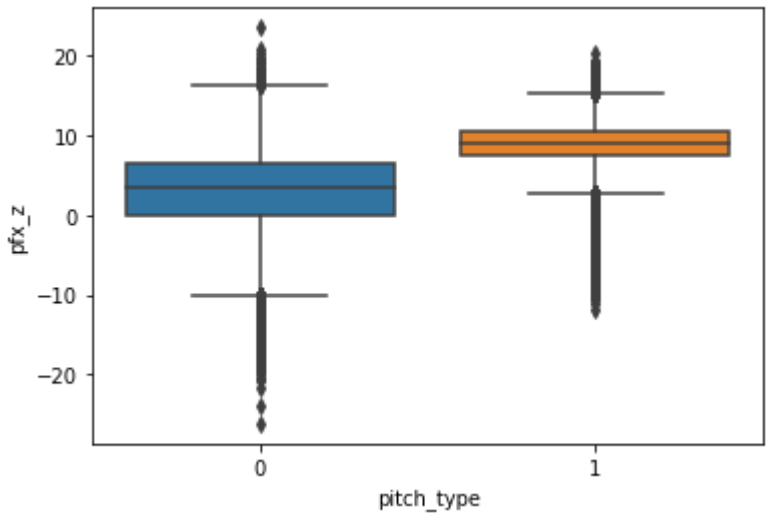
FF since this vairable is positively correlated with it.

In [12]:
```
az = sns.boxplot(x = FF['pitch_type'],
            y = FF['az'])
```



In [13]:
```
sns.boxplot(x = FF['pitch_type'],
            y = FF['pfx_z'])
```

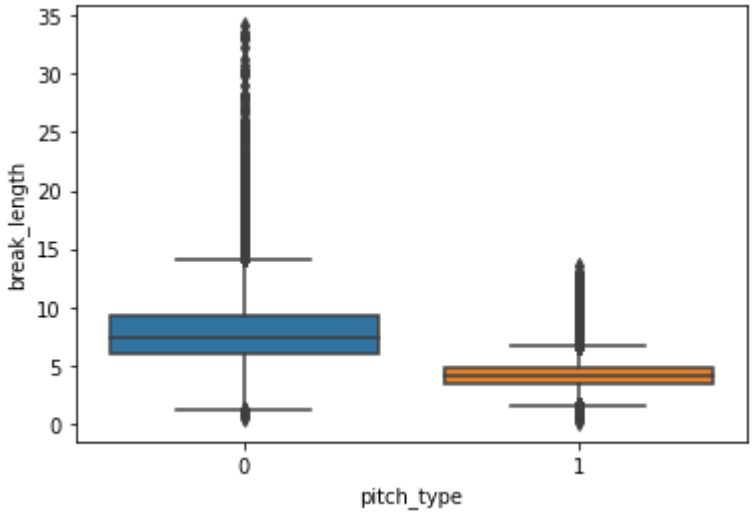
Out[13]:
```
<AxesSubplot:xlabel='pitch_type', ylabel='pfx_z'>
```



The Boxplots below show that variable break_length and vy0 does contribute information towards indicating if a pitch is a fastball (FF) or not as it helps identify the types of pitches that are not FF since this vairable is negatively correlated with it.

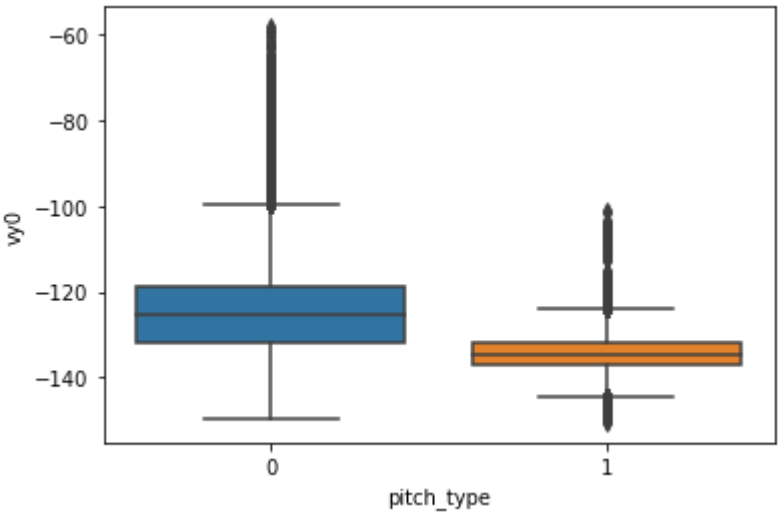
In [14]:
```
sns.boxplot(x = FF['pitch_type'],
            y = FF['break_length'])
```

