Statistical Analysis and Visualization

CIT660

Spring 2023

Group H

Post-weaning diarrhea I

Aya Mohamed Hassan

221002289

Supervised by:

Dr. Mohamed Maysara

Eng. Marium Oweda

Eng. Menna Abdelsattar

Table of Content:

Abstract:

That project aim to investigate the impact of various treatments, including ZnO, nutraceuticals, and vaccination, on the average daily weight gain of piglets during different post-weaning periods. The findings of this study will contribute to our understanding of effective control strategies for post-weaning diarrhea, considering the emerging concerns regarding antimicrobial resistance and the potential legislative restrictions on the use of zinc oxide.

Introduction:

Post-weaning diarrhea (PWD) is a prevalent and economically significant disease affecting pigs in piggeries worldwide. This condition leads to various detrimental effects, including increased mortality rates, weight loss, retarded growth, elevated treatment costs, and augmented usage of antibiotics. The primary cause of PWD is believed to be Enterotoxigenic Escherichia coli (ETEC), a type of bacteria commonly found in the intestinal tracts of affected pigs.

Traditionally, antimicrobial agents have been employed to control the disease; however, the emergence of antimicrobial resistance in E. coli has raised concerns and necessitated the exploration of alternative strategies. For instance, researchers have investigated the inclusion of additional dietary fiber, reduction of crude protein levels, and the use of zinc oxide (ZnO) as potential solutions. Notably, the inclusion of ZnO has shown promising effects in mitigating PWD. However, it is worth mentioning that the use of zinc oxide may be restricted by European Union legislation starting from 2022.

Another potential strategy to combat PWD is vaccination of the piglets. In the present study, we aim to investigate the efficacy of vaccination compared to the addition of ZnO and nutraceuticals (such as dietary fibers) to the piglets' feed. The study examines five distinct treatment groups, each consisting of 128 piglets distributed among 40 pens, with 16 piglets per pen. The treatments include variations in feed composition and vaccination protocols.

To evaluate the effectiveness of the different treatments, we focus on three specific outcome measures: average daily weight gain (ADWG) during different post-weaning periods. ADWG0021 represents the average daily weight gain between 0 and 21 days post-weaning, ADWG2150 represents the average daily weight gain between day 21 and day 50 post-weaning, and ADWG0050 represents the average daily weight gain between 0 and 50 days post-weaning.

Reading Data:

PWD.RData contains 9 features which are:

Pen	Treatment	Feeder	Sex	W0	P0	ADWG0021	ADWG2150	ADWG0050
# of pens which contain 16 piglet in one pen	Type of treatment	Feeder ID	Gender type of piglet	Piglet's weight	# of piglets in each pen	Avg daily weight gain(g/day) in the period between 0 to 21 days post-weaning	Avg daily weight gain(g/day) in the period between 21 to 50 days post-weaning	Avg daily weight gain(g/day) in the period between 0 to 50 days post-weaning
1 ,2,…	A,B,C,D,E	1 , 2 , … 20	1 => male 2 => female	Continuous values	16	Continuous values	Continuous values	Continuous values

There are features that we want to focus on which they are Treatment, Sex, W0, ADWG0021, ADWG2150, ADWG0050 as they will give a valuable insights

Descriptive Statistics:

	Pen	Treatment	Feeder	Sex	W0	P0	ADWG0021	ADWG2150	ADWG0050
Min	1	A: 8	1	1	87.5	16	102.7	375	275.6
Max	40	B: 8	20	2	113.5	16	178.6	608.1	416.9
Mean	20.5	C: 8	10.5	1.5	99.17	16	143.1	500.9	350.6
Median	20.5	D: 8	10.5	1.5	99	16	144.3	501.1	349.7
1st Q	10.75	E: 8	5.75	1	90.88	16	129.5	471.4	325.7
3rd Q	30.25		15.25	2	106.75	16	155.9	535	374.5

Categorical Data:

Sex	
Male	Female
20	20

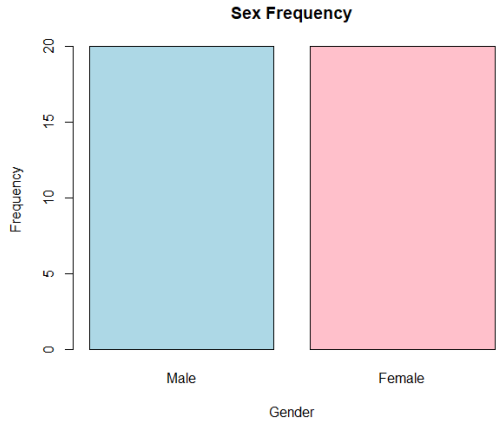
Treatment				
A	B	C	D	E
8	8	8	8	8

Correlation Coefficient:

	ADWG0021 and ADWG2150	ADWG0021 and ADWG0050
Pearson	0.2270356	0.4427165
Spearman	0.2753651	0.4553267

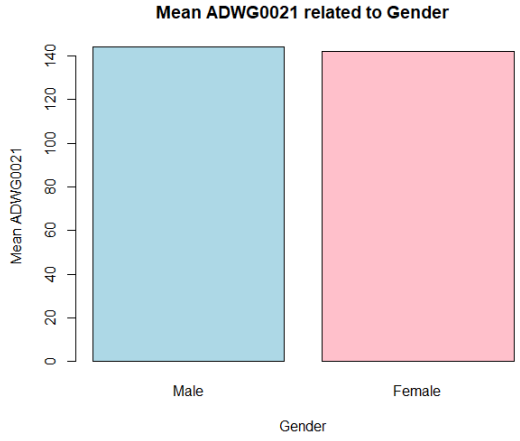
Graphics:

1- Bar Chart of Sex

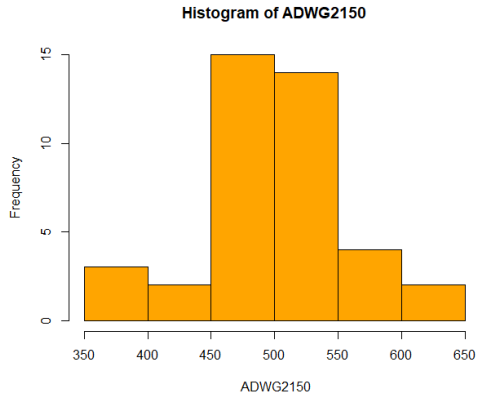


Data is balanced

2- Bar Chart of Mean ADWG0021 in Males and Females

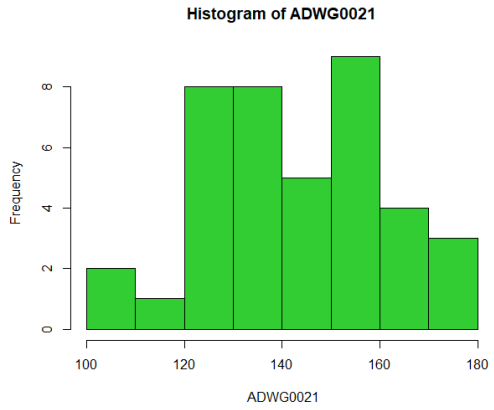


3- Histogram of ADWG2150



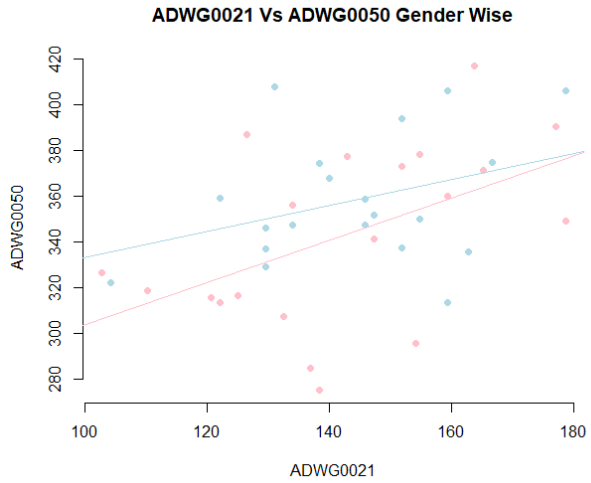
The distribution of data seems to be normal but we will check normality using Shapiro test

4- Histogram of ADWG0021



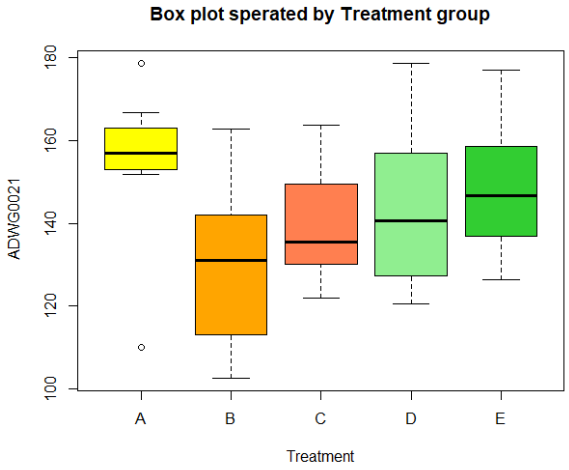
The distribution of data seems to be not normal but we will check normality using Shapiro test

5- Scatter Plot of ADWG0050 Vs ADWG0021 Sex wise



The distribution of the points suggests a negative relationship between ADWG0050 and ADWG0021 for each gender

6- Box plot of ADWG0021 per Treatment



There are 2 outliers in treatment A and show different variance between the treatments and ADWG0021

Outliers Detection:

Exploring data using boxplot for detecting outliers we found

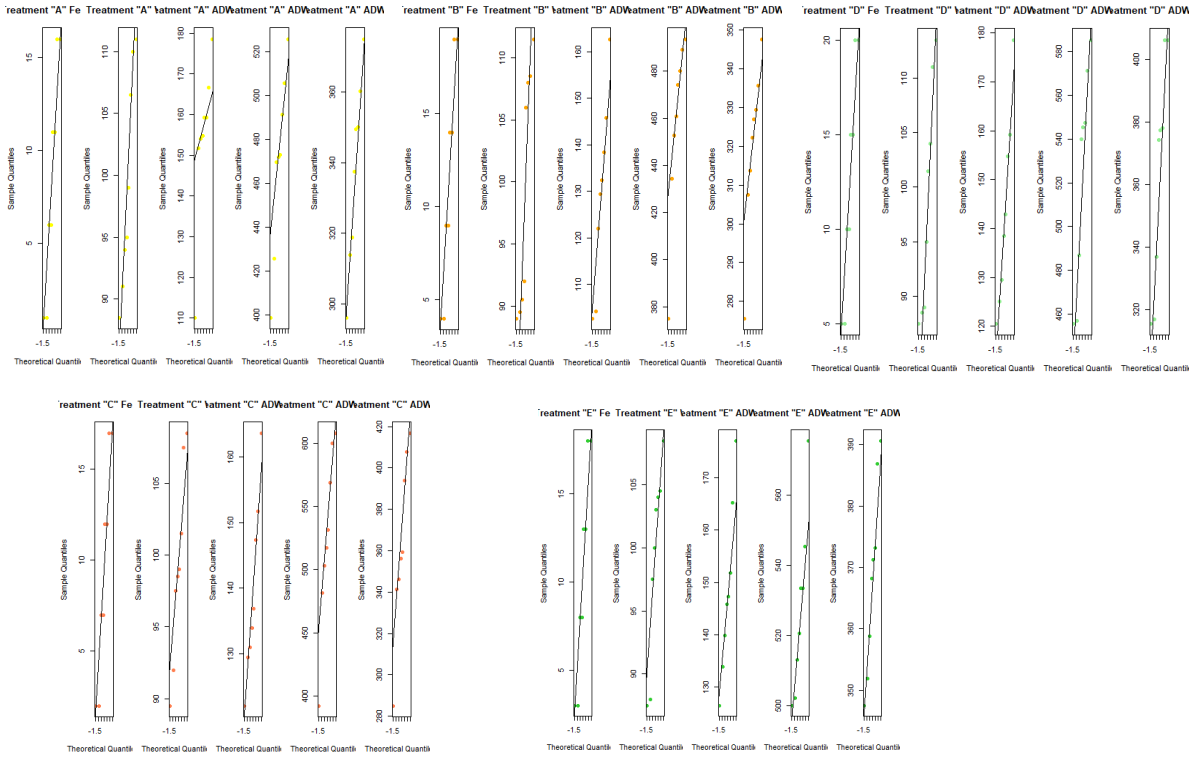
	ADWG0021 and Treatment		ADWG0050 and Treatment	
Category	**A**		**B**	
Outliers Values	178.5714	110.1190	275.625	-

