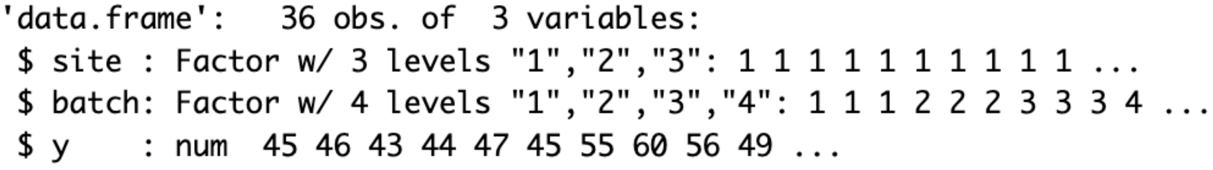
Priyasri Sankaran
Project 3

1)(a) Distinguish between fixed and random effects in modeling experimental data. Explain briefly, when each should be used.

Fixed effect Model	Random Effect Model
A fixed effects is when we are trying to study a specific treatments or groups or conditions that are of direct interest. The explanatory variables represent all levels we are interested in. The study is specific. The variables are fixed, not picked at random. Our goal is to estimate and compare the means across those specific levels. Our conclusion applies only to these particular levels.	Random effects are used when the levels in our study represent a random sample from a larger population. We are not interested in the specific levels themselves, but rather in understanding the variability they introduce and making inference about the population.
When to use: When we are interested in specific levels or groups or treatments in our study, we want to make conclusions about those particular levels.	When to use: when our levels are randomly sampled from a larger population, and we want to generalize our findings beyond the specific-levels in our study to the broader population they represent.
Example: If we select specific schools and want to study about their characteristics, and want to make conclusions about these specific school. We can use fixed effect.	Example: If we want to randomly pick the schools from a school district and want to make a generalize to all schools in the district, we use random effect.

(b) A company is studying the variability in tensile strength of the steel beams that it produces. The beams are produced at different sites, and some of the variability may be due to differences between sites. Another source of variability could be differences between batches of steel used to produce the beams. The company randomly selected 3 sites and then selected four batches at random at each site. From the production of each batch three beams have been selected at random and their tensile strengths have been measured as attached in the end.

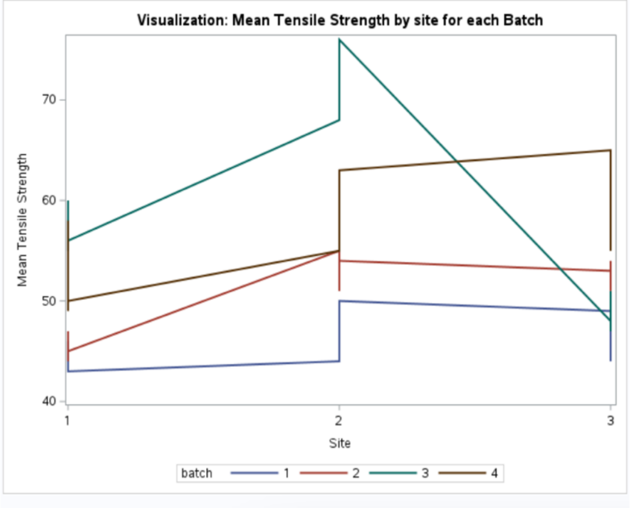
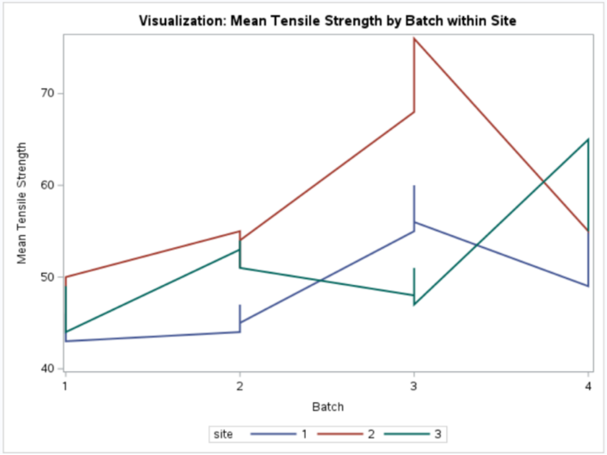
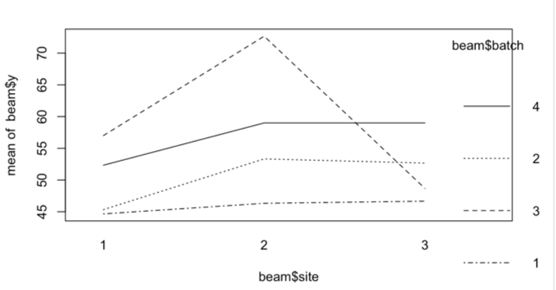
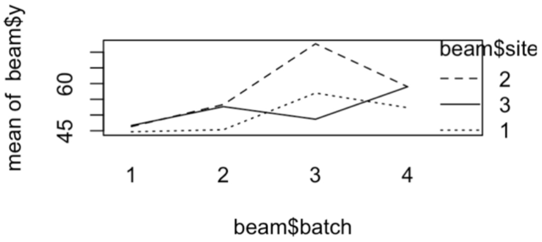
(i) Suggest a model to analyze the data. Interpret each of the terms in this model and state clearly the assumptions needed to conduct an analysis.



```
'data.frame':	36 obs. of  3 variables:
 $ site : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
 $ batch: Factor w/ 4 levels "1","2","3","4": 1 1 1 2 2 2 3 3 3 4 ...
 $ y    : num  45 46 43 44 47 45 55 60 56 49 ...
```