Out[14]:
```
<AxesSubplot:xlabel='pitch_type', ylabel='break_length'>
```



In [15]:
```
sns.boxplot(x = FF['pitch_type'],
            y = FF['vy0'])
```

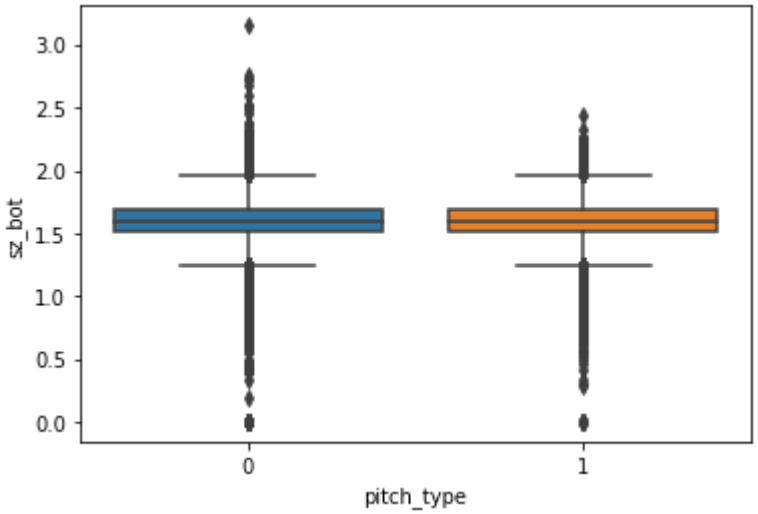
Out[15]:
```
<AxesSubplot:xlabel='pitch_type', ylabel='vy0'>
```



The boxplot below shows that variable sz_bot does not contribute any information towards indicating if a pitch is a fastball or not.

In [16]:
```
sns.boxplot(x = FF['pitch_type'],
            y = FF['sz_bot'])
```

Out[16]:
```
<AxesSubplot:xlabel='pitch_type', ylabel='sz_bot'>
```



The Variables chosen to develop my FF pitch model based on the correlation matrix and boxplots will be the following:

break_length

vy0

vz0

y

break_y

zone

spin_dir

break_angle

pz

spin_rate

ay

end_speed

start_speed

pfx_z

az

In [17]:

```
# Extracting the selected columns from the dataframe
col_list = ['break_length',
            'vy0',
            'vz0',
            'y',
```

```
            'break_y',
            'zone',
            'spin_dir',
            'break_angle',
            'pz',
            'spin_rate',
            'ay',
            'end_speed',
            'start_speed',
            'pfx_z',
            'az',
            'pitch_type']
FF = FF[col_list]
print(FF)
```

```
        break_length      vy0     vz0       y  break_y  zone  spin_dir  \
26               2.8 -127.336  -9.248  163.19     23.9   8.0   183.148
27               1.9 -132.458  -8.133  142.47     23.8  12.0   187.663
28               2.3 -131.189 -10.574  171.83     23.9  14.0   179.643
29               2.1 -132.437  -7.546  138.15     23.8   3.0   184.623
30               0.7 -135.449 -10.658  155.42     23.8   6.0   182.338
...              ...      ...     ...     ...      ...   ...       ...
718956           2.6 -143.374  -6.112  140.74     23.7   2.0   197.537
718957           6.1 -133.947  -5.564  170.10     23.8   9.0   164.636
718958           3.4 -142.543  -7.265  155.42     23.7   4.0   204.835
718959           7.2 -136.340  -4.042  158.01     23.8   8.0   253.646
718960           3.5 -142.847  -3.044  129.52     23.7   1.0   207.372

        break_angle     pz  spin_rate      ay  end_speed  start_speed  pfx_z  \
26             -0.7  1.746   2519.455  22.579       81.4         87.2  13.21
27              6.9  2.666   2838.803  26.928       84.0         90.9  14.34
28            -12.4  1.436   2701.919  24.831       83.8         90.0  13.82
29              0.1  2.814   2683.280  26.271       84.0         90.7  13.60
30            -11.3  2.030   3352.205  27.663       85.9         92.9  16.68
...             ...    ...        ...     ...        ...          ...    ...
718956         26.6  2.689   2312.186  36.837       89.1         97.9  10.56
718957         -4.7  1.446    697.763  29.860       84.3         91.4   3.39
718958         28.7  2.122   2162.620  38.935       88.0         97.3   9.53
718959         33.8  2.053   2180.650  32.209       85.3         93.1   3.05
718960         28.4  3.704   1996.857  38.276       88.3         97.4   8.56

            az  pitch_type
26     -10.094           1
27      -6.487           1
28      -7.742           1
29      -7.759           1
30      -0.903           1
...        ...         ...
718956 -10.386           1
718957 -25.962           0
718958 -12.891           1
718959 -26.404           0
718960 -14.714           1

[716681 rows x 16 columns]
```

In [18]:

```
# Preparing the data for model
feature_cols = ['break_length',
            'vy0',
```

```
            'vz0',
            'y',
            'break_y',
            'zone',
            'spin_dir',
            'break_angle',
            'pz',
            'spin_rate',
            'ay',
            'end_speed',
            'start_speed',
            'pfx_z',
            'az']
X = FF[feature_cols] # Features
y = FF.pitch_type # Target Variable
```

In [19]:
```
# Splitting the data into Training and Testing having 75% of the data to be used for tr
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.25,random_state=0)
```

In [20]:
```
# Initializing Logistic Regression model
logreg = LogisticRegression(max_iter=1000)
# Fitting the model with the data
logreg.fit(X_train,y_train)
```

Out[20]:
```
LogisticRegression(max_iter=1000)
```

In [21]:
```
# Making Predictions on test set
y_pred=logreg.predict(X_test)
# Predicting Probabilities on test set
Fastpred = logreg.predict_proba(X_test)
print(Fastpred)
```

```
[[9.99999999e-01 9.13516655e-10]
 [8.84569102e-01 1.15430898e-01]
 [9.99229194e-01 7.70805944e-04]
 ...
 [9.25730456e-03 9.90742695e-01]
 [5.58641088e-01 4.41358912e-01]
 [3.25884895e-02 9.67411510e-01]]
```

