Statistics Assignment 6

- Joshua Boryer 41497475

1a)

Loading the data in and summarising it shows the following variables and their data types.

```
   PostType SocialMedia Fashion     Engagement
   <chr>    <chr>       <chr>            <dbl>
 1 text     S.Media A   fast             0.248
 2 text     S.Media A   sustainable      0.822
 3 image    S.Media A   fast             0.620
 4 image    S.Media A   sustainable      0.682
 5 video    S.Media A   fast             0.850
 6 video    S.Media A   sustainable      0.878
 7 text     S.Media B   fast             0.346
 8 text     S.Media B   sustainable      0.243
 9 image    S.Media B   fast             0.408
10 image    S.Media B   sustainable      0.487
# ℹ 32 more rows
# ℹ Use `print(n = ...)` to see more rows
```

PostType contains “text”, “image”, or “video”.

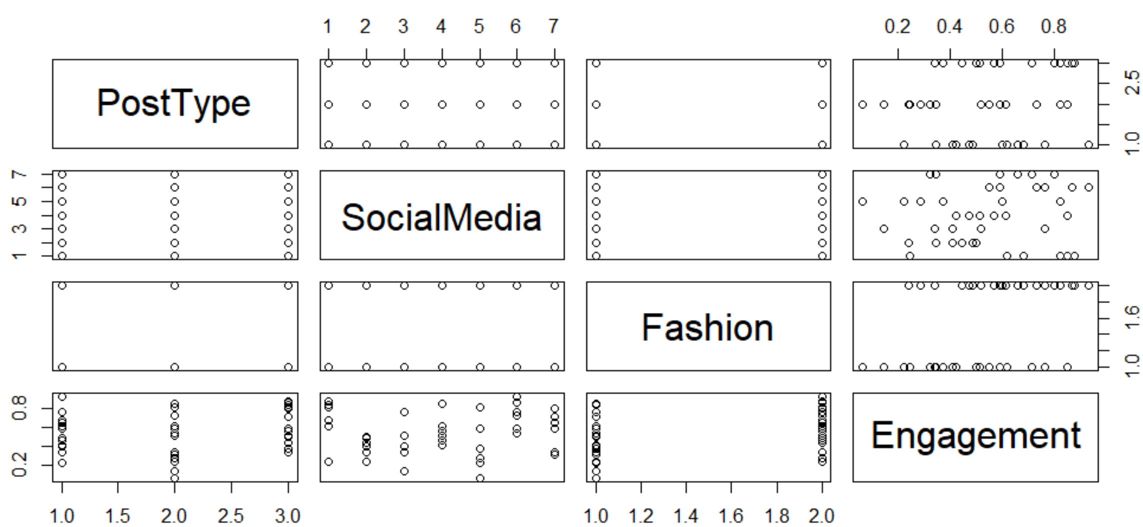
SocialMedia contains “S.Media A”, “S.Media B” … “S.Media G”.

Fashion contains “fast” or “sustainable”.

Engagement contains a decimal point number.

All variables with an exception to “Engagement” have to be loaded as factors as they have textual representations for their values.

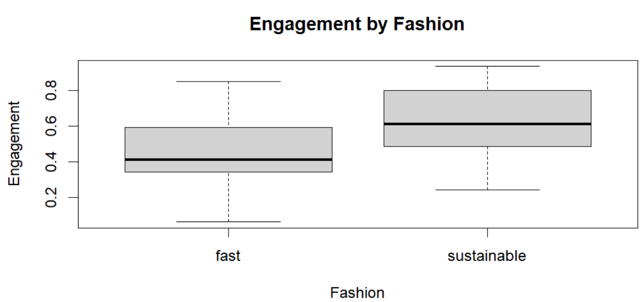
Exploring the data, a matrix plot is created to look for any correlations between variables.