Exploratory Data Visualization:

Before proceeding the hypothesis testing, we created visualization plots to explore patterns in tensile strength variability across sites and batches.

 Tensile Strength by site for each Batch	This plot reveals substantial variation in how batches behave across different sites. Batch3(green line) shows an unusual peak with tensile strength increasing sharply from site 1 to site 2, reaching ~75, then decreasing at site 3 ~48. Batch 4(brown line) demonstrates more stability across sites but still shows an upward trend. Batches 1 and 2 (blue and red lines) display more moderate variations between sites, with batch 1 maintaining the most consistent values around ~44-50. We noticed across all batches an increase from site 1 to site 2 suggesting site 2 may produce high tensile strength values. The lack of parallel lines suggesting different batches respond differently at each site.
	This plot reveals how tensile strength varies across batches within each site. Site 2(red line) shows a peak at batch 3~around 75. Higher than the other batches at this site. Site 1(blue line) shows a moderate increase at batch 3 followed by a decrease at batch 4. Site 3(green line) shows a unique pattern with a dip at batch 3, then a sharp increase at batch 4. The substantial difference in tensile strength patterns across batches within the same site suggest batch-to-batch variation is an important factor.
R result 	

Given: Company randomly selects 3 sites and then selects four batched at random at each site. This is a balanced two-factor random- nested model(B nested within A).

The batch1 at site 1 is completely different from batch 1 at site 2. For this reason traditional interactions doesn't make any sense here in site * batch.

The correct nesting structure is batch(site). Batches are nested within sites. From the problem statement, we can justify this. The two factors, site and batch are not crossed. We ignore the interaction term. We can conclude that the batch(site) nested model.

The Model:

$$Y_{ijk} = \mu_{..} + \alpha_i + \beta_{j(i)} + \epsilon_{ijk}$$

Where:

Y_{ijk} =tensile strength of the kth beam from the jth batch at the ith site

$\mu_{..}$=overall mean tensile strength

α_i= random effect of site i(i=1,2,3)

$\beta_{j(i)}$= random effect of batch j nested within site i(j=1,2,3,4)

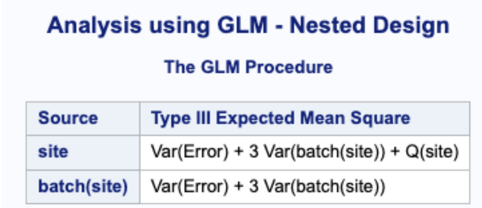
ε_{ijk}= random error term(k=1,2,3)

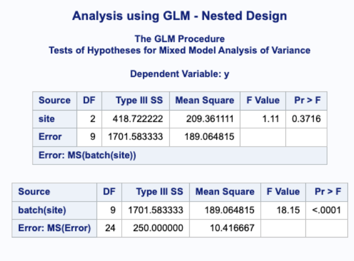
Assumptions:

- Site effects α_i main effect
- α_i and $\beta_{j(i)}$are independently distributed and $N(0,\sigma^2_\alpha)$ and as $N(0,\sigma^2_{\beta(\alpha)})$
- Error terms ε_{ijk} are independently distributed as $N(0, \sigma^2)$
- All random effects are pairwise independent of each other.

