Assignment4

Question1

#a)

animals <- read.csv("animals.csv")

head(animals)

tail(animals)

Text

Description automatically generated with low confidence

#b)

boxplot(Weight~Feed, data = animals, pch = 19)

Chart, box and whisker chart

Description automatically generated

From the boxplot above we can see Feed stock C has the highest median value of weight while Feed stock A has the lowest median value of weight. Different feed stock has different animal weights. We can also observe that the variance of each group is almost the same while Feedstock B doesn’t have upper limit, but it is because of the small amount of data points so overall we can see the variance of each feed stock is the almost the same.

#c)

Text, letter

Description automatically generated

From the summary we can observe the p-value of FeedB, FeedC and FeedD are all much smaller than 0.05, which means we could reject null hypothesis. There is a difference between FeedA(baseline) and FeedB, FeedA and FeedC, FeedB and FeedD.

From the anova we can observe that the F value is large while the p-value of F-test is small. We could say feed has an impact to the weights.

d)  
Calendar

Description automatically generated

From Residuals vs Fitted above, we can observe that the residuals is constantly scattering around 0 and there is no pattern. Also, the red line nearly fits the dash line. Overall we can say the model meets the linearity.

Chart, scatter chart

Description automatically generated

From the Normal Q-Q graph above, we can see there is not too much deviation between residuals since most of the residuals are nearly fitting the dash line. Overall we could say the model meets the normality.

Chart, histogram

Description automatically generated

From the Cook’s distance above, we can observe that there are 3 data points (3, 8 and 20 respectively) away from our datasets. But their Cook’s distance is below 0.4, we wouldn’t consider them as high influence points.

e)  
Text

Description automatically generated

Text

Description automatically generated

from the TukeyHSD we can see none of them contains 0 in their lower and upper limit, so we can say they are all different. From the boxplot we can see feedC can make the biggest amount of weight. From TukeyHSD, we have the same conclusion. As we can see B is greater than A by 8.56, C is greater than A by 39.66, D is greater than A by 25.7, so B, C, D are greater than A. If we continue observing, C is greater than B by 31.10, D is greater than B by 17.14. And D is smaller than C by 13.96. We can see FeedC will cause a highest weight. So feed stock C is recommended.

Question2

a)  
Table

Description automatically generated

b)  
Chart, line chart

Description automatically generated

from the interaction plot above, we can observe that the recovery days in Facility South is greater than the recovery days in Facility North. And in both facilities, we can see treatmentA has the lowest recovery days and treatmentC has the highest recovery days.

Also, we can observe that except the intercepts are different between different facilities, the slopes of increase in treatmentB and treatment in different facilities are the same.

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

From the boxplot, we can observe treatment A in North facility has the lowest recovery days while the treatment C in South Facility has the highest recovery day.

Overall, treatmentC has the highest recovery day no matter in which facility.

With first graph, we can see boxplots are in facilities order, North facility first and then South facility. While in the second graph, we can see the box plots are in treatment order, while treatmentA first, and then treatment, and then treatmentC.

The difference between North and South facilities depends on the different treatments.

c)  
Text

Description automatically generated

From the summary, we can observe that the interaction between treatmentB and facility South and the interaction between treatmentC and facility South are not significant as their p-value is greater than 0.05, while facilitySouth is significant with baseline treatment A. Also, we observe that treatmentC is significant. These observations mean treatment and facilities are significant to the model. But the interaction between facilities and treatments is not significant. So we can reduce the model by reducing the interaction into the additive model. Although from the summary, we observe treatmentB is not significant as well. But after the reduction, it might be significant.

From the anova, it confirms the same thing. The p-value of the interaction between treatment and facility is 0.9054 which is greater than 0.05, which means it is not significant to our model. So we can reduce our model to a simpler model by remove the interaction between facilities and treatments.

d)  
health.lm2<-lm(Days~Treatment+Facility, data = health)

e)

Text

Description automatically generated

From the summary of the new model, we can observe the treatmentC and FacilitySouth are significant to our model, which means there is a relationship between treatmentA and treatmentC. treatment However, treatmentB is not significant.

From the anova of the model, we can observe that treatment as a whole variable is significant to the model and same as facility.

If we have a look at the previous graph, we can notice that the intercept is different for different facilities, but the slope is the same as treatments change no matter in what facility.

The difference between the treatments mean is the same for different facility.

f)

Chart, scatter chart

Description automatically generated

From the Residuals vs Fitted graph, we can observe that the residuals are scattered around the zero and there is no patter appears. And the red line is almost fitted to the zero dash line. These shows the model meets the linearity.

Chart, scatter chart

Description automatically generated

From the Normal Q-Q graph we can see all the observations almost fit the zero dash line and there is not too much deviation between the residual and the zero dash line. So we can say the model meets the normality.

Chart, histogram

Description automatically generated

From the Cook’s distance graph, we can observe that there are 3 points(26, 28 and 29 respectively) which might be more significant to the model. But their Cook’s distances are all below 0.4, so they are not high influence points.

g)

From the anova result above, we can observe that both facility and treatment are significant to our model, which means there is at least one treatment is significant, (which is treatmentC). Although we can see from summary, treatmentB is not that significant, as a whole treatment group is significant to our model. So we can’t remove the model further more.

h)

Text

Description automatically generated

From the TukeyHSD we can see there is no difference between treatment B and treatment A, which means they are similar. While there is difference between treatment C and treatment A, treatment C and treatment B. For facility, we can observe there is difference between South and North. Overall, I would recommend treatment A as it has the lowest revocery days.

Code:

#question1

#a)

animals <- read.csv("animals.csv")

head(animals)

tail(animals)

#b)

animals$Feed

boxplot(Weight~Feed, data = animals, pch = 19)

#c)

animals.lm1<-lm(Weight~Feed, data = animals)

summary(animals.lm1)

anova(animals.lm1)

#d)

plot(animals.lm1, which = 1)

plot(animals.lm1, which = 2)

plot(animals.lm1, which = 4)

#e)

animals.lm2<-aov(Weight~Feed, data = animals)

anova(animals.lm1)

anova(animals.lm2)

TukeyHSD(animals.lm2)

# I am confused about the models,are lm1 and lm2 the same, then why do we need to compare

# question2

#a)

health <- read.csv("health.csv")

head(health)

tail(health)

#b)

interaction.plot(health$Treatment, health$Facility, health$Days)

boxplot(Days~Treatment\*Facility, col=c("white", "steelblue"), data

=health)

boxplot(Days~Facility\*Treatment, col=c("white", "steelblue"), data

=health)

#c)

health.lm1<-lm(Days~Treatment\*Facility, data = health)

summary(health.lm1)

anova(health.lm1)

#d)

health.lm2<-lm(Days~Treatment+Facility, data = health)

#e)

summary(health.lm2)

anova(health.lm2)

#f)

plot(health.lm2, which = 1)

plot(health.lm2, which = 2)

plot(health.lm2, which = 4)

#g)

# Can you reduce your model ever further by removing either of the main effects (i.e. can you

# remove Facility or Treatment)?

#h)

health.lm3<-aov(Days~Treatment+Facility, data = health)

anova(health.lm3)

TukeyHSD(health.lm3)