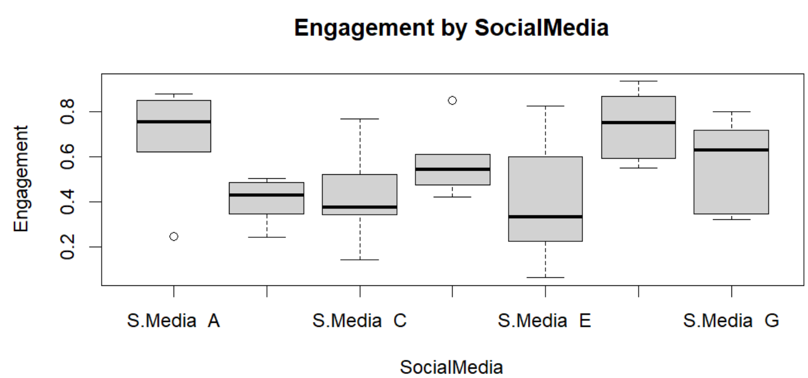


On the matrix plot the visual observations are…

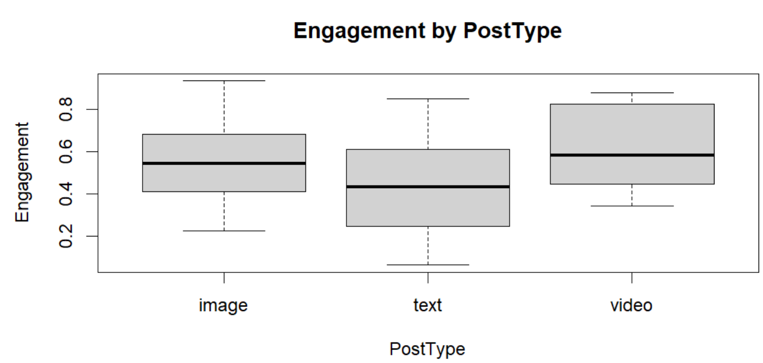
- SocialMedia is correlated with Engagement
- Fashion is moderately correlated with Engagement
- PostType may have a weak correlation with Engagement

Since Engagement is the response variable in our analysis assuming the observations are independent. A histogram of Engagement shows that it is roughly normally distributed.





Checking all the variables against the response variable to look for any visual trend or correlations. Visually, different social media changes the amount of engagement. With social media B having a clearly lower average engagement in comparison to Social media A and Social Media F higher in comparison to the



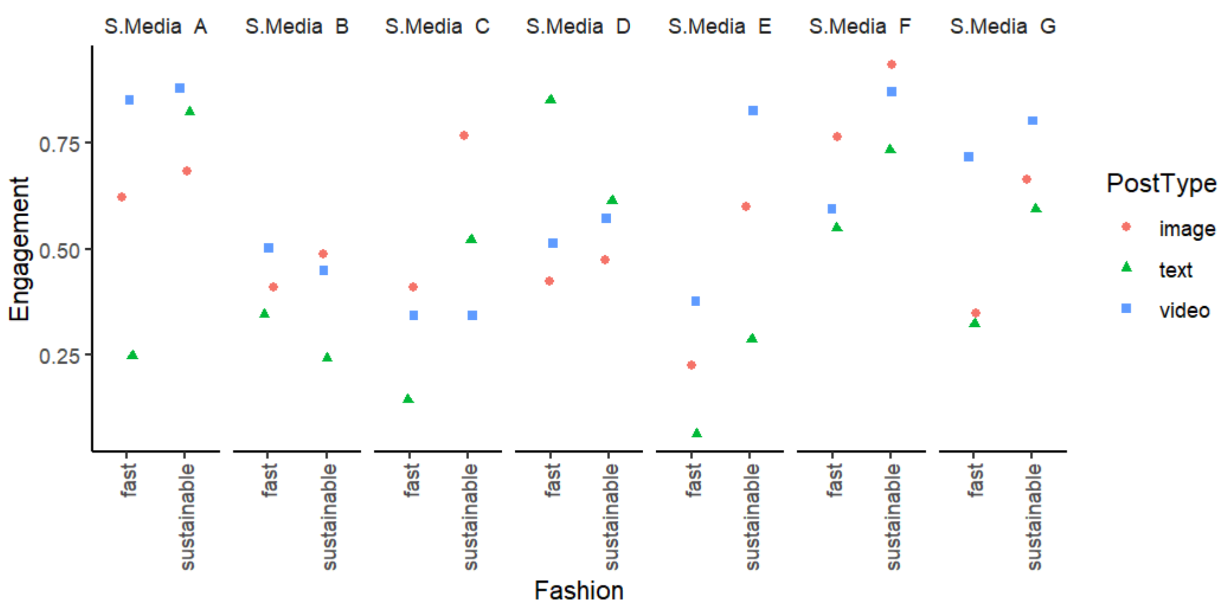
lower C, D, and G. Residual outliers are seen in Social Media A and Social Media D.

Plotting Fashion against Engagement in a box plot shows that fast has a slightly lower average than sustainable.