(ii) Complete this analysis and report on the importance of the two possible sources of variability.

Your report should contain details of how any necessary estimates have been made and of any hypotheses that have been tested.

<table>
<tr>
<td>
<b>Analysis using GLM - Nested Design</b>
<b>The GLM Procedure</b>
<table>
<tr><th>Source</th><th>Type III Expected Mean Square</th></tr>
<tr><td>site</td><td>Var(Error) + 3 Var(batch(site)) + Q(site)</td></tr>
<tr><td>batch(site)</td><td>Var(Error) + 3 Var(batch(site))</td></tr>
</table>
</td>
<td>Expected Mean Square table:
The difference between these 2 are Q(site) term.</td>
</tr>
<tr>
<td>$H_0 : \sigma^2_\alpha =0$ (there is no variation between sites)
$Ha : \sigma^2_\alpha >0$ (There is significant variation between sites)</td>
<td>Site Variance: σ^2_α
P-Value=0.46>0.05 alpha level significance.
The site effect is not significant(F=1.11, P-value=0.3716 >0.05
$F_{critical(.05,2,9)}$=4.26
Fstat<Fcritical
We fail to reject the null hypothesis. There is lack of evidence to conclude that variation between sites is significantly different from zero.</td>
</tr>
<tr>
<td>$H_0 : \sigma^2_{\beta(\alpha)} =0$(there is no variation between sites)
$Ha : \sigma^2\sigma^2_{\beta(\alpha)} >0$ (There is significant variation between sites)</td>
<td>Batch(site)Variance: $\sigma^2_{\beta(\alpha)}$
P-value =0.023<0.05 alpha level of significance.
batch(site)effect F=18.15, p-value=0.0001 is highly significant.
We reject the null hypothesis. There is a strong evidence that batches within sites vary significantly. We reject the null hypothesis.</td>
</tr>
</table>



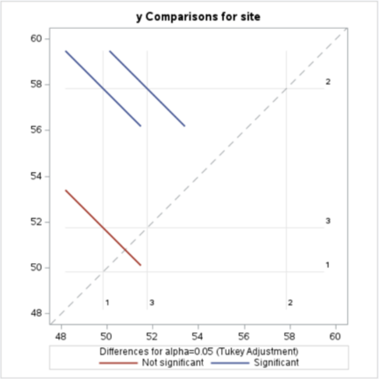
Analysis using GLM - Nested Design

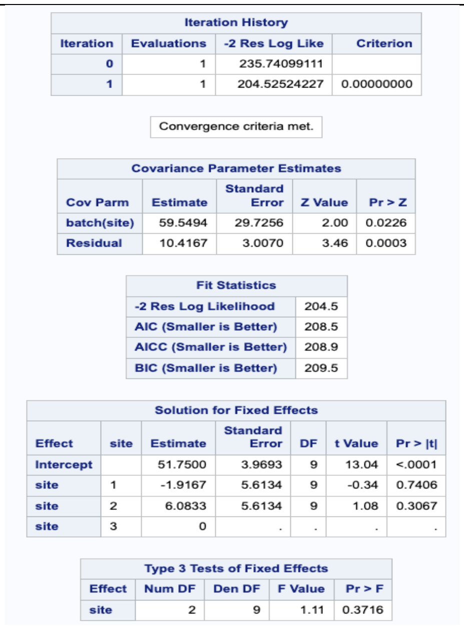
The GLM Procedure
Tests of Hypotheses for Mixed Model Analysis of Variance

Dependent Variable: y

Source	DF	Type III SS	Mean Square	F Value	Pr > F
site	2	418.722222	209.361111	1.11	0.3716
Error	9	1701.583333	189.064815		
Error: MS(batch(site))					

Source	DF	Type III SS	Mean Square	F Value	Pr > F
batch(site)	9	1701.583333	189.064815	18.15	<.0001
Error: MS(Error)	24	250.000000	10.416667		





Iteration History			
Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	235.74099111	
1	1	204.52524227	0.00000000

Convergence criteria met.

