

**Nile University**

**Core Requirement**

**Post-weaning diarrhea I**

**Project C-Report**

**Team Members:**

Shrouk Hesham 19106271

Reham Galal 19105790

Omar Khaled 19105294

Mohaned Atef 19106161

Beshoy Reda 19104978

**Under Supervision of:** Dr. Mohamed Maysara, Eng. Mariam Oweda

Table of Contents

1. [Introduction 3](#_Toc105235181)
2. [Data Reading 4](#_Toc105235182)
3. [Descriptive statistics 5](#_Toc105235183)
4. [Graphics 6](#_Toc105235184)
5. [Outlier detection 7](#_Toc105235185)
6. [Testing for normality/ homoscedasticity 9](#_Toc105235186)
7. [Statistical Inference 10](#_Toc105235187)
8. [Hypothesis testing 11](#_Toc105235188)
9. [Linear model 13](#_Toc105235189)
10. [Conclusion 18](#_Toc105235190)

# Introduction

Post-weaning diarrhea (PWD) is a worldwide economically important disease in pigs in piggeries. The disease is characterized by increased mortality, weight loss, retarded growth, increased costs, and higher use of antibiotics. Enterotoxigenic Escherichia coli is the most important cause of the disease.

Currently the disease is often controlled by using antimicrobials, but the emergence of antimicrobial resistance in E. coli urges the need for alternative control strategies. For example, inclusion of additional dietary fiber and reduction of crude protein levels, but also the addition of zinc oxide (ZnO) has been demonstrated to have beneficial effects. However, by 2022 this zinc may no longer be used (EU legislation).

Another strategy is to vaccinate the piglets. In this study we are interested in the effect of vaccination as compared to the addition of ZnO and nutraceuticals (e.g. fibers) to the feed. In particular the following treatments are considered (in the dataset this variables is names Treatment):

**A: normal feed + ZnO**

**B: normal feed + nutraceuticals**

**C: vaccination + high energy/protein in phases 2 and 3 (time periods)**

**D: vaccination + high energy/protein in phases 1, 2 and 3**

**E: vaccination + high energy/protein in phases 1, 2 and 3 + nutraceutics**

We are interested in the following outcomes:

**ADWG0021, ADWG2150, ADWG0050:** average daily weight gain (g/day) in the period between 0 and 21 days post-weaning, between day 21 and day 50 post-weaning and in the period between 0 and 50 days post-weaning, respectively. Thus, as biostatisticians, we can perform some analysis using R language to determine whether or not treatments are effective.

