

TCGA Clinical Data Analysis

Preparing clinical data to analysis following Hadley Wickham's Tidyverse manifesto.

1. Importing data

```
kirc_clin_raw <- read_delim("data/kirc_tcga_clinical_data.tsv", "\t",
                           escape_double = FALSE,
                           trim_ws = TRUE)

## Parsed with column specification:
## cols(
##   .default = col_character(),
##   `Diagnosis Age` = col_double(),
##   `Neoplasm American Joint Committee on Cancer Clinical Group Stage` = col_logical(),
##   `Neoplasm American Joint Committee on Cancer Clinical Regional Lymph Node N Stage` = col_logical(),
##   `Neoplasm American Joint Committee on Cancer Clinical Primary Tumor T Stage` = col_logical(),
##   `Days to Sample Collection.` = col_double(),
##   `Last Alive Less Initial Pathologic Diagnosis Date Calculated Day Value` = col_double(),
##   `Days to Sample Procurement` = col_logical(),
##   `Disease Free (Months)` = col_double(),
##   `Performance Status` = col_double(),
##   `Lymphomatous Extranodal Site Involvement Indicator` = col_logical(),
##   `Fraction Genome Altered` = col_double(),
##   `Year Cancer Initial Diagnosis` = col_double(),
##   `Karnofsky Performance Score` = col_double(),
##   `Longest Dimension` = col_double(),
##   `Lymph nodes examined positive` = col_double(),
##   `Lymph Node(s) Examined Number` = col_double(),
##   `First Pathologic Diagnosis Biospecimen Acquisition Method Type` = col_logical(),
##   `Mutation Count` = col_double(),
##   `Oct embedded` = col_logical(),
##   `Overall Survival (Months)` = col_double()
##   # ... with 15 more columns
## )

## See spec(...) for full column specifications.

class(kirc_clin_raw)

## [1] "spec_tbl_df" "tbl_df"      "tbl"         "data.frame"

dim(kirc_clin_raw)

## [1] 538  89

names(kirc_clin_raw)

## [1] "Study ID"
## [2] "Patient ID"
```

```

## [3] "Sample ID"
## [4] "Diagnosis Age"
## [5] "Neoplasm American Joint Committee on Cancer Clinical Distant Metastasis M Stage"
## [6] "American Joint Committee on Cancer Metastasis Stage Code"
## [7] "Neoplasm Disease Lymph Node Stage American Joint Committee on Cancer Code"
## [8] "Neoplasm Disease Stage American Joint Committee on Cancer Code"
## [9] "American Joint Committee on Cancer Publication Version Type"
## [10] "American Joint Committee on Cancer Tumor Stage Code"
## [11] "Cancer Type"
## [12] "Cancer Type Detailed"
## [13] "Neoplasm American Joint Committee on Cancer Clinical Group Stage"
## [14] "Neoplasm American Joint Committee on Cancer Clinical Regional Lymph Node N Stage"
## [15] "Neoplasm American Joint Committee on Cancer Clinical Primary Tumor T Stage"
## [16] "Days to Sample Collection."
## [17] "Last Alive Less Initial Pathologic Diagnosis Date Calculated Day Value"
## [18] "Days to Sample Procurement"
## [19] "Disease Free (Months)"
## [20] "Disease Free Status"
## [21] "Disease code"
## [22] "Performance Status"
## [23] "Erythrocyte sedimentation rate"
## [24] "Ethnicity Category"
## [25] "Lymphomatous Extranodal Site Involvement Indicator"
## [26] "Form completion date"
## [27] "Fraction Genome Altered"
## [28] "Neoplasm Histologic Grade"
## [29] "Hemoglobin level"
## [30] "Neoplasm Histologic Type Name"
## [31] "Neoadjuvant Therapy Type Administered Prior To Resection Text"
## [32] "Prior Cancer Diagnosis Occurrence"
## [33] "ICD-10 Classification"
## [34] "International Classification of Diseases for Oncology, Third Edition ICD-O-3 Histology Code"
## [35] "International Classification of Diseases for Oncology, Third Edition ICD-O-3 Site Code"
## [36] "Idh level"
## [37] "Informed consent verified"
## [38] "Year Cancer Initial Diagnosis"
## [39] "Is FFPE"
## [40] "Karnofsky Performance Score"
## [41] "Primary Tumor Laterality"
## [42] "Longest Dimension"
## [43] "Primary Lymph Node Presentation Assessment Ind-3"
## [44] "Lymph nodes examined positive"
## [45] "Lymph Node(s) Examined Number"
## [46] "First Pathologic Diagnosis Biospecimen Acquisition Method Type"
## [47] "Mutation Count"
## [48] "New Neoplasm Event Post Initial Therapy Indicator"
## [49] "Oct embedded"
## [50] "Oncotree Code"
## [51] "Overall Survival (Months)"
## [52] "Overall Survival Status"
## [53] "Specimen Collection Method"
## [54] "Other Patient ID"
## [55] "Other Sample ID"
## [56] "Pathology Report File Name"

```

```
## [57] "Pathology report uuid"
## [58] "Performance Status Assessment Timepoint Category"
## [59] "Platelet count"
## [60] "Project code"
## [61] "Tissue Prospective Collection Indicator"
## [62] "Race Category"
## [63] "Did patient start adjuvant postoperative radiotherapy?"
## [64] "Tissue Retrospective Collection Indicator"
## [65] "Number of Samples Per Patient"
## [66] "Sample Initial Weight"
## [67] "Sample Type"
## [68] "Sample type id"
## [69] "Serum calcium level"
## [70] "Sex"
## [71] "Shortest Dimension"
## [72] "Tumor Tissue Site"
## [73] "Person Cigarette Smoking History Pack Year Value"
## [74] "Started Smoking Year"
## [75] "Stopped Smoking Year"
## [76] "Specimen Current Weight"
## [77] "Specimen Freezing Means"
## [78] "Specimen Second Longest Dimension"
## [79] "Stage Other"
## [80] "Adjuvant Postoperative Targeted Therapy Administered Indicator"
## [81] "Time between clamping and freezing"
## [82] "Time between excision and freezing"
## [83] "Tissue Source Site"
## [84] "Patient Smoking History Category"
## [85] "Primary Therapy Outcome Success Type"
## [86] "Person Neoplasm Status"
## [87] "Vial number"
## [88] "Patient's Vital Status"
## [89] "WBC"
```

```
glimpse(kirc_clin_raw)
```

```
## Rows: 538
## Columns: 89
## $ `Study ID`
## $ `Patient ID`
## $ `Sample ID`
## $ `Diagnosis Age`
## $ `Neoplasm American Joint Committee on Cancer Clinical Distant Metastasis M Stage`
## $ `American Joint Committee on Cancer Metastasis Stage Code`
## $ `Neoplasm Disease Lymph Node Stage American Joint Committee on Cancer Code`
## $ `Neoplasm Disease Stage American Joint Committee on Cancer Code`
## $ `American Joint Committee on Cancer Publication Version Type`
## $ `American Joint Committee on Cancer Tumor Stage Code`
## $ `Cancer Type`
## $ `Cancer Type Detailed`
## $ `Neoplasm American Joint Committee on Cancer Clinical Group Stage`
## $ `Neoplasm American Joint Committee on Cancer Clinical Regional Lymph Node N Stage`
## $ `Neoplasm American Joint Committee on Cancer Clinical Primary Tumor T Stage`
## $ `Days to Sample Collection.`
## $ `Last Alive Less Initial Pathologic Diagnosis Date Calculated Day Value`
```

## \$ `Days to Sample Procurement`	<lg1>
## \$ `Disease Free (Months)`	<dbl>
## \$ `Disease Free Status`	<chr>
## \$ `Disease code`	<chr>
## \$ `Performance Status`	<dbl>
## \$ `Erythrocyte sedimentation rate`	<chr>
## \$ `Ethnicity Category`	<chr>
## \$ `Lymphomatous Extranodal Site Involvement Indicator`	<lg1>
## \$ `Form completion date`	<chr>
## \$ `Fraction Genome Altered`	<dbl>
## \$ `Neoplasm Histologic Grade`	<chr>
## \$ `Hemoglobin level`	<chr>
## \$ `Neoplasm Histologic Type Name`	<chr>
## \$ `Neoadjuvant Therapy Type Administered Prior To Resection Text`	<chr>
## \$ `Prior Cancer Diagnosis Occurence`	<chr>
## \$ `ICD-10 Classification`	<chr>
## \$ `International Classification of Diseases for Oncology, Third Edition ICD-O-3 Histology Code`	<chr>
## \$ `International Classification of Diseases for Oncology, Third Edition ICD-O-3 Site Code`	<chr>
## \$ `Idh level`	<chr>
## \$ `Informed consent verified`	<chr>
## \$ `Year Cancer Initial Diagnosis`	<dbl>
## \$ `Is FFPE`	<chr>
## \$ `Karnofsky Performance Score`	<dbl>
## \$ `Primary Tumor Laterality`	<chr>
## \$ `Longest Dimension`	<dbl>
## \$ `Primary Lymph Node Presentation Assessment Ind-3`	<chr>
## \$ `Lymph nodes examined positive`	<dbl>
## \$ `Lymph Node(s) Examined Number`	<dbl>
## \$ `First Pathologic Diagnosis Biospecimen Acquisition Method Type`	<lg1>
## \$ `Mutation Count`	<dbl>
## \$ `New Neoplasm Event Post Initial Therapy Indicator`	<chr>
## \$ `Oct embedded`	<lg1>
## \$ `Oncotree Code`	<chr>
## \$ `Overall Survival (Months)`	<dbl>
## \$ `Overall Survival Status`	<chr>
## \$ `Specimen Collection Method`	<lg1>
## \$ `Other Patient ID`	<chr>
## \$ `Other Sample ID`	<chr>
## \$ `Pathology Report File Name`	<chr>
## \$ `Pathology report uuid`	<chr>
## \$ `Performance Status Assessment Timepoint Category`	<chr>
## \$ `Platelet count`	<chr>
## \$ `Project code`	<chr>
## \$ `Tissue Prospective Collection Indicator`	<chr>
## \$ `Race Category`	<chr>
## \$ `Did patient start adjuvant postoperative radiotherapy?`	<chr>
## \$ `Tissue Retrospective Collection Indicator`	<chr>
## \$ `Number of Samples Per Patient`	<dbl>
## \$ `Sample Initial Weight`	<dbl>
## \$ `Sample Type`	<chr>
## \$ `Sample type id`	<dbl>
## \$ `Serum calcium level`	<chr>
## \$ Sex	<chr>
## \$ `Shortest Dimension`	<dbl>

```
## $ `Tumor Tissue Site` <chr>
## $ `Person Cigarette Smoking History Pack Year Value` <dbl>
## $ `Started Smoking Year` <dbl>
## $ `Stopped Smoking Year` <dbl>
## $ `Specimen Current Weight` <lgl>
## $ `Specimen Freezing Means` <lgl>
## $ `Specimen Second Longest Dimension` <dbl>
## $ `Stage Other` <lgl>
## $ `Adjuvant Postoperative Targeted Therapy Administered Indicator` <chr>
## $ `Time between clamping and freezing` <lgl>
## $ `Time between excision and freezing` <lgl>
## $ `Tissue Source Site` <chr>
## $ `Patient Smoking History Category` <dbl>
## $ `Primary Therapy Outcome Success Type` <chr>
## $ `Person Neoplasm Status` <chr>
## $ `Vial number` <chr>
## $ `Patient's Vital Status` <chr>
## $ WBC <chr>
```

```
skim(kirc_clin_raw)
```

