Inferential Statistical Analysis

This section loos at statistical significance on observations made and thoughts about the findings during the EDA section. This is an essential step to understanding whether the differences between Match Winners and Losers are unlikely to have happened by chance alone.

The focus lies primarily on three categories: 1) Distribution between Match Winners and Match Losers, 2) Correlation with the Target Variable (i.e., Win or Non-Win), and 3) Collinearity between Features.

Challenges

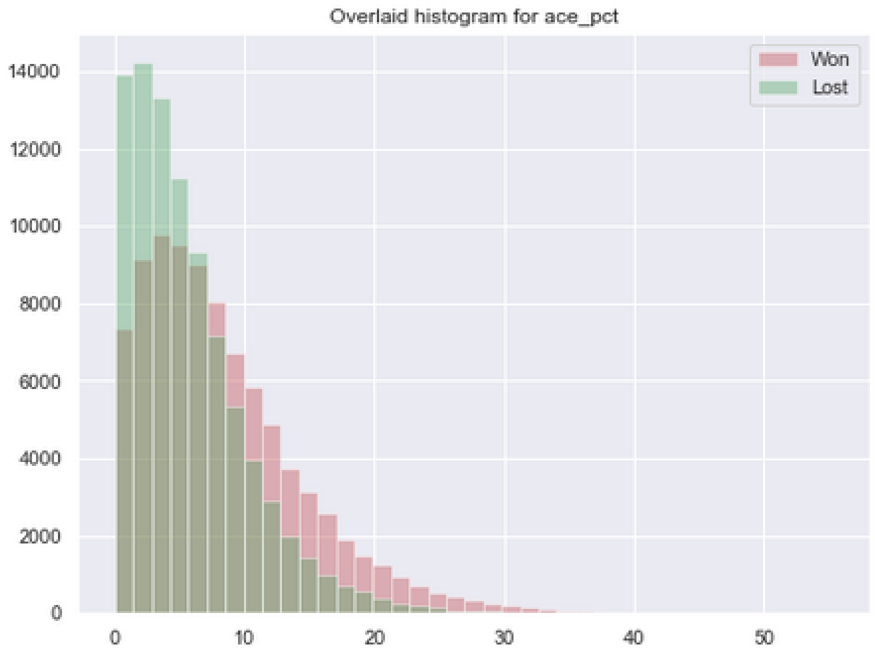
Normality. A challenge for this project was the lack of normally distributed data (see Q-Q plots below). Normally distributed data are often a requirement for classical statistical tests. Fortunately, the Central Limit Theorem allows us to use the Z-test to compare sample distribution differences. (Central Limit Theorem states that the sampling distribution of the sample means approaches a normal distribution as the sample size gets larger — no matter what the shape of the population distribution. This is especially true for sample sizes over 30. In other words, as you take more samples, especially large ones, the graph of the sample means will look more like a normal distribution.)

Statistical and Practical Significance. Due to the relatively large sample sizes, even small differences can be considered statistically significant, but may not allow us to use a particular feature since this may put a limitation on that feature's predicative qualities.

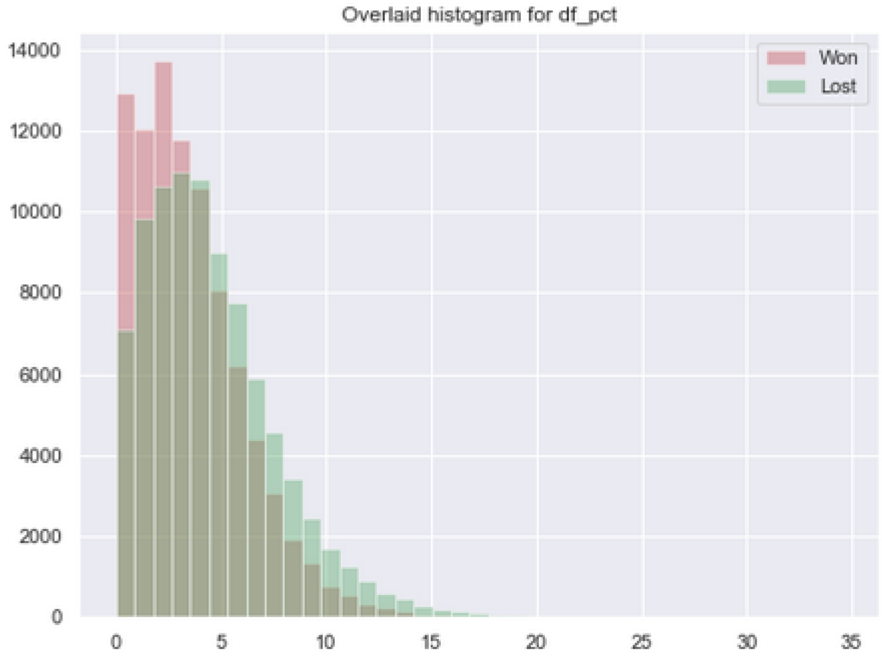
Continuous Features. In many ways, prediction match winners is about finding the subtle differences and similarities between Wins and Non-Wins. Continuous features tend to be much more valuable in uncovering those differences and trends across time. This machine learning model only uses continuous features.

Distributions. To understand whether the differences between Wins and Non-Wins observed are significant, I conducted Z-tests for the distributions on the following: Ace Percent, Double Fault Perfect, First Serves In Percent, First Serve Points Won Percent, Second Serve Points Won Percent, Service Points Won Percent, Return Points Won Percent, Break Points Won Percent and Total Points Won Percent.

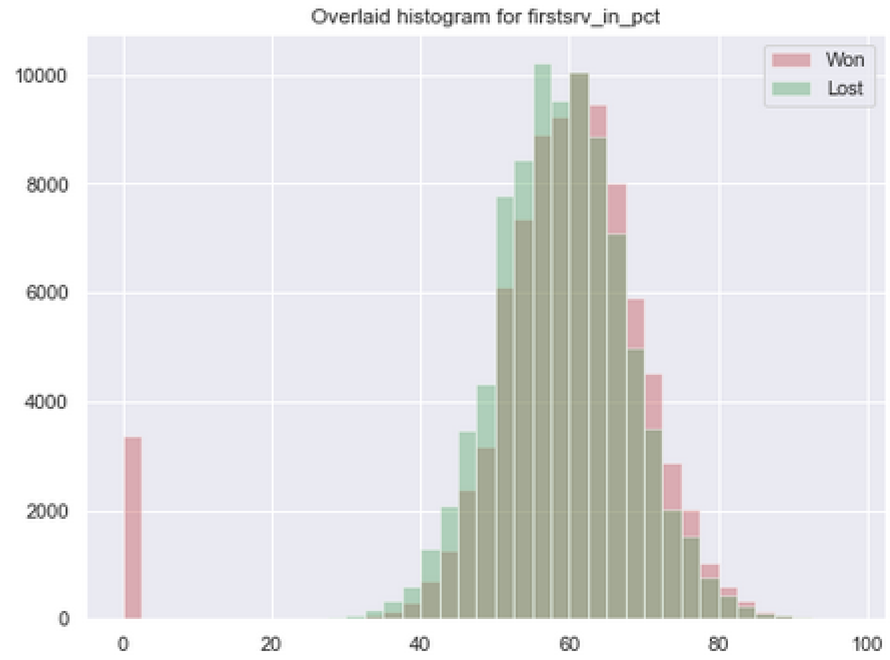
The method used was to compare mean differences across 10,000 permutations and then check whether the mean difference of the observed distributions would fall into the realm of significant possibilities. The results are shown below.