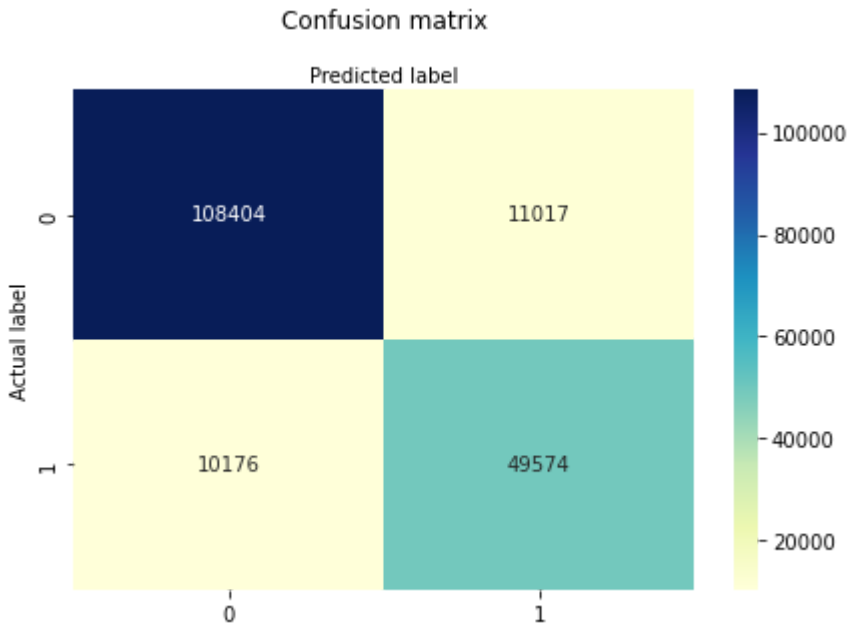
In [22]:
```
# Building Confusion Matrix based on results
cnf_matrix = metrics.confusion_matrix(y_test, y_pred)
cnf_matrix
```

Out[22]:
```
array([[108404,  11017],
       [ 10176,  49574]], dtype=int64)
```

In [23]:
```
# Formatting Confusion Matrix
class_names=[0,1]
fig, ax = plt.subplots()
tick_marks = np.arange(len(class_names))
plt.xticks(tick_marks, class_names)
plt.yticks(tick_marks, class_names)
sns.heatmap(pd.DataFrame(cnf_matrix), annot=True, cmap="YlGnBu" ,fmt='g')
```

```
ax.xaxis.set_label_position("top")
plt.tight_layout()
plt.title('Confusion matrix', y=1.1)
plt.ylabel('Actual label')
plt.xlabel('Predicted label')
```

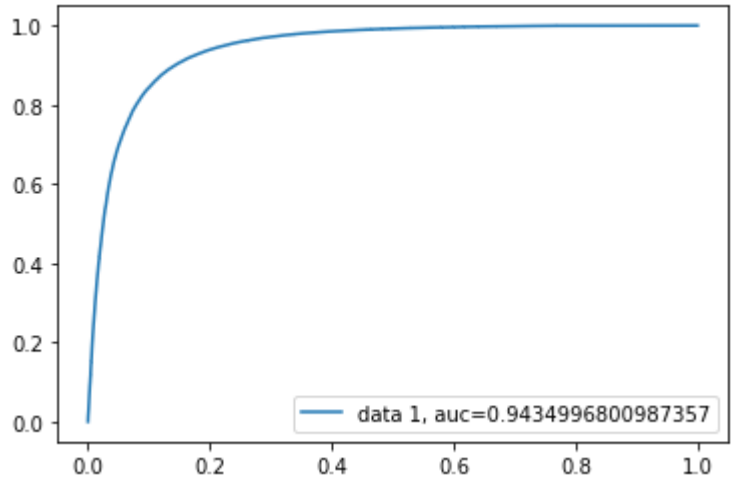
Out[23]:
```
Text(0.5, 257.44, 'Predicted label')
```



In [24]:
```
# Printing Evaluation Metrics
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
print("Precision:",metrics.precision_score(y_test, y_pred))
print("Recall:",metrics.recall_score(y_test, y_pred))
```

```
Accuracy: 0.8817163491859732
Precision: 0.818174316317605
Recall: 0.8296903765690377
```

In [25]:
```
# Creating the ROC Curve and getting the AUC
y_pred_proba = logreg.predict_proba(X_test)[::,1]
fpr, tpr, _ = metrics.roc_curve(y_test,  y_pred_proba)
auc = metrics.roc_auc_score(y_test, y_pred_proba)
plt.plot(fpr,tpr,label="data 1, auc="+str(auc))
plt.legend(loc=4)
plt.show()
```



Setting up Slider Model

In [26]:
```
# Cleaning Slider data
SL = SL.drop(['year', 'y0'], axis = 1)
SL = SL.drop(SL.loc[:,'runner1_id':'modified_by'].columns, axis = 1)
```

In [27]:
```
# Setting up correlation matrix to see what variables show correlation with the pitch t
corr_SL = SL.corr()
corr_SL.style.background_gradient(cmap='coolwarm')
```

