[Tutorial] Gene expression profiles in cancer patients

1. Introduction

Lung cancer one of major cancers, accounting for 2.09 million deaths out of the 9.6 million total cancer deaths in 2018. There are two main types of lung cancer: small cell lung carcinoma (SCLC) and non-small cell lung carcinoma (NSCLC). NSCLC is responsible for 85–90% of lung cancer cases and its two largest subtypes are lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC).

LUAD develops in the periphery of the lungs and may be associated with smoking, but is the most common lung cancer type among non-smokers. In contrast, LUSC accounts for 25–30% of all total lung cancer cases while LUAD accounts for 40% of all total lung cancer cases. LUSC are likely to be found in the middle of the lungs and is associated with smoking. In this tutorial, we will explore the sample dataset of LUSC and LUAD from the The Cancer Genome Atlas (TCGA), a public resource for the genomic dataset. Here we use two types of lung cancers - LUAD and LUSC and will examine which information we can use for genomic analyses of lung cancers.

2. Obtain the gene expression profile dataset

To access the gene expression data, there are two ways

1. Formal way (but slow): you can download the data as described in 2.1
2. Simple ways: get the file from my dropbox link where I downloaded and save to R object as described in 2.2

2.1. Download the data from GDC portal

You will need to install the TCGAbiolinks package to obtain the lung cancer gene expression profiles from TCGA.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
# Install the TCGAbiolinks package if necessary
#if (!requireNamespace("BiocManager", quietly = TRUE))    #install.packages("BiocManager")
#BiocManager::install("TCGAbiolinks")

#library(TCGAbiolinks)
```

Then, Obtain the TCGA dataset from the GDC portal. You first download the gene expression dataset for Lung Adenocarcinoma (LUAD).

```
#query <- GDCquery(project = "TCGA-LUAD",
                   #data.category = "Transcriptome Profiling",
                   #data.type = "Gene Expression Quantification",
                   #workflow.type = "HTSeq - FPKM-UQ")
#GDCdownload(query)
#d_luad0 <- GDCprepare(query)
#d_luad = as.data.frame(d_luad0@colData)
```

Now we are downloading the dataset for Lung Squamous Cell Carcinoma (LUSC).

```
#query <- GDCquery(project = "TCGA-LUSC",
                   #data.category = "Transcriptome Profiling",
                   #data.type = "Gene Expression Quantification",
                   #workflow.type = "HTSeq - FPKM-UQ")
#GDCdownload(query)
#d_lusc0 <- GDCprepare(query)
#d_lusc = as.data.frame(d_lusc0@colData)
```

For simplicity, we are choosing only genes on the chromosome 1. Please save to the Rdata into your working folder.

```
#library(SummarizedExperiment)
#e_luad = assay(d_luac0)
#e_lusc = assay(d_lusc0)

#g_luad = d_luad0@rowRanges %>% as.data.frame() %>% filter(seqnames=='chr1') %>% pull(ensembl_gene_id)
#g_lusc = d_lusc0@rowRanges %>% as.data.frame() %>% filter(seqnames=='chr1') %>% pull(ensembl_gene_id)
#e_luad = e_luad[g_luad,]
#e_lusc = e_lusc[g_lusc,]

#save(d_luad, d_lusc, e_luad, e_lusc, #file='data.TCGA_LUAD_LUSC.gene_expression.Rdata')
```

2.2. Get the file from the link

You can download the data from this link. Please save this file to the working folder. Then, load the objects to your workspace.

```
load('data.TCGA_LUAD_LUSC.gene_expression.Rdata')
```

3. Explore the input dataset

Let's explore which columns are provided in the LUAD dataset.

```
colnames(d_luad)
```

```
##  [1] "barcode"
##  [2] "patient"
##  [3] "sample"
##  [4] "shortLetterCode"
##  [5] "definition"
##  [6] "sample_submitter_id"
##  [7] "sample_type_id"
##  [8] "sample_id"
##  [9] "sample_type"
## [10] "days_to_collection"
## [11] "state"
## [12] "initial_weight"
## [13] "intermediate_dimension"
## [14] "pathology_report_uuid"
## [15] "submitter_id"
## [16] "shortest_dimension"
## [17] "oct_embedded"
## [18] "longest_dimension"
## [19] "is_ffpe"
## [20] "tissue_type"
## [21] "synchronous_malignancy"
## [22] "ajcc_pathologic_stage"
## [23] "tumor_stage"
## [24] "days_to_diagnosis"
## [25] "treatments"
## [26] "last_known_disease_status"
## [27] "tissue_or_organ_of_origin"
## [28] "days_to_last_follow_up"
## [29] "age_at_diagnosis"
## [30] "primary_diagnosis"
## [31] "prior_malignancy"
## [32] "year_of_diagnosis"
## [33] "prior_treatment"
## [34] "ajcc_staging_system_edition"
## [35] "ajcc_pathologic_t"
## [36] "morphology"
## [37] "ajcc_pathologic_n"
## [38] "ajcc_pathologic_m"
## [39] "classification_of_tumor"
## [40] "diagnosis_id"
## [41] "icd_10_code"
## [42] "site_of_resection_or_biopsy"
## [43] "tumor_grade"
## [44] "progression_or_recurrence"
## [45] "cigarettes_per_day"
## [46] "alcohol_history"
## [47] "exposure_id"
## [48] "years_smoked"
## [49] "pack_years_smoked"
## [50] "gender"
## [51] "ethnicity"
```

```
## [52] "race"
## [53] "vital_status"
## [54] "age_at_index"
## [55] "days_to_birth"
## [56] "year_of_birth"
## [57] "demographic_id"
## [58] "days_to_death"
## [59] "year_of_death"
## [60] "bcr_patient_barcode"
## [61] "primary_site"
## [62] "disease_type"
## [63] "project_id"
## [64] "releasable"
## [65] "name"
## [66] "released"
## [67] "paper_patient"
## [68] "paper_Sex"
## [69] "paper_Age.at.diagnosis"
## [70] "paper_T.stage"
## [71] "paper_N.stage"
## [72] "paper_Tumor.stage"
## [73] "paper_Smoking.Status"
## [74] "paper_Survival"
## [75] "paper_Transversion.High.Low"
## [76] "paper_Nonsilent.Mutations"
## [77] "paper_Nonsilent.Mutations.per.Mb"
## [78] "paper_Oncogene.Negative.or.Positive.Groups"
## [79] "paper_Fusions"
## [80] "paper_expression_subtype"
## [81] "paper_chromosome.affected.by.chromothripsis"
## [82] "paper_iCluster.Group"
## [83] "paper_CIMP.methylation.signature."
## [84] "paper_MTOR.mechanism.of.mTOR.pathway.activation"
## [85] "paper_Ploidy.ABSOLUTE.calls"
## [86] "paper_Purity.ABSOLUTE.calls"
```

Do we see the same columns for the LUSC dataset?

```
colnames(d_lusc)
```

```
##  [1] "barcode"                       "patient"
##  [3] "sample"                        "shortLetterCode"
##  [5] "definition"                    "sample_submitter_id"
##  [7] "sample_type_id"                "sample_id"
##  [9] "sample_type"                   "days_to_collection"
## [11] "state"                         "initial_weight"
## [13] "intermediate_dimension"        "pathology_report_uuid"
## [15] "submitter_id"                  "shortest_dimension"
## [17] "oct_embedded"                  "longest_dimension"
## [19] "is_ffpe"                       "tissue_type"
## [21] "synchronous_malignancy"        "ajcc_pathologic_stage"
## [23] "tumor_stage"                   "days_to_diagnosis"
## [25] "treatments"                    "last_known_disease_status"
```

```
## [27] "tissue_or_organ_of_origin"       "days_to_last_follow_up"
## [29] "primary_diagnosis"               "age_at_diagnosis"
## [31] "prior_malignancy"                "year_of_diagnosis"
## [33] "prior_treatment"                 "ajcc_staging_system_edition"
## [35] "ajcc_pathologic_t"               "morphology"
## [37] "ajcc_pathologic_n"               "ajcc_pathologic_m"
## [39] "classification_of_tumor"         "diagnosis_id"
## [41] "icd_10_code"                     "site_of_resection_or_biopsy"
## [43] "tumor_grade"                     "progression_or_recurrence"
## [45] "pack_years_smoked"               "cigarettes_per_day"
## [47] "alcohol_history"                 "exposure_id"
## [49] "years_smoked"                    "race"
## [51] "ethnicity"                       "gender"
## [53] "vital_status"                    "age_at_index"
## [55] "days_to_birth"                   "year_of_birth"
## [57] "demographic_id"                  "days_to_death"
## [59] "year_of_death"                   "bcr_patient_barcode"
## [61] "primary_site"                    "project_id"
## [63] "disease_type"                    "name"
## [65] "releasable"                      "released"
## [67] "paper_patient"                   "paper_Sex"
## [69] "paper_Age.at.diagnosis"          "paper_T.stage"
## [71] "paper_N.stage"                   "paper_M.stage"
## [73] "paper_Smoking.Status"            "paper_Pack.years"
## [75] "paper_Nonsilent.Mutatios"        "paper_Nonsilent.Mutatios.per.Mb"
## [77] "paper_Selected.Mutation.Summary" "paper_High.Level.Amplifications"
## [79] "paper_Homozygous.Deletions"      "paper_Expression.Subtype"
```

3.1. Which tissues or samples are available for your analysis?

Lung cancer samples in the dataset are provided in tumor or normal tissues. The column `shortLetterCode` contains Sample Type Codes to describe the type of tissues collected in the dataset.

TP: Primary solid Tumor TR: Recurrent solid Tumor *NT: Solid Tissue Normal

Check which samples are included in the LUAD dataset.

