This project is to make a model for xWhiff and in turn xWhiff+ from MLB pitch data for each kind of pitch, fastballs, offspeed, and breaking. xWhiff is the probability of a whiff, and xWhiff+ is a comparison statistic where 100 is league average, and value above or below 100 is the percentage better or worse than average. The dataset is of 25000 pitches that resulted in swings from the MLB for each pitch type from 2025, 2024, and 2023, all sourced from BaseballSavant.com . I cleaned the data in python beforehand using python, skimming it down to the rows I needed, as release speed, release position on the z axis, spin rate, extension, IVB, HB, and the dependent variable, whiffs, where whiff is either a 1 for a whiff or 0 for no whiff. I then downloaded these as separate CSV files as Breaking_Final, FF_Final, and Offspeed_Final. These were loaded into rust using the read_csv method of the DataFrame struct I recycled from my homework 8 assignment. That method takes in the path and a vector of types of columns as an input, then reads the header for labels and iterates over the row and parses it into the corresponding column value enum variant based on the provided type.

- **enum ColumnVal:** Represents the different types of data that can be stored within a DF struct DataFrame: Represents a table structure similar to python, outputs new df instance
- **read_csv** : given reference to the df to be populated, a path to a csv file, and the types of each column, populates new df by reading headers and iterating through rows corresponding to the csv.
- **restrict_columns:** given a reference to the original df and a slice of a string representing the column labels to keep creates a new df with only the columns passed.
- **add_xwoba_plus_column**: given a reference to a df, the label of the new column and the league average float for xwhiff writes a new column of xwhiff to the dataframe.
- **struct PitchRankData** : struct to hold info about each pitch for ranking purposes later for top/bottom 5 pitches. Holds values for pitch_id, pitch_name, xwhiff, and xwhiff_plus
- **standardize_features** : passing in a slice of vectors representing the feature, it standardizes a set of features to have no mean or variance between them by finding the mean and SD of each feature and then modifying it by applying a standardization formula. Outputs a vector of tuples where each tuple has the new mean and SF for the feature.
- **sigmoid** : Takes in a float z and applies the sigmoid activation function to give the sigmoid value, output as another float.
- **predict_probability** : Given features, weights, and a bias, calculates the linear combination of the features and weights, adds the bias, and then applies the sigmoid function to return a probability of a whiff.
- **gradient**: Given features, predictions from training samples, and the targets (either whiff or contact), iterates through data, calculates the error between the predictions and targets and then adds the gradients for each weight and the corresponding bias. Outputs a tuple with the gradient of the weights and the gradient of the bias.
- **train_logistic_regression**: Given features, targets, a learning rate for the gradient descent, and the number of epochs, iteratively updates the weights and bias at the learning rate for the number of epochs, and returns a tuple with the new weights and bias. I chose 2000 epochs because it is around where the error plateaus.

- **calculate_xwhiff_and_write** : given a df, the output path, the new weights and bias, and the feature stats, it will go through the dataframe and predict the xwhiff for each pitch and calculates the xwhiff+ by taking the xwhiff value, dividing it by the league average and multiplying it by 100, and then append that as a column
- **train_and_predict**: Given the df, output file path, and column types it needs, it will prepare the data by standardizing and getting columns ready and then run the regression model and write the new csv by calling calculate_xwhiff_and_write.
- **analyze_top_bottom_pitches**: Given the path to the csv and the type of pitch, it will convert the csv into a new df, finds the columns needed, and then prints the top and bottom 5 xwhiff+ values for that kind of pitch.
- The main function then creates a DF for each CSV file, running train and predict for each function, which creates new CSV files for each one with xWhiff and xWhiff+ columns. It then runs analyze_top_bottom_pitches over those new ones to give the top and bottom 5 xWhiff+ pitches.