Table 1: Data summary

Name	kirc_clin_raw
Number of rows	538
Number of columns	89
Column type frequency:	
character	54
logical	13
numeric	22
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate
Study ID	0	1.00
Patient ID	0	1.00
Sample ID	0	1.00
Neoplasm American Joint Committee on Cancer Clinical Distant Metastasis M Stage	504	0.06
American Joint Committee on Cancer Metastasis Stage Code	2	1.00
Neoplasm Disease Lymph Node Stage American Joint Committee on Cancer Code	0	1.00
Neoplasm Disease Stage American Joint Committee on Cancer Code	3	0.99
American Joint Committee on Cancer Publication Version Type	367	0.32
American Joint Committee on Cancer Tumor Stage Code	0	1.00
Cancer Type	0	1.00
Cancer Type Detailed	0	1.00
Disease Free Status	99	0.82
Disease code	534	0.01
Erythrocyte sedimentation rate	527	0.02
Ethnicity Category	152	0.72
Form completion date	0	1.00

skim_variable	n_missing	complete_rate
Neoplasm Histologic Grade	3	0.99
Hemoglobin level	83	0.85
Neoplasm Histologic Type Name	0	1.00
Neoadjuvant Therapy Type Administered Prior To Resection Text	0	1.00
Prior Cancer Diagnosis Occurence	0	1.00
ICD-10 Classification	0	1.00
International Classification of Diseases for Oncology, Third Edition ICD-O-3 Histology Code	0	1.00
International Classification of Diseases for Oncology, Third Edition ICD-O-3 Site Code	0	1.00
Idh level	451	0.16
Informed consent verified	0	1.00
Is FFPE	1	1.00
Primary Tumor Laterality	0	1.00
Primary Lymph Node Presentation Assessment Ind-3	7	0.99
New Neoplasm Event Post Initial Therapy Indicator	503	0.07
Oncotree Code	0	1.00
Overall Survival Status	0	1.00
Other Patient ID	0	1.00
Other Sample ID	1	1.00
Pathology Report File Name	1	1.00
Pathology report uuid	1	1.00
Performance Status Assessment Timepoint Category	428	0.20
Platelet count	93	0.83
Project code	534	0.01
Tissue Prospective Collection Indicator	20	0.96
Race Category	7	0.99
Did patient start adjuvant postoperative radiotherapy?	506	0.06
Tissue Retrospective Collection Indicator	18	0.97
Sample Type	0	1.00
Serum calcium level	172	0.68
Sex	0	1.00
Tumor Tissue Site	0	1.00
Adjuvant Postoperative Targeted Therapy Administered Indicator	506	0.06
Tissue Source Site	0	1.00
Primary Therapy Outcome Success Type	507	0.06
Person Neoplasm Status	35	0.93
Vial number	1	1.00
Patient's Vital Status	3	0.99
WBC	96	0.82

Variable type: logical

skim_variable	n_missing	complete_rate	m
Neoplasm American Joint Committee on Cancer Clinical Group Stage	538	0.00	N
Neoplasm American Joint Committee on Cancer Clinical Regional Lymph Node N Stage	538	0.00	N
Neoplasm American Joint Committee on Cancer Clinical Primary Tumor T Stage	538	0.00	N
Days to Sample Procurement	538	0.00	N
Lymphomatous Extranodal Site Involvement Indicator	538	0.00	N
First Pathologic Diagnosis Biospecimen Acquisition Method Type	538	0.00	N
Oct embedded	503	0.07	C
Specimen Collection Method	538	0.00	N
Specimen Current Weight	538	0.00	N

skim_variable	n_missing	complete_rate	m
Specimen Freezing Means	538	0.00	N
Stage Other	538	0.00	N
Time between clamping and freezing	538	0.00	N
Time between excision and freezing	538	0.00	N

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd
Diagnosis Age	0	1.00	60.58	12.14
Days to Sample Collection.	503	0.07	545.26	566.08
Last Alive Less Initial Pathologic Diagnosis Date Calculated Day Value	0	1.00	0.00	0.00
Disease Free (Months)	99	0.82	40.30	31.64
Performance Status	399	0.26	0.24	0.48
Fraction Genome Altered	9	0.98	0.17	0.17
Year Cancer Initial Diagnosis	0	1.00	2006.02	2.76
Karnofsky Performance Score	490	0.09	88.33	20.56
Longest Dimension	36	0.93	1.66	0.66
Lymph nodes examined positive	407	0.24	0.34	1.44
Lymph Node(s) Examined Number	404	0.25	6.30	5.97
Mutation Count	87	0.84	73.85	127.76
Overall Survival (Months)	0	1.00	44.33	32.26
Number of Samples Per Patient	0	1.00	1.00	0.06
Sample Initial Weight	503	0.07	296.29	366.21
Sample type id	0	1.00	1.01	0.17
Shortest Dimension	36	0.93	0.38	0.21
Person Cigarette Smoking History Pack Year Value	516	0.04	28.55	15.77
Started Smoking Year	525	0.02	1978.38	17.35
Stopped Smoking Year	525	0.02	1994.77	15.12
Specimen Second Longest Dimension	36	0.93	0.94	0.31
Patient Smoking History Category	450	0.16	1.91	1.19

```
#View(kirc_clin_raw)
```

2. Cleaning data

Select variables based on NA count (> 50% complete is a good choice!).

```
NA_fifty <- dim(kirc_clin_raw)[1]/2

NA_sum <- colSums(is.na(kirc_clin_raw))
NA_sum <- as.data.frame(NA_sum)
NA_sum <- tibble::rownames_to_column(NA_sum, "variables")
NA_sum <- NA_sum %>%
  filter(NA_sum < NA_fifty)

kirc_clean <- kirc_clin_raw %>%
  select(one_of(NA_sum$variables))
```

Remove duplicate observations:

```
kirc_clean0 <- kirc_clean %>%
  distinct_at('Patient ID', .keep_all = TRUE)
```

Remove numeric variables with unique observations:

```
skim(kirc_clean0)
```

Table 5: Data summary

Name	kirc_clean0
Number of rows	537
Number of columns	55
Column type frequency:	
character	43
numeric	12
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate
Study ID	0	1.00
Patient ID	0	1.00
Sample ID	0	1.00
American Joint Committee on Cancer Metastasis Stage Code	2	1.00
Neoplasm Disease Lymph Node Stage American Joint Committee on Cancer Code	0	1.00
Neoplasm Disease Stage American Joint Committee on Cancer Code	3	0.99
American Joint Committee on Cancer Tumor Stage Code	0	1.00
Cancer Type	0	1.00
Cancer Type Detailed	0	1.00
Disease Free Status	99	0.82
Ethnicity Category	152	0.72
Form completion date	0	1.00
Neoplasm Histologic Grade	3	0.99
Hemoglobin level	83	0.85
Neoplasm Histologic Type Name	0	1.00
Neoadjuvant Therapy Type Administered Prior To Resection Text	0	1.00
Prior Cancer Diagnosis Occurence	0	1.00
ICD-10 Classification	0	1.00
International Classification of Diseases for Oncology, Third Edition ICD-O-3 Histology Code	0	1.00
International Classification of Diseases for Oncology, Third Edition ICD-O-3 Site Code	0	1.00
Informed consent verified	0	1.00
Is FFPE	0	1.00
Primary Tumor Laterality	0	1.00
Primary Lymph Node Presentation Assessment Ind-3	7	0.99
Oncotree Code	0	1.00
Overall Survival Status	0	1.00
Other Patient ID	0	1.00
Other Sample ID	0	1.00
Pathology Report File Name	0	1.00
Pathology report uuid	0	1.00
Platelet count	93	0.83

skim_variable	n_missing	complete_rate
Tissue Prospective Collection Indicator	20	0.96
Race Category	7	0.99
Tissue Retrospective Collection Indicator	18	0.97
Sample Type	0	1.00
Serum calcium level	172	0.68
Sex	0	1.00
Tumor Tissue Site	0	1.00
Tissue Source Site	0	1.00
Person Neoplasm Status	35	0.93
Vial number	0	1.00
Patient's Vital Status	3	0.99
WBC	96	0.82

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd
Diagnosis Age	0	1.00	60.59	12.15
Last Alive Less Initial Pathologic Diagnosis Date Calculated Day Value	0	1.00	0.00	0.00
Disease Free (Months)	99	0.82	40.24	31.66
Fraction Genome Altered	9	0.98	0.17	0.17
Year Cancer Initial Diagnosis	0	1.00	2006.02	2.76
Longest Dimension	35	0.93	1.66	0.66
Mutation Count	86	0.84	73.85	127.76
Overall Survival (Months)	0	1.00	44.26	32.25
Number of Samples Per Patient	0	1.00	1.00	0.04
Sample type id	0	1.00	1.00	0.00
Shortest Dimension	35	0.93	0.38	0.21
Specimen Second Longest Dimension	35	0.93	0.94	0.31

```
kirc_clean1 <- kirc_clean0 %>%
  select(!c('Last Alive Less Initial Pathologic Diagnosis Date Calculated Day Value',
            'Number of Samples Per Patient',
            'Sample type id'))
```

Remove character variables with unique observations:

```
skim(kirc_clean1)
```

Table 8: Data summary

Name	kirc_clean1
Number of rows	537
Number of columns	52
Column type frequency:	
character	43
numeric	9
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate
Study ID	0	1.00
Patient ID	0	1.00
Sample ID	0	1.00
American Joint Committee on Cancer Metastasis Stage Code	2	1.00
Neoplasm Disease Lymph Node Stage American Joint Committee on Cancer Code	0	1.00
Neoplasm Disease Stage American Joint Committee on Cancer Code	3	0.99
American Joint Committee on Cancer Tumor Stage Code	0	1.00
Cancer Type	0	1.00
Cancer Type Detailed	0	1.00
Disease Free Status	99	0.82
Ethnicity Category	152	0.72
Form completion date	0	1.00
Neoplasm Histologic Grade	3	0.99
Hemoglobin level	83	0.85
Neoplasm Histologic Type Name	0	1.00
Neoadjuvant Therapy Type Administered Prior To Resection Text	0	1.00
Prior Cancer Diagnosis Occurence	0	1.00
ICD-10 Classification	0	1.00
International Classification of Diseases for Oncology, Third Edition ICD-O-3 Histology Code	0	1.00
International Classification of Diseases for Oncology, Third Edition ICD-O-3 Site Code	0	1.00
Informed consent verified	0	1.00
Is FFPE	0	1.00
Primary Tumor Laterality	0	1.00
Primary Lymph Node Presentation Assessment Ind-3	7	0.99
Oncotree Code	0	1.00
Overall Survival Status	0	1.00
Other Patient ID	0	1.00
Other Sample ID	0	1.00
Pathology Report File Name	0	1.00
Pathology report uuid	0	1.00
Platelet count	93	0.83
Tissue Prospective Collection Indicator	20	0.96
Race Category	7	0.99
Tissue Retrospective Collection Indicator	18	0.97
Sample Type	0	1.00
Serum calcium level	172	0.68
Sex	0	1.00
Tumor Tissue Site	0	1.00
Tissue Source Site	0	1.00
Person Neoplasm Status	35	0.93
Vial number	0	1.00
Patient's Vital Status	3	0.99
WBC	96	0.82

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
Diagnosis Age	0	1.00	60.59	12.15	26.00	52.00	61.00	70.00	79.00
Disease Free (Months)	99	0.82	40.24	31.66	-11.79	13.43	36.20	61.00	79.00

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
Fraction Genome Altered	9	0.98	0.17	0.17	0.00	0.06	0.12	0.19	0.25
Year Cancer Initial Diagnosis	0	1.00	2006.02	2.76	1998.00	2004.00	2006.00	2007.00	2008.00
Longest Dimension	35	0.93	1.66	0.66	0.40	1.20	1.50	1.80	2.50
Mutation Count	86	0.84	73.85	127.76	1.00	34.00	48.00	63.00	100.00
Overall Survival (Months)	0	1.00	44.26	32.25	0.00	18.10	38.96	63.00	100.00
Shortest Dimension	35	0.93	0.38	0.21	0.10	0.20	0.30	0.40	0.50
Specimen Second Longest Dimension	35	0.93	0.94	0.31	0.30	0.70	0.90	1.20	1.50

```
kirc_clean2 <- kirc_clean1 %>%
  select(!c('Study ID', 'Cancer Type', 'Cancer Type Detailed',
            'Neoplasm Histologic Type Name', 'ICD-10 Classification',
            'International Classification of Diseases for Oncology, Third Edition ICD-O-3 Site Code',
            'Informed consent verified', 'Is FFPE', 'Oncotree Code', 'Sample Type', 'Tumor Tissue Site'))
```

Remove character variables with similar information - check each one!

```
skim(kirc_clean2)
```

Table 11: Data summary

Name	kirc_clean2
Number of rows	537
Number of columns	41
Column type frequency:	
character	32
numeric	9
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate
Patient ID	0	1.00
Sample ID	0	1.00
American Joint Committee on Cancer Metastasis Stage Code	2	1.00
Neoplasm Disease Lymph Node Stage American Joint Committee on Cancer Code	0	1.00
Neoplasm Disease Stage American Joint Committee on Cancer Code	3	0.99
American Joint Committee on Cancer Tumor Stage Code	0	1.00
Disease Free Status	99	0.82
Ethnicity Category	152	0.72
Form completion date	0	1.00
Neoplasm Histologic Grade	3	0.99
Hemoglobin level	83	0.85
Neoadjuvant Therapy Type Administered Prior To Resection Text	0	1.00
Prior Cancer Diagnosis Occurrence	0	1.00
International Classification of Diseases for Oncology, Third Edition ICD-O-3 Histology Code	0	1.00
Primary Tumor Laterality	0	1.00
Primary Lymph Node Presentation Assessment Ind-3	7	0.99
Overall Survival Status	0	1.00
Other Patient ID	0	1.00

skim_variable	n_missing	complete_rate
Other Sample ID	0	1.00
Pathology Report File Name	0	1.00
Pathology report uuid	0	1.00
Platelet count	93	0.83
Tissue Prospective Collection Indicator	20	0.96
Race Category	7	0.99
Tissue Retrospective Collection Indicator	18	0.97
Serum calcium level	172	0.68
Sex	0	1.00
Tissue Source Site	0	1.00
Person Neoplasm Status	35	0.93
Vial number	0	1.00
Patient's Vital Status	3	0.99
WBC	96	0.82

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
Diagnosis Age	0	1.00	60.59	12.15	26.00	52.00	61.00	70.00	79.00
Disease Free (Months)	99	0.82	40.24	31.66	-11.79	13.43	36.20	60.00	66.00
Fraction Genome Altered	9	0.98	0.17	0.17	0.00	0.06	0.12	0.17	0.17
Year Cancer Initial Diagnosis	0	1.00	2006.02	2.76	1998.00	2004.00	2006.00	2007.00	2007.00
Longest Dimension	35	0.93	1.66	0.66	0.40	1.20	1.50	1.66	1.66
Mutation Count	86	0.84	73.85	127.76	1.00	34.00	48.00	63.00	63.00
Overall Survival (Months)	0	1.00	44.26	32.25	0.00	18.10	38.96	63.00	63.00
Shortest Dimension	35	0.93	0.38	0.21	0.10	0.20	0.30	0.38	0.38
Specimen Second Longest Dimension	35	0.93	0.94	0.31	0.30	0.70	0.90	0.94	0.94

```
table(kirc_clean2$`Overall Survival Status`, exclude = NULL)

##
## DECEASED    LIVING
##      177      360

table(kirc_clean2$`Patient's Vital Status`, exclude = NULL)

##
## Alive  Dead  <NA>
##   360   174     3

kirc_clean3 <- kirc_clean2 %>%
  select(!c('Sample ID', 'Overall Survival Status', 'Other Patient ID', 'Other Sample ID',
            'Pathology Report File Name', 'Pathology report uuid'))

# removing other variables not directly related to patient - check each one!
kirc_clean4 <- kirc_clean3 %>%
  select(!c('Form completion date', 'International Classification of Diseases for Oncology, Third Edition'))
```

3. Changing variables names

Using snake_style

```
kirc_clean4 <- kirc_clean4 %>%
  rename(patient_id = 'Patient ID',
         age = 'Diagnosis Age',
         metastasis_stg = 'American Joint Committee on Cancer Metastasis Stage Code',
         neoplasm_ln_stg = 'Neoplasm Disease Lymph Node Stage American Joint Committee on Cancer Code',
         neoplasm_stg = 'Neoplasm Disease Stage American Joint Committee on Cancer Code',
         tumor_stg = 'American Joint Committee on Cancer Tumor Stage Code',
         disease_free_mth = 'Disease Free (Months)',
         disease_free_stt = 'Disease Free Status',
         ethnicity = 'Ethnicity Category',
         frac_genome_alter = 'Fraction Genome Altered',
         histology_grd = 'Neoplasm Histologic Grade',
         hemoglobin = 'Hemoglobin level',
         neoadj_therapy = 'Neoadjuvant Therapy Type Administered Prior To Resection Text',
         prior_cancer = 'Prior Cancer Diagnosis Occurrence',
         year_diagnose = 'Year Cancer Initial Diagnosis',
         tumor_lateral = 'Primary Tumor Laterality',
         long_dim = 'Longest Dimension',
         primer_ln_ind3 = 'Primary Lymph Node Presentation Assessment Ind-3',
         mutation_cnt = 'Mutation Count',
         over_surv_mth = 'Overall Survival (Months)',
         platelet = 'Platelet count',
         tissue_prospect = 'Tissue Prospective Collection Indicator',
         race = 'Race Category',
         tissue_retrospect = 'Tissue Retrospective Collection Indicator',
         serum_ca = 'Serum calcium level',
         sex = 'Sex',
         short_dim = 'Shortest Dimension',
         second_long_dim = 'Specimen Second Longest Dimension',
         tissue_site = 'Tissue Source Site',
         person_neoplasm_stt = 'Person Neoplasm Status',
         vital_stt = "Patient's Vital Status",
         wbc = 'WBC')
```

4. Taming data

Use lubridate for dates

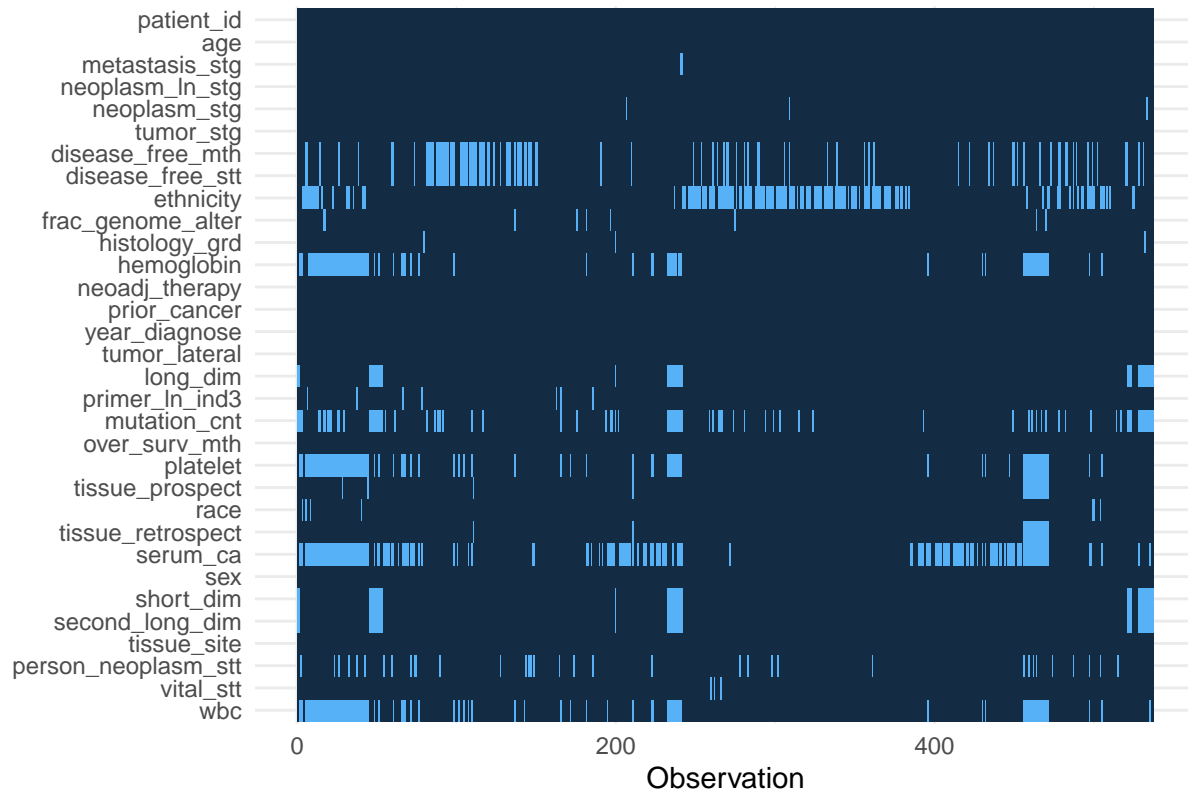
```
kirc_clean4 <- kirc_clean4 %>%
  mutate_if(is.character, as.factor) %>%
  mutate(patient_id = as.character(patient_id))
```

5. Checking NA patterns

Check distincts types of NAs: MCAR, MAR, MNAR

```
kirc_clean4 %>%
  missing_plot()
```

Missing values map



```
missing_glimpse(kirc_clean4)
```

##	label	var_type	n	missing_n	missing_percent
##	patient_id	patient_id	<chr> 537	0	0.0
##	age	age	<dbl> 537	0	0.0
##	metastasis_stg	metastasis_stg	<fct> 535	2	0.4
##	neoplasm_ln_stg	neoplasm_ln_stg	<fct> 537	0	0.0
##	neoplasm_stg	neoplasm_stg	<fct> 534	3	0.6
##	tumor_stg	tumor_stg	<fct> 537	0	0.0
##	disease_free_mth	disease_free_mth	<dbl> 438	99	18.4
##	disease_free_stt	disease_free_stt	<fct> 438	99	18.4
##	ethnicity	ethnicity	<fct> 385	152	28.3
##	frac_genome_alter	frac_genome_alter	<dbl> 528	9	1.7
##	histology_grd	histology_grd	<fct> 534	3	0.6
##	hemoglobin	hemoglobin	<fct> 454	83	15.5
##	neoadj_therapy	neoadj_therapy	<fct> 537	0	0.0
##	prior_cancer	prior_cancer	<fct> 537	0	0.0
##	year_diagnose	year_diagnose	<dbl> 537	0	0.0
##	tumor_lateral	tumor_lateral	<fct> 537	0	0.0
##	long_dim	long_dim	<dbl> 502	35	6.5
##	primer_ln_ind3	primer_ln_ind3	<fct> 530	7	1.3
##	mutation_cnt	mutation_cnt	<dbl> 451	86	16.0
##	over_surv_mth	over_surv_mth	<dbl> 537	0	0.0
##	platelet	platelet	<fct> 444	93	17.3
##	tissue_prospect	tissue_prospect	<fct> 517	20	3.7
##	race	race	<fct> 530	7	1.3
##	tissue_retrospect	tissue_retrospect	<fct> 519	18	3.4

## serum_ca	serum_ca	<fct>	365	172	32.0
## sex	sex	<fct>	537	0	0.0
## short_dim	short_dim	<dbl>	502	35	6.5
## second_long_dim	second_long_dim	<dbl>	502	35	6.5
## tissue_site	tissue_site	<fct>	537	0	0.0
## person_neoplasm_stt	person_neoplasm_stt	<fct>	502	35	6.5
## vital_stt	vital_stt	<fct>	534	3	0.6
## wbc	wbc	<fct>	441	96	17.9

6. Checking numeric variables

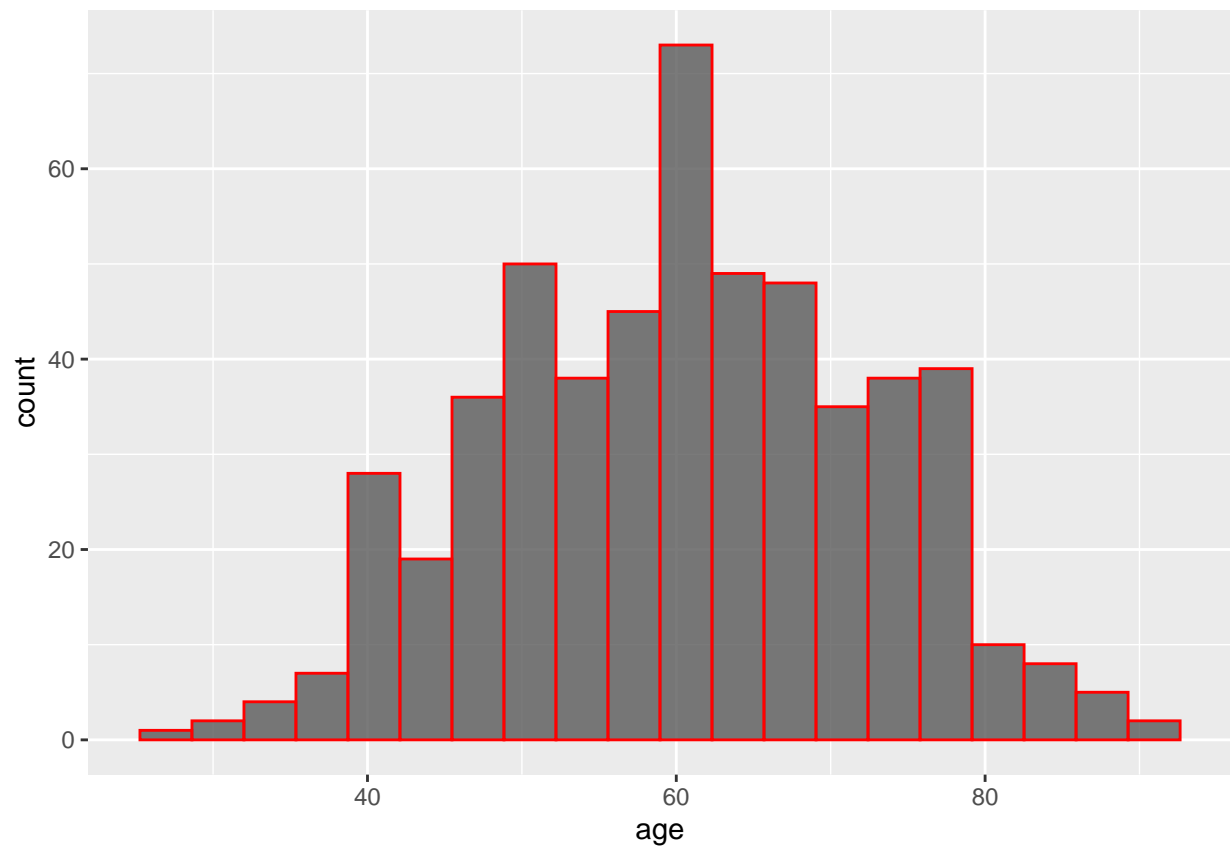
Check data distribution, plausible ranges, outliers;

Thinking about deleting outliers from dataset? Need to evaluate carefully each one!

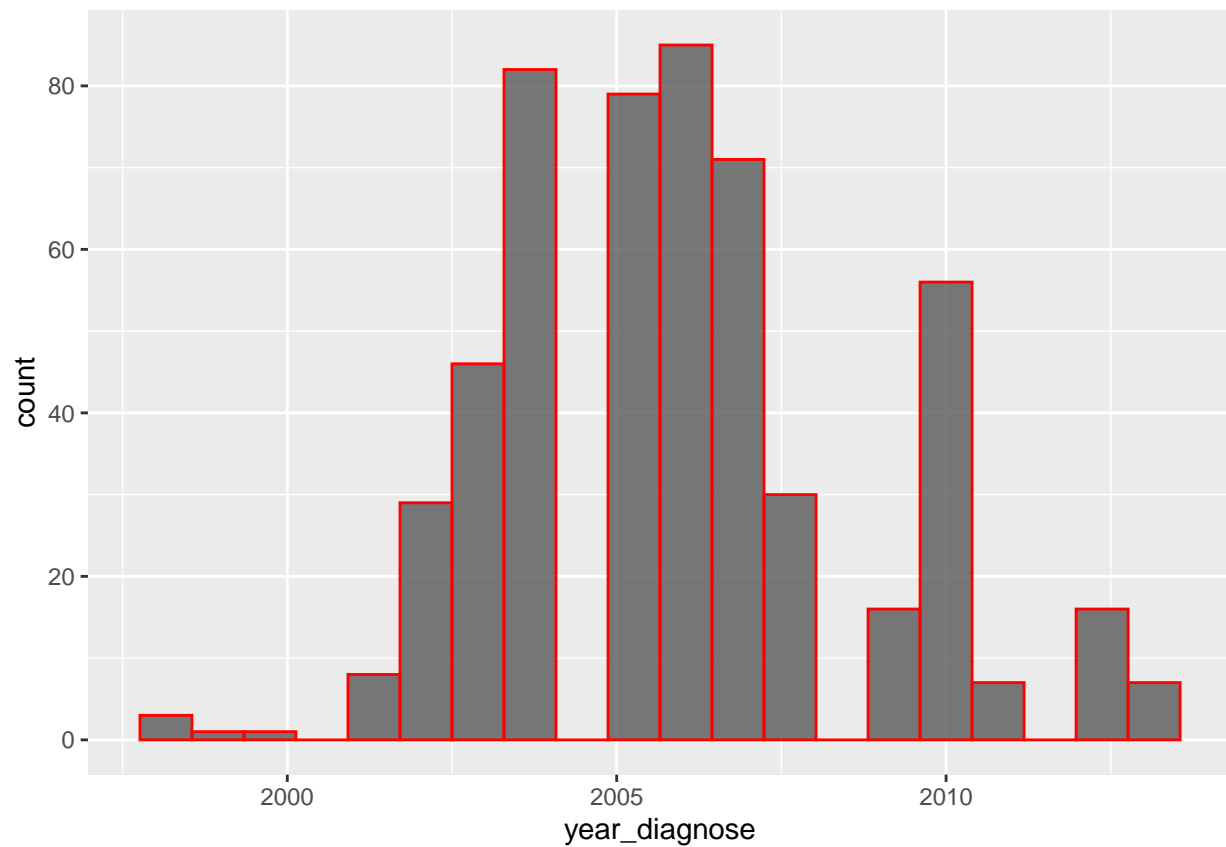
```
kirc_clean4 %>%
  select_if(is.numeric) %>%
  summary()
```

##	age	disease_free_mth	frac_genome_alter	year_diagnose
##	Min. :26.00	Min. :-11.79	Min. :0.00000	Min. :1998
##	1st Qu.:52.00	1st Qu.: 13.43	1st Qu.:0.06295	1st Qu.:2004
##	Median :61.00	Median : 36.20	Median :0.12065	Median :2006
##	Mean :60.59	Mean : 40.24	Mean :0.17016	Mean :2006
##	3rd Qu.:70.00	3rd Qu.: 60.51	3rd Qu.:0.20885	3rd Qu.:2007
##	Max. :90.00	Max. :133.84	Max. :0.94770	Max. :2013
##		NA's :99	NA's :9	
##	long_dim	mutation_cnt	over_surv_mth	short_dim
##	Min. :0.400	Min. : 1.00	Min. : 0.00	Min. :0.1000
##	1st Qu.:1.200	1st Qu.: 34.00	1st Qu.: 18.10	1st Qu.:0.2000
##	Median :1.500	Median : 48.00	Median : 38.96	Median :0.3000
##	Mean :1.662	Mean : 73.85	Mean : 44.26	Mean :0.3759
##	3rd Qu.:2.000	3rd Qu.: 65.50	3rd Qu.: 63.21	3rd Qu.:0.5000
##	Max. :4.000	Max. :1392.00	Max. :149.05	Max. :1.0000
##	NA's :35	NA's :86		NA's :35
##	second_long_dim			
##	Min. :0.3000			
##	1st Qu.:0.7000			
##	Median :0.9000			
##	Mean :0.9368			
##	3rd Qu.:1.1000			
##	Max. :2.0000			
##	NA's :35			

```
ggplot(kirc_clean4, aes(age)) +
  geom_histogram(bins = 20, alpha = 0.8, color = "red")
```



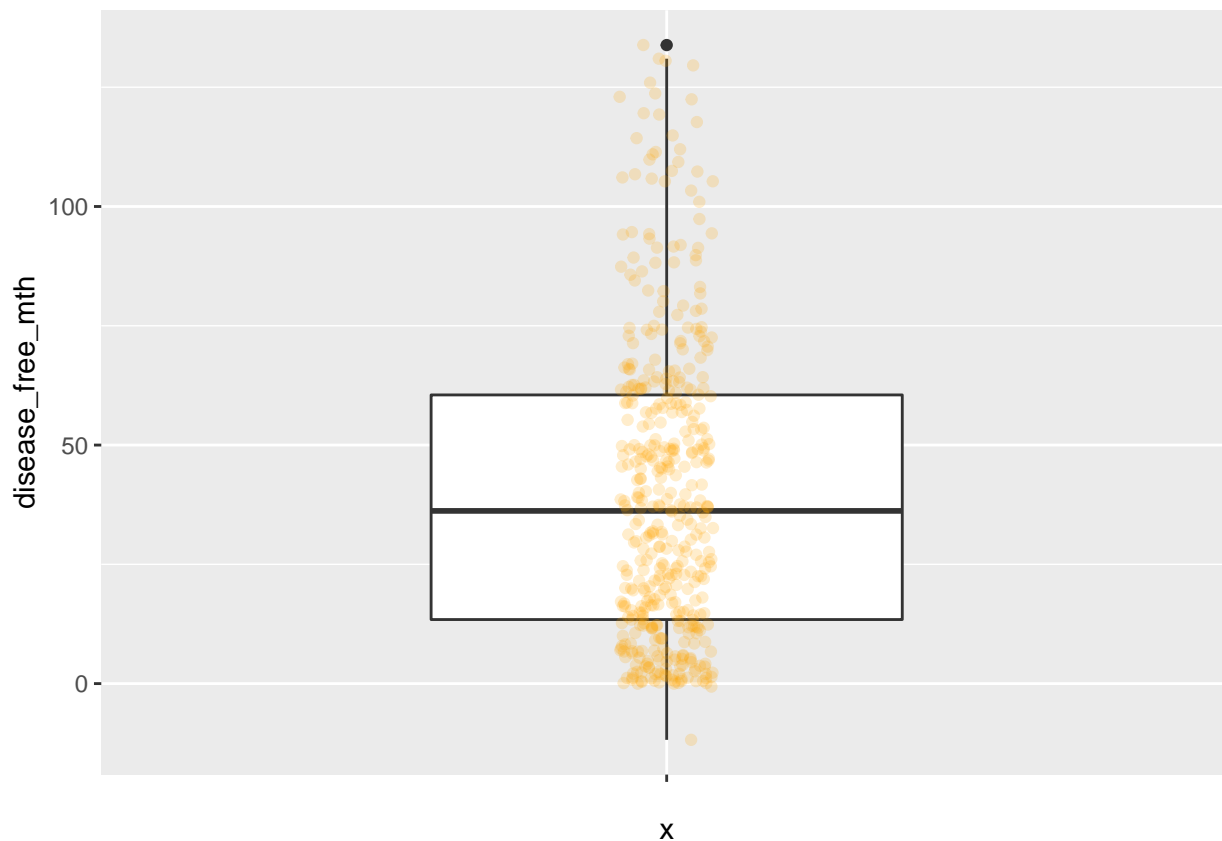
```
ggplot(kirc_clean4, aes(year_diagnose)) +  
  geom_histogram(bins = 20, alpha = 0.8, color = "red")
```

```
ggplot(kirc_clean4, aes(x='', y=disease_free_mth)) +  
  geom_boxplot(width = .5) +  
  geom_jitter(width = 0.05, alpha = 0.2, color = "orange")
```

```
## Warning: Removed 99 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 99 rows containing missing values (geom_point).
```



```
boxplot.stats(kirc_clean4$disease_free_mth)
```

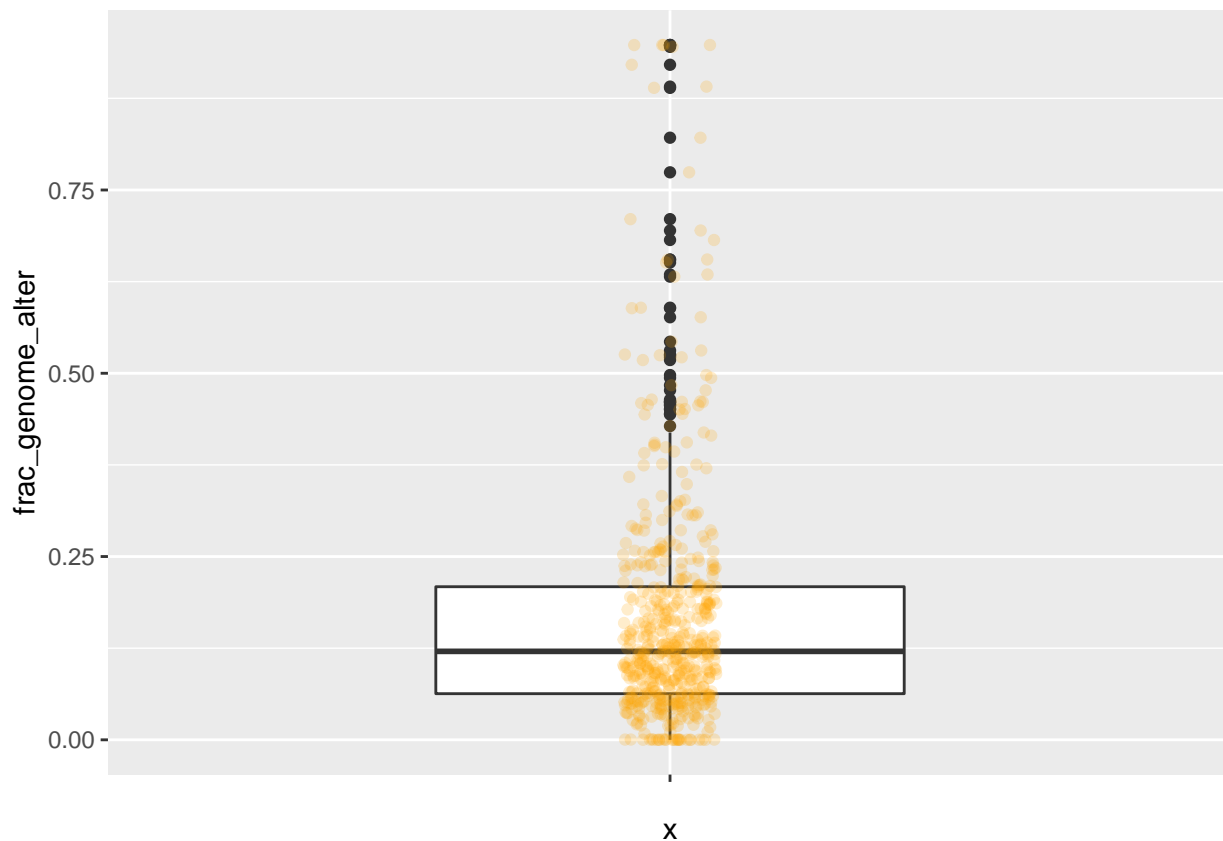
```
## $stats
## [1] -11.79  13.40  36.20  60.55 130.98
##
## $n
## [1] 438
##
## $conf
## [1] 32.6404 39.7596
##
## $out
## [1] 133.84
```

```
# filter(disease_free_mth >= 0)
```

```
ggplot(kirc_clean4, aes(x = '', y=frac_genome_alter)) +
  geom_boxplot(width = .5) +
  geom_jitter(width = 0.05, alpha = 0.2, color = "orange")
```

```
## Warning: Removed 9 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 9 rows containing missing values (geom_point).
```



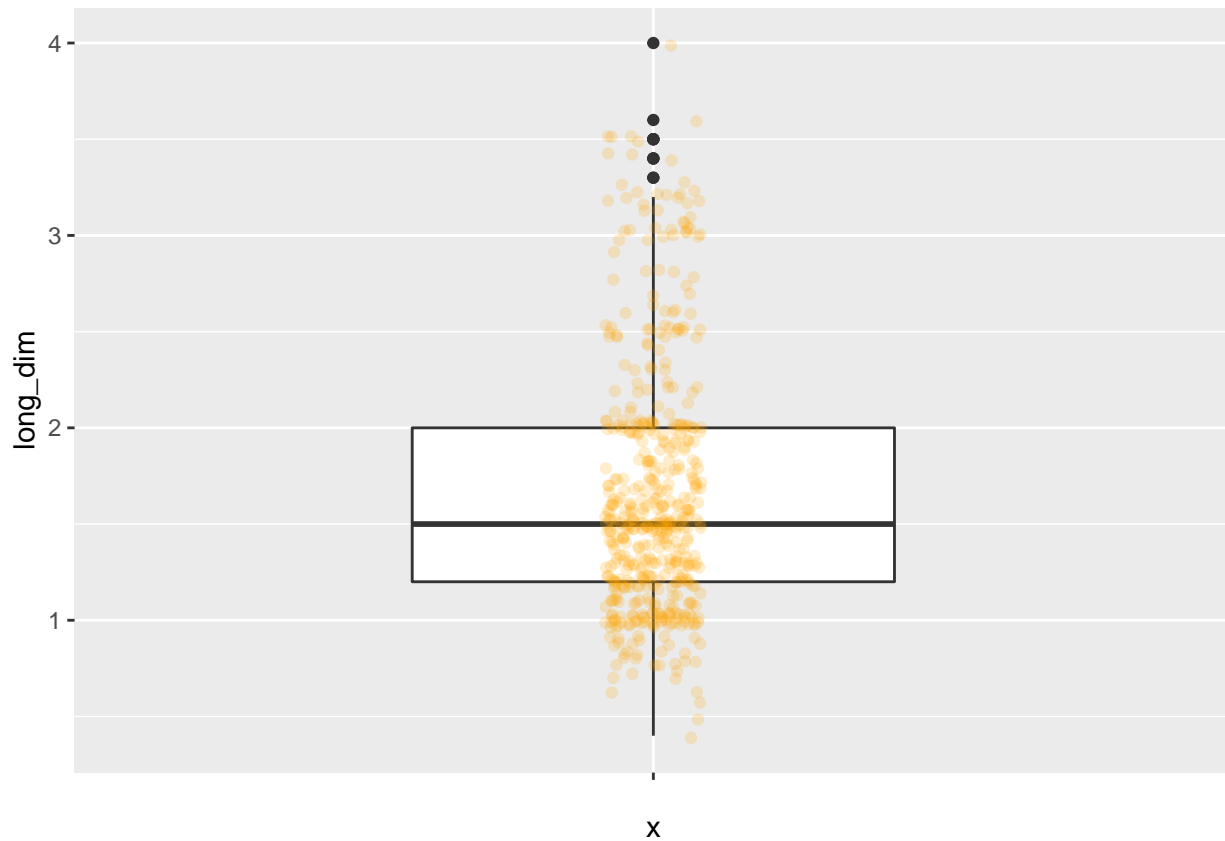
```
boxplot.stats(kirc_clean4$frac_genome_alter)
```

```
## $stats
## [1] 0.00000 0.06290 0.12065 0.20920 0.42800
##
## $n
## [1] 528
##
## $conf
## [1] 0.1105903 0.1307097
##
## $out
## [1] 0.8213 0.6552 0.4608 0.9477 0.5888 0.9208 0.7741 0.4837 0.9477 0.4610
## [11] 0.6549 0.6511 0.5180 0.8910 0.8893 0.9477 0.5246 0.4568 0.4937 0.9477
## [21] 0.4438 0.6947 0.5218 0.4768 0.4593 0.4447 0.9452 0.6347 0.5311 0.4562
## [31] 0.4617 0.5256 0.6318 0.5430 0.4506 0.5764 0.7102 0.4641 0.5894 0.4976
## [41] 0.4513 0.6818
```

```
ggplot(kirc_clean4, aes(x='', y=long_dim)) +
  geom_boxplot(width = .5) +
  geom_jitter(width = 0.05, alpha = 0.2, color = "orange")
```

```
## Warning: Removed 35 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 35 rows containing missing values (geom_point).
```



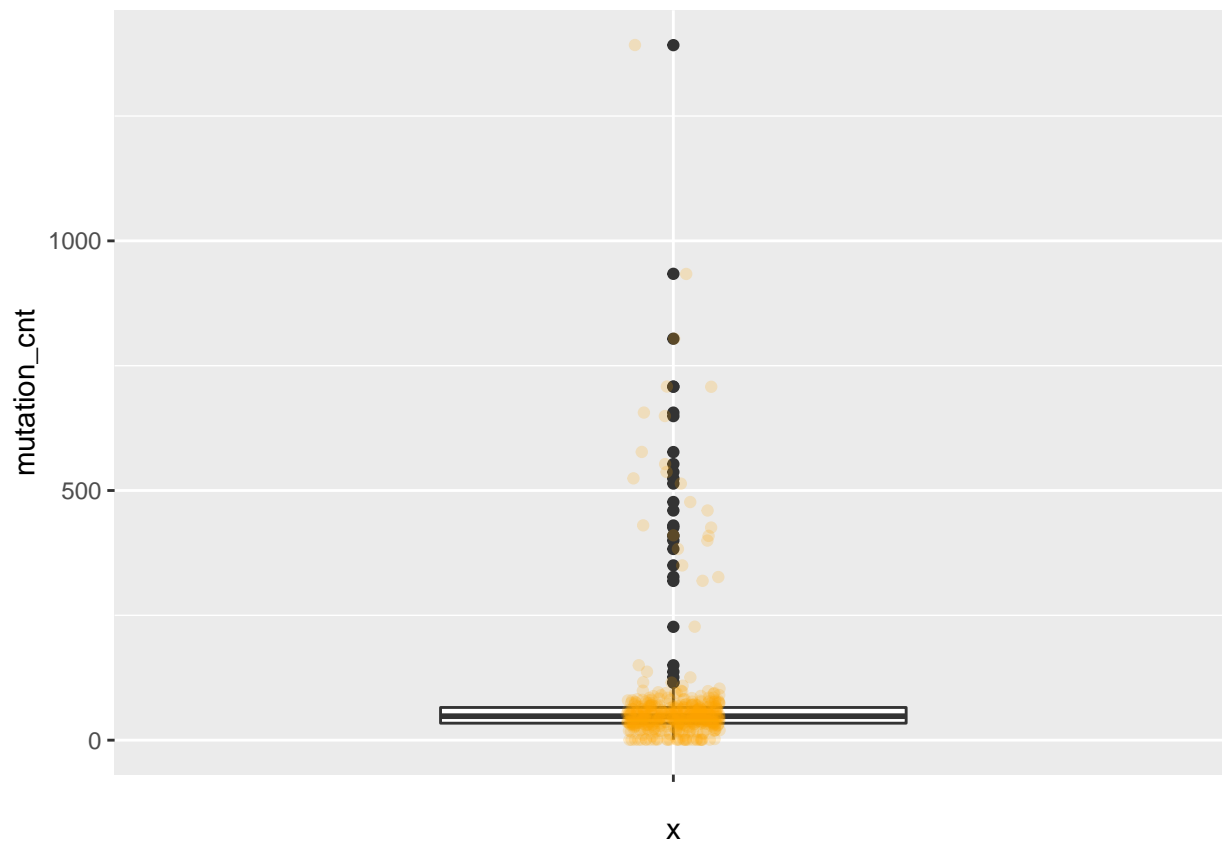
```
boxplot.stats(kirc_clean4$long_dim)
```

```
## $stats
## [1] 0.4 1.2 1.5 2.0 3.2
##
## $n
## [1] 502
##
## $conf
## [1] 1.443585 1.556415
##
## $out
## [1] 3.3 4.0 3.3 3.5 3.4 3.5 3.5 3.4 3.4 3.5 3.6
```

```
ggplot(kirc_clean4, aes(x='', y=mutation_cnt)) +
  geom_boxplot(width = .5) +
  geom_jitter(width = 0.05, alpha = 0.2, color = "orange")
```

```
## Warning: Removed 86 rows containing non-finite values (stat_boxplot).
```

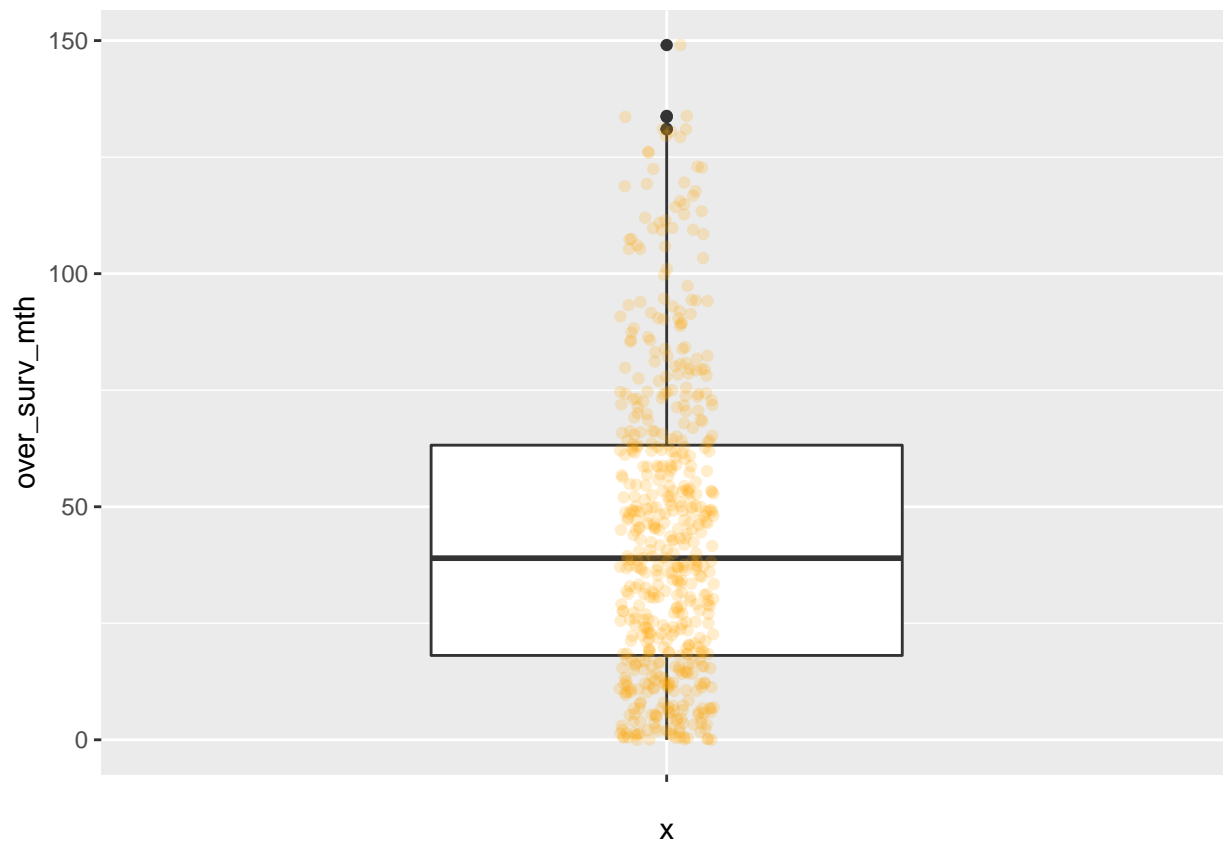
```
## Warning: Removed 86 rows containing missing values (geom_point).
```



```
boxplot.stats(kirc_clean4$mutation_cnt)
```

```
## $stats
## [1]  1.0  34.0  48.0  65.5 109.0
##
## $n
## [1] 451
##
## $conf
## [1] 45.65642 50.34358
##
## $out
## [1]  514  656  577  537  477  150  137  708 1392  460  327  934  409  383  804
## [16]  319  524  426  227  553  400  350  410  430  708  649  126  116  115
```

```
ggplot(kirc_clean4, aes(x='', y=over_surv_mth)) +
  geom_boxplot(width = .5) +
  geom_jitter(width = 0.05, alpha = 0.2, color = "orange")
```



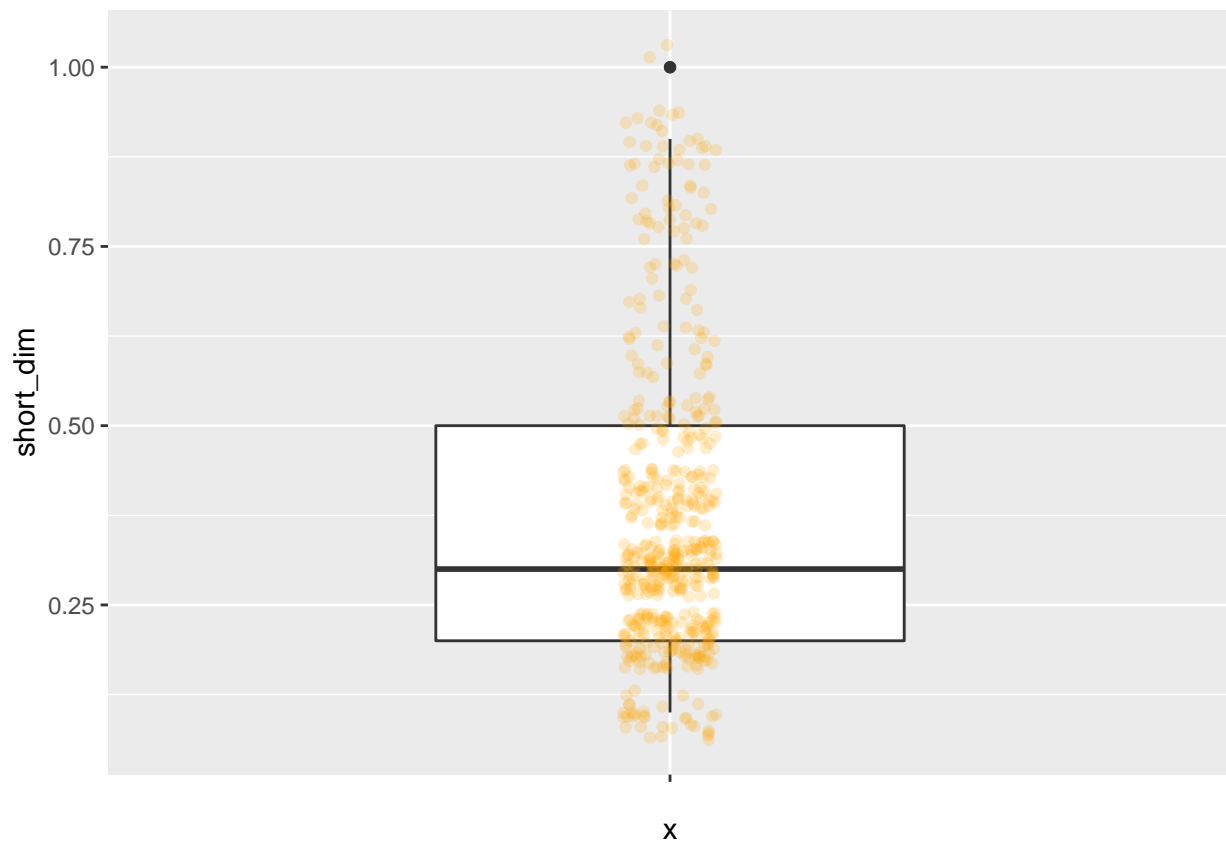
```
boxplot.stats(kirc_clean4$over_surv_mth)
```

```
## $stats
## [1]  0.00  18.10  38.96  63.21 130.55
##
## $n
## [1] 537
##
## $conf
## [1] 35.88431 42.03569
##
## $out
## [1] 133.84 149.05 131.04 130.98 133.61
```

```
ggplot(kirc_clean4, aes(x='', y=short_dim)) +
  geom_boxplot(width = .5) +
  geom_jitter(width = 0.05, alpha = 0.2, color = "orange")
```

```
## Warning: Removed 35 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 35 rows containing missing values (geom_point).
```



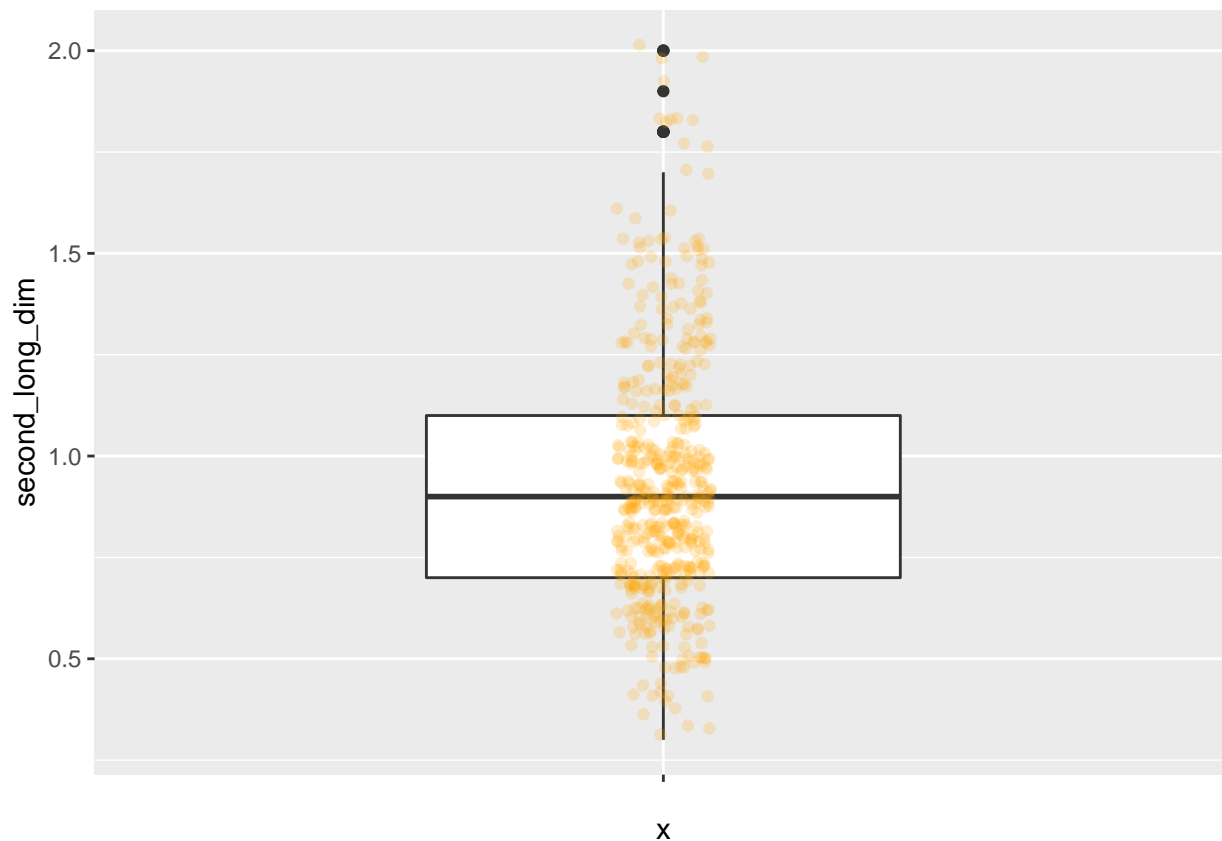
```
boxplot.stats(kirc_clean4$short_dim)
```

```
## $stats
## [1] 0.1 0.2 0.3 0.5 0.9
##
## $n
## [1] 502
##
## $conf
## [1] 0.2788443 0.3211557
##
## $out
## [1] 1 1
```

```
ggplot(kirc_clean4, aes(x='', y=second_long_dim)) +
  geom_boxplot(width = .5) +
  geom_jitter(width = 0.05, alpha = 0.2, color = "orange")
```

```
## Warning: Removed 35 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 35 rows containing missing values (geom_point).
```



```
boxplot.stats(kirc_clean4$second_long_dim)
```

```
## $stats
## [1] 0.3 0.7 0.9 1.1 1.7
##
## $n
## [1] 502
##
## $conf
## [1] 0.8717925 0.9282075
##
## $out
## [1] 1.8 2.0 1.8 1.9 1.8 2.0 2.0 1.8 1.8 1.8 1.8
```

7. Checking categorical variables

Check frequency, labels and levels

```
kirc_clean4 %>%
  select_if(is.factor) %>%
  summary()
```

```
## metastasis_stg neoplasm_ln_stg   neoplasm_stg   tumor_stg
## M0 :426        N0:240          Stage I  :269   T1a    :142
## M1 : 79        N1: 17          Stage II : 57   T3a    :122
## MX : 30        NX:280          Stage III:125  T1b    :111
## NA's: 2                Stage IV : 83   T2     : 55
```



```

##          NA's      : 3   T3b      : 53
##                               T1      : 22
##                               (Other): 32
##      disease_free_stt      ethnicity      histology_grd
## DiseaseFree      :311   HISPANIC OR LATINO      : 26   G1      : 14
## Recurred/Progressed:127   NOT HISPANIC OR LATINO:359   G2      :230
## NA's      : 99   NA's      :152   G3      :207
##                               G4      : 78
##                               GX      : 5
##                               NA's: 3
##
##      hemoglobin      neoadj_therapy
## Elevated: 5   No :519
## Low      :263   Yes: 18
## Normal   :186
## NA's     : 83
##
##
##
##          prior_cancer      tumor_lateral
## No          :459   Bilateral: 1
## Yes         : 72   Left      :253
## Yes, History of Prior Malignancy      : 2   Right      :283
## Yes, History of Synchronous/Bilateral Malignancy: 4
##
##
##
##      primer_ln_ind3      platelet      tissue_prospect      race
## NO :395      Elevated: 38   NO :465      ASIAN      : 8
## YES :135      Low      : 46   YES : 52      BLACK OR AFRICAN AMERICAN: 56
## NA's: 7      Normal   :360   NA's: 20      WHITE      :466
##                               NA's      : 93      NA's      : 7
##
##
##
##      tissue_retrospect      serum_ca      sex      tissue_site
## NO : 53      Elevated: 10   Female:191   BP      :142
## YES :466      Low      :204   Male :345   B0      :107
## NA's: 18      Normal   :151   MALE : 1   CJ      : 71
##                               NA's      :172   A3      : 52
##                               CZ      : 40
##                               B8      : 33
##                               (Other): 92
##
##      person_neoplasm_stt      vital_stt      wbc
## TUMOR FREE:361      Alive:360      Elevated:164
## WITH TUMOR:141      Dead :174      Low      : 9
## NA's      : 35      NA's : 3      Normal   :268
##                               NA's      : 96
##
##
##

```

```

# agregating levels
kirc_clean5 <- kirc_clean4 %>%

```

```

mutate(tumor_stg = fct_collapse(tumor_stg,
                                T1 = c('T1', 'T1a', 'T1b'),
                                T2 = c('T2', 'T2a', 'T2b'),
                                T3 = c('T3', 'T3a', 'T3b', 'T3c')))
kirc_clean5 <- kirc_clean4 %>%
  mutate(prior_cancer = fct_collapse(prior_cancer,
                                      Yes = c('Yes', 'Yes, History of Prior Malignancy', 'Yes, History of Synchronous/Bilateral')))

kirc_clean5 <- kirc_clean4 %>%
  mutate(sex = fct_collapse(sex, Male = c('MALE', 'Male')))

kirc_clean5 <- kirc_clean4 %>%
  mutate(tissue_site = fct_collapse(tissue_site,
                                    A = c('A3', 'AK', 'AS'),
                                    B = c('B0', 'B2', 'B4', 'B8', 'BP'),
                                    C = c('CJ', 'CW', 'CZ'),
                                    G = c('G6', 'GK'),
                                    M = c('MM', 'MW')))

# dropping levels
kirc_clean5 <- kirc_clean4 %>%
  mutate(race = fct_recode(race, NULL = 'ASIAN'))

# kirc_clean5 <- kirc_clean4 %>%
#   mutate(race = fct_drop(race, only = 'ASIAN'))

# recoding levels
# OBS: It can be done latter, for regression analysis
# kirc_clean4 %>%
#   select_if(is.factor) %>%
#   summary()
#
# kirc_clean5 <- kirc_clean4 %>%
#   mutate(sex = fct_recode(sex, '1'='Male', '2'='Female'))
#
# kirc_clean5 <- kirc_clean4 %>%
#   mutate(sex = if_else(sex %in% c('Male', 'Female'), 1, 0))

```

8. Saving dataset

```

write_csv(kirc_clean5, path = "data/kirc_clinical_clean.csv")

rm(kirc_clean4, kirc_clean3, kirc_clean2, kirc_clean1, kirc_clean0, kirc_clean)

# table(kirc_clean4$metastasis_stg, exclude = NULL)
# table(kirc_clean4$neoplasm_ln_stg, exclude = NULL)
# table(kirc_clean4$neoplasm_stg, exclude = NULL)
# table(kirc_clean4$tumor_stg, exclude = NULL)
# table(kirc_clean4$disease_free_stt, exclude = NULL)
# table(kirc_clean4$ethnicity, exclude = NULL)
# table(kirc_clean4$histology_grd, exclude = NULL)
# table(kirc_clean4$hemoglobin, exclude = NULL)

```

```
# table(kirc_clean4$neoadj_therapy, exclude = NULL)
# table(kirc_clean4$prior_cancer, exclude = NULL)
# table(kirc_clean4$tumor_lateral, exclude = NULL)
# table(kirc_clean4$primer_ln_ind3, exclude = NULL)
# table(kirc_clean4$platelet, exclude = NULL)
# table(kirc_clean4$tissue_prospect, exclude = NULL)
# table(kirc_clean4$race, exclude = NULL)
# table(kirc_clean4$tissue_retrospect, exclude = NULL)
# table(kirc_clean4$serum_ca, exclude = NULL)
# table(kirc_clean4$sex, exclude = NULL)
# table(kirc_clean4$tissue_site, exclude = NULL)
# table(kirc_clean4$person_neoplasm_stt, exclude = NULL)
# table(kirc_clean4$wbc, exclude = NULL)
```

```
sessionInfo()
```

```
## R version 3.6.3 (2020-02-29)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 18.04.4 LTS
##
## Matrix products: default
## BLAS: /usr/lib/x86_64-linux-gnu/openblas/libblas.so.3
## LAPACK: /usr/lib/x86_64-linux-gnu/libopenblas-p-r0.2.20.so
##
## locale:
##  [1] LC_CTYPE=pt_BR.UTF-8      LC_NUMERIC=C
##  [3] LC_TIME=pt_BR.UTF-8      LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=pt_BR.UTF-8  LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=pt_BR.UTF-8     LC_NAME=C
##  [9] LC_ADDRESS=C             LC_TELEPHONE=C
## [11] LC_MEASUREMENT=pt_BR.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats      graphics  grDevices  utils      datasets  methods    base
##
## other attached packages:
## [1] finalfit_1.0.1  skimr_2.1.1    forcats_0.5.0  stringr_1.4.0
## [5] dplyr_0.8.5     purrr_0.3.4    readr_1.3.1    tidyr_1.0.3
## [9] tibble_3.0.1    ggplot2_3.3.0  tidyverse_1.3.0
##
## loaded via a namespace (and not attached):
## [1] Rcpp_1.0.4.6    lubridate_1.7.8  lattice_0.20-41  assertthat_0.2.1
## [5] digest_0.6.25  utf8_1.1.4      R6_2.4.1        cellranger_1.1.0
## [9] repr_1.1.0     backports_1.1.6  reprex_0.3.0    evaluate_0.14
## [13] httr_1.4.1     highr_0.8       pillar_1.4.4    rlang_0.4.6
## [17] readxl_1.3.1   rstudioapi_0.11  Matrix_1.2-18   rmarkdown_2.1
## [21] labeling_0.3   splines_3.6.3    munsell_0.5.0   broom_0.5.6
## [25] compiler_3.6.3 modelr_0.1.7     xfun_0.13       pkgconfig_2.0.3
## [29] base64enc_0.1-3 htmltools_0.4.0  tidyselect_1.1.0 fansi_0.4.1
## [33] crayon_1.3.4   dbplyr_1.4.3     withr_2.2.0     grid_3.6.3
## [37] nlme_3.1-147   jsonlite_1.6.1   gtable_0.3.0    lifecycle_0.2.0
## [41] DBI_1.1.0      magrittr_1.5     scales_1.1.1    cli_2.0.2
## [45] stringi_1.4.6  farver_2.0.3     fs_1.4.1        mice_3.8.0
## [49] xml2_1.3.2     ellipsis_0.3.0   generics_0.0.2  vctrs_0.3.0
```

```
## [53] boot_1.3-25      tools_3.6.3      glue_1.4.0       hms_0.5.3
## [57] survival_3.1-12  yaml_2.2.1       colorspace_1.4-1 rvest_0.3.5
## [61] knitr_1.28       haven_2.2.0
```