Out[27]:

	uid	game_pk	team_id_b	team_id_p	inning	top	at_bat_num	pcount_
uid	1.000000	0.390780	0.019080	0.019230	-0.007139	-0.000892	0.001742	0.0
game_pk	0.390780	1.000000	0.071178	0.073782	-0.004490	-0.001576	0.000299	0.0
team_id_b	0.019080	0.071178	1.000000	-0.073066	-0.002456	-0.005154	-0.006554	-0.0
team_id_p	0.019230	0.073782	-0.073066	1.000000	0.003914	0.003741	-0.001520	-0.0
inning	-0.007139	-0.004490	-0.002456	0.003914	1.000000	0.040169	0.976341	0.0
top	-0.000892	-0.001576	-0.005154	0.003741	0.040169	1.000000	-0.051980	-0.0
at_bat_num	0.001742	0.000299	-0.006554	-0.001520	0.976341	-0.051980	1.000000	0.0
pcount_at_bat	0.002171	0.000916	-0.001440	-0.003043	0.001987	-0.002008	0.002824	1.0
pcount_pitcher	-0.012253	-0.016561	0.004429	0.010576	0.011561	-0.003248	-0.010417	0.0
balls	-0.005117	-0.002658	0.003328	-0.006731	0.000590	-0.006637	0.002917	0.8
strikes	0.005940	0.003369	-0.004021	0.005034	-0.000823	0.003657	-0.001847	0.7
fouls	0.006703	0.003333	-0.007033	-0.006135	0.012520	0.001433	0.012091	0.6
outs	0.001811	-0.000789	0.001415	0.004695	0.006771	-0.000118	0.055724	0.0
is_final_pitch	-0.000431	0.000122	0.000706	0.000362	0.000162	0.000773	-0.000296	0.3
final_balls	-0.010316	-0.002899	0.007661	-0.009953	0.002893	-0.012730	0.006964	0.3
final_strikes	0.010756	0.005459	-0.008023	0.014904	0.009145	0.010341	0.007473	0.2
final_outs	-0.000249	-0.001402	0.001317	0.009067	0.001277	0.005850	0.042191	-0.0

	uid	game_pk	team_id_b	team_id_p	inning	top	at_bat_num	pcount_
start_tfs	-0.041532	-0.027894	-0.002767	-0.003658	-0.201844	0.011827	-0.202907	0.(
batter_id	0.066375	-0.003258	0.002937	-0.009212	0.007382	0.003876	0.006802	0.(
pitcher_id	0.027437	-0.011546	-0.003897	-0.115061	0.039526	0.005108	0.041579	0.(
away_team_runs	0.026047	0.013042	-0.016549	-0.014822	0.488752	-0.038259	0.582259	0.(
home_team_runs	0.028451	0.022960	-0.019408	-0.016573	0.478855	-0.015593	0.569420	0.(
pitch_id	0.003006	0.001331	-0.009104	-0.004432	0.964648	-0.050919	0.995346	0.(
pitch_tfs	-0.040438	-0.020530	-0.004019	-0.004009	-0.197081	0.011635	-0.197378	0.(
x	-0.010647	-0.004231	0.007697	-0.003205	0.002184	0.001073	0.001569	-0.(
y	-0.009731	0.004193	0.019807	0.030795	-0.014789	-0.006033	-0.014862	0.(
start_speed	0.032115	0.021043	-0.011282	-0.003862	0.060975	0.007507	0.061263	0.(
end_speed	0.054650	0.029973	-0.013044	-0.003899	0.055342	0.007583	0.055790	0.(
sz_top	0.033549	0.019338	-0.021110	0.007204	-0.016973	0.018257	-0.016469	-0.(
sz_bot	0.021399	0.011670	-0.011309	0.003799	-0.012171	-0.001194	-0.009573	0.(
pfx_x	0.010359	0.004256	-0.011228	0.059711	-0.023553	-0.002222	-0.021485	0.(
pfx_z	-0.010501	0.008258	-0.011585	-0.019493	-0.025919	0.006518	-0.025768	-0.(
px	0.011032	0.006416	-0.008223	0.003028	-0.005013	-0.001935	-0.004309	0.(
pz	0.008976	-0.000102	-0.012683	-0.025179	0.012942	0.006758	0.012882	-0.(
x0	0.011112	0.002959	-0.006118	0.086777	-0.052607	-0.010203	-0.051102	0.(
z0	0.017892	0.025436	-0.006460	0.024459	-0.137384	0.003359	-0.140359	-0.(
vx0	-0.005844	-0.001067	0.004899	-0.087047	0.048075	0.007537	0.046556	-0.(
vz0	-0.009883	-0.023671	0.004008	-0.016471	0.052252	-0.003004	0.053018	0.(
vy0	-0.032322	-0.021063	0.011329	0.003127	-0.061138	-0.007689	-0.061357	-0.(
ax	0.008789	0.004565	-0.011317	0.067020	-0.028223	-0.002782	-0.026102	0.(
az	-0.004819	0.013524	-0.011578	-0.017629	-0.016758	0.007082	-0.016473	-0.(
ay	-0.067526	-0.019262	0.000625	0.006032	0.077492	0.004551	0.076771	0.(
break_length	-0.014105	-0.011245	0.016611	0.014910	-0.004461	-0.007634	-0.004803	0.(
break_y	0.124168	0.048114	-0.008871	0.002568	-0.041883	-0.000762	-0.041053	0.(
break_angle	-0.007035	0.000098	0.010295	-0.067054	0.034019	0.005419	0.031633	-0.(
pitch_type	0.003359	-0.005580	-0.007499	-0.008639	0.065466	0.000610	0.068363	0.(
type_confidence	0.020910	0.020328	0.005219	-0.081324	-0.008850	0.011183	-0.015331	0.(
zone	0.000313	0.003654	0.003802	0.011479	0.000009	-0.004409	0.000777	0.(
nasty	-0.002411	-0.001969	-0.000778	0.002520	-0.017771	0.000639	-0.018668	-0.:
spin_dir	-0.006149	-0.004168	0.007052	-0.032804	-0.001996	-0.003261	-0.002730	-0.(

	uid	game_pk	team_id_b	team_id_p	inning	top	at_bat_num	pcount_
spin_rate	-0.029124	0.013427	0.004885	0.003599	-0.009530	0.002892	-0.010834	-0.(
on_1b	0.065175	-0.005285	-0.007411	-0.013273	0.022826	0.015284	0.018019	-0.(
on_2b	0.061079	-0.013664	-0.017993	-0.005855	0.016400	0.016719	0.014921	0.(
on_3b	0.077527	-0.012314	-0.000785	-0.017172	0.023869	0.013940	0.022160	0.(

In [28]:
```
# Sorting correlation values to pitch_type in order to determine which are the best ind
corr_SL['pitch_type'].sort_values()
```

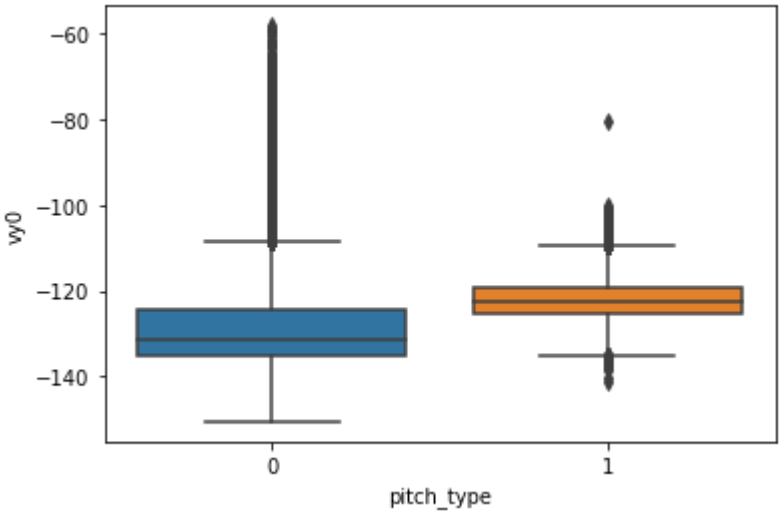
Out[28]:
```
spin_rate         -0.575832
az                -0.345496
pfx_z             -0.308124
ay                -0.299677
start_speed       -0.275061
end_speed         -0.241479
spin_dir          -0.234409
break_angle       -0.178478
pz                -0.148398
x                 -0.093780
nasty             -0.068472
x0                -0.034679
final_balls       -0.029389
balls             -0.023004
pcount_pitcher    -0.021755
start_tfs         -0.021115
pitch_tfs         -0.020987
vx0               -0.011472
team_id_p         -0.008639
z0                -0.008424
sz_top            -0.007842
team_id_b         -0.007499
game_pk           -0.005580
sz_bot            -0.005196
top                0.000610
on_2b              0.000614
batter_id          0.002103
uid                0.003359
on_1b              0.004119
pitcher_id         0.010732
is_final_pitch     0.013296
on_3b              0.014004
outs               0.027245
final_outs         0.036889
home_team_runs     0.038607
away_team_runs     0.041244
pcount_at_bat      0.043118
fouls              0.051077
zone               0.062542
inning             0.065466
type_confidence    0.066555
at_bat_num         0.068363
pitch_id           0.068509
final_strikes      0.077006
px                 0.091226
```

```
strikes           0.099650
y                 0.141859
vz0               0.165498
pfx_x             0.186211
ax                0.187365
break_y           0.240606
break_length      0.259531
vy0               0.273558
pitch_type        1.000000
Name: pitch_type, dtype: float64
```

Using the boxplot charts below we see that variable vy0 and break_length does contribute information towards indicating if a pitch is a slider (SL) or not as it helps identify the types of pitches that are SL since this vairable is positively correlated with it.

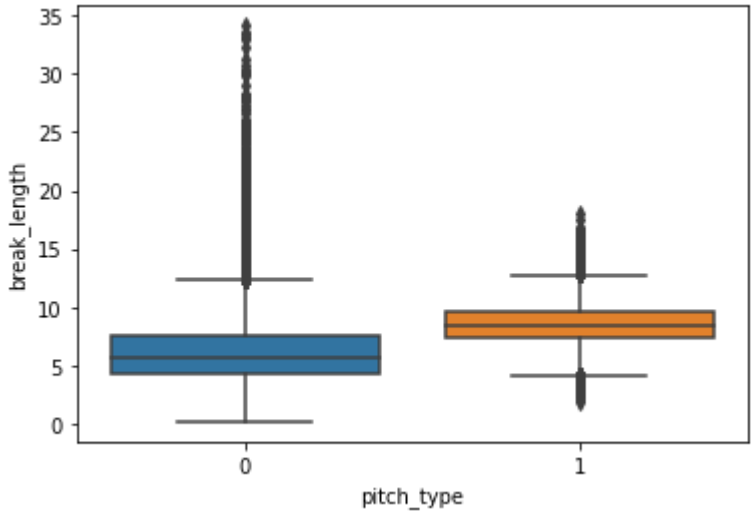
In [29]:
```
sns.boxplot(x = SL['pitch_type'],
            y = SL['vy0'])
```

Out[29]:
```
<AxesSubplot:xlabel='pitch_type', ylabel='vy0'>
```



In [30]:
```
sns.boxplot(x = SL['pitch_type'],
            y = SL['break_length'])
```

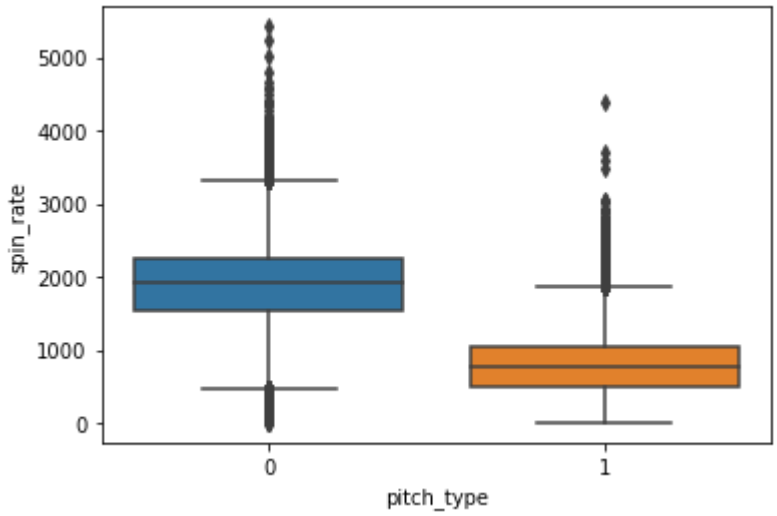
Out[30]:
```
<AxesSubplot:xlabel='pitch_type', ylabel='break_length'>
```



The Boxplots below show that variable spin_rate and az does contribute information towards indicating if a pitch is a slider (SL) or not as it helps identify the types of pitches that are not SL since this vairable is negatively correlated with it.

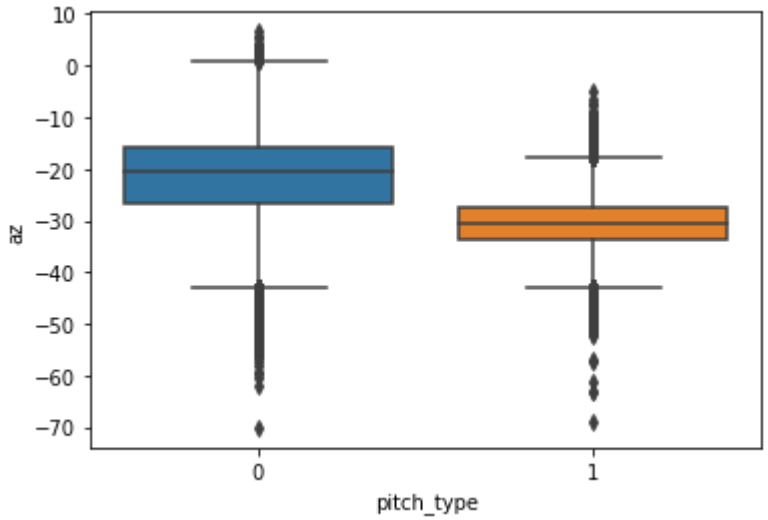
In [31]:
```
sns.boxplot(x = SL['pitch_type'],
            y = SL['spin_rate'])
```

Out[31]:
```
<AxesSubplot:xlabel='pitch_type', ylabel='spin_rate'>
```



In [32]:
```
sns.boxplot(x = SL['pitch_type'],
            y = SL['az'])
```

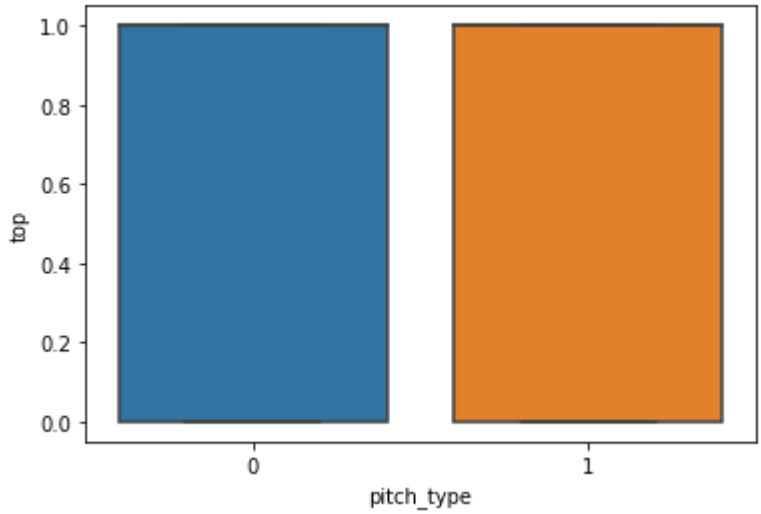
Out[32]:
```
<AxesSubplot:xlabel='pitch_type', ylabel='az'>
```



The boxplot below shows that variable top does not contribute any information towards indicating if a pitch is a slider or not.