```
library(tidyverse)
d_luad %>% count(shortLetterCode)
```

```
##   shortLetterCode   n
## 1              NT  59
## 2              TP 533
## 3              TR   2
```

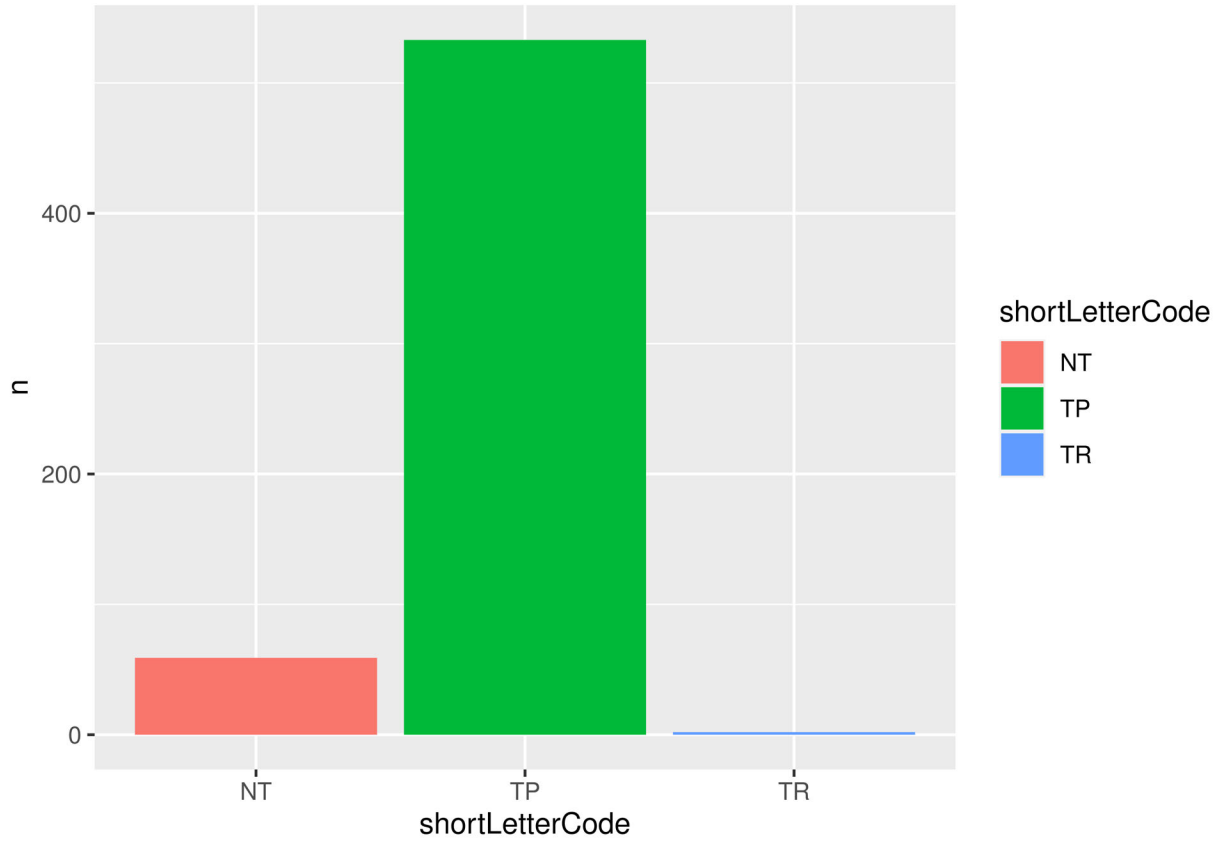
Check which samples are included in the LUSC dataset.

```
d_lusc %>% count(shortLetterCode)
```

```
##   shortLetterCode   n
## 1              NT  49
## 2              TP 502
```

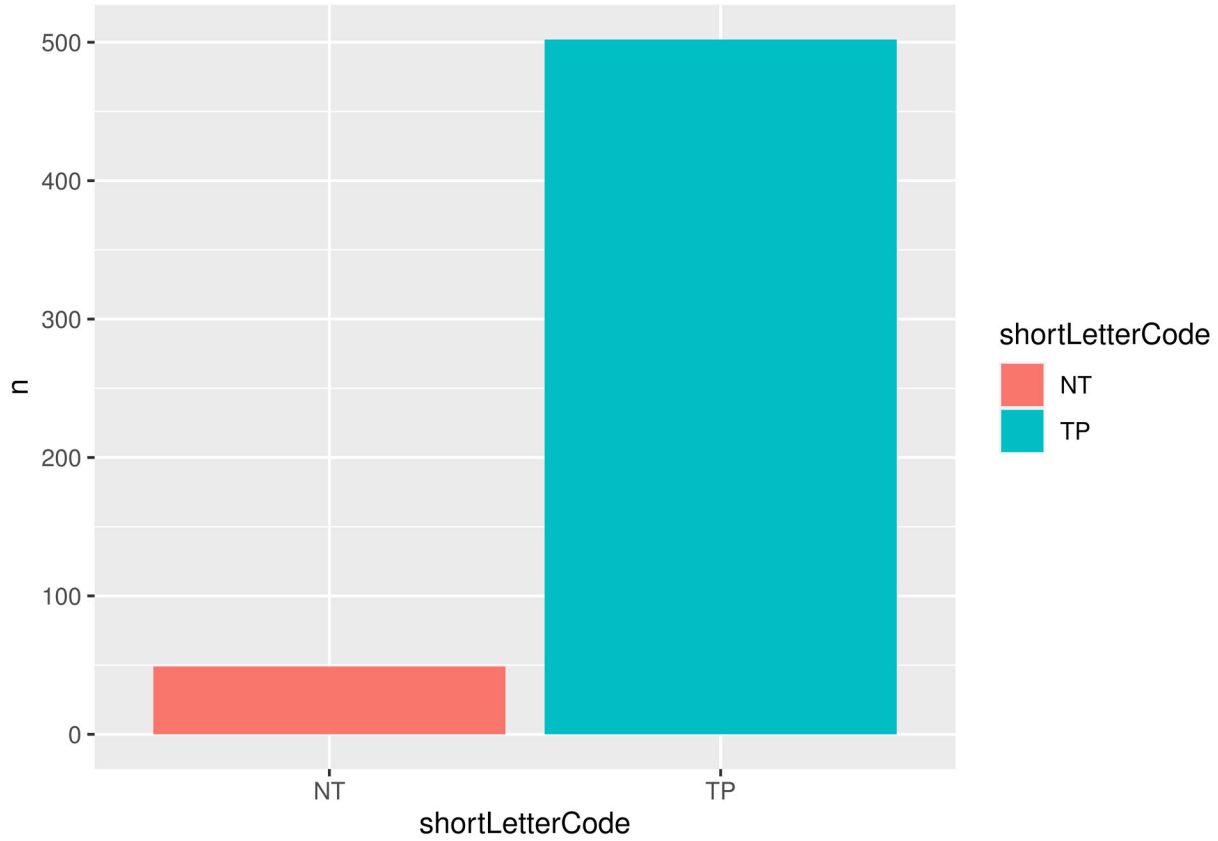
Let's make a simple bar plot to compare the number of tissues between two tissue types. We will look at the LUAD samples first.

