In [819…
```
#Toronto Blue Jays Right Handed Pitcher Predictions

import pandas as pd
import seaborn as sns
import numpy as np
from scipy import stats
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
```

In [820…
```
df= pd.read_csv('/Users/idelsontaveras/Downloads/Toronto Blue Jays Research Intern Technical Exercise/deploy.csv')
```

In [821…
```
df.head(5)
```

Out[821…

	Velo	SpinRate	HorzBreak	InducedVertBreak
0	94.72	2375.0	3.10	18.15
1	95.25	2033.0	11.26	14.50
2	92.61	2389.0	11.00	21.93
3	94.94	2360.0	6.84	18.11
4	97.42	2214.0	16.70	13.38

In [822…
```
df.shape
```

Out[822…
```
(10000, 4)
```

In [823…
```
df.describe()
```

Out[823…

	Velo	SpinRate	HorzBreak	InducedVertBreak
count	10000.000000	9987.000000	10000.000000	10000.000000
mean	93.924134	2240.582958	9.501652	14.162905
std	2.608429	195.113163	5.032214	4.593760
min	56.740000	1114.000000	-6.700000	-9.280000
25%	92.510000	2111.000000	5.657500	11.310000
50%	94.030000	2240.000000	9.370000	15.195000
75%	95.600000	2368.000000	13.580000	17.590000
max	102.500000	3357.000000	23.480000	25.640000

In [824…
```
df.isnull().sum()
```

Out[824…
```
Velo               0
SpinRate          13
HorzBreak          0
InducedVertBreak   0
dtype: int64
```

In [825…
```
df['SpinRate'] = df['SpinRate'].fillna(df['SpinRate'].mean())
```

In [826…
```
df.isnull().sum()
```

Out[826…
```
Velo               0
SpinRate           0
HorzBreak          0
InducedVertBreak   0
dtype: int64
```

In [827…
```
np.random.seed(42)
```

In [828…
```
InPlayRand = np.random.randint(2, size=len(df))
```

In [829…
```
df['InPlayRand'] = InPlay
```

In [830…
```
print(df.head())
```
```
    Velo  SpinRate  HorzBreak  InducedVertBreak  InPlayRand
0  94.72    2375.0       3.10             18.15           0
1  95.25    2033.0      11.26             14.50           1
2  92.61    2389.0      11.00             21.93           0
3  94.94    2360.0       6.84             18.11           0
4  97.42    2214.0      16.70             13.38           0
```

In [831…
```
df.corr()['Velo']
```

Out[831…
```
Velo               1.000000
SpinRate           0.323363
HorzBreak          0.003784
InducedVertBreak   0.084603
InPlayRand         0.009573
Name: Velo, dtype: float64
```

In [832…
```
df.corr()['SpinRate']
```

Out[832…
```
Velo               0.323363
SpinRate           1.000000
HorzBreak         -0.264320
InducedVertBreak   0.383966
InPlayRand        -0.012671
Name: SpinRate, dtype: float64
```

In [833…
```
df.corr()['HorzBreak']
```

Out[833…
```
Velo               0.003784
SpinRate          -0.264320
HorzBreak          1.000000
InducedVertBreak  -0.578438
InPlayRand        -0.006832
Name: HorzBreak, dtype: float64
```

In [834…
```
df.corr()['InducedVertBreak']
```

Out[834…
```
Velo               0.084603
SpinRate           0.383966
HorzBreak         -0.578438
InducedVertBreak   1.000000
InPlayRand        -0.000936
Name: InducedVertBreak, dtype: float64
```

In [835…
```
SpinRate = df['SpinRate']
df=df.drop(['HorzBreak'], axis=1)
df['SpinRate']= SpinRate
df
```