In [33]:
```
sns.boxplot(x = SL['pitch_type'],
            y = SL['top'])
```

Out[33]:
```
<AxesSubplot:xlabel='pitch_type', ylabel='top'>
```



The Variables chosen to develop my slider pitch model will be the following:

spin_rate

az

pfx_z

ay

start_speed

end_speed

spin_dir

break_angle

pz

x

px

strikes

y

vz0

pfx_x

ax

break_y

break_length

vy0

In [34]:
```
# Extracting the selected columns from the dataframe
col_list = ['spin_rate',
            'az',
            'pfx_z',
            'ay',
            'start_speed',
            'end_speed',
            'spin_dir',
            'break_angle',
            'pz',
            'x',
            'px',
            'strikes',
            'y',
            'vz0',
            'pfx_x',
            'ax',
            'break_y',
            'break_length',
            'vy0',
            'pitch_type']
SL = SL[col_list]
print(SL)
```

```
        spin_rate      az  pfx_z      ay  start_speed  end_speed  spin_dir  \
26       2519.455 -10.094  13.21  22.579         87.2       81.4   183.148
27       2838.803  -6.487  14.34  26.928         90.9       84.0   187.663
28       2701.919  -7.742  13.82  24.831         90.0       83.8   179.643
29       2683.280  -7.759  13.60  26.271         90.7       84.0   184.623
30       3352.205  -0.903  16.68  27.663         92.9       85.9   182.338
...           ...     ...    ...     ...          ...        ...       ...
718956   2312.186 -10.386  10.56  36.837         97.9       89.1   197.537
718957    697.763 -25.962   3.39  29.860         91.4       84.3   164.636
718958   2162.620 -12.891   9.53  38.935         97.3       88.0   204.835
718959   2180.650 -26.404   3.05  32.209         93.1       85.3   253.646
718960   1996.857 -14.714   8.56  38.276         97.4       88.3   207.372

        break_angle     pz       x     px  strikes       y     vz0  pfx_x  \
26             -0.7  1.746  104.72 -0.081        0  163.19  -9.248  -0.73
27              6.9  2.666   51.50  1.489        1  142.47  -8.133  -1.94
28            -12.4  1.436   62.66  1.160        1  171.83 -10.574   0.09
29              0.1  2.814   82.40  0.542        1  138.15  -7.546  -1.10
30            -11.3  2.030   93.56  0.258        2  155.42 -10.658  -0.68
...             ...    ...     ...    ...      ...     ...     ...    ...
718956         26.6  2.689  102.15 -0.069        1  140.74  -6.112  -3.35
718957         -4.7  1.446   90.99  0.275        2  170.10  -5.564   0.94
718958         28.7  2.122  109.01 -0.261        2  155.42  -7.265  -4.43
718959         33.8  2.053  105.58 -0.136        0  158.01  -4.042 -10.53
718960         28.4  3.704  102.15 -0.561        1  129.52  -3.044  -4.45

           ax  break_y  break_length      vy0  pitch_type
26     -1.215     23.9           2.8 -127.336           0
27     -3.457     23.8           1.9 -132.458           0
28      0.153     23.9           2.3 -131.189           0
29     -1.975     23.8           2.1 -132.437           0
30     -1.278     23.8           0.7 -135.449           0
...       ...      ...           ...      ...         ...
```

```
718956  -6.886     23.7          2.6 -143.374          0
718957   1.707     23.8          6.1 -133.947          0
718958  -8.925     23.7          3.4 -142.543          0
718959 -19.664     23.8          7.2 -136.340          0
718960  -9.040     23.7          3.5 -142.847          0

[716681 rows x 20 columns]
```

In [35]:
```
# Preparing the data for model
feature_cols = ['spin_rate',
                'az',
                'pfx_z',
                'ay',
                'start_speed',
                'end_speed',
                'spin_dir',
                'break_angle',
                'pz',
                'x',
                'px',
                'strikes',
                'y',
                'vz0',
                'pfx_x',
                'ax',
                'break_y',
                'break_length',
                'vy0']
Xsl = SL[feature_cols] # Features
ysl = SL.pitch_type # Target Variable
```

In [36]:
```
# Splitting the data into Training and Testing having 75% of the data to be used for tr
X_train,X_test,y_train,y_test=train_test_split(Xsl,ysl,test_size=0.25,random_state=0)
```

In [37]:
```
# Initializing Logistic Regression Model
logreg = LogisticRegression(max_iter=5000)
# Fitting the model with the data
logreg.fit(X_train,y_train)
```

Out[37]:
```
LogisticRegression(max_iter=5000)
```

In [38]:
```
# Making Predictions on test set
y_pred=logreg.predict(X_test)
# Predicting Probabilities on test set
Slidpred = logreg.predict_proba(X_test)
print(Slidpred)
```

