

UFC RESEARCH QUESTION 1

Due 8th December, 2024

```
ufc <- read.csv("ufc-master.csv")

# View the data
head(ufc)
```

```
# data cleaning cell
colSums(is.na(ufc))
```

RedFighter	BlueFighter
0	0
RedOdds	BlueOdds
219	219
RedExpectedValue	BlueExpectedValue
219	219
Date	Location
0	0
Country	Winner
0	0
TitleBout	WeightClass
0	0
Gender	NumberOfRounds
0	0
BlueCurrentLoseStreak	BlueCurrentWinStreak
0	0
BlueDraws	BlueAvgSigStrLanded
0	930
BlueAvgSigStrPct	BlueAvgSubAtt
765	832
BlueAvgTDLanded	BlueAvgTDPct
833	842

BlueLongestWinStreak	0	BlueLosses	0
BlueTotalRoundsFought	0	BlueTotalTitleBouts	0
BlueWinsByDecisionMajority	0	BlueWinsByDecisionSplit	0
BlueWinsByDecisionUnanimous	0	BlueWinsByKO	0
BlueWinsBySubmission	0	BlueWinsByTKODoctorStoppage	0
BlueWins	0	BlueStance	0
BlueHeightCms	0	BlueReachCms	0
BlueWeightLbs	0	RedCurrentLoseStreak	0
RedCurrentWinStreak	0	RedDraws	0
RedAvgSigStrLanded	455	RedAvgSigStrPct	357
RedAvgSubAtt	357	RedAvgTDLanded	357
RedAvgTDPct	367	RedLongestWinStreak	0
RedLosses	0	RedTotalRoundsFought	0
RedTotalTitleBouts	0	RedWinsByDecisionMajority	0
RedWinsByDecisionSplit	0	RedWinsByDecisionUnanimous	0
RedWinsByKO	0	RedWinsBySubmission	0
RedWinsByTKODoctorStoppage	0	RedWins	0
RedStance	0	RedHeightCms	0
RedReachCms	0	RedWeightLbs	0
RedAge	0	BlueAge	0
LoseStreakDif	0	WinStreakDif	0
LongestWinStreakDif		WinDif	

	0	0
LossDif		TotalRoundDif
	0	0
TotalTitleBoutDif		KODif
	0	0
SubDif		HeightDif
	0	0
ReachDif		AgeDif
	0	0
SigStrDif		AvgSubAttDif
	0	0
AvgTDDif		EmptyArena
	0	1436
BMatchWCRank		RMatchWCRank
	5289	4716
RWFlyweightRank		RWFeatherweightRank
	6382	6469
RWStrawweightRank		RWBantamweightRank
	6334	6324
RHeavyweightRank		RLightHeavyweightRank
	6295	6296
RMiddleweightRank		RWelterweightRank
	6296	6290
RLightweightRank		RFeatherweightRank
	6295	6303
RBantamweightRank		RFlyweightRank
	6299	6292
RPFPRank		BWFlyweightRank
	6228	6406
BWFeatherweightRank		BWStrawweightRank
	6477	6380
BWBantamweightRank		BHeavyweightRank
	6371	6332
BLightHeavyweightRank		BMiddleweightRank
	6360	6341
BWelterweightRank		BLightweightRank
	6360	6359
BFeatherweightRank		BBantamweightRank
	6355	6360
BFlyweightRank		BPFPRank
	6348	6411
BetterRank		Finish
	0	0

FinishDetails	FinishRound
0	622
FinishRoundTime	TotalFightTimeSecs
0	622
RedDecOdds	BlueDecOdds
1077	1107
RSubOdds	BSubOdds
1326	1350
RKOdds	BKOdds
1324	1351

```
# removing the data which has way too many missing values
```

```
ufc = subset(ufc, select = -c(BMatchWCRank, RMatchWCRank, RWFlyweightRank,
RWFeatherweightRank, RWStrawweightRank, RWBantamweightRank,
RHeavyweightRank, RLIGHTHeavyweightRank, RMiddleweightRank,
RWelterweightRank, RLIGHTweightRank, RFeatherweightRank,
RBantamweightRank, RFlyweightRank, RPFPRank, BWFlyweightRank,
BWFeatherweightRank, BWStrawweightRank, BWBantamweightRank,
BHeavyweightRank, BLIGHTHeavyweightRank, BMiddleweightRank,
BWelterweightRank, BLIGHTweightRank, BFeatherweightRank,
BBantamweightRank, BFlyweightRank, BPFPRank))
```

```
library(dplyr)
```

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

```
filter, lag
```

The following objects are masked from 'package:base':

```
intersect, setdiff, setequal, union
```

```
# removing all missing value rows from the columns of interest
ufc_clean <- ufc %>%
  filter(
    !is.na(RedAvgSubAtt),
    !is.na(BlueAvgSubAtt),
```

```

!is.na(BlueReachCms),
!is.na(RedReachCms),
!is.na(BlueAvgSigStrLanded),
!is.na(RedAvgSigStrLanded),
!is.na(TotalFightTimeSecs),
!is.na(WeightClass)
)
nrow(ufc_clean)

```

[1] 4895

```

filtered_ufc_blue <- ufc_clean[c("BlueReachCms", "BlueAvgSigStrLanded",
                                "WeightClass", "BlueHeightCms",
                                "BlueCurrentWinStreak")]
colnames(filtered_ufc_blue) <- c("ReachCms", "AvgSigStrLanded", "WeightClass",
                                 "Height", "WinStreak")
filtered_ufc_red <- ufc_clean[c("RedReachCms", "RedAvgSigStrLanded",
                                 "WeightClass", "RedHeightCms", "RedCurrentWinStreak")]
colnames(filtered_ufc_red) <- c("ReachCms", "AvgSigStrLanded",
                               "WeightClass", "Height", "WinStreak")

# appending the two data sets
ufc_q1 <- rbind(filtered_ufc_blue, filtered_ufc_red)

# exclude outlier(one observation with 0 cm reach)
ufc_q1 <- ufc_q1[ufc_q1$ReachCms > 0,]
ufc_q1 <- ufc_q1[ufc_q1$AvgSigStrLanded > 0,]

# check missing value: no missing
colSums(is.na(ufc_q1))

```

ReachCms	AvgSigStrLanded	WeightClass	Height	WinStreak
0	0	0	0	0

```

# Log-transform the variables
ufc_q1$LogAvgSigStrLanded <- log(ufc_q1$AvgSigStrLanded)
ufc_q1$LogReachCms <- log(ufc_q1$ReachCms)
model_q1 <- lm(AvgSigStrLanded ~ ReachCms + WeightClass + Height + WinStreak,
                 data = ufc_q1)
summary(model_q1)

```

Call:

```
lm(formula = AvgSigStrLanded ~ ReachCms + WeightClass + Height +  
    WinStreak, data = ufc_q1)
```

Residuals:

Min	1Q	Median	3Q	Max
-32.933	-16.378	-5.291	12.299	130.359

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	72.069136	7.569175	9.521	< 2e-16 ***
ReachCms	-0.293536	0.044100	-6.656	2.96e-11 ***
WeightClassCatch Weight	-10.877927	2.362018	-4.605	4.17e-06 ***
WeightClassFeatherweight	2.574233	0.883784	2.913	0.00359 **
WeightClassFlyweight	-2.177334	1.082890	-2.011	0.04439 *
WeightClassHeavyweight	3.866656	1.324240	2.920	0.00351 **
WeightClassLight Heavyweight	6.003263	1.236674	4.854	1.23e-06 ***
WeightClassLightweight	4.238484	0.842526	5.031	4.97e-07 ***
WeightClassMiddleweight	3.005096	1.062903	2.827	0.00470 **
WeightClassWelterweight	5.807064	0.946702	6.134	8.90e-10 ***
WeightClassWomen's Bantamweight	-1.066585	1.309501	-0.814	0.41538
WeightClassWomen's Featherweight	-14.146466	3.414904	-4.143	3.46e-05 ***
WeightClassWomen's Flyweight	-11.644051	1.243821	-9.362	< 2e-16 ***
WeightClassWomen's Strawweight	-2.079263	1.207181	-1.722	0.08503 .
Height	-0.004203	0.057011	-0.074	0.94124
WinStreak	0.939495	0.121104	7.758	9.51e-15 ***

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'
	0.1 '	'	1	

Residual standard error: 19.86 on 9719 degrees of freedom

Multiple R-squared: 0.03332, Adjusted R-squared: 0.03183

F-statistic: 22.33 on 15 and 9719 DF, p-value: < 2.2e-16

```
# Load necessary libraries  
library(car) # For VIF
```

Loading required package: carData

Attaching package: 'car'

The following object is masked from 'package:dplyr':

recode

```
library(ggplot2)      # For residual plots

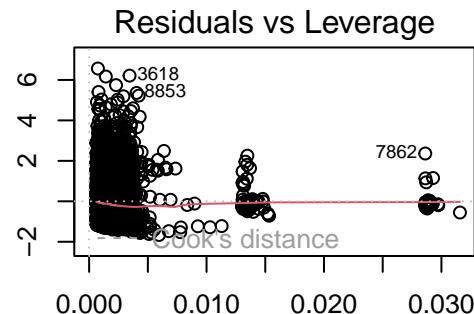
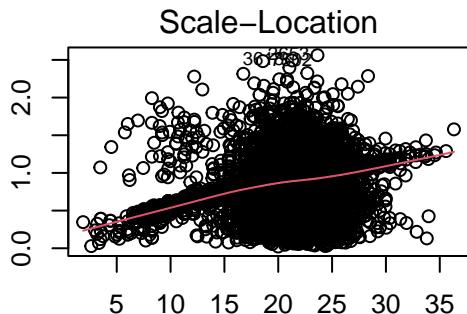
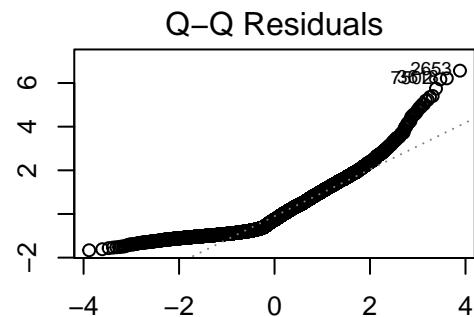
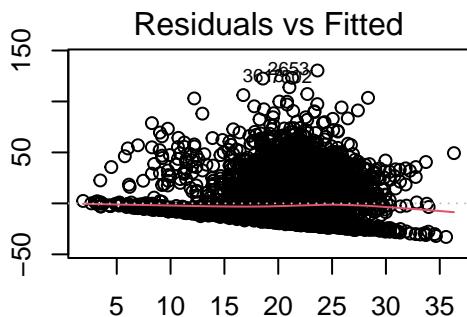
# 1. Check Variance Inflation Factor (VIF) for collinearity
vif_values <- vif(model_q1)
print("Variance Inflation Factor (VIF):")
```

[1] "Variance Inflation Factor (VIF):"

```
print(vif_values)
```

	GVIF	Df	GVIF ^{(1/(2*Df))}
ReachCms	5.796169	1	2.407523
WeightClass	4.223489	12	1.061866
Height	6.625134	1	2.573934
WinStreak	1.004542	1	1.002268

```
# 2. Residuals vs Fitted Plot for Linearity
par(mfrow = c(2, 2), mar = c(2, 2, 2, 2))  # Set plotting layout
plot(model_q1)
```



```

# 3. Normal Q-Q Plot for Normality of Residuals
qqnorm(residuals(model_q1))
qqline(residuals(model_q1))

# 4. Scale-Location Plot for Homoscedasticity
plot(model_q1, which = 3)

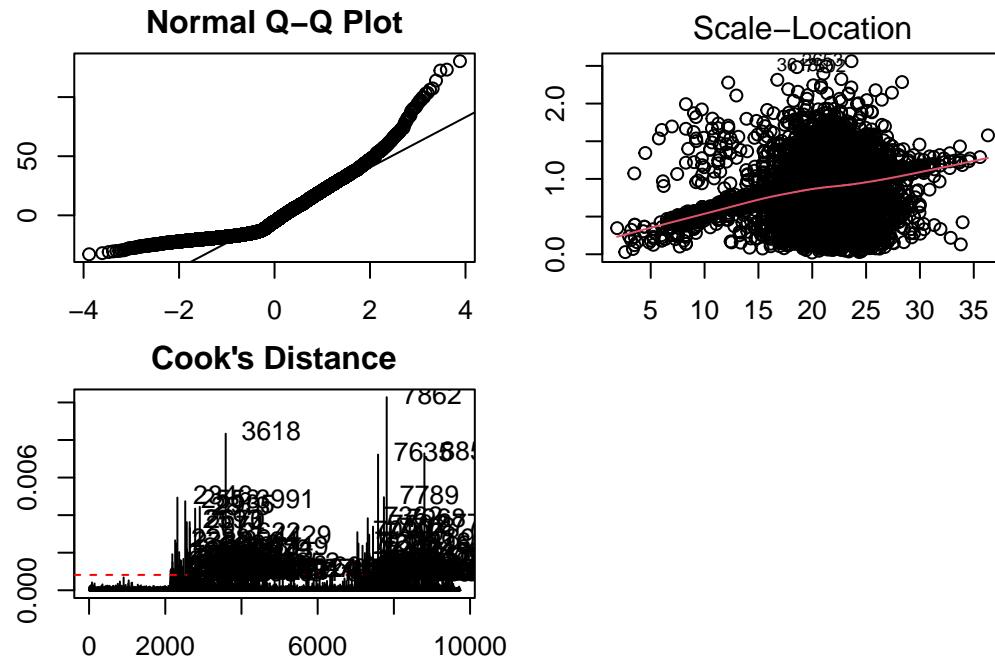
# 5. Check for influential points using Cook's Distance
cooksrd <- cooks.distance(model_q1)
plot(cooksrd, type = "h", main = "Cook's Distance", ylab = "Cook's Distance")

# Highlight observations with Cook's Distance > threshold
threshold <- 4 / nrow(ufc_clean)
influential <- which(cooksrd > threshold)
abline(h = threshold, col = "red", lty = 2)
text(x = influential, y = cooksrd[influential], labels = names(cooksrd[influential]), pos = 4)

# 6. R-squared value
r_squared <- summary(model_q1)$r.squared
cat("R-squared:", r_squared, "\n")

```

R-squared: 0.033321



Linearity assumption is not being met as the residuals show a curved pattern or non-linearity.
 Polynomial does not work
 Removing influential points

OPTION: REMOVING INFLUENTIAL POINTS AND TAKING LOG

```
cooks_d <- cooks.distance(model_q1)
influential <- which(cooks_d > (4 / nrow(ufc_q1)))
ufc_q1_clean <- ufc_q1[-influential, ]

model_clean <- lm(AvgSigStrLanded ~ LogReachCms + WeightClass + log(Height) + WinStreak, data = ufc_q1_clean)
summary(model_clean)
```

Call:
`lm(formula = AvgSigStrLanded ~ LogReachCms + WeightClass + log(Height) + WinStreak, data = ufc_q1_clean)`

Residuals:

Min	1Q	Median	3Q	Max
-26.901	-14.184	-3.669	11.283	60.174

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	400.7096	34.9873	11.453	< 2e-16 ***
LogReachCms	-48.2990	6.9390	-6.961	3.62e-12 ***
WeightClassCatch Weight	-12.0364	2.1284	-5.655	1.60e-08 ***
WeightClassFeatherweight	2.9853	0.7562	3.948	7.94e-05 ***
WeightClassFlyweight	-4.9225	0.9423	-5.224	1.79e-07 ***
WeightClassHeavyweight	6.4464	1.1245	5.733	1.02e-08 ***
WeightClassLight Heavyweight	8.4262	1.0546	7.990	1.51e-15 ***
WeightClassLightweight	6.5209	0.7185	9.075	< 2e-16 ***
WeightClassMiddleweight	5.7734	0.9060	6.372	1.95e-10 ***
WeightClassWelterweight	8.5157	0.8094	10.521	< 2e-16 ***
WeightClassWomen's Bantamweight	-7.0528	1.1701	-6.027	1.73e-09 ***
WeightClassWomen's Featherweight	-15.4604	3.0825	-5.016	5.39e-07 ***
WeightClassWomen's Flyweight	-16.2649	1.0927	-14.885	< 2e-16 ***
WeightClassWomen's Strawweight	-12.3818	1.0955	-11.303	< 2e-16 ***
log(Height)	-25.9182	8.7938	-2.947	0.00321 **

```

WinStreak          0.8518      0.1135    7.505 6.73e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

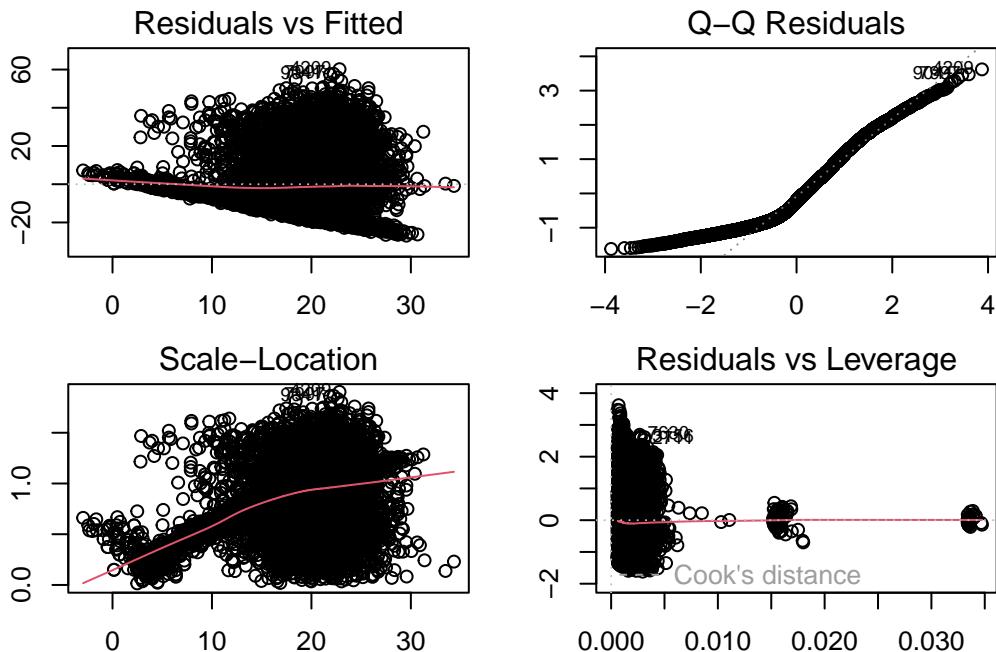
Residual standard error: 16.63 on 9298 degrees of freedom
Multiple R-squared:  0.07784,   Adjusted R-squared:  0.07635
F-statistic: 52.32 on 15 and 9298 DF,  p-value: < 2.2e-16

```

```

par(mfrow = c(2, 2), mar = c(2,2,2,2))
plot(model_clean)

```



This proved somewhat helpful.

Linearity is still an issue.

Addressing the high VIF in the original model:

```

model_simple <- lm(LogAvgSigStrLanded ~ LogReachCms + WeightClass + log(Height) + WinStreak,
vif(model_simple)

```

	GVIF	Df	GVIF ^{1/(2*Df)}
LogReachCms	5.894761	1	2.427913
WeightClass	4.294365	12	1.062602

```

log(Height) 6.695866 1      2.587637
WinStreak    1.004200 1      1.002098

```

```
vif(model_clean)
```

	GVIF	Df	GVIF^(1/(2*Df))
LogReachCms	5.735358	1	2.394861
WeightClass	4.195464	12	1.061571
log(Height)	6.624430	1	2.573797
WinStreak	1.004118	1	1.002057

OPTION : INTERACTION TERM - after removing outliers

```

model_clean_int <- lm(AvgSigStrLanded ~ LogReachCms + WinStreak * WeightClass + log(Height),
summary(model_clean_int)

```

Call:

```
lm(formula = AvgSigStrLanded ~ LogReachCms + WinStreak * WeightClass +
log(Height), data = ufc_q1_clean)
```

Residuals:

Min	1Q	Median	3Q	Max
-27.795	-14.171	-3.582	11.262	60.200

Coefficients:

	Estimate	Std. Error	t value
(Intercept)	403.625326	35.088642	11.503
LogReachCms	-48.698879	6.949576	-7.007
WinStreak	0.827893	0.357816	2.314
WeightClassCatch Weight	-12.006880	3.037479	-3.953
WeightClassFeatherweight	3.154362	0.937613	3.364
WeightClassFlyweight	-4.595719	1.207744	-3.805
WeightClassHeavyweight	5.799845	1.287057	4.506
WeightClassLight Heavyweight	8.541489	1.206885	7.077
WeightClassLightweight	6.487728	0.878628	7.384
WeightClassMiddleweight	5.921978	1.041420	5.686
WeightClassWelterweight	8.306257	0.959762	8.654
WeightClassWomen's Bantamweight	-8.091215	1.453752	-5.566
WeightClassWomen's Featherweight	-14.751683	4.156747	-3.549

WeightClassWomen's Flyweight	-15.154280	1.349614	-11.229
WeightClassWomen's Strawweight	-12.085748	1.352000	-8.939
log(Height)	-26.078586	8.810044	-2.960
WinStreak:WeightClassCatch Weight	-0.006436	1.800246	-0.004
WinStreak:WeightClassFeatherweight	-0.140130	0.494172	-0.284
WinStreak:WeightClassFlyweight	-0.326980	0.708211	-0.462
WinStreak:WeightClassHeavyweight	0.585823	0.537363	1.090
WinStreak:WeightClassLight Heavyweight	-0.050494	0.531143	-0.095
WinStreak:WeightClassLightweight	0.048469	0.447468	0.108
WinStreak:WeightClassMiddleweight	-0.083407	0.453502	-0.184
WinStreak:WeightClassWelterweight	0.215612	0.456110	0.473
WinStreak:WeightClassWomen's Bantamweight	1.013347	0.830416	1.220
WinStreak:WeightClassWomen's Featherweight	-0.937783	3.616108	-0.259
WinStreak:WeightClassWomen's Flyweight	-1.017894	0.710903	-1.432
WinStreak:WeightClassWomen's Strawweight	-0.370837	0.823802	-0.450
Pr(> t)			
(Intercept)	< 2e-16 ***		
LogReachCms	2.60e-12 ***		
WinStreak	0.020704 *		
WeightClassCatch Weight	7.78e-05 ***		
WeightClassFeatherweight	0.000771 ***		
WeightClassFlyweight	0.000143 ***		
WeightClassHeavyweight	6.68e-06 ***		
WeightClassLight Heavyweight	1.58e-12 ***		
WeightClassLightweight	1.67e-13 ***		
WeightClassMiddleweight	1.34e-08 ***		
WeightClassWelterweight	< 2e-16 ***		
WeightClassWomen's Bantamweight	2.68e-08 ***		
WeightClassWomen's Featherweight	0.000389 ***		
WeightClassWomen's Flyweight	< 2e-16 ***		
WeightClassWomen's Strawweight	< 2e-16 ***		
log(Height)	0.003083 **		
WinStreak:WeightClassCatch Weight	0.997148		
WinStreak:WeightClassFeatherweight	0.776749		
WinStreak:WeightClassFlyweight	0.644309		
WinStreak:WeightClassHeavyweight	0.275662		
WinStreak:WeightClassLight Heavyweight	0.924264		
WinStreak:WeightClassLightweight	0.913746		
WinStreak:WeightClassMiddleweight	0.854082		
WinStreak:WeightClassWelterweight	0.636425		
WinStreak:WeightClassWomen's Bantamweight	0.222387		
WinStreak:WeightClassWomen's Featherweight	0.795383		
WinStreak:WeightClassWomen's Flyweight	0.152226		

```

WinStreak:WeightClassWomen's Strawweight    0.652611
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

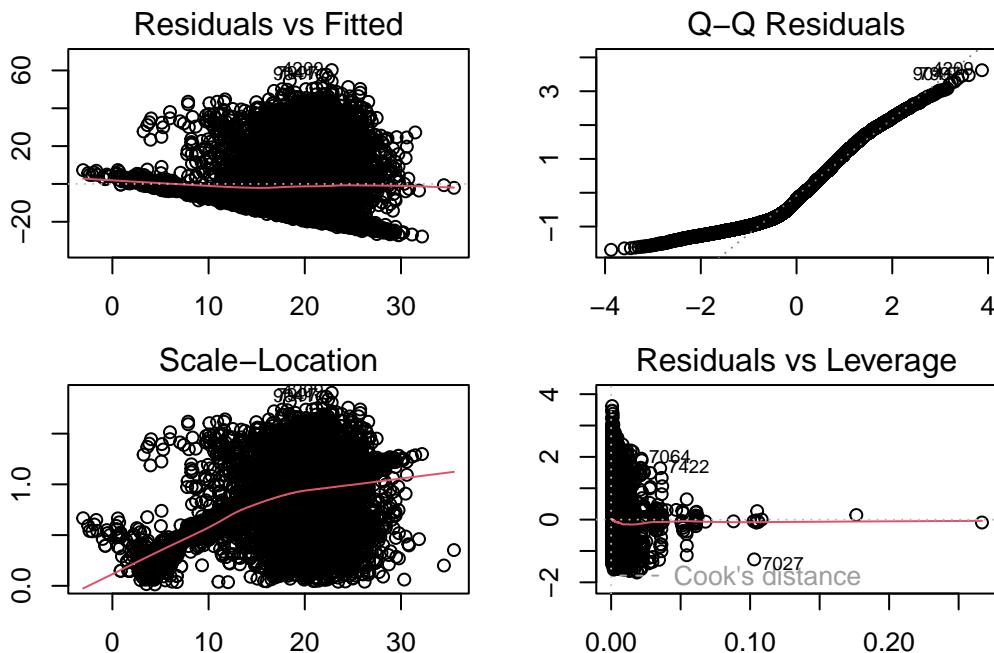
Residual standard error: 16.63 on 9286 degrees of freedom
Multiple R-squared:  0.07864,   Adjusted R-squared:  0.07596
F-statistic: 29.36 on 27 and 9286 DF,  p-value: < 2.2e-16

```

```

par(mfrow = c(2, 2), mar = c(2,2,2,2))
plot(model_clean_int)

```



```
vif(model_clean_int)
```

there are higher-order terms (interactions) in this model
consider setting type = 'predictor'; see ?vif

	GVIF	Df	GVIF^(1/(2*Df))
LogReachCms	5.750458	1	2.398011
WinStreak	9.975393	1	3.158385
WeightClass	1011.233303	12	1.334142
log(Height)	6.646197	1	2.578022
WinStreak:WeightClass	2303.903508	12	1.380711

OPTION: INTERACTION TERM WITH THE OG DATA

```
model_int <- lm(AvgSigStrLanded ~ LogReachCms + WinStreak * WeightClass + log(Height), data = ufc_q1)
summary(model_int)
```

Call:

```
lm(formula = AvgSigStrLanded ~ LogReachCms + WinStreak * WeightClass +
log(Height), data = ufc_q1)
```

Residuals:

Min	1Q	Median	3Q	Max
-40.784	-16.289	-5.243	12.177	130.334

Coefficients:

	Estimate	Std. Error	t value
(Intercept)	297.33703	40.58299	7.327
LogReachCms	-53.01281	8.06506	-6.573
WinStreak	0.86960	0.41266	2.107
WeightClassCatch Weight	-9.07458	3.14289	-2.887
WeightClassFeatherweight	1.95325	1.08759	1.796
WeightClassFlyweight	-4.61619	1.33938	-3.447
WeightClassHeavyweight	3.19591	1.49429	2.139
WeightClassLight Heavyweight	6.28582	1.38881	4.526
WeightClassLightweight	4.69386	1.02357	4.586
WeightClassMiddleweight	3.02318	1.22080	2.476
WeightClassWelterweight	6.14137	1.11582	5.504
WeightClassWomen's Bantamweight	-0.88583	1.60375	-0.552
WeightClassWomen's Featherweight	-12.80681	3.96684	-3.228
WeightClassWomen's Flyweight	-10.01930	1.54192	-6.498
WeightClassWomen's Strawweight	-5.36021	1.49938	-3.575
log(Height)	-0.68261	10.15598	-0.067
WinStreak:WeightClassCatch Weight	-1.41349	1.63281	-0.866
WinStreak:WeightClassFeatherweight	0.55490	0.54428	1.020
WinStreak:WeightClassFlyweight	1.97514	0.66901	2.952
WinStreak:WeightClassHeavyweight	0.38553	0.60656	0.636
WinStreak:WeightClassLight Heavyweight	-0.32441	0.56484	-0.574
WinStreak:WeightClassLightweight	-0.34887	0.49941	-0.699
WinStreak:WeightClassMiddleweight	-0.05175	0.52543	-0.098
WinStreak:WeightClassWelterweight	-0.27948	0.51026	-0.548
WinStreak:WeightClassWomen's Bantamweight	-0.23272	0.87818	-0.265

```

WinStreak:WeightClassWomen's Featherweight -1.30626   1.96116 -0.666
WinStreak:WeightClassWomen's Flyweight     -1.58388   0.82670 -1.916
WinStreak:WeightClassWomen's Strawweight    3.04131   0.84388  3.604
                                         Pr(>|t|)
(Intercept)                         2.55e-13 ***
LogReachCms                          5.18e-11 ***
WinStreak                            0.035118 *
WeightClassCatch Weight               0.003894 **
WeightClassFeatherweight             0.072534 .
WeightClassFlyweight                 0.000570 ***
WeightClassHeavyweight                0.032481 *
WeightClassLight Heavyweight          6.08e-06 ***
WeightClassLightweight                4.58e-06 ***
WeightClassMiddleweight               0.013289 *
WeightClassWelterweight              3.81e-08 ***
WeightClassWomen's Bantamweight       0.580721
WeightClassWomen's Featherweight      0.001249 **
WeightClassWomen's Flyweight          8.54e-11 ***
WeightClassWomen's Strawweight         0.000352 ***
log(Height)                          0.946414
WinStreak:WeightClassCatch Weight    0.386687
WinStreak:WeightClassFeatherweight   0.307986
WinStreak:WeightClassFlyweight        0.003161 **
WinStreak:WeightClassHeavyweight     0.525053
WinStreak:WeightClassLight Heavyweight 0.565752
WinStreak:WeightClassLightweight      0.484834
WinStreak:WeightClassMiddleweight    0.921548
WinStreak:WeightClassWelterweight    0.583894
WinStreak:WeightClassWomen's Bantamweight 0.791016
WinStreak:WeightClassWomen's Featherweight 0.505387
WinStreak:WeightClassWomen's Flyweight  0.055407 .
WinStreak:WeightClassWomen's Strawweight 0.000315 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

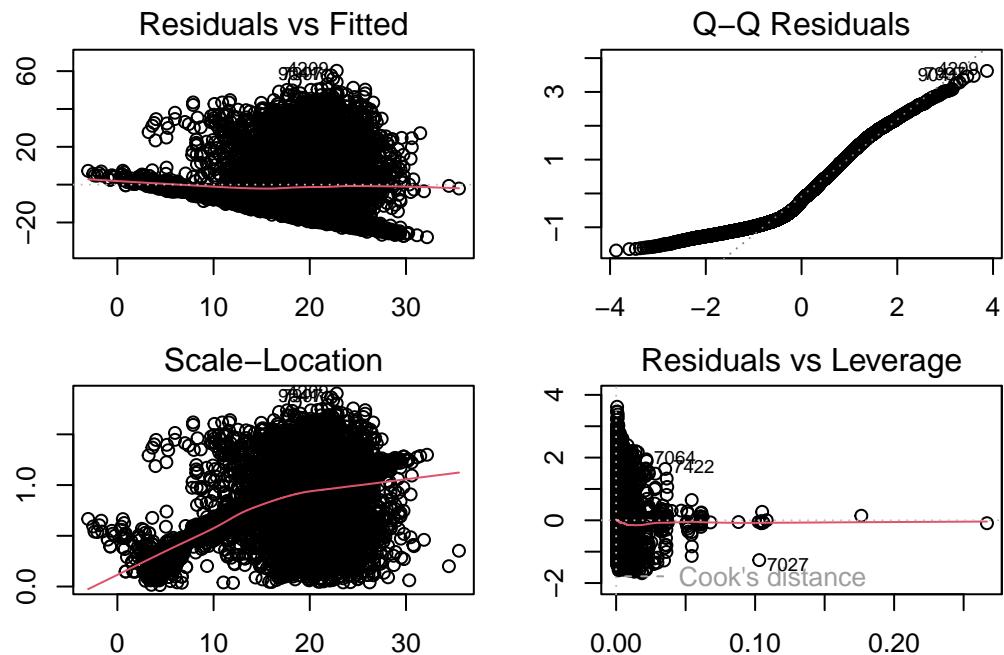
```

Residual standard error: 19.83 on 9707 degrees of freedom
Multiple R-squared: 0.03775, Adjusted R-squared: 0.03508
F-statistic: 14.11 on 27 and 9707 DF, p-value: < 2.2e-16

```

par(mfrow = c(2, 2), mar = c(2,2,2,2))
plot(model_clean_int)

```



```
vif(model_int)
```

there are higher-order terms (interactions) in this model
consider setting type = 'predictor'; see ?vif

	GVIF	Df	GVIF ^{(1/(2*Df))}
LogReachCms	5.919067	1	2.432913
WinStreak	11.703034	1	3.420970
WeightClass	483.945667	12	1.293798
log(Height)	6.727039	1	2.593654
WinStreak:WeightClass	1247.012930	12	1.345844