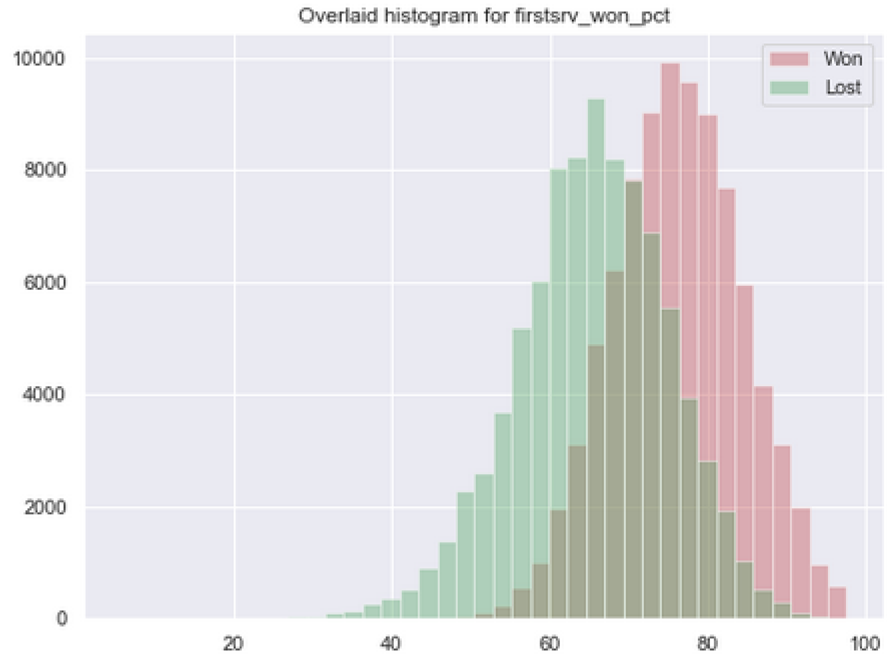
```
Wins Mean:
8.340683934072723
No-Wins Mean:
5.599402213071454
Mean Diff:
2.741281721001269
H0 Diff: 0

p: 0.0
CI: [-0.05431697
0.05277014]
ME:
0.053402597518318703
```



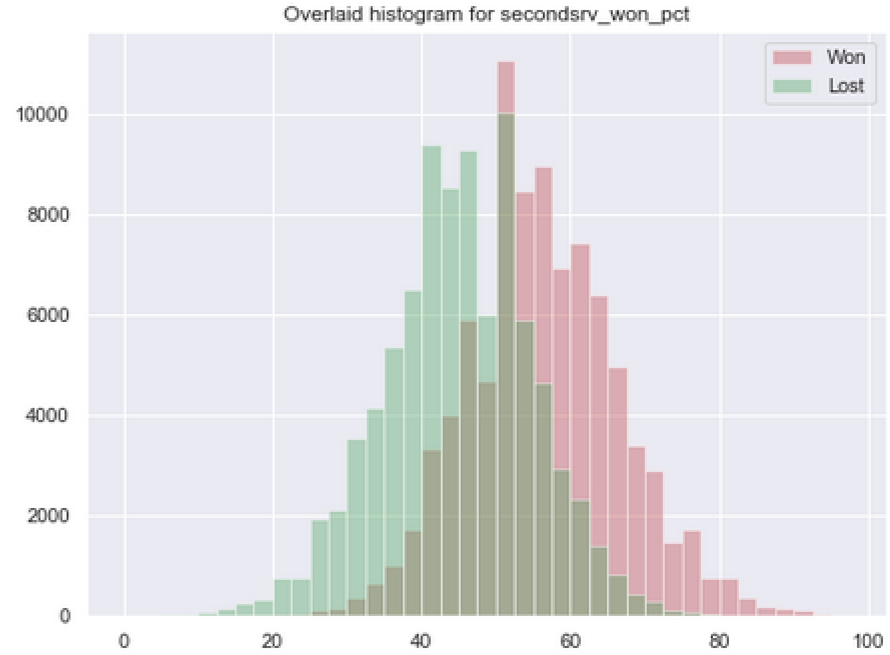
```
Wins Mean:
3.471326834959865
No-Wins Mean:
4.511129954525473
Mean Diff: -
1.0398031195656081
H0 Diff: 0

p: 0.0
CI: [-0.02795768
0.02747843]
ME: 0.0275627659420
```