```
d_luad %>%
  count(shortLetterCode) %>%
  ggplot(., aes(shortLetterCode, n, fill=shortLetterCode)) +
  geom_bar(stat="identity", , position=position_dodge())
```



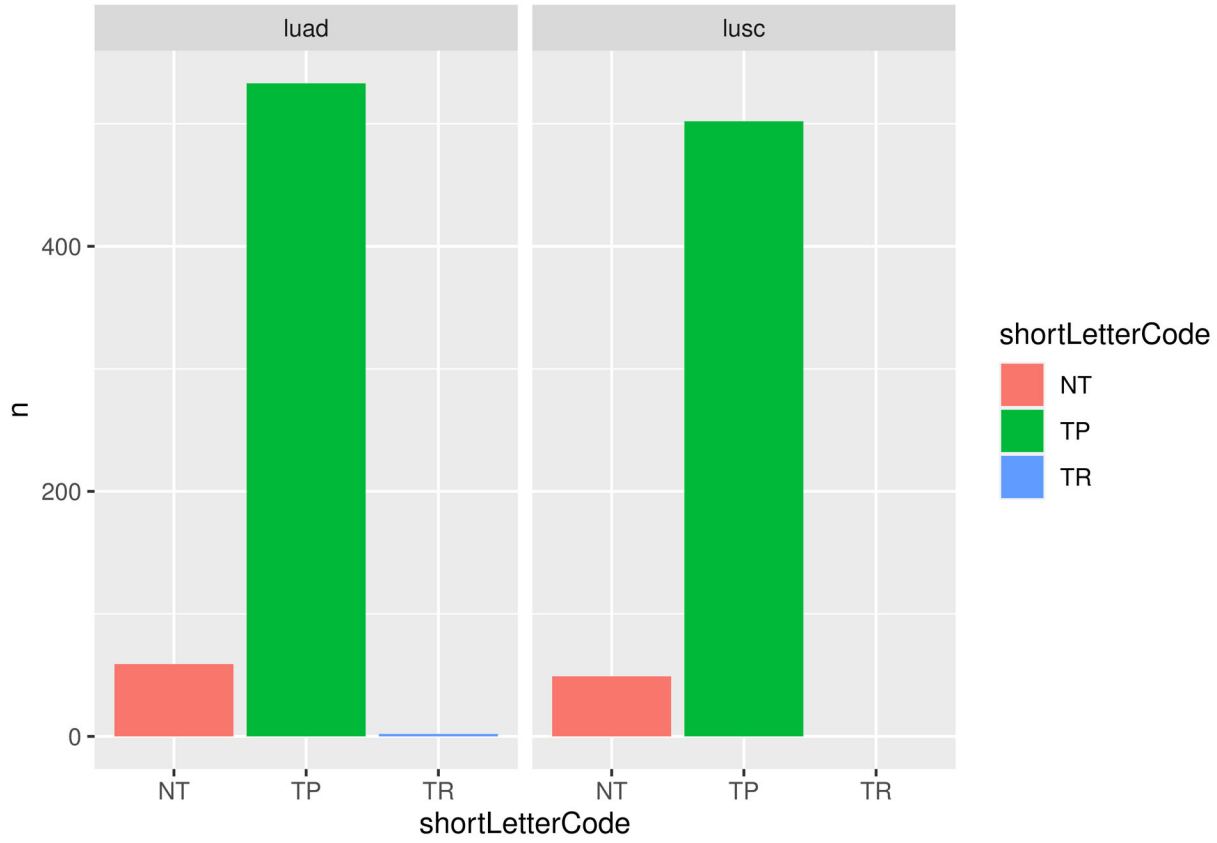
Now we can plot the LUSC samples.

```
d_lusc %>%
  count(shortLetterCode) %>%
  ggplot(., aes(shortLetterCode, n, fill=shortLetterCode)) +
  geom_bar(stat="identity", position=position_dodge())
```



We made two separate plots for the LUAD and LUSC dataset. It would be great if we can merge them into one plot. Here I put the code for this. It would be good practice if you comment or delete the line that you don't understand fully, you will compare the difference between outcomes. Also, please figure out what Qs do in this plot.

```
bind_rows(d_luad %>%
            mutate(type='luad') %>%
            select(type, shortLetterCode, tumor_stage),
          d_lusc %>%
            mutate(type='lusc') %>%
            select(type, shortLetterCode, tumor_stage)) %>%
  count(shortLetterCode, type) %>%
  complete(type, shortLetterCode, fill = list(n = 0)) %>% # Q1: What is this?
  ggplot(., aes(shortLetterCode, n, fill=shortLetterCode)) +
  geom_bar(stat="identity", position=position_dodge()) +
  facet_wrap(~type, scales = 'free_x', ncol=5) # Q2: What is this?
```

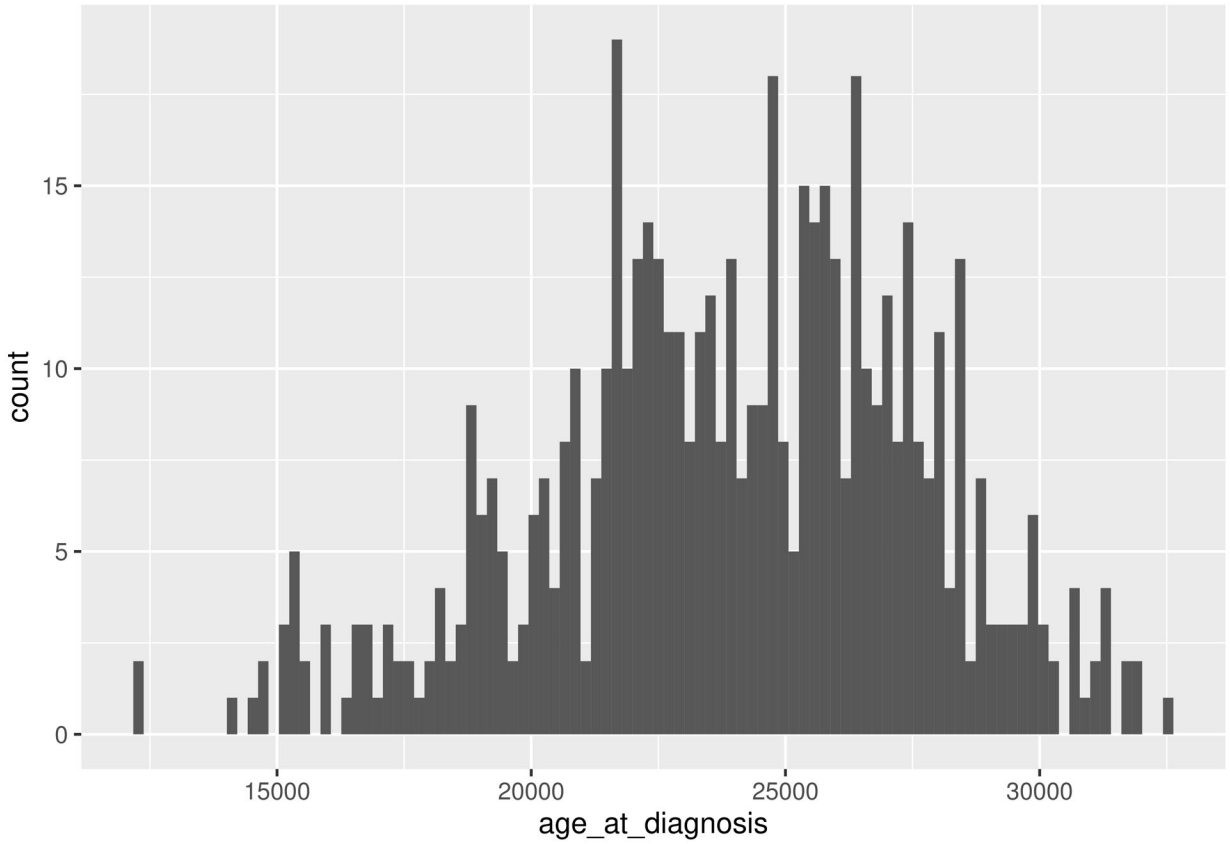


3.2. Distribution of clinical variables

We will plot the distribution of age at diagnosis using histogram. Few things you can check: - Is the distribution continuous? - Is it following the normal distribution? - What is the scale on the x-axis? - If the distribution is stratified, which makes it?

```
# Plot histogram
ggplot(d_luad, aes(age_at_diagnosis)) + geom_histogram(bins=100)
```

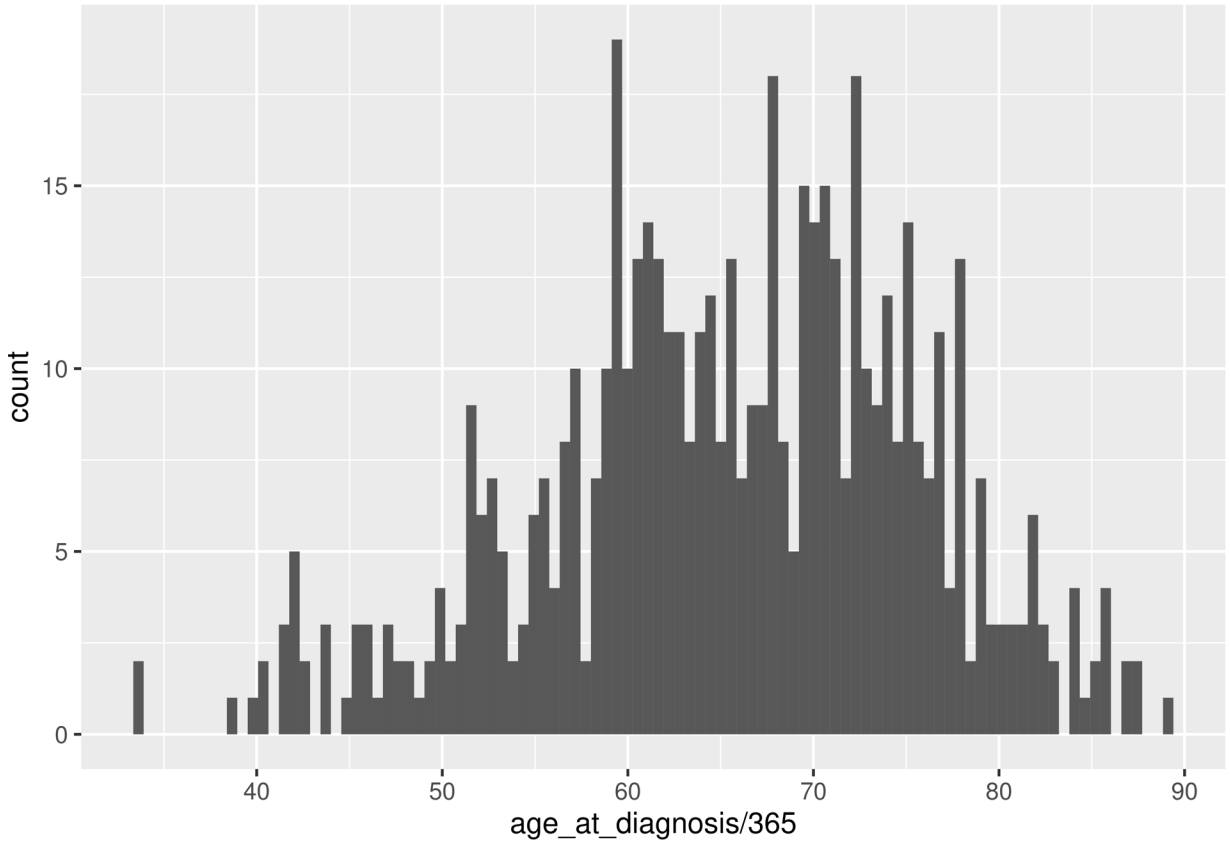
```
## Warning: Removed 37 rows containing non-finite values (stat_bin).
```



The x-axis is a day scale. Let's convert it to year.

```
# Change the axis
ggplot(d_luad, aes(age_at_diagnosis/365)) + geom_histogram(bins=100)
```

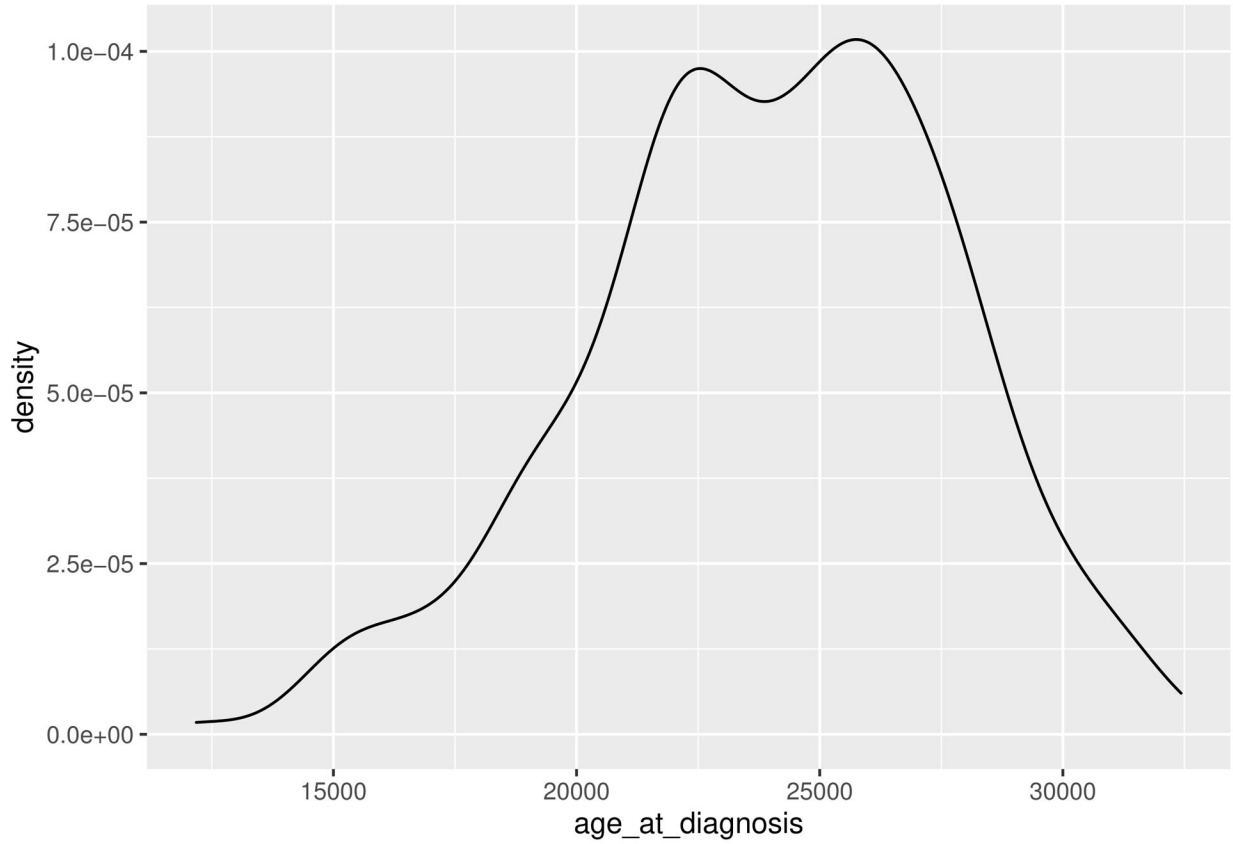
```
## Warning: Removed 37 rows containing non-finite values (stat_bin).
```



To create a smooth density, we use the geom_density

```
ggplot(d_luad, aes(age_at_diagnosis)) + geom_density()
```

```
## Warning: Removed 37 rows containing non-finite values (stat_density).
```

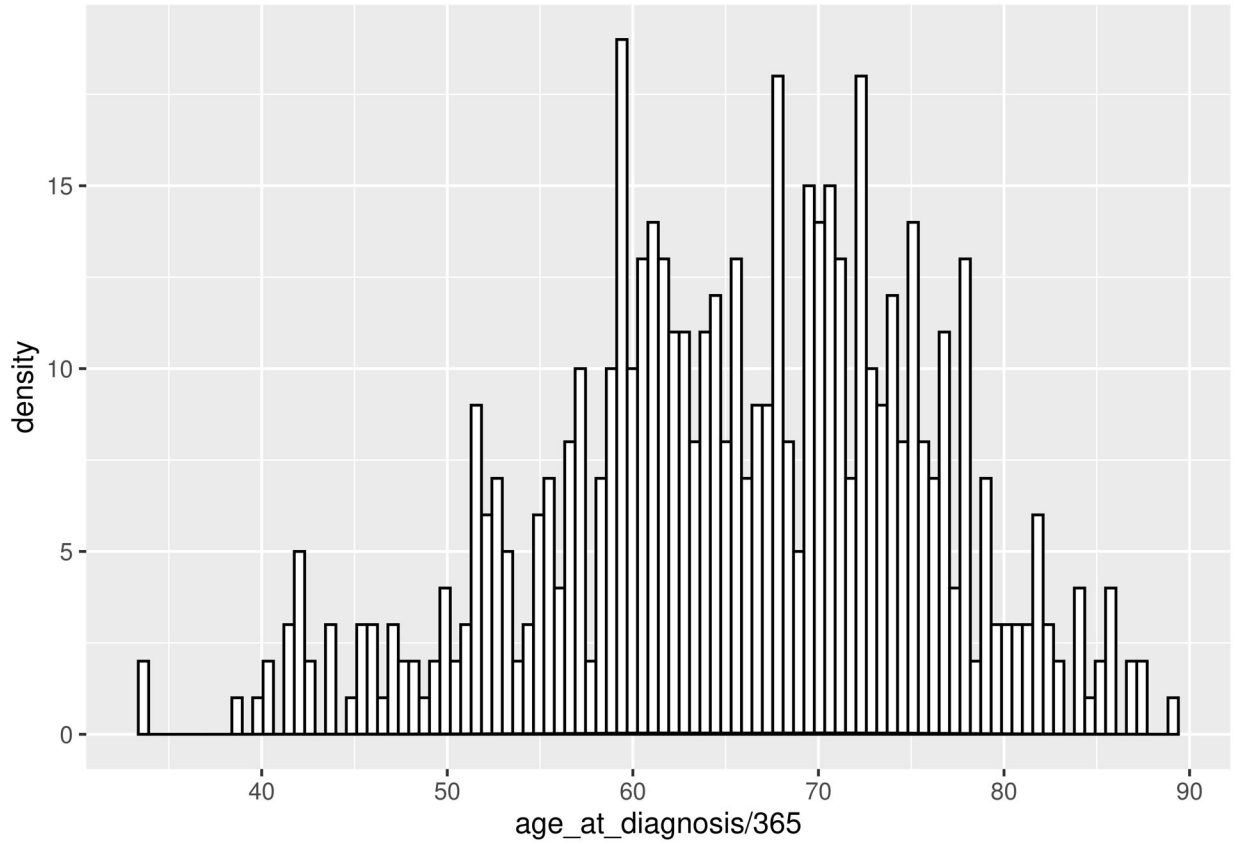


More plot types on distribution. First we can try the plot for both histogram and density.

```
# Plot both histogram and density plot
ggplot(d_luad, aes(age_at_diagnosis/365)) +
  geom_histogram(bins=100, colour="black", fill="white") +
  geom_density(alpha=.2, fill="#FF6666")
```

```
## Warning: Removed 37 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 37 rows containing non-finite values (stat_density).
```



What did you miss from this?

```
#The y-axis units are different for histogram (count) and density (percentage) so density plot cannot b
```

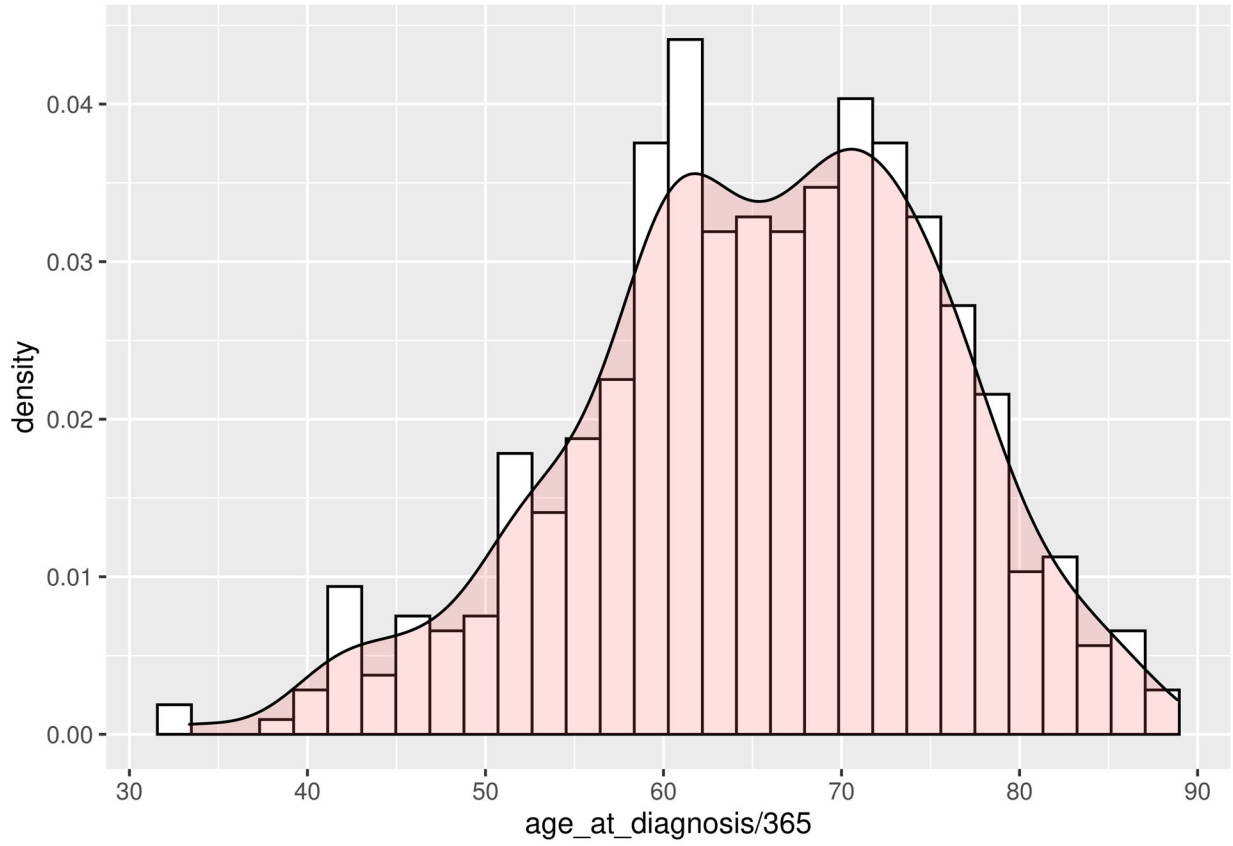
Let's plot a bit different version. We will replace the y-axis of the histogram with the y-axis of the density plot.

```
# Histogram with density plot
ggplot(d_luad, aes(age_at_diagnosis/365)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")
```

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

```
## Warning: Removed 37 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 37 rows containing non-finite values (stat_density).
```

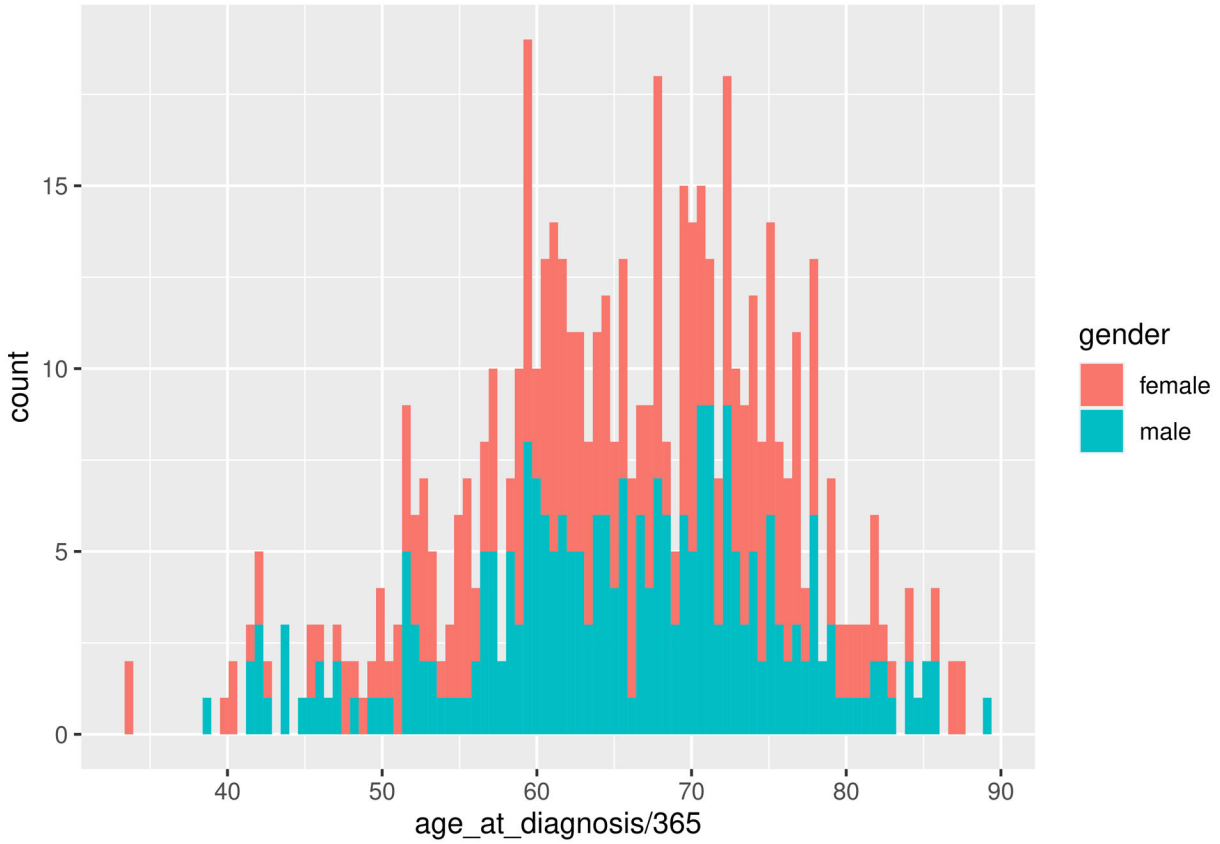


Q: Do you think whether it is good visualization for the distribution?

From the plot above, we still found the bump around the center. What would contribute to this stratification? Let's look at some variables from other columns. First we can try information from the gender column.

```
ggplot(d_luad, aes(age_at_diagnosis/365, fill=gender)) + geom_histogram(bins=100)
```

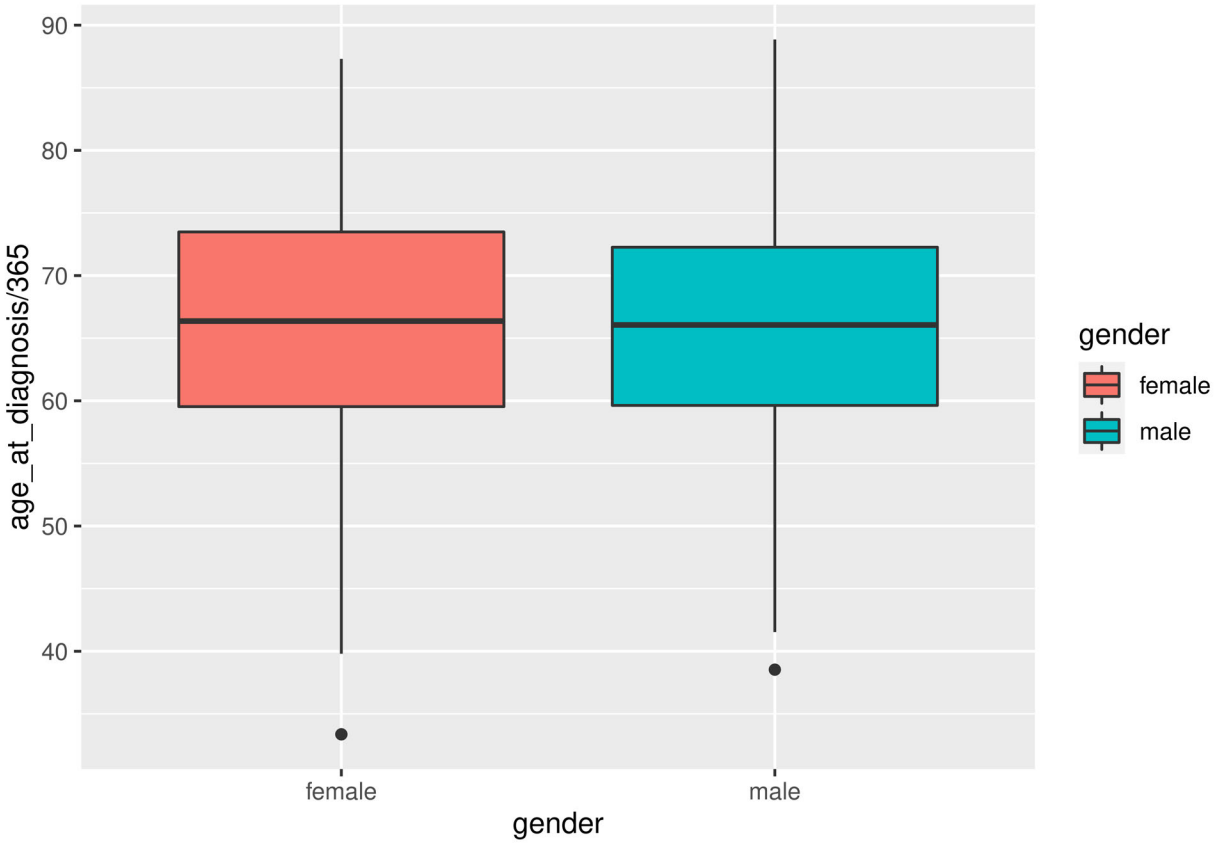
```
## Warning: Removed 37 rows containing non-finite values (stat_bin).
```



Histogram would be okay but not best visualization. What else we can try?

```
ggplot(d_luad, aes(gender, age_at_diagnosis/365, fill=gender)) +
  geom_boxplot()
```

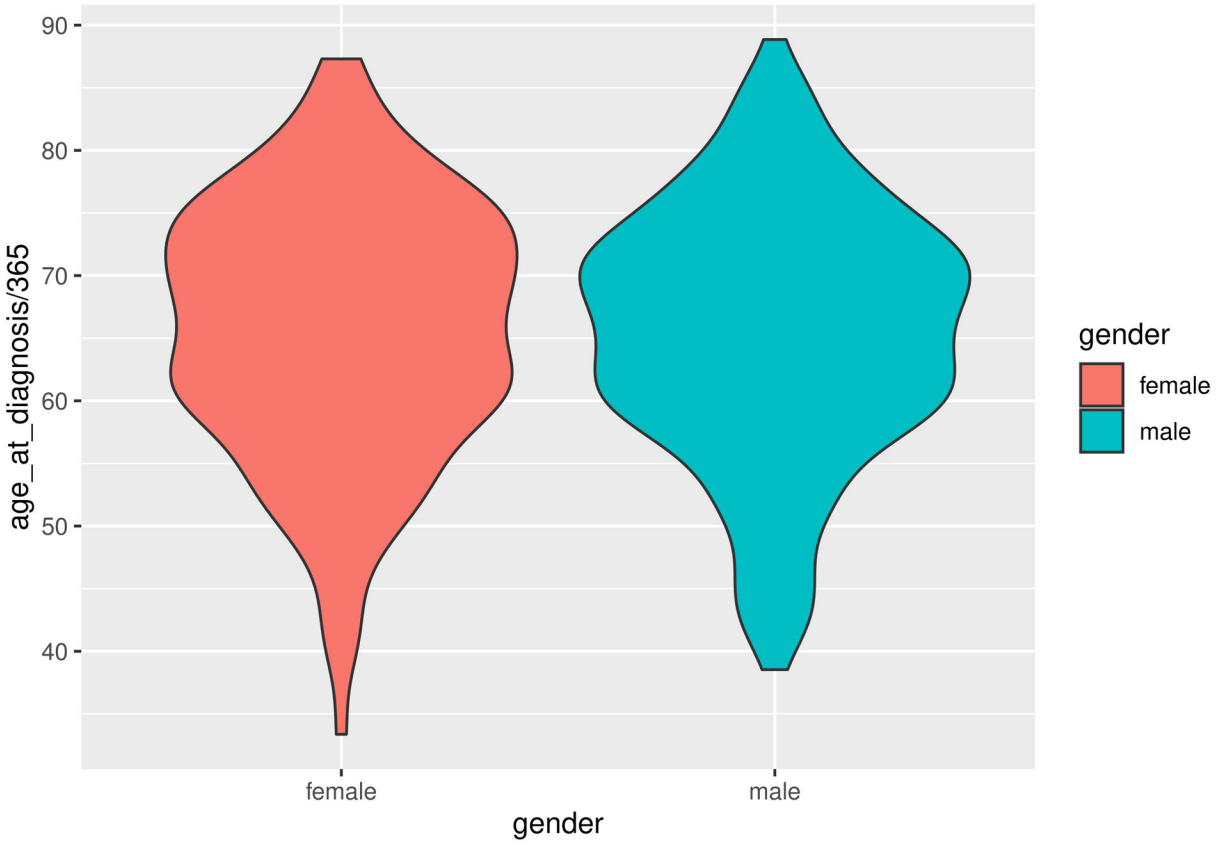
```
## Warning: Removed 37 rows containing non-finite values (stat_boxplot).
```



Can you tell the difference between boxplot and density plot? We can try a violin plot.

```
ggplot(d_luad, aes(gender, age_at_diagnosis/365, fill=gender)) +
  geom_violin()
```

```
## Warning: Removed 37 rows containing non-finite values (stat_ydensity).
```



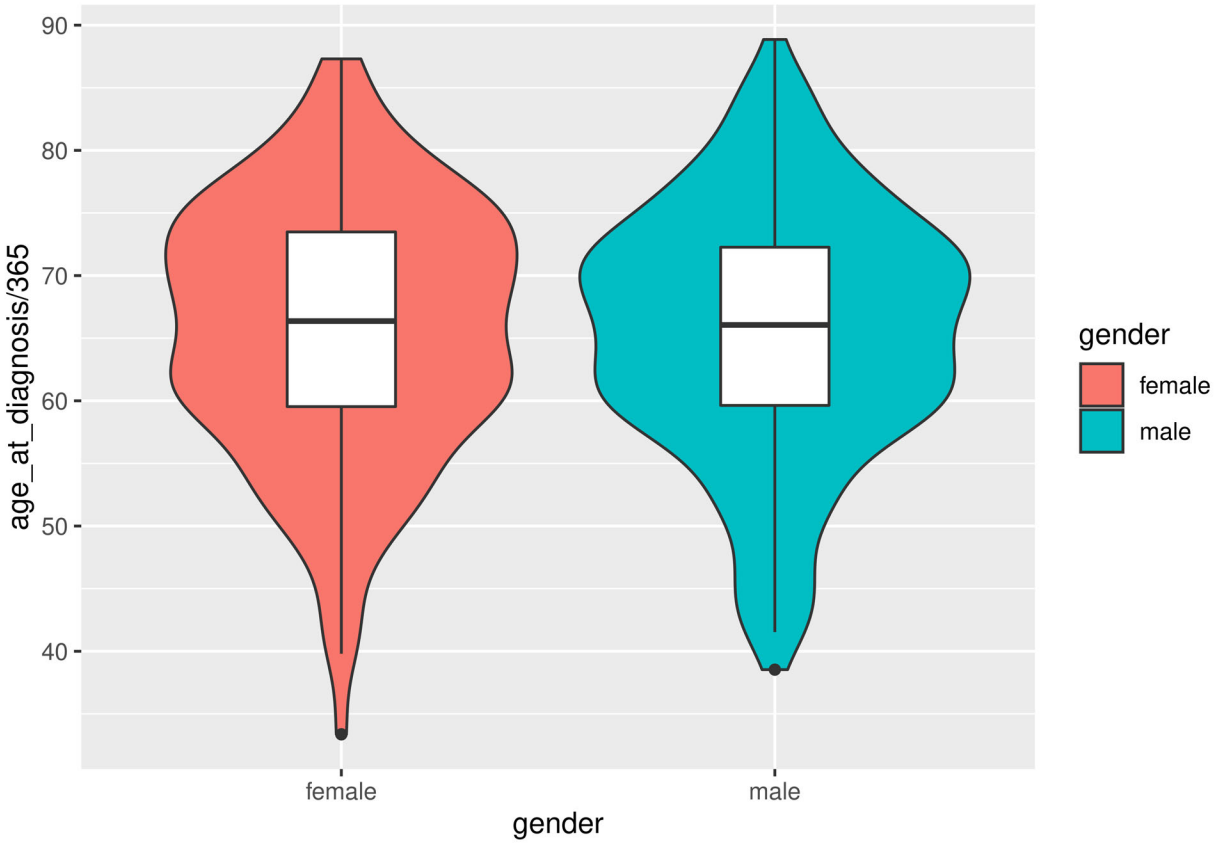
Can you tell the difference between boxplot and violin plot?

```
#violin plot reveals the more detailed distribution like bimodality.
```

```
#ggplot(d_luad, aes(gender, age_at_diagnosis/365, fill=gender)) +
#  geom_boxplot() +
#  geom_violin()

# ggplot(d_luad, aes(gender, age_at_diagnosis/365, fill=gender)) +
#  geom_violin() +
#  geom_boxplot()

# ggplot(d_luad, aes(gender, age_at_diagnosis/365)) +
#   geom_violin(aes(fill=gender)) +
#   geom_boxplot(fill='white')

ggplot(d_luad, aes(gender, age_at_diagnosis/365)) +
  geom_violin(aes(fill=gender)) +
  geom_boxplot(fill='white', width=0.25)
```

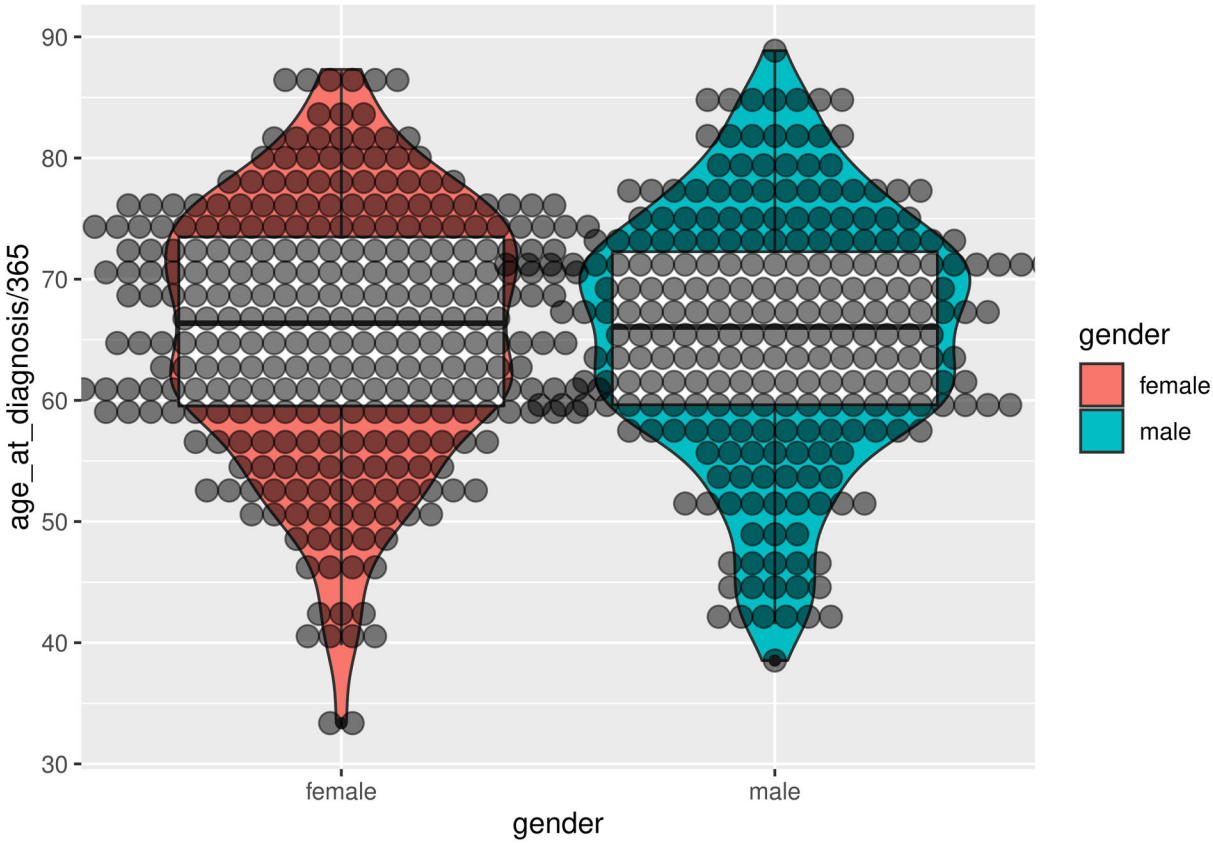
```
## Warning: Removed 37 rows containing non-finite values (stat_ydensity).
```

```
## Warning: Removed 37 rows containing non-finite values (stat_boxplot).
```




```
ggplot(d_luad, aes(gender, age_at_diagnosis/365)) +
  geom_violin(aes(fill=gender))  +
  geom_boxplot(fill='white') +
  geom_dotplot(binaxis='y', stackdir='center', dotsize=1, alpha=0.5)
```

```
## Warning: Removed 37 rows containing non-finite values (stat_ydensity).