(EACH CSV STILL RETAINS THE 25000 LINES)

Here is the data for fastballs:

```
Analyzing top and bottom pitches for Fastball from FF_Final_xwhiff_separate.csv
Reading from FF_Final_xwhiff_separate.csv
Column count in types: 10
Columns in FF_Final_xwhiff_separate.csv: ["player_name", "release_speed", "release_pos_z", "release_spin_rate", "release_extension", "pfx_x", "pfx_z", "whiff", "xWhiff", "xWhiff+"]
Column indices - pitch_id: 0, xWhiff: 8, xWhiff+: 9

TOP 5 FASTBALL PITCHES BY xWhiff+:
Pitch ID            xWhiff          xWhiff+
------------------------------------------------------------
#1  Scott, Tanner     0.2751          157.43
    Features: release_speed: 95.20, release_spin_rate: 2790.00, pfx_x: 0.55, pfx_z: 1.45
#2  Díaz, Alexis      0.2746          157.14
    Features: release_speed: 94.20, release_spin_rate: 2900.00, pfx_x: -0.90, pfx_z: 1.27
#3  Imanaga, Shota    0.2741          156.86
    Features: release_speed: 91.70, release_spin_rate: 2654.00, pfx_x: 1.24, pfx_z: 1.78
#4  Scott, Tanner     0.2686          153.71
    Features: release_speed: 96.60, release_spin_rate: 2666.00, pfx_x: 0.60, pfx_z: 1.67
#5  Díaz, Alexis      0.2653          151.82
    Features: release_speed: 94.00, release_spin_rate: 2780.00, pfx_x: -0.69, pfx_z: 1.31

BOTTOM 5 FASTBALL PITCHES BY xWhiff+:
Pitch ID            xWhiff          xWhiff+
------------------------------------------------------------
#1  Dollander, Chase  0.0346          19.80
    Features: release_speed: 88.80, release_spin_rate: 21.00, pfx_x: 0.47, pfx_z: 0.37
#2  Schreiber, John   0.0360          20.60
    Features: release_speed: 92.70, release_spin_rate: 163.00, pfx_x: -1.29, pfx_z: -0.03
#3  Gray, Sonny       0.0367          21.00
    Features: release_speed: 91.50, release_spin_rate: 66.00, pfx_x: -1.21, pfx_z: 0.94
#4  Gray, Sonny       0.0371          21.23
    Features: release_speed: 93.00, release_spin_rate: 59.00, pfx_x: -1.16, pfx_z: 1.03
#5  Kremer, Dean      0.0413          23.63
    Features: release_speed: 94.60, release_spin_rate: 68.00, pfx_x: -0.75, pfx_z: 1.43
```

Here is the new CSV the program made with xWhiff and xWhiff+ values for fastballs:

```
player_name,release_speed,release_pos_z,release_spin_rate,release_extension,pfx_x,pfx_z,whiff,xWhiff,xWhiff+
"Patrick, Chad",88,5.4,2411,6.5,0.45,0.78,false,0.1873,107.18593716456992
"Patrick, Chad",87.7,5.68,2479,6.3,0.75,0.62,false,0.1863,106.61366841302389
"Patrick, Chad",86.4,5.63,2491,5.9,0.83,0.68,false,0.1908,109.18887779498098
"Patrick, Chad",92.9,5.81,2225,6.2,-1.13,1.1,false,0.1593,91.1624121212813
"Patrick, Chad",87.3,5.52,2480,6.1,0.17,0.88,false,0.1929,110.39064217322765
"Patrick, Chad",87.6,5.55,2416,6.1,0.12,0.99,false,0.1888,108.04434029188896
"Patrick, Chad",86.7,5.6,2438,6.1,0.6,0.58,false,0.1797,102.83669465282014