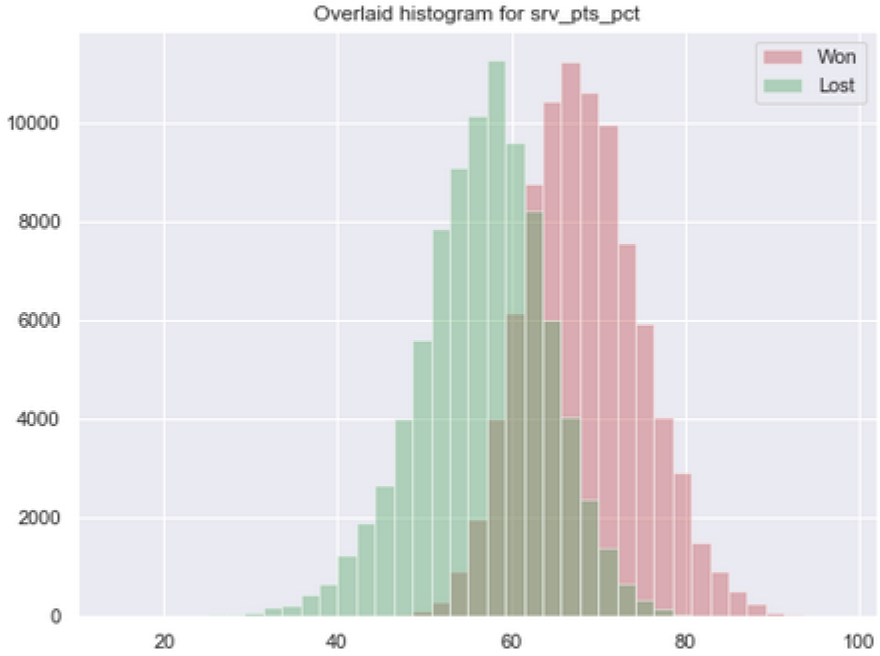
```
Wins Mean:
58.35125231156482
No-Wins Mean:
58.8646494896052
Mean Diff: -
1.0398031195656081
H0 Diff: 0

p: 0.0
CI: [-0.11174051
0.10695299]
ME:
0.10793662133669706
```



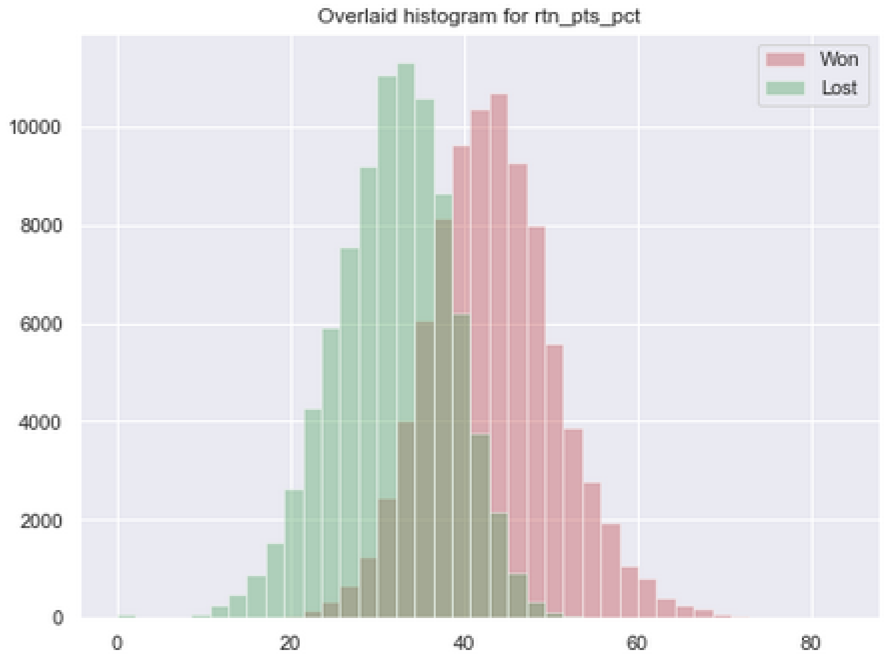
```
Wins Mean:
76.23225218749391
No-Wins Mean:
65.35481512219901
Mean Diff:
10.877437065294899
H0 Diff: 0

p: 0.0
CI: [-0.09521021
0.09905118]
ME:
0.09875387007053657
```