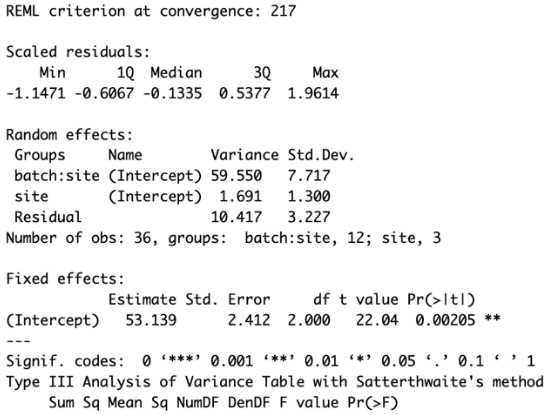
Covariance Parameter Estimates				
Cov Parm	Estimate	Standard Error	Z Value	Pr > Z
batch(site)	59.5494	29.7256	2.00	0.0226
Residual	10.4167	3.0070	3.46	0.0003

Fit Statistics	
-2 Res Log Likelihood	204.5
AIC (Smaller is Better)	208.5
AICC (Smaller is Better)	208.9
BIC (Smaller is Better)	209.5

Solution for Fixed Effects						
Effect	site	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		51.7500	3.9693	9	13.04	<.0001
site	1	-1.9167	5.6134	9	-0.34	0.7406
site	2	6.0833	5.6134	9	1.08	0.3067
site	3	0	.	.	.	.

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
site	2	9	1.11	0.3716

Same result is obtained from R



```
REML criterion at convergence: 217

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.1471 -0.6067 -0.1335  0.5377  1.9614

Random effects:
 Groups     Name        Variance Std.Dev.
 batch:site (Intercept) 59.550   7.717
 site       (Intercept)  1.691   1.300
 Residual               10.417   3.227
Number of obs: 36, groups:  batch:site, 12; site, 3

Fixed effects:
            Estimate Std. Error     df t value Pr(>|t|)
(Intercept)   53.139      2.412  2.000   22.04  0.00205 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Type III Analysis of Variance Table with Satterthwaite's method
     Sum Sq Mean Sq NumDF DenDF F value Pr(>F)
```

