R Notebook

Alex Baker, James Kitch, Jackson Smith, Matthew Sheridan

Introduction:

Data Cleaning, EDA:

```
#DO EDA HERE
data_2017 = clean(data17)
data_2018 = clean(data18)
data_2019 = clean(data19)
data_2020 = clean(data20)
data_2021 = clean(data21)
```

Initial Modeling:

```
lm1 = lm(Score_Dif~., data=data_2019[,seq(3, ncol(data_2019))])

RF1 = randomForest::randomForest(Score_Dif~., data=data_2019[,seq(2, ncol(data_2019))])

step1 <- step(lm(Score_Dif~1, data=data_2019), scope=formula(lm1), direction="forward",trace=0)
```

```
rmse_2021 = c(RMSE(data_2021$Score_Dif, predict(lm1, newdata=data_2021)),
              RMSE(data_2021$Score_Dif, predict(step1, newdata=data_2021)),
              RMSE(data_2021$Score_Dif, predict(RF1, newdata=data_2021)))

rmse_2020 = c(RMSE(data_2020$Score_Dif, predict(lm1, newdata=data_2020)),
              RMSE(data_2020$Score_Dif, predict(step1, newdata=data_2020)),
              RMSE(data_2020$Score_Dif, predict(RF1, newdata=data_2020)))

rmse_2019 = c(RMSE(data_2019$Score_Dif, predict(lm1)),
              RMSE(data_2019$Score_Dif, predict(step1)),
              RMSE(data_2019$Score_Dif, predict(RF1)))

rmse_2018 = c(RMSE(data_2018$Score_Dif, predict(lm1, newdata=data_2018)),
              RMSE(data_2018$Score_Dif, predict(step1, newdata=data_2018)),
              RMSE(data_2018$Score_Dif, predict(RF1, newdata=data_2018)))

rmse_2017 = c(RMSE(data_2017$Score_Dif, predict(lm1, newdata=data_2017)),
              RMSE(data_2017$Score_Dif, predict(step1, newdata=data_2017)),
              RMSE(data_2017$Score_Dif, predict(RF1, newdata=data_2017)))

rmse = data.frame(rmse_2017, rmse_2018, rmse_2019, rmse_2020, rmse_2021, row.names = c("Linear", "Step"
rmse
```

```
##        rmse_2017 rmse_2018 rmse_2019 rmse_2020 rmse_2021
```

```
## Linear  9.992705  9.833446  9.919896  9.924871  10.16560
## Step    9.986854  9.799207  9.939556  9.923551  10.14859
## RF1    10.746631 10.313824 10.905360 10.514303  11.08069
```

Conclusion: