SURVIVAL RATE PREDICTION IN BREAST CANCER

BY: K. POOLADVAND

ABSTRACT

Breast cancer, the most commonly diagnosed cancer among women in the USA, demands accurate survival rate estimation. This critical information plays a pivotal role in guiding patients, their families, and healthcare professionals toward informed decisions, financial planning, and personalized treatment strategies. Our project involved meticulous curation of data from over 300,000 patients (spanning 2011-2015) and rigorous exploratory analyses.

During this process, we conducted correlation, Fisher, and Chi-square analyses to identify the most crucial parameters. From the original 36 features extracted using SEER *STAT software, we pinpointed 16 essential variables. The data underwent thorough examination and preparation, including specialized encoding techniques such as one-hot. Dealing with large databases, handling missing values, and selecting the right tools were some of the challenges we encountered and valuable lessons we learned.

Subsequently, we divided the cleaned and prepared dataset and applied machine learning techniques, including random forest, logistic regression, and deep neural network. Our approach addressed complexities related to both categorical and numerical variables, as well as handling missing data. Random Forest demonstrated better accuracy and acceptable speed in predicting breast cancer patient survival rates.

While our model is not intended for clinical use and lacks the rigor of a scientific investigation, it showcases remarkable success in estimating survival rates based on the training data. Furthermore, we extended its application to previously unseen databases from 2019-2022, comprising over 133,000 cases, allowing us to compare survival predictions against the next available dataset in the SEER program.

PROBLEM STATEMENT: PREDICTING BREAST CANCER SURVIVAL RATES

Leveraging the SEER database (2011-2015), our mission is to train a machine learning model with ambitious objectives:

- Exceptional Accuracy: Our aim is not only to surpass the 75% survival rate benchmark but also achieve accuracy levels exceeding 96%.
- Critical Parameter Exploration: We will delve into 16 influential factors (such as race and cancer type) that impact survival. These parameters will be instrumental in training our model effectively.

This project serves as a vital link between data science and patient care, empowering informed decisions.

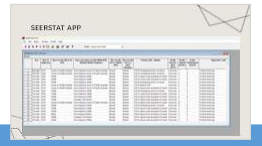
Our approach involves statistical analyses to identify crucial parameters and evaluate their dependencies.

INTRO – WHAT & WHY

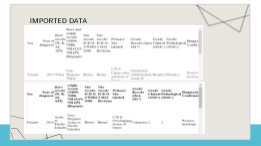
- 670,000 have lost their life to breast cancers globally in 2022 (43,700 in USA)
- Breast cancer was the most common cancer in women in 157 countries out of 185 in 2022
- Breast cancer is the most diagnosed cancer among U.S. women.
- It is also personal to me.

- Accurate survival prediction is critical for patient management
- Precise survival estimates help allocate healthcare resources effectively
- Patients and their families can make informed decisions about treatment options.
- Quality of Life Considerations
- And Economic consideration.

HOW - FLOWCHART

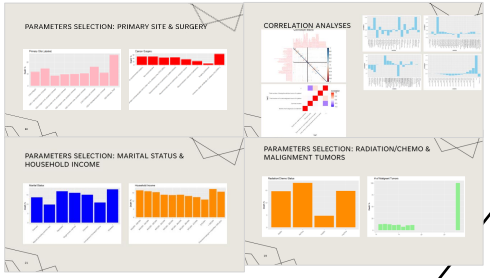


Choose the relevant parameters



Import datasets from SEER

Perform EDA for the datasets (two 2011-2015 and 2019-2022)



Find Statistically critical parameters among (16)

Data preparation

Modeling

Evaluate

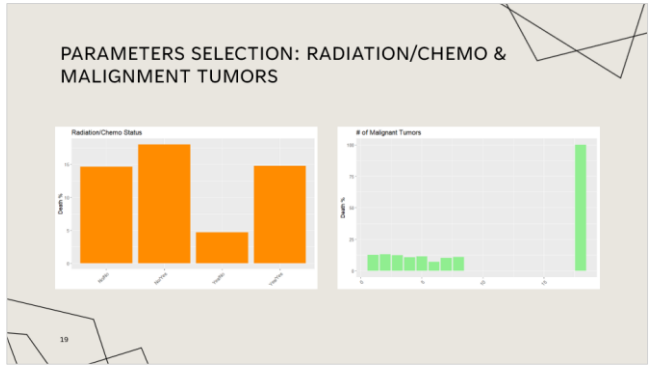
DATA DISPLAY AND PRELIMINARY ANALYSES

- **Survival rate 2011-2015 of 300K: 75%**

COD	count	Total Count	Population
Alive	228221	303557	75
Breast	38472	303557	13
Other	36864	303557	12

- **Marital status and survival**

Marital status at diagnosis	Event Population	Population	Group % in total	Death %
Divorced	4399	32214	10.61	13.66
Married (including common law)	15694	160551	52.89	9.78
Separated	544	3225	1.06	16.87
Single (never married)	7161	44678	14.72	16.03
Unknown	2774	18481	6.09	15.01
Unmarried or Domestic Partner	110	1014	0.33	10.85
Widowed	7790	43394	14.30	17.95



Principal Parameter selection

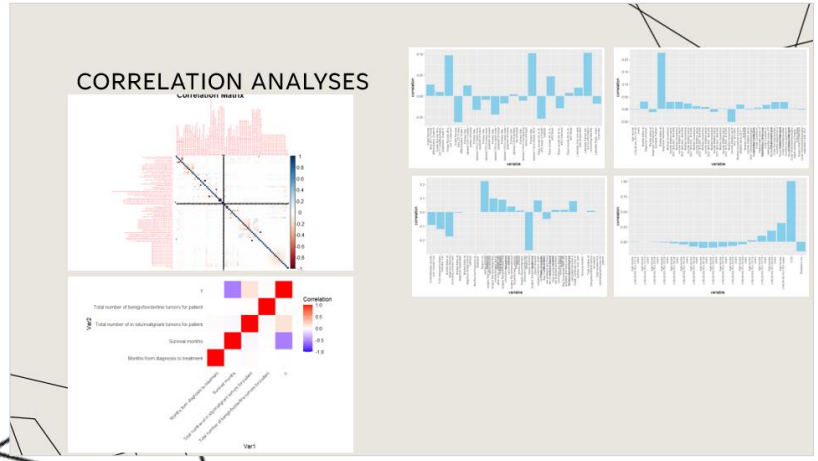
- **Evaluation the data, empty, correlation analyses, normality analyses,**
- **Among 36 parameters 16 found critical: treatment (chemo/radiation, surgery), # tumors, months from diagnosis to treatment, race, grade, and type**
- **Data were analyzed using bootstrapping and random sampling to avoid the size using Chis-square and Fisher exact test**
- **Many collinearities were identified and removed like sex, Origin Recode, Diagnostic confirmation, ...**

variable	p_value
Race recode (W, B, AI, API)	0.0004998
Primary Site - labeled	0.0004998
Grade Recode (thru 2017)	0.0004998
Laterality	0.0004998
Chemotherapy recode (yes, no/unk)	0.0004998
Reason no cancer-directed surgery	0.0004998
Survival months flag	0.0004998
First malignant primary indicator	0.0004998
Marital status at diagnosis	0.0004998
Median household income inflation adj to 2021	0.0004998
Age recode (<60,60-69,70+)	0.0004998
COD	0.0004998
Radiation	0.0004998
Rural-Urban Continuum Code	0.0011294

Fisher Exact with simulation .p.value=TRUE due to size of the sample!

HYPOTHESIS/CORRELATION EVALUATION

- **Summary**
  - **Many collinearity have found and remove from the data**
  - **Stepwise correlation analyses was performed to removed unimportant and poorly correlated parameters such as**
  - **Both numerical and categorical data were analyzed for correlation analyses**
  - **Dealing with categorical data was challenging specifically encoding the data for analyses**

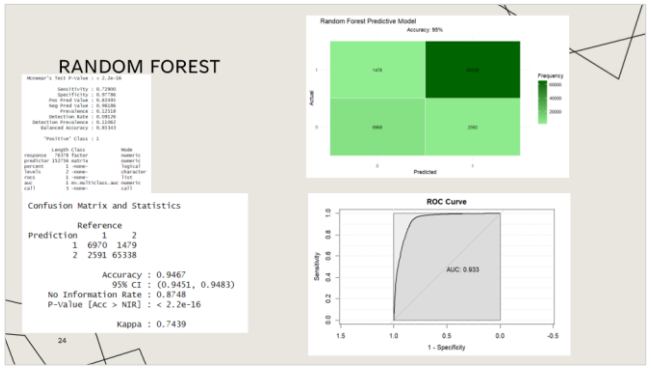


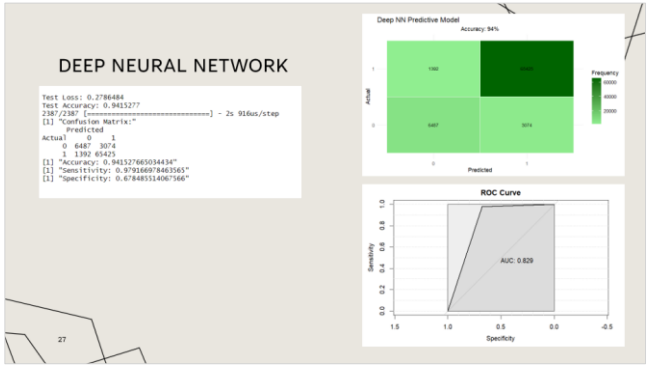
Unimportant (Total of 16)	Important (total of 16)
Different race related indicators were in the data base (all removed expect one)	(+)Months from diagnosis to treatment
Sex	(-) Total number of in situ/malignant tumors for patient
Patient ID	(-) Total number of benign/borderline tumors for patient
Year of death (the survival month exists)	Age
Sequence number	Race (W, B, AI, API)
Year of diagnosis	Primary Label
Rural urban continuum	Grade
	Treatment (Radiation/Surgery/Chemo)
	Income
	Marital Status

DATA ANALYSES, EVALUATION, & SUMMARY

- **Summary**
  - **Random Forest was the easiest to apply and reasonably fast**
  - **Logistic Regression was the fast, but still accurate**
  - **The accuracy of 94% and 93% for RF and Logistic Regression were achieved.**
  - **Logistic regression was easier to apply compared to Random Forrest.**

Method	Accuracy (Survival)
Logistic Regression	93%
Random Forest	94%
Deep Neural Network	94%



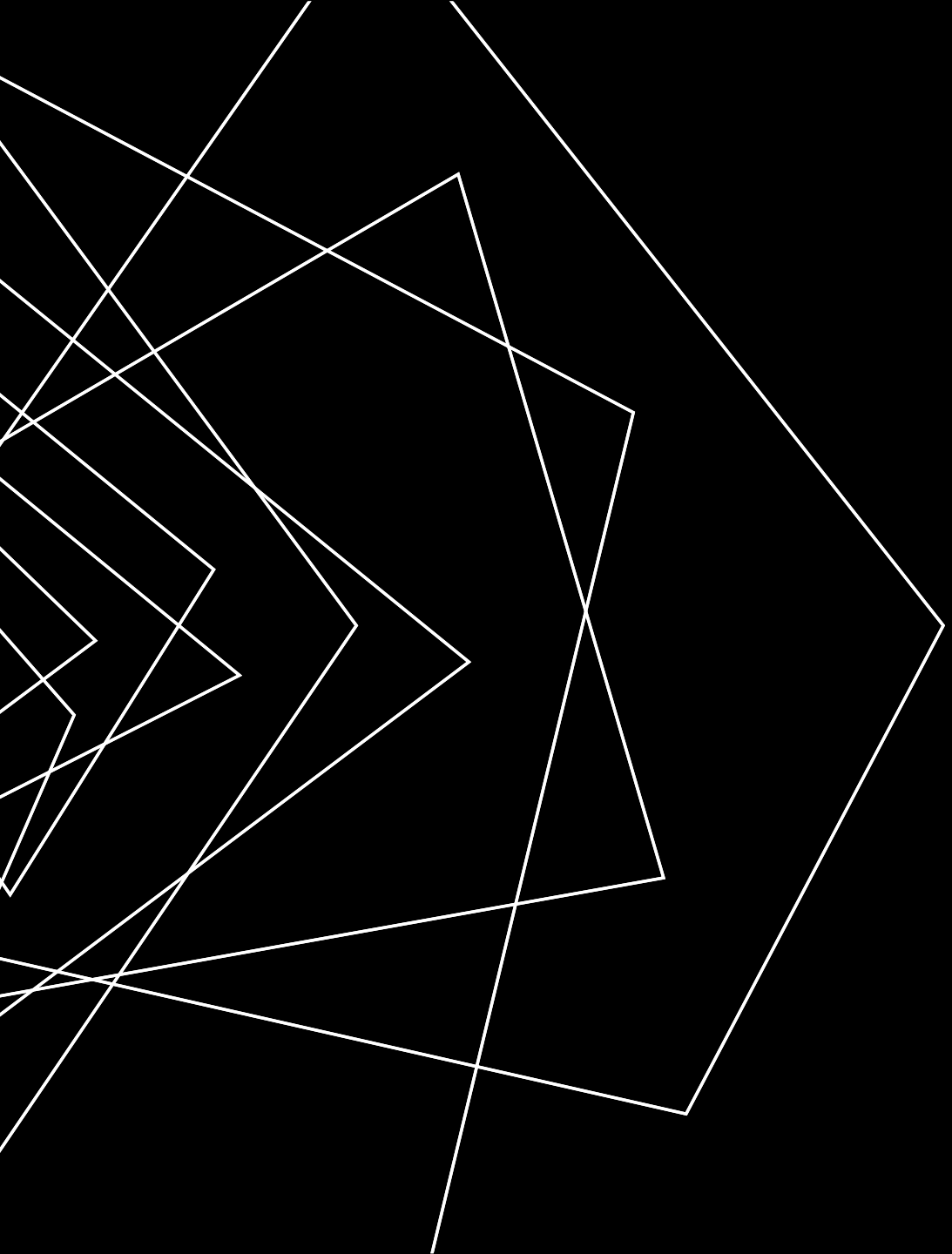


CONCLUSION

- A combined categorical and numerical data collected by SEER can be used for survival prediction of breast cancer with acceptable accuracy of 94%.
- Random Forest Simplicity: Opt for user-friendly models like Random Forest.
- Logistic regression is a powerful and reliable tool for a quick and relatively accurate survival rate estimate.
- Correlation analyses of the large database including categorical data is challenging.
- For this sort of problem Neural Network did not offer a better result than random Forrest.

REFERENCES:

- [1] SEER (https://seer.cancer.gov/data/access.html)
- [2] zgalochkina/SEER_solid_tumor: R code for SEER data analysis of solid tumor in different populations (github.com)
- [3] XAI_Healthcare_eXplainable_AI_in_Healthcare.pdf (upc.edu)
- [4] Pargen, F., Pfisterer, F., Thomas, J., Bischl, B.: Regularized target encoding out performs traditional methods in supervised machine learning with high cardinality features. Computational Statistics 37(5), 2671–2692 (Nov 2022)
- [5] American Cancer Society - Breast Cancer Survival Rates
- GitHub: koohpi/DATA606_Fianl_Project: DATA606 Final Project (github.com)
- Rpub: RPubs - DATA606 - Breast Cancer Survival Rate Estimate



THANK YOU

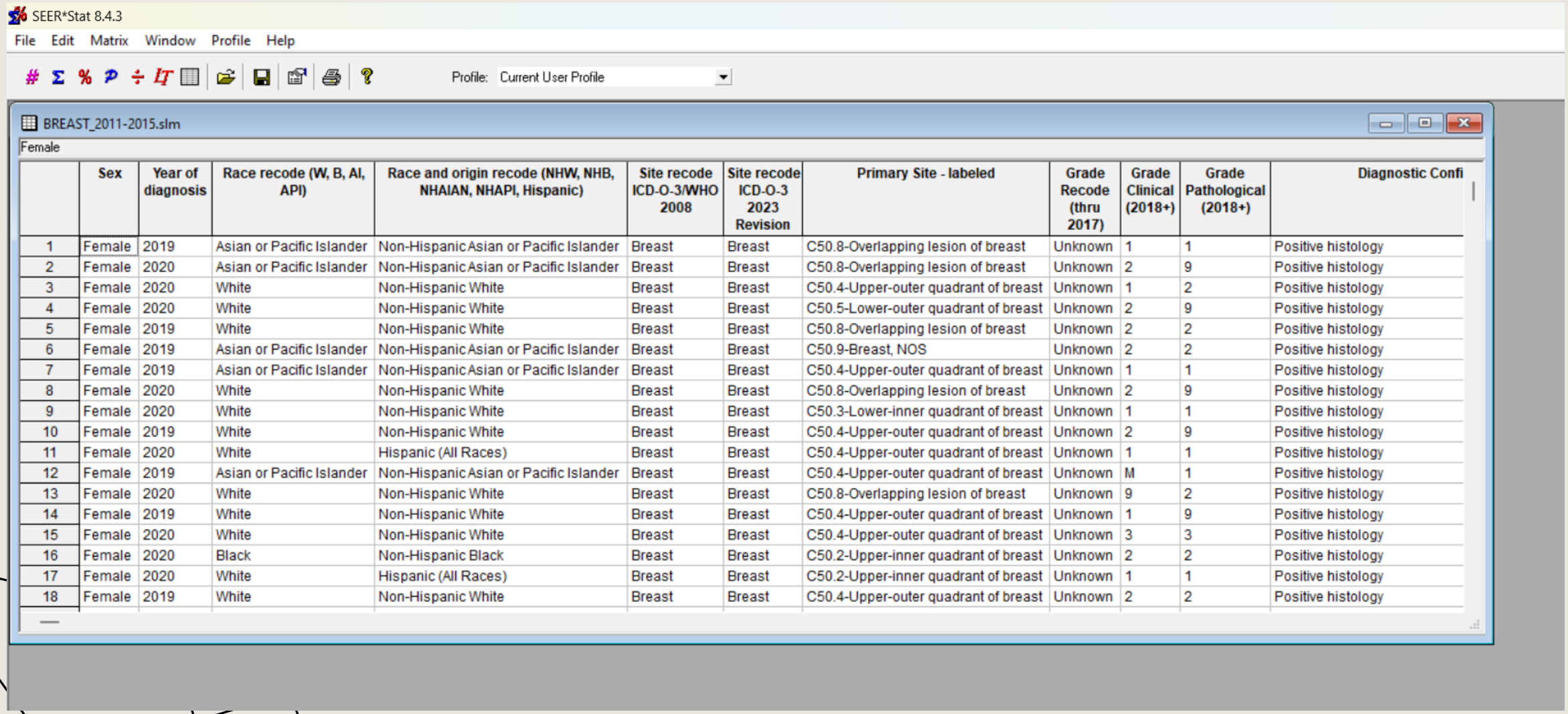
Koohyar Pooladvand
koohyar.pooladvand69@cuny.spsmail.edu
?

IMPORTED DATA

Sex	Year of diagnosis	Race recode (W, B, AI, API)	Race and origin recode (NHW, NHB, NHAIAN, NHAPI, Hispanic)	Site recode ICD-O-3/WHO 2008	Site recode ICD-O-3 2023 Revision	Primary Site - labeled	Grade Recode (thru 2017)	Grade Clinical (2018+)	Grade Pathological (2018+)	Diagno Confir
Female	2015	White	Non-Hispanic White	Breast	Breast	C50.4-Upper-outer quadrant of breast	Moderately differentiated; Grade II	Blank(s)	Blank(s)	Positive histolog

Sex	Year of diagnosis	Race recode (W, B, AI, API)	origin recode (NHW, NHB, NHAIAN, NHAPI, Hispanic)	Site recode ICD-O-3/WHO 2008	Site recode ICD-O-3 2023 Revision	Primary Site - labeled	Grade Recode (thru 2017)	Grade Clinical (2018+)	Grade Pathological (2018+)	Diagnostic Confirmat
Female	2019	Asian or Pacific Islander	Non-Hispanic Asian or Pacific Islander	Breast	Breast	C50.8-Overlapping lesion of breast	Unknown	1	1	Positive histology

SEERSTAT APP

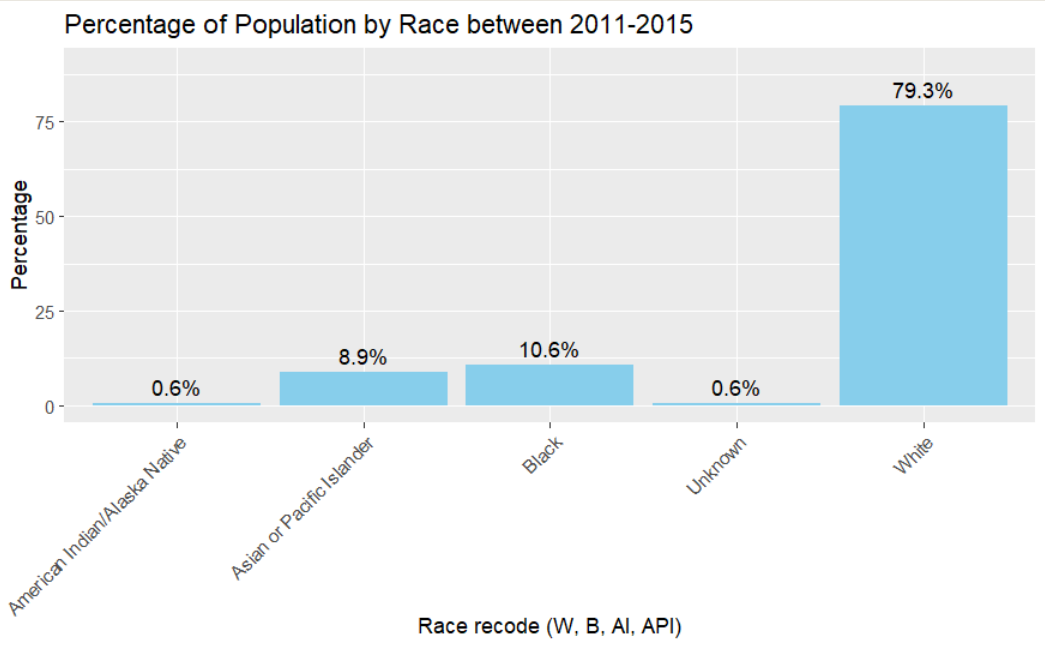


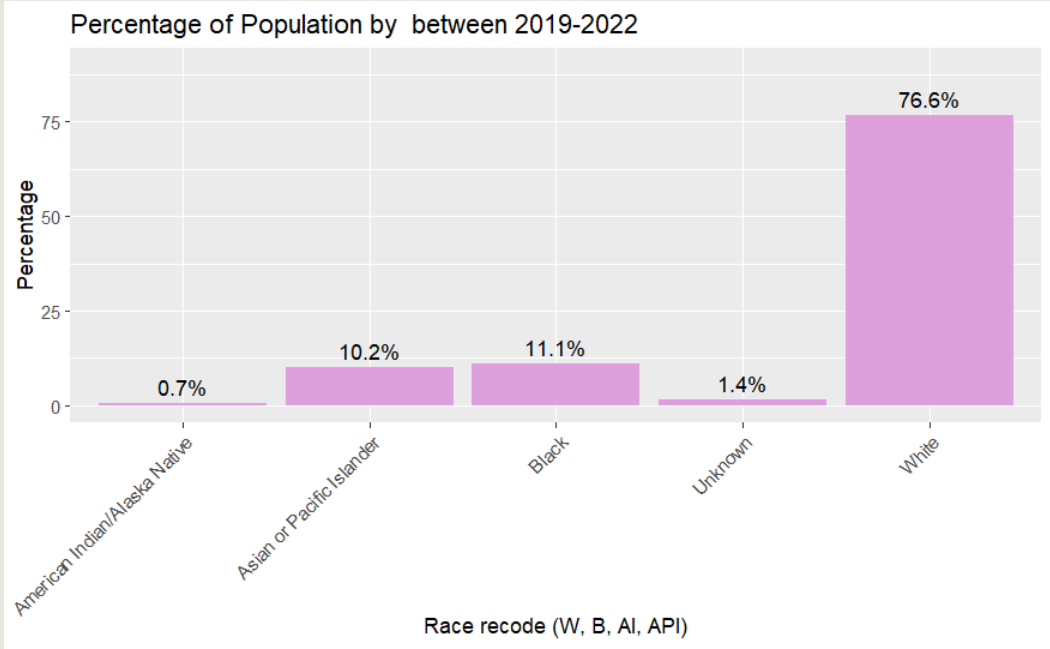
	Sex	Year of diagnosis	Race recode (W, B, AI, API)	Race and origin recode (NHW, NHB, NHAIAN, NHAPI, Hispanic)	Site recode ICD-O-3/WHO 2008	Site recode ICD-O-3 2023 Revision	Primary Site - labeled	Grade Recode (thru 2017)	Grade Clinical (2018+)	Grade Pathological (2018+)	Diagnostic Confi
1	Female	2019	Asian or Pacific Islander	Non-Hispanic Asian or Pacific Islander	Breast	Breast	C50.8-Overlapping lesion of breast	Unknown	1	1	Positive histology
2	Female	2020	Asian or Pacific Islander	Non-Hispanic Asian or Pacific Islander	Breast	Breast	C50.8-Overlapping lesion of breast	Unknown	2	9	Positive histology
3	Female	2020	White	Non-Hispanic White	Breast	Breast	C50.4-Upper-outer quadrant of breast	Unknown	1	2	Positive histology
4	Female	2020	White	Non-Hispanic White	Breast	Breast	C50.5-Lower-outer quadrant of breast	Unknown	2	9	Positive histology
5	Female	2019	White	Non-Hispanic White	Breast	Breast	C50.8-Overlapping lesion of breast	Unknown	2	2	Positive histology
6	Female	2019	Asian or Pacific Islander	Non-Hispanic Asian or Pacific Islander	Breast	Breast	C50.9-Breast, NOS	Unknown	2	2	Positive histology
7	Female	2019	Asian or Pacific Islander	Non-Hispanic Asian or Pacific Islander	Breast	Breast	C50.4-Upper-outer quadrant of breast	Unknown	1	1	Positive histology
8	Female	2020	White	Non-Hispanic White	Breast	Breast	C50.8-Overlapping lesion of breast	Unknown	2	9	Positive histology
9	Female	2020	White	Non-Hispanic White	Breast	Breast	C50.3-Lower-inner quadrant of breast	Unknown	1	1	Positive histology
10	Female	2019	White	Non-Hispanic White	Breast	Breast	C50.4-Upper-outer quadrant of breast	Unknown	2	9	Positive histology
11	Female	2020	White	Hispanic (All Races)	Breast	Breast	C50.4-Upper-outer quadrant of breast	Unknown	1	1	Positive histology
12	Female	2019	Asian or Pacific Islander	Non-Hispanic Asian or Pacific Islander	Breast	Breast	C50.4-Upper-outer quadrant of breast	Unknown	M	1	Positive histology
13	Female	2020	White	Non-Hispanic White	Breast	Breast	C50.8-Overlapping lesion of breast	Unknown	9	2	Positive histology
14	Female	2019	White	Non-Hispanic White	Breast	Breast	C50.4-Upper-outer quadrant of breast	Unknown	1	9	Positive histology
15	Female	2020	White	Non-Hispanic White	Breast	Breast	C50.4-Upper-outer quadrant of breast	Unknown	3	3	Positive histology
16	Female	2020	Black	Non-Hispanic Black	Breast	Breast	C50.2-Upper-inner quadrant of breast	Unknown	2	2	Positive histology
17	Female	2020	White	Hispanic (All Races)	Breast	Breast	C50.2-Upper-inner quadrant of breast	Unknown	1	1	Positive histology
18	Female	2019	White	Non-Hispanic White	Breast	Breast	C50.4-Upper-outer quadrant of breast	Unknown	2	2	Positive histology

CLEANED/TIDY DATA SUMMARY

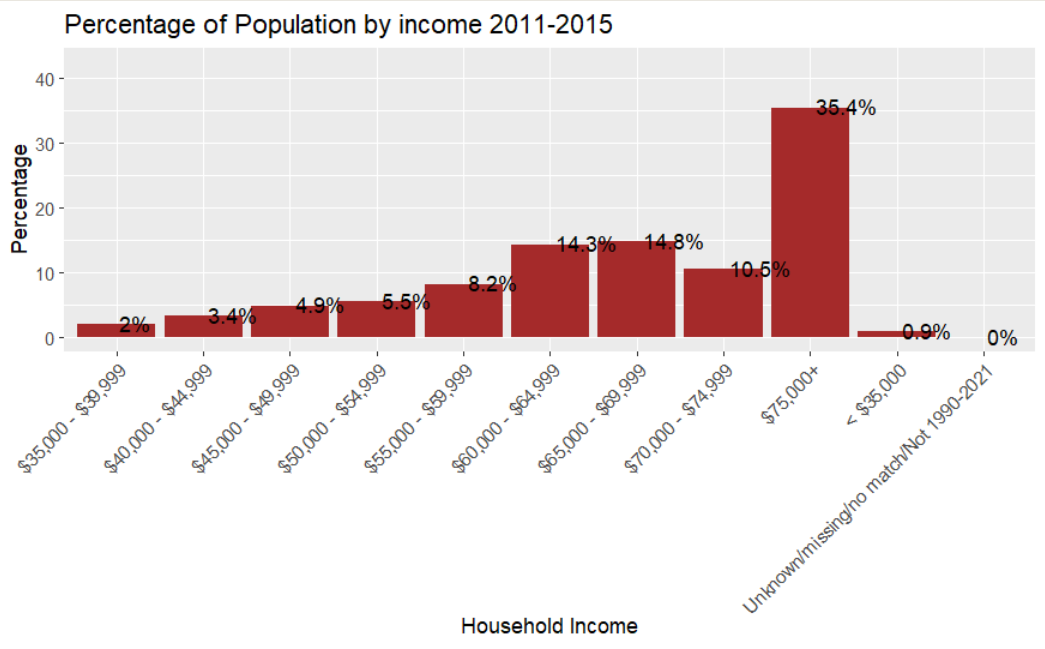
```
── Data Summary ────────────────────────
                           Values
Name                       BREAST_DF_surv_clean
Number of rows             303557
Number of columns          18
___________________________
Column type frequency:
  factor                   14
  numeric                  4
___________________________
Group variables            None
── Data Summary ────────────────────────
                           Values
Name                       BREAST_DF_eval_clean
Number of rows             131395
Number of columns          17
___________________________
Column type frequency:
  factor                   13
  numeric                  4
___________________________
Group variables            None
```

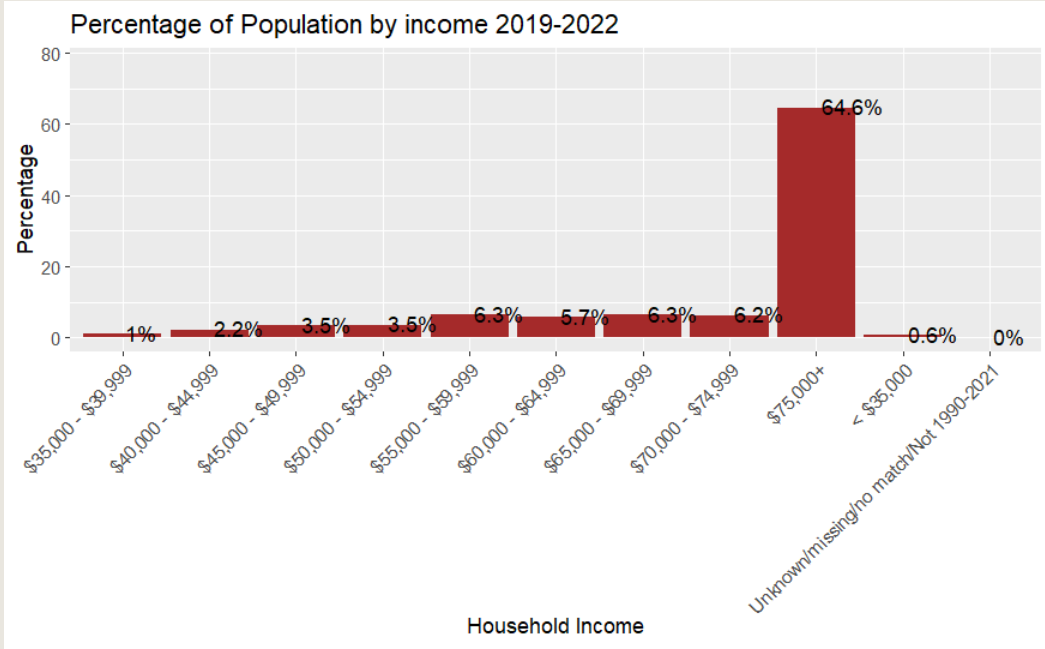
RACE


Percentage of Population by Race between 2011-2015
Percentage
75
50
25
0
0.6%
8.9%
10.6%
0.6%
79.3%
American Indian/Alaska Native
Asian or Pacific Islander
Black
Unknown
White
Race recode (W, B, AI, API)

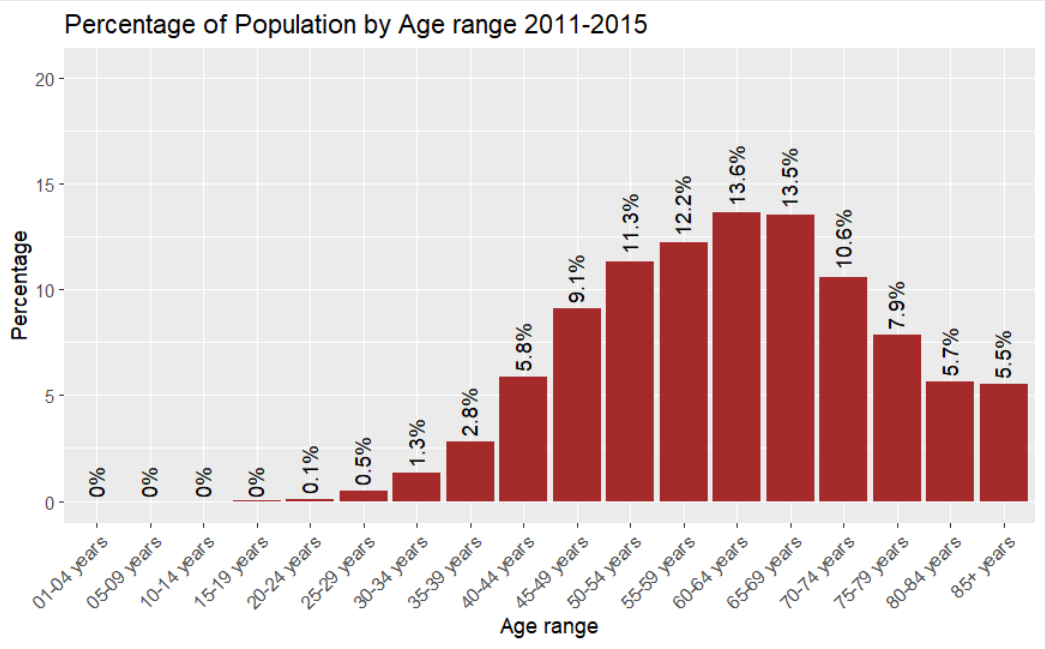

Percentage of Population by between 2019-2022
Percentage
75
50
25
0
0.7%
10.2%
11.1%
1.4%
76.6%
American Indian/Alaska Native
Asian or Pacific Islander
Black
Unknown
White
Race recode (W, B, AI, API)

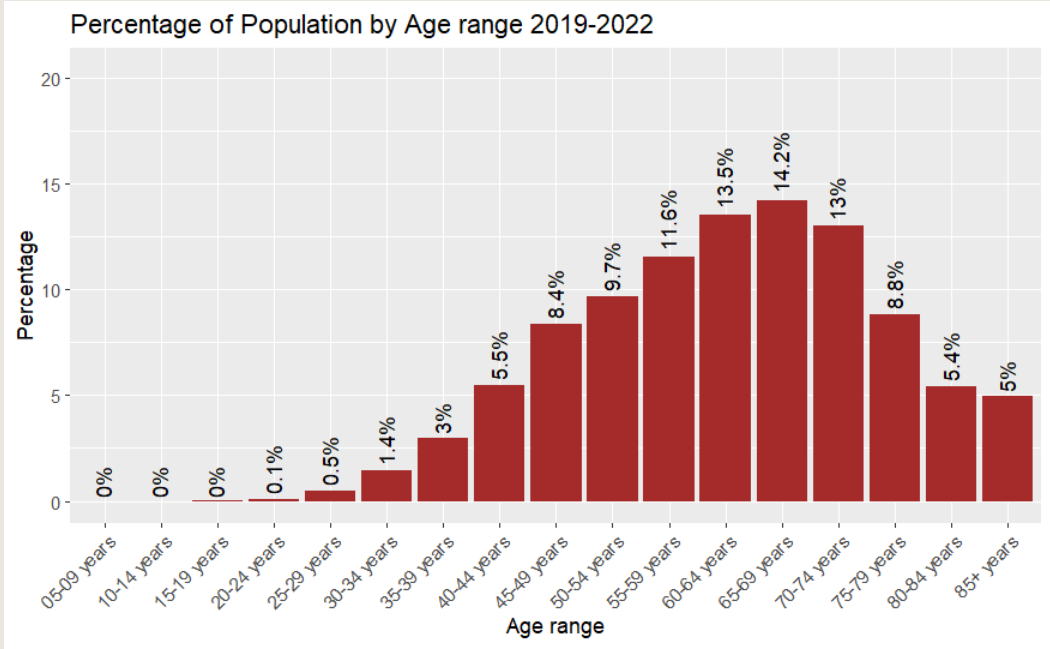
HOUSEHOLDS INCOME



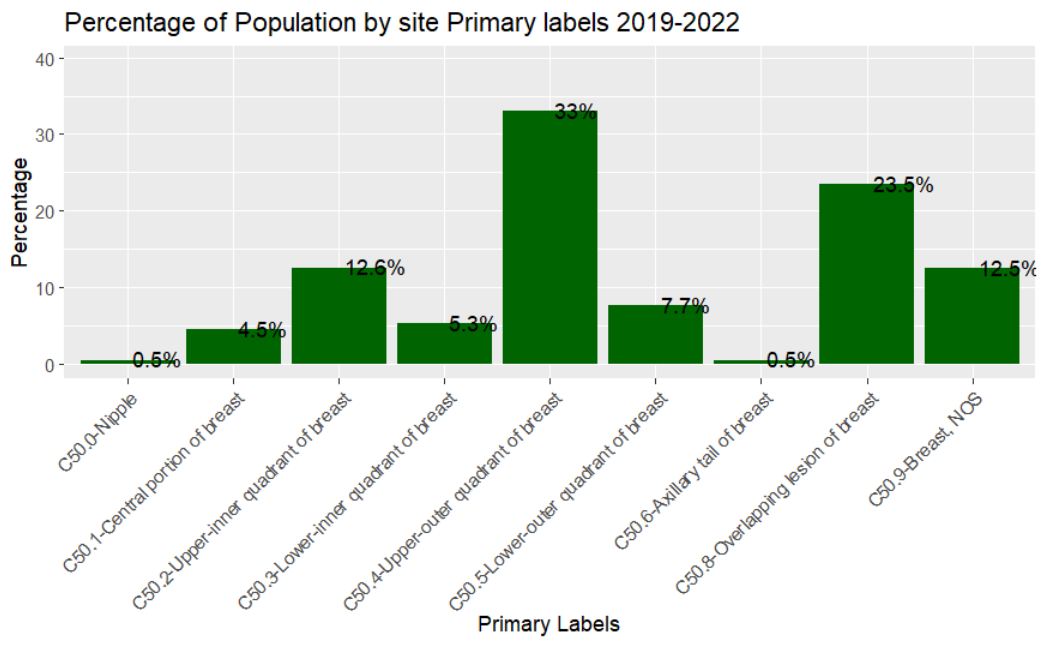


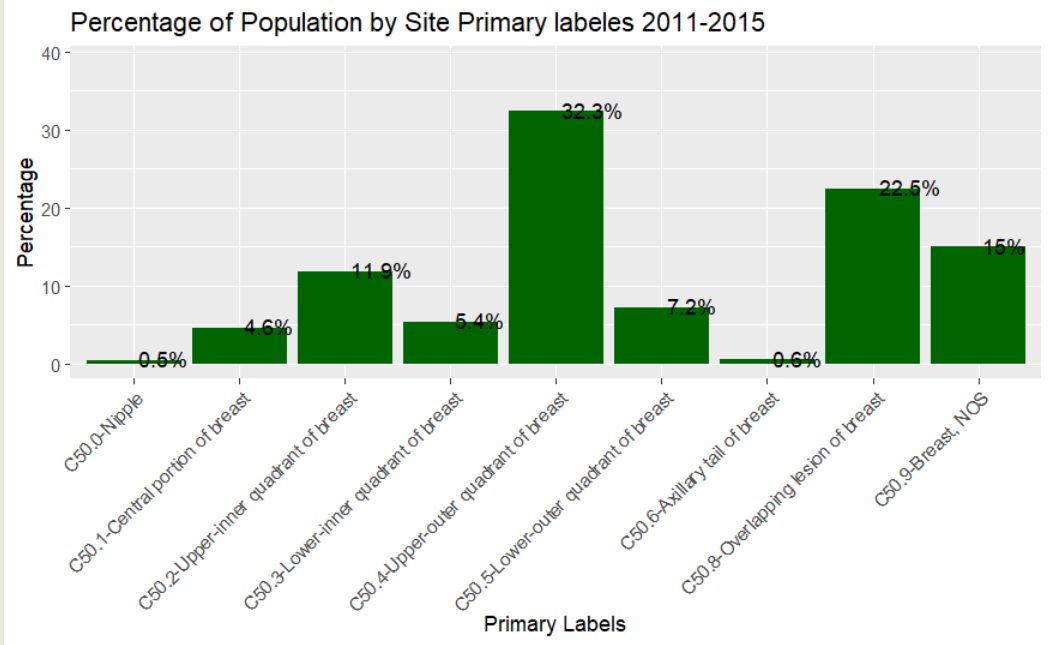
CANCER AGE DISTRIBUTION



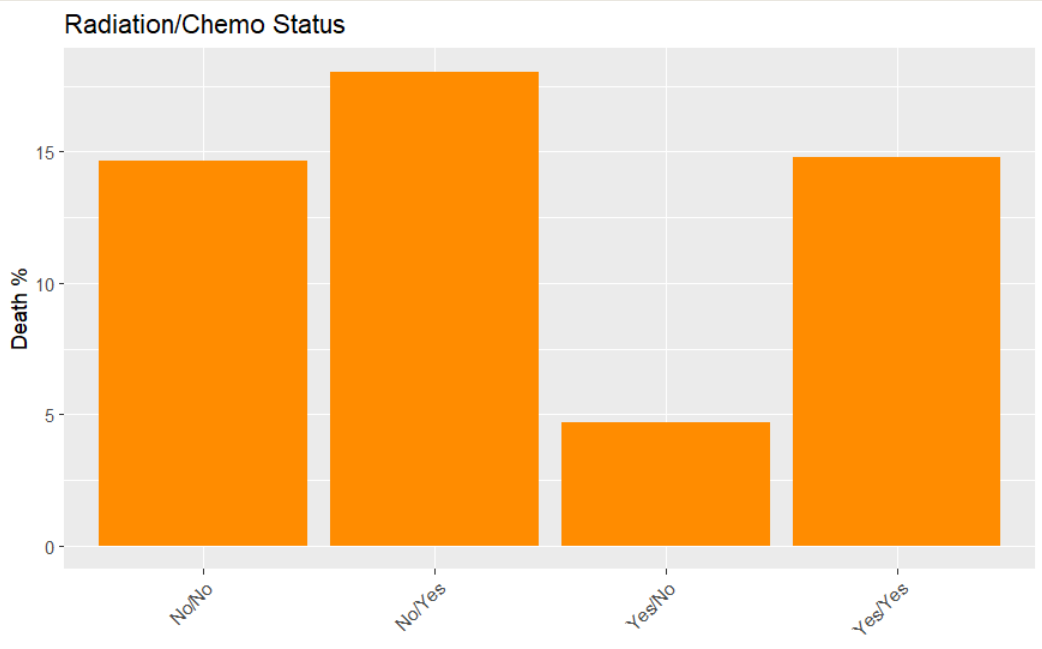


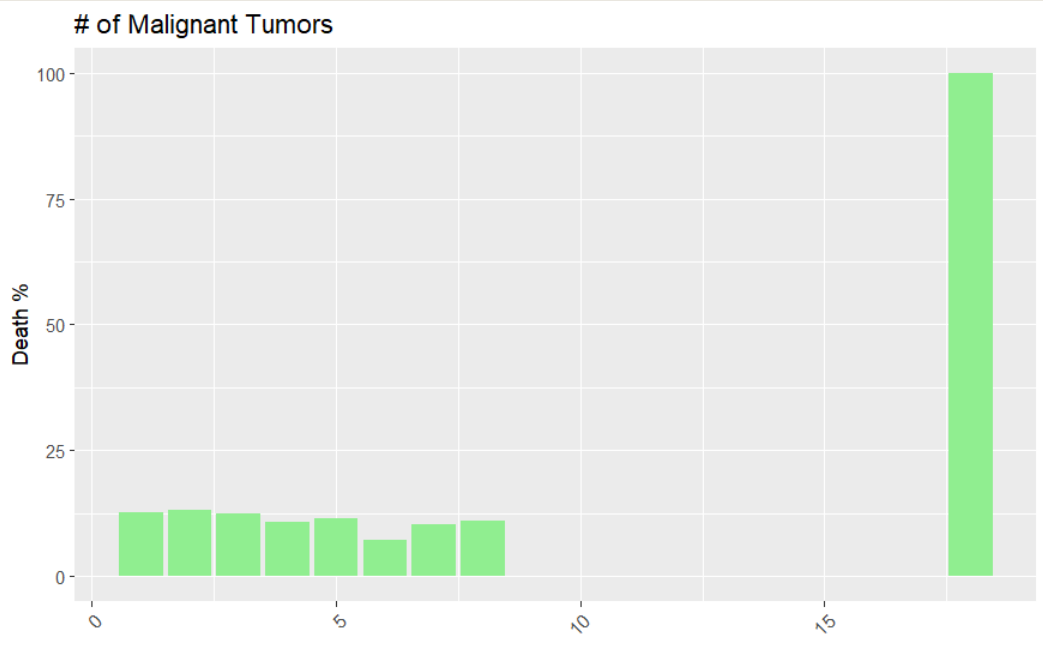
CANCER PRIMARY SITE LABELS



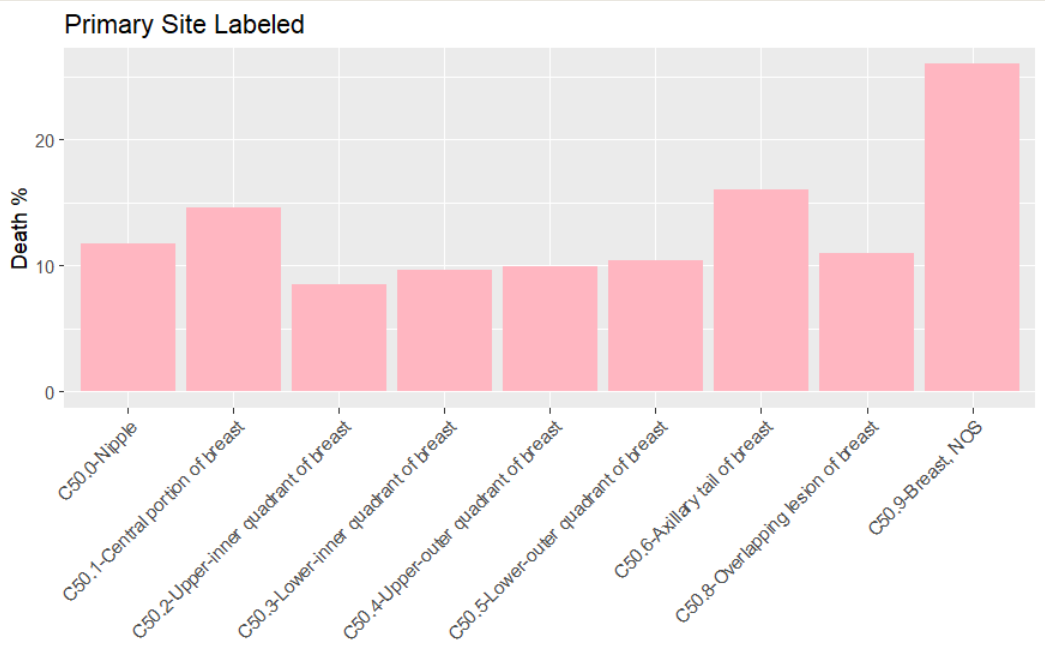


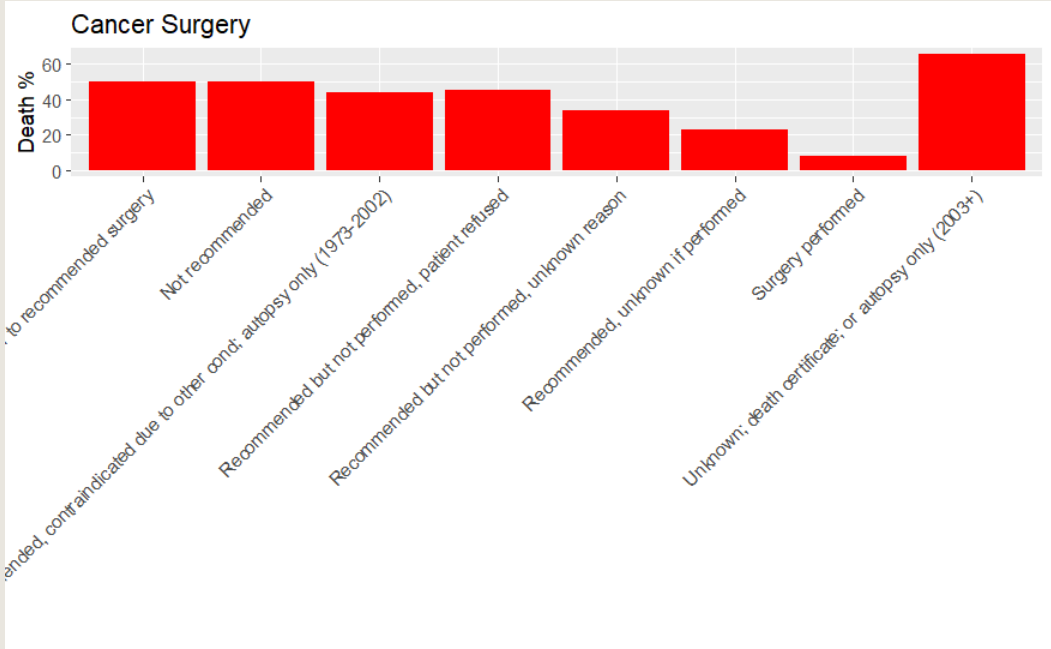
