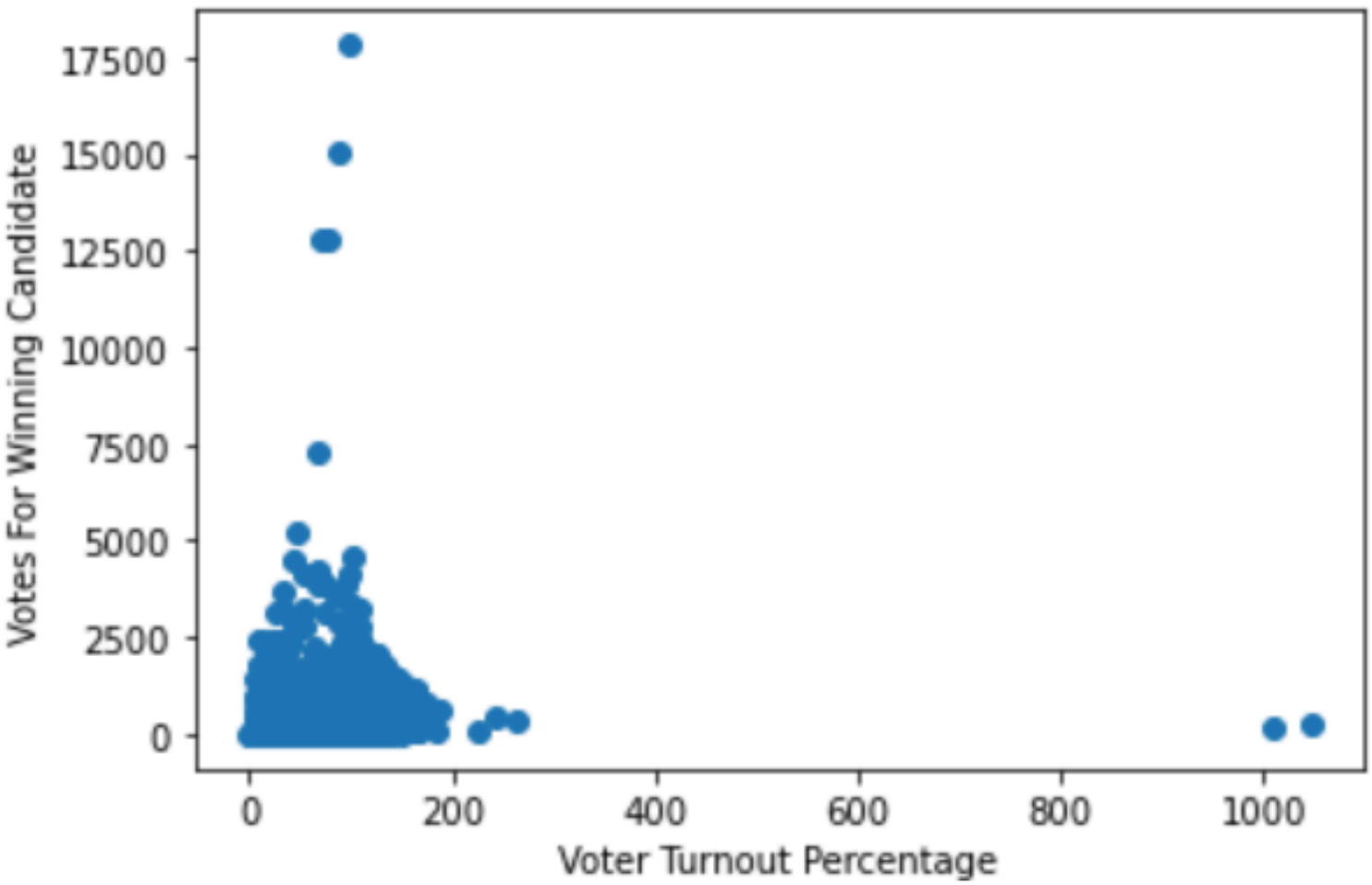
CS 649 Big Data: Tools and Methods
Spring Semester, 2022

