Jeffrey Dean

Python Methodologies Final Project

UFC Statistics and Methodologies Walkthrough

Introduction:

As discussed in the project brief, I will be developing deliverables that connect data science methodologies to real life occurrences. This walkthrough will elaborate more thoroughly on my methods, the result and a copy of the output visuals that will support the analysis. From there, I will deliver my conclusions and reflect on further efforts that can be developed to expand this effort in the future. This work will accompany a set of PowerPoint slides for a presentation and the code written in python.

Hypothesis:

I support that out of 20 + provided attributes of a fighter in the UFC, statistical analysis can be used to find 5 attributes that are the most significant to determine whether or not a fighter wins a fight.

Procedure:

The code and analysis will include four parts that can support my conclusion to the hypothesis presented.

Code portions:

1) Basis statistics (mean, standard deviation, 95% Confidence Interval)
2) ANOVA analysis (t-testing, f-testing)
3) Visual graphics
4) Fighter Profile

Benchmark:

5% statistical significance level (t-test, f-test)

Initial Factors to Explore:

1. Submission attempts
2. Takedown accuracy percentage
3. Total striking accuracy
4. Total rounds fought
5. Significant strike accuracy

6. Win by TKO/KO
7. Win by Unanimous Decision
8. Height Advantage
9. Weight Advantage
10. Stance
11. Reach Advantage
12. Longest Win Streak
13. Age Advantage
14. Ground Attempts Landed
15. Body Shot Accuracy

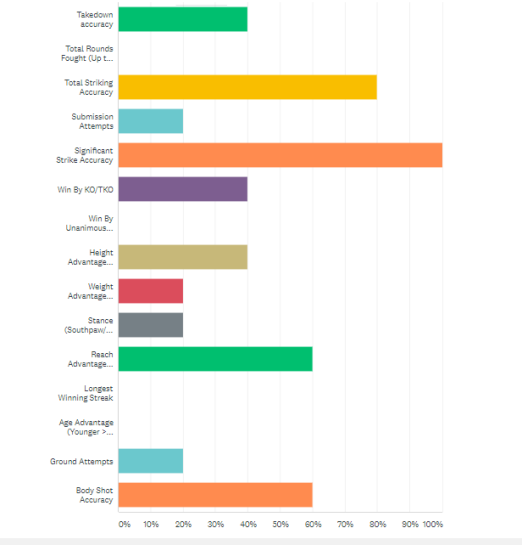
These three parts provide varied information to give a robust picture of what is occurring. I will begin with 15 prominent statistics that have been collected for review and I will narrow this down to 5 statistics that have the most correlation to success. Basic statistics and ANOVA analysis will be developed in addition to linear regression models. Each factor will be measured against the win percentages.

The model will use a benchmark p-value of 5% for the ANOVA portion of the tests. 5% as a p-value criterion is in line with industry standards. Similar measures of simple statistics such as median, mode, and variance are not a necessary extension for my work as it does not further a conclusion than what otherwise would exist.

As a fun extension, I will also develop a poll that asks consumers of UFC content what they believe the 5 most significant factors are amongst the same list of 15 attributes and the poll results will be compared to the actual findings.

Results:

First, I viewed the poll created to see what the layman believes most impacts the fight performance in UFC. The data has a small sample size, but it was good to see how most people measure these results.



Some of the data matches what I found to be valuable, but is dispersed enough not to see any discernable pattern. Just useful to see that there is no consensus among fans.

Next, I found the five factors I wanted to explore further as they had the highest correlation to win percentages. The output in the code had 10 other factors with lower correlations between variables.

```
Takedown Accuracy Correlation to Winning
[[1.         0.24212566]
 [0.24212566 1.        ]]
Unanimous Win Correlation to Winning
[[1.         0.24392116]
 [0.24392116 1.        ]]
Total Win by KO or TKO Correlation to Winning
[[1.         0.21973337]
 [0.21973337 1.        ]]
Total Win Streak Correlation to Winning
[[1.         0.46653342]
 [0.46653342 1.        ]]
Ground Attempt Percentage Correlation to Winning
[[1.         0.21217688]
 [0.21217688 1.        ]]
```

Below is the result of simple statistics and confidence intervals for each of the factors I found of value.

