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                   BlueLosses
                          0                            0
     BlueTotalRoundsFought         BlueTotalTitleBouts
                          0                            0
 BlueWinsByDecisionMajority     BlueWinsByDecisionSplit
                          0                            0
BlueWinsByDecisionUnanimous                BlueWinsByKO
                          0                            0
       BlueWinsBySubmission BlueWinsByTKODoctorStoppage
                          0                            0
                  BlueWins                   BlueStance
                          0                            0
             BlueHeightCms                BlueReachCms
                          0                            0
             BlueWeightLbs        RedCurrentLoseStreak
                          0                            0
        RedCurrentWinStreak                    RedDraws
                          0                            0
         RedAvgSigStrLanded             RedAvgSigStrPct
                        455                          357
               RedAvgSubAtt              RedAvgTDLanded
                        357                          357
               RedAvgTDPct         RedLongestWinStreak
                        367                            0
                 RedLosses        RedTotalRoundsFought
                          0                            0
        RedTotalTitleBouts   RedWinsByDecisionMajority
                          0                            0
    RedWinsByDecisionSplit  RedWinsByDecisionUnanimous
                          0                            0
               RedWinsByKO         RedWinsBySubmission
                          0                            0
 RedWinsByTKODoctorStoppage                     RedWins
                          0                            0
                RedStance                 RedHeightCms
                          0                            0
               RedReachCms                 RedWeightLbs
                          0                            0
                    RedAge                      BlueAge
                          0                            0
             LoseStreakDif                 WinStreakDif
                          0                            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
RKOOdds                      BKOOdds
         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(LogAvgSigStrLanded ~ LogReachCms * WeightClass + Height + WinStreak, data = u
summary(model_q1)
```

```
Call:
lm(formula = LogAvgSigStrLanded ~ LogReachCms * WeightClass +
    Height + WinStreak, data = ufc_q1)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2411 -1.0121  0.1562  0.9527  3.0385

Coefficients:
                                            Estimate Std. Error t value
(Intercept)                                20.299645   5.245888   3.870
LogReachCms                                -3.366757   1.035627  -3.251
WeightClassCatch Weight                    -9.847861  12.912028  -0.763
WeightClassFeatherweight                   16.755691   7.550510   2.219
WeightClassFlyweight                       57.696594   9.120653   6.326
WeightClassHeavyweight                    -13.523345   7.987021  -1.693
WeightClassLight Heavyweight              -14.172206   8.268269  -1.714
WeightClassLightweight                     -0.404508   6.949375  -0.058
WeightClassMiddleweight                    -5.138435   7.523066  -0.683
WeightClassWelterweight                    -1.955368   6.981548  -0.280
WeightClassWomen's Bantamweight           -10.914377  11.779043  -0.927
WeightClassWomen's Featherweight           -1.750844  31.478949  -0.056
WeightClassWomen's Flyweight              -23.948120  10.169350  -2.355
WeightClassWomen's Strawweight            -42.488577   9.154564  -4.641
Height                                     -0.003163   0.003207  -0.986
WinStreak                                   0.041084   0.006774   6.065
LogReachCms:WeightClassCatch Weight         1.781285   2.487705   0.716
LogReachCms:WeightClassFeatherweight       -3.188399   1.458418  -2.186
LogReachCms:WeightClassFlyweight          -11.267528   1.772449  -6.357
LogReachCms:WeightClassHeavyweight          2.646695   1.527269   1.733
LogReachCms:WeightClassLight Heavyweight    2.794438   1.582216   1.766
LogReachCms:WeightClassLightweight          0.154980   1.341615   0.116
LogReachCms:WeightClassMiddleweight         1.050072   1.444678   0.727
LogReachCms:WeightClassWelterweight         0.474180   1.344581   0.353
LogReachCms:WeightClassWomen's Bantamweight 2.093198   2.290004   0.914
LogReachCms:WeightClassWomen's Featherweight 0.143693  6.098032   0.024
LogReachCms:WeightClassWomen's Flyweight    4.511495   1.979945   2.279
LogReachCms:WeightClassWomen's Strawweight  8.272935   1.790830   4.620
                                           Pr(>|t|)
(Intercept)                                0.00011 ***
LogReachCms                                0.00115 **
WeightClassCatch Weight                    0.44567
```