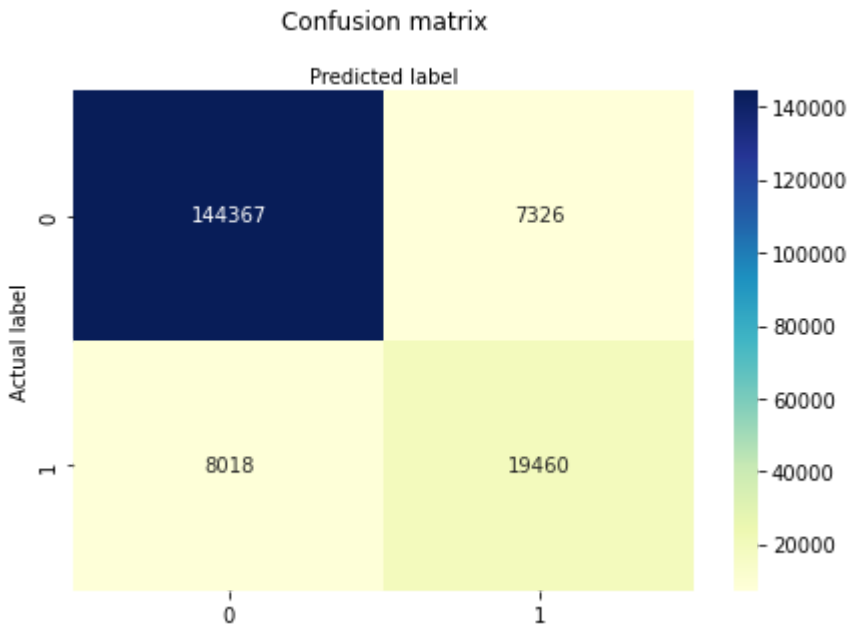
```
[[9.97527854e-01 2.47214635e-03]
 [9.99999910e-01 9.03727746e-08]
 [9.74175468e-01 2.58245320e-02]
 ...
 [9.99883553e-01 1.16447071e-04]
 [9.99920389e-01 7.96108879e-05]
 [9.90219884e-01 9.78011618e-03]]
```

In [39]:
```
# Building Confusion Matrix based on results
cnf_matrix = metrics.confusion_matrix(y_test, y_pred)
cnf_matrix
```

Out[39]:
```
array([[144367,   7326],
       [  8018,  19460]], dtype=int64)
```

In [40]:
```
# Formating Confusion Matrix
class_names=[0,1]
fig, ax = plt.subplots()
tick_marks = np.arange(len(class_names))
plt.xticks(tick_marks, class_names)
plt.yticks(tick_marks, class_names)
sns.heatmap(pd.DataFrame(cnf_matrix), annot=True, cmap="YlGnBu" ,fmt='g')
ax.xaxis.set_label_position("top")
plt.tight_layout()
plt.title('Confusion matrix', y=1.1)
plt.ylabel('Actual label')
plt.xlabel('Predicted label')
```

Out[40]:
```
Text(0.5, 257.44, 'Predicted label')
```

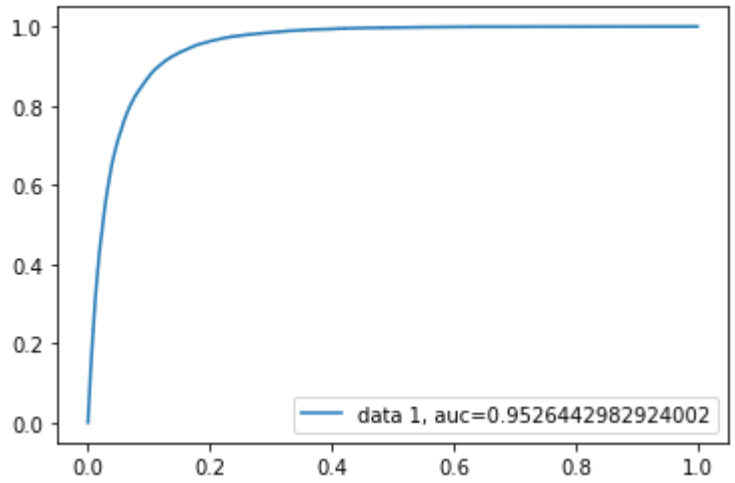


In [41]:
```
# Printing Evaluation Metrics
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
print("Precision:",metrics.precision_score(y_test, y_pred))
print("Recall:",metrics.recall_score(y_test, y_pred))
```

```
Accuracy: 0.914361141032868
Precision: 0.7264989173448817
Recall: 0.7082029259771454
```

In [42]:
```
# Creating the ROC Curve and getting the AUC
y_pred_proba = logreg.predict_proba(X_test)[::,1]
fpr, tpr, _ = metrics.roc_curve(y_test,  y_pred_proba)
auc = metrics.roc_auc_score(y_test, y_pred_proba)
plt.plot(fpr,tpr,label="data 1, auc="+str(auc))
```

```
plt.legend(loc=4)
plt.show()
```



Future Work

The future steps I would take with this project would be implement 10-fold cross validation to reduce the bias in the predictions, do some more investigation into wheter scaling the data beforehand would help in the models performance, and use other models such as KNN, Penalized Logistic Regression, Naïve Bayes, Random Forest, Boosted Trees, LDA, QDA, and SVM to see which model is best at predicting the probability of a given pitch. Another future step would be to see if any other information can be extracted from the dataframe and getting a better understanding at what the data is telling us. Also, with exploring new models I would also look at the tuning parameters for each of them and find the best ones to use in order to optimize the performance of the model. In order to measure the success of these models I would use metrics such as accuracy, AUROC, TPR, FPR, and Precision and use all these metrics to make a decision on which one works best. With the logistic regression model performance in I am optimistic that in implementing these steps the perfomance of these models can be improved.