CASE STUDY: PREDICTORS OF CERVICAL CANCER RELAPSE IN ADULTS AFTER TREATMENT

Agenda

- Introduction
- Exploratory Data Analysis
- Model Selection
- Model Validation
- Discussion/Conclusion

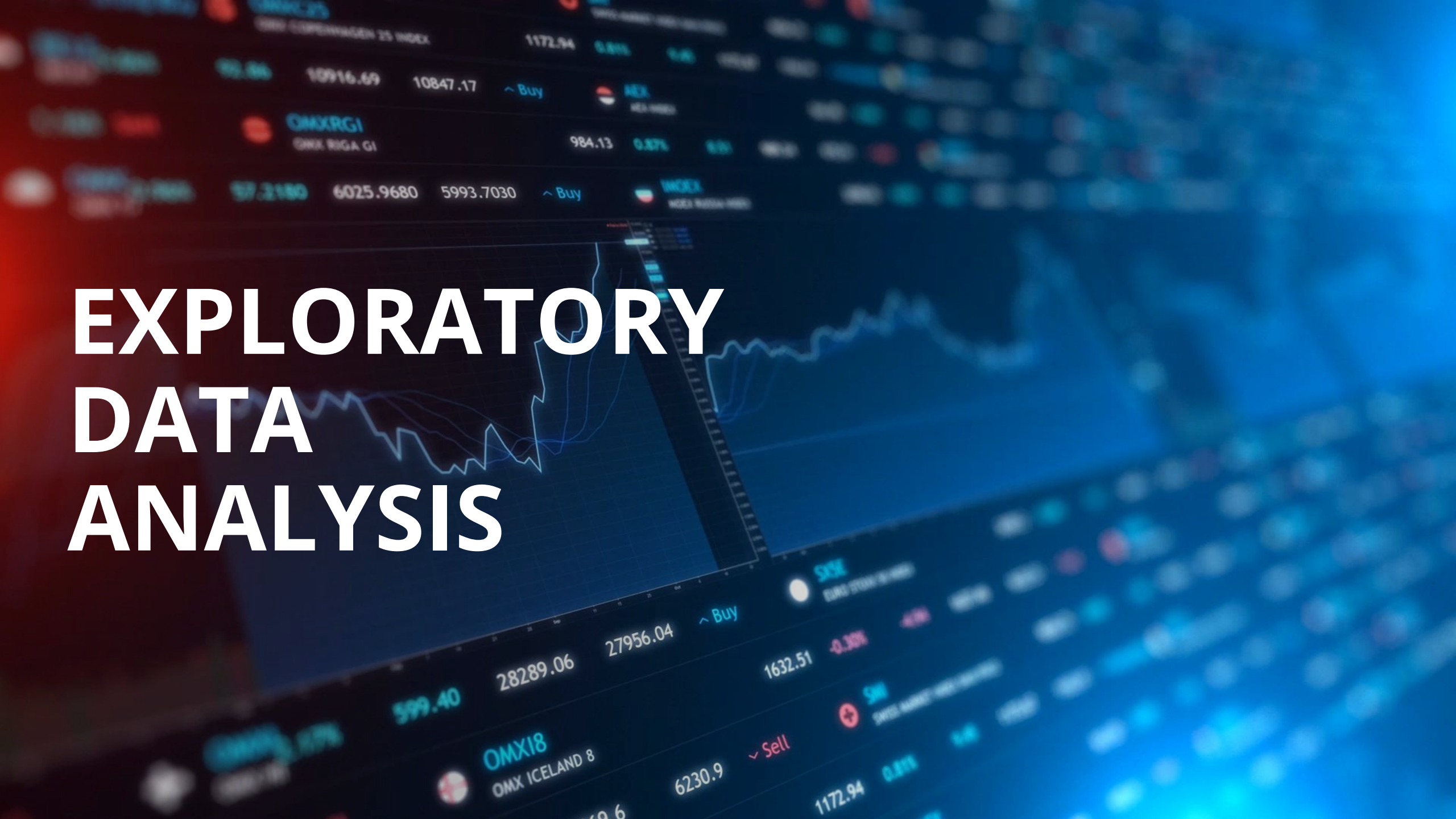
Introduction

- Background Research
- Response Variable:
  - Relapse
- Explanatory Variables:
  - Age
  - Radiation
  - disease status
  - cell differentiation
  - Depth
  - Size
  - pelvis involvement
  - CLS
- Goal of the Study:
  - Evaluate covariates such as the ones above to determine if they cause relapse in adults after treatment

Statistical Procedures

- Tidying Data
- Modifying Covariates
- Summarize Data
- Create Visualizations
- Variable Significance
- Multicollinearity Check
- Model Selection
- Model Validation:
  - H-L test
  - ROC Curve
  - Classification Table

EXPLORATORY DATA ANALYSIS



Tidying Data

1. Remove patient entries with no follow-up date (FU_DATE)
2. Remove patient entries that are not adults (ages 25 - 64)
3. Remove all patient entries who died of unrelated reasons or complications with no disease present
4. Remove all patient entries with missing values in any column
5. Create new data frame with desired covariates

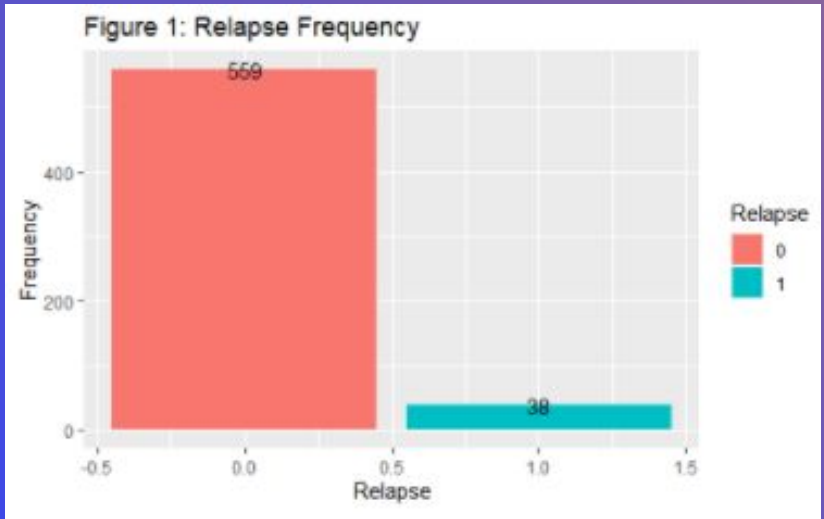
SUMMARY OF DATA

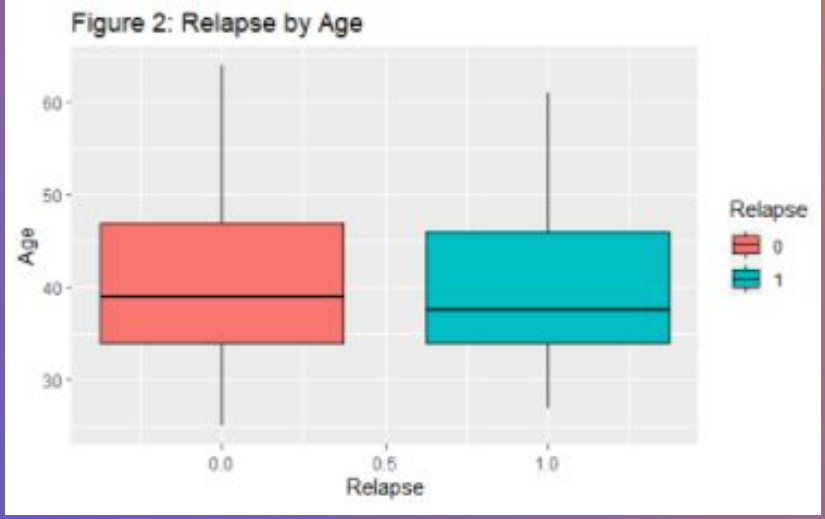
```
      size
 Min.   : 0.000
 1st Qu.: 0.000
 Median : 0.000
 Mean   : 7.462
 3rd Qu.:15.000
 Max.   :70.000
```

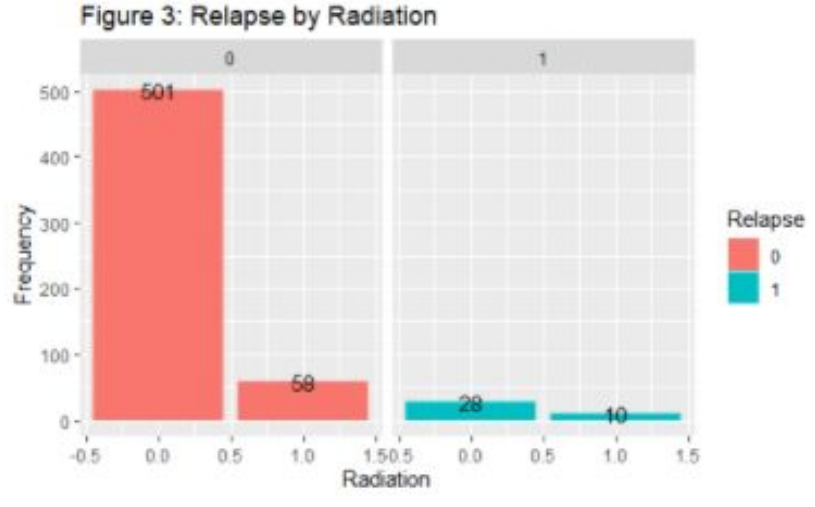
```
    maxDepth
 Min.   : 0.000
 1st Qu.: 3.000
 Median : 5.000
 Mean   : 7.304
 3rd Qu.:10.000
 Max.   :50.000
```

```
       age
 Min.   :25.00
 1st Qu.:34.00
 Median :39.00
 Mean   :40.85
 3rd Qu.:47.00
 Max.   :64.00
```

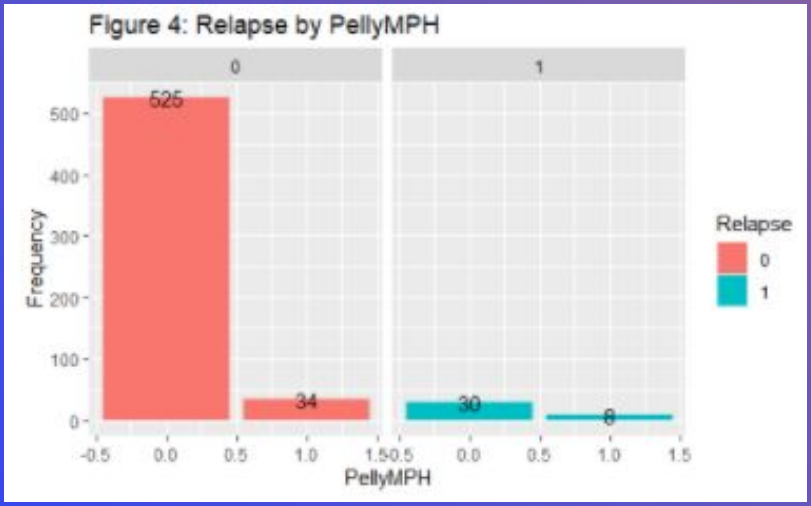
VISUALIZATION OF DATA

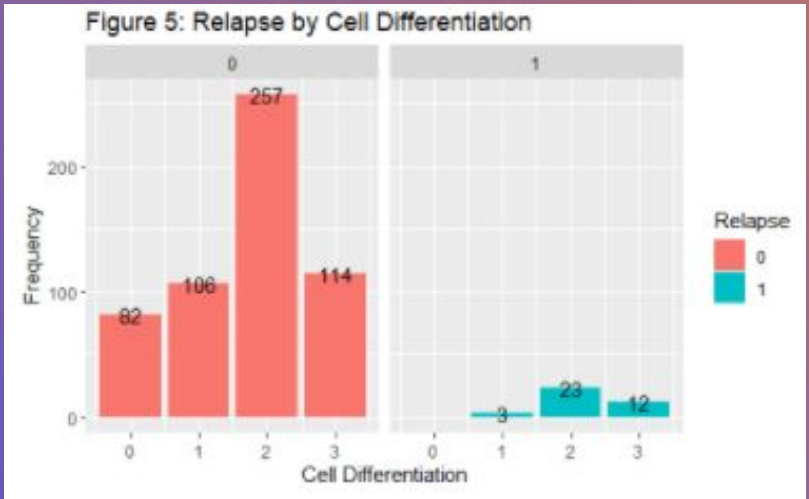


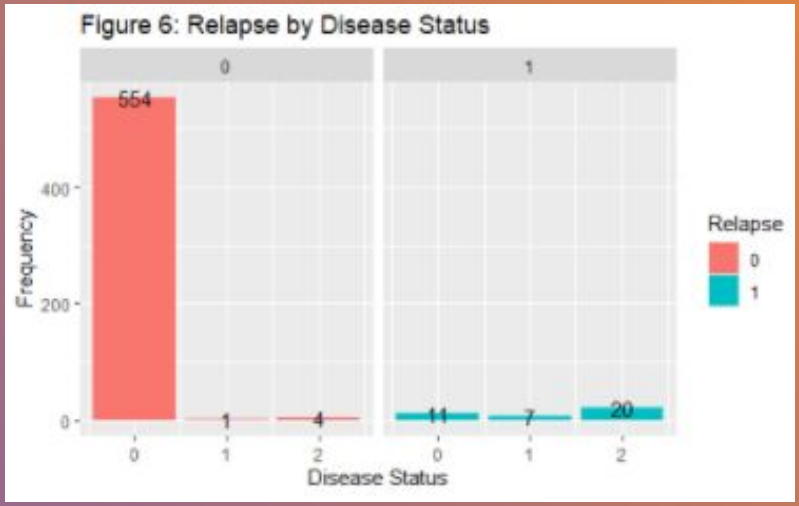




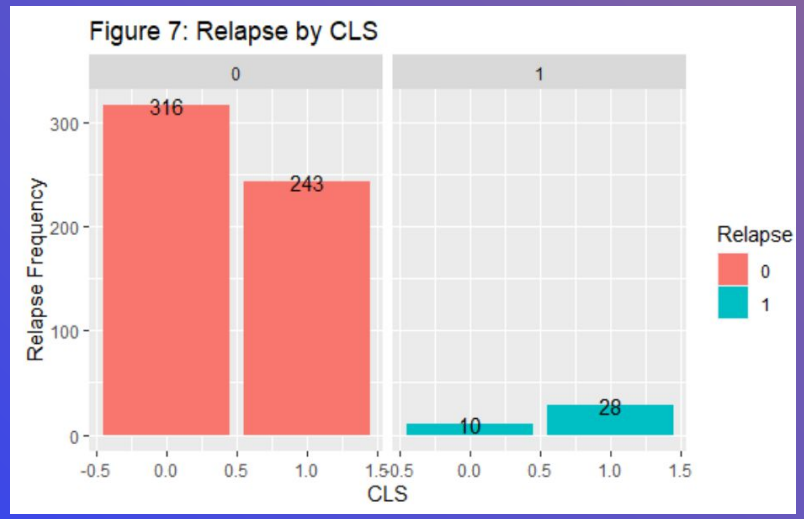
VISUALIZATION OF DATA

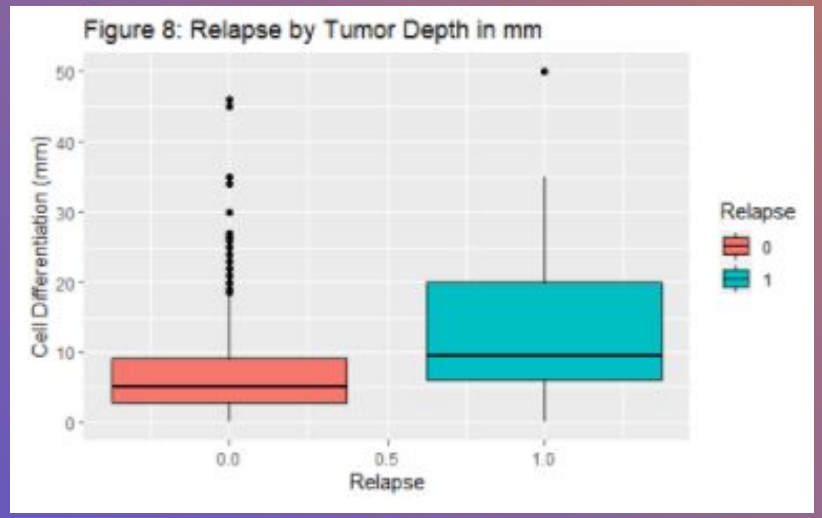


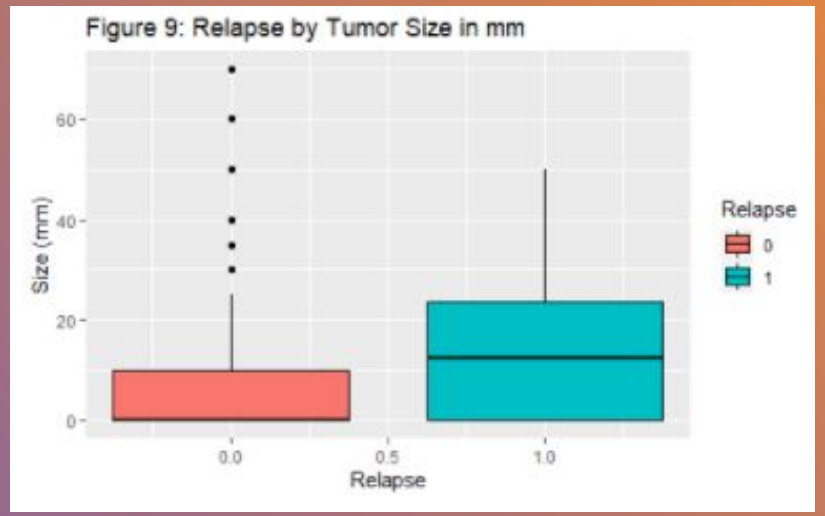




VISUALIZATION OF DATA







VARIABLE SIGNIFICANCE

- **Used vglm() to make a simple model of covariates and used LRT to check the p-values.**

Variables	P-Values	Significance
Age	**0.09584**	**< α = 0.1, fair predictor**
disSTA	**< 2.2 x 10^-16**	**< α = 0.001, very good predictor**
maxDepth	**0.0178**	**< α = 0.01, good predictor**
size	0.8337	< α = 1, poor predictor
radiation	0.842	< α = 1, poor predictor
pellyMPH	0.2579	< α, poor predictor
grad	0.4044	< α, poor predictor

MULTICOLLINEARITY

	age	maxDepth	size
age	1	0.1186659	0.01536332
maxDepth	0.1186659	1	0.38520779
Size	0.01536332	0.38520779	1

- No strong correlation (> 0.80) between any of these numerical covariates

MODEL SELECTION

MODEL SELECTION

- To predict the strength and significance
- Binomial GLM with response variable relapse
- Interaction variable between size and the depth of the tumor
- disSTA, maxDepth, size, and maxDepth:size are significant

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -4.006846   1.380508  -2.902 0.003703 **
age           -0.043369   0.029511  -1.470 0.141677
radiation      0.124326   0.765185   0.162 0.870929
pellyMPH      -1.305302   1.111585  -1.174 0.240286
disSTA         3.390985   0.444527   7.628 2.38e-14 ***
grad           0.114930   0.344657   0.333 0.738787
maxDepth       0.151124   0.039923   3.785 0.000153 ***
size           0.078939   0.035519   2.222 0.026255 *
cls            0.343449   0.562006   0.611 0.541125
maxDepth:size -0.005601   0.002369  -2.365 0.018045 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

General linear-regression model

Model Selection

```
Step:  AIC=139.01
relapse ~ age + disSTA + maxDepth + size + maxDepth:size

               Df Deviance    AIC     LRT  Pr(>Chi)
<none>              127.00 139.00
- age           1   129.18 139.18   2.173  0.140461
- maxDepth:size 1   133.86 143.86   6.851  0.008858 **
- disSTA        1   256.16 266.16 129.152 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Call:  glm(formula = relapse ~ age + disSTA + maxDepth + size + maxDepth:size,
    family = binomial(), data = coVariates)

Coefficients:
  (Intercept)           age         disSTA       maxDepth           size  maxDepth:size
    -3.661985      -0.041491       3.356980       0.144841       0.077010      -0.005459

Degrees of Freedom: 596 Total (i.e. Null);  591 Residual
Null Deviance:      282.9
Residual Deviance: 127  AIC: 139
```

Step-backward selection model

- **Running the step( ) function**
- **Determine which has lowest AIC value**
- **AIC: 139.01**
- **Null Deviance: 282.9**
- **Residual Deviance: 127**

```
Start:  AIC=145.02
relapse ~ age + radiation + pellyMPH + disSTA + grad + maxDepth +
    size * maxDepth + cls


Call:  glm(formula = relapse ~ age + radiation + pellyMPH + disSTA +
    grad + maxDepth + size * maxDepth + cls, family = binomial(),
    data = coVariates)

Coefficients:
  (Intercept)           age      radiation       pellyMPH         disSTA           grad       maxDepth
    -4.006846      -0.043369       0.124326      -1.305302       3.390985       0.114930       0.151124
         size           cls  maxDepth:size
     0.078939       0.343449      -0.005601

Degrees of Freedom: 596 Total (i.e. Null);  587 Residual
Null Deviance:      282.9
Residual Deviance: 125  AIC: 145
```

Step-forward selection model

Model Selection

- **Running the step( ) function**
- **Determine which has lowest AIC value**
- **AIC: 145.02**
- **Null Deviance: 282.9**
- **Residual Deviance: 125**

Model Comparison

Backwards step model:

AIC: 139.01
Model: $y = -3.661985 - 0.041491age + 3.356980disSta + 0.144841maxDepth + 0.077010size - 0.005459maxDepth:size$

Forwards step model:

AIC: 145.02
Model: $y = -4.006846 - 0.043369age + 0.124326radiation - 1.305302pellyMPH + 3.390985disSta + 0.114930grad + 0.151124maxDepth + 0.078939size + 0.343449cls - 0.00560maxDepth:size$

MODEL VALIDATION

Hosmer and Lemeshow Goodness of Fit Test

- **Model with lowest AIC must be validated**
- **Ungrouped data, hence we use Hosmer-Lemeshow test**
- **P-Value = 0.2768 > α = 0.05, fail to reject null**
- **Indicates a good fit model**

```
    Hosmer and Lemeshow goodness of fit (GOF) test

data:  mod.2$y, fitted(mod.2)
X-squared = 5.1035, df = 4, p-value = 0.2768
```

Classification Table

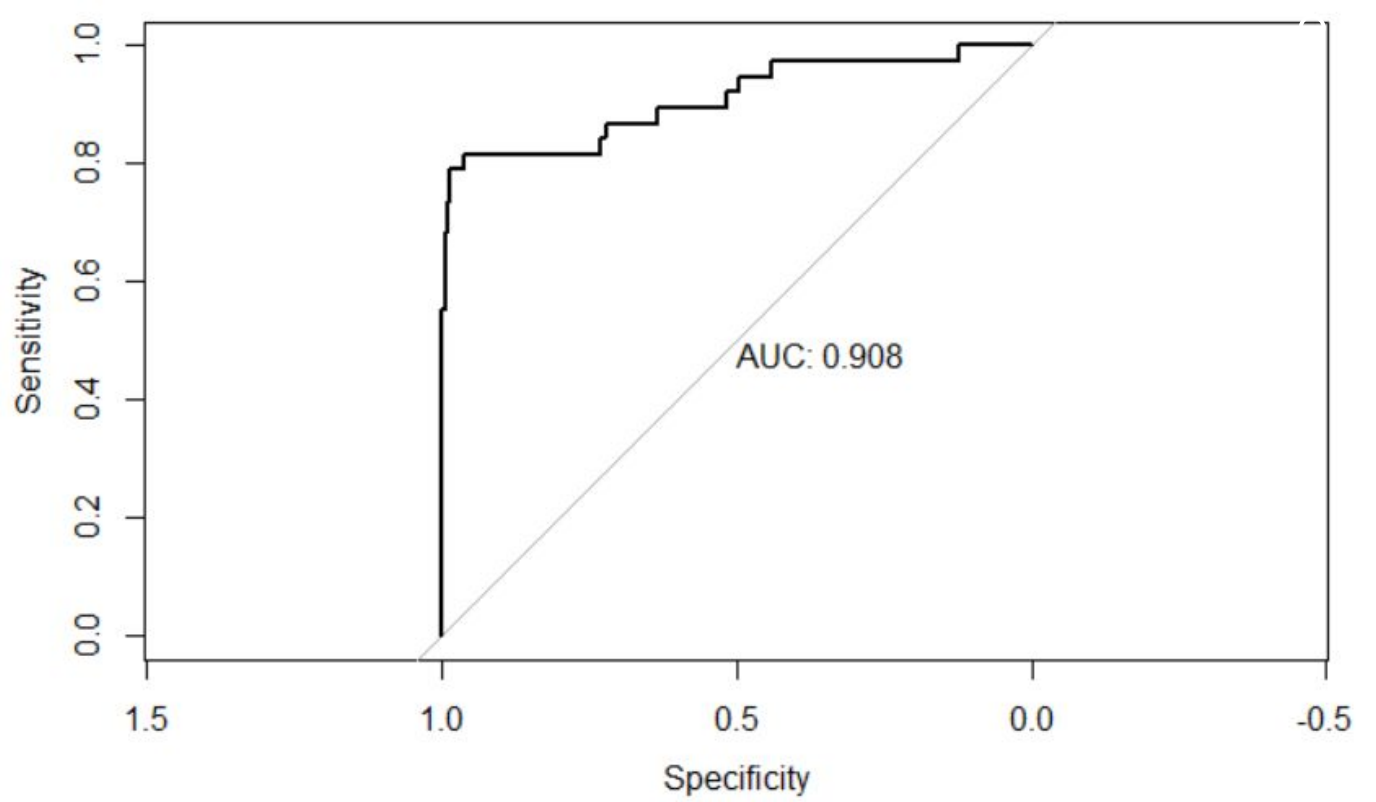
Actual/Predicted	Y=0	Y=1
y = 0	555	4
y = 1	16	22

- **Sensitivity = P(Y = 1 | y = 1) = 0.5789**
- **Specificity = P(Y = 0 | y = 0) = 0.9928**
- **Concordance Rate = (a + d)/n = 0.9665**

- **Shows us # of successes predicted by the model**
- **We have 4x4 matrix**
- **Focus on 2 outcomes**
- **Sensitivity (True Positive) and Specificity (True Negative)**

ROC CURVE

- **Estimates probability that both the predictions and the outcome are concordant**
- **Bigger area under ROC = better the model**
- **Area under ROC is 0.908 = 90.8%**
- **0.908 is very close to 1**
- **Excellent model**



Receiver Operating Characteristic (ROC) Curve

CONCLUSION

Summary of Findings

- **Goal**: Determine which covariates cause relapse in adults after treatment
- **Final Model**: includes age, disSTA, maxDepth, and maxDepth:size
- **Validation**:

AIC	**139.01**
Hosmer and Lemeshow p-value	0.28
Sensitivity	0.58
Specificity	0.99
ROC (AUC)	90.8%

Limitations

- A lot of incomplete and missing data entries
- Different methods of analysis could lead to different results
  - Grouping data by relapse and checking importance before splitting could result in different/better model
  - Using classification trees to expose structure of data which could provide a better way to visualize impact of variables

Further Research

- Expand this research to find new methods to detect cervical cancer in early stages
- Other factors can be analyzed such as:
  - Smoking history
  - Number of full-term pregnancies
  - Usage of contraceptives
- Expand this research by using such factors to help distinguish between similar conditions such as Prolapsed Uterine Fibroid (PUB) which has similar symptoms of abnormal bleeding and large mass

All models are wrong, but some are useful.

— George E. P. Box —