```
WeightClassFeatherweight                         0.02650 *
WeightClassFlyweight                            2.63e-10 ***
WeightClassHeavyweight                           0.09046 .
WeightClassLight Heavyweight                     0.08655 .
WeightClassLightweight                           0.95358
WeightClassMiddleweight                          0.49461
WeightClassWelterweight                          0.77942
WeightClassWomen's Bantamweight                  0.35416
WeightClassWomen's Featherweight                 0.95565
WeightClassWomen's Flyweight                     0.01855 *
WeightClassWomen's Strawweight                  3.51e-06 ***
Height                                           0.32408
WinStreak                                       1.37e-09 ***
LogReachCms:WeightClassCatch Weight              0.47399
LogReachCms:WeightClassFeatherweight             0.02882 *
LogReachCms:WeightClassFlyweight                2.15e-10 ***
LogReachCms:WeightClassHeavyweight               0.08313 .
LogReachCms:WeightClassLight Heavyweight         0.07740 .
LogReachCms:WeightClassLightweight               0.90804
LogReachCms:WeightClassMiddleweight              0.46733
LogReachCms:WeightClassWelterweight              0.72435
LogReachCms:WeightClassWomen's Bantamweight      0.36071
LogReachCms:WeightClassWomen's Featherweight     0.98120
LogReachCms:WeightClassWomen's Flyweight         0.02271 *
LogReachCms:WeightClassWomen's Strawweight      3.89e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.108 on 9707 degrees of freedom
Multiple R-squared:  0.05948,   Adjusted R-squared:  0.05686
F-statistic: 22.74 on 27 and 9707 DF,  p-value: < 2.2e-16
```

```r
# 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)
```

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

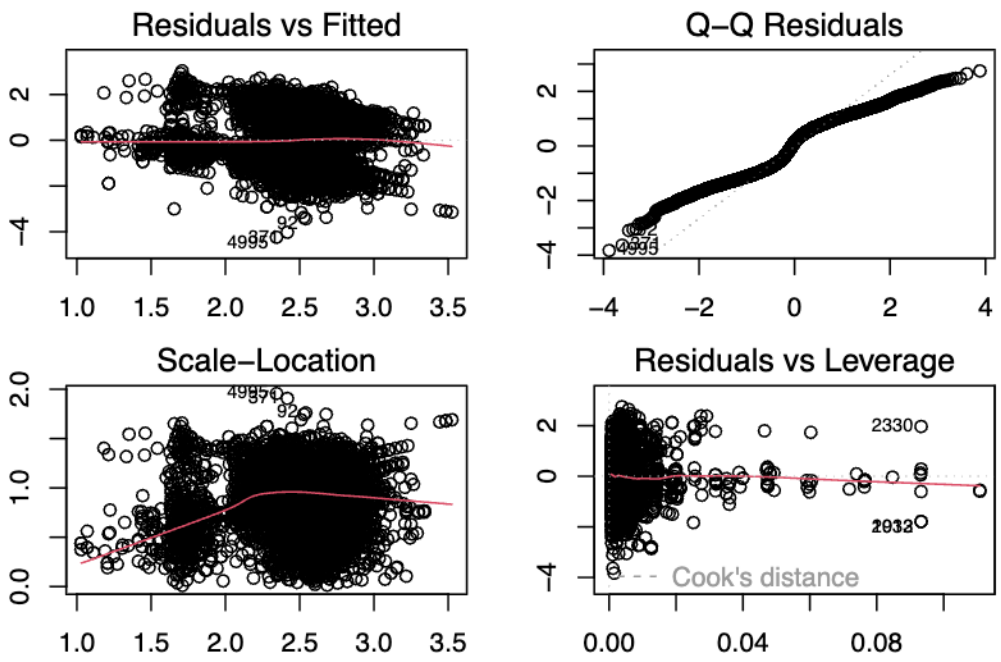
```
print("Variance Inflation Factor (VIF):")
```

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

