Introduction
The database is based on fictional data from an upstream manufacturing process in a biopharmaceutical company. The variables in the dataset are VCD(Viable cell density), Viability, Glucose, and lactate.

Objective
The objective of this project is to find the best predictor for VCD between the 3 predictor variables (Viability, Glucose, and lactate). Objective number 2 is to create a model to predict what VCD will be when the predictor variables are at various levels.

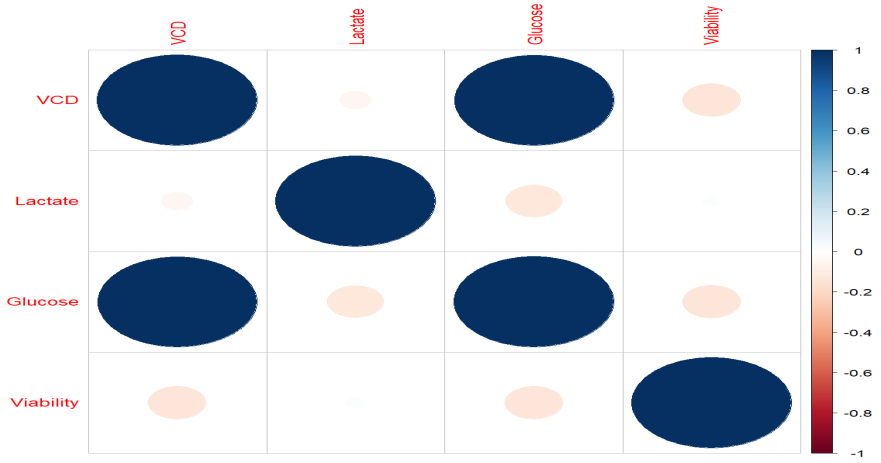
Part 1
Firstly, I read the Excel dataset into r using the readxl package. I used the str function to check the dataset was read in the correct format, i.e., numeric isn't uploaded as a character. All the variables were in the correct numeric variable. I used the summary function to analyze the spread of the data, the summary function shows the minimum for each variable, and the 1st quarter number shows that 25% of the data falls below the number outlined. For example(see image 1), for VCD, 25% of the data in that variable falls below 21.22.
The Median is the middle data, and The 3rd quartile shows 25% of the data is greater than 21.24, Max is the maximum number within the variable.

Image 1 Summary of Biomass Data

```
> summary(Biomass_data)
      VCD            Lactate             Glucose          Viability
 Min.   :21.21   Min.   :0.007728   Min.   :0.1805   Min.   :89.71
 1st Qu.:21.22   1st Qu.:0.009121   1st Qu.:0.1928   1st Qu.:91.32
 Median :21.23   Median :0.009756   Median :0.2000   Median :93.17
 Mean   :21.23   Mean   :0.009803   Mean   :0.1993   Mean   :92.87
 3rd Qu.:21.24   3rd Qu.:0.010499   3rd Qu.:0.2055   3rd Qu.:94.09
 Max.   :21.26   Max.   :0.011623   Max.   :0.2198   Max.   :96.83
>
```

I used the cor and corrplot functions to determine which variable(s) correlate with the VCD variable.

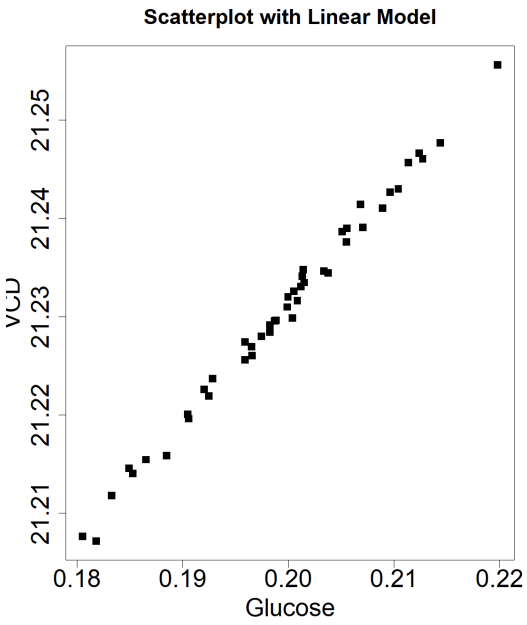
Image 2: Corrplot/Cor




```
> cor(Biomass_data)
                VCD     Lactate    Glucose   Viability
VCD         1.00000000 -0.04111483  0.9963982 -0.13303926
Lactate    -0.04111483  1.00000000 -0.1255034  0.01297057
Glucose     0.99639823 -0.12550340  1.0000000 -0.13324508
Viability  -0.13303926  0.01297057 -0.1332451  1.00000000
```

It's clear from image 3 that the greatest correlation to VCD is Glucose. The size and colour of the circles represent the level of correlation. In addition, there doesn't seem to be any multicollinearity present; The numbers in the cor table represent the level of correlation, 1 represents perfect correlation, and the only variables that have a close correlation to 1 are Glucose and VCD with a cor number of 0.99639823.

Image 3: Scatterplot



I used the plot function to see if the correlation was linear. Image 3 shows a clear linear correlation between glucose and VCD.

Image 4: Model output

```
Call:
lm(formula = Biomass_data$VCD ~ Biomass_data$Glucose)

Residuals:
      Min         1Q     Median         3Q        Max
-0.0021755 -0.0006998 -0.0001201  0.0006732  0.0019046

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        20.989881   0.003125 6716.21   <2e-16 ***
Biomass_data$Glucose  1.206831   0.015662   77.05   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.000952 on 43 degrees of freedom
Multiple R-squared:  0.9928,    Adjusted R-squared:  0.9926
F-statistic:  5937 on 1 and 43 DF,  p-value: < 2.2e-16
```

The residuals are the difference between the actual values and the predicted values. Looking at the data numbers given by the residuals, I can infer that the residuals are behaving normally because the 1Q and 3Q are of similar magnitudes and the min(the residual closest to the line) and the Max(residual furthest from the line) have a similar magnitude. The median is close to zero, which represents symmetric residuals.

The coefficient estimate is the expected value given for VCD when glucose is 0.
The slope is saying for every 1 increase in glucose there will be an increase of 1.206831 increase in VCD.

The standard error is the average amount the estimate varies from the actual value, it is therefore ideal that the standard error is a low number of 0.015662.

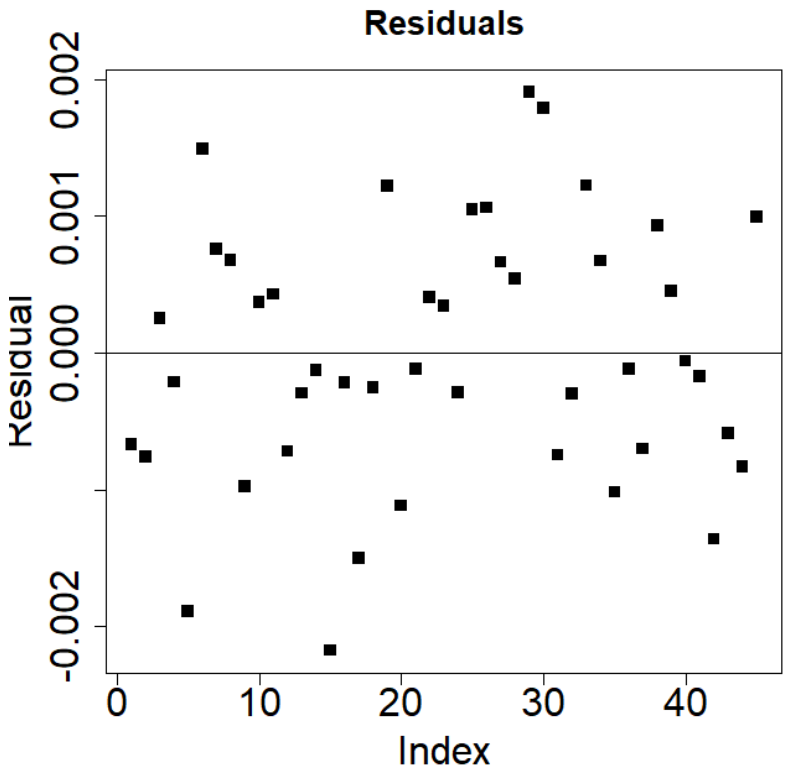
T-value
Is a measure of how many standard deviations there are between the estimate and 0. In general, if the estimate has a high magnitude then the coefficient is statistically significant, in this case, the T value is high at 77.05.

Pr

The P=value is showing 2e-16 which is well over the benchmark of 0.05, this result signifies that the correlation between VCD and glucose is unlikely due to chance.

It is important to note That the T value and P value can only be used if the residuals are normal.

Image 5: Residual Index plot



It is clear from the residuals index plot that the residuals are normal because the spread of data points along the line is even.

Image 6: Histogram of Residuals

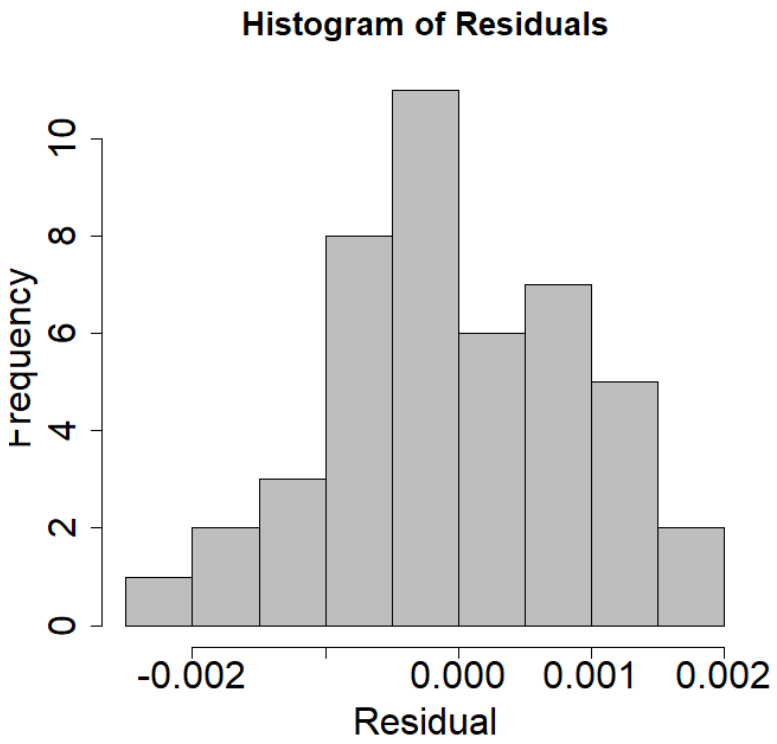
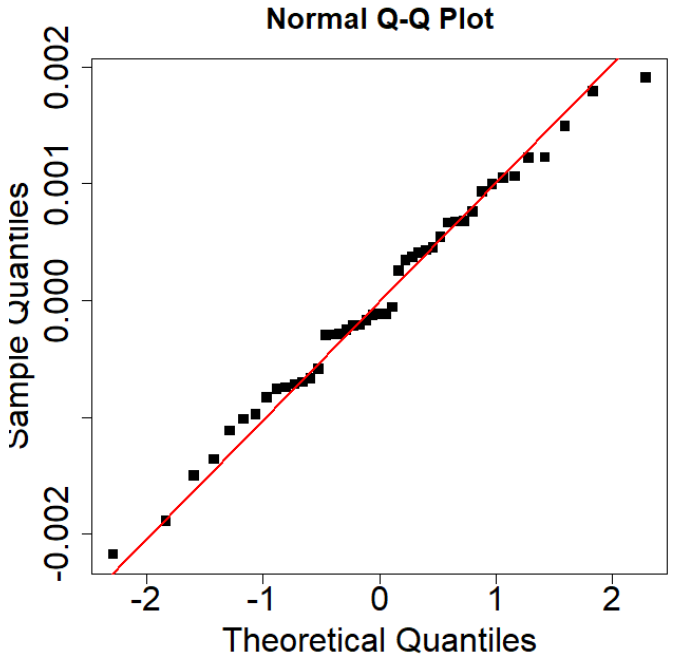


Image 7: QQ plot



The qqplot(Image 7) shows normal distribution as the residuals are forming a symmetric scatter along the redline(line of best fit); this symmetric shape is better represented in the histogram(Image 6). Additionally, the line represents how well the model predicts VCD because of how close the residuals are to the line.

The normality of the residuals can be mathematically confirmed using the Shapiro-Wilk test. The Shapiro-Wilk test null hypothesis says that the residuals are normal. If the value obtained is less than the allocated alpha value(0.05) then the null hypothesis must be rejected. In this case, the residuals are well over 0.05 with a value of 0.926, we can therefore assume the residuals are normal because we fail to reject the null hypothesis.

Residual standard Error

This tells you how much the actual value will differ from the value that you're trying to predict. In this case, the residual standard error is very low at 0.000952 which is indicating that the model is accurate at predicting VCD.

Degrees of Freedom

This is the number of data points that were added to the estimation, so it will be a number of rows subtracted by the number of variables being estimated including the intercept. In this study, the degrees of freedom was just 43.

Multiple R-squared values

This is the percentage of variance from the response variable that can be explained by the predictive variables. The multiple R -squared value is 0.9928 which indicates that 99% of the variance in the response variable can be explained using the predictor variable glucose.

Adjusted R-squared

This controls for the overfitting of the model. This will add penalties when a new predictor is added because generally, each predictor will explain a portion of the response. If there is a large difference between the R squared and Adjusted R-squared then it is likely that the model has been overfitted. In this case, the Multiple and R-squared values have similar values as expected because there is only one predictor variable used in this model.

F-statistics

A general indicator if there is a relationship between response variables and the predictor variables, the ideal result is a number that is further away from 1. In this case, the F-statistic is telling me that there is a great correlation between my predicted variable and my predictor variable because the F-statistic is 5937.

P -value

The P value as a whole will determine if the model is statistically significant. A value of 2.2e-16 is much lower than the benchmark of 0.05 which means we fail to approve the null hypothesis that states that there is no significant relationship between the predictor variable and the predicted variable.

Part 2

Predict using the Regression model

I created a model that would allow me to predict VCD when the Viability, Glucose, and lactate are at various levels.

I created the model by using the lm function to add the variables and I called the new model model2(Image 8).

Image 8:Model2 lm Output

```
Call:
lm(formula = Biomass_data$VCD ~ Biomass_data$Glucose + Biomass_data$Lactate +
    Biomass_data$Viability)

Residuals:
       Min         1Q     Median         3Q        Max
-6.700e-05 -4.586e-05 -2.178e-05  2.776e-05  2.245e-04

Coefficients:
                      Estimate Std. Error   t value Pr(>|t|)
(Intercept)           2.098e+01  5.930e-04 35375.137   <2e-16 ***
Biomass_data$Glucose  1.220e+00  1.096e-03  1112.743   <2e-16 ***
Biomass_data$Lactate  9.912e-01  1.043e-02    95.065   <2e-16 ***
Biomass_data$Viability 2.845e-07  5.468e-06     0.052    0.959
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.552e-05 on 41 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 4.208e+05 on 3 and 41 DF,  p-value: < 2.2e-16
```

As expected Glucose is showing 3 significant astrix. Lactate is also showing 3 significant astrix values. I didn't expect the Viability to have 0 significant astrix based on the relationship Viability and VCD have in a real-world manufacturing process. The residual standard error is better in model2 than model1(0.0000952) because it is lower at 0.00006552. The multiple R-squared and adjusted R-squared values are both 1 which is an indication that all the variability in the response variable can be predicted by the predictor variable; this is expected because the data provided is sparse at just 41, which is indicated by the degrees of freedom. I suspect the R-squared value would be reduced if more data was provided. The F-statistic should always be taken while considering the P-value. In model2 the F value is large(4.208e+05) and the P value is lower than the 0.05 benchmark so we can assume that the correlation between the response variable and the predictors is significant.

So I want to determine the VCD when my predictor variables are adjusted to different levels. Firstly I gave my predictor variable random values;

Vialbiliy = .97
Glucose = .015
Lactate= 0.02

Then I predicted the results for each variable by adding the intercept estimate and all the predictors, then I multiplied by the value given for the variable of interest, please see Image 9 below

Image 9 Prediction formulas

```
##VCD when Viability is .97###
VCD_Viability= 2.098e+01  +  1.220e+00 +9.912e-01 +  2.845e-07 * Vialbiliy
VCD_Viability
##VCD when Glucose is 0.15###

VCD_Glucose= 2.098e+01  +  1.220e+00 * Glucose +9.912e-01 +  2.845e-07
VCD_Glucose


##VCD when Lactate is 0.02####

VCD_Lactate= 2.098e+01  +  1.220e+00 +9.912e-01* Lactate +  2.845e-07
VCD_Lactate
```

Conclusion

In this study, glucose was the best predictor for VCD, and Viability was a weak predictor for VCD.