Outliers may provide valuable insights, highlight data quality issues, or indicate rare events in my opinion we don't remove Outliers but investigate them further to understand their nature and potential impact on the analysis instead of removing outliers we could change their values to max. value using boxplot whisker

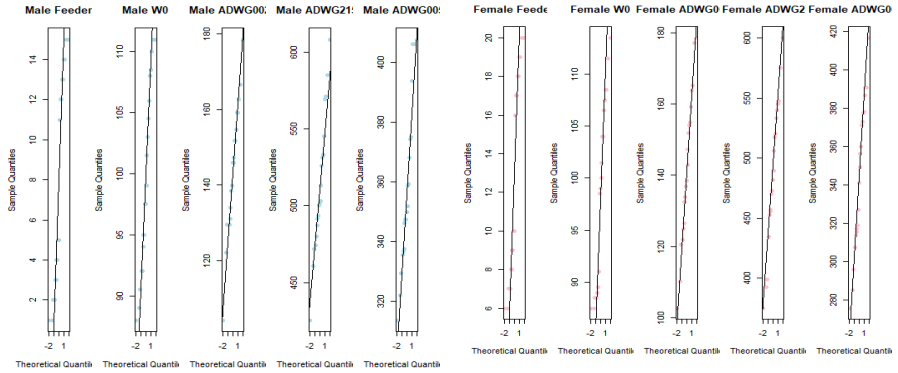
Testing Normality

Q-Q Plot

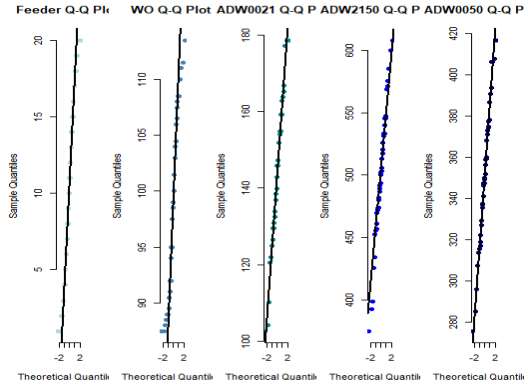
1- Treatment with all the features



2- Sex with all the features



3- Numerical Data



Shapiro Test

Null Hypothesis: Data is normally distributed

Alternative Hypothesis: Data is not normally distributed

1- Sex

P-Value	Feeder	W0	ADWG0021	ADWG0050	ADWG2150	Null Hypothesis	Result
Male	0.005621	0.1768	0.977	0.1895	0.849	Have enough evidence to reject	Not Normal
Female	0.005621	0.02526	0.9513	0.7312	0.9276	Have enough evidence to reject	Not Normal
Conclusion	Sex column is not Normally distributed						

2- Treatment

P-Value	Feeder	W0	ADWG0021	ADWG0050	ADWG2150	Null Hypothesis	Result
A	0.2738	0.3348	0.0395	0. 8976	0.6584	Have enough evidence to reject	Not Normal
B	0.2738	0.03048	0.8132	0.5031	0. 1098	Have enough evidence to reject	Not Normal
C	0.2738	0.6739	0.6954	0.5882	0.6256	Don't have enough evidence to reject	Normal
D	0.2738	0.2784	0.7126	0.1702	0.2102	Don't have enough evidence to reject	Normal
E	0.2738	0.211	0.86	0.7315	0.5288	Don't have enough evidence to reject	Normal
Conclusion	Treatment A & B is not normally distributed while C, D & E are normally distributed						

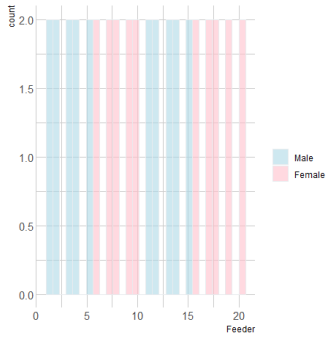
3- Numeric Data

	Feeder	W0	ADWG0021	ADWG0050	ADWG2150
P-Value	0.08873	0.009664	0.7305	0.9276	0.9086
Null Hypothesis	Have enough evidence to reject	Don't have enough evidence to reject	Have enough evidence to reject	Have enough evidence to reject	Have enough evidence to reject
Result	Normal	Not Normal	Normal	Normal	Normal
Conclusion	W0 column is not normally distributed while Feeder, ADWG0021, ADWG0050, ADWG2150 are normally distributed				

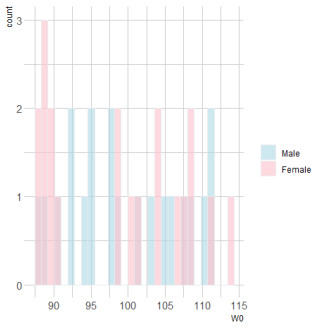
Histogram:

1- Sex

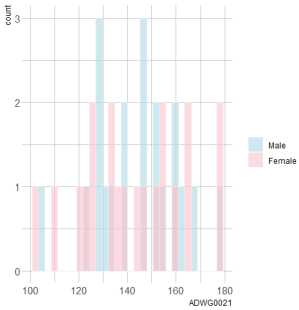
- Feeder



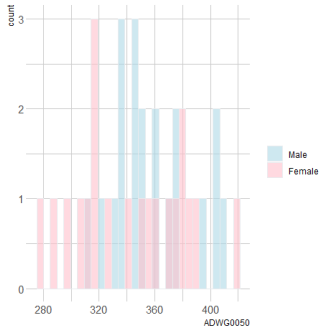
- W0



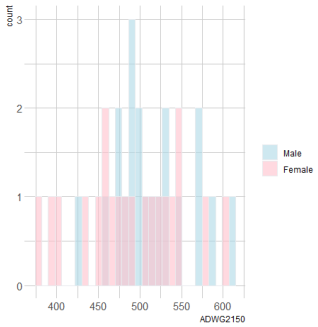
- ADWG0021



- ADWG0050

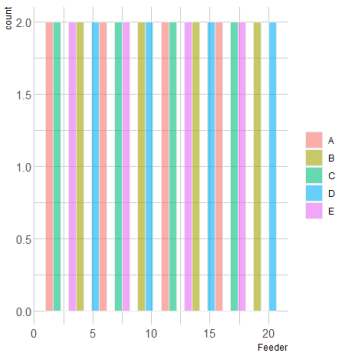


- ADWG2150

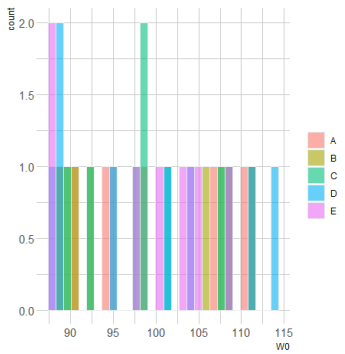


2- Treatment

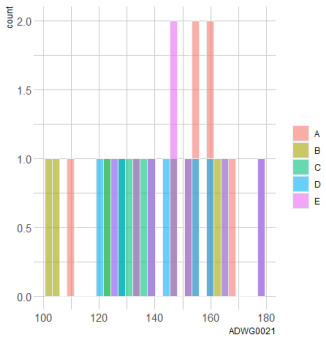
- Feeder



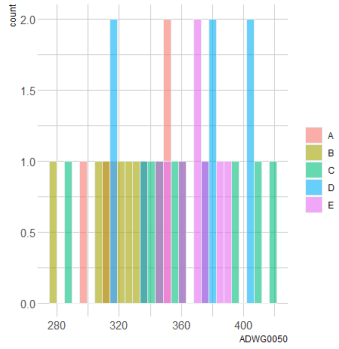
- W0



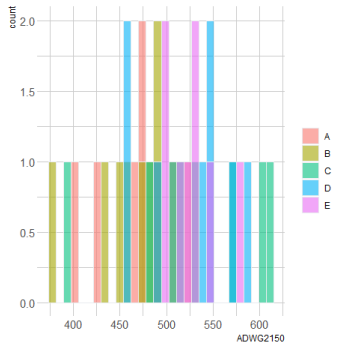
- ADWG0021



- ADWG0050

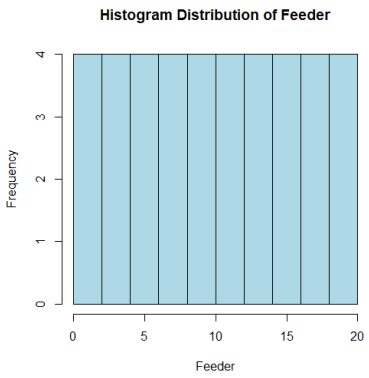


- ADWG2150

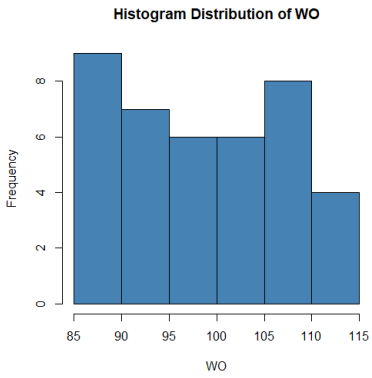


3- Numerical Data

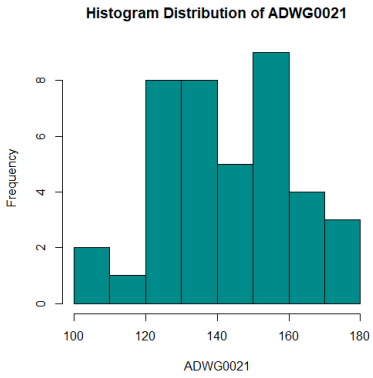
- Feeder



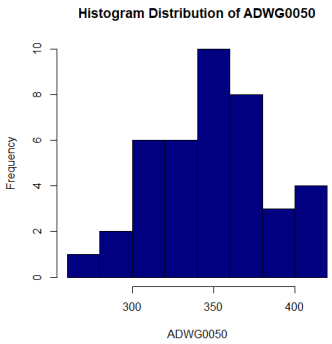
- W0



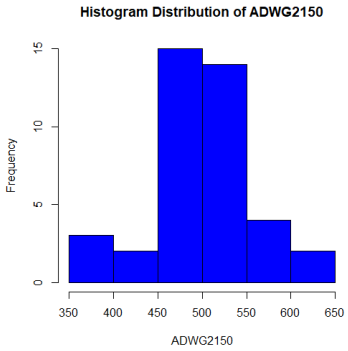
- ADWG0021



- ADWG0050



- ADWG2150

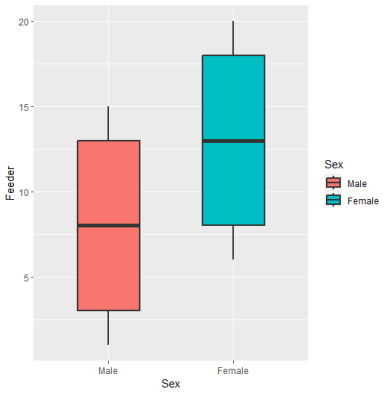


Testing Homoscedasticity

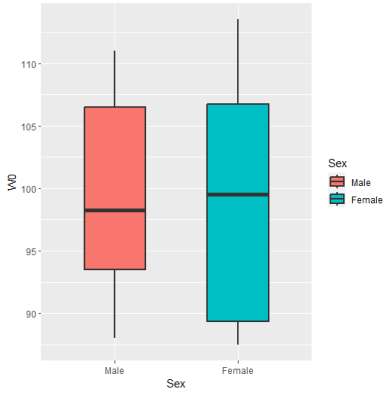
Box Plot

1- Sex

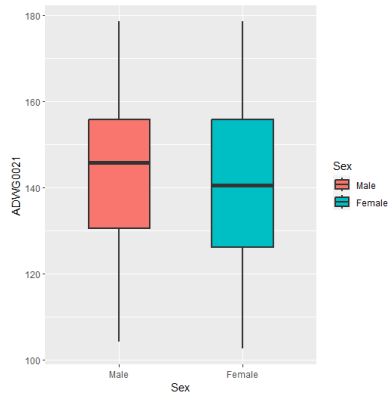
- Feeder



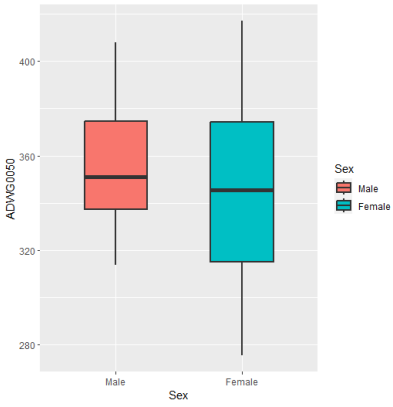
- W0



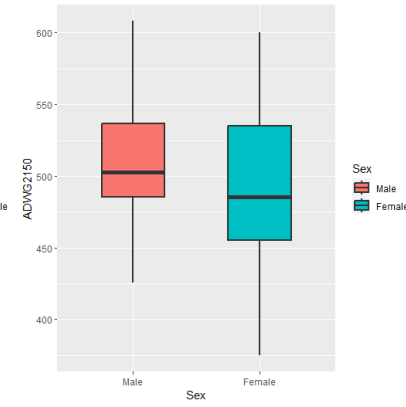
- ADWG0021



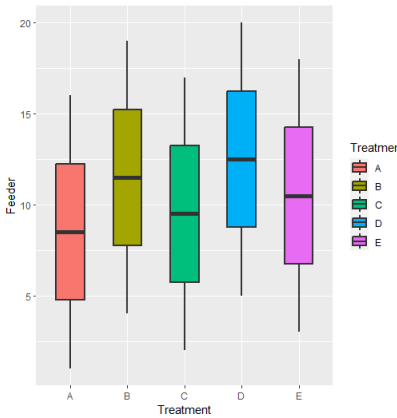
- ADWG0050

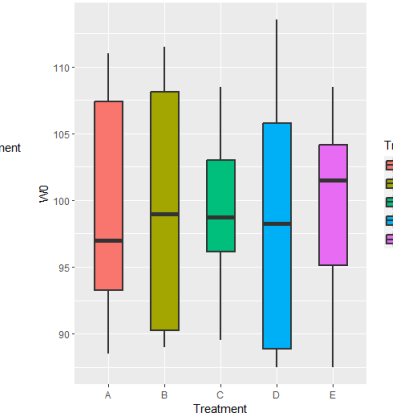


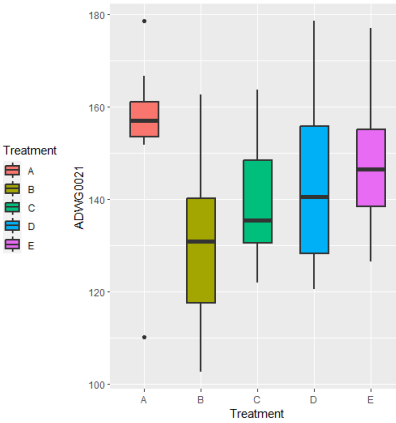
- ADWG2150

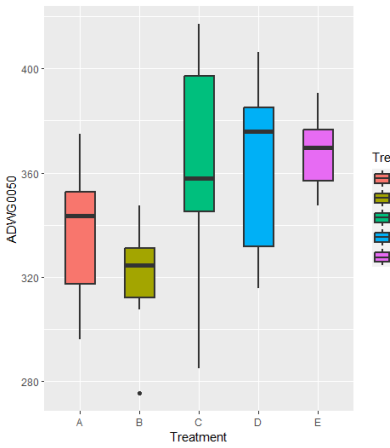


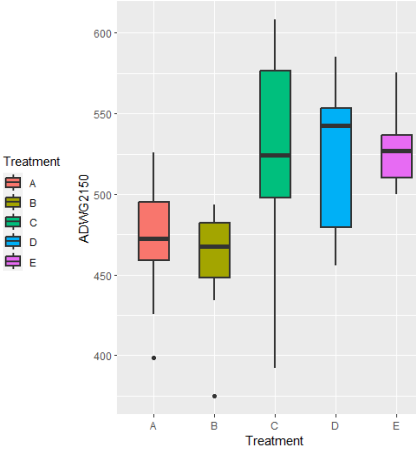
2- Treatment











Levene Test

Null Hypothesis: the variance of the groups being compared are equal (Homogeneity of variance)
Alternative Hypothesis: the variance of the groups being compared are Not equal (Heterogeneity of variance)

1- **Treatment**

	Feeder	W0	ADWG0021	ADWG0050	ADWG2150
p-value	1	0.2265	0.8968	0.2478	0.2864
Null Hypothesis	Don't have enough evidence to reject	Don't have enough evidence to reject	Don't have enough evidence to reject	Don't have enough evidence to reject	Don't have enough evidence to reject
Result	Homo-variance	Homo-variance	Homo-variance	Homo-variance	Homo-variance
Conclusion	all features are Homoscedastic as all results of p-value are greater than 0.05				

2- **Sex**

	Feeder	W0	ADWG0021	ADWG0050	ADWG2150
p-value	1	0.3397	0.3099	0.06383	0.17
Null Hypothesis	Don't have enough evidence to reject	Don't have enough evidence to reject	Don't have enough evidence to reject	Don't have enough evidence to reject	Don't have enough evidence to reject
Result	Homo-variance	Homo-variance	Homo-variance	Homo-variance	Homo-variance
Conclusion	all features are Homoscedastic as all results of p-value are greater than 0.05				