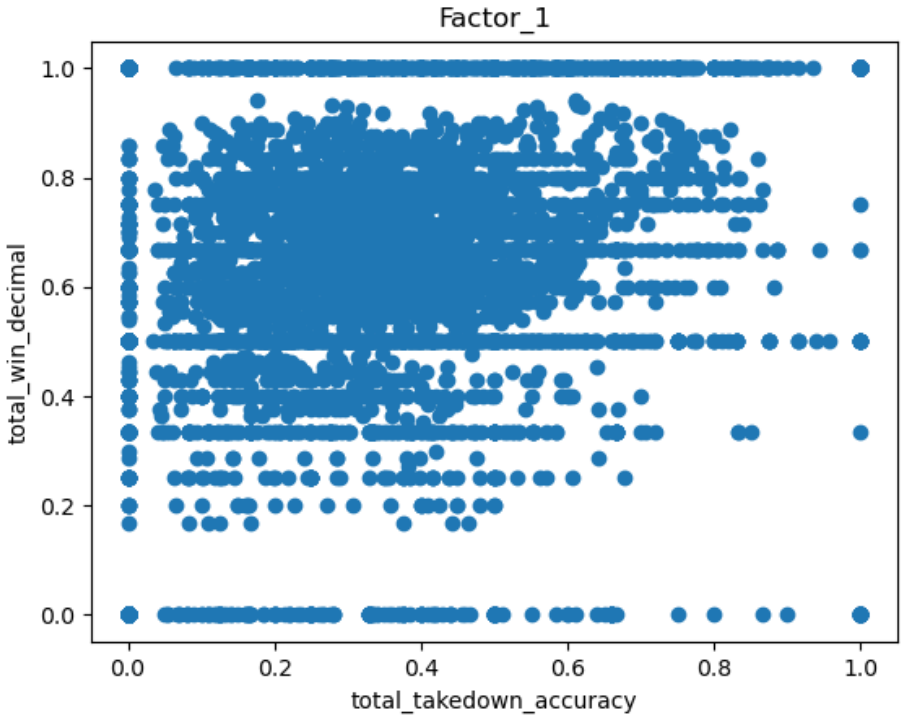
```
Mean Of Each Factor

total_takedown_accuracy    0.324142
total_win_unanimous        1.380492
total_win_tko_ko           1.485510
total_winning_streak       2.600187
ground_attempts_pct        0.627433
dtype: float64

Standard Deviation of Each Factor

total_takedown_accuracy    0.232974
total_win_unanimous        1.630646
total_win_tko_ko           1.898350
total_winning_streak       1.966609
ground_attempts_pct        0.227711
dtype: float64


95% Confidence Interval Lower and Upper Bound

total_takedown_accuracy    0.316081
total_win_unanimous        1.324073
total_win_tko_ko           1.419827
total_winning_streak       2.532143
ground_attempts_pct        0.619555
dtype: float64

total_takedown_accuracy    0.332203
total_win_unanimous        1.436912
total_win_tko_ko           1.551192
total_winning_streak       2.668231
ground_attempts_pct        0.635312
dtype: float64
```

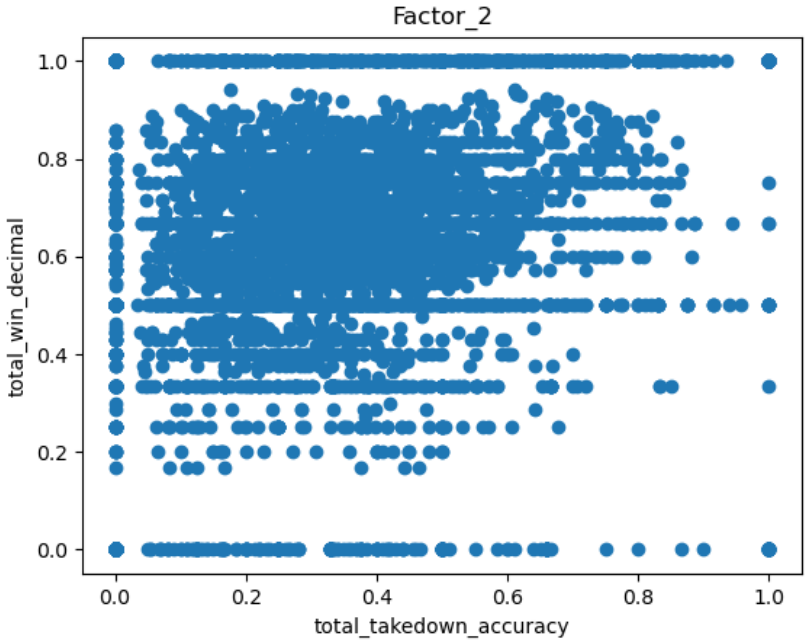
Further below is the ANOVA table and regression results that create a clearer interpretation. Each individual OLS model shows a similar picture but the combined model is more important to show impact of interactions between the five factors. Overall, each t-statistic and the F-statistic are very significant at any level. This means this confirms that these five factors to be valuable to finding a winning UFC fighter.