PARAMETERS SELECTION: RADIATION/CHEMO & MALIGNMENT TUMORS



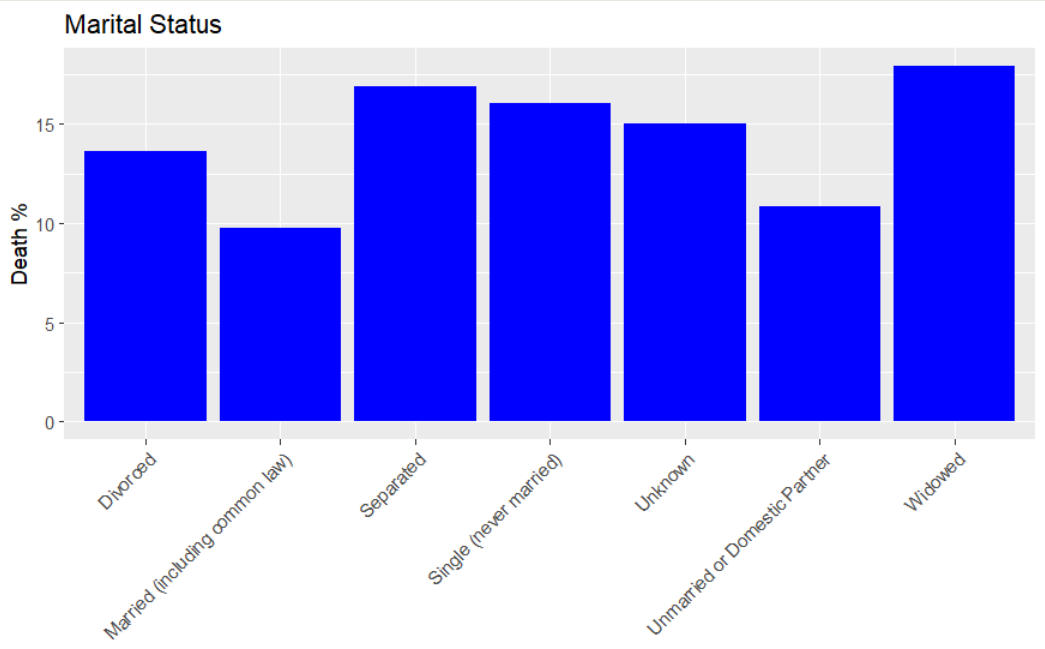


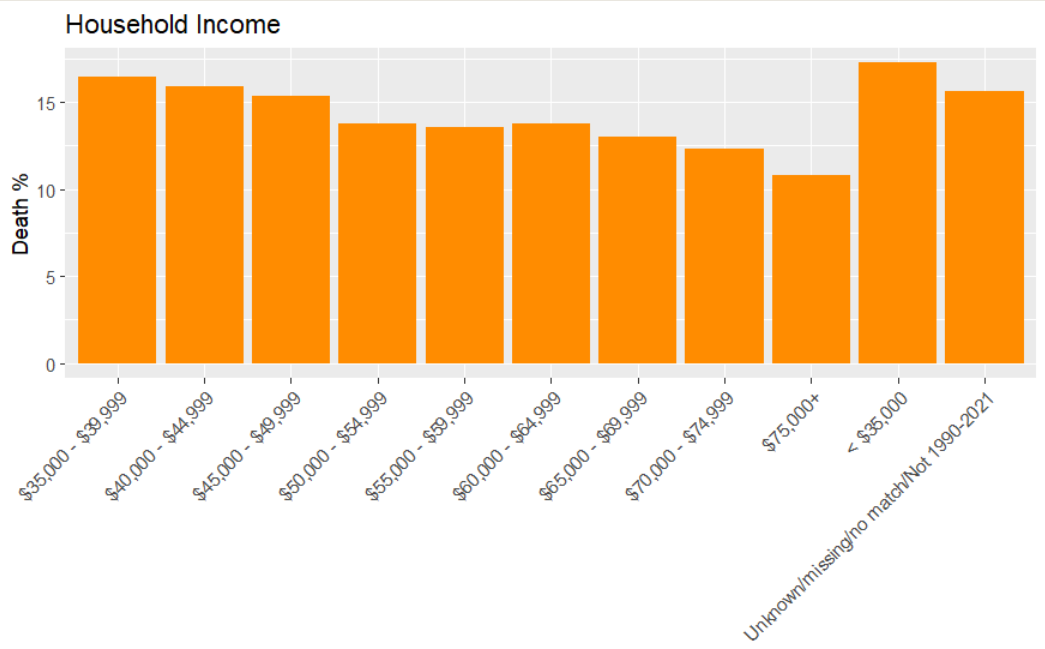
PARAMETERS SELECTION: PRIMARY SITE & SURGERY



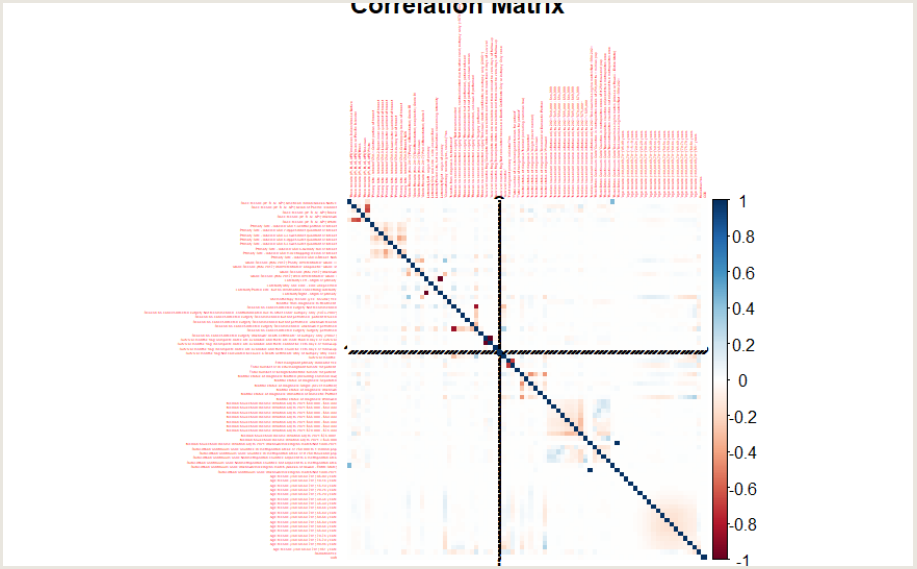


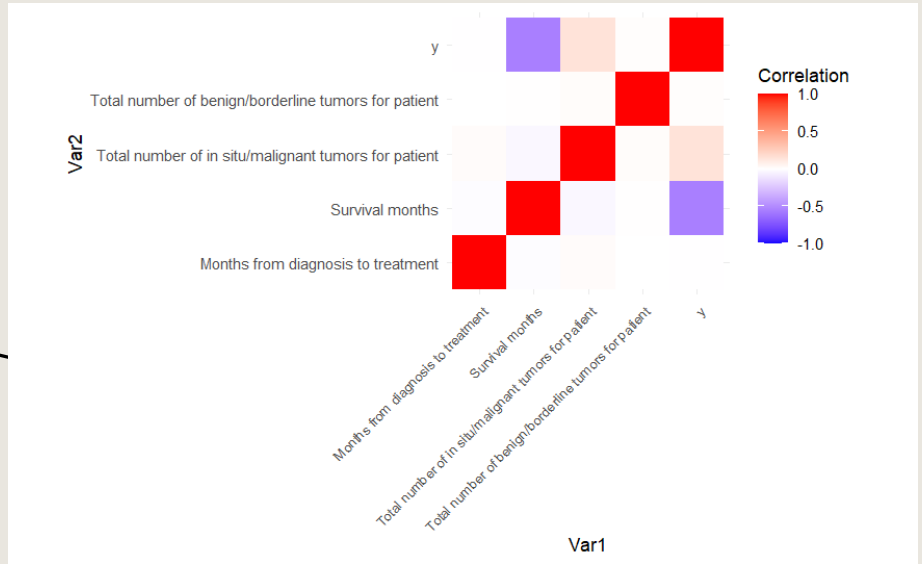
PARAMETERS SELECTION: MARITAL STATUS & HOUSEHOLD INCOME

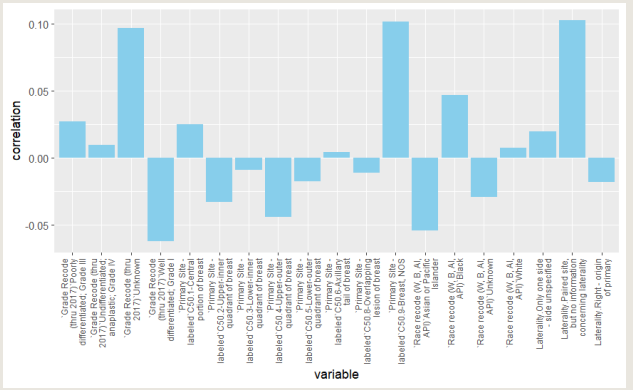


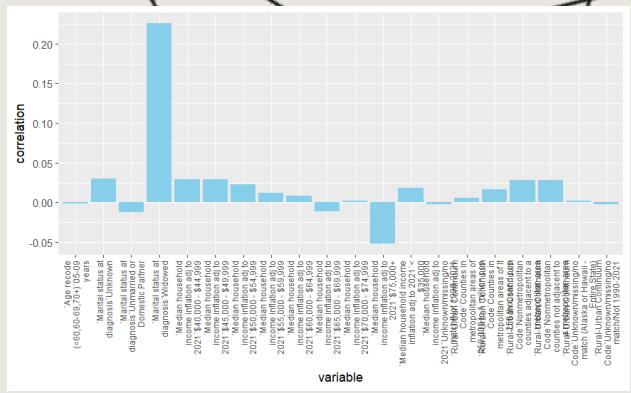


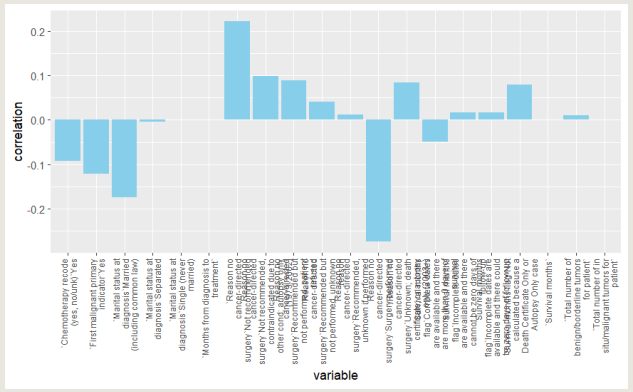
CORRELATION ANALYSES

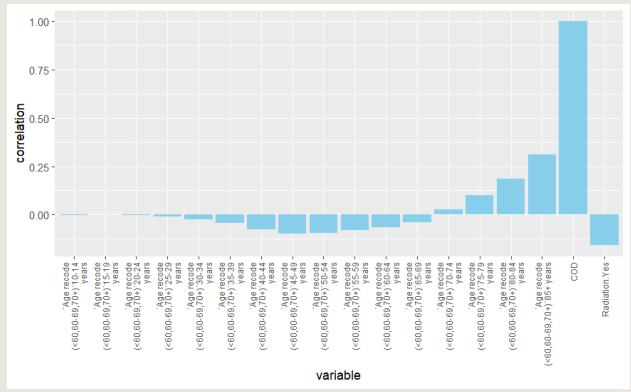










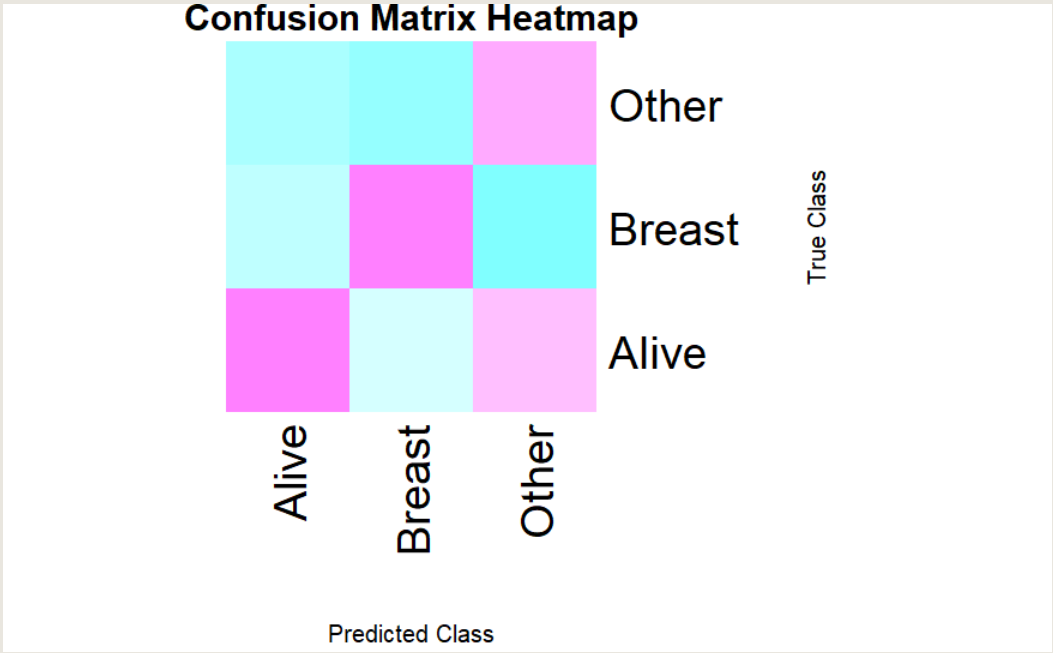


RANDOM FOREST

```
Confusion Matrix and Statistics

          Reference
Prediction  Alive Breast  Other
    Alive  173595   6943  10682
    Breast   2662  14334   4729
    Other    1923   4220  11079
```

```
Overall Statistics

               Accuracy : 0.8646
                 95% CI : (0.8632, 0.866)
    No Information Rate : 0.7741
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5992

 Mcnemar's Test P-Value : < 2.2e-16
```

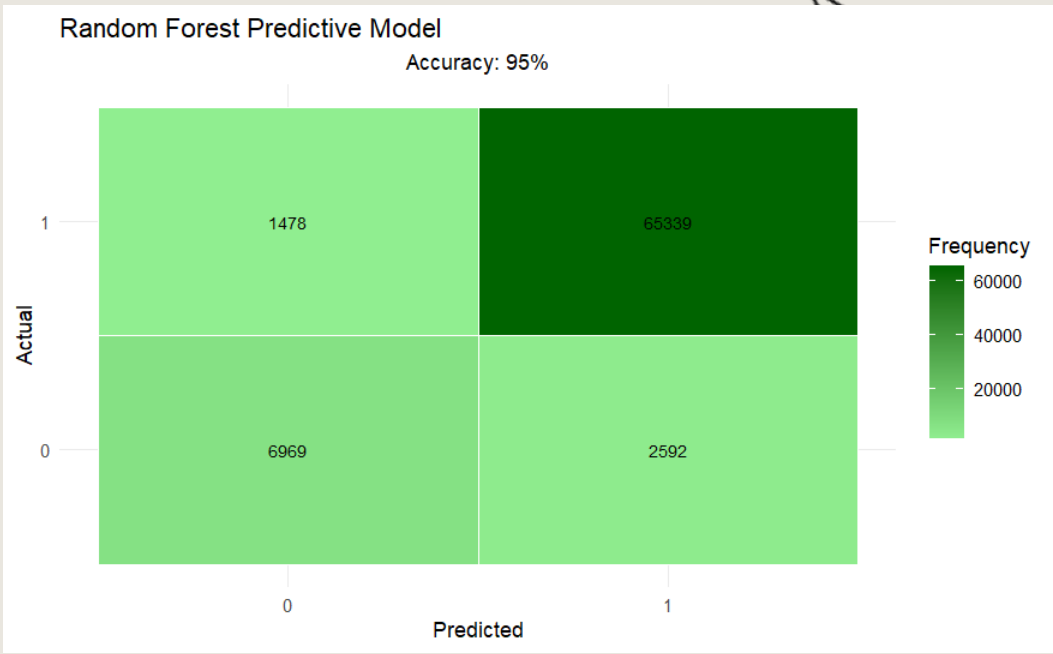
```
Statistics by Class:

                     Class: Alive Class: Breast Class: Other
Sensitivity                0.9743        0.56218       0.41823
Specificity                0.6610        0.96389       0.96984
Pos Pred Value             0.9078        0.65979       0.64331
Neg Pred Value             0.8823        0.94645       0.92763
Prevalence                 0.7741        0.11078       0.11509
Detection Rate             0.7542        0.06228       0.04813
Detection Prevalence       0.8308        0.09439       0.07482
Balanced Accuracy          0.8176        0.76304       0.69404
```

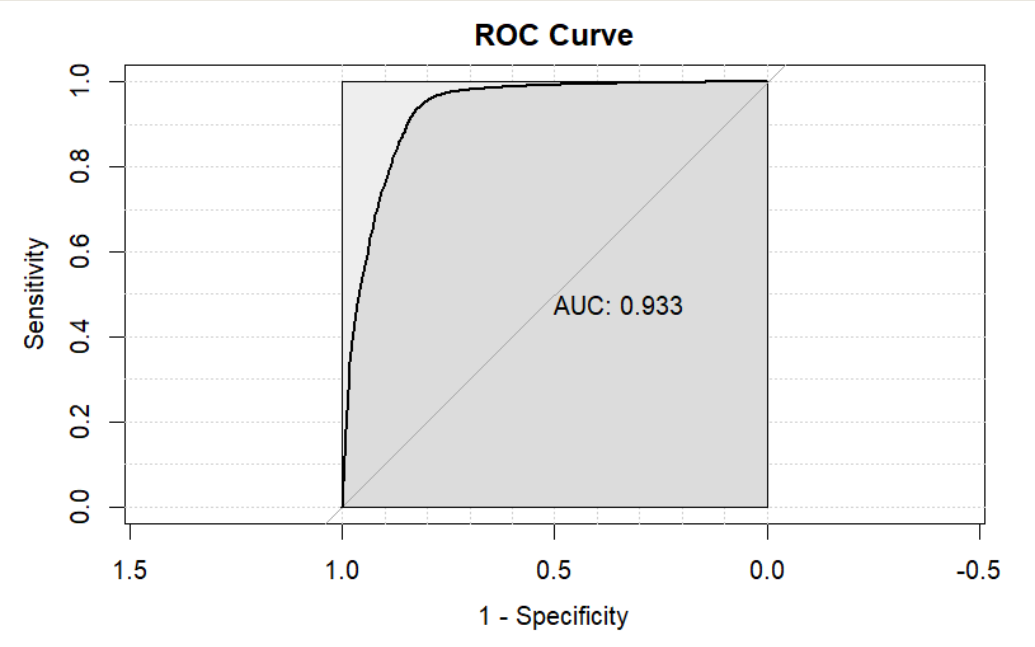


RANDOM FOREST

```
Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.72900
            Specificity : 0.97786
         Pos Pred Value : 0.82495
         Neg Pred Value : 0.96186
             Prevalence : 0.12518
         Detection Rate : 0.09126
   Detection Prevalence : 0.11062
      Balanced Accuracy : 0.85343

       'Positive' Class : 1

           Length Class             Mode
response    76378 factor            numeric
predictor  152756 matrix            numeric
percent         1 -none-            logical
levels          2 -none-            character
rocs            1 -none-            list
auc             1 mv.multiclass.auc numeric
call            3 -none-            call
```

```
Confusion Matrix and Statistics

          Reference
Prediction     1     2
         1  6970  1479
         2  2591 65338

               Accuracy : 0.9467
                 95% CI : (0.9451, 0.9483)
    No Information Rate : 0.8748
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7439
```

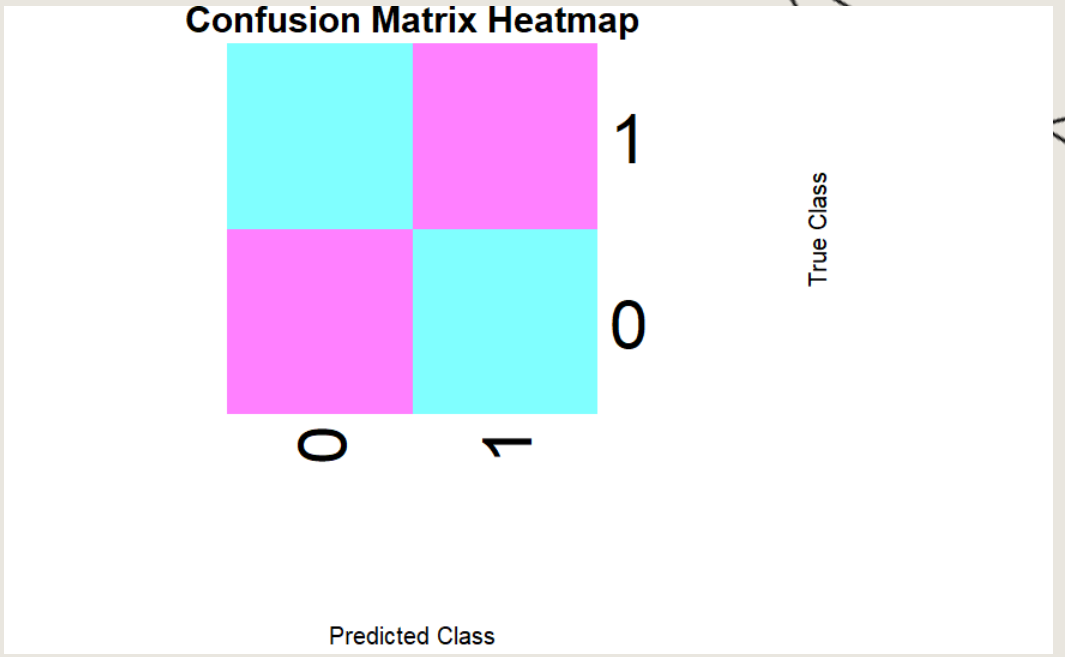


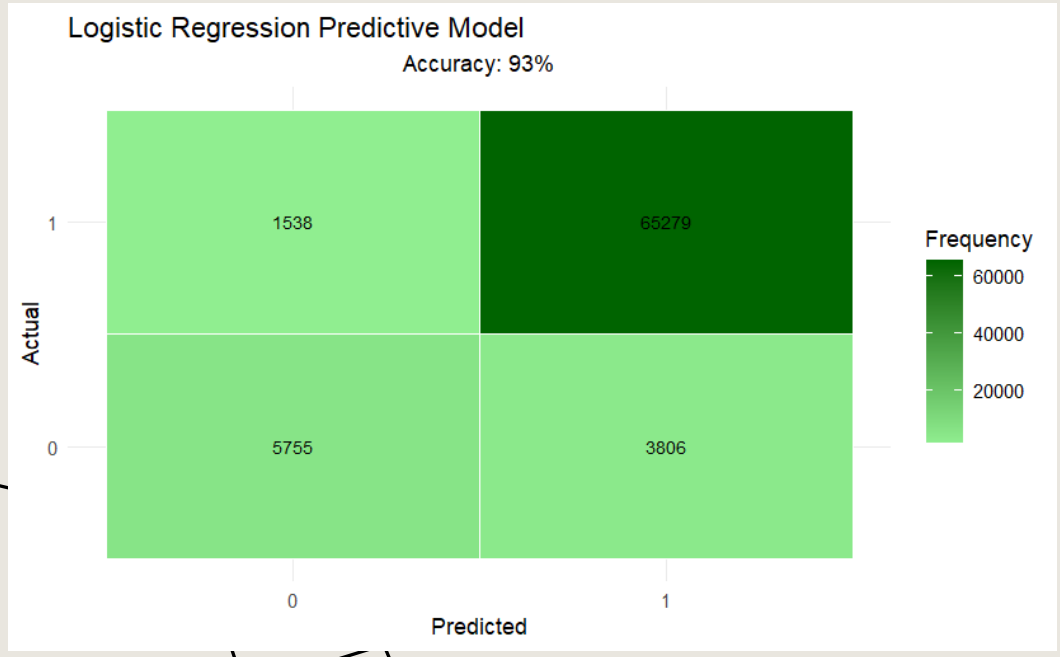


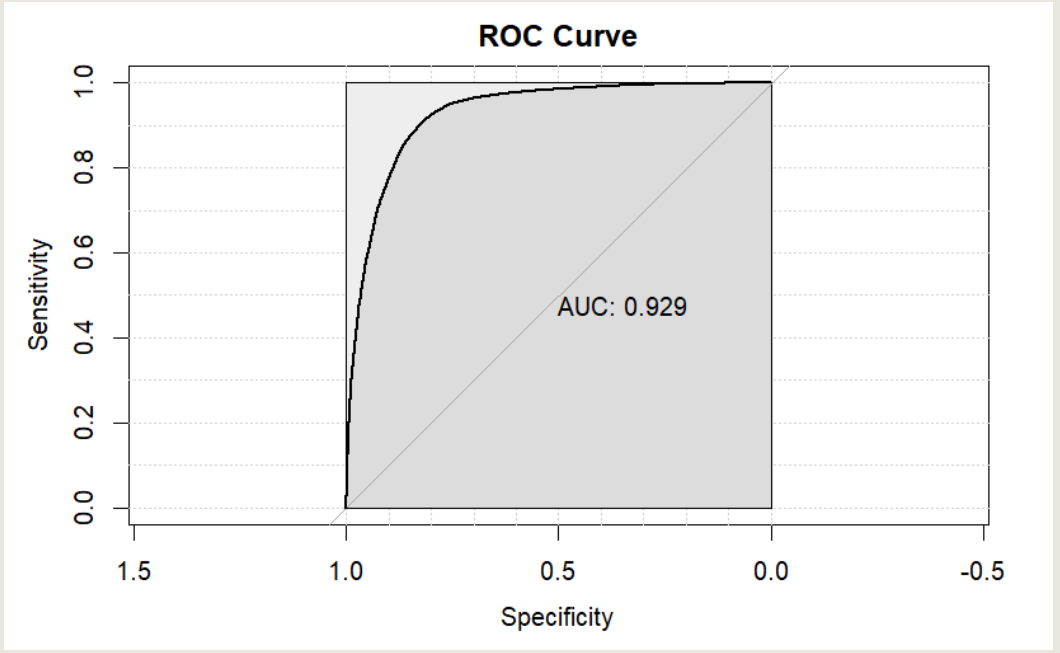
LOGISTIC REGRESSION

```
predicted_class     0     1
              0  5755  1538
              1  3806 65279
[1] "Accuracy: 0.9300322082275"
```

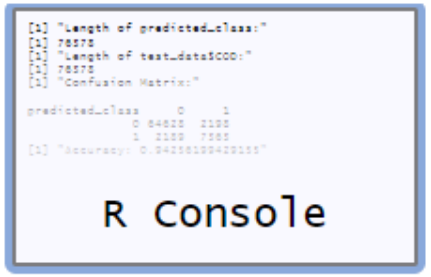
```
Call:
roc.default(response = test_data$COD, predictor = predictions_logistic)

Data: predictions_logistic in 9561 controls (test_data$COD 0) < 66817 cases
(test_data$COD 1).
Area under the curve: 0.9291
```

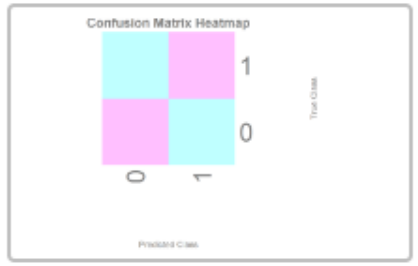




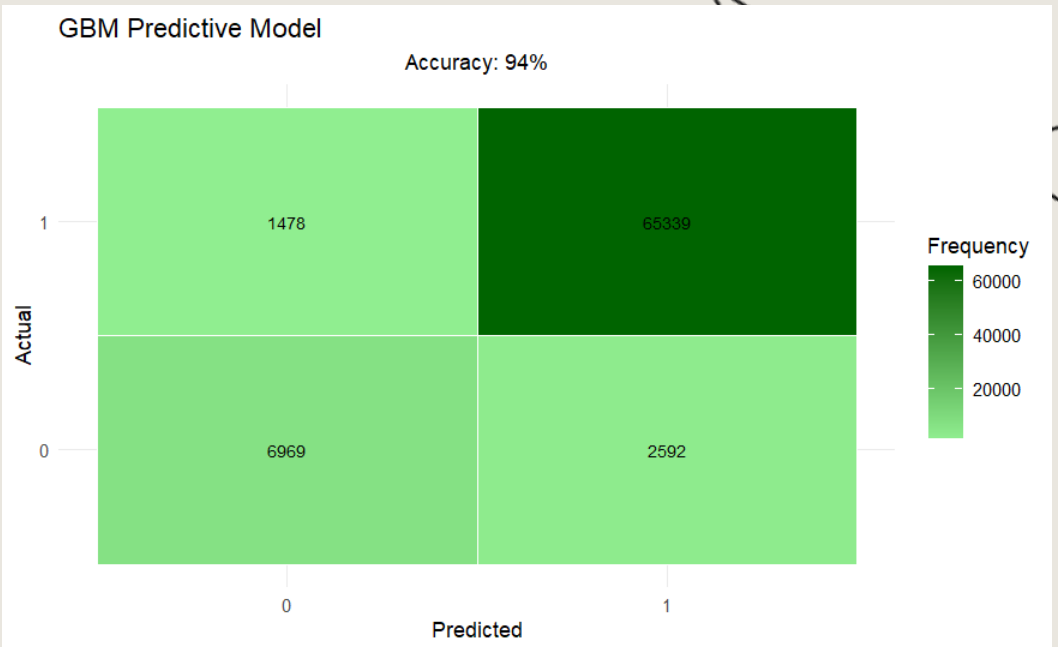


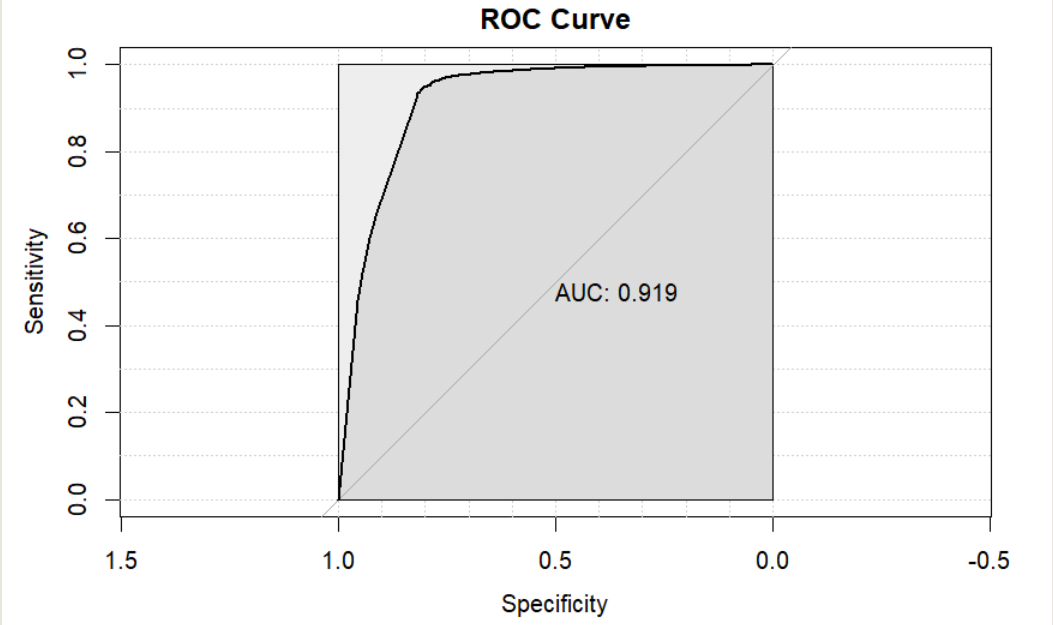
GRADIENT BOOSTING






```
[1] "Length of predicted_class:"
[1] 76378
[1] "Length of test_data$COD:"
[1] 76378
[1] "Confusion Matrix:"

predicted_class     0     1
              0 64628  2198
              1  2189  7363
[1] "Accuracy: 0.94256199429155"
```





DEEP NEURAL NETWORK

```
Test Loss: 0.2786484
Test Accuracy: 0.9415277
2387/2387 [==============================] - 2s 916us/step
[1] "Confusion Matrix:"
      Predicted
Actual     0     1
     0  6487  3074
     1  1392 65425
[1] "Accuracy: 0.941527665034434"
[1] "Sensitivity: 0.979166978463565"
[1] "Specificity: 0.678485514067566"
```