Bartlett test

Null Hypothesis: the variance of the groups being compared are equal (Homogeneity of variance)
Alternative Hypothesis: the variance of the groups being compared are Not equal (Heterogeneity of variance)

1- Treatment

	Feeder	W0	ADWG0021	ADWG0050	ADWG2150
p-value	1	0.8015	0.8498	0.09321	0.1251
Null Hypothesis	Don't have enough evidence to reject	Don't have enough evidence to reject	Don't have enough evidence to reject	Don't have enough evidence to reject	Don't have enough evidence to reject
Result	Homo-variance	Homo-variance	Homo-variance	Homo-variance	Homo-variance
Conclusion	all features are Homoscedastic as all results of p-value are greater than 0.05				

2- Sex

	Feeder	W0	ADWG0021	ADWG0050	ADWG2150
p-value	1	0.4835	0.4192	0.1621	0.209
Null Hypothesis	Don't have enough evidence to reject	Don't have enough evidence to reject	Don't have enough evidence to reject	Don't have enough evidence to reject	Don't have enough evidence to reject
Result	Homo-variance	Homo-variance	Homo-variance	Homo-variance	Homo-variance
Conclusion	all features are Homoscedastic as all results of p-value are greater than 0.05				

Statistical Inference

Equation:

$$CI = \bar{x} \pm z\frac{s}{\sqrt{n}}$$

CI = confidence interval
$\bar{x}$ = sample mean
z = confidence level value
s = sample standard deviation
n = sample size

	Sex	Lower Interval	Upper Interval
90% confidence	Male	139.5	148.6
	Female	136.6	147.7
95% confidence	Male	138.65	149.5
	Female	135.6	148.7
99% confidence	Male	134.98	153.16
	Female	131.17	153.12

Conclusion: When the confidence interval increases the width increase that means 99% confidence have the wider width then 95% lastly with thinner width 90%

Hypothesis testing

a) We hypothesis that ADWG0021 is different between male vs female. Assuming normality and homoscedasticity, can you test this hypothesis using statistical hypothesis framework?

Statistical Q: Is the mean different between Male vs Female in ADWG0021?

Null Hypothesis: The mean of groups Male and Female are equal

Alternative Hypothesis: The mean of groups Male and Female is different

As in this case normality and homoscedasticity were assumed so we will use standard two sample t test directly according to these assumptions, here our data is two sided because it is different in general not greater or smaller than a specific value, also data is independent

Result: P-value = 0.7557 > alpha(0.05)

we do not have enough evidence to reject the null hypothesis in support of alternative hypothesis

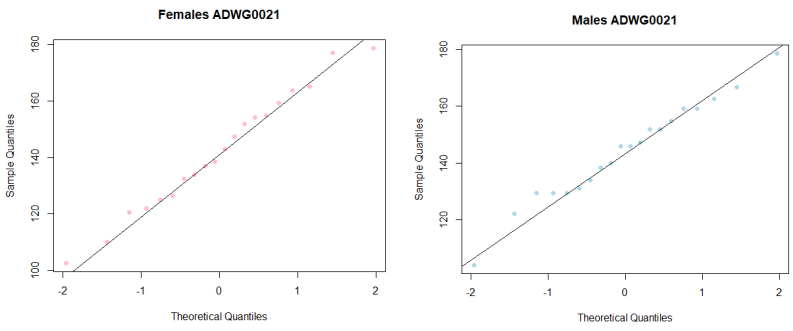
conclusion: the mean of the 2 groups males and females is not different from each other

"ADWG0021 is not different between male vs female".

b) Now we will assess whether the previous test assumptions have been meet for the test

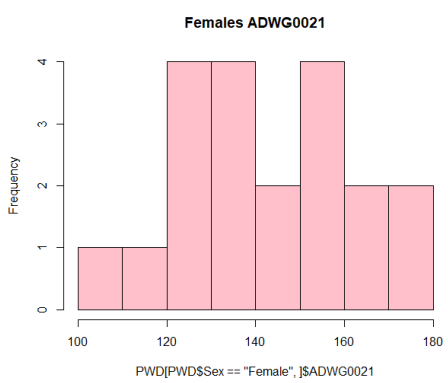
Testing Normality:

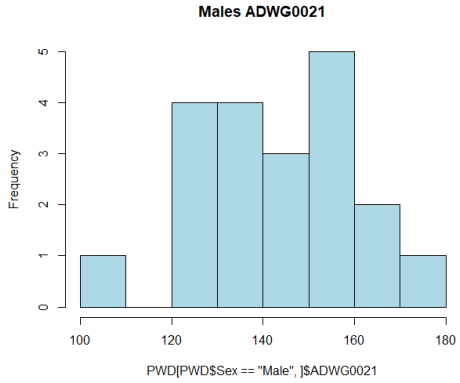
- **Q-Q Plot:**



the line has no deviations from the data points so the data of males is normally distributed

- **Histogram**





- **Shapiro Test**
  p-value Male = 0.9771
  p-value Female = 0.9513
  Conclusion: we do not have enough evidence to reject the null hypothesis in support of alternative hypothesis “Data is normally distributed”

Testing Homoscedasticity

- **Levene Test**
  p-value = 0.3099
  Conclusion: we do not have enough evidence to reject the null hypothesis in support of alternative hypothesis “Data is Homo-variance”
- **F Test**
  p-value = 0.4193
  Conclusion: we do not have enough evidence to reject the null hypothesis in support of alternative hypothesis “Data is Homo-variance”

c) We hypothesis that ADWG0021is "different" in the group receiving Treatment A (normal feed + ZnO) compared to the Treatment B (normal feed + nutraceuticals). Can you test this hypothesis assuming heteroscedasticity?

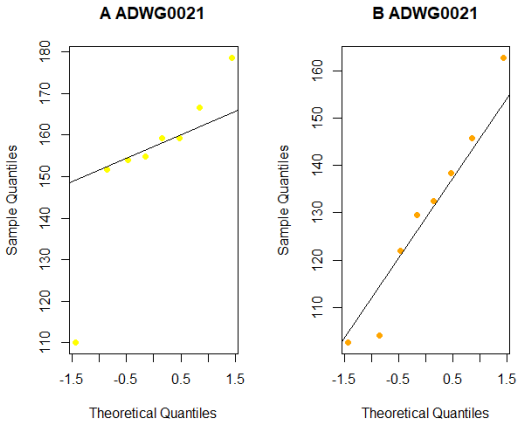
Statistical Q: Is the mean different between treatments A and B in ADWG0021?

Null Hypothesis: mean of treatment A equal to mean of treatment B

Alternative Hypothesis: mean of treatment A different from mean of treatment B

Testing Normality:

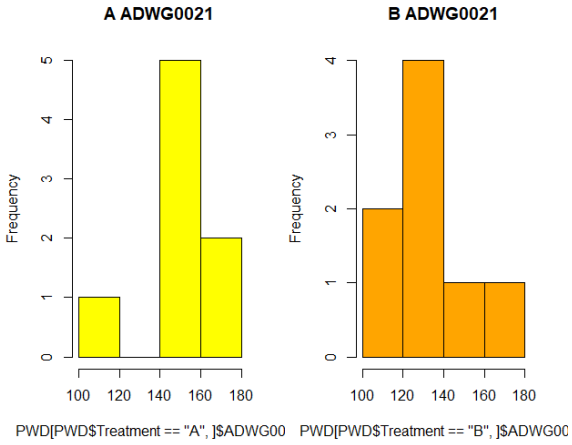
Q-Q Plot



In A, the line has deviations from the data points so the data of treatment A is not normally distributed.

In B, the line has no deviations from the data points so the data of treatment B is normally distributed

Histogram



Shapiro Test

p-value A = 0.0395
Conclusion: we have enough evidence to reject the null hypothesis in support of alternative hypothesis "Data is Not normally distributed"

p-value B = 0.8132
Conclusion: we do not have enough evidence to reject the null hypothesis in support of alternative hypothesis "Data is normally distributed"
So A is not normally distributed while B is normally distributed

when we have at least one of our data not normal so we assume that the data is not normally distributed so here we will use man Whitny test (Welcox test) instead of standard two sample t test assuming non normality of data (non parametric)

Welcox Test

p-value A = 0.0312
Conclusion: we have enough evidence to reject the null hypothesis in support of alternative hypothesis "ADWG0021 is "different" in the group receiving Treatment A (normal feed + ZnO) compared to the Treatment B (normal feed + nutraceuticals)."

d) **Assess the previous test assumption.**

Test Hetero-variance of data

p-value A = 0.955

Conclusion: we don't have enough evidence to reject the null hypothesis in support of alternative hypothesis "Data is not heteroscedastic (they have the same variance)."

e) We hypothesis that ADWG0021is different between the different Treatments. Can you perform comparison between the different groups, after assessing the assumptions and performing post-hoc testing (assuming normality and homoscedasticity)?

Checking Homoscedasticity

Levene Test:

p-value = 0.8968

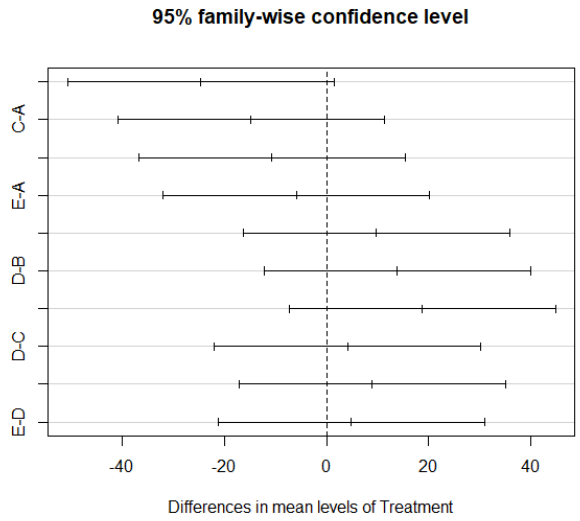
Conclusion: we do not have enough evidence to reject the null hypothesis in support of alternative hypothesis "Data is Homo-variance"

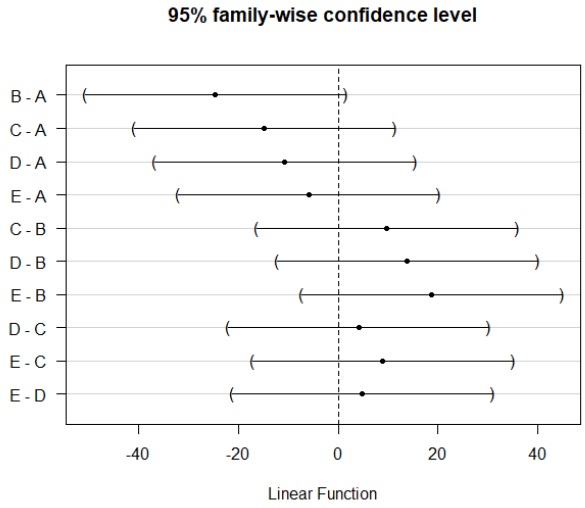
Assuming normality and homoscedasticity as in the question to see whether the assumptions met the test or not so we going to use standard ANOVA test, post-hoc is performed after ANOVA using tukey honest test assuming normality and homoscedasticity (includes p value correction to be p adjusted).

```
> summary(AnovaModel)
            Df Sum Sq Mean Sq F value Pr(>F)
Treatment    4   2771   692.7   2.105  0.101
Residuals   35  11519   329.1
> coef(AnovaModel)
(Intercept)  TreatmentB  TreatmentC  TreatmentD  TreatmentE
 154.303075  -24.590774  -14.794147  -10.701885   -5.865575
> report(AnovaModel)
The ANOVA (formula: ADWG0021 ~ Treatment) suggests that:

  - The main effect of Treatment is statistically not significant and large (F(4, 35) = 2.10, p = 0.101;
Eta2 = 0.19, 95% CI [0.00, 1.00])

Effect sizes were labelled following Field's (2013) recommendations.
```

Post-hoc Test

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

Fit: aov(formula = ADWG0021 ~ Treatment, data = PWD, var.equal = TRUE, alternative = "two.sided")

$Treatment
          diff        lwr       upr     p adj
B-A -24.590774 -50.670281  1.488734 0.0724945
C-A -14.794147 -40.873654 11.285361 0.4884256
D-A -10.701885 -36.781392 15.377622 0.7626339
E-A  -5.865575 -31.945083 20.213932 0.9661386
C-B   9.796627 -16.282880 35.876134 0.8154751
D-B  13.888889 -12.190619 39.968396 0.5499153
E-B  18.725198  -7.354309 44.804706 0.2580810
D-C   4.092262 -21.987246 30.171769 0.9910644
E-C   8.928571 -17.150936 35.008079 0.8605295
E-D   4.836310 -21.243198 30.915817 0.9832666
```





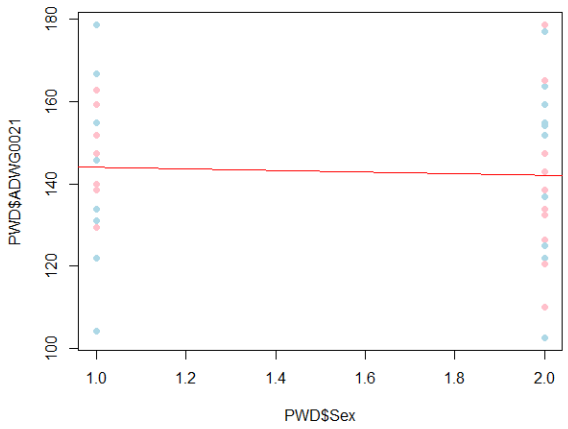

```
        Pairwise comparisons using t tests with pooled SD

data:  PWD$ADWG0021 and PWD$Treatment

  A    B    C    D
B 0.10 -    -    -
C 1.00 1.00 -    -
D 1.00 1.00 1.00 -
E 1.00 0.46 1.00 1.00

P value adjustment method: bonferroni
```

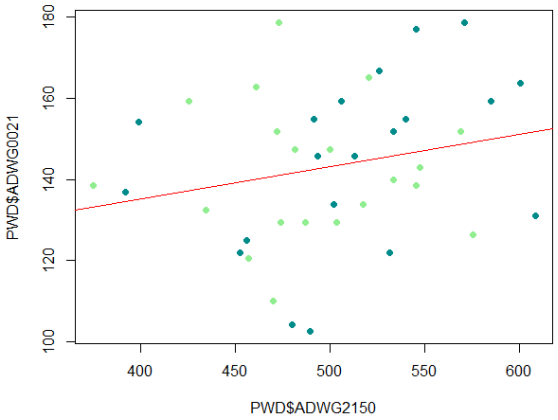
the results of tukey test with adjusted p values with benferoni gives a p adjusted value 0.0437 between treatment A and B (A mixed with B) which is lower than 0.05 so we have evidence to reject null so this leads to non normality

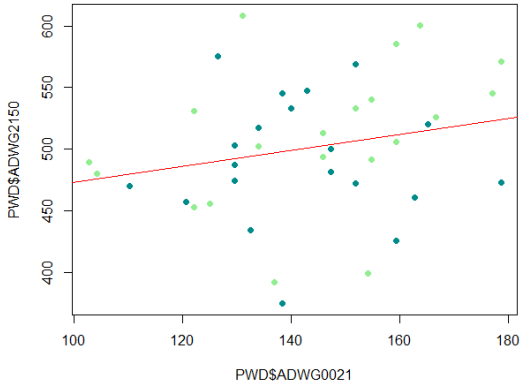
of data(also then ADWG0021 is different in A with B after adjusting p value with benferoni) while the p value of other treatments with each other is greater than 0.05 so we do not have evidence to reject the null so this leads to normality of other treatments (also ADWG0021 is not different between other treatment with each others(except A with B)) also the mean of each pair of treatment is shown in the results.

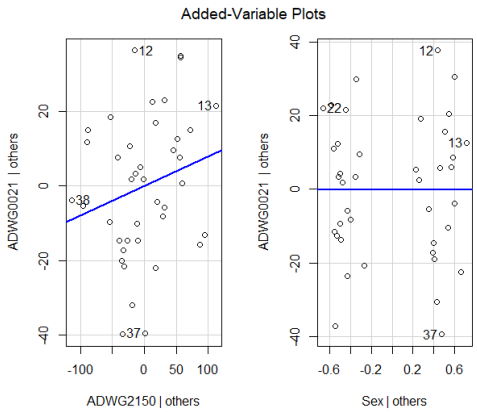
Linear Model

```
Call:
lm(formula = ADWG0021 ~ Sex, data = PWD)

Residuals:
    Min      1Q  Median      3Q     Max
-39.906 -14.608   1.233  13.245  36.419

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  145.992      9.684  15.076   <2e-16 ***
Sex           -1.920      6.125  -0.313    0.756
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.37 on 38 degrees of freedom
Multiple R-squared:  0.002579,  Adjusted R-squared:  -0.02367
F-statistic: 0.09824 on 1 and 38 DF,  p-value: 0.7557
```









The results of multiple linear regression between the 3 variables ADWG0021(y) and ADWG2150+Sex gives also a bad model as the residuals median is too large(2.854) and the p value of ADWG2150 is equal to 0.17519 which is much greater than 0.05 so we do not have enough evidence to reject the null hypothesis so y(ADWG0021) can't be predicted by x(ADWG2150) in the existence of Sex as there is no linearity between them also p value of sex equal0.98652 which is also greater than 0.05 so do not reject the null so sex is not significant and has no effect on the model, also the regression line when fitted shows that it is horizontally so that means that the slope is close to zero (null hypothesis)

Also the also here the adjusted R squared is more accurate than the R squared and the adjusted gives a very small value which is 0.0002851(0.02%) which means that the regression model explains(capture) only 0.02% of the total variation of y(ADGW2150) which is not good at all, so there is no linearity