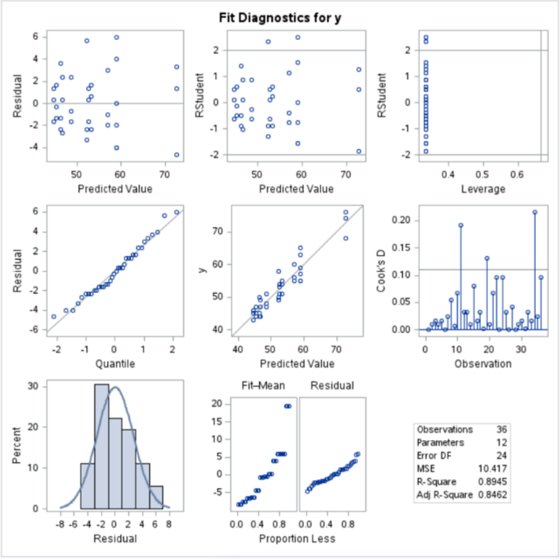
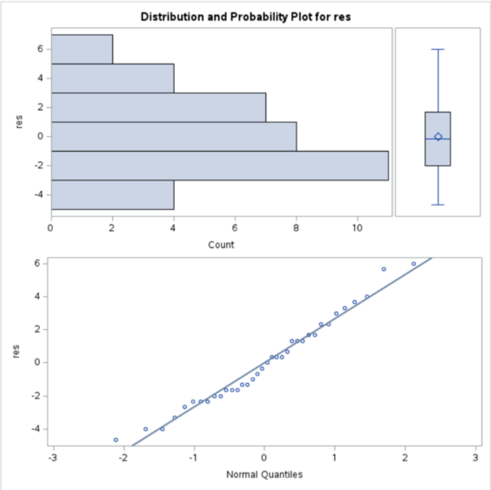
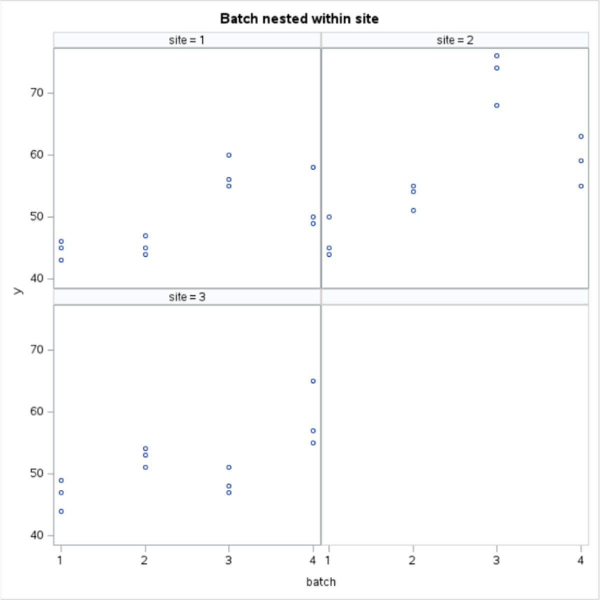
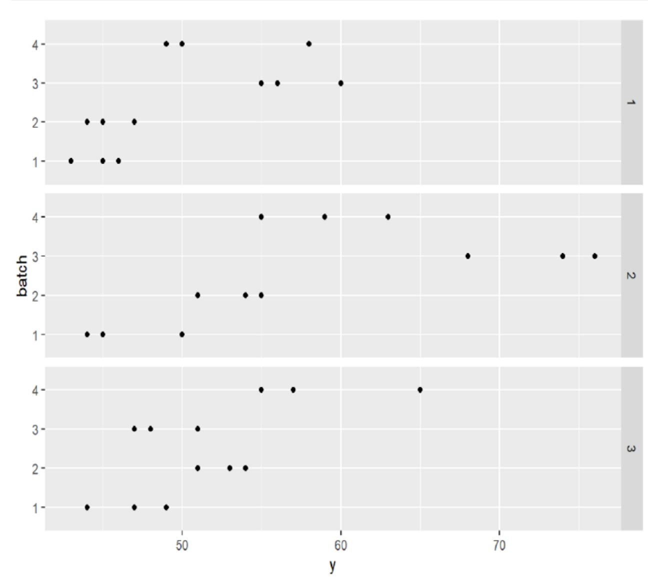
The final model we have is:

$y_{ijk} = \mu_{..} + \beta_{j(i)} + \varepsilon_{ijk}$ where i=1,2,3; j=1,2,3,4; k=1,2,3

This indicates that the main source of variability in tensile strength is differences between batches within sites, while the differences between sites do not contribute significantly to the variability.

Diagnostic plot: On the residual plot, we don’t see any pattern. The residuals are randomly scattered, suggesting the model assumptions are valid.

Q-Q-Plot- The Q-Q Plot shows that the data points are along the line. We do not see any outliers. Normality hold. The histogram also shows that it’s almost normal.

<table>
<tr>
<td>
</td>
<td>
</td>
</tr>
<tr>
<td colspan="2">On the plot to the left shows a visual evidence supporting a potential batch effect. There is variablity in tensile strength across sites. Different batched show different trends, which indicates that batch effects play a role in the variablity.</td>
</tr>
<tr>
<td>SAS code</td>
<td>R code</td>
</tr>
<tr>
<td>
</td>
<td>
</td>
</tr>
</table>

2) An experiment is to be conducted involving five treatments A-E, and there are enough units available to replicate each treatment five times. However, the experimenter can only deal with five units each day and therefore intends to spend five days on the experiment. There may be systematic differences between days, and also differences due to the order in which treatments are carried out each day.

(a) Explain how a Latin square design can be used to eliminate systematic variation, and write down the linear model that is the basis for analyzing data from this experiment, stating the properties of each term in it.

In this experiment, there are 5 treatments(A,B,C,D,E), 5 days, and five slots (orders) on each day. The experimenter wants to control for potential systematic variation caused by both the day on which the treatment is applied and the order in which the treatments are applied each day.