The boxplot PostType against Engagement shows that text posts may generally have a lower engagement level.

1b)

Plotting the graphs



Observations:

Social media A: For fast fashion the PostType engagement has a larger spread whereas sustainable fashion engagement is high (above 0.5) regardless of PostType.

Social media B: Roughly similar engagements for all PostTypes and types of fashion.

Social media C: Fast fashion has noticeably less engagement than sustainable fashion with sustainable image posts having the highest engagement.

Social media D: Fast fashion textual posts have the highest engagement.

Social media E: Fast fashion textual posts have little to 0 engagement while sustainable fashion posts have a large spread with videos being the highest engagement.

Social media F: Generally high engagement for fast fashion and sustainable fashion with sustainable drawing the most engagement.

Social media G: Videos on this social media type have the most engagement, text and video posts have more engagement with sustainable fashion.

Fixed effects: Fashion

Random effects: SocialMediaType, PostType

Reponse: Engagement

PostType is nested within SocialMedia

1c)

Appropriate model random slopes random intercepts:
Engagement ~ Fashion + (1 + PostType | SocialMedia)

Summary of the model

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
Formula: Engagement ~ Fashion + (1 + PostType | SocialMedia)
   Data: marketing_df

REML criterion at convergence: -16.4

Scaled residuals:
     Min       1Q   Median       3Q      Max
-1.60205 -0.63267  0.01804  0.48381  1.89376

Random effects:
 Groups      Name          Variance Std.Dev. Corr
 SocialMedia (Intercept)   0.01296  0.1139
             PostTypetext  0.02062  0.1436    0.12
             PostTypevideo 0.01278  0.1130   -0.07 -0.36
 Residual                  0.01993  0.1412
Number of obs: 42, groups:  SocialMedia, 7

Fixed effects:
                   Estimate Std. Error       df t value Pr(>|t|)
(Intercept)         0.47447    0.05546  8.30470   8.555 2.14e-05 ***
Fashionsustainable  0.16810    0.04356 20.00015   3.859 0.000978 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr)
Fashnsstnbl -0.393
```

The linear mixed model fitted to this data using random intercepts and random slopes shows high significance for both sustainable fashion (0.0009) and fast fashion (2.14e-05). The significance shows that the estimates are accurate and meaningful so Sustainable fashion has more engagement on average by 0.17 units.

Research question: Which type of fashion (fast or sustainable) results in the highest engagement?

Sustainable fashion results in the highest engagement.

1d)

```
                       (Intercept)                        PostTypetext
                        0.45715757                         -0.09626071
                      PostTypevideo                 Fashionsustainable
                        0.09912473                          0.20076250
 PostTypetext:Fashionsustainable PostTypevideo:Fashionsustainable
                       -0.01711678                         -0.08086606
```

The estimated engagement score for sustainable fashion is 0.45715757 + 0.20076250 = 0.65792007 = 0.66

```
$SocialMedia
              (Intercept) PostTypetext PostTypevideo
S.Media  A   0.0838925091  0.007732382    0.08165089
S.Media  B  -0.0889996756 -0.043727723   -0.02344228
S.Media  C  -0.0473504193 -0.048205184   -0.11981146
S.Media  D  -0.0012797389  0.150522994   -0.03110357
S.Media  E  -0.1062794133 -0.095479939    0.05088351
S.Media  F   0.1590841337  0.022594708   -0.03416874
S.Media  G   0.0009326042  0.006562763    0.07599165

with conditional variances for "SocialMedia"
```

The most engaging PostType on average is dependent on the Social Media type:

Video is most engaging on Social Media A, E, and G

Text is most engaging on Social Media D, and F

Image is most engaging on Social Media B, and C

2a)