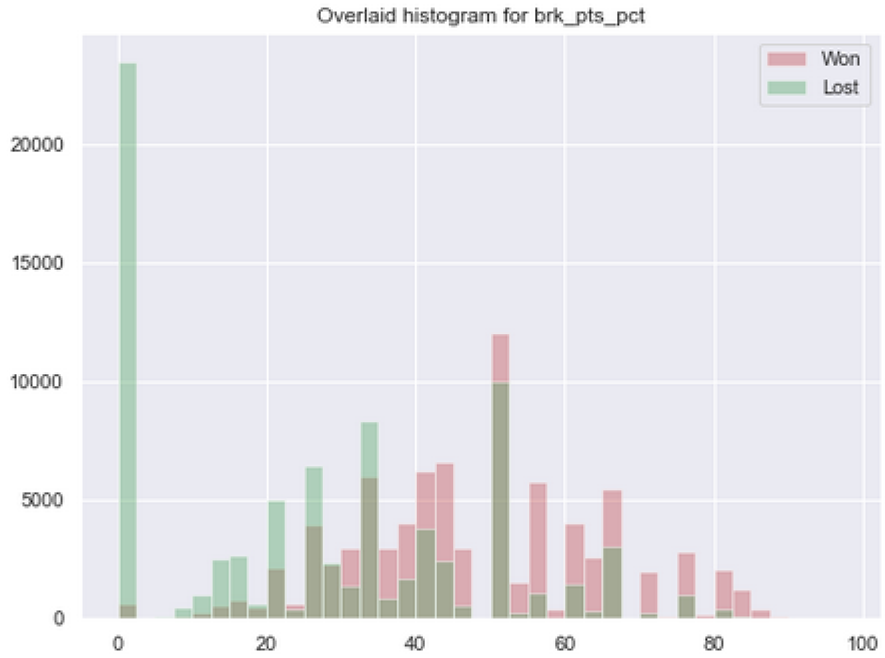
```
Wins Mean:
56.11935795826883
No-Wins Mean:
44.72161921208218
Mean Diff:
11.397738746186647
H0 Diff: 0

p: 0.0
CI: [-0.11144302
0.10846935]
ME:
0.10890684540854724
```



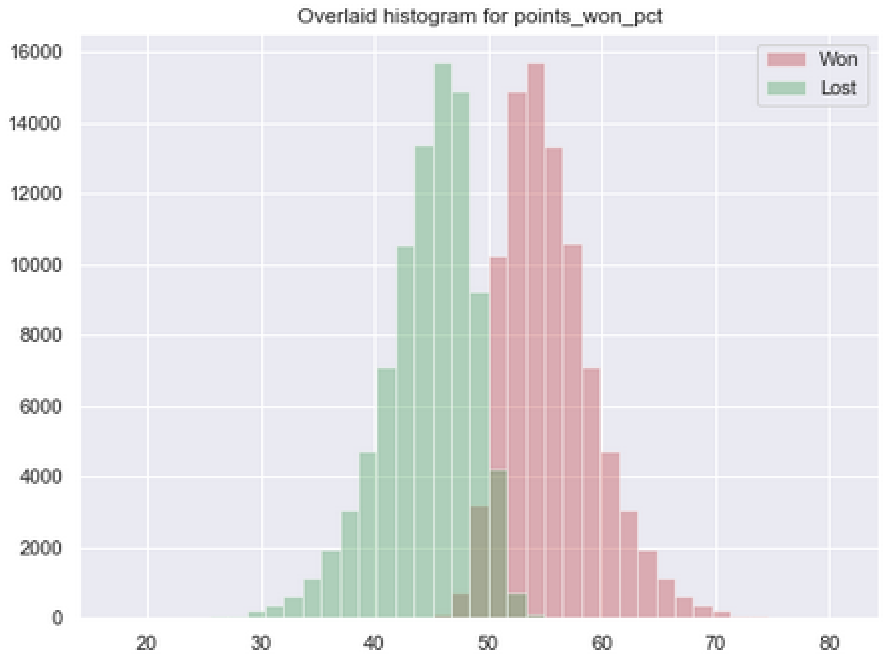
```
Wins Mean:
68.22250043301648
No-Wins Mean:
56.694481221203695
Mean Diff:
11.528019211812783
H0 Diff: 0

p: 0.0
CI: [-0.08612867
0.08654083]
ME:
0.0869216829786314
```



```
Wins Mean:
43.3193882825373
No-Wins Mean:
31.80259127897296
Mean Diff:
11.516797003564335
H0 Diff: 0

p: 0.0
CI: [-0.08571761
0.08432219]
ME:
0.08365446555536553
```



```
Wins Mean:
49.53499443350504
No-Wins Mean:
31.58515297066974
Mean Diff:
17.949841462835302
H0 Diff: 0

p: 0.0
CI: [-0.2415653
0.24239284]
ME:
0.24322333891412118
```



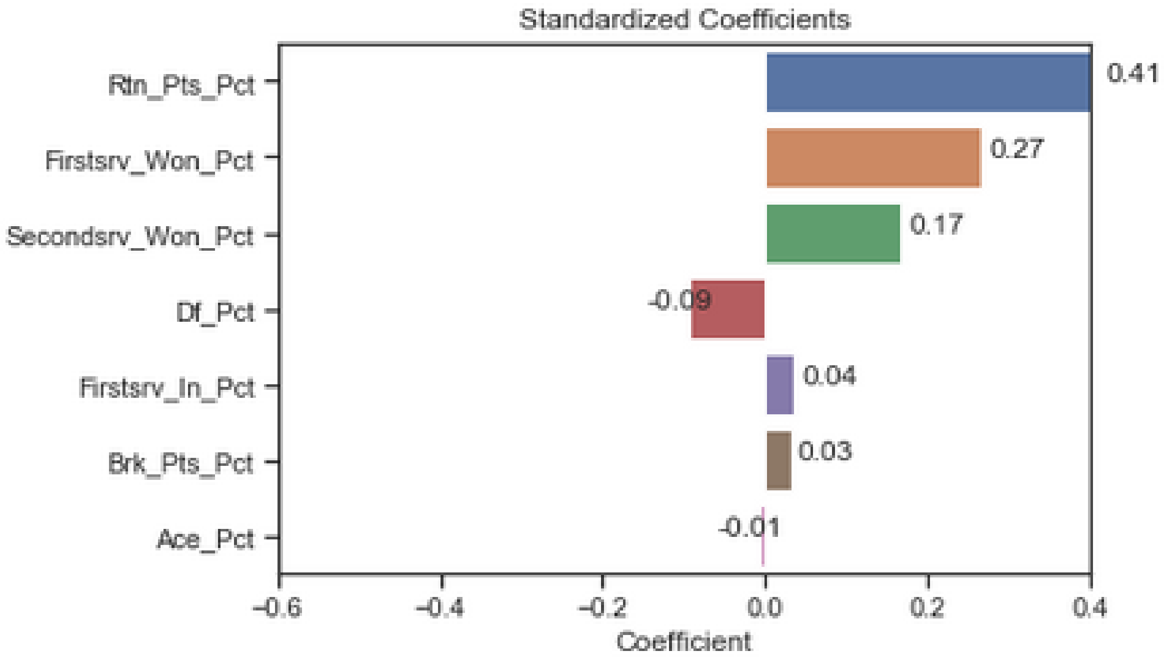
```
Wins Mean:
55.467522564258786
No-Wins Mean:
44.53247743574097
Mean Diff:
10.935045128517814
H0 Diff: 0

p: 0.0
CI: [-0.06302145
0.06394967]
ME:
0.06362357961431575
```

For all features above, the probability (p-value) was <0.001, allowing me to reject the null hypothesis (H0) that Wins and Non-Wins were equally distributed for the respective feature. In the next step I'm going to look at the correlation of these features with each other and with the target variable.

Please note that conducting a multitude of Z-tests increases the likelihood of a Type 1 Error. Unfortunately, use of ANOVA or Chi-squared is not applicable when using continuous data to predict a binary target. ANOVA would be helpful if we had a multitude of categorical data and a continuous target variable while Chi-squared allows to compare categorical data.

Correlation. As pointed out above, exploring relationships between continuous variables with binary outcomes comes with a few challenges especially when we're attempting to use popular statistical tools (Pearson's R, ANOVA etc.). Due to this issue for this project we've used logistic regression instead.

As mentioned earlier, because Total Points Won can be broken into Serve Points and Return Points Won, I removed it from the correlation table. Likewise, Serve Points Won can be broken into First and Second Serve Points and was removed for the same reason.



Using Logistic Regression's Beta based on standardized values allowed us to evaluate the relative importance of the features used. We can see at the top are three features to detect Wins: Return Points Won Percent, First Serve Points Won Percent and Second Serve Points Won Percent. The bottom three features, First Serve In Percent, Break Points Converted Percent and Aces Percent seem to have very little influence on whether a player wins the match. Hence, they will be dropped from the feature list.