"Woods Richardson, Simeon",93.2,6.47,2192,6.8,-0.54,1.38,false,0.1586,90.76182399519908
"Mize, Casey",94,6.04,2036,6.9,-1.31,1.04,false,0.1353,77.42796208417678
"Lugo, Seth",92.7,5.3,2430,6.2,-1.23,0.59,false,0.1717,98.25854464045197
"Burke, Sean",94.4,6.45,2712,6.9,0.06,1.6,false,0.2344,134.13979536238756
"Lively, Ben",92,5.09,2077,6.9,-1.23,1.04,false,0.1540,88.12938773808739
"Lugo, Seth",92.8,5.53,2603,6.1,0.09,1.04,false,0.2177,124.582907211569
"Patrick, Chad",92.1,5.76,2132,6.1,-1.15,0.77,true,0.1408,80.57544021767991
"Burke, Sean",94.9,6.39,2607,7,-0.19,1.33,false,0.2089,119.546942197964
"Freeland, Kyle",90.6,5.93,2402,6.7,0.63,1.3,false,0.1985,113.59534718188537
"Lively, Ben",84.4,4.94,1946,6.8,-0.45,0.74,false,0.1371,78.45804583695963
"Mahle, Tyler",93.1,5.96,2324,6.5,-0.89,1.39,false,0.1788,102.32165277642873
"Patrick, Chad",93.7,5.87,2331,6.2,-0.51,1.29,false,0.1822,104.2673665316852
"Kikuchi, Yusei",94.7,5.37,2033,6.8,0.74,1.17,false,0.1688,96.59896526096851
"Mize, Casey",93,6.02,1899,7.2,-1.38,0.97,true,0.1212,69.35897268737787
"Baz, Shane",95.7,5.62,2387,6.4,-0.6,1.44,false,0.1999,114.3965234340498
"Kikuchi, Yusei",95,5.31,2111,6.9,0.83,1.44,false,0.1889,108.10156716704356
"Berríos, José",94,5.65,1991,6.4,-1.45,0.73,false,0.1280,73.25040019789083
"Mahle, Tyler",92.8,5.96,2382,6.4,-0.84,1.53,false,0.1905,109.01719716951719
"Lively, Ben",90,5.15,2109,7,-0.78,1.42,false,0.1711,97.91518338952436
"Houck, Tanner",93.7,5.55,1989,6.1,-1.45,0.false,0.1109,63.46460454645384
```

Offspeed:

```
Analyzing top and bottom pitches for Offspeed from Offspeed_final_xwhiff_separate.csv
Reading from Offspeed_final_xwhiff_separate.csv
Column count in types: 10
Columns in Offspeed_final_xwhiff_separate.csv: ["player_name", "release_speed", "release_pos_z", "release_spin_rate", "release_extension", "pfx_x", "pfx_z", "whiff", "xWhiff", "xWhiff+"]
Column indices - pitch_id: 0, xWhiff: 8, xWhiff+: 9

TOP 5 OFFSPEED PITCHES BY xWhiff+:
Pitch ID            xWhiff         xWhiff+
------------------------------------------------------------
#1  Sasaki, Roki      0.4474          144.39
    Features: release_speed: 83.90, release_spin_rate: 441.00, pfx_x: -54.72, pfx_z: -125.28
#2  Sasaki, Roki      0.4436          143.17
    Features: release_speed: 84.80, release_spin_rate: 462.00, pfx_x: 15.84, pfx_z: -99.36
#3  Sasaki, Roki      0.4425          142.81
    Features: release_speed: 86.60, release_spin_rate: 382.00, pfx_x: -8.64, pfx_z: -87.84
#4  Sasaki, Roki      0.4410          142.33
    Features: release_speed: 87.10, release_spin_rate: 504.00, pfx_x: -41.76, pfx_z: -119.52
#5  Sasaki, Roki      0.4368          140.97
    Features: release_speed: 86.80, release_spin_rate: 391.00, pfx_x: -76.32, pfx_z: -73.44

BOTTOM 5 OFFSPEED PITCHES BY xWhiff+:
Pitch ID            xWhiff         xWhiff+
------------------------------------------------------------
#1  Alvarez, Eddy     0.2121          68.45
    Features: release_speed: 73.80, release_spin_rate: 2010.00, pfx_x: -126.72, pfx_z: 188.64
#2  Castillo, Luis    0.2164          69.84
    Features: release_speed: 88.40, release_spin_rate: 2063.00, pfx_x: -201.60, pfx_z: 31.68
#3  Jacob, Alek       0.2190          70.68
    Features: release_speed: 73.60, release_spin_rate: 2094.00, pfx_x: -231.84, pfx_z: 87.84
#4  Castillo, Luis    0.2197          70.91
    Features: release_speed: 89.20, release_spin_rate: 2059.00, pfx_x: -233.28, pfx_z: 41.76
#5  Castillo, Luis    0.2202          71.07
    Features: release_speed: 88.00, release_spin_rate: 1846.00, pfx_x: -213.12, pfx_z: 11.52
```

```
player_name,release_speed,release_pos_z,release_spin_rate,release_extension,pfx_x,pfx_z,whiff,xWhiff,xWhiff+
"Mize, Casey",88.3,5.98,1301,7,-169.92000000000002,46.08,false,0.3407,109.95803390771579
"Mize, Casey",88.5,5.99,1417,7,-192.96000000000004,47.519999999999996,true,0.3376,108.95753521351585
"Mize, Casey",87.9,5.99,1398,6.8,-201.59999999999997,89.28,false,0.3177,102.53497907978075
"Mahle, Tyler",87.4,6.03,1895,6.4,-180,133.92000000000002,false,0.2863,92.40089553207817
"Patrick, Chad",87.5,5.62,2102,6.2,-106.55999999999999,102.24,false,0.2690,86.817467335414
"Senga, Kodai",83.9,5.78,1089,6.4,-108,41.75999999999999,false,0.3226,104.11641249964518
"Skenes, Paul",94.2,5.64,1773,6.8,-129.60000000000002,82.08,false,0.2903,93.69186158911035
"Baz, Shane",88.5,5.55,1832,6.5,-197.28000000000003,86.39999999999999,true,0.2789,90.01260832656864
"Senga, Kodai",85.1,5.71,1142,6.3,-110.88,-23.04,false,0.3338,107.73111745933528
"Senga, Kodai",85.6,5.78,1155,6.3,-105.12,12.96,false,0.3251,104.9232662852903
"Senga, Kodai",83.4,5.74,1377,6.4,-110.88,50.39999999999999,false,0.3138,101.2762871741744
"Woods Richardson, Simeon",82,6.24,2412,6.8,-178.56,112.32,true,0.3169,102.27678586837435
"Skenes, Paul",93.2,5.66,1752,6.7,-144,73.44,false,0.2918,94.17597386049742
"Mahle, Tyler",80.6,5.71,1656,6.4,-180,56.16,false,0.3071,99.1139190286455
"Woods Richardson, Simeon",81.5,6.25,2386,6.5,-152.64000000000001,122.39999999999999,true,0.3062,98.82345166581327
"Skenes, Paul",92.2,5.74,1657,6.6,-131.04,28.800000000000004,false,0.3098,99.98532111714222
"Mahle, Tyler",82.1,6.09,1608,6.2,-168.48,83.51999999999998,false,0.3105,100.21124017712285
"Walker, Taijuan",89.2,6.29,1668,6.1,-167.03999999999996,51.84,false,0.3167,102.21223756552271
"Freeland, Kyle",86.4,5.99,1717,6.6,211.68,103.68,false,0.3206,103.4709294711291
"Mahle, Tyler",83.8,6.05,1712,6.3,-171.35999999999999,113.76,false,0.2982,96.2415195517489
"Skenes, Paul",86.3,5.63,1804,6.7,-207.36,87.84,true,0.2919,94.20824801192322
