NBA Dataset Assignment 1

Kael Villa

Sheridan College

PROG25211 AI and ML - Python

Soleimani, Ahmad

June 9, 2025

Will Chris Paul make the Hall of Fame (HOF)?

2. CHOOSE A DATASET. You can either go to Kaggle or to another site that contains a set of data. You may want to choose a dataset from a topic that you like, as it will make the assignment easier and more enjoyable. Provide a link in your report to your dataset.

https://www.kaggle.com/code/devraai/nba-player-analysis-hof-prediction

V2: NBA Player Database

3. ASK A QUESTION. In your report, ask a logistic question about your data. If you think about what we did in class, the question might be “Would I survive the Titanic?”. The question you ask will be the focus of your assignment. It should be part of the title of your assignment, and what your end goal is.

Will Chris Paul make the Hall of Fame (HOF)?

4. CLEAN YOUR DATA. You will need to go through your data removing null or empty values, removing columns that are not relevant to your question, and changing data so that the ML algorithm can process it. You will need to provide an explanation for each row you are deleting or altering.

0	Abdelnaby	1991	1995	'Center']	82	240.0	1968	['Duke']	False	False	256	5.7	3.3	0.3	50.2	0.0	70.1	50.2	13.0	4.8
1	Zaid Abdul-Aziz	1969	1978	['Center', 'Forward']	81	235.0	April 7, 1946	['Iowa State']	False	False	505	9.0	8.0	1.2	42.8	NaN	72.8	NaN	15.1	17.5
2	Kareem Abdul-Jabbar	1970	1989	['Center']	86	225.0	April 16, 1947	['UCLA']	True	False	1560	24.6	11.2	3.6	55.9	5.6	72.1	55.9	24.6	273.4
3	Mahmoud Abdul-Rauf	1991	2001	['Guard']	73	162.0	March 9, 1969	['LSU']	False	False	586	14.6	1.9	3.5	44.2	35.4	90.5	47.2	15.4	25.2
4	Tariq Abdul-Wahad	1998	2003	['Forward']	78	223.0	November 3, 1974	['Michigan', ' San Jose State']	False	False	236	7.8	3.3	1.1	41.7	23.7	70.3	42.2	11.4	3.5

```
if ( 'Birthday' in NBA ):
    NBA.drop( columns=[ 'Birthday', 'Active', 'Debut', 'Final', 'PER' ] )
    #Their birthday, active, when they played and PER doesn't effect their Chances of making the HOF
NBA.dropna(axis=0, inplace=True)
NBA[ 'HOF' ].replace( {'False':0, 'True':1}, inplace=True )
NBA.head()
```

✓ 0.0s Python

if ('Birthday' in NBA):

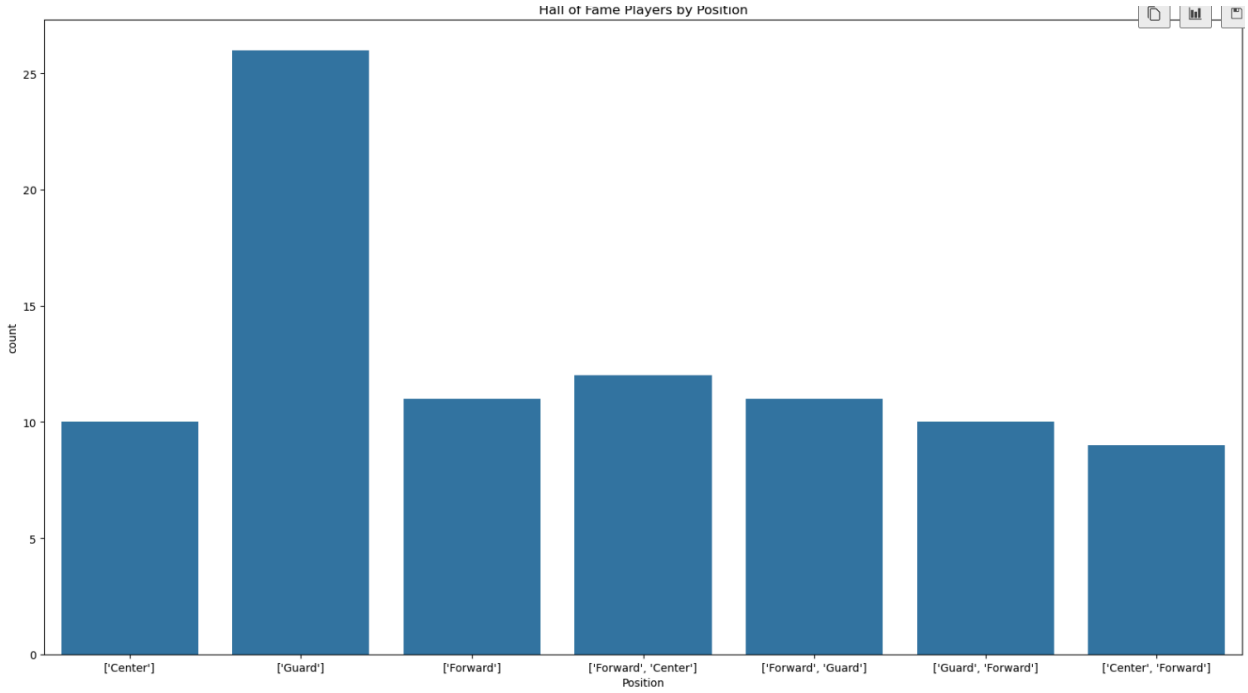
NBA.drop(columns=['Birthday', 'Active', 'Debut', 'Final', 'PER'])HOF