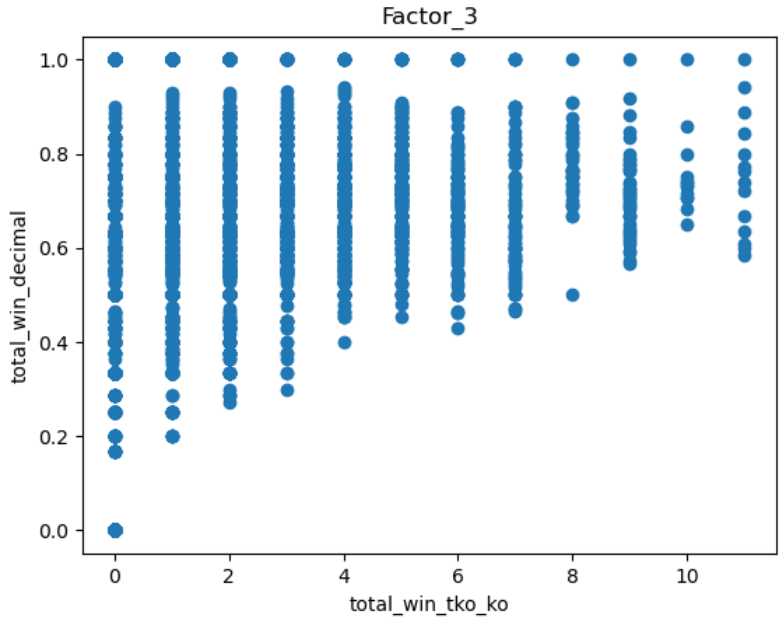
Combined OLS Model

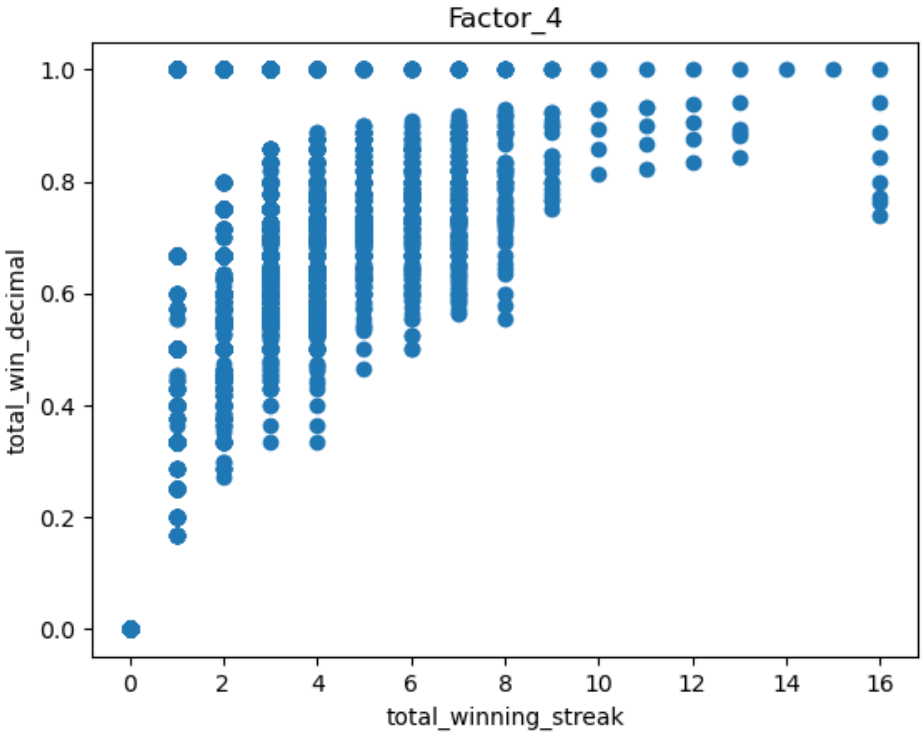
	sum_sq	df	F	PR(>F)
ground_attempts_pct	5.558939	1.0	104.209280	2.789701e-24
total_winning_streak	57.360810	1.0	1075.300325	3.579836e-218
total_win_tko_ko	2.978205	1.0	55.830188	8.954053e-14
total_win_unanimous	2.993681	1.0	56.120313	7.735462e-14
total_takedown_accuracy	13.060742	1.0	244.839991	3.447325e-54
Residual	342.041666	6412.0	NaN	NaN