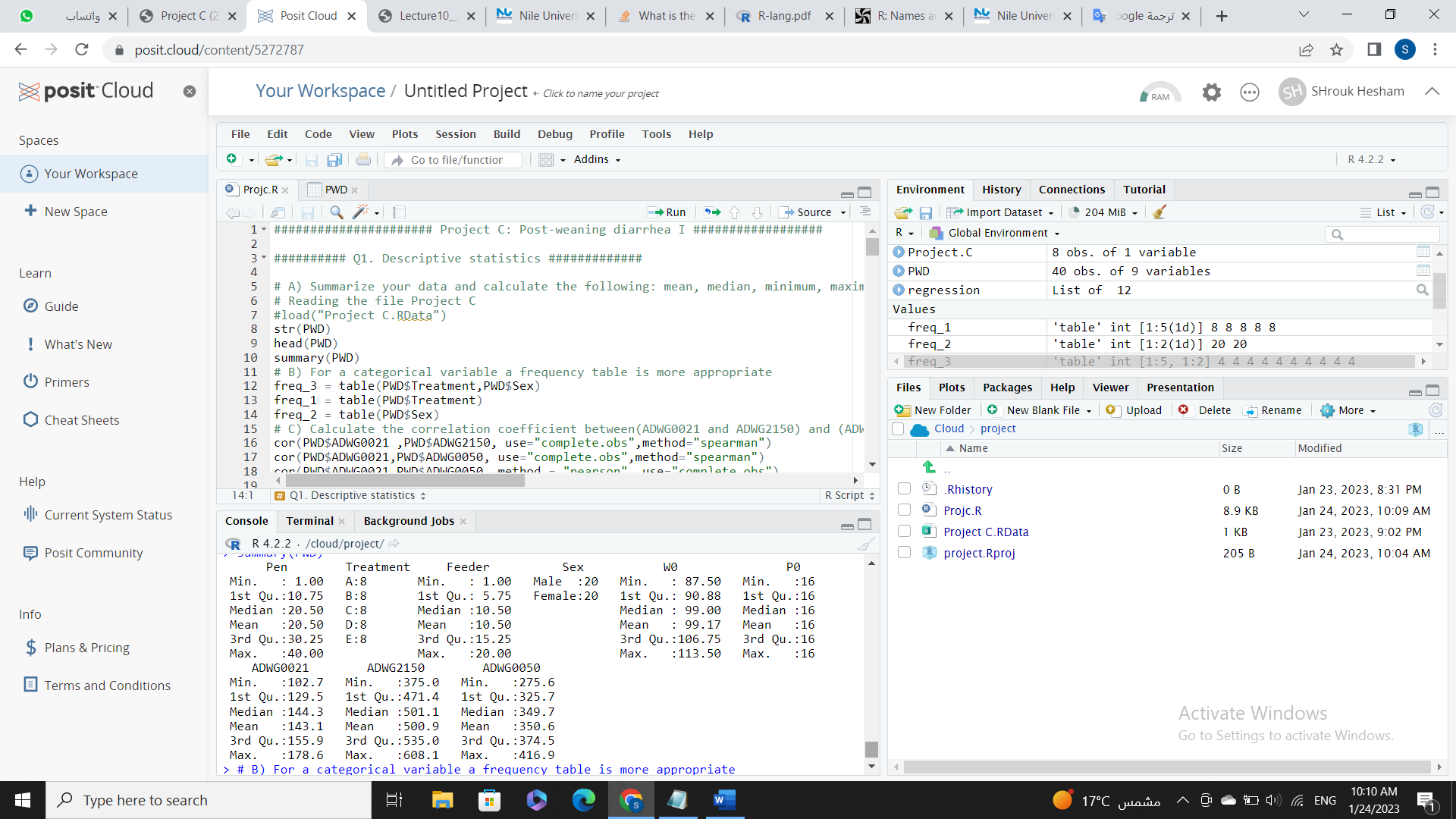
# Data Reading

We read the data by importing it into R, then proceeding it. Nine columns were identified, and they are as following:

* **Pen** (number of pens which contain 16 piglets in one pen)
* **Treatment** (contain 5 treatments (A, B, C,D,E), each treatment group consists of 128 piglets (thus 8 pens of 16 piglets))
* **Feeder** (food)
* **Sex** (male, female)
* **W0** (weight)
* **P0** (number of piglets in each pen)
* **ADWG0021** (average daily weight gain (g/day) in the period between 0 and 21 days post-weaning)
* **ADWG2150** (between day 21 and day 50 post-weaning)
* **ADWG0050** (period between 0 and 50 days post-weaning)

# Descriptive statistics

1. After reading our data file “Project C. RData” , then we summarize our data.
2. We Calculated the following: mean, median, minimum, maximum, first and third quartile (for each variable).



1. For the categorical variable existing, calculate a frequency table, We found that There are categorical variable (sex and treatment).
2. Calculate the correlation coefficient (ADWG0021 and ADWG2150) and (ADWG0021 and ADWG0050), to know how strongly two different variables, x and y, are related linearly. So, we employed two techniques: Pearson and Spearman correlation.
3. **Graphics**

* Generate a bar chart of a categorical variable for the gender (Sex parameter)

Chart, box and whisker chart

Description automatically generated

Fig1 shows: our data contains 20 male and 20 female.

* Generate a bar chart graph with mean ADWG0021 in males and females.