- Each treatment appears exactly once in each row(representing day) and exactly once in each column(representing order within the day)

The design structure ensures that the effects of each day and order are balanced and orthogonal to the treatment effects, which minimizes confounding and improves the accuracy of treatment comparisons.

The Linear model for a Latin square design is:

$$y_{ijk} = \mu\ldots+\rho_i+k_j+\tau_k+\varepsilon_{ijk}$$

Where:

- y_{ijk} is the observed response when for the ith day, jth order and the kth treatment.
- μ is the overall mean effect.
- ρ_i represents the effects of row(day), for i=1,2…,5
- k_j represents the effect of order, for j=1,2…,5
- τ_k represents the effects of treatment, for k=1,2…,5
- ε_{ijk} is the random error term, assumed to follow a normal distribution with mean 0 and variance $\sigma^{2.}$

  In this model the treatment effects are estimated while controlling for variations due to day and order.

(b) The experimenter tries to write down a plan for the week's work and asks you how to construct the necessary design. You show him a table of standard 5*5 Latin squares and he says some of those look rather "systematic". Explain carefully how to choose a square at random from all possible 5*5 Latin squares.

Latin squares appearing "systematic" comes from their structured arrangement. To randomly select a 5X5 Latin square. We can do these steps to avoid unintended patterns that might impact the experiment’s outcome.

- Finding a valid 5x5 Latin square
- Randomizing the selection-instead of picking a Latin Square that looks “too systematic”, you can shuffle and randomly select one from the list.
- Rearrange rows and columns-after selecting a square, you can randomly reorder the rows and columns.

(c) The experiment is finally carried out using the following plan, which also shows the measurement y obtained from each unit.

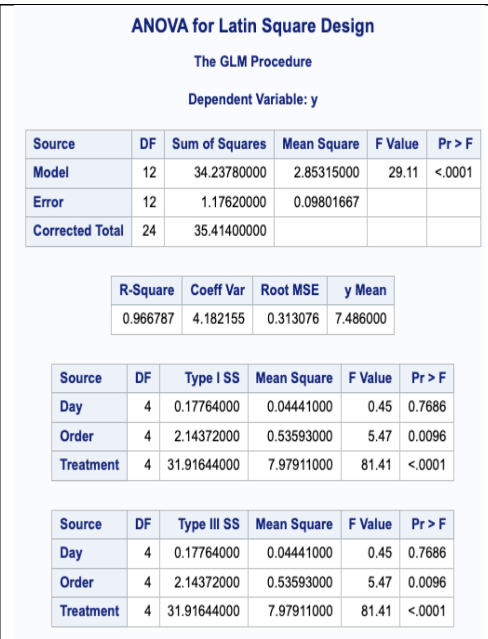
i) Construct the analysis of variance for these data.

Anova Results:

1)Overall Model: Highly significant F=29.11, P-value=0.0001<0.05, R^2=0.967, which suggests that 96.7% of the variation in responses is explained by the model.

2)Main effects:

- Day: not significant (F=0.45, p-value=0.77>0.05, there are no significant difference between days.
- Order: Significant(F=5.47, P-value=0.0096), which is moderate influence due to the position
- Treatment:Highly significant(F=81.41,P-value<0.0001). it's our interest.



ANOVA for Latin Square Design

The GLM Procedure

Dependent Variable: y

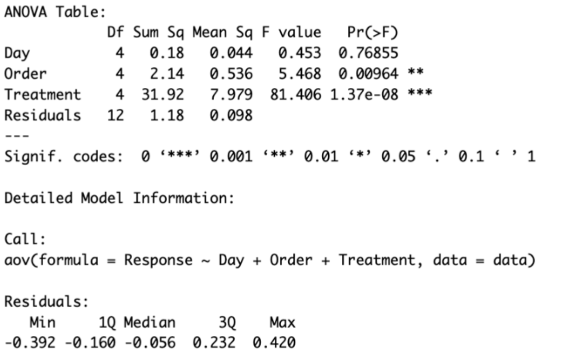
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	34.23780000	2.85315000	29.11	<.0001
Error	12	1.17620000	0.09801667		
Corrected Total	24	35.41400000			