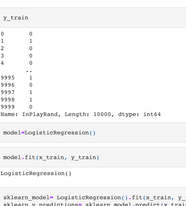
Out[835…

	Velo	SpinRate	InducedVertBreak	InPlayRand
0	94.72	2375.0	18.15	0
1	95.25	2033.0	14.50	1
2	92.61	2389.0	21.93	0
3	94.94	2360.0	18.11	0
4	97.42	2214.0	13.38	0
...	...	...	...	...
9995	92.32	2148.0	16.70	1
9996	94.96	2420.0	14.13	0
9997	92.83	2132.0	18.40	1
9998	97.12	2436.0	15.87	1
9999	96.00	2350.0	18.22	0

10000 rows × 4 columns

In [836…
```
x = df['Velo']
y = df['SpinRate']
coefficients = np.polyfit(x, y, 1)
m = coefficients[0]
b = coefficients[1]
y_pred = m * x + b
plt.scatter(x, y, label='Data Points')
plt.plot(x, y_pred, color='red', label='Trend Line')
plt.xlabel('Velo')
plt.ylabel('SpinRate')
plt.legend()

plt.show()
```

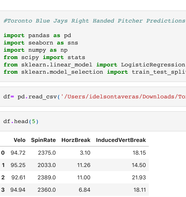


In [837…
```
x = df['Velo']
y = df['InducedVertBreak']
coefficients = np.polyfit(x, y, 1)
m = coefficients[0]
b = coefficients[1]
y_pred = m * x + b
plt.scatter(x, y, label='Data Points')
plt.plot(x, y_pred, color='red', label='Trend Line')
plt.xlabel('Velo')
plt.ylabel('InducedVertBreak')
plt.legend()

plt.show()
```


In [838…
```
x = df['SpinRate']
y = df['InducedVertBreak']
coefficients = np.polyfit(x, y, 1)
m = coefficients[0]
b = coefficients[1]
y_pred = m * x + b
plt.scatter(x, y, label='Data Points')
plt.plot(x, y_pred, color='red', label='Trend Line')
plt.xlabel('SpinRate')
plt.ylabel('InducedVertBreak')
plt.legend()

plt.show()
```



In [839…
```
x_train=df[['Velo','SpinRate','InducedVertBreak']]
y_train= df['InPlayRand']
x_train.shape
```

Out[839…
```
(10000, 3)
```

In [840…
```
x_train
```

Out[840…

	Velo	SpinRate	InducedVertBreak
0	94.72	2375.0	18.15
1	95.25	2033.0	14.50
2	92.61	2389.0	21.93
3	94.94	2360.0	18.11
4	97.42	2214.0	13.38
...	...	...	...
9995	92.32	2148.0	16.70
9996	94.96	2420.0	14.13
9997	92.83	2132.0	18.40
9998	97.12	2436.0	15.87
9999	96.00	2350.0	18.22

10000 rows × 3 columns

In [841…
```
y_train
```

Out[841…
```
0       0
1       1
2       0
3       0
4       0
       ..
9995    1
9996    0
9997    1
9998    1
9999    0
Name: InPlayRand, Length: 10000, dtype: int64
```

In [842…
```
model=LogisticRegression()
```

In [843…
```
model.fit(x_train, y_train)
```

Out[843…
```
LogisticRegression()
```

In [844…
```
sklearn_model= LogisticRegression().fit(x_train, y_train)
sklearn_y_predictions= sklearn_model.predict(x_train)
sklearn_y_predictions
```

Out[844…
```
array([0, 1, 0, ..., 1, 0, 0])
```

In [845…
```
predictions = model.predict(x_train)
```

In [846…
```
df['InPlayPredict'] = predictions
```

In [847…
```
print(df.tail())
```
```
      Velo  SpinRate  InducedVertBreak  InPlayRand  InPlayPredict
9995  92.32    2148.0             16.70           1              1
9996  94.96    2420.0             14.13           0              0
9997  92.83    2132.0             18.40           1              1
9998  97.12    2436.0             15.87           1              0
9999  96.00    2350.0             18.22           0              0
```

In [848…
```
df.to_csv('/Users/idelsontaveras/Downloads/Toronto Blue Jays Research Intern Technical Exercise/Predictions.csv', index=False)
```

In [ ]: