

# Predicting Tennis Match Duration



# Response and Predictor Variables

## Response Variable:

- **Match Duration** – total time of the match in minutes

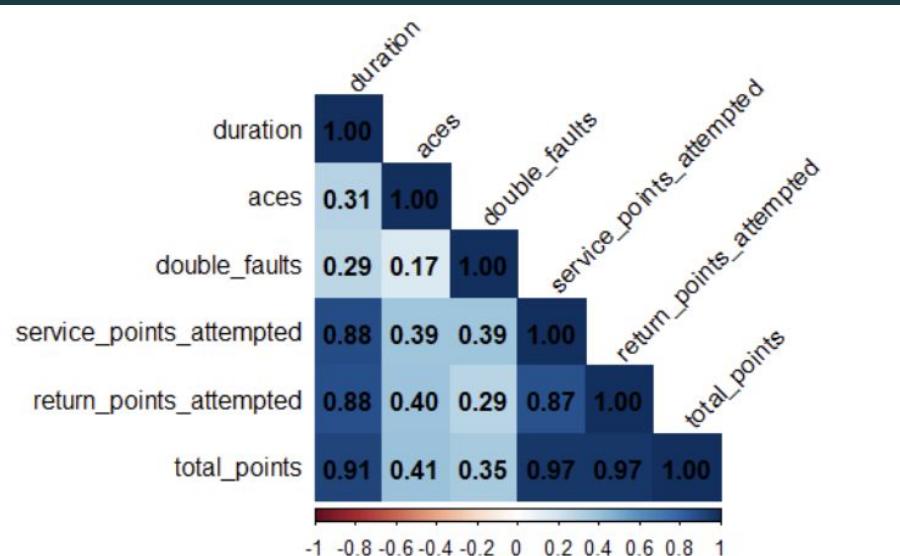
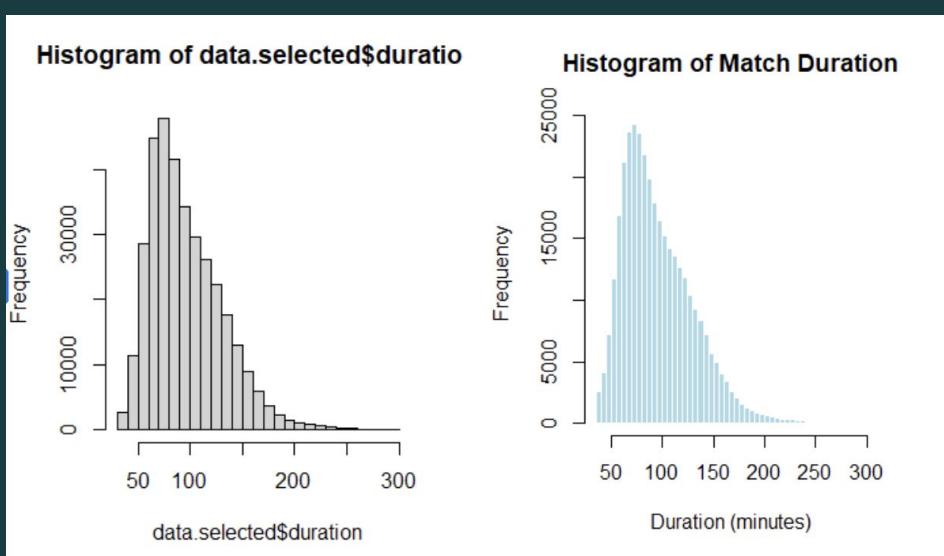
## Predictor Variables:

- **Aces** – total number of aces served
- **Double Faults** – total number of double faults
- **Service Points Attempted** – total number of points played on serve
- **Return Points Attempted** – total number of points played on return
- **Total Points** – total number of points played in the match
- **Court Surface** – type of surface (hard, clay, grass)

## Data Summary:

- The dataset includes **345,974 matches** from professional tournaments

# Data Cleaning and Filtering



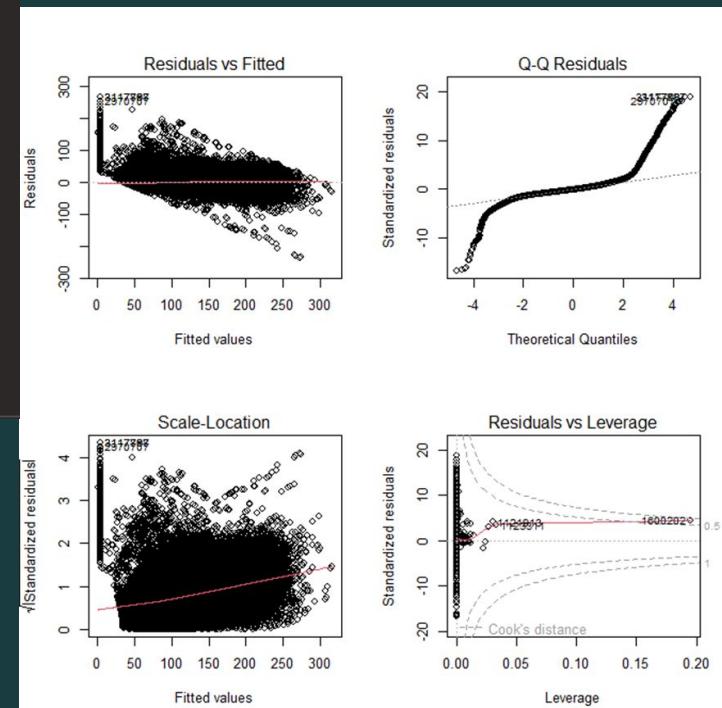
- Most matches are short; the data is right skewed
- We kept matches between 35 and 300 minutes
- Used a transformation to make the data more normal
- Point-based variables are strongly related
- Adjusted for overlap between predictors to improve the model

- Strong correlations among point-based variables (Total, Service, Return)
- Aces and Double Faults show weaker correlations

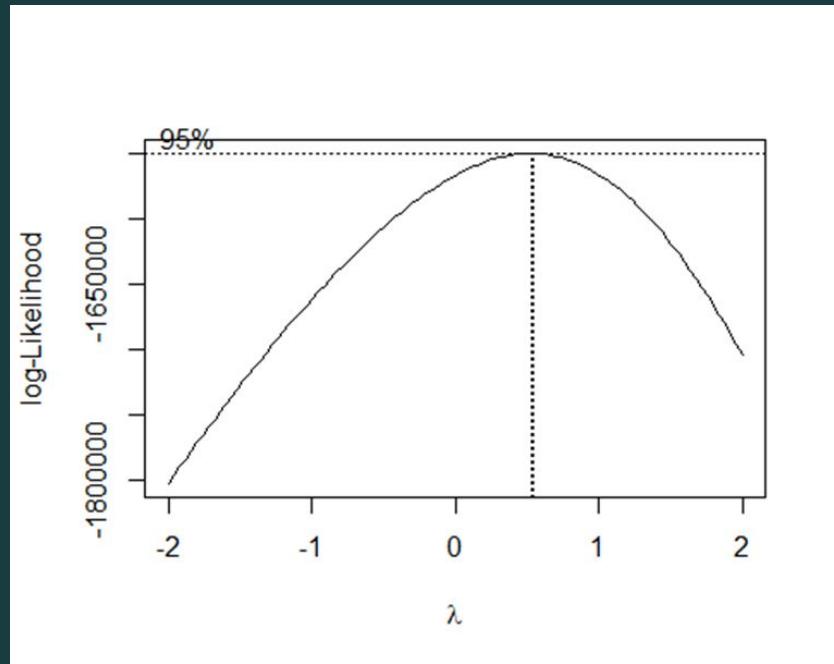
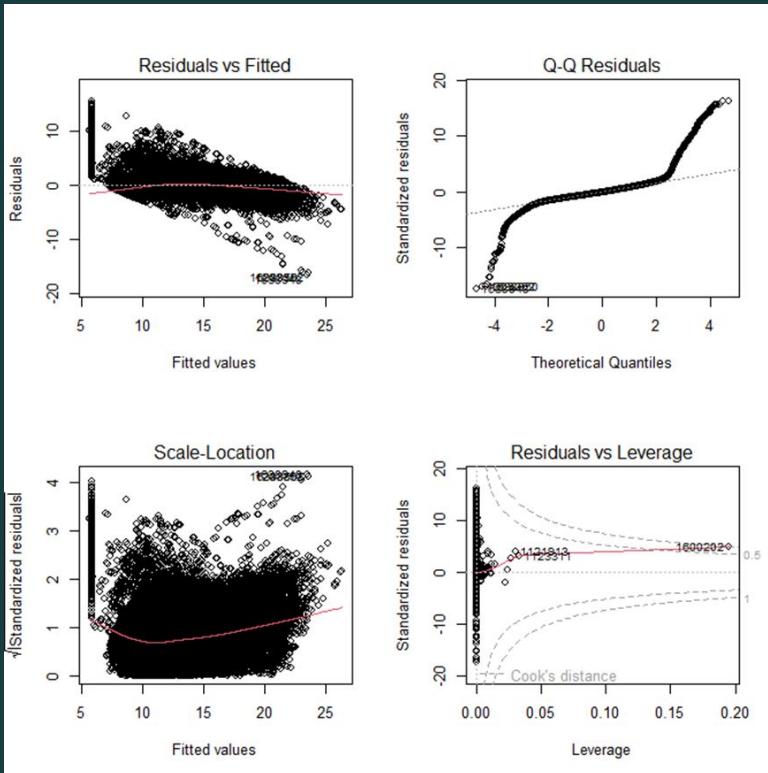
# First Order Model

```
Estimate Std. Error t value Pr(>|t|)  
(Intercept) 3.791843 0.074656 50.791 < 2e-16 ***  
aces -0.563498 0.005982 -94.201 < 2e-16 ***  
double_faults -0.553068 0.010727 -51.560 < 2e-16 ***  
service_points_attempted 0.242930 0.024826 9.785 < 2e-16 ***  
return_points_attempted 0.195791 0.024686 7.931 2.17e-15 ***  
total_points 0.444346 0.024694 17.994 < 2e-16 ***  
---  
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 14.19 on 346338 degrees of freedom  
(3943600 observations deleted due to missingness)  
Multiple R-squared: 0.8349, Adjusted R-squared: 0.8349  
F-statistic: 3.502e+05 on 5 and 346338 DF, p-value: < 2.2e-16
```

- All the Predictors variables are statistically significant, such as aces, faults, service/return/total points
- Adjusted  $R^2 = 0.835$ , Residual SE = 14.19 minutes
- Aces and faults negatively associated with duration
- Residuals show non-linearity and non-constant variance



# Transformed Model



- We used 0.545 as our Lambda Value for our response variable transformation
- Predictors: same as first-order
- Adjusted R<sup>2</sup> = 0.829, Residual SE = 0.95
- All predictors statistically significant
- Aces/faults impact the duration of the match

# Interaction Model and Stepwise Regression

```
(Intercept) < 2e-16 ***
aces.c < 2e-16 ***
double_faults.c < 2e-16 ***
service_points_attempted.c < 2e-16 ***
return_points_attempted.c 5.38e-15 ***
total_points.c < 2e-16 ***
aces.c:double_faults.c 0.000315 ***
aces.c:service_points_attempted.c 7.86e-11 ***
aces.c:total_points.c 5.49e-10 ***
double_faults.c:service_points_attempted.c < 2e-16 ***
double_faults.c:total_points.c < 2e-16 ***
service_points_attempted.c:return_points_attempted.c < 2e-16 ***
return_points_attempted.c:total_points.c 7.87e-10 ***

---
signif. codes: 0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9476 on 346331 degrees of freedom
Multiple R-squared:  0.8309,    Adjusted R-squared:  0.8309
F-statistic: 1.418e+05 on 12 and 346331 DF,  p-value: < 2.2e-16
```

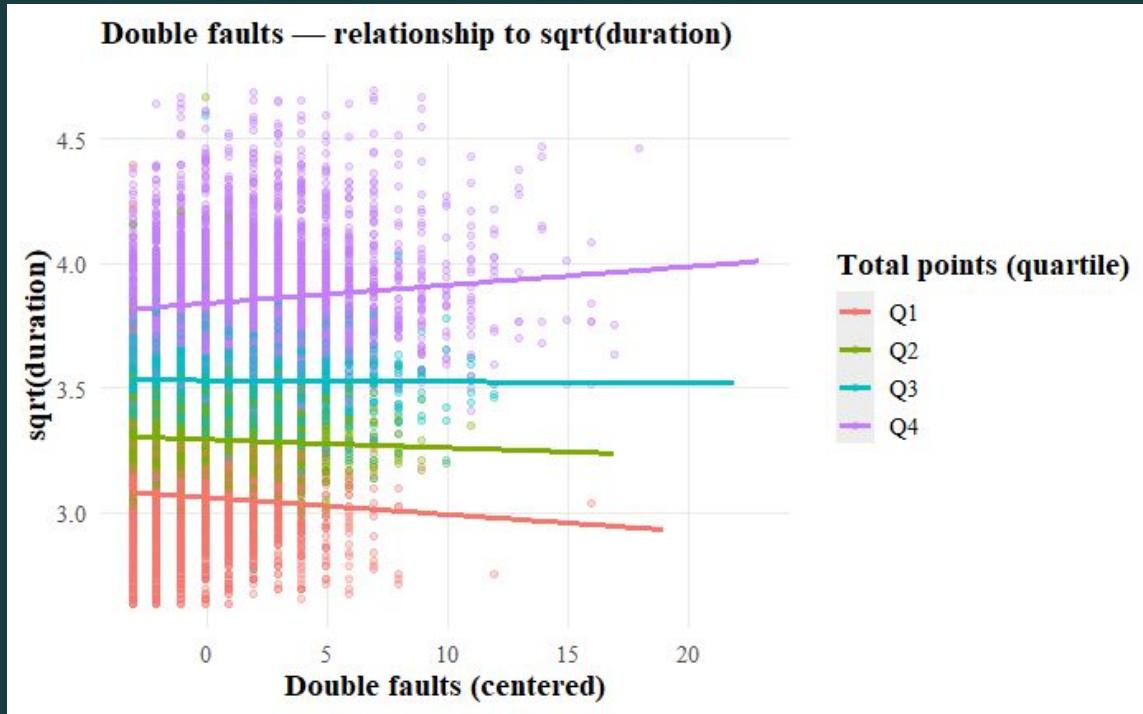
- Added two-way interactions between all quantitative variables
- Final model includes 13 terms selected via AIC stepwise regression
- Adjusted  $R^2 = 0.8309$ , Residual SE = 0.947
- There are 7 Significant interactions in the final model: such as aces  $\times$  service points and return  $\times$  total points
- Improved fit and clarity over non-interaction mode

# Model Comparison – Test MSE

Model	Train_MSE	Test_MSE
<chr>	<dbl>	<dbl>
Full First	0.9070824	0.9071574
Stepwise First	0.9070824	0.9071574
Full Second	0.8979662	0.8983528
Stepwise Second	0.8980007	0.8982778

