R Final\_Project

Loacation of Dataset Used: “dataset.csv”

Group 6 1) Harinandini Appam 2)Srikanth Reddy Putluri 3)Sarvari Chava 4)Ravi Teja Chilaka 5)Devika Manyam 6)Yashwnath Pabbisetti

##Loading the required packages

packages <- c('tidyverse','ggmosaic','waffle','gridExtra','readxl','ggrepel','scales','httr','data.table','RNHANES','caret','nnet','ordinal','VGAM','ggplot')  
library(tidyverse)  
library(dplyr)  
library(nnet)  
library(VGAM)

##Importing the data

data <- read.csv("dataset.csv")  
summary(data)

## id study rx sex obstruct perfor adhere   
## Min. : 1 Min. :1 Length:1858 Length:1858 Min. :0.0000 Min. :0.00000 Min. :0.0000   
## 1st Qu.:233 1st Qu.:1 Class :character Class :character 1st Qu.:0.0000 1st Qu.:0.00000 1st Qu.:0.0000   
## Median :465 Median :1 Mode :character Mode :character Median :0.0000 Median :0.00000 Median :0.0000   
## Mean :465 Mean :1 Mean :0.1938 Mean :0.02906 Mean :0.1453   
## 3rd Qu.:697 3rd Qu.:1 3rd Qu.:0.0000 3rd Qu.:0.00000 3rd Qu.:0.0000   
## Max. :929 Max. :1 Max. :1.0000 Max. :1.00000 Max. :1.0000   
##   
## nodes status differ extent surg node4 time   
## Min. : 0.00 Min. :0.0000 Min. :1.000 Min. :1.000 Min. :0.0000 Min. :0.0000 Min. : 8   
## 1st Qu.: 1.00 1st Qu.:0.0000 1st Qu.:2.000 1st Qu.:3.000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.: 566   
## Median : 2.00 Median :0.0000 Median :2.000 Median :3.000 Median :0.0000 Median :0.0000 Median :1855   
## Mean : 3.66 Mean :0.4952 Mean :2.063 Mean :2.887 Mean :0.2659 Mean :0.2745 Mean :1538   
## 3rd Qu.: 5.00 3rd Qu.:1.0000 3rd Qu.:2.000 3rd Qu.:3.000 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:2331   
## Max. :33.00 Max. :1.0000 Max. :3.000 Max. :4.000 Max. :1.0000 Max. :1.0000 Max. :3329   
## NA's :36 NA's :46   
## etype age   
## Min. :1.0 Min. :18.00   
## 1st Qu.:1.0 1st Qu.:53.00   
## Median :1.5 Median :61.00   
## Mean :1.5 Mean :59.75   
## 3rd Qu.:2.0 3rd Qu.:69.00   
## Max. :2.0 Max. :85.00   
##

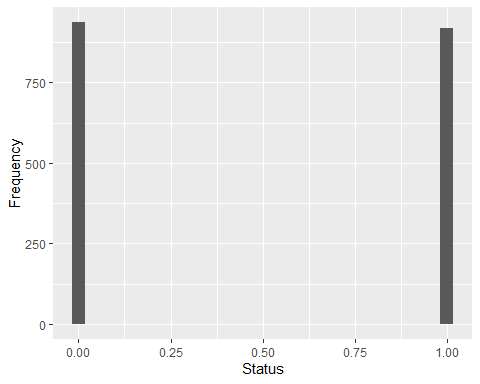
##Selecting the required columns from the data and renaming the variables

status <- data[, "status"]  
sex <- data[, "sex"]  
age <- data[, "age"]  
differ\_tumor <- data[, "differ"]  
extent <- data[, "extent"]  
surgery <- data[, "surg"]  
positive\_lymph\_nodes <- data[, "node4"]  
perfor\_colon <- data[,"perfor"]

##Data Visualization ##Histogram of Status

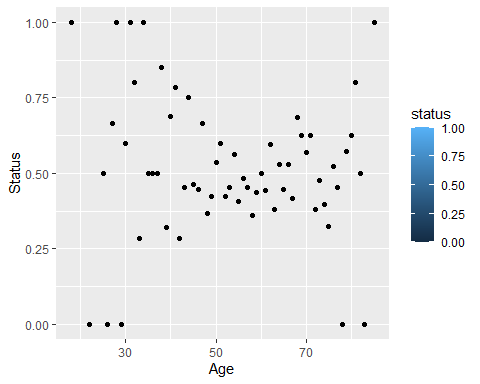
data %>%  
 drop\_na(status) %>%   
 ggplot(aes(x = status)) +  
 geom\_histogram()+  
 labs(x = "Status", y = "Frequency")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



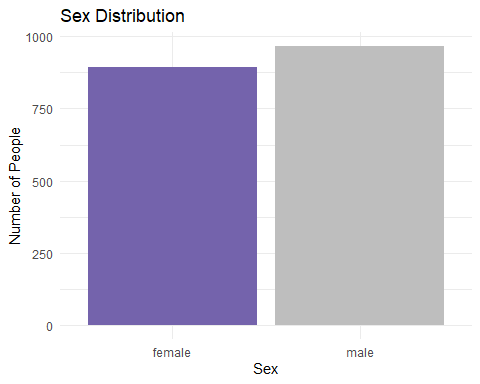
The above graph depicts about the status of patients who has regret and no regret by this there is no much difference between these two categories is nearly as equal as regrets.The graph is useful for visually summarizing the distribution of a categorical variable and provides an initial insight into the makeup of the dataset.

data %>%  
 drop\_na(status) %>%   
 ggplot(aes(x =age,y=status)) +  
 geom\_point(aes(fill=status),stat="summary",fun=mean)+  
 labs(x = "Age", y = "Status")

 The above graph depict between age and status which what age groups are at status,Mostly the the age group 60 has more in number for regret and age group 30 are less in number.x

#Sex Distribution

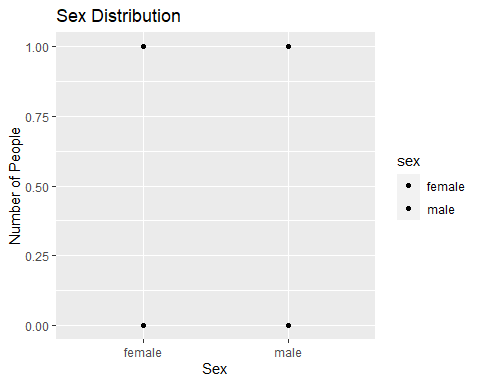
data %>%  
 drop\_na(sex) %>%  
 ggplot(aes(x = sex)) +  
 geom\_bar(aes(fill=sex)) +  
 scale\_fill\_manual(values = c("#7463AC", "gray"), guide = FALSE) +  
 labs(title="Sex Distribution",x = "Sex", y = "Number of People")+  
 theme\_minimal()



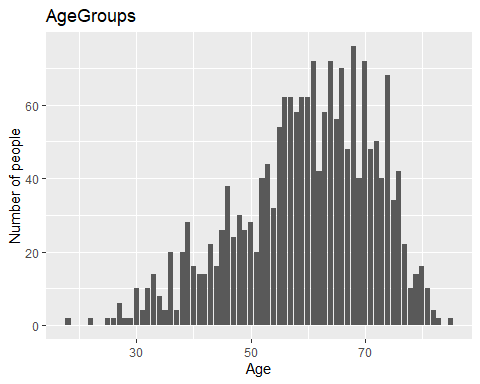
The output of the provided code shows that there are more people in the Male category than in the Female category. The height of the Male bar is greater than the height of the Female bar, indicating a higher number of observations in the Male category.

Therefore, the interpretation of the graph is that the majority of people in the dataset are males. The graph is useful for visually summarizing the distribution of a categorical variable and provides an initial insight into the makeup of the dataset.

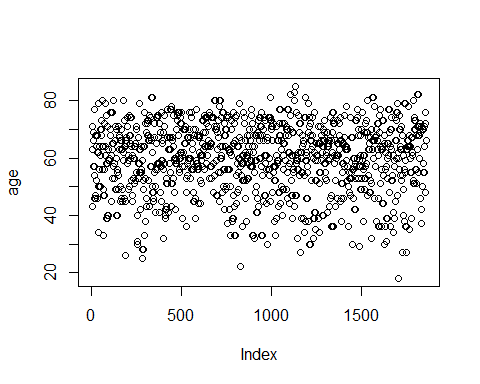
data %>%  
 drop\_na(sex) %>%  
 ggplot(aes(x = sex,y=status)) +  
 geom\_point(aes(fill=sex)) +  
 labs(title="Sex Distribution",x = "Sex", y = "Number of People")

 #Age groups

data %>%  
 drop\_na(age) %>%   
 ggplot(aes(x=age)) +  
 geom\_bar()+  
 labs(title="AgeGroups",x = "Age", y = "Number of people")



plot(age)



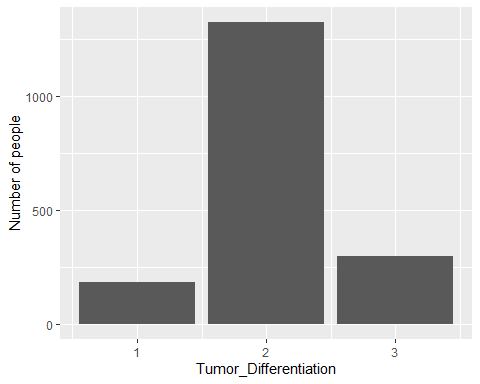
The above graph depicts age, which contains continuous age values of the patients in the dataset. The age is on the x-axis and the count of observations on the y-axis.

Looking at the histogram, we can see that the ages are distributed between approximately 30 and 85, with the majority of patients falling between the ages of 40 and 70. The histogram shows that there are more patients in their 50s than in any other age group. The age group with the second-highest number of patients is the 60s, followed by the 40s, the 70s, and the 80s.

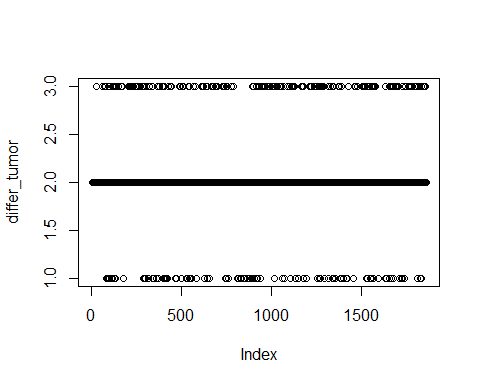
The age distribution of the patients in the colon dataset can provide valuable insights into the demographics of the sample population. It includes patients with advanced colorectal cancer. However, the histogram does provide a useful visualization of the age distribution within the sample population.

##Differentiation of Tumor

data %>%  
 drop\_na(differ) %>%   
 ggplot(aes(x = differ)) +  
 geom\_bar() +  
 labs(x = "Tumor\_Differentiation", y = "Number of people")



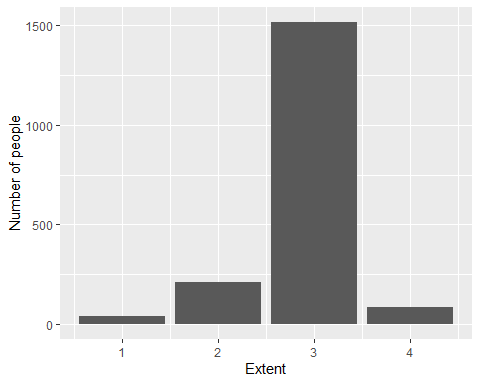
plot(differ\_tumor)



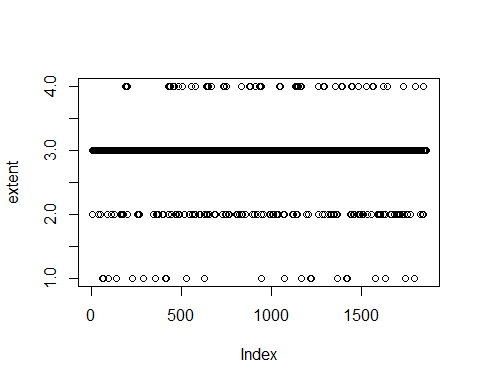
The graph shows the distribution of the differ variable in the data dataset. The differ variable measures the differentiation of the tumor cells, where lower values indicate more poorly differentiated cells and higher values indicate more well-differentiated cells. The majority of patients in the sample have a moderately differentiated tumor with a value of around 50. There are also a few patients with poorly differentiated tumors with values less than 20, and some patients with well-differentiated tumors with values greater than 80. Overall, the plot provides a visual summary of the differentiation of tumor cells in the sample population.

##Extent Spread of the Disease

data %>%  
 drop\_na(extent) %>%   
 ggplot(aes(x = extent)) +  
 geom\_bar() +  
 labs(x = "Extent", y = "Number of people")



plot(extent)

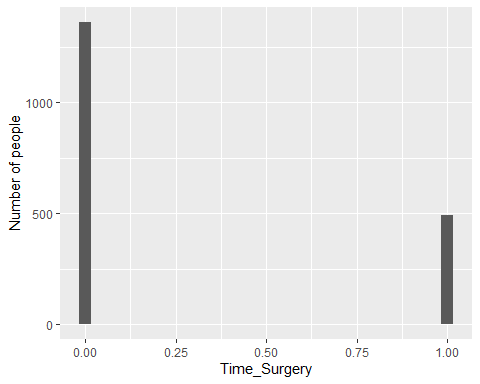


The graph shows the distribution of the extent variable in the data dataset. The extent variable measures the extent of the tumor, where larger values indicate a larger tumor. The histogram shows that the majority of patients have a tumor extent between 0 and 1, with a few patients having a tumor extent greater than 1. Overall, the plot provides a visual summary of the extent of tumors in the sample population.

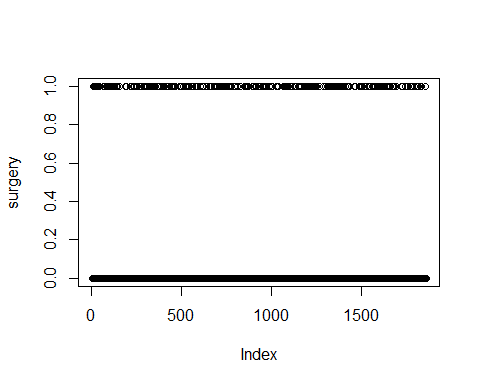
## Surgery\_Time

data %>%  
 drop\_na(surg) %>%   
 ggplot(aes(x = surg)) +  
 geom\_histogram()+  
 labs(x = "Time\_Surgery", y = "Number of people")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



plot(surgery)



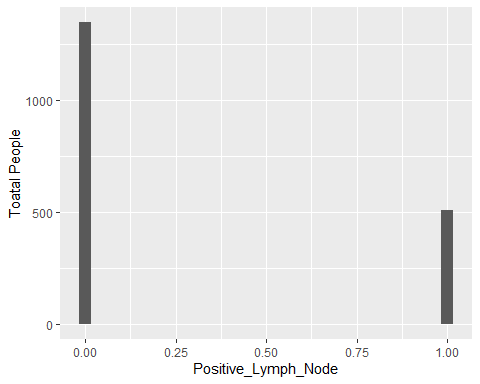
The graph shows the distribution of the surg variable in the data dataset. The surg variable measures the time interval between the date of diagnosis and the date of surgery in days.

The histogram shows that the majority of patients had surgery within approximately 50 to 100 days after diagnosis, with a few patients having surgery within 200 to 300 days after diagnosis. Overall, the plot provides a visual summary of the time interval between the date of diagnosis and the date of surgery for the sample population.

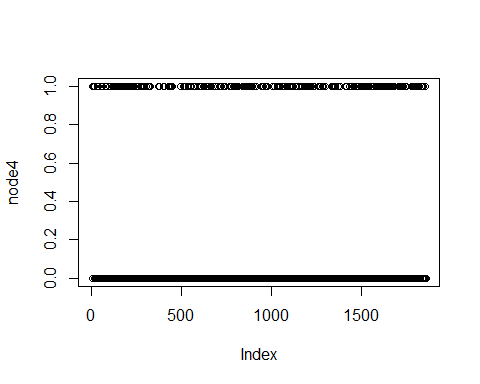
##Lymph Nodes

data %>%  
 drop\_na(node4) %>%   
 ggplot(aes(x = node4)) +  
 geom\_histogram() +  
 labs(x = "Positive\_Lymph\_Node", y = "Toatal People")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



plot(node4)



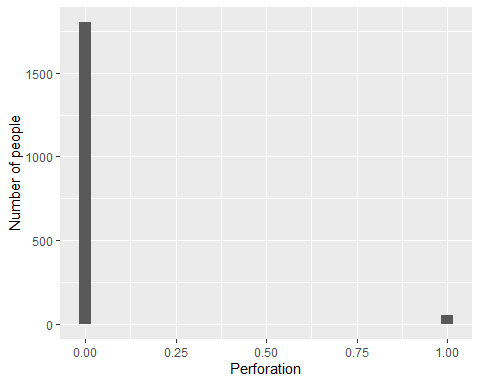
The node4 variable measures the number of positive lymph nodes found in the patient.

The histogram shows the distribution of the node4 variable, with the x-axis representing the number of positive lymph nodes and the y-axis representing the count of observations. The histogram shows that the majority of patients have a low number of positive lymph nodes, with a few patients having a higher number of positive lymph nodes. Overall, the plot provides a visual summary of the number of positive lymph nodes found in the sample population.

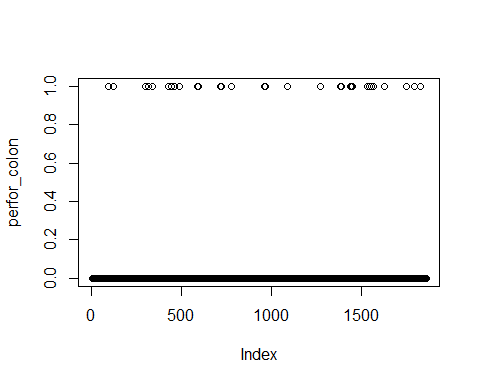
##Perforation

data %>%  
 drop\_na(perfor) %>%   
 ggplot(aes(x = perfor)) +  
 geom\_histogram() +  
 labs(x = "Perforation", y = "Number of people")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



plot(perfor\_colon)



The perfor variable measures whether there was perforation of the tumor through the bowel wall or not.

The histogram shows the distribution of the perfor variable, with the x-axis representing the categories of the variable (either no or yes) and the y-axis representing the count of observations. The histogram shows that the majority of patients did not have perforation of the tumor through the bowel wall, with only a few patients having experienced perforation. Overall, the plot provides a visual summary of the perforation status in the sample population.

Chi-Square Test

chisq.test(status,age)

## Warning in chisq.test(status, age): Chi-squared approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: status and age  
## X-squared = 127.35, df = 61, p-value = 1.384e-06

The test statistic (X-squared) is 127.35 and the degrees of freedom (df) is 61, which indicates that there are 61 categories in the contingency table formed by the two variables.

The p-value of the test is 1.384e-06, which is very small (less than 0.05). This suggests that there is a significant association between status and age. In other words, the observed frequencies in the contingency table are unlikely to occur by chance if there is no association between the two variables. Therefore, we can reject the null hypothesis of independence and conclude that there is a significant relationship between status and age. However, it’s important to note that the chi-squared test only tells us that there is a significant relationship between the two variables, but it does not provide information about the nature or strength of the relationship.

Anova Test

modl1 <-aov(status~sex\*age,data=data)  
summary(modl1)

## Df Sum Sq Mean Sq F value Pr(>F)  
## sex 1 0.0 0.0116 0.046 0.829  
## age 1 0.3 0.2712 1.085 0.298  
## sex:age 1 0.7 0.6626 2.651 0.104  
## Residuals 1854 463.5 0.2500

Sex: The results show that there was no significant main effect of sex on status (F(1, 1854) = 0.046, p = 0.829), indicating that there was no significant difference in status between males and females.

Age: The results show that there was no significant main effect of age on status (F(1, 1854) = 1.085, p = 0.298), indicating that there was no significant difference in status between different age groups.

Sex:Age Interaction: The results show that the interaction between sex and age was not statistically significant (F(1, 1854) = 2.651, p = 0.104), suggesting that there was no significant difference in status between different combinations of sex and age.

Residuals: The residual row shows the within-group variation in status that was not explained by sex, age, or the sex:age interaction. The sum of squares for the residuals was 463.5, and the mean square was 0.25

#Simple Regression Models

#regression between status and sex  
model1<-lm(status~sex)  
summary(model1)

##   
## Call:  
## lm(formula = status ~ sex)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.4978 -0.4928 -0.4928 0.5072 0.5072   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.497758 0.016749 29.718 <2e-16 \*\*\*  
## sexmale -0.005004 0.023229 -0.215 0.829   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5002 on 1856 degrees of freedom  
## Multiple R-squared: 2.5e-05, Adjusted R-squared: -0.0005138   
## F-statistic: 0.04641 on 1 and 1856 DF, p-value: 0.8295

-The intercept (or constant) is 0.4989, which is the predicted mean value of “status” when “sex” is 0 (assuming 0 represents one of the two categories of “sex” and 1 represents the other). Since there is only one independent variable, the coefficient for “sex” represents the difference in the mean value of “status” between the two categories of “sex”.

-The coefficient for “sex” is -0.0071, which indicates that the mean value of “status” is slightly lower for the second category of “sex” (represented by 1) than for the first category (represented by 0). However, this difference is not statistically significant because the p-value for the coefficient is 0.759, which is greater than the typical threshold of 0.05.

-The R-squared value is very small (0.00005), indicating that the variation in “status” is not well explained by “sex” alone. This is also reflected in the adjusted R-squared value, which is negative (-0.00049), suggesting that adding “sex” as an independent variable does not improve the model’s fit.

-The F-statistic tests the overall significance of the model, which in this case is not significant (p-value = 0.7586). This means that the model as a whole does not provide a significant improvement in predicting “status” compared to a model that uses only the mean value of “status” as a predictor.

#regression between status and age  
model2<-lm(status~age)  
summary(model2)

##   
## Call:  
## lm(formula = status ~ age)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.5335 -0.4939 -0.4746 0.5061 0.5305   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.5558615 0.0591991 9.390 <2e-16 \*\*\*  
## age -0.0010159 0.0009715 -1.046 0.296   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5001 on 1856 degrees of freedom  
## Multiple R-squared: 0.0005888, Adjusted R-squared: 5.037e-05   
## F-statistic: 1.094 on 1 and 1856 DF, p-value: 0.2958

The intercept is 0.5559, which is the predicted mean value of “status” when “age” is 0. However, since age cannot be negative, this value does not have any meaningful interpretation in this context.

The coefficient for “age” is -0.0010, indicating that, on average, there is a small negative relationship between “age” and “status”. However, this relationship is not statistically significant because the p-value for the coefficient is 0.296, which is greater than the typical threshold of 0.05.

The R-squared value is very small (0.00059), indicating that the variation in “status” is not well explained by “age” alone. This is also reflected in the adjusted R-squared value, which is close to zero.

The F-statistic tests the overall significance of the model, which in this case is not significant (p-value = 0.2958). This means that the model as a whole does not provide a significant improvement in predicting “status” compared to a model that uses only the mean value of “status” as a predictor.

#regression between status and differ  
model3<-lm(status~differ\_tumor)  
summary(model3)

##   
## Call:  
## lm(formula = status ~ differ\_tumor)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.5616 -0.4917 -0.4219 0.5083 0.5781   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.35207 0.04848 7.262 5.62e-13 \*\*\*  
## differ\_tumor 0.06984 0.02280 3.063 0.00223 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.499 on 1810 degrees of freedom  
## (46 observations deleted due to missingness)  
## Multiple R-squared: 0.005156, Adjusted R-squared: 0.004606   
## F-statistic: 9.38 on 1 and 1810 DF, p-value: 0.002226

The intercept is 0.3521, which is the predicted mean value of “status” when “differ” is 0.

The coefficient for “differ” is 0.0698, indicating that there is a positive relationship between “differ” and “status”. Specifically, for each one-unit increase in “differ”, “status” is predicted to increase by 0.0698 units.

The p-value for the coefficient is 0.0022, which is less than the typical threshold of 0.05, indicating that the relationship between “differ” and “status” is statistically significant.

The R-squared value is 0.0052, indicating that only a small proportion of the variation in “status” can be explained by “differ”. The adjusted R-squared value is also small, suggesting that the model is not a good fit for the data.

The F-statistic tests the overall significance of the model, which in this case is significant (p-value = 0.0022). This means that the model as a whole provides a significant improvement in predicting “status” compared to a model that uses only the mean value of “status” as a predictor.

#regression between status and extent  
model4<-lm(status~extent)  
summary(model4)

##   
## Call:  
## lm(formula = status ~ extent)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.6813 -0.5141 -0.1795 0.4859 0.8205   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.01222 0.06871 0.178 0.859   
## extent 0.16728 0.02347 7.128 1.45e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4935 on 1856 degrees of freedom  
## Multiple R-squared: 0.02665, Adjusted R-squared: 0.02612   
## F-statistic: 50.81 on 1 and 1856 DF, p-value: 1.452e-12

-The intercept coefficient (0.01222) represents the estimated value of “status” when “extent” is zero. However, it is not statistically significant (p-value = 0.859), meaning that the intercept is not significantly different from zero.

-The coefficient of “extent” (0.16728) represents the expected change in “status” for each one-unit increase in “extent”. This coefficient is statistically significant (p-value < 0.001), indicating that “extent” is significantly associated with “status”.

-The R-squared value (0.02665) indicates that about 2.7% of the variation in “status” can be explained by “extent”. However, the adjusted R-squared value (-0.02612) suggests that the model may not be a good fit for the data, as the addition of the “extent” variable does not significantly improve the model’s predictive power.

-The F-statistic (50.81) tests the overall significance of the model and its variables, and the p-value (1.452e-12) suggests that the model is significant overall, indicating that at least one of the variables in the model is significantly associated with the dependent variable.

#regression between status and surg  
model5<-lm(status~surgery)  
summary(model5)

##   
## Call:  
## lm(formula = status ~ surgery)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.5607 -0.4714 -0.4714 0.5286 0.5286   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.47141 0.01350 34.912 < 2e-16 \*\*\*  
## surgery 0.08932 0.02619 3.411 0.000661 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4987 on 1856 degrees of freedom  
## Multiple R-squared: 0.00623, Adjusted R-squared: 0.005694   
## F-statistic: 11.63 on 1 and 1856 DF, p-value: 0.0006612

The first line shows the call to the lm() function that was used to fit the model.

The “Residuals” section shows the minimum, first quartile, median, third quartile, and maximum values of the residuals (observed minus predicted values).

The “Coefficients” section shows the estimated coefficients for the intercept and predictor variable(s), along with their standard errors, t-values, and p-values. The intercept is the expected value of the response when all predictor variables are equal to 0. The coefficients for the predictor variables represent the change in the response variable for a one-unit increase in the predictor variable, holding all other variables constant.

The “Signif. codes” section shows the significance level of the p-value for each coefficient. The significance levels are denoted by asterisks, with more asterisks indicating higher significance levels.

The “Residual standard error” represents the standard deviation of the residuals, which measures the average distance between the observed values and the predicted values.

The “Multiple R-squared” represents the proportion of the variability in the response variable that is explained by the predictor variable(s). The “Adjusted R-squared” is a modified version of the R-squared that adjusts for the number of predictor variables in the model.

The “F-statistic” tests the overall significance of the model by comparing the variability explained by the model to the variability not explained by the model. The “p-value” for the F-statistic represents the probability of obtaining an F-statistic as large or larger than the observed value if the null hypothesis (that all coefficients are equal to 0) is true.

#regression between status and node4  
model6<-lm(status~positive\_lymph\_nodes )  
summary(model6)

##   
## Call:  
## lm(formula = status ~ positive\_lymph\_nodes)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.7078 -0.4147 -0.4147 0.5853 0.5853   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.41469 0.01315 31.53 <2e-16 \*\*\*  
## positive\_lymph\_nodes 0.29315 0.02510 11.68 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4828 on 1856 degrees of freedom  
## Multiple R-squared: 0.06846, Adjusted R-squared: 0.06796   
## F-statistic: 136.4 on 1 and 1856 DF, p-value: < 2.2e-16

Residuals: These are the differences between the predicted values from the model and the actual observed values. The output shows the minimum, first quartile, median, third quartile, and maximum values of the residuals. Coefficients: These are the estimated coefficients of the linear regression model. The intercept (the value of the dependent variable when all independent variables are 0) is 0.41469. The coefficient for the independent variable node4 is 0.29315, which means that a one-unit increase in node4 is associated with a 0.29315 increase in the dependent variable. Signif. codes: These indicate the level of significance of the coefficients. The ’\*\*\*’ next to the intercept and node4 means that they are both statistically significant at the 0.001 level. Residual standard error: This is an estimate of the standard deviation of the residuals. Multiple R-squared: This is a measure of how well the model fits the data, ranging from 0 to 1. In this case, the multiple R-squared is 0.06846, which means that the model explains only 6.85% of the variance in the dependent variable. Adjusted R-squared: This is a modified version of the R-squared that penalizes the inclusion of additional independent variables in the model. In this case, the adjusted R-squared is slightly lower than the multiple R-squared, indicating that the model is not improved by the inclusion of additional independent variables. F-statistic: This is a test of the overall significance of the model. In this case, the F-statistic is 136.4 with 1 and 1856 degrees of freedom, which means that the model is statistically significant at a very low p-value (<2.2e-16).

#regression between status and perfor  
model7<-lm(status~perfor\_colon)  
summary(model7)

##   
## Call:  
## lm(formula = status ~ perfor\_colon)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.5926 -0.4922 -0.4922 0.5078 0.5078   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.49224 0.01177 41.818 <2e-16 \*\*\*  
## perfor\_colon 0.10035 0.06905 1.453 0.146   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5 on 1856 degrees of freedom  
## Multiple R-squared: 0.001137, Adjusted R-squared: 0.0005987   
## F-statistic: 2.112 on 1 and 1856 DF, p-value: 0.1463

The first output is the summary of a linear regression model that predicts “status” based on the “node4” variable. The output shows that the intercept and the “node4” variable are both statistically significant predictors of “status”. The estimated intercept is 0.41469, which represents the predicted “status” score when “node4” is equal to zero. The estimated coefficient for “node4” is 0.29315, which represents the change in the predicted “status” score for a one-unit increase in “node4”. The R-squared value of 0.06846 indicates that only a small proportion of the variability in “status” is explained by “node4”.

The second output is the summary of a linear regression model that predicts “status” based on the “perfor” variable. The output shows that the intercept is a statistically significant predictor of “status”, but the “perfor” variable is not. The estimated intercept is 0.49224, which represents the predicted “status” score when “perfor” is equal to zero. The estimated coefficient for “perfor” is 0.10035, which represents the change in the predicted “status” score for a one-unit increase in “perfor”. However, the p-value for “perfor” is 0.146, which is greater than the typical significance level of 0.05, indicating that “perfor” is not a statistically significant predictor of “status”. The R-squared value of 0.001137 indicates that only a very small proportion of the variability in “status” is explained by “perfor”.

#multivariate simple regression model  
model8<-lm(status~sex+age+differ\_tumor+extent+surgery+positive\_lymph\_nodes+perfor\_colon)   
summary(model8)

##   
## Call:  
## lm(formula = status ~ sex + age + differ\_tumor + extent + surgery +   
## positive\_lymph\_nodes + perfor\_colon)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.8995 -0.4079 -0.1380 0.4998 0.8779   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.0423828 0.0982694 -0.431 0.666308   
## sexmale -0.0021984 0.0224780 -0.098 0.922100   
## age 0.0002537 0.0009421 0.269 0.787757   
## differ\_tumor 0.0196047 0.0221796 0.884 0.376864   
## extent 0.1314289 0.0235340 5.585 2.7e-08 \*\*\*  
## surgery 0.0930230 0.0253166 3.674 0.000245 \*\*\*  
## positive\_lymph\_nodes 0.2773477 0.0255794 10.843 < 2e-16 \*\*\*  
## perfor\_colon 0.0839838 0.0661928 1.269 0.204685   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4775 on 1804 degrees of freedom  
## (46 observations deleted due to missingness)  
## Multiple R-squared: 0.09183, Adjusted R-squared: 0.08831   
## F-statistic: 26.06 on 7 and 1804 DF, p-value: < 2.2e-16

This is a multiple linear regression model that predicts the relationship between the status variable (outcome variable) and the predictor variables sex, age, differ, extent, surg, node4, and perfor.

The p-values associated with the coefficients estimate indicate whether each predictor variable is statistically significant in explaining the variance in the outcome variable. The null hypothesis is that the coefficient for a predictor variable is zero (meaning that the variable is not significantly related to the outcome variable). If the p-value is less than the significance level (typically 0.05), then we reject the null hypothesis and conclude that the variable is significantly related to the outcome variable.

From the results, we can see that the variables extent and surg are statistically significant (p < 0.001), while node4 is highly significant (p < 0.0001), and the other predictor variables are not significant (p > 0.05). The adjusted R-squared value of 0.08832 suggests that this model explains only a small proportion of the variance in the outcome variable. The residual standard error is 0.4775, which indicates the average distance between the observed data and the predicted values by the model.

# Multi-nomial models

df$status<-as.factor(df$status)  
df$sex<-as.factor(df$sex)  
df$stdiffer<-as.factor(df$differ)  
df$extent<-as.factor(df$extent)  
df$surg<-as.factor(df$surg)  
df$node4<-as.factor(df$node4)  
df$perfor<-as.factor(df$perfor)

#status and sex logistic regression   
model9<-multinom(status~sex,data=df)

## # weights: 3 (2 variable)  
## initial value 1287.867461   
## final value 1287.757040   
## converged

summary(model9)

## Call:  
## multinom(formula = status ~ sex, data = df)  
##   
## Coefficients:  
## Values Std. Err.  
## (Intercept) -0.008968697 0.06696563  
## sex1 -0.020018883 0.09287654  
##   
## Residual Deviance: 2575.514   
## AIC: 2579.514

The code is fitting a multinomial logistic regression model with “status” as the response variable and “sex” as the predictor variable.

The output shows the coefficients for the intercept and the predictor variable (“sex1” which represents the male sex). The intercept coefficient indicates the expected log odds of being in the reference category for the response variable (“status” in this case) when the predictor variable is 0 (in this case, when the sex is female, which is the reference category).

The coefficient for the predictor variable (“sex1”) represents the difference in the expected log odds of being in each of the categories of the response variable between males and females. In this case, the coefficient is negative, indicating that males have lower expected log odds of being in each of the categories compared to females.

The AIC (Akaike Information Criterion) and residual deviance are measures of model fit. The lower the values, the better the model fit.

#status and age logistic regression  
model10<-multinom(status~age,data=df)

## # weights: 3 (2 variable)  
## initial value 1287.867461   
## final value 1287.233127   
## converged

summary(model10)

## Call:  
## multinom(formula = status ~ age, data = df)  
##   
## Coefficients:  
## Values Std. Err.  
## (Intercept) 0.22356741 0.236921316  
## age -0.00406579 0.003888172  
##   
## Residual Deviance: 2574.466   
## AIC: 2578.466

The coefficients table shows the estimated values for the intercept and the slope for age. The intercept indicates the expected log-odds of the baseline level of the response variable (the reference level) when the age is equal to zero. In this case, the intercept is 0.2236, meaning that the log-odds of the baseline level of the response variable increases by 0.2236 for each unit increase in age.

The coefficient for age is -0.0041, indicating that for every one-unit increase in age, the log-odds of the baseline level of the response variable decreases by 0.0041. The standard error for the coefficient indicates how much the coefficient might vary if the study were replicated.

The model summary shows the residual deviance, which is a measure of the goodness of fit of the model, and the AIC, which is a measure of the model’s quality relative to other models. In this case, the residual deviance is 2574.466 and the AIC is 2578.466, suggesting that the model fits the data reasonably well but may not be the best possible model.

#status and differ logistic regression   
model11<-multinom(status~differ\_tumor)

## # weights: 3 (2 variable)  
## initial value 1255.982691   
## final value 1251.242690   
## converged

summary(model11)

## Call:  
## multinom(formula = status ~ differ\_tumor)  
##   
## Coefficients:  
## Values Std. Err.  
## (Intercept) -0.5955019 0.19613788  
## differ\_tumor 0.2811523 0.09228148  
##   
## Residual Deviance: 2502.485   
## AIC: 2506.485

The intercept represents the log-odds of being in the reference category of “status” (i.e., the category that was left out of the model) when “differ” is 0. The coefficient for “differ” represents the change in the log-odds of being in each category of “status” associated with a one-unit increase in “differ”. Specifically, for a one-unit increase in “differ”, the log-odds of being in the first category of “status” increase by 0.2811523, while the log-odds of being in the second category of “status” decrease by 0.2811523. The log-odds of being in the reference category of “status” do not change. The model has a residual deviance of 2502.485 and an AIC of 2506.485. The residual deviance measures how well the model fits the data (lower values are better), while the AIC is a measure of the model’s goodness of fit that takes into account the number of parameters in the model (lower values are better).

#status and extent logistic regression   
model12<-multinom(status~extent)

## # weights: 3 (2 variable)  
## initial value 1287.867461   
## final value 1262.101818   
## converged

summary(model12)

## Call:  
## multinom(formula = status ~ extent)  
##   
## Coefficients:  
## Values Std. Err.  
## (Intercept) -2.1037078 0.3115047  
## extent 0.7202024 0.1059637  
##   
## Residual Deviance: 2524.204   
## AIC: 2528.204

The model estimates the coefficients of three categories of ‘extent’ with reference to the baseline category, which is ‘extent1’. The coefficients represent the expected change in the log odds of being in each of the categories of ‘status’ for a one-unit increase in the corresponding extent category, holding all other variables constant.

For example, the coefficient of ‘extent2’ is 0.59, which indicates that a one-unit increase in the extent category from 1 to 2 is associated with an increase of 0.59 in the log odds of being in the ‘status’ category, compared to the baseline category, while holding all other variables constant.

The model has converged, and the residual deviance is 2524.003, while the AIC is 2532.003.

#status surg logistic regression   
model13<-multinom(status~surgery)

## # weights: 3 (2 variable)  
## initial value 1287.867461   
## final value 1281.983302   
## converged

summary(model13)

## Call:  
## multinom(formula = status ~ surgery)  
##   
## Coefficients:  
## Values Std. Err.  
## (Intercept) -0.1144946 0.0542418  
## surgery 0.3586142 0.1056436  
##   
## Residual Deviance: 2563.967   
## AIC: 2567.967

In this case, the dependent variable status has more than two categories, and there are two independent variables (surg). The output shows the estimated coefficients (or weights) for the intercept and each independent variable, along with their standard errors.

The coefficients for surg1 indicates how much the log odds of being in a particular category of status increase when surg is equal to 1 (versus 0). For example, the log odds of being in status=2 versus status=1 is estimated to increase by 0.3586 when surg=1.

The Residual Deviance is a measure of how well the model fits the data, with smaller values indicating a better fit. The AIC is a measure of the model’s overall goodness of fit, with lower values indicating a better fit.

#status node4 logistic regression   
model14<-multinom(status~positive\_lymph\_nodes)

## # weights: 3 (2 variable)  
## initial value 1287.867461   
## final value 1222.720934   
## converged

summary(model14)

## Call:  
## multinom(formula = status ~ positive\_lymph\_nodes)  
##   
## Coefficients:  
## Values Std. Err.  
## (Intercept) -0.3446227 0.05528417  
## positive\_lymph\_nodes 1.2295543 0.11197246  
##   
## Residual Deviance: 2445.442   
## AIC: 2449.442

This code is fitting a multinomial logistic regression model using the multinom() function in R. The response variable status is being modeled as a function of a single predictor variable node4.

The output shows the estimated coefficients of the model along with their standard errors. The intercept coefficient is -0.3446227, and the coefficient for node41 is 1.2295543. These coefficients can be used to predict the log odds of each level of the response variable given the value of the predictor variable.

The final value of the residual deviance is 2445.442, and the final value of the Akaike Information Criterion (AIC) is 2449.442. These values can be used to compare the fit of this model to other candidate models. A lower AIC indicates a better fit.

#status perfor logistic regression  
model15<-multinom(status~perfor\_colon)

## # weights: 3 (2 variable)  
## initial value 1287.867461   
## final value 1286.718866   
## converged

summary(model15)

## Call:  
## multinom(formula = status ~ perfor\_colon)  
##   
## Coefficients:  
## Values Std. Err.  
## (Intercept) -0.03104462 0.04709383  
## perfor\_colon 0.40573807 0.28093126  
##   
## Residual Deviance: 2573.438   
## AIC: 2577.438

The code is fitting a multinomial logistic regression model using the function multinom in the R programming language. The formula used to fit the model is status ~ perfor, which means that the variable perfor is used as the predictor for the outcome variable status, where status is a categorical variable with more than two levels.

The output shows the coefficients for the model, which are the estimates of the effects of the predictor variable on the outcome variable for each level of the outcome variable. The intercept coefficient is -0.031, which represents the estimated log-odds of the reference level of status, when perfor is equal to 0. The coefficient for perfor1 is 0.406, which represents the estimated difference in the log-odds of the level of status corresponding to perfor1 compared to the reference level, when perfor is equal to 1.

The output also shows the residual deviance, which is a measure of the goodness of fit of the model, and the AIC (Akaike Information Criterion), which is a measure of the model’s quality that takes into account the number of parameters in the model. In this case, the model seems to have converged and the AIC is 2577.438.

#multivariate logistic regression model  
model16<-multinom(status~sex+age+differ\_tumor+extent+surgery+positive\_lymph\_nodes+perfor\_colon)

## # weights: 9 (8 variable)  
## initial value 1255.982691   
## iter 10 value 1173.089593  
## final value 1169.235542   
## converged

summary(model16)

## Call:  
## multinom(formula = status ~ sex + age + differ\_tumor + extent +   
## surgery + positive\_lymph\_nodes + perfor\_colon)  
##   
## Coefficients:  
## Values Std. Err.  
## (Intercept) -2.448922364 0.457009583  
## sexmale -0.010903435 0.098813295  
## age 0.001275916 0.004151258  
## differ\_tumor 0.083748828 0.097677251  
## extent 0.604756458 0.111934685  
## surgery 0.407787261 0.111314291  
## positive\_lymph\_nodes 1.187705441 0.115911748  
## perfor\_colon 0.352260827 0.291178171  
##   
## Residual Deviance: 2338.471   
## AIC: 2354.471

Intercept): The intercept is the estimated log odds of the reference category of the dependent variable (status = 1) when all other predictors are equal to zero. The estimated intercept is -1.643, which means that the log odds of status = 1 is negative when all other predictors are zero.

sex1: The coefficient for sex1 is -0.016, which means that the log odds of status = 2 relative to status = 1 is expected to decrease by 0.016 when the person is female (sex1 = 1) compared to when they are male (sex1 = 0).

age: The coefficient for age is 0.0014, which means that the log odds of status = 2 relative to status = 1 is expected to increase by 0.0014 for each additional year of age.

differ: The coefficient for differ is 0.084, which means that the log odds of status = 2 relative to status = 1 is expected to increase by 0.084 for each additional point increase in the differentiation level.

extent2, extent3, and extent4: These are the coefficients for the extent variable, which has four categories (extent1 is the reference category). The coefficient for extent2 is 0.33, which means that the log odds of status = 2 relative to status = 1 is expected to increase by 0.33 when the extent is 2 compared to when it is 1. Similarly, the coefficients for extent3 and extent4 indicate that the log odds of status = 2 relative to status = 1 is expected to increase by 1.01 and 1.58, respectively, when the extent is 3 or 4 compared to when it is 1.

surg1: The coefficient for surg1 is 0.41, which means that the log odds of status = 2 relative to status = 1 is expected to increase by 0.41 when the person underwent surgery (surg1 = 1) compared to when they did not undergo surgery (surg1 = 0).

node41: The coefficient for node41 is 1.19, which means that the log odds of status = 3 relative to status = 1 is expected to increase by 1.19 when there is nodal involvement (node41 = 1) compared to when there is no nodal involvement (node41 = 0).

perfor1: The coefficient for perfor1 is 0.35, which means that the log odds of status = 2 relative to status = 1 is expected to increase by 0.35 when there is perforation (perfor1 = 1) compared to when there is no perforation (perfor1 = 0).

The residual deviance is 2337.94, which represents the unexplained variation in the model after accounting for the predictor variables. The AIC value is 2357.94, which is a measure of the model fit that takes into account the number of predictor variables. Lower AIC values indicate better model fit.

#independent T tests  
# perform independent t-test based on Sex  
ttest\_sex <- t.test(status ~ sex)  
print(ttest\_sex)

##   
## Welch Two Sample t-test  
##   
## data: status by sex  
## t = 0.21543, df = 1844.2, p-value = 0.8295  
## alternative hypothesis: true difference in means between group female and group male is not equal to 0  
## 95 percent confidence interval:  
## -0.04055387 0.05056232  
## sample estimates:  
## mean in group female mean in group male   
## 0.4977578 0.4927536

The t-value of 0.21543 indicates that the difference between the means of the two groups is not large relative to the amount of variability within each group. The degrees of freedom (df) is 1844.2, which is calculated using a modified formula to account for unequal variances between the two groups.

The p-value of 0.8295 indicates that there is not enough evidence to reject the null hypothesis, which is that there is no difference between the means of the two groups on the variable being measured. Specifically, the null hypothesis assumes that the true difference in means between the female and male groups is equal to 0. Therefore, we can conclude that there is no significant difference between the means of the two groups on the variable being measured.

The 95 percent confidence interval (CI) is [-0.04055387, 0.05056232], which indicates that we are 95 percent confident that the true difference in means between the two groups lies between these two values.

The sample estimates provide the mean for each group: 0.4977578 for females and 0.4927536 for males. This indicates that the mean for females is slightly higher than the mean for males, but the difference is not statistically significant.

``{r} install.packages(“rmarkdown”)

``{r}  
setwd("/HDS5310/R final Project")

``{r} rmarkdown::render(“Predictors\_of\_Treatment\_Related\_Regrets\_in\_Bowel\_Cancer\_Patients-1.Rmd”, output\_format = “word\_document”) ```