```
   Individual  Week VO2max
        <dbl> <dbl>  <dbl>
 1          1     0   21.4
 2          1     4   25.1
 3          1     8   28.1
 4          1    12   28.7
 5          1    16   28.9
 6          2     0   20.6
 7          2     4   25.1
 8          2     8   24.6
 9          2    12   28
10          2    16   32.2
# ℹ 125 more rows
```

Having the data loaded the variables to work with are…

Individual: an identifier containing an integer value.

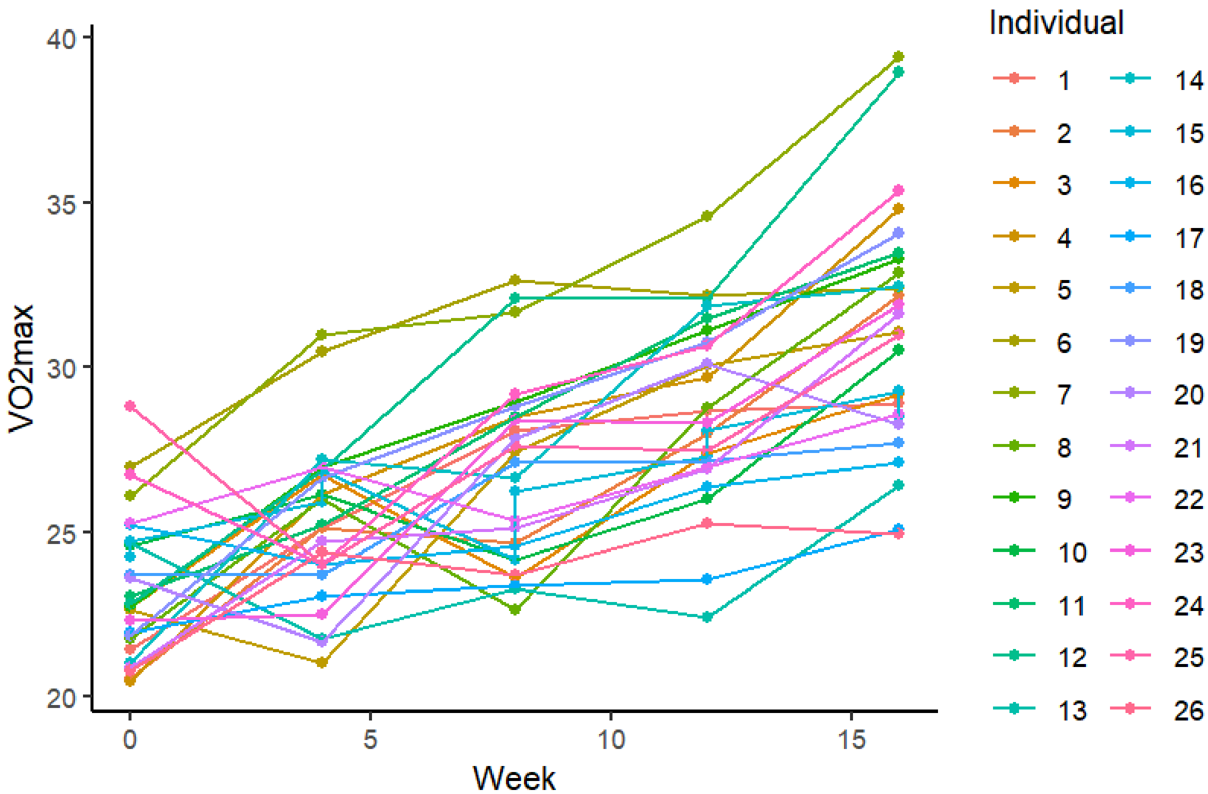
Week: When the individuals cardiovascular fitness was measured, number unit.

Vo2max: Oxygen uptake containing a float or integer.

Individual needs to be changed as a factor since it's a categorial identifier, week needs to be changed to a factor since it's also a categorical identifier that's measuring the result.

fitness_df$Individual<-as.factor(fitness_df$Individual)

fitness_df$Week<-as.factor(fitness_df$Week)



2b)

Fixed effects: Week

Random effects: Individual

Response: Vo2max

Scale the variables

```
scaled_df <- fitness_df %>%
  mutate(across(where(is.numeric), scale))
```

Random slope model:

```
fitness.m1 <- lmer(VO2max ~ 1 + Week + (1 | Individual), data=fitness_df)
```

Random intercept model:

fitness.m2 <- lmer(VO2max ~ 1 + Week + (0 + Week | Individual), data=fitness_df)

Random slope and intercept model:

fitness.m3 <- lmer(VO2max~ 1 + Week + (1 + Week | Individual), data=fitness_df)

1c)

```
Scaled residuals:
     Min       1Q   Median       3Q      Max 
-1.60849 -0.35547 -0.06369  0.36066  1.53313 