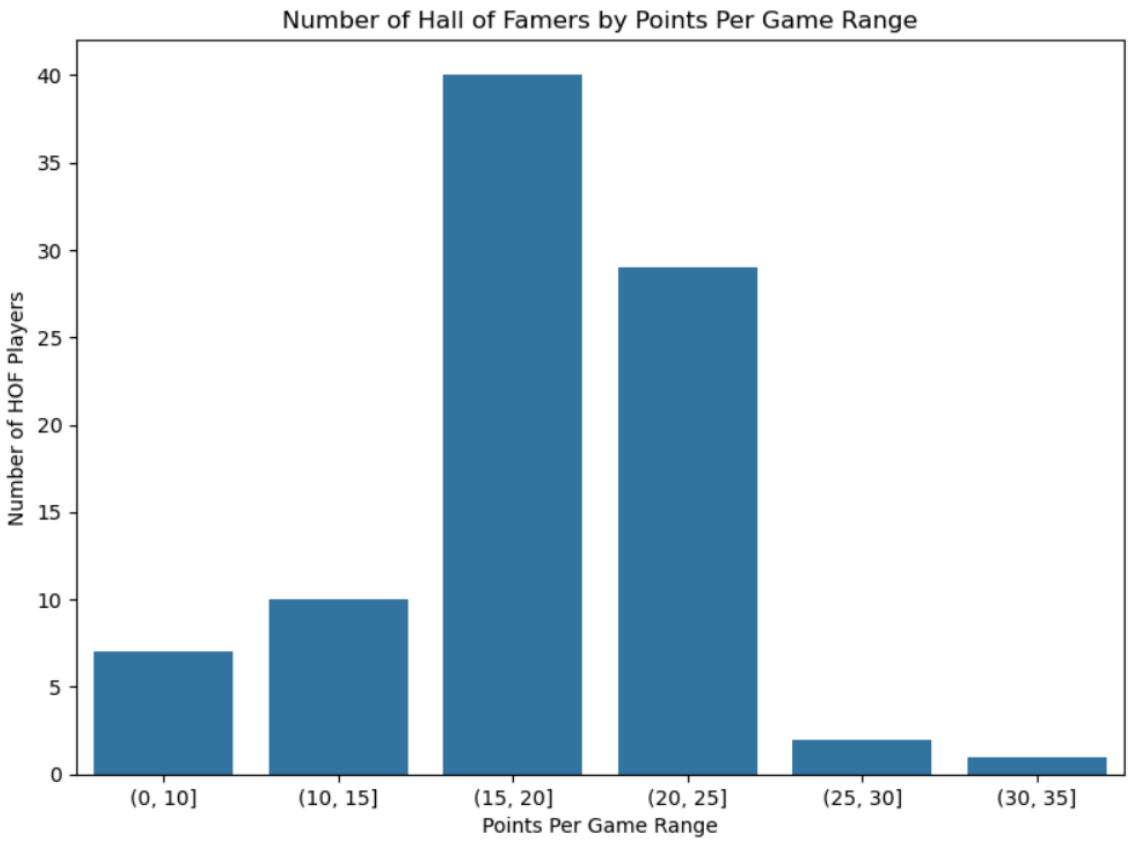
NBA.dropna(axis=0, inplace=True)

