```
In [1]:
         # Import our dependencies
         from sklearn.model_selection import train_test_split
         from sklearn.preprocessing import StandardScaler,OneHotEncoder, MinMaxScaler
         import pandas as pd
         import tensorflow as tf
         import numpy as np
         # Import our input dataset
         df = pd.read_csv('./pitcher_salaries_cleaned.csv')
         df.head()
Out[1]:
                                                                                               Batters
                                                     Earned Strike
                                                                                         Outs
                                                                   Home
                                                                                                Faced
                                                                                                        Games
                                   Salary ERA Hits
                                                                                                               Weight I
            Vear
                    Full Name Age
                                                                          Wins Losses
                                                                                       Pitched
                                                                                                  by
                                                       Runs
                                                             Outs
                                                                   Runs
                                                                                                      Finished
```

Pitcher 1990 AbbottJim 185000 4.51 105 200 23 246 106 16 10 14 635 925 0 0 5 **1** 1990 AbbottPaul 100000 5.97 37 23 25 0 104 162 0 185 **2** 1990 AldredScott 100000 3.77 13 6 7 0 1 2 43 63 0 195 82 20 7 18 797 0 **3** 1990 AndersonAllan 26 300000 4.53 214 95 566 178 1990 23 100000 2.76 127 13 12 557 784 1 180 **AppierKevin** 57

Create Salary Brackets

```
# Look at distribution of salaries (suppressing scientific notation)
df['Salary'].describe().apply(lambda x: format(x, 'f'))
```

```
Out[2]: count
                      4937.000000
         mean
                   3011304.443387
                   4265619.190449
         std
        min
                    100000.000000
         25%
                    327000.000000
         50%
                    980000.000000
         75%
                   4000000.000000
                  33000000.000000
         max
         Name: Salary, dtype: object
```

```
In [3]:
# create salary brackets and labels
bins = [0, 499999, 4999999, 34999999]
labels = ['low', 'mid', 'high', 'top']
```

```
In [4]:
# apply salary brackets
df['salBin'] = pd.cut(df['Salary'], bins=bins, labels=labels)
df
```

Out[4]:		Year	Full Name	Age	Salary	ERA	Hits	Earned Runs	Strike Outs	Home Runs	Wins	Losses	Outs Pitched	Batters Faced by Pitcher	Games Finished
	0	1990	AbbottJim	23	185000	4.51	246	106	105	16	10	14	635	925	0
	1	1990	AbbottPaul	23	100000	5.97	37	23	25	0	0	5	104	162	0
	2	1990	AldredScott	22	100000	3.77	13	6	7	0	1	2	43	63	0
	3	1990	AndersonAllan	26	300000	4.53	214	95	82	20	7	18	566	797	0
	4	1990	AppierKevin	23	100000	2.76	179	57	127	13	12	8	557	784	1

	Year	Full Name	Age	Salary	ERA	Hits	Earned Runs	Strike Outs	Home Runs	Wins	Losses	Outs Pitched	Faced by Pitcher	Games Finished
4932	2016	WorleyVance	29	2600000	3.53	84	34	56	11	2	2	260	365	13
4933	2016	WrightMike	26	510500	5.79	81	48	50	12	3	4	224	328	5
4934	2016	WrightSteven	32	514500	3.33	138	58	127	12	13	6	470	656	0
4935	2016	YoungChris	37	4250000	6.19	104	61	94	28	3	9	266	406	7
4936	2016	ZimmermannJordan	30	18000000	4.87	118	57	66	14	9	7	316	450	1

4937 rows × 20 columns

```
In [5]:
### Drop unnecessary columns
df= df.drop(["Full Name","Team","League","Age","Year","Salary"],1)
df.head()
```

C:\Users\alyss\anaconda3\envs\mlenv\lib\site-packages\ipykernel_launcher.py:2: FutureWarning: In a future version of pandas all arguments of DataFrame.drop except for the argument 'labels' will be keyword-only