Random effects:
 Groups     Name        Variance Std.Dev. Corr             
 Individual (Intercept) 0.27697  0.5263                    
            Week4       0.50085  0.7077   -0.59            
            Week8       0.51965  0.7209   -0.41  0.57      
            Week12      0.68950  0.8304   -0.55  0.75  0.93
            Week16      1.08105  1.0397   -0.44  0.69  0.83  0.91
 Residual               0.05128  0.2264                    
Number of obs: 135, groups:  Individual, 26
```

Correlation for random effects:

Week0 (baseline)

-0.59 correlation with week 4
-0.41 correlation with week 8
-0.55 correlation with week 12
-0.44 correlation with week 16

Week4:

0.57 correlation with week 8
0.75 correlation with week 12 - Early gain of VO2max continues to increase
0.69 correlation with week 16

Week8:

0.93 correlation with week 12 - Strong consistency in mid to late week gain
0.83 correlation with week 16

Week12:

0.91 correlation with week 16 - late stage gain is strongly connected

These results show that people with a lower baseline VO2max tend to improve more on average, improvements in VO2max are consistent in the later weeks.

2d)

fitness.m4 <- lmer(VO2max ~ 1 +Week + (1 | Individual) + (0 + Week | Individual), data=scaled_df)

2e)

```
> AIC(fitness.m1, fitness.m2, fitness.m3, fitness.m4)
           df      AIC
fitness.m1  7 279.0604
fitness.m2 21 258.0054
fitness.m3 21 258.0054
fitness.m4 22 260.0055
```

From the AIC results the lowest AIC models are fitness.m2, fitness.m3, therefore either of these models are okay, picking the simplest model gives fitness.m2 with random slopes only.

Therefore model fitness.m2 has the best fit. Although I will choose model fitness.m3 as it has random intercepts and random slopes as I believe that people start at different fitness levels and also improve at different rates.

2f)

```
> fixef(fitness.m3)
(Intercept)       Week4       Week8      Week12      Week16
 -0.9667571   0.5098189   0.9205591   1.3913889   2.0399117
```

This model suggests that overtime the VO2max is increasing from week0 all the way through till week16.

The intercept is a negative number which is due to centering or standardizing meaning -0.97 represents the average standardized VO2 at week0

```
> ranef(fitness.m3)
$Individual
   (Intercept)        Week4       Week8       Week12     Week16
1  -0.36112420  0.266805544  0.48925337  0.387222371 -0.1017763
2  -0.64281605  0.596011178  0.15240238  0.492251674  0.8163262
3  -0.17617225  0.453039469 -0.51947852 -0.243410581 -0.3186212
4  -0.58552771  0.770355904  0.86690621  1.048935184  1.4351726
5  -0.20545461 -0.653883976  0.43176131  0.308714129  0.2118381
6   0.90002850  0.235073844  0.36077269  0.054914013 -0.4099208
7   0.61588957  0.748646005  0.68815875  0.878221061  1.4676404
8  -0.45576405  0.656800895 -0.28599935  0.287069508  0.7988061
9  -0.11498699  0.507075655  0.61753914  0.704495907  0.7051414
10  0.24375678 -0.095301187 -0.87148480 -0.789363743 -0.5068365
11 -0.08741984  0.137130106  0.56890093  0.640661410  0.7090336
12 -0.05334990  0.464578490  1.20293144  1.205113240  1.9134671
13  0.27956938 -1.160119158 -1.28835268 -1.614607721 -1.6186902
14 -0.54280934  1.009328037  0.64912332  1.062555407  0.9700923
15  0.25501212 -0.004594736 -0.63324438 -0.576977061 -0.8133631
16  0.37374286 -0.690644895 -0.89402590 -1.055286865 -1.3702092
17 -0.29562555 -0.331398301 -0.72444126 -0.857881570 -1.2561797
18  0.10395827 -0.510330551 -0.17243120 -0.461931733 -0.9140296
19 -0.30403109  0.620343892  0.74092453  0.878753060  1.0453169
20  0.02478291 -0.770860050  0.26526958 -0.001644596 -0.6121327
21 -0.54203847  0.375957122  0.06478306  0.254303761  0.5564445
22  0.42598776 -0.112737487 -0.79299870 -0.841744382 -1.0575294
23 -0.19900176 -0.418157756  0.44266175  0.236029084  0.3376070
24  0.70507902 -0.880098758 -0.02590935 -0.235425881  0.2607960
25  1.18026122 -1.436635654 -0.97722476 -1.432153425 -1.3001579
26 -0.54194660  0.223616369 -0.35579756 -0.328812250 -0.9482356
```

Individual #17 has the large negative change from week 0 to week 16
(-0.296), (-1.256) the slope estimate for this individual is −0.0865 meaning a decrease of
-0.0865 VO2max units a week from the group average.