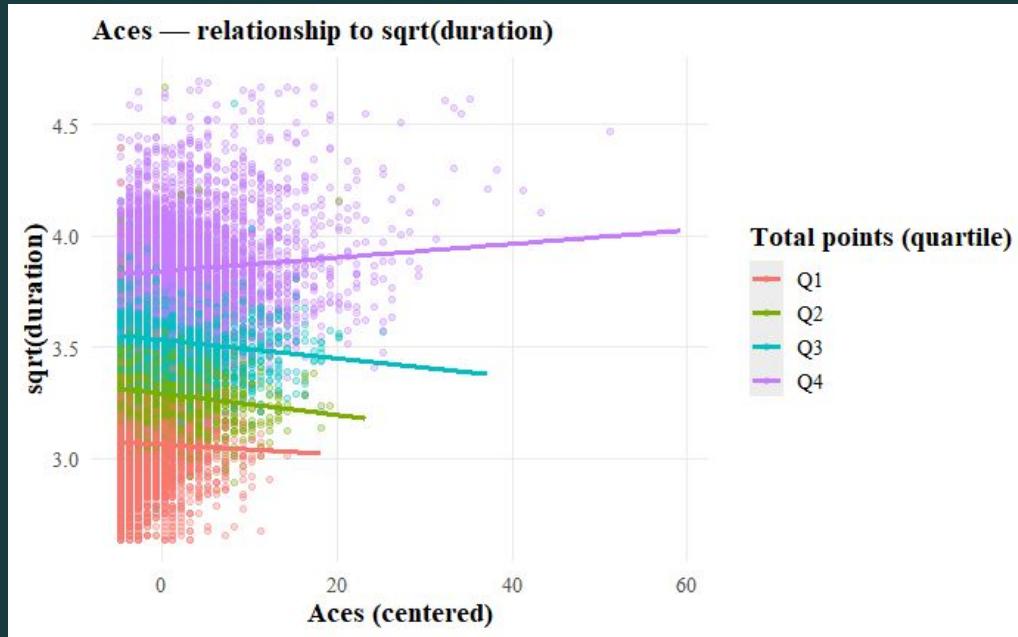
- Compared multiple models using **Test Mean Squared Error (MSE)**
- First-order model: **Test MSE  $\approx 0.907$**
- Transformed model with Box-Cox: **Test MSE  $\approx 0.898$**
- Interaction model (full): **Test MSE  $\approx 0.898$**
- Stepwise interaction model: **lowest Test MSE  $\approx 0.897$**
- The last model (Stepwise Second) has the lowest MSE value, therefore it will be the model we will use as our final model .

# Interaction Plots - Final Model



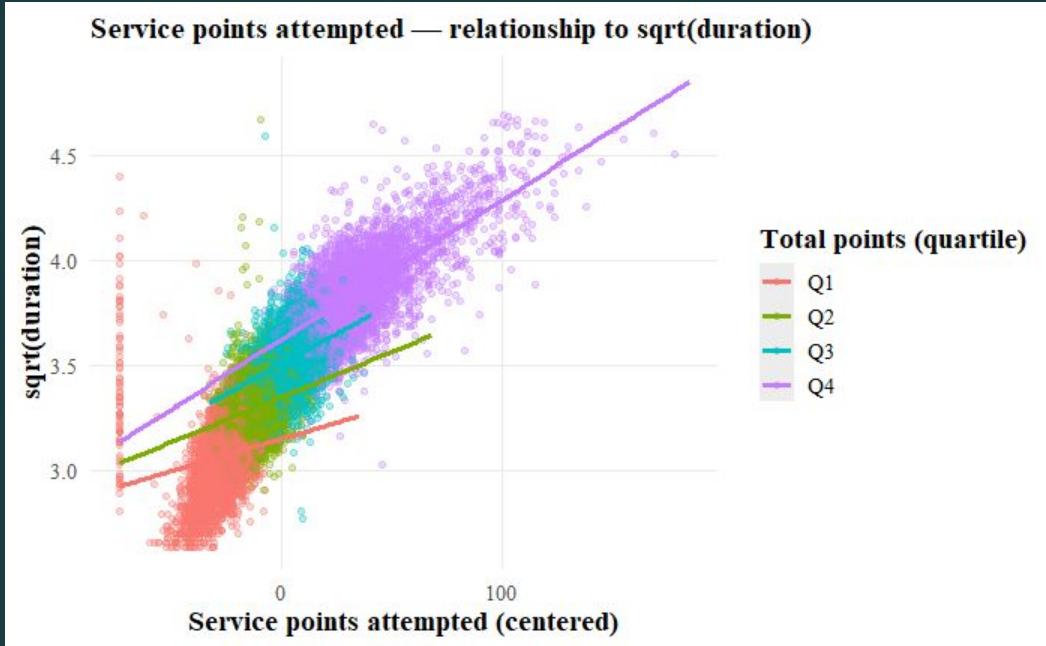
- We can see how the relationship between double fault and duration is slightly weaker for low number of total points compared to medium and high number