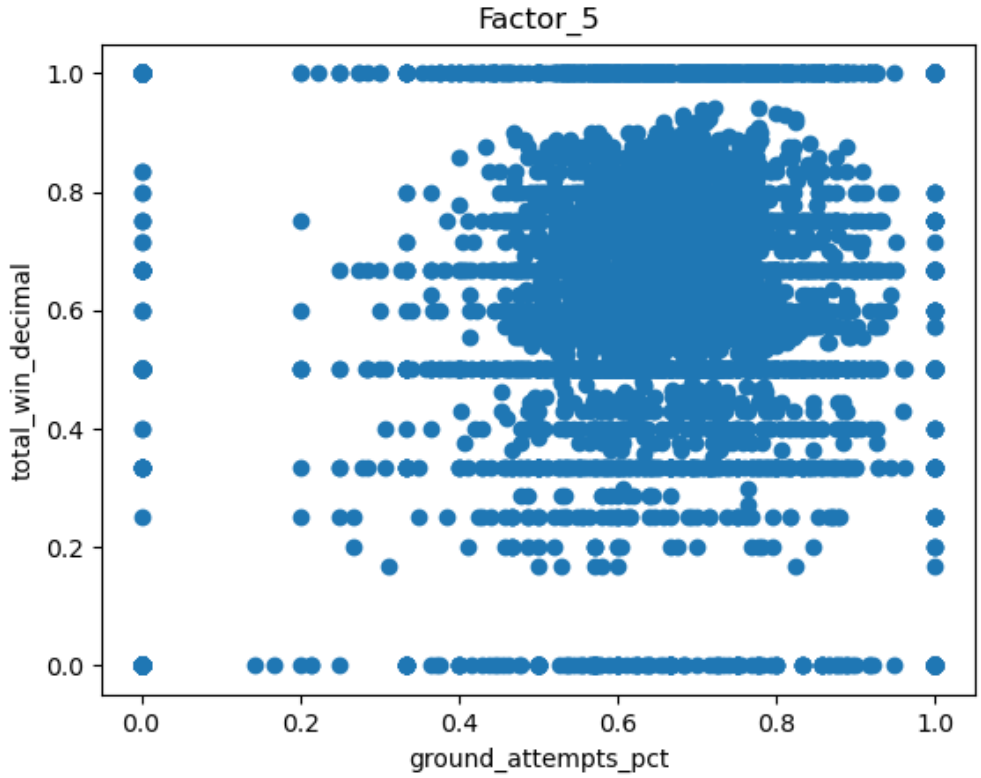
Further below we see the graphs of each factor, there is no trend line as there is more of a logistic curve rather than a linear relationship.

