Big data analytics - Final report

B215284

06/12/23

Introduction

Problem statement

Breast cancer, the most common cancer in women, also has the highest cancer-related mortality rate. Recent advances in the treatment of breast cancer have improved survival rates. The most common clinical question asked is prognosis and survival rates of the cancer. The survival rates are also helpful to assess the advantages recent advances in the cancer treatment and efficacy of the policy making, screen and healthcare spending. Current prognostication of breast cancer is typically based on the patient's clinical characteristics, tumor stage and histology, and treatment type. Factoring in the genetic data, can we predict the survival outcomes more accurately?

The data:

Source: Collected by a combined effort from researchers from the UK and Canada (Cancer Research UK, the British Columbia Cancer Foundation and Canadian Breast Cancer Foundation BC/Yukon). The data is hosted at cBioPortal.

Contents:

```
Information about around 1980 breast cancer patients

the first 31 columns contain clinical details and survival information

the next 331 columns contain m-RNA levels reflecting gene expression by the tumour. the values are stand

the next 175 columns contain information of gene mutations.
```

Goal of the project:

The aim of this project is to develop a statistical model for prediction of survival outcome of a patient with given clinical and general sequencing information.

The data is explored with tables and graphs to arrive at initial impressions. The data is then prepared to model fitting by filling the missing values.

A logistic regression model is fitted into the data to predict the outcomes. Theoretically, all the dependent variables could be fit into the model to assess the most important variable for fine tuning. as the dataset is large, selected few dependent variables are used for model fitting.

The model will be evaluated with metrics like accuracy, area under curve and True and false positive/negative rates.

Importing libraries

```
library(sparklyr)
library(dplyr)
library(ggplot2)
library(knitr)
library(corrr)
library(dbplot)
library(reshape)
```

Creating a spark connection

The next block creates a spark connection to the master instance of the clusters. This Spark master node is responsible for delegating the computation to the executor nodes. The data is imported as a spark table with the spark_read_csv function. The connection acts as an SQL database. The dplyr operation acts as SQL operations under the hood. Most of the operations are carried out on the spark table. The data is imported as R tibble when absolutely necessary.

```
sc <- spark_connect(
  master="spark://spark.eidf071:7077",
)

project_path <- '/work/eidf071/eidf071/shared/Projects/Breast cancer gene expression/'

metabric_RNA_data <- spark_read_csv(
  sc, name='diabetes_data',
  path=sprintf('file:///%s/METABRIC_RNA_Mutation.csv', project_path),
)
```

Data overview

```
# number of rows
count(metabric_RNA_data)
```

```
## # Source: spark<?> [?? x 1]
##       n
##   <dbl>
## 1  1904
```

```
# number of columns
ncol(metabric_RNA_data)
```

```
## [1] 693
```

```
# glimpse of data. printing first 10 columns
metabric_RNA_data %>%
  select(1:10) %>%  # select only first 10 columns
  glimpse()
```

```
## Rows: ??
## Columns: 10
## Database: spark_connection
## $ patient_id                <int> 0, 2, 5, 6, 8, 10, 14, 22, 28, 35, 36, 39, 4~
## $ age_at_diagnosis          <dbl> 75.65, 43.19, 48.87, 47.68, 76.97, 78.77, 56~
## $ type_of_breast_surgery    <chr> "MASTECTOMY", "BREAST CONSERVING", "MASTECTO~
## $ cancer_type               <chr> "Breast Cancer", "Breast Cancer", "Breast Ca~
## $ cancer_type_detailed      <chr> "Breast Invasive Ductal Carcinoma", "Breast ~
## $ cellularity               <chr> NA, "High", "High", "Moderate", "High", "Mod~
## $ chemotherapy              <int> 0, 0, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 1,~
## $ pam50__claudinlow_subtype <chr> "claudin-low", "LumA", "LumB", "LumB", "LumB~
## $ cohort                    <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ er_status_measured_by_ihc <chr> "Positve", "Positve", "Positve", "Positve", ~
```

Number of missing values

```
missing_values <- metabric_RNA_data %>%
  mutate_all(is.na) %>%   # convert all the values to 1 or 0 based on if they are NA or not
  mutate_all(as.numeric) %>%    # convert them to numerical
  summarise_all(sum) %>%   # sum of all the na (or 1s if they are na)
  collect()  # concert to tibble

transposed_data <- data.frame(t(missing_values),row.names = names(missing_values))   # transposing the
```

```
transposed_data %>%
  arrange(desc(t.missing_values.)) %>%
  head(15) %>%
  mutate(percentage = round(t.missing_values. / 1904 * 100, 2)) %>%
   kable(caption = "Summary of missing vaules")
```

Table 1: Summary of missing vaules

	t.missing_values.	percentage
tumor_stage	501	26.31
3gene_classifier_subtype	204	10.71
primary_tumor_laterality	106	5.57
neoplasm_histologic_grade	72	3.78
cellularity	54	2.84
mutation_count	45	2.36
er_status_measured_by_ihc	30	1.58
type_of_breast_surgery	22	1.16
tumor_size	20	1.05
cancer_type_detailed	15	0.79
tumor_other_histologic_subtype	15	0.79
oncotree_code	15	0.79
death_from_cancer	1	0.05
patient_id	0	0.00
age_at_diagnosis	0	0.00

The tumour stage, 3gene classification type and primary tumour laterality are the most common type of missing information. Rest of the columns have percentage of missing values less than 5%.

Summary of survival outcomes

```
metabric_RNA_data %>%
  group_by(overall_survival) %>%
  tally() %>%
  mutate(percentage = round(n / sum(n) * 100, 2)) %>%  # creating a percentage column and rounding the
  kable()
```

overall_survival	n	percentage
1	801	42.07
0	1103	57.93

Explorative data analysis

Relationship between the clinical attributes and the outcome

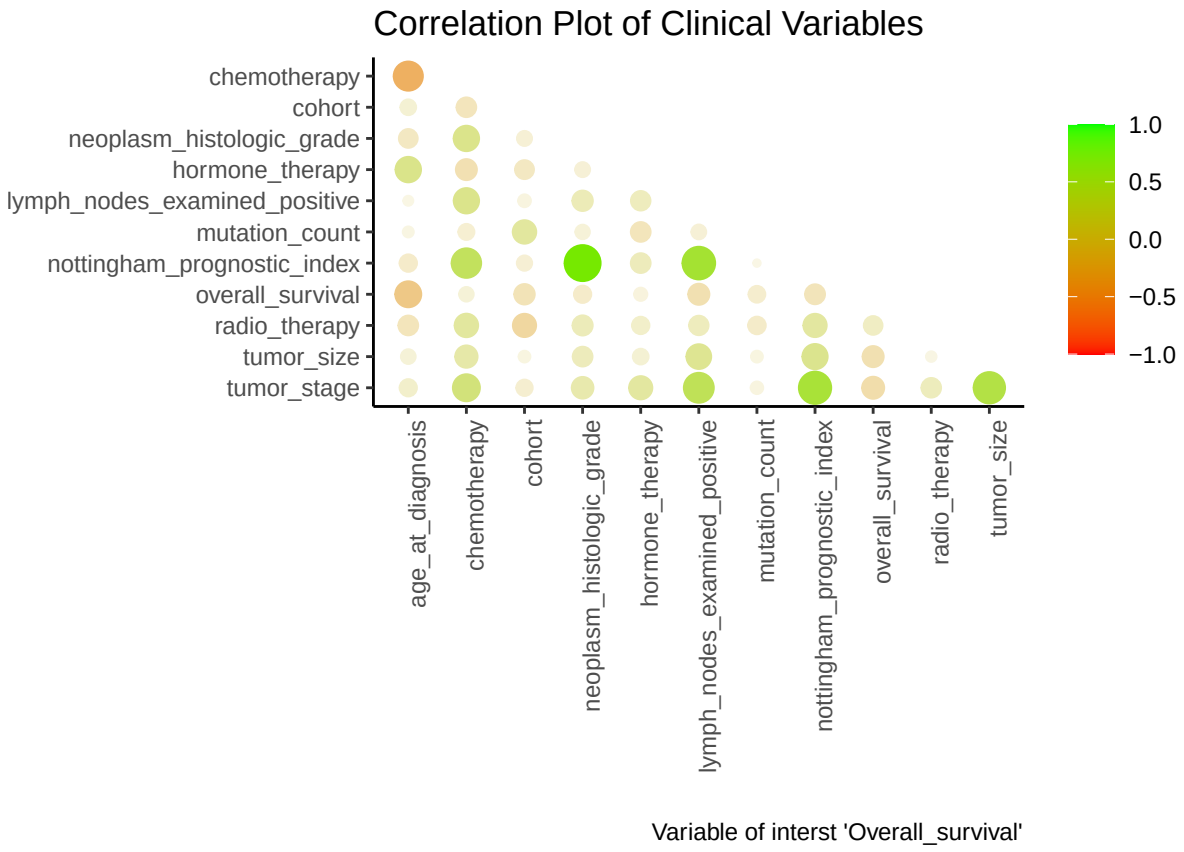
```
# Caching the clinical info in to a new spark table
clinical_info <- metabric_RNA_data %>%
  select(1:31) %>%
  compute('clinical_info')
```

The caching is creating another spark data frame with memory across the cluster nodes. The cached data frame acts similar to the spark data frame with same connection.

Correlation plot

```
clinical_info %>%
  collect() %>%  # Converting to R tibble
  select_if(is.numeric) %>%  # selecting only numerical columns. I tried to included all the variables,
  select(-patient_id, -overall_survival_months) %>%  # patient id is not relevant. survival month is on
  correlate(method = "pearson") %>%
  shave() %>%  # remove the half of the correlation matrix
  rplot(shape = 19, colors = c("red", "green"), legend = TRUE) +
  ggtitle("Correlation Plot of Clinical Variables") +
  labs(caption = "Variable of interst 'Overall_survival'",
       fill = "Correlation") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

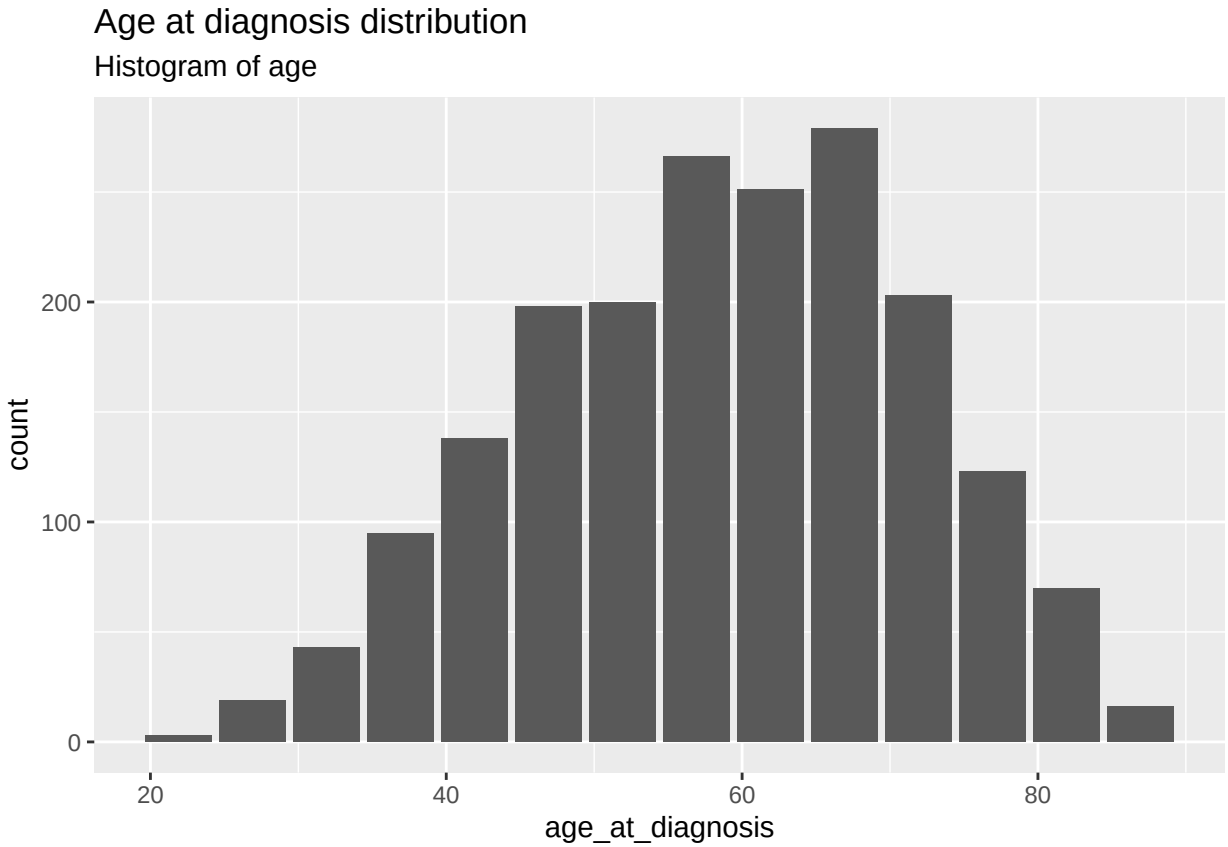
```
## Correlation computed with
## * Method: 'pearson'
## * Missing treated using: 'pairwise.complete.obs'
```



The survival is negatively correlated with age at diagnosis, number of positive lymph nodes, Nottingham prognostic index, tumour size, and tumour stage.

Age

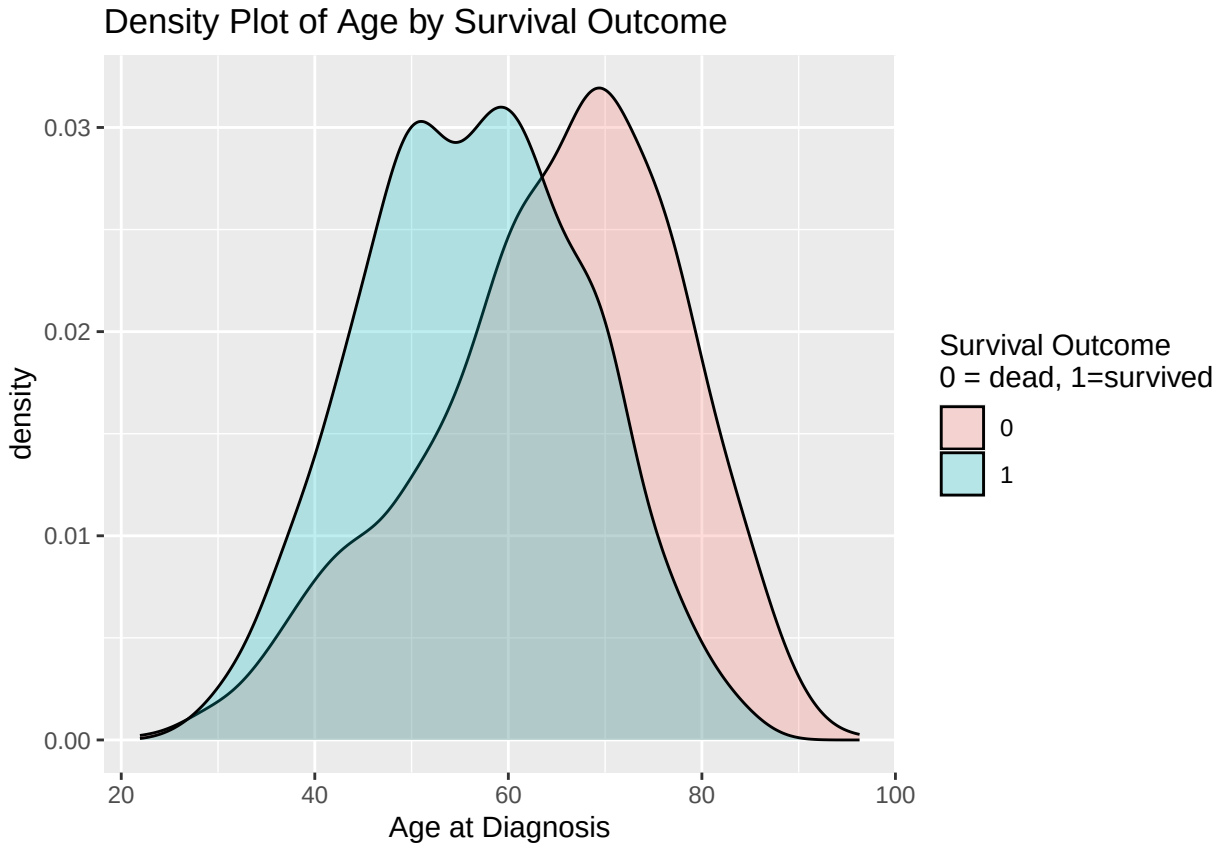
```
clinical_info %>%
dbplot_histogram(age_at_diagnosis, binwidth = 5) +
labs(title = "Age at diagnosis distribution",
     subtitle = "Histogram of age")
```



The Dbplot is developed to plot spark data frames. The Dbplot operated directly with spark taking advantage of distributive computation. There is no need to collect the data for a plot which is in the case of ggplot. The above plot shows the average of the age is around 60s at the time of diagnosis with a slight neagtive skew.