"Houck, Tanner",91.5,5.61,1844,6.2,-156.96000000000004,-36,false,0.3070,99.08164487721969
"Baz, Shane",89.9,5.64,2148,6.4,-194.40000000000003,72,false,0.2781,89.75441511516222
"Mahle, Tyler",82,6.05,1718,6.2,-162.71999999999997,139.68,false,0.2897,93.49821668055553
"Skenes, Paul",86.8,5.64,1681,6.7,-204.48,60.480000000000004,true,0.3022,97.53248560878109
"Abbott, Andrew",84.5,6.08,1647,6.3,180,129.60000000000002,false,0.3092,99.79167620858738
"Berríos, José",87.7,5.47,1537,6.6,-190.07999999999998,46.08,true,0.2962,95.59603652323283
"Abbott, Andrew",85.4,6.09,1730,6.4,154.07999999999998,149.76,false,0.3032,97.85522712303913
"Walker, Taijuan",89.2,6.27,1438,6.1,-175.68,61.92,false,0.3157,101.88949605126467
"Mahle, Tyler",83.8,6.04,1792,6.3,-156.96000000000004,108,false,0.2989,96.46743861172953
```

Breaking:

```
Analyzing top and bottom pitches for Breaking from Breaking_Final_xwhiff_separate.csv
Reading from Breaking_Final_xwhiff_separate.csv
Column count in types: 10
Columns in Breaking_Final_xwhiff_separate.csv: ["player_name", "release_speed", "release_pos_z", "release_spin_rate", "release_extension", "pfx_x", "pfx_z", "whiff", "xWhiff", "xWhiff+"]
Column indices - pitch_id: 0, xWhiff: 8, xWhiff+: 9

TOP 5 BREAKING PITCHES BY xWhiff+:
Pitch ID            xWhiff          xWhiff+
------------------------------------------------------------
#1  Díaz, Edwin       0.4926          157.91
    Features: release_speed: 91.80, release_spin_rate: 2362.00, pfx_x: -0.60, pfx_z: 4.68
#2  Díaz, Alexis      0.4783          153.32
    Features: release_speed: 90.60, release_spin_rate: 2828.00, pfx_x: 3.24, pfx_z: 4.92
#3  Díaz, Alexis      0.4740          151.95
    Features: release_speed: 90.20, release_spin_rate: 2701.00, pfx_x: 1.80, pfx_z: 5.76
#4  Díaz, Edwin       0.4709          150.95
    Features: release_speed: 91.90, release_spin_rate: 2371.00, pfx_x: 1.32, pfx_z: 2.76
#5  Díaz, Alexis      0.4678          149.96
    Features: release_speed: 90.30, release_spin_rate: 2577.00, pfx_x: 0.12, pfx_z: 5.16

BOTTOM 5 BREAKING PITCHES BY xWhiff+:
Pitch ID            xWhiff          xWhiff+
------------------------------------------------------------
#1  Duran, Ezequiel   0.0231          7.40
    Features: release_speed: 40.60, release_spin_rate: 810.00, pfx_x: 14.04, pfx_z: 16.68
#2  Duran, Ezequiel   0.0240          7.69
    Features: release_speed: 41.70, release_spin_rate: 993.00, pfx_x: 21.48, pfx_z: 18.60
#3  Duran, Ezequiel   0.0242          7.76
    Features: release_speed: 40.80, release_spin_rate: 944.00, pfx_x: 0.96, pfx_z: 16.32
#4  Pereda, Jhonny    0.0303          9.71
    Features: release_speed: 42.60, release_spin_rate: 839.00, pfx_x: -7.32, pfx_z: 13.56
#5  Hernández, Enrique 0.0315          10.10
    Features: release_speed: 48.50, release_spin_rate: 1099.00, pfx_x: 16.68, pfx_z: 18.48
```