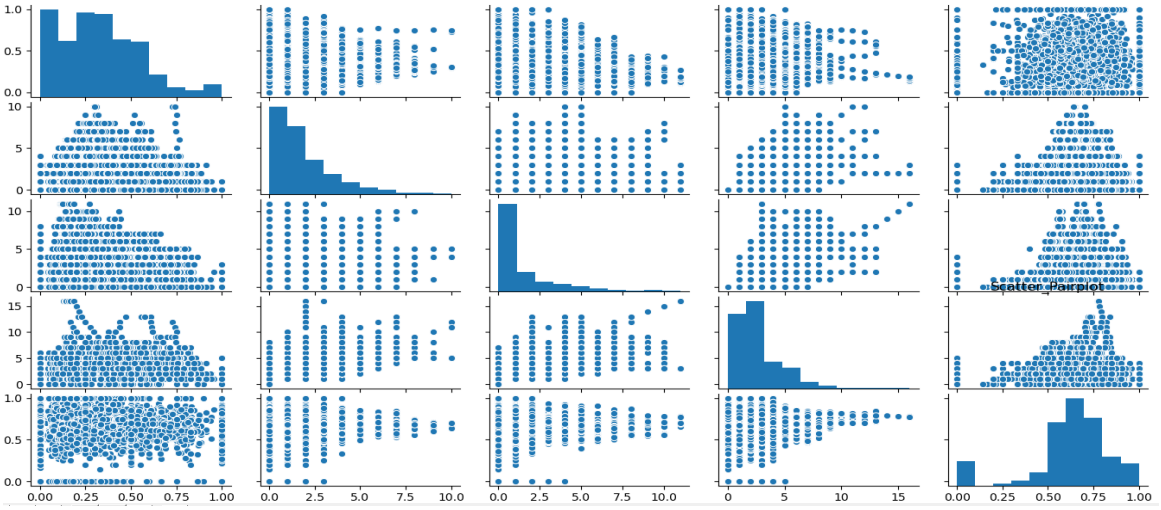



Factor_2
total_win_decimal
total_takedown_accuracy
0.0
0.2
0.4
0.6
0.8
1.0


Factor_3
total_win_decimal
total_win_tko_ko
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10


Factor_4
total_win_decimal
total_winning_streak
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
12
14
16


Factor_5
total_win_decimal
ground_attempts_pct
0.0
0.2
0.4
0.6
0.8
1.0



The data is rather messy, but this is due to unexplored factors and variance in data that can occur. A good fighter could change their style of game, or just have strong seasons and later weak ones. Differences in weight class, age, and gender can occur so further analysis can gauge the nuances of these factors in a later study.

The last portion shows how a user can put data into the code and save the data for later use. Sites that track specific fighters allow this so a user can see only data that applies to what they want to know. The tools for this are below.

```
class Basic(object):
    def __init__(self, name, gender, age):
        self.name = name
        self.gender = gender
        self.age = age

    def getName(self):
        return self.name

    def getGender(self):
        return self.gender

    def getAge(self):
        return self.age
```

```
class Performance(object):
    def __init__(self, height, weight, stance, wins, losses, ties):
        self.height = height
        self.weight = weight
        self.stance = stance
        self.wins = wins
        self.losses = losses
        self.ties = ties

    def getHeight(self):
        return self.height

    def getWeight(self):
        return self.weight

    def getStance(self):
        return self.stance

    def getWins(self):
        return self.wins

    def getLosses(self):
        return self.losses

    def getTies(self):
        return self.ties
```

```
class Fighter(Basic, Performance):
    def getProfile(self):
        print("Name :", self.name)
        print("Gender :", self.gender)
        print("Age :", self.age)
        print("Height :", self.height)
        print("Weight :", self.weight)
        print("Stance :", self.stance)
        print("Wins :", self.wins)
        print("Losses :", self.losses)
        print("Ties :", self.ties)

    def isFighter(self):
        if self.name == name:
            return TRUE
        else:
            return FALSE
```

Conclusions:

Given the above code and visuals, we can see there is significance for five factors out of the total 15 and the graphics support a positive trend with variance in the spread. There is more that can be explored by subdividing the factors given or by introducing other factors that may be useful. The subsequent code for Fighter profile will be a good, flexible tool that is in the hands of the user. Overall, the polling results were not wrong but showed that it is hard to gauge what factors matter the most. Unlike product data, sports have no easy relationships to discern. Fighters evolve and trends in training camps can change what works over time. The data is from 1993 to 2019 and the UFC has come a long way.