```
clinical_info %>%
  select(age_at_diagnosis, overall_survival) %>%
  collect() %>%
  mutate(overall_survival = as.factor(overall_survival)) %>%
  melt() %>%
  ggplot(aes(x=value, group=overall_survival, fill=overall_survival)) +
  geom_density(alpha=.25) +
  labs(
    title = "Density Plot of Age by Survival Outcome",
    x = "Age at Diagnosis",
    fill = "Survival Outcome \n0 = dead, 1=survived")
```

```
## Using overall_survival as id variables
```

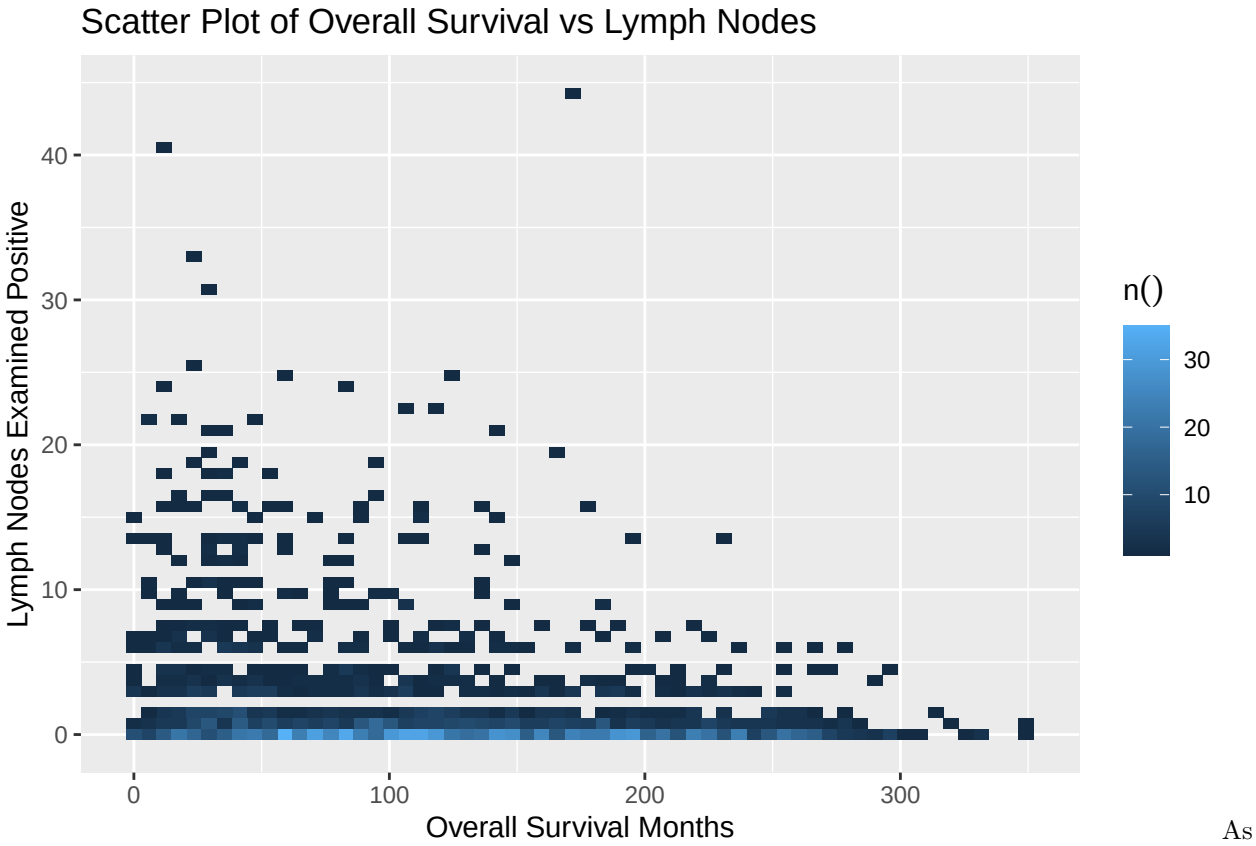


Age is an important predictor of survival in most of the medical conditions. Age affects the survival outcomes in various ways including clinical decision of offering radical treatment or with underlying co-morbidities.

The overlapping density plot conveys the same information in different visual. The range age range of these two groups appears same but the density plot of the deceased group is negatively skewed.

Number of positive lymph nodes

```
dbplot_raster(clinical_info, overall_survival_months, lymph_nodes_examined_positive, resolution = 60) +
  labs(
    title = "Scatter Plot of Overall Survival vs Lymph Nodes",
    x = "Overall Survival Months",
    y = "Lymph Nodes Examined Positive"
  )
```



As the Dbplot uses a continuous stream of data to plot the graph, the data points are pixelated. The plot shows a slight negative correlation with survival outcomes.

Feature extration

The below code creates a column showing proportion of the survived according tumor stage. Standard deviation for each is calculated.

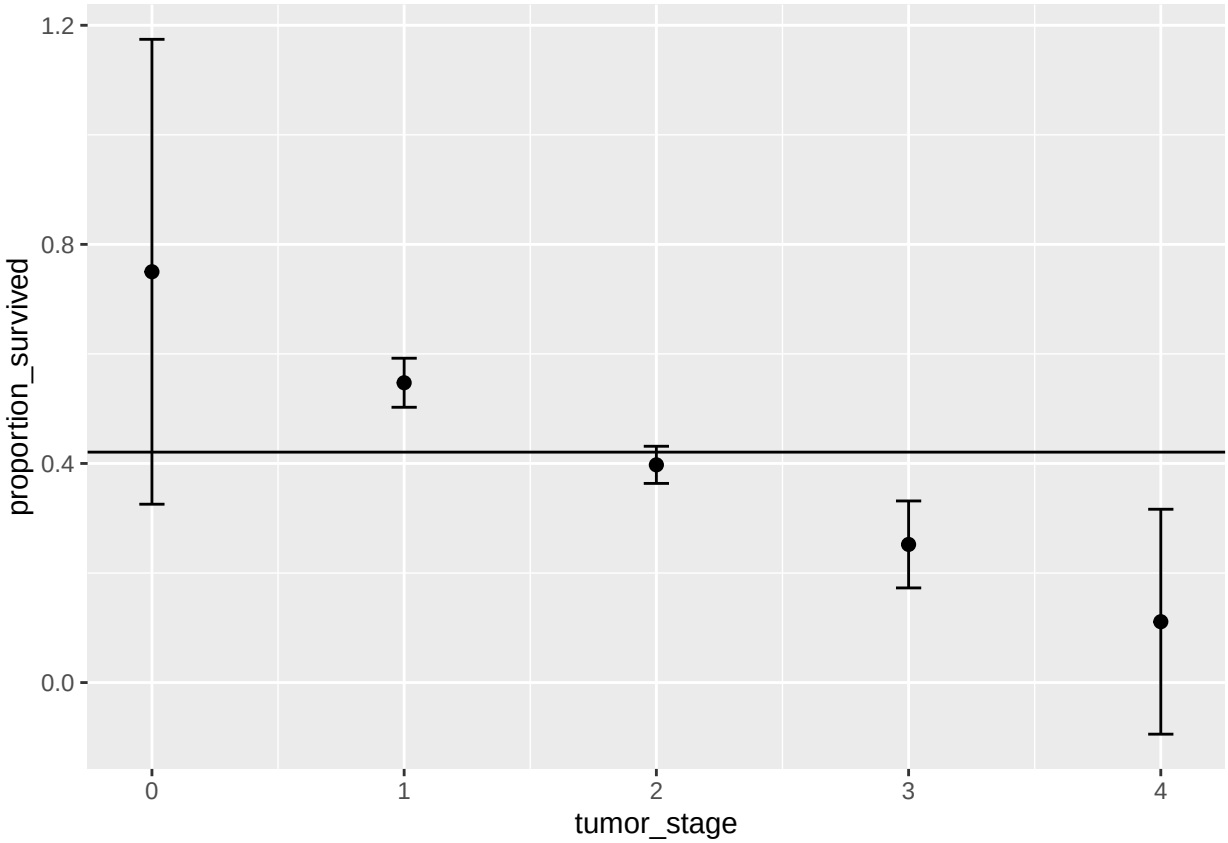
```
prop_data <- clinical_info %>%
  group_by(tumor_stage, overall_survival) %>%
  tally() %>%
  group_by(tumor_stage) %>%
  summarize(count = sum(n),
            proportion_survived = sum(overall_survival * n) / sum(n)) %>%
   mutate(se = sqrt(proportion_survived * (1 - proportion_survived) / count)) %>%
  arrange(tumor_stage) %>%
  collect()

prop_data
```

```
## # A tibble: 6 x 4
##   tumor_stage count proportion_survived     se
##         <dbl> <dbl>               <dbl>  <dbl>
## 1          NA   501               0.379 0.0217
## 2           0     4               0.75  0.217
## 3           1   475               0.547 0.0228
## 4           2   800               0.398 0.0173
## 5           3   115               0.252 0.0405
## 6           4     9               0.111 0.105
```

Plotting the proportions

```
prop_data %>%
  ggplot(aes(x = tumor_stage, y = proportion_survived)) +
  geom_point(size = 2) +
  geom_errorbar(aes(ymin = proportion_survived - 1.96 * se,
                    ymax = proportion_survived + 1.96 * se),
                width = .1) +
  geom_hline(yintercept = sum(prop_data$proportion_survived * prop_data$count) /
                          sum(prop_data$count))
```



When tumour stages are plotted against the proportion of survival, a decrease in survival proportion with higher tumour stage is seen. The confidence interval is highest for stage 0 which is also crossing the average horizontal line indicating that when taken alone, stage 0 is not a good predictor of high or low survival rate than average.

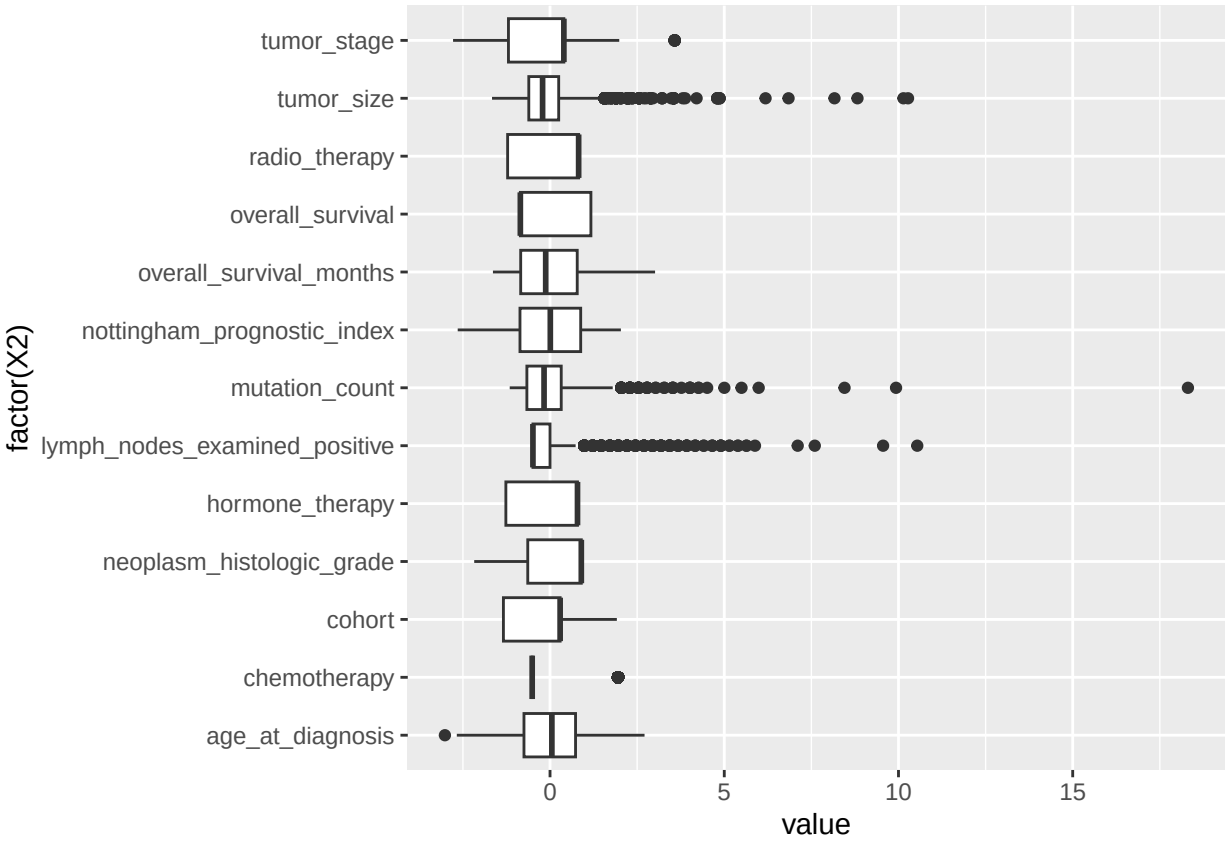
Visualising the distribution of clinical variables

```
scaled_clinical_info <- clinical_info %>%
  select_if(is.numeric) %>%
  select(-patient_id) %>%
  collect() %>%
  scale()
```

```
melt_data <- melt(scaled_clinical_info)

ggplot(melt_data, aes(factor(X2), value)) +
  geom_boxplot() +
```

```
  coord_flip()
```



Horizontal box plot of normalized values of clinical variables shows the mutation count, tumour size, number of positive lymph nodes carry high extreme values in positive side.

mrna expression

```
mrna_data <- metabric_RNA_data %>%
  select(25, 32:362) %>%
  compute('mrna_data')
```

Getting correltaion data

```
mrna_corr <- mrna_data %>%
  ml_corr() %>%
  slice(1) %>%
  collect()
```

```
## New names:
## * `` -> `...1`
## * `` -> `...2`
## * `` -> `...3`
## * `` -> `...4`
## * `` -> `...5`
## * `` -> `...6`
## * `` -> `...7`
## * `` -> `...8`
```

```
## * `` -> `...9`
## * `` -> `...10`
## * `` -> `...11`
## * `` -> `...12`
## * `` -> `...13`
## * `` -> `...14`
## * `` -> `...15`
## * `` -> `...16`
## * `` -> `...17`
## * `` -> `...18`
## * `` -> `...19`
## * `` -> `...20`
## * `` -> `...21`
## * `` -> `...22`
## * `` -> `...23`
## * `` -> `...24`
## * `` -> `...25`
## * `` -> `...26`
## * `` -> `...27`
## * `` -> `...28`
## * `` -> `...29`
## * `` -> `...30`
## * `` -> `...31`
## * `` -> `...32`
## * `` -> `...33`
## * `` -> `...34`
## * `` -> `...35`
## * `` -> `...36`
## * `` -> `...37`
## * `` -> `...38`
## * `` -> `...39`
## * `` -> `...40`
## * `` -> `...41`
## * `` -> `...42`
## * `` -> `...43`
## * `` -> `...44`
## * `` -> `...45`
## * `` -> `...46`
## * `` -> `...47`
## * `` -> `...48`
## * `` -> `...49`
## * `` -> `...50`
## * `` -> `...51`
## * `` -> `...52`
## * `` -> `...53`
## * `` -> `...54`
## * `` -> `...55`
## * `` -> `...56`
## * `` -> `...57`
## * `` -> `...58`
## * `` -> `...59`
## * `` -> `...60`
## * `` -> `...61`
## * `` -> `...62`
```

```
## * `` -> `...63`
## * `` -> `...64`
## * `` -> `...65`
## * `` -> `...66`
## * `` -> `...67`
## * `` -> `...68`
## * `` -> `...69`
## * `` -> `...70`
## * `` -> `...71`
## * `` -> `...72`
## * `` -> `...73`
## * `` -> `...74`
## * `` -> `...75`
## * `` -> `...76`
## * `` -> `...77`
## * `` -> `...78`
## * `` -> `...79`
## * `` -> `...80`
## * `` -> `...81`
## * `` -> `...82`
## * `` -> `...83`
## * `` -> `...84`
## * `` -> `...85`
## * `` -> `...86`
## * `` -> `...87`
## * `` -> `...88`
## * `` -> `...89`
## * `` -> `...90`
## * `` -> `...91`
## * `` -> `...92`
## * `` -> `...93`
## * `` -> `...94`
## * `` -> `...95`
## * `` -> `...96`
## * `` -> `...97`
## * `` -> `...98`
## * `` -> `...99`
## * `` -> `...100`
## * `` -> `...101`
## * `` -> `...102`
## * `` -> `...103`
## * `` -> `...104`
## * `` -> `...105`
## * `` -> `...106`
## * `` -> `...107`
## * `` -> `...108`
## * `` -> `...109`
## * `` -> `...110`
## * `` -> `...111`
## * `` -> `...112`
## * `` -> `...113`
## * `` -> `...114`
## * `` -> `...115`
## * `` -> `...116`
```

```
## * `` -> `...117`
## * `` -> `...118`
## * `` -> `...119`
## * `` -> `...120`
## * `` -> `...121`
## * `` -> `...122`
## * `` -> `...123`
## * `` -> `...124`
## * `` -> `...125`
## * `` -> `...126`
## * `` -> `...127`
## * `` -> `...128`
## * `` -> `...129`
## * `` -> `...130`
## * `` -> `...131`
## * `` -> `...132`
## * `` -> `...133`
## * `` -> `...134`
## * `` -> `...135`
## * `` -> `...136`
## * `` -> `...137`
## * `` -> `...138`
## * `` -> `...139`
## * `` -> `...140`
## * `` -> `...141`
## * `` -> `...142`
## * `` -> `...143`
## * `` -> `...144`
## * `` -> `...145`
## * `` -> `...146`
## * `` -> `...147`
## * `` -> `...148`
## * `` -> `...149`
## * `` -> `...150`
## * `` -> `...151`
## * `` -> `...152`
## * `` -> `...153`
## * `` -> `...154`
## * `` -> `...155`
## * `` -> `...156`
## * `` -> `...157`
## * `` -> `...158`
## * `` -> `...159`
## * `` -> `...160`
## * `` -> `...161`
## * `` -> `...162`
## * `` -> `...163`
## * `` -> `...164`
## * `` -> `...165`
## * `` -> `...166`
## * `` -> `...167`
## * `` -> `...168`
## * `` -> `...169`
## * `` -> `...170`
```

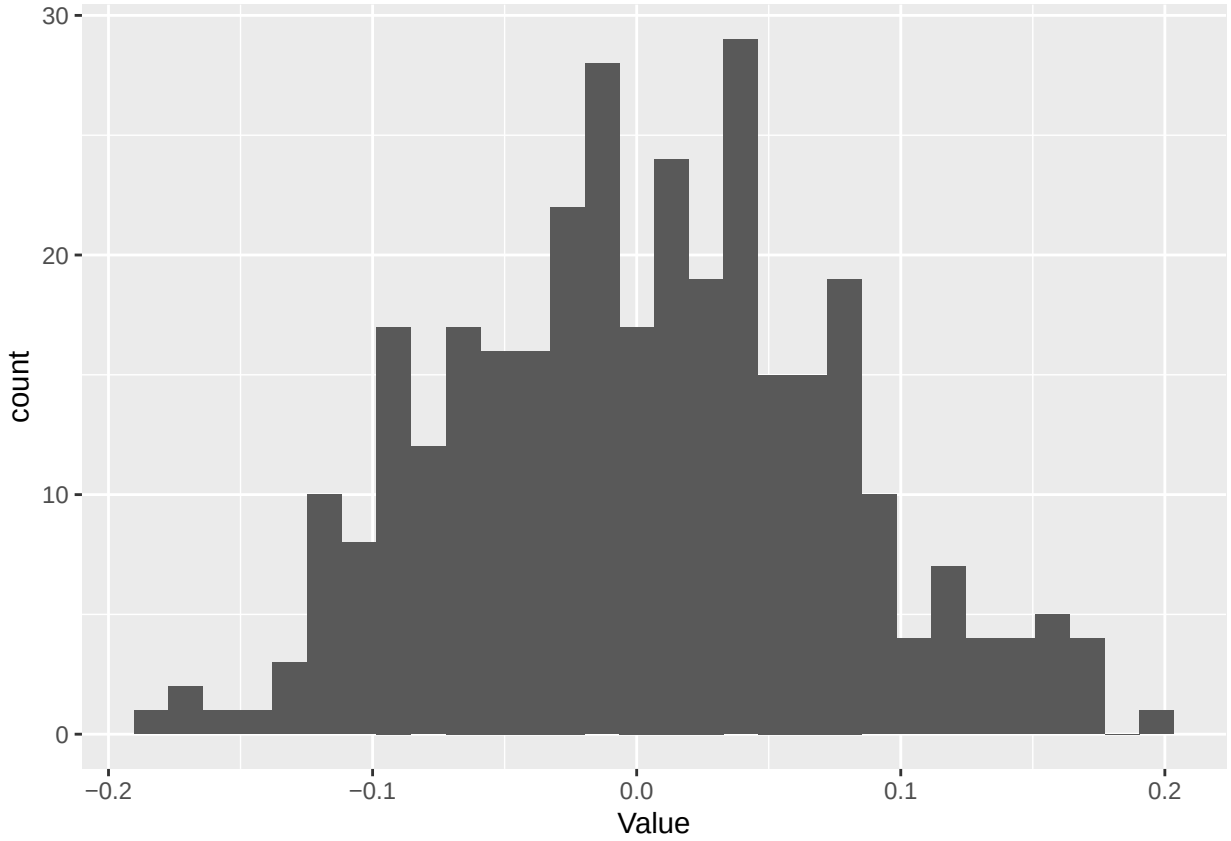
```
## * `` -> `...171`
## * `` -> `...172`
## * `` -> `...173`
## * `` -> `...174`
## * `` -> `...175`
## * `` -> `...176`
## * `` -> `...177`
## * `` -> `...178`
## * `` -> `...179`
## * `` -> `...180`
## * `` -> `...181`
## * `` -> `...182`
## * `` -> `...183`
## * `` -> `...184`
## * `` -> `...185`
## * `` -> `...186`
## * `` -> `...187`
## * `` -> `...188`
## * `` -> `...189`
## * `` -> `...190`
## * `` -> `...191`
## * `` -> `...192`
## * `` -> `...193`
## * `` -> `...194`
## * `` -> `...195`
## * `` -> `...196`
## * `` -> `...197`
## * `` -> `...198`
## * `` -> `...199`
## * `` -> `...200`
## * `` -> `...201`
## * `` -> `...202`
## * `` -> `...203`
## * `` -> `...204`
## * `` -> `...205`
## * `` -> `...206`
## * `` -> `...207`
## * `` -> `...208`
## * `` -> `...209`
## * `` -> `...210`
## * `` -> `...211`
## * `` -> `...212`
## * `` -> `...213`
## * `` -> `...214`
## * `` -> `...215`
## * `` -> `...216`
## * `` -> `...217`
## * `` -> `...218`
## * `` -> `...219`
## * `` -> `...220`
## * `` -> `...221`
## * `` -> `...222`
## * `` -> `...223`
## * `` -> `...224`
```

```
## * `` -> `...225`
## * `` -> `...226`
## * `` -> `...227`
## * `` -> `...228`
## * `` -> `...229`
## * `` -> `...230`
## * `` -> `...231`
## * `` -> `...232`
## * `` -> `...233`
## * `` -> `...234`
## * `` -> `...235`
## * `` -> `...236`
## * `` -> `...237`
## * `` -> `...238`
## * `` -> `...239`
## * `` -> `...240`
## * `` -> `...241`
## * `` -> `...242`
## * `` -> `...243`
## * `` -> `...244`
## * `` -> `...245`
## * `` -> `...246`
## * `` -> `...247`
## * `` -> `...248`
## * `` -> `...249`
## * `` -> `...250`
## * `` -> `...251`
## * `` -> `...252`
## * `` -> `...253`
## * `` -> `...254`
## * `` -> `...255`
## * `` -> `...256`
## * `` -> `...257`
## * `` -> `...258`
## * `` -> `...259`
## * `` -> `...260`
## * `` -> `...261`
## * `` -> `...262`
## * `` -> `...263`
## * `` -> `...264`
## * `` -> `...265`
## * `` -> `...266`
## * `` -> `...267`
## * `` -> `...268`
## * `` -> `...269`
## * `` -> `...270`
## * `` -> `...271`
## * `` -> `...272`
## * `` -> `...273`
## * `` -> `...274`
## * `` -> `...275`
## * `` -> `...276`
## * `` -> `...277`
## * `` -> `...278`
```

```
## * `` -> `...279`
## * `` -> `...280`
## * `` -> `...281`
## * `` -> `...282`
## * `` -> `...283`
## * `` -> `...284`
## * `` -> `...285`
## * `` -> `...286`
## * `` -> `...287`
## * `` -> `...288`
## * `` -> `...289`
## * `` -> `...290`
## * `` -> `...291`
## * `` -> `...292`
## * `` -> `...293`
## * `` -> `...294`
## * `` -> `...295`
## * `` -> `...296`
## * `` -> `...297`
## * `` -> `...298`
## * `` -> `...299`
## * `` -> `...300`
## * `` -> `...301`
## * `` -> `...302`
## * `` -> `...303`
## * `` -> `...304`
## * `` -> `...305`
## * `` -> `...306`
## * `` -> `...307`
## * `` -> `...308`
## * `` -> `...309`
## * `` -> `...310`
## * `` -> `...311`
## * `` -> `...312`
## * `` -> `...313`
## * `` -> `...314`
## * `` -> `...315`
## * `` -> `...316`
## * `` -> `...317`
## * `` -> `...318`
## * `` -> `...319`
## * `` -> `...320`
## * `` -> `...321`
## * `` -> `...322`
## * `` -> `...323`
## * `` -> `...324`
## * `` -> `...325`
## * `` -> `...326`
## * `` -> `...327`
## * `` -> `...328`
## * `` -> `...329`
## * `` -> `...330`
## * `` -> `...331`
## * `` -> `...332`
```

```
mrna_corr %>%
 pivot_longer(cols = -overall_survival, names_to = "Variable", values_to = "Value") %>%
 select(-overall_survival) %>%
 ggplot(aes(x=Value)) +
 geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The plot shows most of the mRNA carry a weak correlation with survival outcomes ranging from 0.19 to –0.15.

```
mrna_corr_arranged <- mrna_corr %>%
 pivot_longer(cols = -overall_survival, names_to = "Variable", values_to = "Value") %>%
 select(-overall_survival) %>%
 arrange(desc(Value))
```

```
mrna_corr_arranged %>% head(5)
```

```
## # A tibble: 5 x 2
##   Variable Value
##   <chr>    <dbl>
## 1 jak1     0.194
## 2 casp8    0.168
## 3 tgfbr2   0.166
## 4 abcb1    0.165
## 5 kit      0.164
```

```
mrna_corr_arranged %>% tail(5)
```

```
## # A tibble: 5 x 2
##   Variable  Value
##   <chr>     <dbl>
## 1 pdgfb    -0.145
## 2 tsc2     -0.162
## 3 map4     -0.165
## 4 kmt2c    -0.172
## 5 gsk3b    -0.186
```

Data cleaning

The missing values can be handled as

`Dropping the rows with the missing values`

`Dropping the column containing the missing values`

`Replacing the missing values as new category as "missing value" as analysis`

`Impute the values with mean, mode or advanced algorithms like K-nearest neighbour (KNN) imputation`

Due to limited computational resources and time for model fitting, a selected number of variables are chosen for model fitting. 1. Age, number of postive lymph nodes, nottigam grade, tumor size. tumpr stage is dropped due to high number of missing vaules 2. various types of treatments. It should be noted that correlation may not be causation with regards to treatment and outcome 3. Most postively correlated and most negatively correlated gene expression

~~~~~~~##

1. Imputing type of surgery with most common type of surgery
2. Imputing the tumor size with mean tumor size

Rest of the varibles do not have missing values

```
metabric_RNA_data %>%
  count(type_of_breast_surgery)
```

```
## # Source: spark<?> [?? x 2]
##   type_of_breast_surgery     n
##   <chr>                  <dbl>
## 1 MASTECTOMY              1127
## 2 <NA>                      22
## 3 BREAST CONSERVING        755
```

```
metabric_RNA_data %>%
  summarise(mean_size = mean(tumor_size))
```

```
## # Source: spark<?> [?? x 1]
##   mean_size
##       <dbl>
## 1      26.2
```

Imputation

```
metabric_RNA_data <- metabric_RNA_data %>%
  mutate(tumor_size = ifelse(is.na(tumor_size), 26.23, tumor_size)) %>%
  mutate(type_of_breast_surgery = ifelse(is.na(type_of_breast_surgery), "MASTECTOMY", type_of_breast_su
```
~~~~~~~

Creating a table containing only columns of intrest

```
metabric_reduced <- metabric_RNA_data %>%
  select(overall_survival, age_at_diagnosis, lymph_nodes_examined_positive, nottingham_prognostic_index
         type_of_breast_surgery, hormone_therapy, chemotherapy, radio_therapy, jak1, pdgfb) %>%
  mutate(surgery_mastectomy = ifelse(type_of_breast_surgery == "MASTECTOMY", 1, 0)) %>%
  select(-type_of_breast_surgery)
```

```
metabric_reduced %>% glimpse()
```

```
## Rows: ??
## Columns: 11
## Database: spark_connection
## $ overall_survival              <int> 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0~
## $ age_at_diagnosis              <dbl> 75.65, 43.19, 48.87, 47.68, 76.97, 78.77~
## $ lymph_nodes_examined_positive <dbl> 10, 0, 1, 3, 8, 0, 1, 1, 1, 0, 0, 0, 3, ~
## $ nottingham_prognostic_index   <dbl> 6.044, 4.020, 4.030, 4.050, 6.080, 4.062~
## $ tumor_size                    <dbl> 22, 10, 15, 25, 40, 31, 10, 29, 16, 28, ~
## $ hormone_therapy               <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1~
## $ chemotherapy                  <int> 0, 0, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0~
## $ radio_therapy                 <int> 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1~
## $ jak1                          <dbl> 1.1097, 0.9804, 1.5835, 0.6194, 0.0461, ~
## $ pdgfb                         <dbl> -0.0349, -0.3739, -1.6093, -0.4251, -1.0~
## $ surgery_mastectomy            <dbl> 1, 0, 1, 1, 1, 1, 0, 0, 0, 1, 1, 0, 1, 1~
```

```
gen_linear_model <- ml_generalized_linear_regression(metabric_reduced,
                                    overall_survival ~ age_at_diagnosis + lymph_nodes_examined_positi
                                    family = "binomial")

gen_linear_model <- tidy(gen_linear_model)
```

```
# metabric_RNA_data <- metabric_RNA_data %>%
#   select(-patient_id, -tumor_stage, -'3gene_classifier_subtype', -primary_tumor_laterality, -overall_
#   mutate(neoplasm_histologic_grade = ifelse(is.na(neoplasm_histologic_grade), "missing", neoplasm_his
#   mutate(cellularity = ifelse(is.na(cellularity), "missing", cellularity)) %>%
#   mutate(mutation_count = ifelse(is.na(mutation_count), "missing", mutation_count)) %>%
#   mutate(er_status_measured_by_ihc = ifelse(is.na(er_status_measured_by_ihc), "missing", er_status_me
#   mutate(type_of_breast_surgery = ifelse(is.na(type_of_breast_surgery), "missing", type_of_breast_sur
#   mutate(tumor_size = ifelse(is.na(tumor_size), "missing", tumor_size)) %>%
#   mutate(cancer_type_detailed = ifelse(is.na(cancer_type_detailed), "missing", cancer_type_detailed))
#   mutate(tumor_other_histologic_subtype = ifelse(is.na(tumor_other_histologic_subtype), "missing", tu
#   mutate(oncotree_code = ifelse(is.na(oncotree_code), "missing", oncotree_code)) %>%
#   mutate(death_from_cancer = ifelse(is.na(death_from_cancer), "missing", death_from_cancer))
```

```
# metabric_RNA_data %>% distinct(type_of_breast_surgery) to get unique values of

# metabric_RNA_data_reduced <- metabric_RNA_data %>%
#   select(overall_survival, age_at_diagnosis, lymph_nodes_examined_positive, nottingham_prognostic_ind
#   mutate(type_of_breast_surgery = ifelse(is.na(type_of_breast_surgery), "missing", type_of_breast_sur
#   mutate(surgery_mastectomy = ifelse(type_of_breast_surgery == "MASTECTOMY", 1, 0),
#          surgery_bcs = ifelse(type_of_breast_surgery == "BREAST CONSERVING", 1, 0),
#          surgery_missing = ifelse(type_of_breast_surgery == "missing", 1, 0),)
# ```
# ```{r}
# metabric_RNA_data_reduced %>% glimpse()
```

model fitting

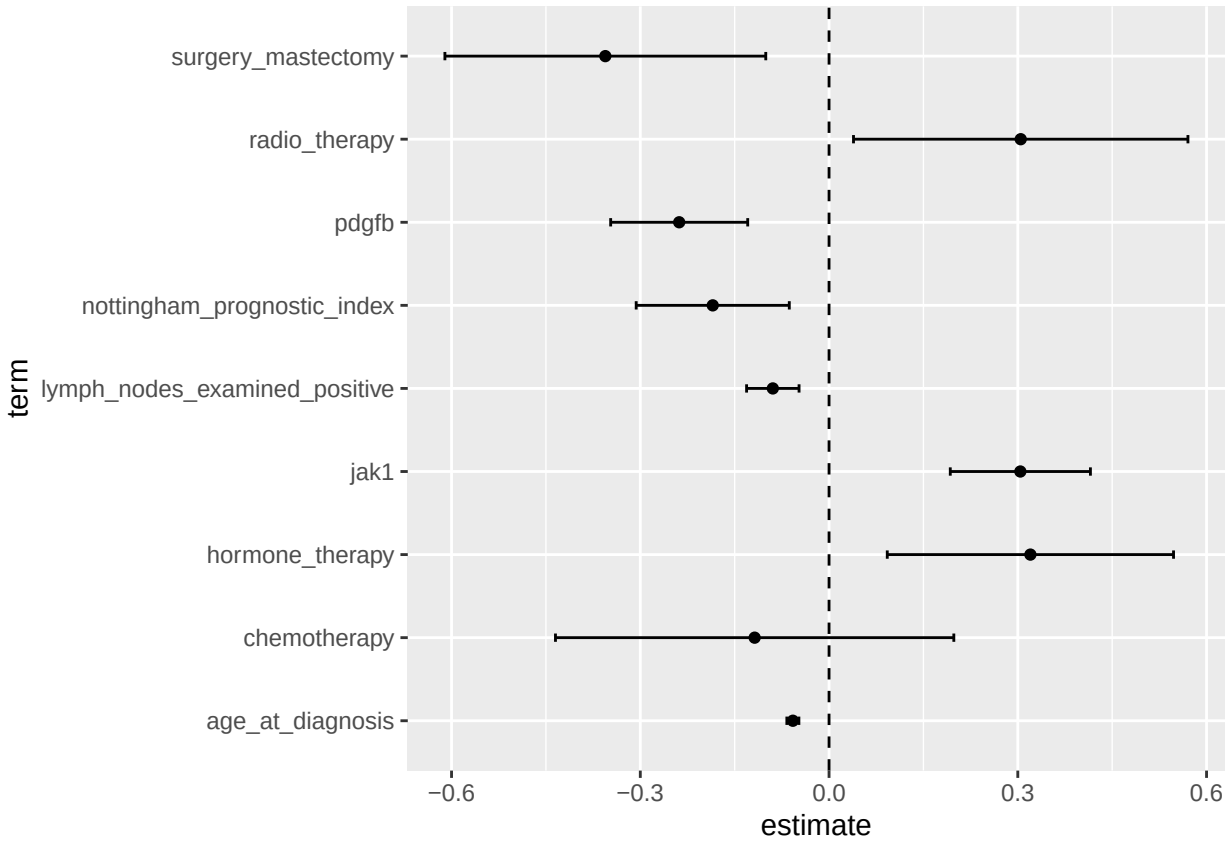
```
# linear_model <- ml_logistic_regression(
#   metabric_RNA_data_reduced, overall_survival ~ age_at_diagnosis + lymph_nodes_examined_positive + no
#   fit_intercept = FALSE)
# linear_model
```

```
# validation_summary <- ml_evaluate(linear_model, metabric_RNA_data_reduced)
```

```
# validation_summary
# ```
# ```{r}
# roc <- validation_summary$roc() %>%
#   collect()
#
# ggplot(roc, aes(x = FPR, y = TPR)) +
#   geom_line() + geom_abline(lty = "dashed")
```

```
# gen_linear_model <- ml_generalized_linear_regression(
#   metabric_RNA_data_reduced,
#   overall_survival ~ age_at_diagnosis + lymph_nodes_examined_positive + nottingham_prognostic_index +
#   family = "binomial"
# )
#
# gen_linear_model <- tidy(gen_linear_model)
```

```
gen_linear_model %>%
  slice_tail(n=-1) %>%
  ggplot(aes(x = term, y = estimate)) +
  geom_point() +
  geom_errorbar(
    aes(ymin = estimate - 1.96 * std.error,
        ymax = estimate + 1.96 * std.error, width = .1)
  ) +
  coord_flip() +
  geom_hline(yintercept = 0, linetype = "dashed")
```




```
spark_disconnect(sc)
```