Doc 28 Exam Comments
May 3, 2022


17500
15000
12500
10000
7500
5000
2500
0
Votes For Winning Candidate
0
200
400
600
800
1000
Voter Turnout Percentage

```
gpa_gre_filepath=r'C:/Users/xxx/Downloads/gpa-gre/gpa-gre.csv'
russia_elections_2012_filepath1=r'C:/Users/xxx/Downloads/Exam/Russia2012_1of2.xls'
russia_elections_2012_filepath2=r'C:/Users/xxx/Downloads/Exam/Russia2012_2of2.xls'
```

```
path = "/Users/rwhitney/data/"  ## path which should be changed
gre_gpa_path = path + "gpa-gre.csv"
russia_df1_path = path + "Russia2012_1of2.xls"
russia_df2_path = path + "Russia2012_2of2.xls"
```

```
import pandas as pd
import numpy as np
from sklearn.linear_model import LinearRegression
import matplotlib.pyplot as plt
from sklearn.metrics import r2_score
df = pd.read_csv(path_to_GRE_GPA_File)
df_1 = pd.read_excel(path_to_Russian_1of2)
df_2 = pd.read_excel(path_to_Russian_2of2)
```

```
df = pd.concat([df_1,df_2], ignore_index=True)
```

2.What is the difference between a Spark transformation and a Spark action.

Answer:

Spark transformation: Spark transformation is a function that produces new RDD from an existing RDD. It creates new RDD when transformation applied. It returns dataset, dataframe or RDD only.

Spark action: Spark actions are RDD's operations to execute on a cluster. It returns anything else other than a dataset, dataframe or RDD or nothing at all.

(Resource:Stack overflow)

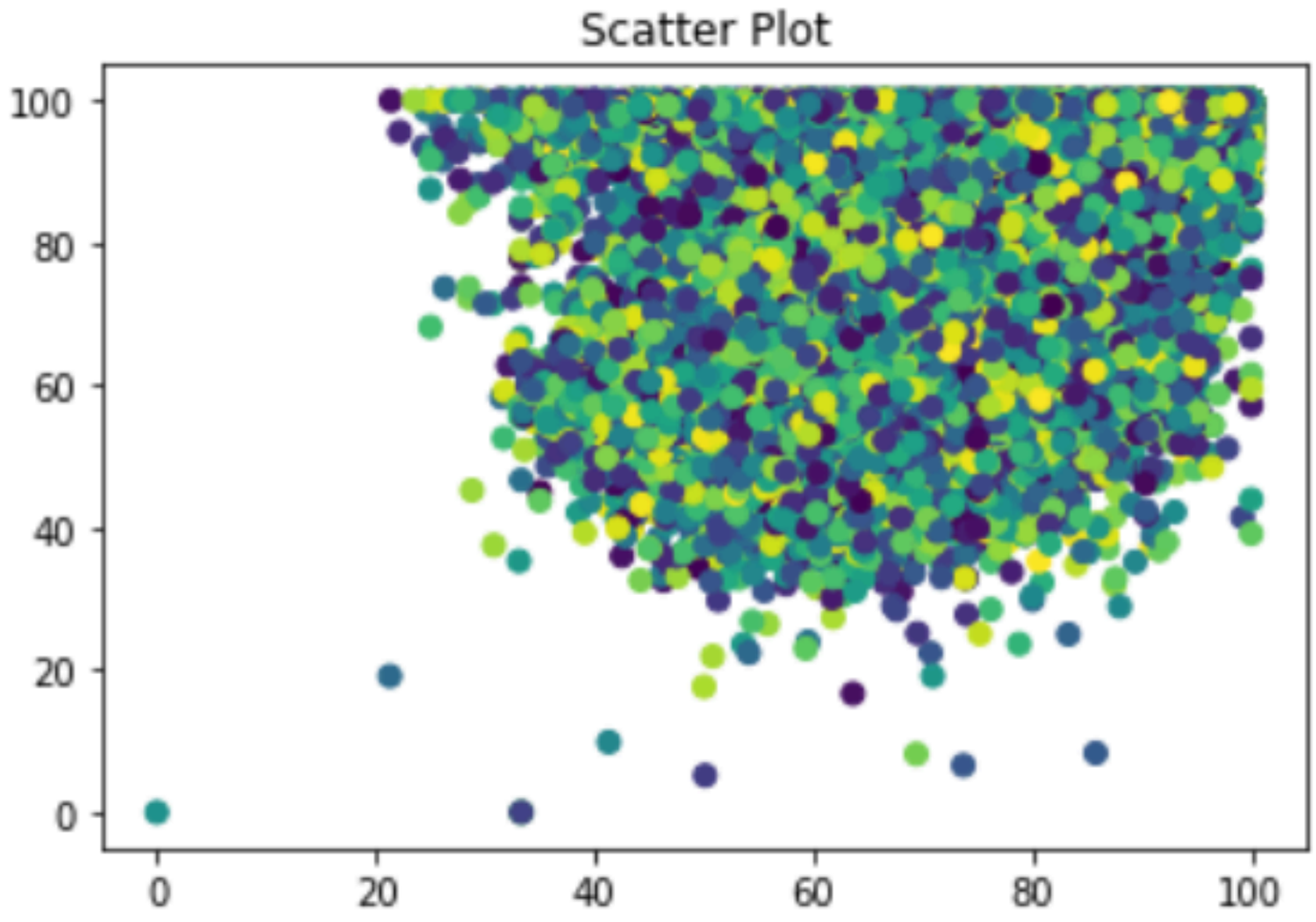
Reference: https://stackoverflow.com/questions/43839786/what-is-difference-between-transformations-and-rdd-functions-in-spark

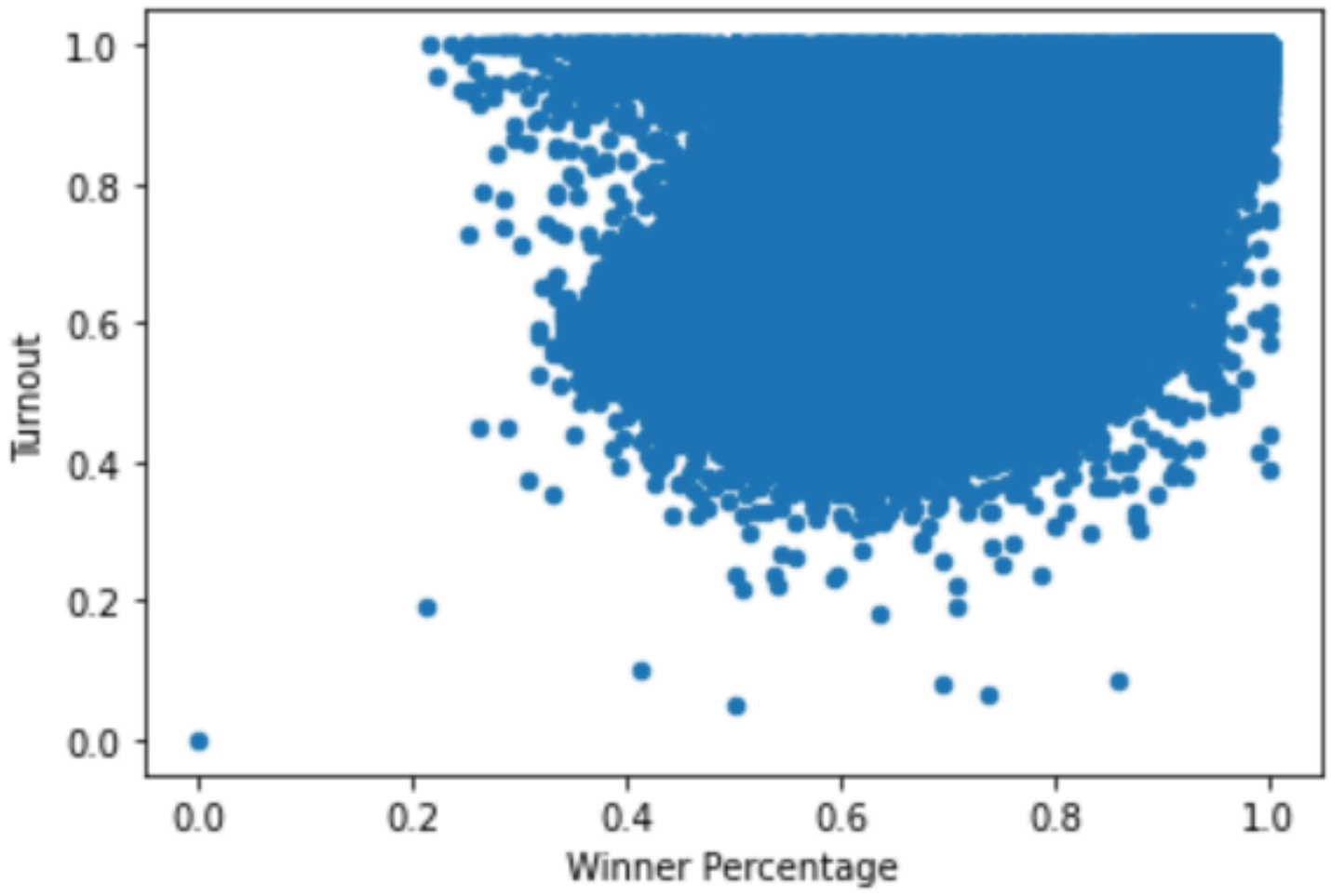
https://medium.com/codex/spark-transformation-and-action-a-deep-dive-f351bce88086

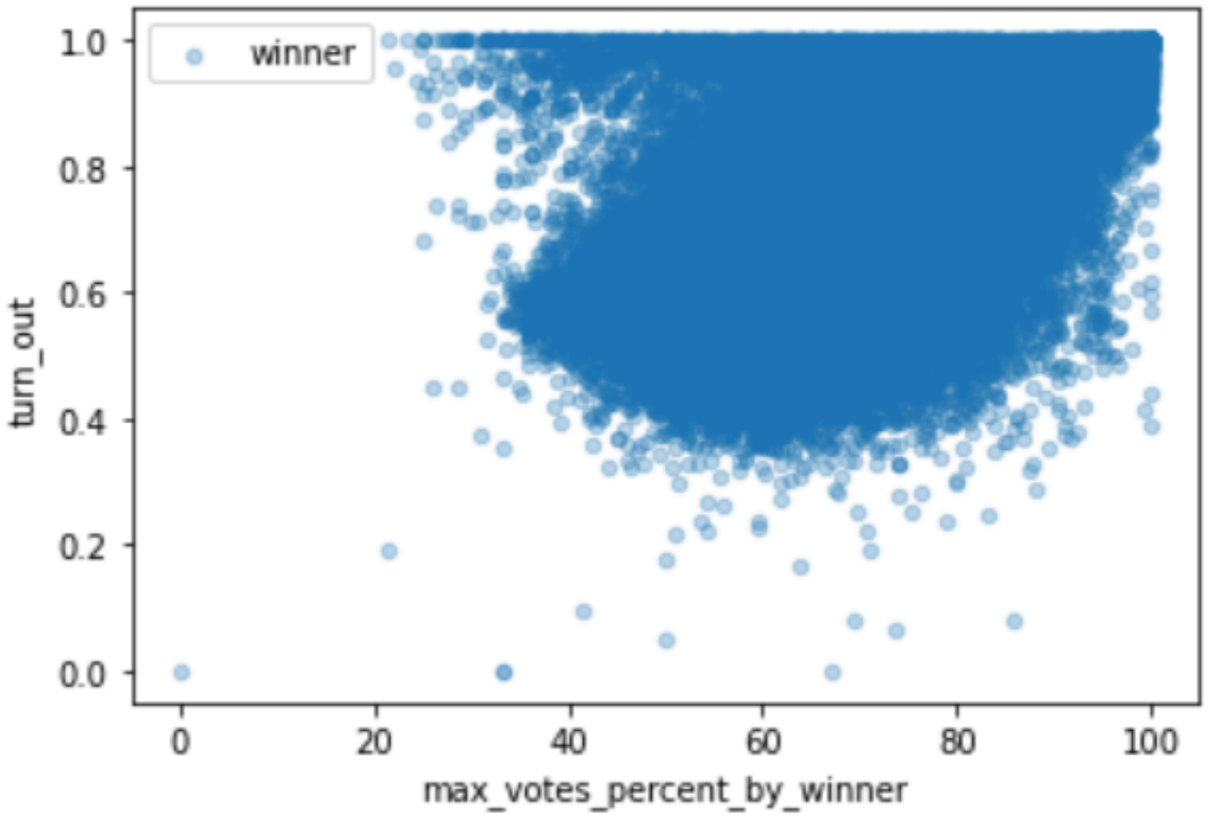
https://sparkbyexamples.com/spark/sparksession-vs-sparkcontext/

https://spark.apache.org/docs/latest/rdd-programming-guide.html

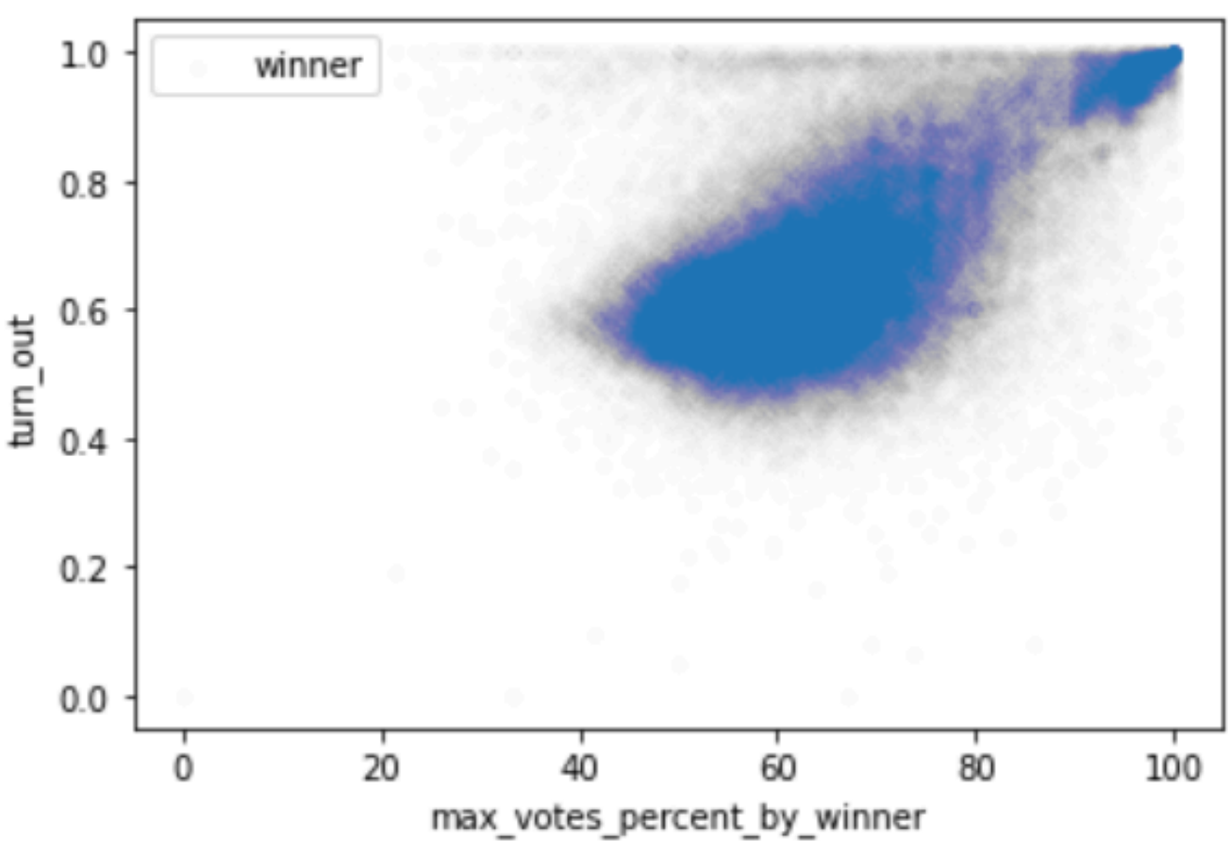
In order to clean the data, replacing the NA values with 0 and the value 'A' with 0. This will help to bring consistency in data and help in proper prediction.