Chart, bar chart

Description automatically generated

Fig2 shows: bar chart graph with mean ADWG0021 in males and females

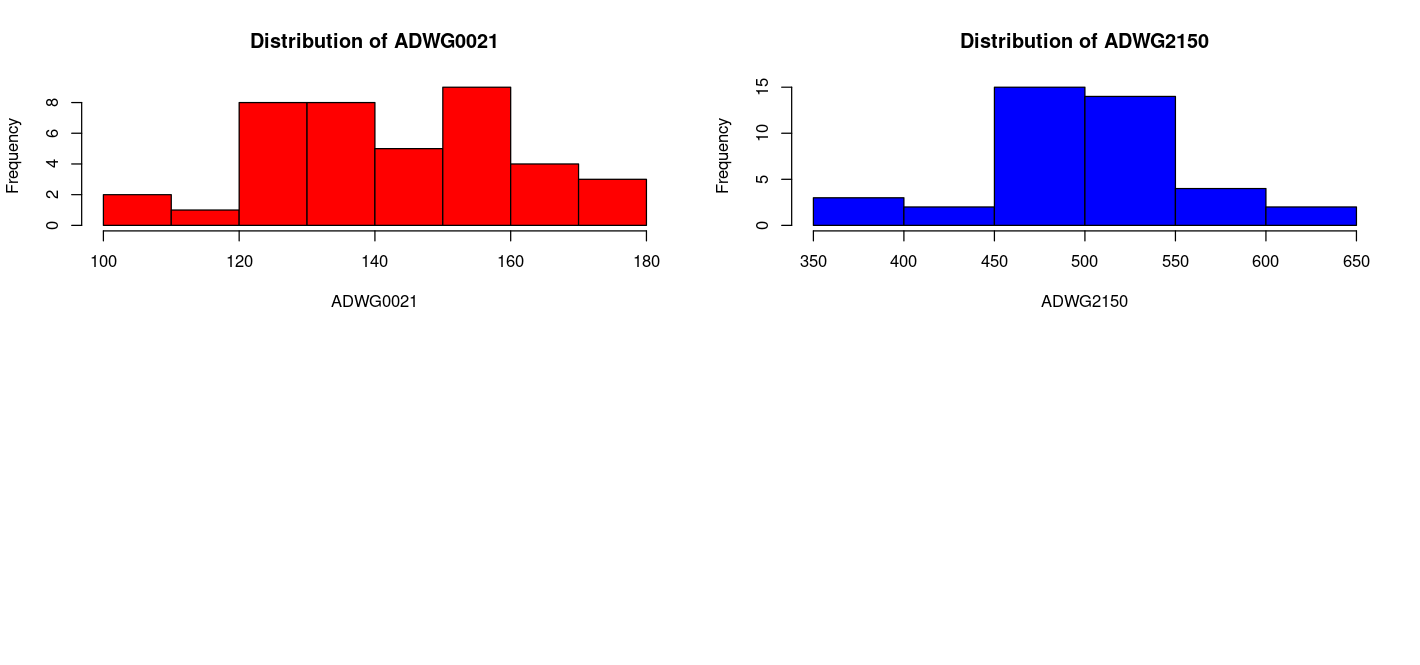
* Make a histogram of a continuous variable: “ADWG2150” as well as “ADWG0021.

Fig3 shows histogram visualization of continuous variable: “ADWG2150” as well as “ADWG0021

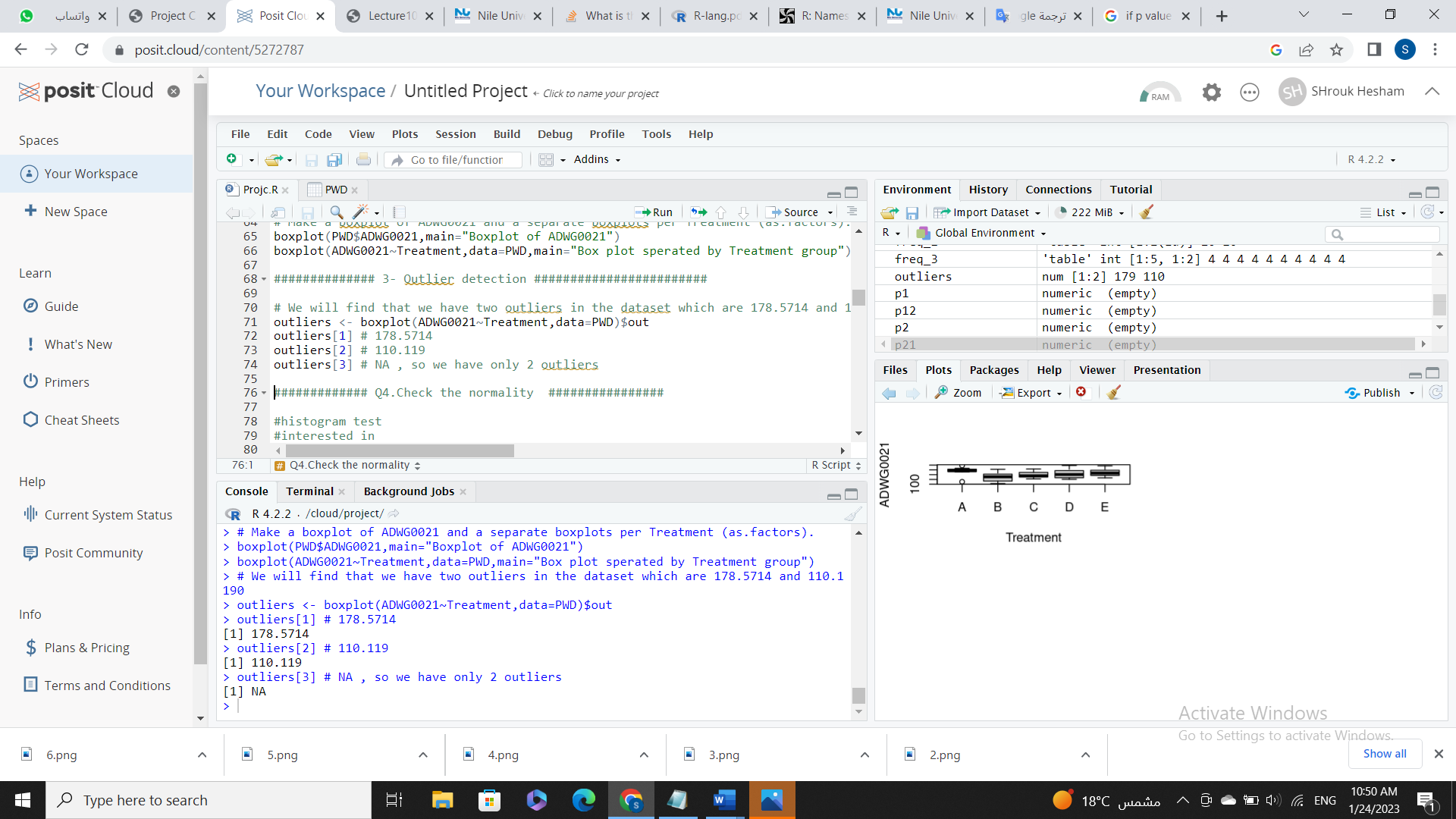
* Make a boxplot of ADWG0021 and a separate boxplots per Treatment (as.factors).

**Diagram

Description automatically generated**

Fig4 shows boxplot visualization ADWG0021 and a separate boxplot per Treatment

1. **Outliers Detection**
2. They're unusual values in a dataset.
3. outlier 1 is 178.5714, outlier 2 is 110.119 and outlier 3 is NA.
4. So, we have only 2 outliers
5. Since outliers are problematic for many statistical analyses because they can cause tests to either miss significant findings or distort real results, then having low outliers is a good indicator about our data



1. **Normality/ Homoscedasticity**
2. **Normality**
3. Data must fall under two assumptions, which are Normal distribution and random sampling
4. We used 2 methods to check normality, Shapiro and QQ-plot
5. According to Shapiro’s test for checking the normality the values gives for ADWG0021, ADWG2150, ADWG0050 are 0.7305, 0.9086, 0.9037 respectively
6. Comparing to the alpha value (0.05), this value assume non-normality of the data
7. The second method is the QQ-plot which supported the Shapiro’s test results
8. **Homoscedasticity**
9. Heteroscedasticity is present when the variance of the residuals is observed to be uneven. However, homoscedasticity is the condition in which the residuals have equal variance.
10. We tested the variance using 3 methods: Boxplot, Bartlett’s test and Levene’s test
11. Boxplot showed huge variation in the data.

Chart

Description automatically generated with low confidence

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Diagram

Description automatically generated

Chart

Description automatically generated

1. **Statistical inference**
2. A confidence interval is the mean of estimate plus and minus the variation in that estimate. This is the range of values to expect estimate to fall between
3. We Calculated the 90%, 95%, 99% confidence interval for the means of ADWG0021per each gender
4. At 90 percent confidence interval for males is: 138.0131-148.2121
5. At 90 percent confidence interval for females is: 138.0131-148.2121
6. At 95 percent confidence interval for males is: 136.9906-149.2346
7. At 95 percent confidence interval for females is: 136.9906-149.2346
8. At 99 percent confidence interval for males is: 134.9167-151.3085
9. At 99 percent confidence interval for females is: 134.9167-151.3085
10. **Hypothesis testing**

**Doing statistical hypothesis testing:** 1) First, converting the research question into a statistical question, which is: Is the mean between male and female in ADWG0021 different?

Then, we state the null and alternative hypotheses, as following:

* The null hypothesis (H0) would be that there is no differences between the mean of groups Male and Female in terms of the mean content.
* The alternative hypothesis (H1) would be that the mean of Male groups will be different from the mean of Female groups.

**Next, we Test Statistic and check the results by comparing the p-value result to our significance level alpha (0.05)**

If the p-value is greater than alpha (0.05), then we don't have enough evidence to reject the null hypothesis in support of alternative hypothesis.

If P-value is smaller or equal than alpha (0.05), then we have enough evidence to reject null hypothesis in support of alternative hypothesis.

When we check Normality with using Shapiro Test

We will find that our p-value = 0.0004679 which is smaller than the significance level alpha(0.05). Then, we have enough evidence to reject the null hypothesis in support of alternative hypothesis. So, we cannot assume normality (not normally distributed)

We will also use a t‐test because Paired tests are used when two measurements are taken from the same individual (say, person).

So here, we are comparing the ADWG0021 between Males and Females. Furthermore, the paired t-test allows us to compare the mean difference between the two measurements before and after to determine whether the difference is statistically significant or not.

**So, we will find that after using t-test there is a difference in means between Male and Female in ADWG0021**

**P-value = 0.7557** which is greater than the significance level alpha (0.05) , So we don't have enough evidence to reject null hypothesis in support of alternative hypothesis. So, we assume normality and it's not significant.

**Checking homogeneity test using levene's test and F-test**

We will check our p-value result if smaller than 0.05 so reject null hypothesis (H0) ,so in this case assume hetero

If p-value greater than alpha (0.05) so do not reject null, so in this case assume homo

Here, we used levene's test as it Robust against many types of non-normality and most powerful. We will find that, our p-value = 0.3009 which is greater than the significance level alpha (0.05). So, we do not have enough evidence to reject the null hypothesis in support of alternative hypothesis, so our data have homo variance or equal variance. In addition, we found that F-Test results a p value of 0.4193 which is also greater than alpha (0.05) so we do not have enough evidence to reject null hypothesis in support of alternative hypothesis, so our data have homo variance and the previous assumptions have been met.

Assuming null: ADWG0021 is equal in the group receiving Treatment A (normal feed + ZnO) and the Treatment B (normal feed + nutraceuticals).

Alternative: ADWG0021 is different in the group receiving Treatment A (normal feed + ZnO) compared to the Treatment B (normal feed + nutraceuticals).

By applying welch t-test we will find that the p-value = 0.02783 which is smaller than alpha = .05

We have enough evidence to reject null which means that ADWG0021 is different in the group receiving Treatment A (normal feed + ZnO) compared to the Treatment B (normal feed + nutraceuticals)

**Checking Normality:**

Null: data is normally distributed

Alternative Hypothesis:

by checking normality in group A & group B after dividing Treatment into A&B by using Shapiro.test in group A & group B we will find that in group A , p-value = 0.0395 < .05, so we have enough evidence to reject null which means that Group A is not normally distributed.

in group B, p-value = p-value = 0.8132 > .05, so we don’t have enough evidence to reject null which means that group B is normally distributed.

Checking variance:

Null: data is equal in variance

Alternative Hypothesis: data is not equal in variance

By using var.test we will find that p-value = 0.955 > .05. We don’t have enough evidence to reject null. That data is homoscedastic.

Null hypothesis: We hypothesis that ADWG0021is different between the different Treatments

Alternative Hypothesis: We hypothesis that ADWG0021is not different between the different Treatments.

By using Anova:

p = 0.101> .05, we don’t have enough evidence to reject null which means that there is difference between groups.

Then we use TukeyHSD(anova) to compare all groups.

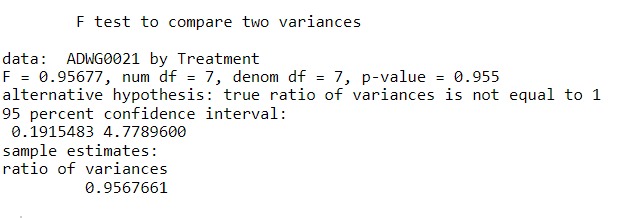
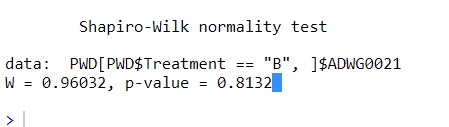
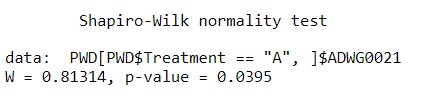
Checking normality:

we will find that all groups are normally disributed except A

Checking variance:

We will find that P-value>.05

we don't have enough evidence to reject null, so treatment is equal variance>> data is homoscedasticity.

Graphical user interface, text

Description automatically generated

Letter

Description automatically generated with low confidenceGraphical user interface, text, application

Description automatically generatedGraphical user interface, application

Description automatically generatedText

Description automatically generatedText

Description automatically generated

1. **Linear Model**
2. Linear models were created for two hypotheses, using the Sex and Treatment regressors respectively.
3. Treatment column was encoded using one-hot encoding because:
   1. It allows the model to receive it as an input
   2. It allows each treatment category to have its own weight
4. ModelS had intercept of 145.992, slope of -1.920, RSE = 19.37, R2 = 0.002579,
5. F-stat 0.09824, p-value = 0.7557
6. ModelS clearly ineffective due to lack if proper linear relationship
7. 95% interval -14.31805 - 10.47877 for sex regressor means that on average Intercept + (-14.31805 - 10.47877) contains true values
8. average change per unit for Rsex is -1.920, do further demonstrate:
   1. Dummy data was generated for both male and female.
   2. As expected model can only output 2 constant values based on sex
   3. constant prediction when male: 144.0724, prediction when female: 142.1528
   4. 144.072 - 142.1528 = 1.9196

|  |  |  |
| --- | --- | --- |
| Variables | ModelS | ModelT |
| RSE | 19.37 | 18.14 |
| Intercept | 145.992 | 154.303 |
| Slope | -1.920 | -24.591  -14.794  -10.702  -5.866 |
| R2 | 0.002579 | 0.1939 |

1. **Conclusion**

Finally, we discovered that treatments A and B significantly differed from the other treatments, indicating that they had an impact on the piglets. Additionally, we found that post-weaning diarrhea (PWD) affects pigs in piggeries all over the world and it's significant.  The most significant contributing factor to the illness is enterotoxigenic Escherichia coli.