R-Square	Coeff Var	Root MSE	y Mean
0.966787	4.182155	0.313076	7.486000

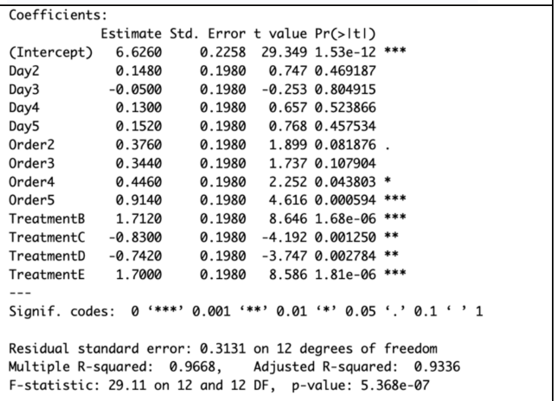
Source	DF	Type I SS	Mean Square	F Value	Pr > F
Day	4	0.17764000	0.04441000	0.45	0.7686
Order	4	2.14372000	0.53593000	5.47	0.0096
Treatment	4	31.91644000	7.97911000	81.41	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Day	4	0.17764000	0.04441000	0.45	0.7686
Order	4	2.14372000	0.53593000	5.47	0.0096
Treatment	4	31.91644000	7.97911000	81.41	<.0001

R code:

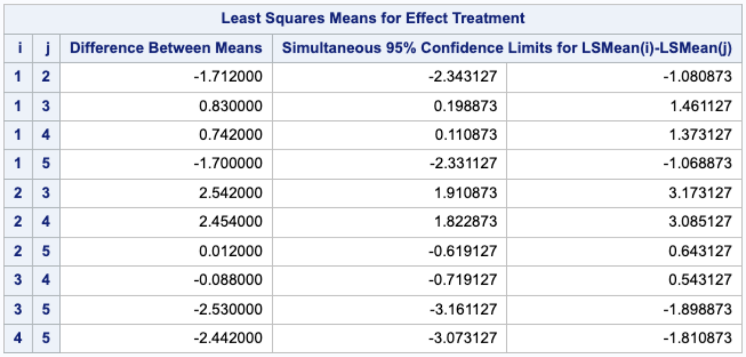


```
ANOVA Table:
            Df Sum Sq Mean Sq F value   Pr(>F)
Day          4   0.18   0.044   0.453  0.76855
Order        4   2.14   0.536   5.468  0.00964 **
Treatment    4  31.92   7.979  81.406 1.37e-08 ***
Residuals   12   1.18   0.098
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Detailed Model Information:

Call:
aov(formula = Response ~ Day + Order + Treatment, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-0.392 -0.160 -0.056  0.232  0.420
```



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.6260     0.2258  29.349 1.53e-12 ***
Day2          0.1480     0.1980   0.747 0.469187
Day3         -0.0500     0.1980  -0.253 0.804915
Day4          0.1300     0.1980   0.657 0.523866
Day5          0.1520     0.1980   0.768 0.457534
Order2        0.3760     0.1980   1.899 0.081876 .
Order3        0.3440     0.1980   1.737 0.107904
Order4        0.4460     0.1980   2.252 0.043803 *
Order5        0.9140     0.1980   4.616 0.000594 ***
TreatmentB    1.7120     0.1980   8.646 1.68e-06 ***
TreatmentC   -0.8300     0.1980  -4.192 0.001250 **
TreatmentD   -0.7420     0.1980  -3.747 0.002784 **
TreatmentE    1.7000     0.1980   8.586 1.81e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3131 on 12 degrees of freedom
Multiple R-squared:  0.9668,    Adjusted R-squared:  0.9336
F-statistic: 29.11 on 12 and 12 DF,  p-value: 5.368e-07
```

ii. Find a familywise 95% confidence interval for $\mu_D - \mu_E$ and $\mu_C - \mu_D$.

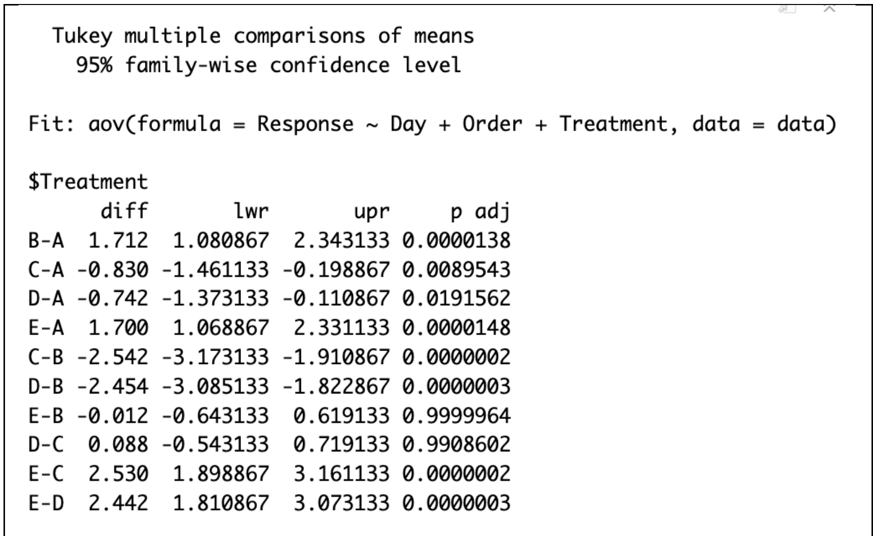


Least Squares Means for Effect Treatment

i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	-1.712000	-2.343127	-1.080873
1	3	0.830000	0.198873	1.461127
1	4	0.742000	0.110873	1.373127
1	5	-1.700000	-2.331127	-1.068873
2	3	2.542000	1.910873	3.173127
2	4	2.454000	1.822873	3.085127
2	5	0.012000	-0.619127	0.643127
3	4	-0.088000	-0.719127	0.543127
3	5	-2.530000	-3.161127	-1.898873
4	5	-2.442000	-3.073127	-1.810873

μ_D-μ_E= -2.44, CI[-3.07,-1.81]This interval does not include 0, so the difference is statistically significant. D is significantly lower than E.

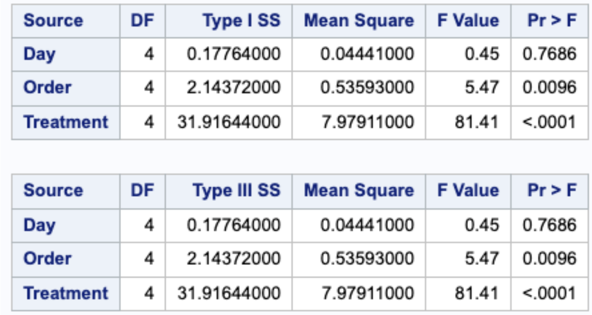
μ_C-μ_D= -0.088, CI:[-0.72,0.54] This confidence interval does include 0. So the difference is not significant. C and D no significant difference.



```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Response ~ Day + Order + Treatment, data = data)

$Treatment
      diff        lwr       upr     p adj
B-A  1.712  1.080867  2.343133 0.0000138
C-A -0.830 -1.461133 -0.198867 0.0089543
D-A -0.742 -1.373133 -0.110867 0.0191562
E-A  1.700  1.068867  2.331133 0.0000148
C-B -2.542 -3.173133 -1.910867 0.0000002
D-B -2.454 -3.085133 -1.822867 0.0000003
E-B -0.012 -0.643133  0.619133 0.9999964
D-C  0.088 -0.543133  0.719133 0.9908602
E-C  2.530  1.898867  3.161133 0.0000002
E-D  2.442  1.810867  3.073133 0.0000003
```

R result

iii. Comment briefly on the Day and Order terms in the analysis.



Source	DF	Type I SS	Mean Square	F Value	Pr > F
Day	4	0.17764000	0.04441000	0.45	0.7686
Order	4	2.14372000	0.53593000	5.47	0.0096
Treatment	4	31.91644000	7.97911000	81.41	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Day	4	0.17764000	0.04441000	0.45	0.7686
Order	4	2.14372000	0.53593000	5.47	0.0096
Treatment	4	31.91644000	7.97911000	81.41	<.0001

Day Effect:
F_{stat}=0.45, P-value=0.77 >0.05, not significant difference across days. The response variable did not systematically vary depending on which the experiment was run.
Order effect:
F_{stat}=5.47, P-value=0.0096<0.05, There is a significant different based on the position. The units might be responding differently as the day progresses.