Scatter Plot
100
80
60
40
20
0
0
20
40
60
80
100

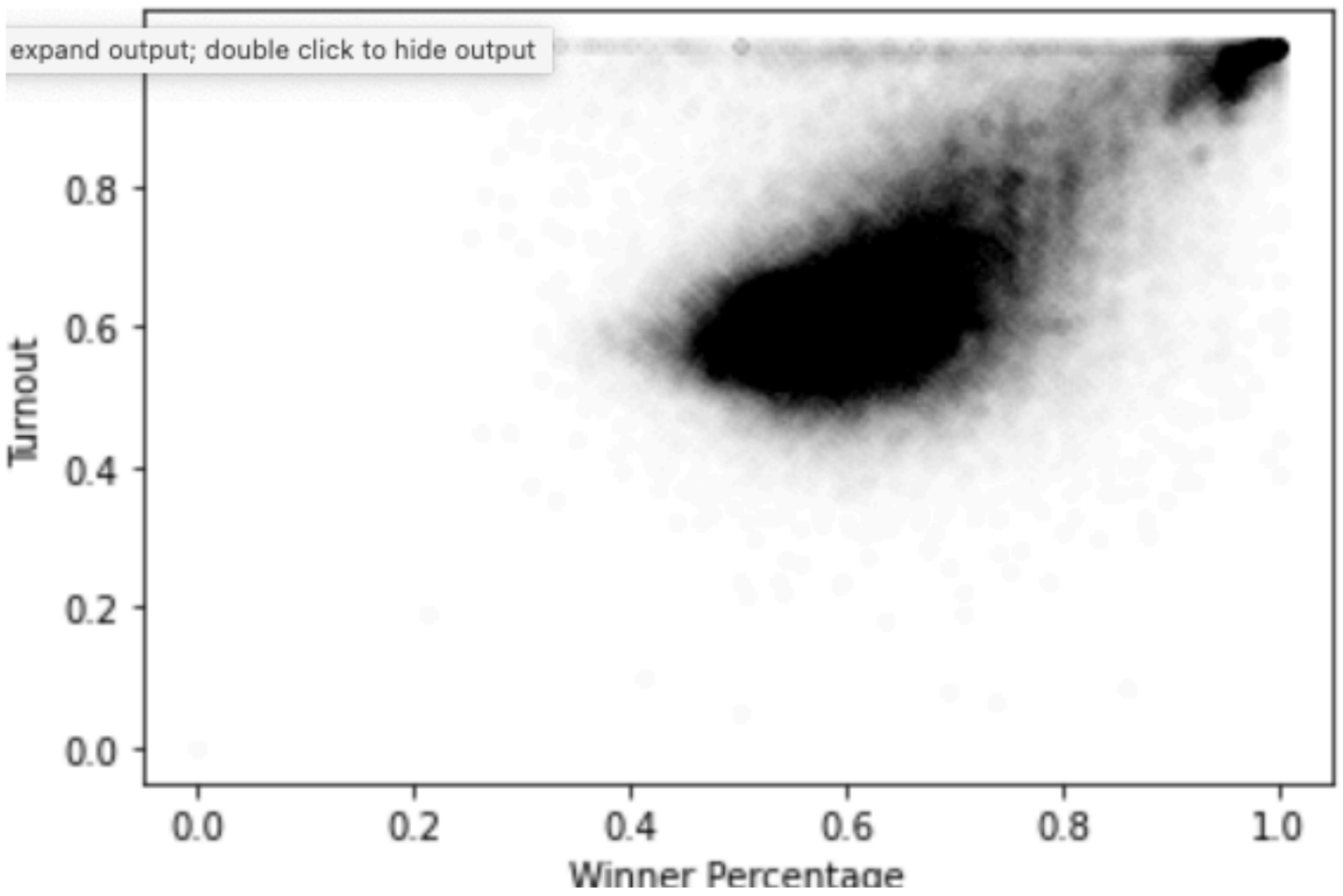

1.0
0.8
0.6
0.4
0.2
0.0
Turnout
0.0
0.2
0.4
0.6
0.8
1.0
Winner Percentage



Alpha 0.3



Alpha 0.002


expand output; double click to hide output
Turnout
0.0
0.2
0.4
0.6
0.8
0.0
0.2
0.4
0.6
0.8
1.0
Winner Percentage