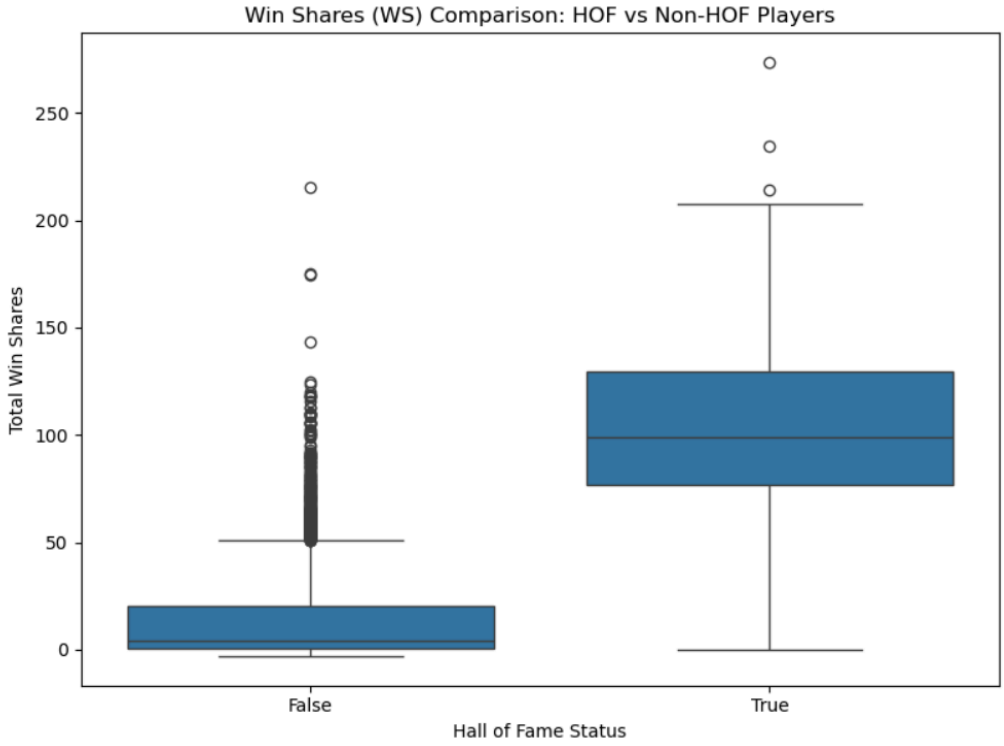
NBA['HOF'].replace({'False':0, 'True':1}, inplace=True)

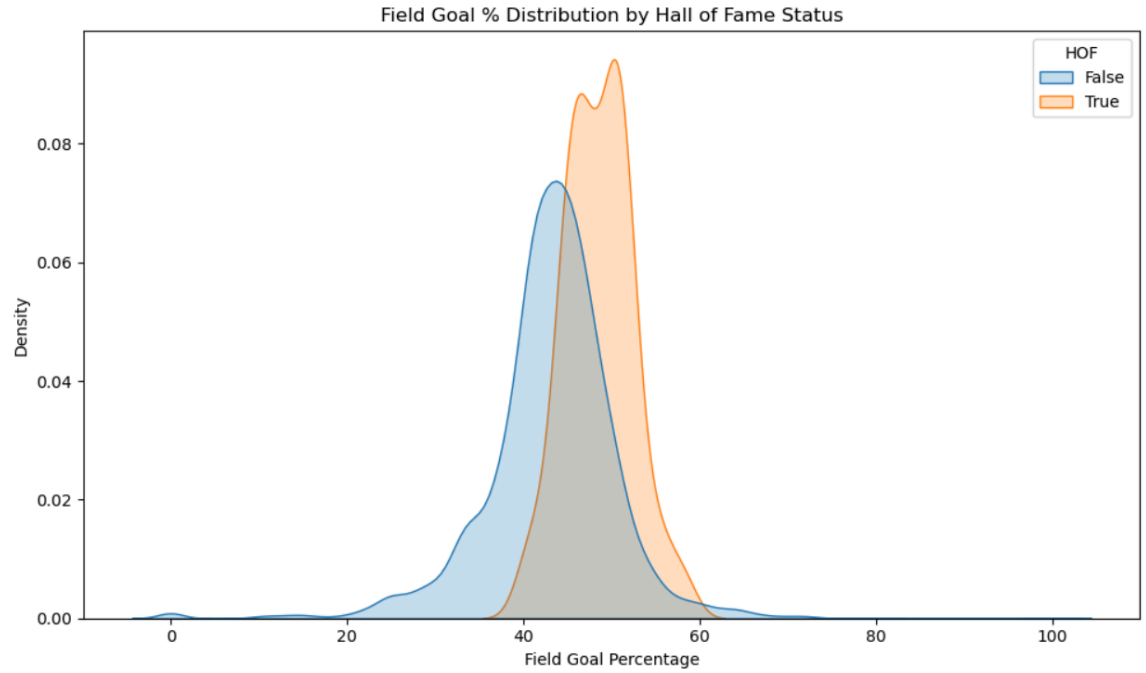
NBA.head()

5. GRAPH YOUR DATA. You need to provide at least 5 different graphs of your data in your report. Include the graphs in both your code and your report. Make sure all graphs are properly labelled with an x and y axis as well as a title. Also add a brief description about what each graph is representing in the report.




Number of Hall of Famers by Points Per Game Range
Number of HOF Players
0
5
10
15
20
25
30
35
40
(0, 10]
(10, 15]
(15, 20]
(20, 25]
(25, 30]
(30, 35]
Points Per Game Range


Win Shares (WS) Comparison: HOF vs Non-HOF Players
Total Win Shares
0
50
100
150
200
250
False
True
Hall of Fame Status



6. TRAIN AND TEST YOUR ALGORITHM. Use the data you collected to train your algorithm with a logistic regression. Make sure to split your data into a training and testing set of appropriate sizes. Include the code (documented) you used to train your data in your report.

```
features = ['G', 'PTS', 'TRB', 'AST', 'FG%', 'FG3%', 'FT%', 'eFG%', 'WS']
X = NBA[features]
y = NBA['HOF']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

model = LogisticRegression(max_iter=1000)
model.fit(X_train, y_train)


y_pred = model.predict(X_test)

print("Accuracy:", accuracy_score(y_test, y_pred))
print("Confusion Matrix:\n", confusion_matrix(y_test, y_pred))
print("Classification Report:\n", classification_report(y_test, y_pred))
```

✓ 0.6s

```
Accuracy: 0.9860681114551083
Confusion Matrix:
 [[630   2]
 [  7   7]]
Classification Report:
              precision    recall  f1-score   support

       False       0.99      1.00      0.99       632
        True       0.78      0.50      0.61        14

    accuracy                           0.99       646
   macro avg       0.88      0.75      0.80       646
weighted avg       0.98      0.99      0.98       646
```

7. EVALUATE YOUR MODEL. Provide an evaluation of your model. Use what you have learned in class to decide if the model is well trained based on your test data. If your model is not within the 70%-90% accuracy range explain why you think it is not accurate. Discuss ideas of how you can improve the accuracy of your model.

```
Accuracy: 0.9860681114551083
Confusion Matrix:
 [[630   2]
 [  7   7]]
Classification Report:
              precision    recall  f1-score   support

       False       0.99      1.00      0.99       632
        True       0.78      0.50      0.61        14

    accuracy                           0.99       646
   macro avg       0.88      0.75      0.80       646
weighted avg       0.98      0.99      0.98       646
```

8. ANSWER YOUR QUESTION AND CONCLUSION. Now that you have trained and tested your algorithm, try to answer your question from part 3. Enter data relevant to the question you asked. Use this part to also provide a conclusion to your report.

THEREFORE Chris Paul will make the HOF