```
print(vif_values)
```

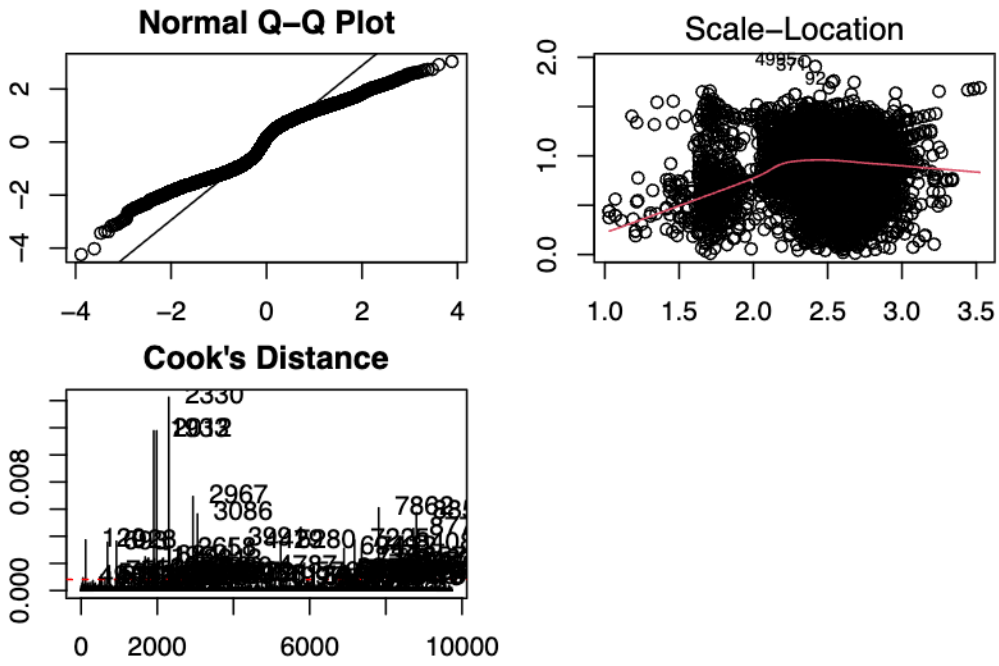
```
                             GVIF Df GVIF^(1/(2*Df))
LogReachCms              3.128002e+01  1        5.592854
WeightClass              2.148248e+52 12      151.531708
Height                   6.742370e+00  1        2.596607
WinStreak                1.010645e+00  1        1.005308
LogReachCms:WeightClass 2.184107e+52 12      151.636265
```

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




```r
# 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
cooksd <- cooks.distance(model_q1)
plot(cooksd, type = "h", main = "Cook's Distance", ylab = "Cook's Distance")

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

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

```
R-squared: 0.05947709
```




```
# Load necessary library
library(knitr)

# Create a summary of the model
model_summary <- summary(model_q1)

# Extract coefficients and format into a data frame
coef_table <- as.data.frame(model_summary$coefficients)
colnames(coef_table) <- c("Estimate", "Std. Error", "t value", "Pr(>|t|)")

# Round to 3 decimal places
coef_table <- round(coef_table, 3)

# Create a kable table
kable(coef_table, caption = "Regression Coefficients for model_log", format = "markdown")
```

Table 1: Regression Coefficients for model_log

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	20.300	5.246	3.870	0.000
LogReachCms	-3.367	1.036	-3.251	0.001

	Estimate	Std. Error	t value	Pr(>|t|)
WeightClassCatch Weight	-9.848	12.912	-0.763	0.446
WeightClassFeatherweight	16.756	7.551	2.219	0.026
WeightClassFlyweight	57.697	9.121	6.326	0.000
WeightClassHeavyweight	-13.523	7.987	-1.693	0.090
WeightClassLight Heavyweight	-14.172	8.268	-1.714	0.087
WeightClassLightweight	-0.405	6.949	-0.058	0.954
WeightClassMiddleweight	-5.138	7.523	-0.683	0.495
WeightClassWelterweight	-1.955	6.982	-0.280	0.779
WeightClassWomen's Bantamweight	-10.914	11.779	-0.927	0.354
WeightClassWomen's Featherweight	-1.751	31.479	-0.056	0.956
WeightClassWomen's Flyweight	-23.948	10.169	-2.355	0.019
WeightClassWomen's Strawweight	-42.489	9.155	-4.641	0.000
Height	-0.003	0.003	-0.986	0.324
WinStreak	0.041	0.007	6.065	0.000
LogReachCms:WeightClassCatch Weight	1.781	2.488	0.716	0.474
LogReachCms:WeightClassFeatherweight	-3.188	1.458	-2.186	0.029
LogReachCms:WeightClassFlyweight	-11.268	1.772	-6.357	0.000
LogReachCms:WeightClassHeavyweight	2.647	1.527	1.733	0.083
LogReachCms:WeightClassLight Heavyweight	2.794	1.582	1.766	0.077
LogReachCms:WeightClassLightweight	0.155	1.342	0.116	0.908
LogReachCms:WeightClassMiddleweight	1.050	1.445	0.727	0.467
LogReachCms:WeightClassWelterweight	0.474	1.345	0.353	0.724
LogReachCms:WeightClassWomen's Bantamweight	2.093	2.290	0.914	0.361
LogReachCms:WeightClassWomen's Featherweight	0.144	6.098	0.024	0.981
LogReachCms:WeightClassWomen's Flyweight	4.511	1.980	2.279	0.023
LogReachCms:WeightClassWomen's Strawweight	8.273	1.791	4.620	0.000