# Interaction Plots - Final Model



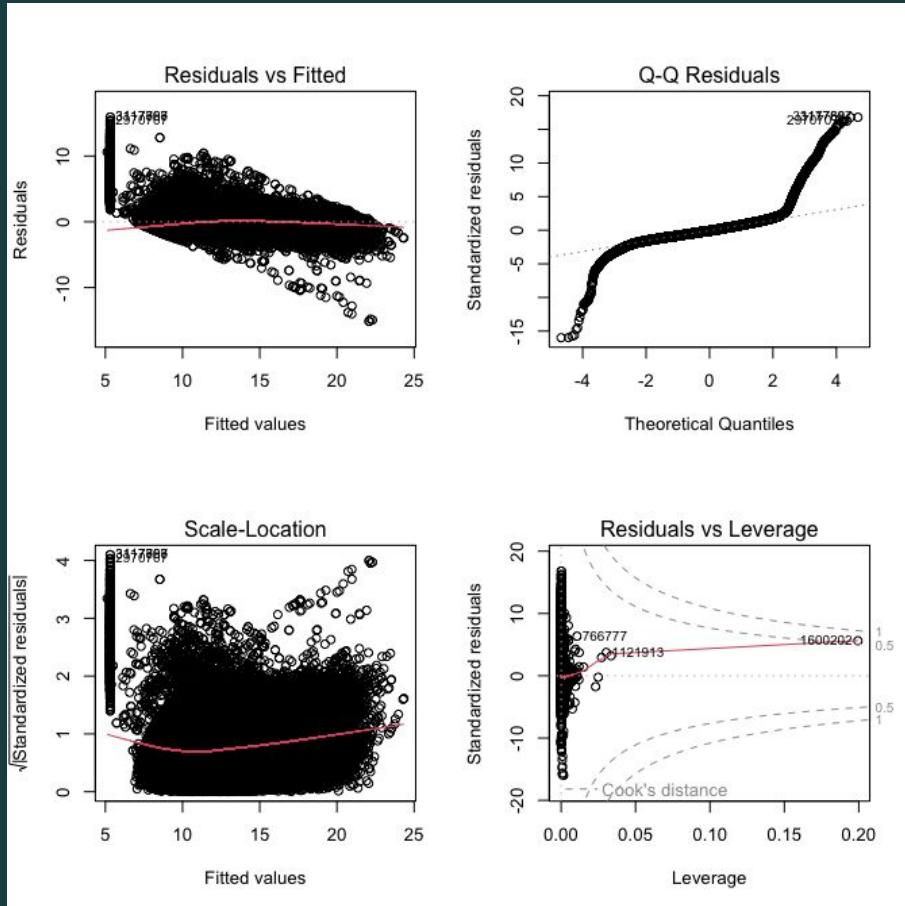
- We can see how the relationship between aces and duration is stronger (steeper) for high number of total points compared to medium and low numbers