Out[5]:		ERA	Hits	Earned Runs	Strike Outs	Home Runs	Wins	Losses	Outs Pitched	Batters Faced by Pitcher	Games Finished	Weight	Height	Games Started	salBin
	0	4.51	246	106	105	16	10	14	635	925	0	200	75	33	low
	1	5.97	37	23	25	0	0	5	104	162	0	185	75	7	low
	2	3.77	13	6	7	0	1	2	43	63	0	195	76	3	low
	3	4.53	214	95	82	20	7	18	566	797	0	178	71	31	low
	4	2.76	179	57	127	13	12	8	557	784	1	180	74	24	low

Reduce number of rows

kept getting error in one-hot encoding, ValueError: Buffer has wrong number of dimensions (expected 1, got 2)

some suggested reducing sample size would solve issue (https://github.com/lmcinnes/umap/issues/496)

-- Update: reducing sample size did not solve issue with one-hot encoding

Encode Salary Bins column

```
In [6]: # use get_dummies to one-hot encode the salarybin column
    encoded_df=pd.get_dummies(df,columns=['salBin'],prefix="salBin")
    encoded_df
```

Out[6]:		ERA	Hits	Earned Runs	Strike Outs	Home Runs	Wins	Losses	Outs Pitched	Batters Faced by Pitcher	Games Finished	Weight	Height	Games Started	salBin_low	salE
_	0	4.51	246	106	105	16	10	14	635	925	0	200	75	33	1	

	ERA	Hits	Earned Runs	Strike Outs	Home Runs	Wins	Losses	Outs Pitched	Faced by Pitcher	Games Finished	Weight	Height	Games Started	salBin_low	salE
1	5.97	37	23	25	0	0	5	104	162	0	185	75	7	1	
2	3.77	13	6	7	0	1	2	43	63	0	195	76	3	1	
3	4.53	214	95	82	20	7	18	566	797	0	178	71	31	1	
4	2.76	179	57	127	13	12	8	557	784	1	180	74	24	1	
4932	3.53	84	34	56	11	2	2	260	365	13	240	74	4	0	
4933	5.79	81	48	50	12	3	4	224	328	5	240	78	12	0	
4934	3.33	138	58	127	12	13	6	470	656	0	215	74	24	0	
4935	6.19	104	61	94	28	3	9	266	406	7	255	82	13	0	
4936	4.87	118	57	66	14	9	7	316	450	1	225	74	18	0	

4937 rows × 17 columns

Split Features/Target & Training/Testing Sets

Split into features and target

- y variable: Our target variables, Salary-Bin_low, Salary-Bin_mid, Salary-Bin_high, Salary-Bin_top
- X variable: Our features

```
In [7]: # Split our preprocessed data into our features and target arrays
y = encoded_df[["salBin_low","salBin_mid","salBin_high","salBin_top"]].values
X = encoded_df.drop(["salBin_low","salBin_mid","salBin_high","salBin_top"],1).values
# Split the preprocessed data into a training and testing dataset
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=1)
```

C:\Users\alyss\anaconda3\envs\mlenv\lib\site-packages\ipykernel_launcher.py:3: FutureWarning: In a future version of pandas all arguments of DataFrame.drop except for the argument 'labels' will be keyword-only This is separate from the ipykernel package so we can avoid doing imports until

Build and Instantiate StandardScaler object, then standardize numerical features

```
In [8]: # Create a StandardScaler instance
scaler = StandardScaler()

# Fit the StandardScaler
X_scaler = scaler.fit(X_train)

# Scale the data
X_train_scaled = X_scaler.transform(X_train)
X_test_scaled = X_scaler.transform(X_test)
```

Build Neural Net Framework

```
number_input_features = len(X_train[0])
hidden_nodes_layer1 = 144
hidden_nodes_layer2 = 64
hidden_nodes_layer3 = 16
nn = tf.keras.models.Sequential()
# First hidden Layer
nn.add(
   tf.keras.layers.Dense(units=hidden_nodes_layer1, input_dim=number_input_features, activation="softmax"
# Second hidden Layer
nn.add(tf.keras.layers.Dense(units=hidden_nodes_layer2, activation="relu"))
# Third hidden Layer
nn.add(tf.keras.layers.Dense(units=hidden_nodes_layer3, activation="relu"))
# Output Layer
nn.add(tf.keras.layers.Dense(units=4, activation="softmax"))
# Check the structure of the model
nn.summary()
```

Model: "sequential_1"

Layer (type)	Output Shape	Param #								
dense_4 (Dense)	(None, 144)	2016								
dense_5 (Dense)	(None, 64)	9280								
dense_6 (Dense)	(None, 16)	1040								
dense_7 (Dense)	(None, 4)	68								
Total params: 12,404 Trainable params: 12,404										

Compile the Model

Non-trainable params: 0

```
# Compile the model
nn.compile(loss="CategoricalCrossentropy", optimizer="adam", metrics=["accuracy"])
```

Train the model

```
In [17]:
     # Train the model
     fit_model = nn.fit(X_train,y_train,epochs=200)
     Epoch 1/200
     Epoch 2/200
     Epoch 3/200
     116/116 [=========== ] - 0s 696us/step - loss: 1.2013 - accuracy: 0.4103
     Epoch 4/200
     116/116 [============ ] - 0s 687us/step - loss: 1.2011 - accuracy: 0.4130
     Epoch 5/200
     116/116 [================ ] - 0s 748us/step - loss: 1.2007 - accuracy: 0.4049
     Epoch 6/200
     Epoch 7/200
     116/116 [=============== ] - 0s 678us/step - loss: 1.2005 - accuracy: 0.4130
```

```
Epoch 8/200
Epoch 9/200
116/116 [=========== ] - 0s 687us/step - loss: 1.2013 - accuracy: 0.4068
Epoch 10/200
Epoch 11/200
116/116 [=========== ] - 0s 704us/step - loss: 1.2009 - accuracy: 0.4122
Epoch 12/200
Epoch 13/200
Epoch 14/200
Epoch 15/200
116/116 [=========== ] - 0s 704us/step - loss: 1.2005 - accuracy: 0.4130
Epoch 16/200
Epoch 17/200
Epoch 18/200
116/116 [=========== ] - 0s 896us/step - loss: 1.2020 - accuracy: 0.4087
Epoch 19/200
Epoch 20/200
116/116 [============ ] - 0s 704us/step - loss: 1.2014 - accuracy: 0.4130
Epoch 21/200
Epoch 22/200
116/116 [=========== ] - 0s 687us/step - loss: 1.2006 - accuracy: 0.4090
Epoch 23/200
Epoch 24/200
116/116 [=========== ] - 0s 678us/step - loss: 1.2011 - accuracy: 0.4130
Epoch 25/200
Fnoch 26/200
Epoch 27/200
Epoch 28/200
Epoch 29/200
116/116 [================= ] - 0s 687us/step - loss: 1.2011 - accuracy: 0.4117
Epoch 30/200
Epoch 31/200
Epoch 32/200
Epoch 33/200
Epoch 34/200
Epoch 35/200
Epoch 36/200
Epoch 37/200
Epoch 38/200
Epoch 39/200
116/116 [=========== ] - 0s 704us/step - loss: 1.2008 - accuracy: 0.4057
Epoch 40/200
Epoch 41/200
Epoch 42/200
Epoch 43/200
116/116 [============ ] - 0s 687us/step - loss: 1.2004 - accuracy: 0.4130
Epoch 44/200
116/116 [================= ] - 0s 687us/step - loss: 1.2009 - accuracy: 0.3971
Epoch 45/200
```

```
Epoch 46/200
Epoch 47/200
116/116 [============ ] - 0s 748us/step - loss: 1.2005 - accuracy: 0.4082
Epoch 48/200
Epoch 49/200
116/116 [=========== ] - 0s 687us/step - loss: 1.2003 - accuracy: 0.4133
Epoch 50/200
Epoch 51/200
Epoch 52/200
Epoch 53/200
116/116 [=========== ] - 0s 678us/step - loss: 1.1997 - accuracy: 0.4187
Epoch 54/200
Epoch 55/200
Epoch 56/200
116/116 [============ ] - 0s 704us/step - loss: 1.1681 - accuracy: 0.4776
Epoch 57/200
Epoch 58/200
116/116 [=========== ] - 0s 687us/step - loss: 1.1673 - accuracy: 0.4795
Epoch 59/200
Epoch 60/200
116/116 [============ ] - 0s 678us/step - loss: 1.1733 - accuracy: 0.4719
Epoch 61/200
Epoch 62/200
116/116 [============ ] - 0s 696us/step - loss: 1.1629 - accuracy: 0.4860
Epoch 63/200
Epoch 64/200
Epoch 65/200
Epoch 66/200
Epoch 67/200
116/116 [================= ] - 0s 696us/step - loss: 1.1587 - accuracy: 0.4887
Epoch 68/200
Epoch 69/200
Epoch 70/200
Epoch 71/200
Epoch 72/200
116/116 [================= ] - 0s 704us/step - loss: 1.1612 - accuracy: 0.4857
Epoch 73/200
Epoch 74/200
Epoch 75/200
Epoch 76/200
Epoch 77/200
116/116 [============ ] - 0s 704us/step - loss: 1.1594 - accuracy: 0.4873
Epoch 78/200
Epoch 79/200
Epoch 80/200
Epoch 81/200
116/116 [============ ] - 0s 678us/step - loss: 1.1624 - accuracy: 0.4822
Epoch 82/200
116/116 [================== ] - 0s 696us/step - loss: 1.1587 - accuracy: 0.4827
Epoch 83/200
```

```
Epoch 84/200
Epoch 85/200
116/116 [============ ] - 0s 696us/step - loss: 1.1584 - accuracy: 0.4838
Epoch 86/200
Epoch 87/200
116/116 [============ ] - 0s 687us/step - loss: 1.1595 - accuracy: 0.4760
Epoch 88/200
Epoch 89/200
Epoch 90/200
Epoch 91/200
116/116 [=========== ] - 0s 696us/step - loss: 1.1554 - accuracy: 0.4830
Epoch 92/200
Epoch 93/200
Epoch 94/200
116/116 [=========== ] - 0s 678us/step - loss: 1.1604 - accuracy: 0.4716
Epoch 95/200
Epoch 96/200
116/116 [============ ] - 0s 704us/step - loss: 1.1539 - accuracy: 0.4884
Epoch 97/200
Epoch 98/200
116/116 [============ ] - 0s 678us/step - loss: 1.1530 - accuracy: 0.4884
Epoch 99/200
Epoch 100/200
116/116 [============] - 0s 687us/step - loss: 1.1579 - accuracy: 0.4692
Epoch 101/200
Fnoch 102/200
Epoch 103/200
Epoch 104/200
Epoch 105/200
Epoch 106/200
Epoch 107/200
Epoch 108/200
Epoch 109/200
Epoch 110/200
116/116 [================= ] - 0s 708us/step - loss: 1.1497 - accuracy: 0.4897
Epoch 111/200
Epoch 112/200
Epoch 113/200
Epoch 114/200
116/116 [================= ] - 0s 713us/step - loss: 1.1470 - accuracy: 0.4841
Epoch 115/200
116/116 [=========== ] - 0s 748us/step - loss: 1.1447 - accuracy: 0.4905
Epoch 116/200
Epoch 117/200
Epoch 118/200
Epoch 119/200
116/116 [============ ] - 0s 678us/step - loss: 1.1470 - accuracy: 0.4830
Epoch 120/200
Epoch 121/200
```

```
Epoch 122/200
Epoch 123/200
116/116 [============ ] - 0s 687us/step - loss: 1.1449 - accuracy: 0.4851
Epoch 124/200
Epoch 125/200
116/116 [============ ] - 0s 696us/step - loss: 1.1391 - accuracy: 0.4876
Epoch 126/200
Epoch 127/200
Epoch 128/200
Epoch 129/200
116/116 [=========== ] - 0s 731us/step - loss: 1.1443 - accuracy: 0.4822
Epoch 130/200
Epoch 131/200
Epoch 132/200
116/116 [============ ] - 0s 748us/step - loss: 1.1446 - accuracy: 0.4824
Epoch 133/200
Epoch 134/200
116/116 [============ ] - 0s 687us/step - loss: 1.1452 - accuracy: 0.4851
Epoch 135/200
Epoch 136/200
116/116 [=========== ] - 0s 948us/step - loss: 1.1458 - accuracy: 0.4806
Epoch 137/200
Epoch 138/200
116/116 [============] - 0s 704us/step - loss: 1.1433 - accuracy: 0.4846
Epoch 139/200
Epoch 140/200
Epoch 141/200
Epoch 142/200
116/116 [============ ] - 0s 696us/step - loss: 1.1395 - accuracy: 0.4824
Epoch 143/200
Epoch 144/200
Epoch 145/200
Epoch 146/200
116/116 [============ ] - 0s 687us/step - loss: 1.1401 - accuracy: 0.4843
Epoch 147/200
Epoch 148/200
Epoch 149/200
Epoch 150/200
Epoch 151/200
Epoch 152/200
Epoch 153/200
116/116 [=========== ] - 0s 713us/step - loss: 1.1404 - accuracy: 0.4819
Epoch 154/200
Epoch 155/200
Epoch 156/200
Epoch 157/200
116/116 [============ ] - 0s 687us/step - loss: 1.1436 - accuracy: 0.4795
Epoch 158/200
Epoch 159/200
```

```
Epoch 160/200
Epoch 161/200
116/116 [============ ] - 0s 696us/step - loss: 1.1416 - accuracy: 0.4843
Epoch 162/200
Epoch 163/200
116/116 [============ ] - 0s 748us/step - loss: 1.1386 - accuracy: 0.4849
Epoch 164/200
Epoch 165/200
116/116 [================= ] - 0s 696us/step - loss: 1.1398 - accuracy: 0.4827
Epoch 166/200
Epoch 167/200
116/116 [=========== ] - 0s 774us/step - loss: 1.1484 - accuracy: 0.4768
Epoch 168/200
Epoch 169/200
Epoch 170/200
116/116 [============ ] - 0s 696us/step - loss: 1.1443 - accuracy: 0.4849
Epoch 171/200
Epoch 172/200
116/116 [=========== ] - 0s 774us/step - loss: 1.1376 - accuracy: 0.4824
Epoch 173/200
Epoch 174/200
116/116 [=========== ] - 0s 713us/step - loss: 1.1407 - accuracy: 0.4827
Epoch 175/200
Epoch 176/200
116/116 [=========== ] - 0s 748us/step - loss: 1.1370 - accuracy: 0.4792
Epoch 177/200
Epoch 178/200
Epoch 179/200
Epoch 180/200
116/116 [============ ] - 0s 713us/step - loss: 1.1355 - accuracy: 0.4838
Epoch 181/200
Epoch 182/200
Epoch 183/200
Epoch 184/200
Epoch 185/200
Epoch 186/200
Epoch 187/200
Epoch 188/200
Epoch 189/200
Epoch 190/200
Epoch 191/200
116/116 [=========== ] - 0s 713us/step - loss: 1.1412 - accuracy: 0.4857
Epoch 192/200
Epoch 193/200
Epoch 194/200
Epoch 195/200
116/116 [============ ] - 0s 730us/step - loss: 1.1369 - accuracy: 0.4838
Epoch 196/200
116/116 [================== ] - 0s 774us/step - loss: 1.1345 - accuracy: 0.4835
Epoch 197/200
```

```
Epoch 198/200
        Epoch 199/200
        116/116 [=========== ] - 0s 691us/step - loss: 1.1404 - accuracy: 0.4833
        Epoch 200/200
        In [18]:
        # Evaluate the model using the test data
        model_loss, model_accuracy = nn.evaluate(X_test_scaled,y_test,verbose=2)
        print(f"Loss: {model_loss*100:.2f}%, Accuracy: {model_accuracy*100:.2f}%")
        39/39 - 0s - loss: 1.2470 - accuracy: 0.4219 - 91ms/epoch - 2ms/step
        Loss: 124.70%, Accuracy: 42.19%
In [19]:
        # Create a DataFrame containing training history
        history df = pd.DataFrame(fit model.history, index=range(1,len(fit model.history["loss"])+1))
        # Plot the loss
        history_df.plot(y="loss")
Out[19]: <AxesSubplot:>
        1.26
                                               loss
        1.24
        1.22
        1.20
        1.18
        1.16
        1.14
             Ò
                 25
                     50
                              100
                                  125
                                       150
                                           175
                                                200
In [20]:
        # Plot the accuracy
        history_df.plot(y="accuracy")
Out[20]: <AxesSubplot:>
                accuracy
        0.48
        0.46
        0.44
        0.42
        0.40
                              100
                     50
                          75
                                  125
                                       150
                                           175
                                                200
In [ ]:
```