```
player_name,release_speed,release_pos_z,release_spin_rate,release_extension,pfx_x,pfx_z,whiff,xWhiff,xWhiff+
"Woods Richardson, Simeon",85.8,6.26,2250,6.8,4.92,6.84,false,0.3004,96.29682948681571
"Woods Richardson, Simeon",85.4,6.24,2234,6.7,5.16,6.959999999999999,true,0.2915,93.44382754795866
"Burke, Sean",86.9,6.18,2634,7.1,4.199999999999999,1.6800000000000002,false,0.3598,115.33821321356956
"Kikuchi, Yusei",85.7,5.13,2161,6.7,-3.5999999999999996,3.12,false,0.3470,111.23501941386502
"Lugo, Seth",81.9,5.74,3294,5.8,11.040000000000001,-14.04,false,0.3260,104.5032170862248
"Kikuchi, Yusei",87.9,5.1,2327,6.9,-2.7600000000000002,3.3600000000000003,false,0.3843,123.1919825958165
"Freeland, Kyle",85,5.56,2410,6.7,-8.040000000000001,-2.52,false,0.3599,115.37026941512975
"Gallen, Zac",81.9,5.8,2291,6.7,2.7600000000000002,-10.08,false,0.3288,105.40079072991013
"Burke, Sean",87.4,6.22,2554,7,4.68,3.7199999999999998,false,0.3474,111.3632442201058
"Lugo, Seth",72,5.21,2446,5.7,14.64,1.32,false,0.1915,61.38762598776702
"Baz, Shane",86.4,5.66,2262,6.5,4.199999999999999,1.6800000000000002,false,0.3243,103.95826165970152
"Lively, Ben",81,5.05,1934,6.9,4.5600000000000005,7.08,false,0.2899,92.93092832299558
"Kikuchi, Yusei",88.6,5.15,2328,6.9,-3.4799999999999995,3,true,0.3912,125.40386050346974
"Freeland, Kyle",85.3,5.44,2450,6.7,-11.64,-4.68,true,0.3796,121.6853411224875
"Berríos, José",82.1,5.51,2365,6.3,17.28,-0.96,false,0.2796,89.62913956229586
"Woods Richardson, Simeon",87,6.38,2224,6.7,4.32,5.4,false,0.3053,97.86758336326511
"Patrick, Chad",86.5,5.54,2392,6.1,6.48,4.08,false,0.3040,97.45085274298262
"Burke, Sean",86.4,6.27,2462,6.8,5.64,6.24,false,0.3143,100.75264150368237
"Mize, Casey",88.2,5.87,2136,7,3.5999999999999996,2.4000000000000004,false,0.3534,113.28661631371727
"Lugo, Seth",82.1,5.6,3298,5.8,11.879999999999999,-13.68,false,0.3302,105.84957755175282
"Baz, Shane",82.4,5.73,2649,6.5,8.16,-11.879999999999999,false,0.3382,108.41407367656817
"Lugo, Seth",78.7,5.51,3210,5.5,14.04,-10.44,false,0.2750,88.15455429052705
"Gallen, Zac",82.8,5.7,2426,6.7,2.04,-14.52,false,0.3596,115.27410081044917
"Mize, Casey",87,5.89,2102,6.9,3.24,7.199999999999999,false,0.3219,103.18891282225692
"Kikuchi, Yusei",84.2,5.04,2125,7,-5.5200000000000005,3.4799999999999995,false,0.3527,113.06222290279595
"Freeland, Kyle",83.8,5.61,2414,6.7,-10.2,-4.199999999999999,true,0.3555,113.95979654648131
"Lugo, Seth",80.1,5.59,3176,5.8,12.96,-12.84,false,0.3050,97.77141475858453
"Gallen, Zac",83.3,5.84,2392,6.6,5.64,-9.84,false,0.3342,107.1318256141605
"Mize, Casey",83.5,5.79,2086,6.8,5.64,-1.7999999999999998,false,0.3082,98.79721320851066
```

To run, have cleaned data given ready, and cargo run –release to run the program. Will add 3 new CSVs to the file when run. Should take ~4 seconds to run.