# Interaction Plots - Final Model

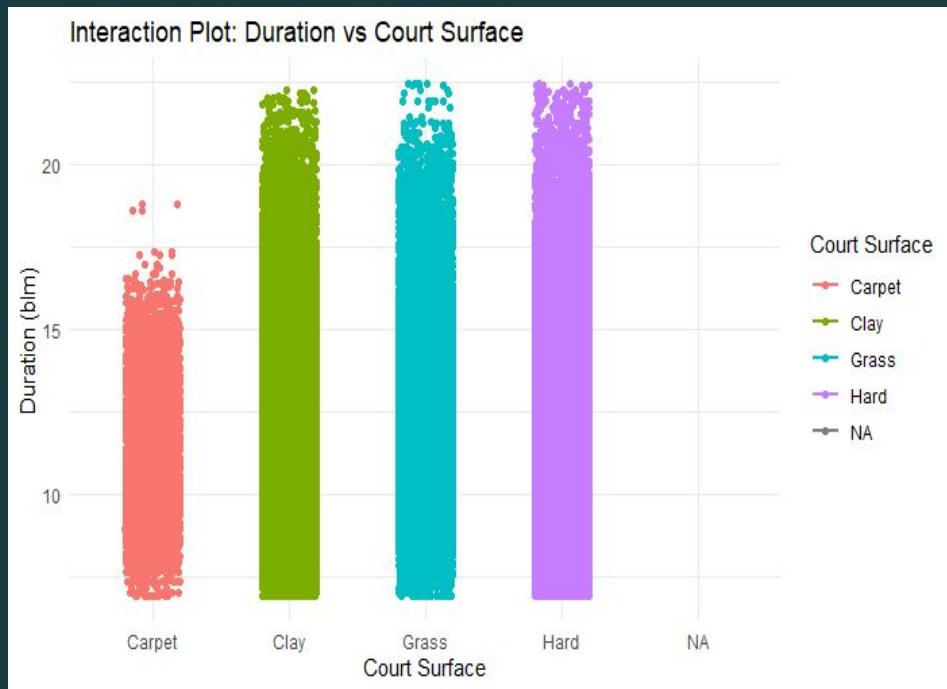
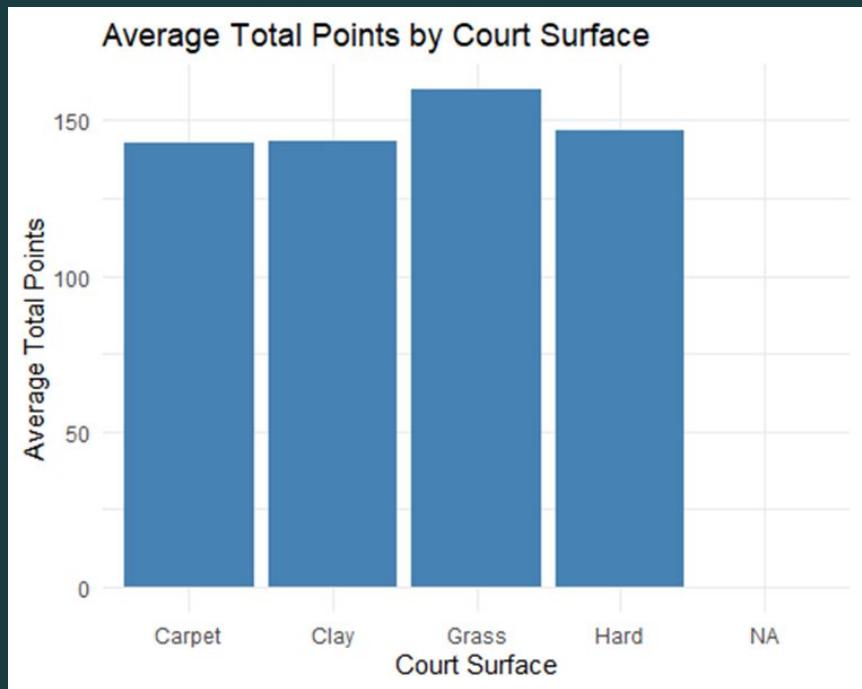


- We can see how the relationship between service points attempts and duration is stronger (steeper) for high number of total points compared to medium and low numbers

# Graph after interaction plots:



# First to Last Comparison



# Conclusion and Limitations

- The final model explains 83.1% of match duration variation
- Serve, return, and point stats are key predictors
- Box-Cox transformation improved fit
- Limitations: Weather, Types of Balls, Humidity, Playing styles

