HW1

Ryo Iwata

16 February, 2024

Nota bene - Use car::Anova unless otherwise specified throughout this course, the reasons for which will be covered in detail in a later lecture. You can do so by using library(car) then Anova or car::Anova, assuming the car package is installed. With our simple models, it isn't different from the normal anova function, but it is a good habit to get into.

Each problem should have its own data loading and carpentry chunk

Problem #1 (40 pts):

Your laboratory is studying the effect of the diffusion of intraventricular drug delivery vehicles (particles) and are currently evaluating the impact of the polymer used. load and prepare data from sheet DataSet1.

```
polymer_data <- readxl::read_excel("./data/HW1_Data.xlsx", sheet="DataSet1")
```

```
## New names:
## * `Polymer Type` -> `Polymer Type...1`
## * `Diffusion Rate` -> `Diffusion Rate...2`
## * `Polymer Type` -> `Polymer Type...3`
## * `Diffusion Rate` -> `Diffusion Rate...4`
```

```
implant_data <- readxl::read_excel("./data/HW1_Data.xlsx", sheet="DataSet2")
bnp_data <- readxl::read_excel("./data/HW1_Data.xlsx", sheet="DataSet3")
```

```
str(polymer_data)
```

```
## tibble [40 x 4] (S3: tbl_df/tbl/data.frame)
##  $ Polymer Type...1  : chr [1:40] "PLGA" "PLGA" "PLGA" "PLGA" ...
##  $ Diffusion Rate...2: num [1:40] 7.32 7.72 8.33 7.31 6.77 ...
##  $ Polymer Type...3  : chr [1:40] "Dextran" "Dextran" "Dextran" "Dextran" ...
##  $ Diffusion Rate...4: num [1:40] 7.92 7.4 8.27 8.13 8.62 ...
```

```
polymer_1 <- polymer_data |>
  select("Polymer Type...1", "Diffusion Rate...2") |>
  rename(polymer_type = "Polymer Type...1") |>
  rename(diffusion_rate = "Diffusion Rate...2")

polymer_2 <- polymer_data |>
  select("Polymer Type...3", "Diffusion Rate...4") |>
  rename(polymer_type = "Polymer Type...3") |>
  rename(diffusion_rate = "Diffusion Rate...4")

polymer_restructured <- bind_rows(polymer_1, polymer_2)
```

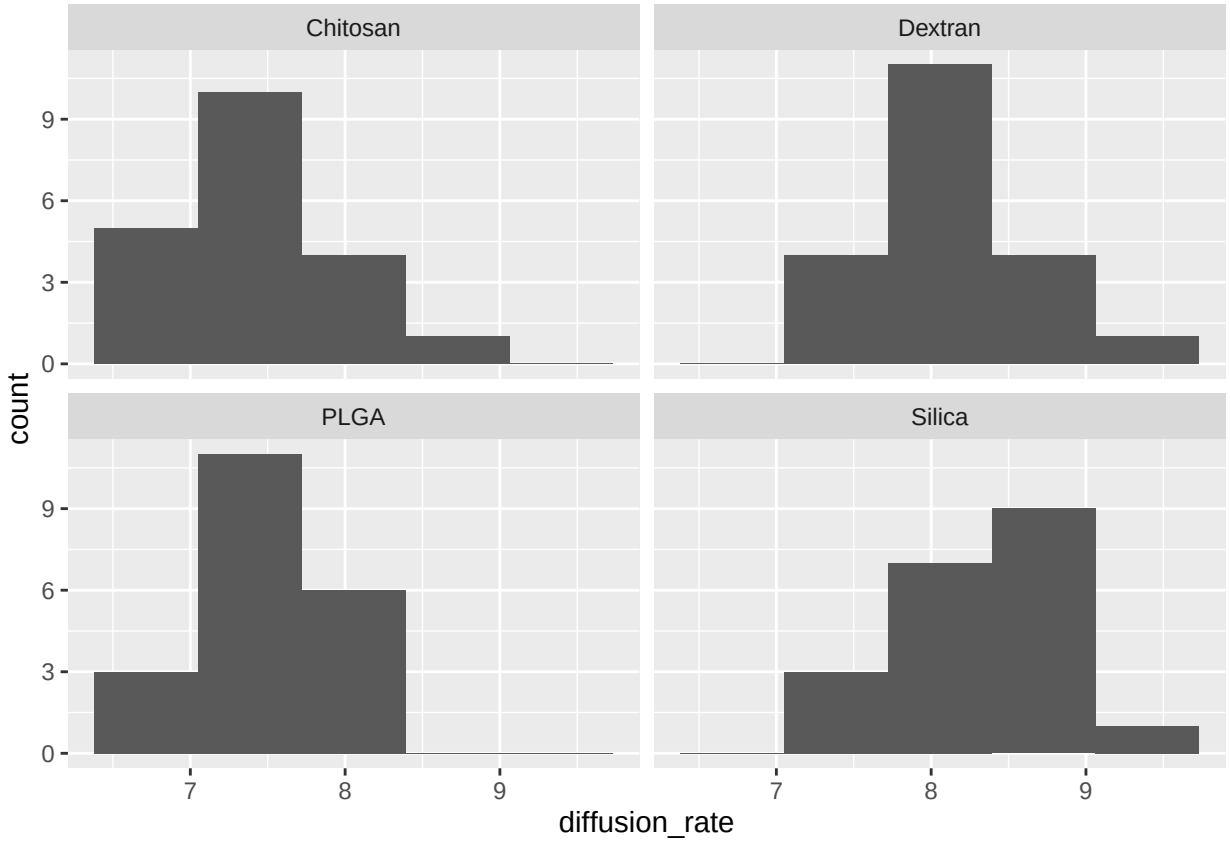
```
implant_restructured <- pivot_longer(implant_data, colnames(implant_data))
```

```
bnp_restructured <- bnp_data |>
  rename(bnp_decrease_with_furosemide="Decrease in BNP After Treatment With Furosemide") |>
  rename(bnp_decrease_with_hydrochlorothiazide="Decrease in BNP After Treatment With Hydrochlorothiazide
  rename(bnp_decrease_with_ethacrynic_acid="Decrease in BNP After Treatment With Ethacrynic Acid") |>
  rename(bnp_decrease_with_metolazone="Decrease in BNP After Treatment With Metolazone")


bnp_restructured <- pivot_longer(bnp_restructured, colnames(bnp_restructured))
```

1) Create a set of plots that allow you to visualize the distribution of the diffusion of each particle type that emphasizes/aids in visual consideration of particular characteristics:

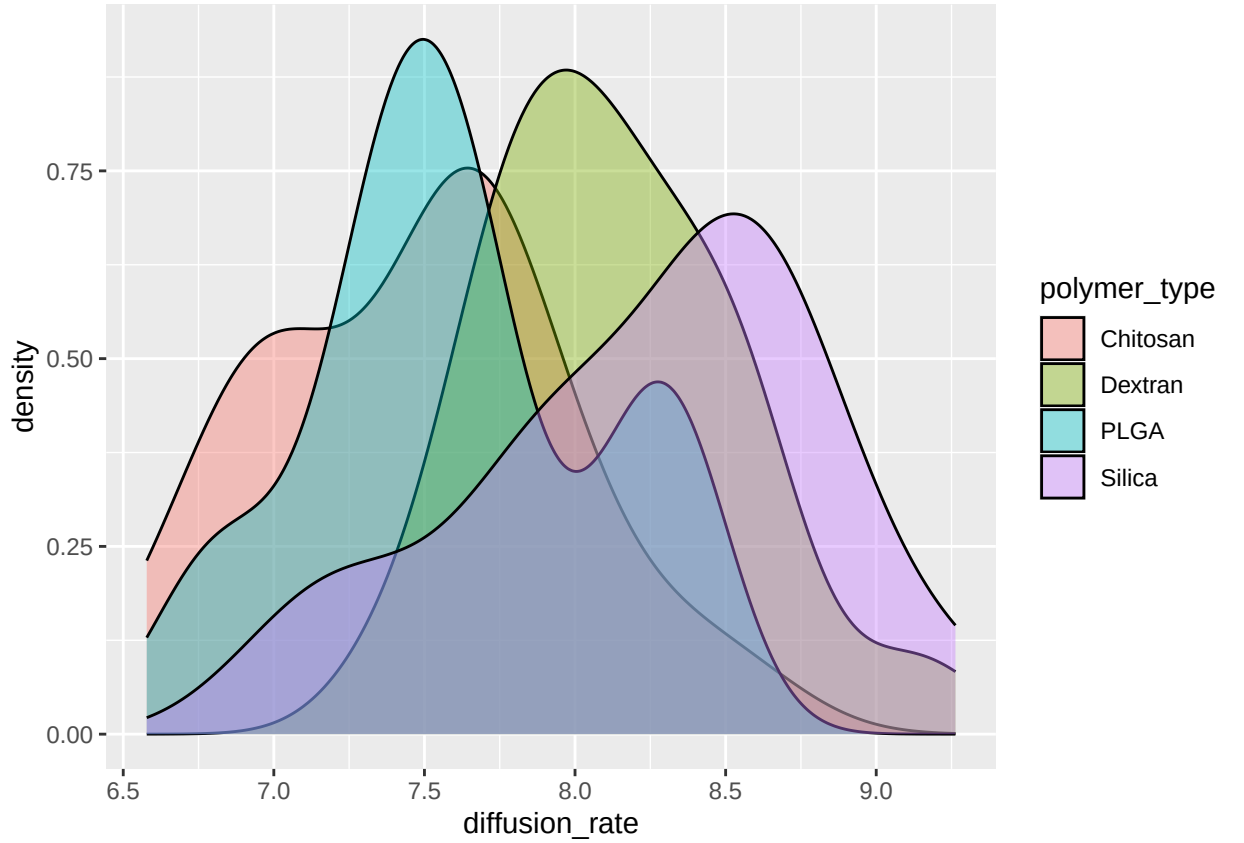
a) Focused on the distribution within polymer type (4 pts)

```
polymer_restructured |> ggplot(aes(x=diffusion_rate)) +
  geom_histogram(bins = 5) +
  facet_wrap(vars(polymer_type))
```



b) Focused on relationship beteen the distributions (4 pts)

```
polymer_restructured |> ggplot(aes(x=diffusion_rate, group=polymer_type, fill=polymer_type)) +
    geom_density(alpha=.4)
```



2) Fit a regression model (lm) to capture the effect of polymer type on diffusion. (2pts)

```
polymer_lm <- lm(diffusion_rate ~ polymer_type, polymer_restructured)
```

a) What are the coefficients for this model (4 pts)?

```
tidy(summary(polymer_lm))
```

```
## # A tibble: 4 x 5
##   term                estimate std.error statistic  p.value
##   <chr>                  <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)            7.47      0.111    67.4   1.59e-69
## 2 polymer_typeDextran    0.657     0.157     4.19  7.48e- 5
## 3 polymer_typePLGA       0.134     0.157     0.854 3.96e- 1
## 4 polymer_typeSilica     0.763     0.157     4.87  6.00e- 6
```

Answer: The coefficients for this model is the mean diffusion rate of every polymer type subtracted by the intercept. One group's mean is the intercept, which does not have a corresponding coefficient.

b) Which group is serving as the intercept (2 pts)?

Answer: In this model, Chitosan is serving as the intercept which can be seen in the output of `summary(polymer_lm)` which shows that Chitosan is not one of the groups that has a coefficient.

c) Using these coefficients (and not the source data), calculate the model's estimated means for each group (4 pts)?

```
tidy_polymer <- tidy(polymer_lm)

mean_chitosan <- pull(filter(tidy_polymer, term=="(Intercept)"), estimate)
mean_dextran <- mean_chitosan + pull(filter(tidy_polymer, term=="polymer_typeDextran"), estimate)
mean_plga <- mean_chitosan + pull(filter(tidy_polymer, term=="polymer_typePLGA"), estimate)
mean_silica <- mean_chitosan + pull(filter(tidy_polymer, term=="polymer_typeSilica"), estimate)

print(paste("Mean diffusion rate of Chitosan: ", mean_chitosan))
```

```
## [1] "Mean diffusion rate of Chitosan:  7.471"
```

```
print(paste("Mean diffusion rate of Dextran: ", mean_dextran))
```

```
## [1] "Mean diffusion rate of Dextran:  8.12805"
```

```
print(paste("Mean diffusion rate of PLGA: ", mean_plga))
```

```
## [1] "Mean diffusion rate of PLGA:  7.60495"
```

```
print(paste("Mean diffusion rate of Silica: ", mean_silica))
```

```
## [1] "Mean diffusion rate of Silica:  8.23445"
```

```
polymer_restructured |>
  group_by(polymer_type) |>
  summarise(avg = mean(diffusion_rate))
```

```
## # A tibble: 4 x 2
##   polymer_type   avg
##   <chr>        <dbl>
## 1 Chitosan      7.47
## 2 Dextran       8.13
## 3 PLGA          7.60
## 4 Silica        8.23
```

d) Calculate (using a tidyverse appraoch) the SS_{total}, $SS_{Regression}$, and SS_{Error} for this model (5 pts)?

```
# Calculating the mean for all the diffusion rates
polymer_restructured <- polymer_restructured |>
  mutate(mean_all = mean(diffusion_rate))

# Calculate SS total
ss_total <- sum((polymer_restructured$diffusion_rate - polymer_restructured$mean_all)^2)

print(paste0("SS total: ", ss_total))
```

```
## [1] "SS total: 27.2672809875"
```

```
# Calculating the predicted values
polymer_restructured <- polymer_restructured |>
  mutate(y_pred = fitted(polymer_lm))

# Calculate SS Regression
ss_regression <- sum((polymer_restructured$y_pred - polymer_restructured$mean_all)^2)
```

```r
print(paste0("SS regression: ", ss_regression))
```

```
## [1] "SS regression: 8.56869013750001"
```

```r
# Calculate SS_Error
ss_error <- sum((polymer_restructured$y_pred - polymer_restructured$diffusion_rate)^2)

print(paste0("SS error: ", ss_error))
```

```
## [1] "SS error: 18.69859085"
```

e) Write out the formula for this model using the beta notation discussed in class (2 pts)?

Answer:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$$

3) Using the estimates and their associated error for the above model, use a t-test to evaluate whether the term is zero or non-zero.

```r
non_zero_terms <- filter(tidy_polymer, p.value < 0.05)$term

zero_terms <- filter(tidy_polymer, p.value >= 0.05)$term

print(paste0("Non-zero term: ", non_zero_terms))
```

```
## [1] "Non-zero term: (Intercept)"         "Non-zero term: polymer_typeDextran"
## [3] "Non-zero term: polymer_typeSilica"
```

```r
print(paste0("Zero term: ", zero_terms))
```

```
## [1] "Zero term: polymer_typePLGA"
```

a) What are the t-statistic and associated p-value. Note-no calculation necessary, we aready have this information (3 pts)?

```r
tstat_chitosan <- pull(filter(tidy_polymer, term=="(Intercept)"), statistic)
tstat_dextran <- mean_chitosan + pull(filter(tidy_polymer, term=="polymer_typeDextran"), statistic)
tstat_plga <- mean_chitosan + pull(filter(tidy_polymer, term=="polymer_typePLGA"), statistic)
tstat_silica <- mean_chitosan + pull(filter(tidy_polymer, term=="polymer_typeSilica"), statistic)

pvalue_chitosan <- pull(filter(tidy_polymer, term=="(Intercept)"), p.value)
pvalue_dextran <- mean_chitosan + pull(filter(tidy_polymer, term=="polymer_typeDextran"), p.value)
pvalue_plga <- mean_chitosan + pull(filter(tidy_polymer, term=="polymer_typePLGA"), p.value)
pvalue_silica <- mean_chitosan + pull(filter(tidy_polymer, term=="polymer_typeSilica"), p.value)

print(paste("T-statistic of Intercept: ", tstat_chitosan, "P-value of Intercept: ", pvalue_chitosan))
```

```
## [1] "T-statistic of Intercept:  67.3590713327174 P-value of Intercept:  1.59326940097543e-69"
```

```r
print(paste("T-statistic of Dextran: ", tstat_dextran, "P-value of Dextran: ", pvalue_dextran))
```

```
## [1] "T-statistic of Dextran:  11.6599075585019 P-value of Dextran:  7.47107478504534"
```

```r
print(paste("T-statistic of PLGA: ", tstat_plga, "P-value of PLGA: ", pvalue_plga))
```

```
## [1] "T-statistic of PLGA:  8.32497483823351 P-value of PLGA:  7.86680267424843"
```

```
print(paste("T-statistic of Silica: ", tstat_silica, "P-value of Silica: ", pvalue_silica))
```

```
## [1] "T-statistic of Silica:  12.3382421817796 P-value of Silica:  7.47100599712551"
```

b) What do these coefficient/estimate terms represent. List and explain each. (3 pt)?

Answer: (Intercept) is the y-value that the model would cross when all other variables are 0. In this case, it is the mean diffusion rate of the Chitosan group.

polymer_typeDextran is the coefficient for the variable of whether a data point is in the Dextran group or not. The coefficient is equal to the mean diffusion rate of the Dextran group.

polymer_typePLGA is the coefficient for the variable of whether a data point is in the PLGA group or not. The coefficient is equal to the mean diffusion rate of the PLGA group.

polymer_typeSilica is the coefficient for the variable of whether a data point is in the Silica group or not. The coefficient is equal to the mean diffusion rate of the Silica group.

c) Which polymers differ from PLGA (ignoring consideration of multiple comparisons/post-hoc testing) (2 pt)?

```
# For pairwise comparison without considering multiple testing corrections
pairwise_t_test_results <- pairwise.t.test(polymer_restructured$diffusion_rate, polymer_restructured$pol

# pull(filter(pairwise_t_test_results, ))
print(pairwise_t_test_results)
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  polymer_restructured$diffusion_rate and polymer_restructured$polymer_type
##
##         Chitosan Dextran PLGA
## Dextran 0.00037  -       -
## PLGA    0.79161  0.00396 -
## Silica  3.6e-05  0.79161 0.00056
##
## P value adjustment method: holm
```

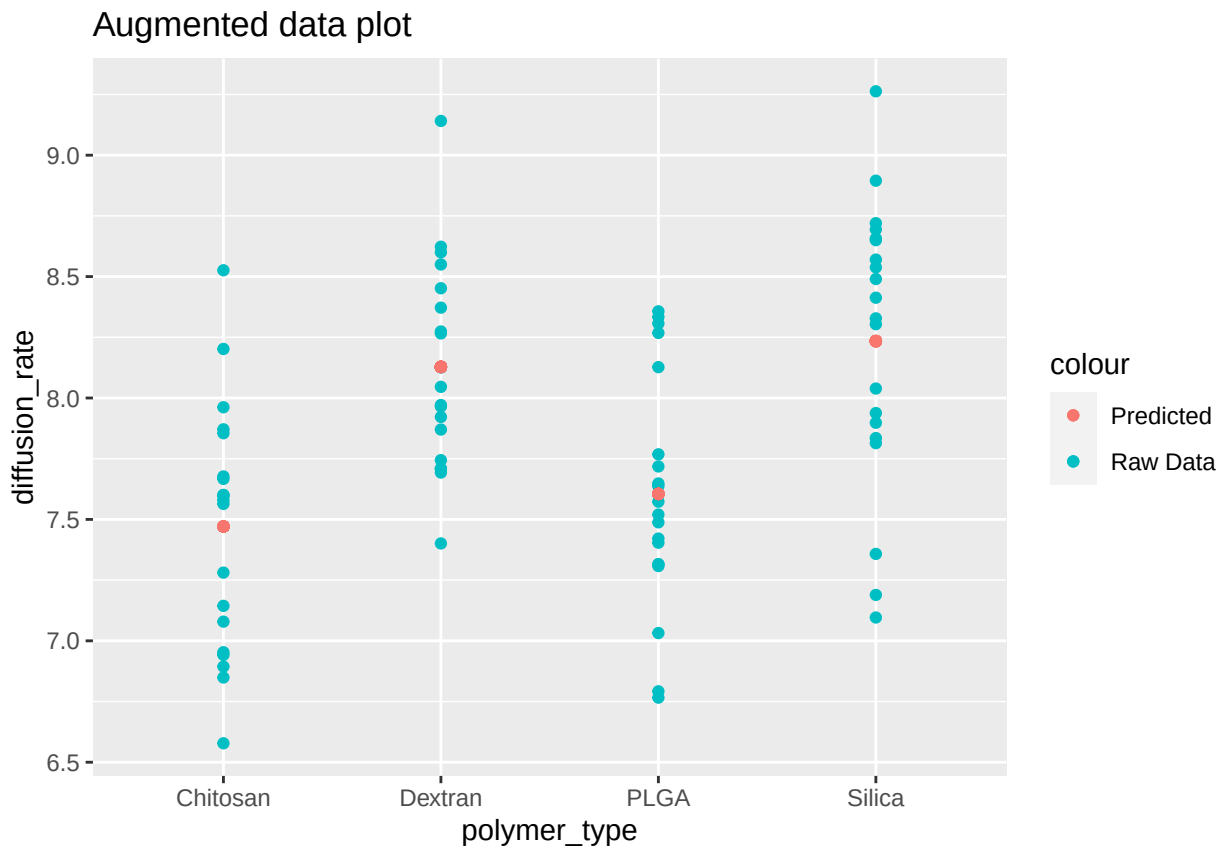
Answer: Based on this pairwise t-test, Dextran is different than PLGA if we assume a p-value threshold of less than 0.05

4) Create two essentially visually identical figures to show both the raw data and the model. 1) from the augmented data (augment function in tidymodels). and 2) Use the outputs of the model fit (lm) to calculate the model's predicted values for each group to include in the plot, and only plotting a single value for each such predicted value for each group (i.e., not using stat_summary or equivalent and with the calculated predicted value in a separate tibble from the data) (7 pt):

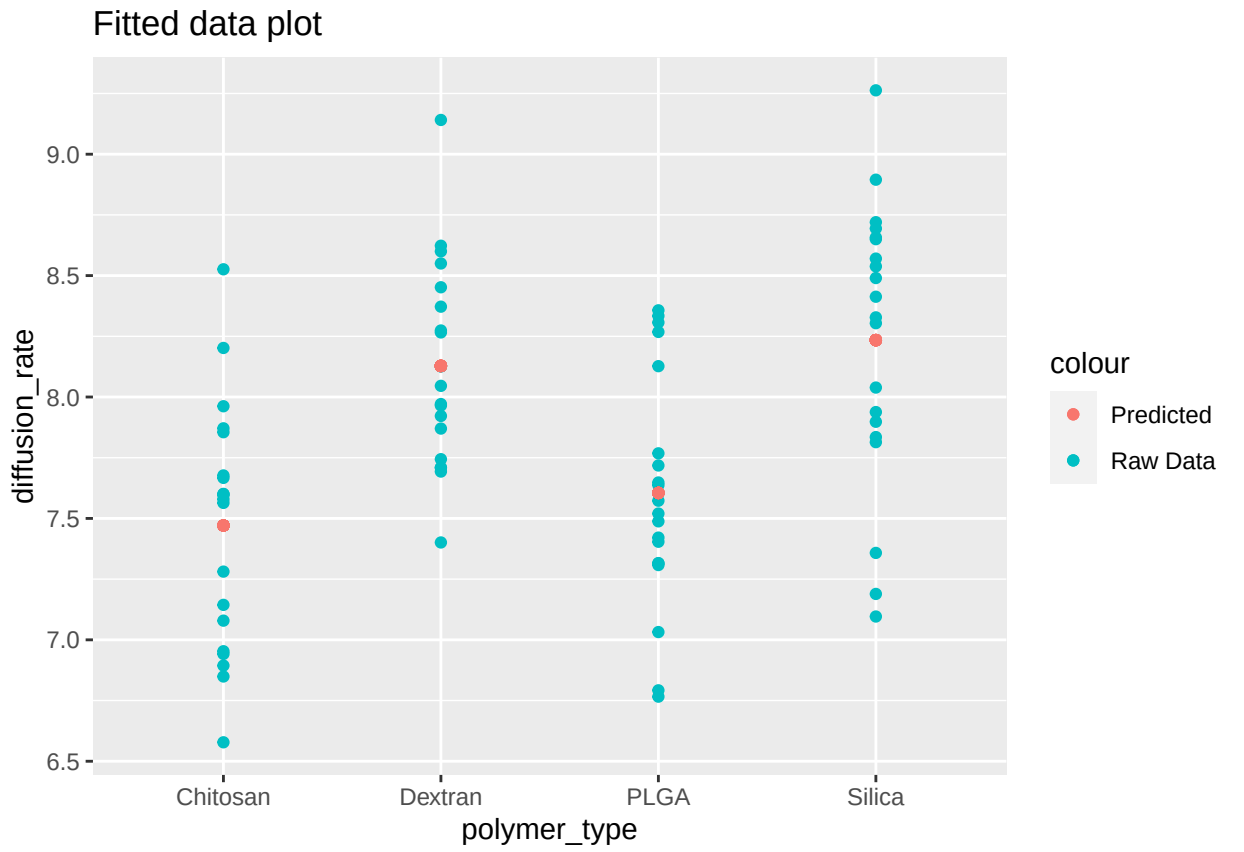
```
augmented_data <- augment(polymer_lm, polymer_restructured)

augmented_data
```

```
## # A tibble: 80 x 10
##    polymer_type diffusion_rate mean_all y_pred .fitted  .resid   .hat .sigma
##    <chr>                 <dbl>    <dbl>  <dbl>   <dbl>   <dbl>  <dbl>  <dbl>
##  1 PLGA                   7.32     7.86   7.60    7.60 -0.290  0.05    0.498
##  2 PLGA                   7.72     7.86   7.60    7.60  0.113  0.0500  0.499
##  3 PLGA                   8.33     7.86   7.60    7.60  0.729  0.0500  0.492
##  4 PLGA                   7.31     7.86   7.60    7.60 -0.297  0.0500  0.498
##  5 PLGA                   6.77     7.86   7.60    7.60 -0.839  0.0500  0.489
##  6 PLGA                   7.42     7.86   7.60    7.60 -0.184  0.0500  0.499
##  7 PLGA                   7.64     7.86   7.60    7.60  0.0320 0.0500  0.499
##  8 PLGA                   8.36     7.86   7.60    7.60  0.752  0.0500  0.491
##  9 PLGA                   7.57     7.86   7.60    7.60 -0.0319 0.0500  0.499
## 10 PLGA                   8.27     7.86   7.60    7.60  0.663  0.0500  0.493
## # i 70 more rows
## # i 2 more variables: .cooksd <dbl>, .std.resid <dbl>
```

```
augmented_data |>
  ggplot(aes(x = polymer_type, y = diffusion_rate)) +
  geom_point(aes(color = "Raw Data")) +
  geom_point(aes(y = .fitted, color = "Predicted")) +
  labs(title = "Augmented data plot")
```




```
polymer_restructured |>
  ggplot(aes(x = polymer_type, y = diffusion_rate)) +
  geom_point(aes(color = "Raw Data")) +
  geom_point(aes(y = y_pred, color = "Predicted")) +
  labs(title = "Fitted data plot")
```

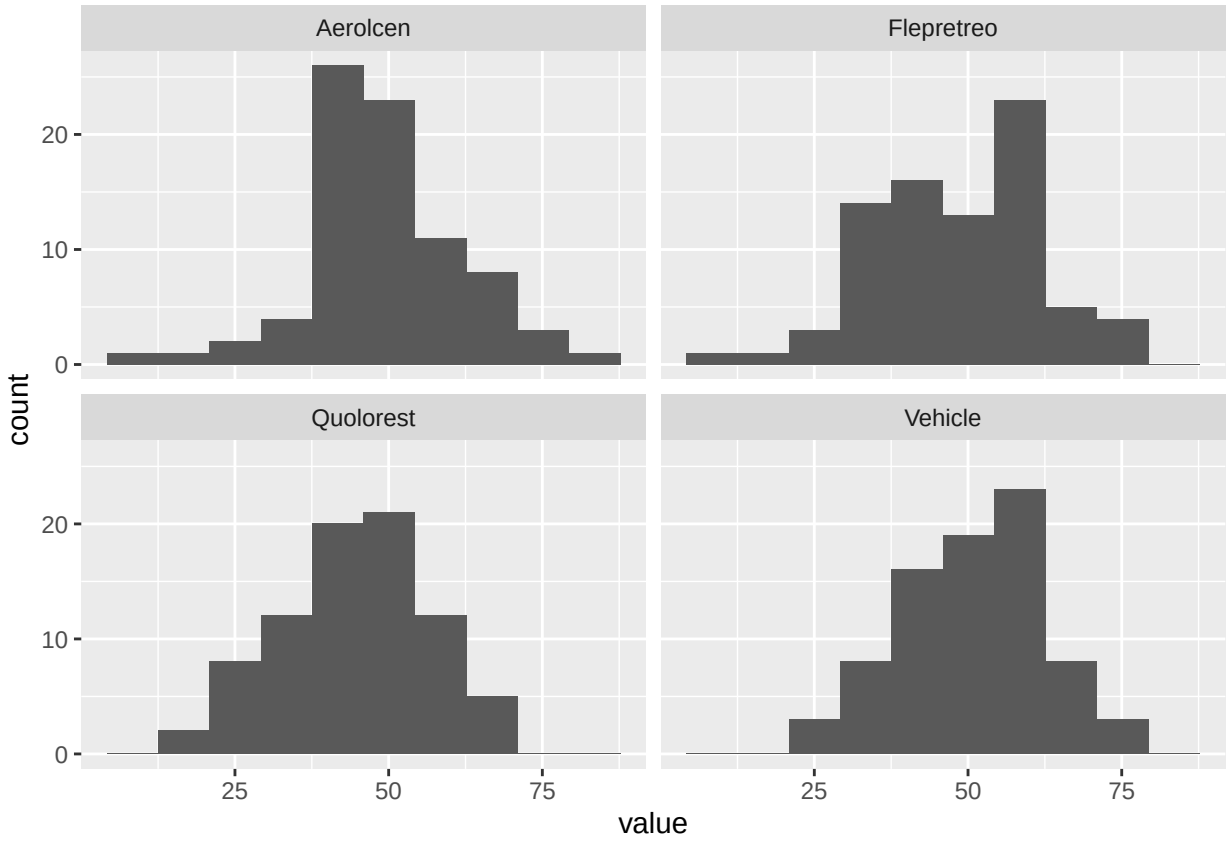


Problem #2 (30 pts): Your group is investigating the encapsulation rate of long-term transcranial implants. In your study, subjects have been implanted with a implants then placed on different therapeutic regimens after implantation. In your study, four different drugs/formulatio ns were used – one control and 3 experimental drugs. The encapsulation level of the implants was evaluated at 1 year after implantation; lower encapsulation is better in this study. Load and prepare data from sheet DataSet2.

a) Visualize your data.

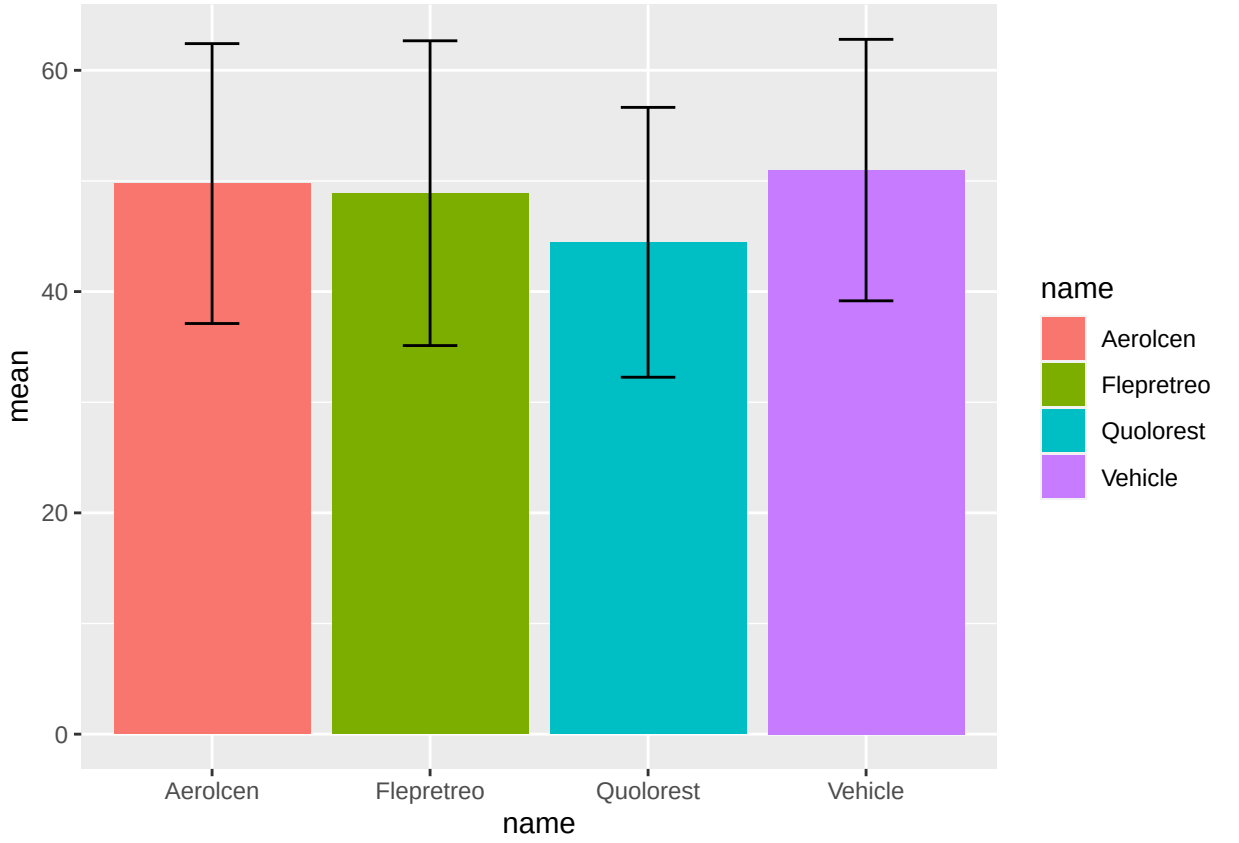
1) Create a histograms for each group in one ggplot call (not multiple separate figures, and be cautious with bins) (4 pts).

```
implant_restructured |> ggplot(aes(x=value)) +
  geom_histogram(bins = 10) +
  facet_wrap(vars(name))
```



2) Create a bar plot that the shows the mean +/- st. dev of each group (4 pts).

```
summarized_implant <- implant_restructured |>
  group_by(name) |>
  summarise(mean = mean(value),
            sd = sd(value),
            lower = mean - sd,
            upper = mean + sd)


summarized_implant |>
  ggplot(aes(x = name, y = mean, fill = name)) +
  geom_bar(stat = "identity") +
  geom_errorbar(aes(ymin = lower, ymax = upper), width = 0.25)
```



b) Describe your regression model

Write out the formula for your regression model in generic terms (i.e., with β terms and no speci

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$$

c) Fit your model

1) What are the coefficients for your model (3 pts)?

```
implant_lm <- lm(value ~ name, implant_restructured)

tidy_implant <- tidy(summary(implant_lm))

tidy_implant
```

```
## # A tibble: 4 x 5
##   term           estimate std.error statistic   p.value
##   <chr>             <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)      49.8        1.41    35.2   1.67e-111
## 2 nameFlepretreo   -0.869      2.00    -0.435 6.64e-  1
## 3 nameQuolorest    -5.31       2.00    -2.66  8.26e-  3
## 4 nameVehicle       1.22       2.00     0.611 5.42e-  1
```

```
mean_aerolcen <- pull(filter(tidy_implant, term=="(intercept)"), estimate)
mean_flepretreo <- pull(filter(tidy_implant, term=="nameFlepretreo"), estimate)
```

```r
mean_quolorest <- pull(filter(tidy_implant, term=="nameQuolorest"), estimate)
mean_vehicle <- pull(filter(tidy_implant, term=="nameVehicle"), estimate)

print(paste0("Coefficient for flepretreo: ", mean_flepretreo))
```

```
## [1] "Coefficient for flepretreo: -0.868605000000001"
```

```r
print(paste0("Coefficient for quolorest: ", mean_quolorest))
```

```
## [1] "Coefficient for quolorest: -5.30788875000002"
```

```r
print(paste0("Coefficient for vehicle: ", mean_vehicle))
```

```
## [1] "Coefficient for vehicle: 1.21940249999999"
```

Answer: The coefficients for this model is the mean encapsulation level of every drug type subtracted by the intercept. One group's mean is the intercept, which does not have a corresponding coefficient.

i. Write out your model with the specific terms replaced by their estimate (as in b) (2pts)

$$y_i = 49.763606 + -0.868605x_{i1} + -5.307889x_{i2} + 1.219402x_{i3} + \epsilon_i$$

2) What is the sum of squares for the drug type in your model (2 pts)?

```r
# Calculating the mean for all the diffusion rates
implant_restructured <- implant_restructured |>
  mutate(mean_all = mean(value)) |>
  mutate(y_pred = fitted(implant_lm))

implant_ss_regression <- sum((implant_restructured$y_pred - implant_restructured$mean_all)^2)

print(paste0("Sum of squares for the drug type: ", implant_ss_regression))
```

```
## [1] "Sum of squares for the drug type: 1941.7529336161"
```

d) Understand/Discuss your model

1) Is the effect of drug type a significant explanation of the total variance in the model? (2 pts)

```r
summary(implant_lm)
```

```
##
## Call:
## lm(formula = value ~ name, data = implant_restructured)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.249  -7.645  -0.071   8.602  32.043
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     49.7636     1.4120  35.243  < 2e-16 ***
## nameFlepretreo  -0.8686     1.9969  -0.435  0.66388
## nameQuolorest   -5.3079     1.9969  -2.658  0.00826 **
## nameVehicle      1.2194     1.9969   0.611  0.54187
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.63 on 316 degrees of freedom
## Multiple R-squared:  0.03709,    Adjusted R-squared:  0.02795
## F-statistic: 4.058 on 3 and 316 DF,  p-value: 0.007508
```

Answer: Yes, the type of drug significantly explains the total variance of the model. This is seen with the p-value of F-statistic for this linear model being less than 0.05 (if below 0.05 is significant).

i. What is the specific statistical hypothesis tested? (2 pts) Answer: The null hypothesis states that there is no significant effect of drug type on the response variable. This could be because the means of the response variable are similar across all drug types

ii. What type of statistic will you use to evaluate this test (2 pts)? Answer: We will use the F-statistic which tests whether the model is a good fit for the data. If it is, then this would suggest that at least one group mean is different from the others.

```
summary(implant_lm)
```

iii. What is the value of the statistic and what is the assiciated p-value (2 pts)?

```
##
## Call:
## lm(formula = value ~ name, data = implant_restructured)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.249  -7.645  -0.071   8.602  32.043
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     49.7636     1.4120  35.243  < 2e-16 ***
## nameFlepretreo  -0.8686     1.9969  -0.435  0.66388
## nameQuolorest   -5.3079     1.9969  -2.658  0.00826 **
## nameVehicle      1.2194     1.9969   0.611  0.54187
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.63 on 316 degrees of freedom
## Multiple R-squared:  0.03709,    Adjusted R-squared:  0.02795
## F-statistic: 4.058 on 3 and 316 DF,  p-value: 0.007508
```

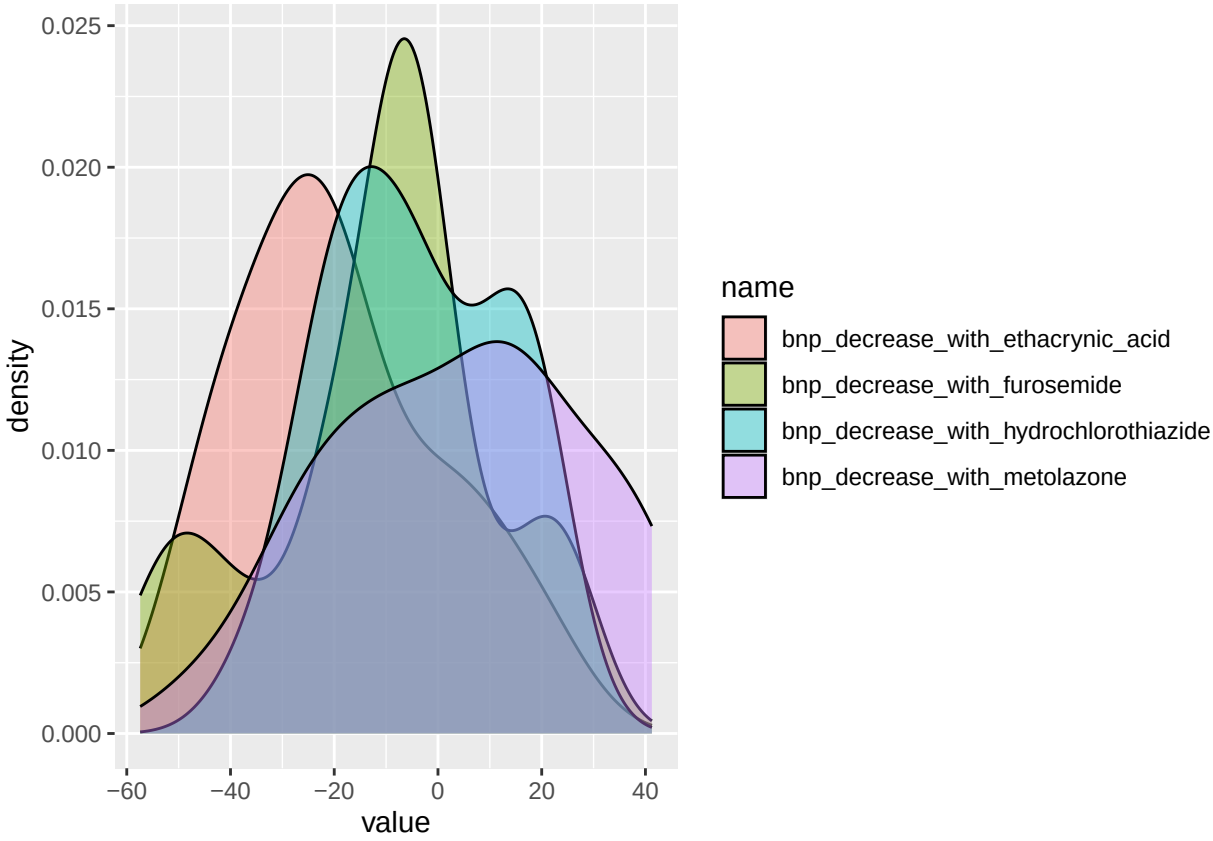
Answer: F-statistic: 4.058; p-value: 0.007508

iv. What can you conclude and what can you not conclude based on this statistical test (4 pts)? Answer: You can conclude that there is a significant difference between one drug types group mean from the others if we assume a p-value threshold of 0.05 which it is currently under. You can not conclude that the drug caused changes in the dependent variable.

Problem #3 (30 pts): Brain natriurectic peptide (BNP) level in blood have been shown to predict heart failure, leading to alterations is cognitive function. In patients with high levels of BNP, your team evaluated the effects of different diuretic drugs on BNP levels. Load and prepare data from sheet DataSet3.

A) Provide a concise set of appropriate visualizations of the data (5 pts)

```
bnp_restructured |> ggplot(aes(x=value, group=name, fill=name)) +
    geom_density(alpha=.4)
```



B) a concise description of your approach to analyzing this data including generic formula given these data (5 pts)

Answer: I would analyze the data by fitting a linear model onto the data with the formula `bnp levels ~ drug type`. Then I could look at the overall model's fit by looking at the p-value of the F-statistic. If it meets my threshold, then I can conclude that I can not reject the null hypothesis that there is no significant difference in the mean BNP levels for any of the drug groups.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$$

C) Fit and interpret your model's summary and Anova tables (15 pts)

```
bnp_lm <- lm(value ~ name, bnp_restructured)

summary(bnp_lm)
```

```
##
## Call:
## lm(formula = value ~ name, data = bnp_restructured)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.625 -12.141   0.669  14.198  41.914
##
## Coefficients:
##                                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                               -18.874      4.171  -4.525 1.73e-05
## namebnp_decrease_with_furosemide            7.624      5.899   1.292 0.199287
## namebnp_decrease_with_hydrochlorothiazide  14.636      5.899   2.481 0.014834
## namebnp_decrease_with_metolazone           22.618      5.899   3.834 0.000225
##
## (Intercept)                               ***
## namebnp_decrease_with_furosemide
## namebnp_decrease_with_hydrochlorothiazide *
## namebnp_decrease_with_metolazone          ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.86 on 96 degrees of freedom
## Multiple R-squared:  0.1438, Adjusted R-squared:  0.117
## F-statistic: 5.372 on 3 and 96 DF,  p-value: 0.001838
```

Answer: The p-value of the F-statistic is 0.001838. Which suggests that the model is a good fit to the data if our threshold is 0.05. The estimate of the model's intercept is -18.874 with a p-value of 1.73e-05. This means that we can reject the null hypothesis that the intercept is 0 assuming a threshold of 0.05 for p-values.

The estimate of the coefficient for the variable of whether the data point is furosemide is 7.624 with a p-value of 0.199287. This means we can not reject the null hypothesis that the coefficient is 0 for the variable of whether the data point is furosemide assuming a threshold of 0.05 for p-values.

The estimate of the coefficient for the variable of whether the data point is hydrochlorothiazide is 14.636 with a p-value of 0.014834. This means we can reject the null hypothesis that the coefficient is 0 for the variable of whether the data point is hydrochlorothiazide assuming a threshold of 0.05 for p-values.

The estimate of the coefficient for the variable of whether the data point is metolazone is 22.618 with a p-value of 0.000225. This means we can reject the null hypothesis that the coefficient is 0 for the variable of whether the data point is metolazone assuming a threshold of 0.05 for p-values.

E) What can you conclude scientifically as a result of your analysis and visualization? be concise but clear (5pts)

Answer: We can scientifically conclude that the mean BNP level is significantly different for one of the diuretic groups compared to all the others.