## Warning: Removed 37 rows containing non-finite values (stat_boxplot).

## Bin width defaults to 1/30 of the range of the data. Pick better value with `binwidth`.

## Warning: Removed 37 rows containing non-finite values (stat_bindot).
```

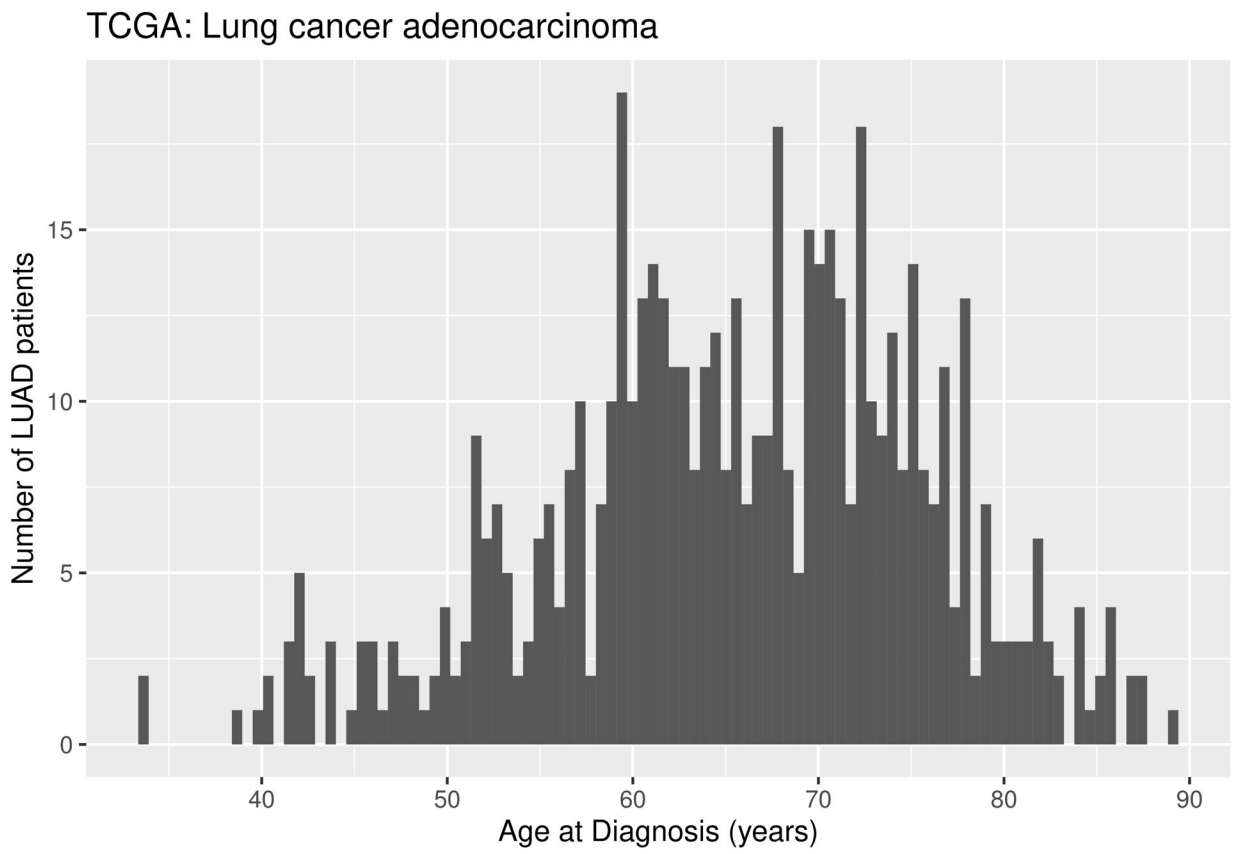



```
# Try different options from the manual
# http://www.sthda.com/english/wiki/ggplot2-violin-plot-quick-start-guide-r-software-and-data-visualiza
```

Do you think gender is the reason? If not, how about tumor stages?

```
ggplot(d_luad, aes(age_at_diagnosis/365)) +
  labs(x ='Age at Diagnosis (years)', y = 'Number of LUAD patients',
       title = 'TCGA: Lung cancer adenocarcinoma') +
  geom_histogram(bins=100)
```

```
## Warning: Removed 37 rows containing non-finite values (stat_bin).
```



More tasks for the class. * Q1. Add labels for the plot. * Q2. Change the color for categories. * Q3. Save to PDF file * Q4. Select the column with continuous information and plot the distribution by yourself.

```
head(d_luad,10)
```

```
##                                                 barcode      patient
## TCGA-55-6982-11A-01R-1949-07 TCGA-55-6982-11A-01R-1949-07 TCGA-55-6982
## TCGA-99-8033-01A-11R-2241-07 TCGA-99-8033-01A-11R-2241-07 TCGA-99-8033
## TCGA-49-6745-01A-11R-1858-07 TCGA-49-6745-01A-11R-1858-07 TCGA-49-6745
## TCGA-49-6767-01A-11R-1858-07 TCGA-49-6767-01A-11R-1858-07 TCGA-49-6767
## TCGA-05-4382-01A-01R-1206-07 TCGA-05-4382-01A-01R-1206-07 TCGA-05-4382
## TCGA-05-5429-01A-01R-1628-07 TCGA-05-5429-01A-01R-1628-07 TCGA-05-5429
## TCGA-38-A44F-01A-11R-A24H-07 TCGA-38-A44F-01A-11R-A24H-07 TCGA-38-A44F
## TCGA-44-8117-01A-11R-2241-07 TCGA-44-8117-01A-11R-2241-07 TCGA-44-8117
## TCGA-71-6725-01A-11R-1858-07 TCGA-71-6725-01A-11R-1858-07 TCGA-71-6725
## TCGA-78-7542-01A-21R-2066-07 TCGA-78-7542-01A-21R-2066-07 TCGA-78-7542
##                                        sample shortLetterCode
## TCGA-55-6982-11A-01R-1949-07 TCGA-55-6982-11A              NT
## TCGA-99-8033-01A-11R-2241-07 TCGA-99-8033-01A              TP
## TCGA-49-6745-01A-11R-1858-07 TCGA-49-6745-01A              TP
## TCGA-49-6767-01A-11R-1858-07 TCGA-49-6767-01A              TP
## TCGA-05-4382-01A-01R-1206-07 TCGA-05-4382-01A              TP
## TCGA-05-5429-01A-01R-1628-07 TCGA-05-5429-01A              TP
## TCGA-38-A44F-01A-11R-A24H-07 TCGA-38-A44F-01A              TP
## TCGA-44-8117-01A-11R-2241-07 TCGA-44-8117-01A              TP
## TCGA-71-6725-01A-11R-1858-07 TCGA-71-6725-01A              TP
```

```
## TCGA-78-7542-01A-21R-2066-07 TCGA-78-7542-01A             TP
##                                        definition sample_submitter_id
## TCGA-55-6982-11A-01R-1949-07 Solid Tissue Normal    TCGA-55-6982-11A
## TCGA-99-8033-01A-11R-2241-07 Primary solid Tumor    TCGA-99-8033-01A
## TCGA-49-6745-01A-11R-1858-07 Primary solid Tumor    TCGA-49-6745-01A
## TCGA-49-6767-01A-11R-1858-07 Primary solid Tumor    TCGA-49-6767-01A
## TCGA-05-4382-01A-01R-1206-07 Primary solid Tumor    TCGA-05-4382-01A
## TCGA-05-5429-01A-01R-1628-07 Primary solid Tumor    TCGA-05-5429-01A
## TCGA-38-A44F-01A-11R-A24H-07 Primary solid Tumor    TCGA-38-A44F-01A
## TCGA-44-8117-01A-11R-2241-07 Primary solid Tumor    TCGA-44-8117-01A
## TCGA-71-6725-01A-11R-1858-07 Primary solid Tumor    TCGA-71-6725-01A
## TCGA-78-7542-01A-21R-2066-07 Primary solid Tumor    TCGA-78-7542-01A
##                                  sample_type_id
## TCGA-55-6982-11A-01R-1949-07             11
## TCGA-99-8033-01A-11R-2241-07             01
## TCGA-49-6745-01A-11R-1858-07             01
## TCGA-49-6767-01A-11R-1858-07             01
## TCGA-05-4382-01A-01R-1206-07             01
## TCGA-05-5429-01A-01R-1628-07             01
## TCGA-38-A44F-01A-11R-A24H-07             01
## TCGA-44-8117-01A-11R-2241-07             01
## TCGA-71-6725-01A-11R-1858-07             01
## TCGA-78-7542-01A-21R-2066-07             01
##                                                sample_id
## TCGA-55-6982-11A-01R-1949-07 ffb185aa-75ac-4696-936b-52f35a345a2d
## TCGA-99-8033-01A-11R-2241-07 5e686ec3-9100-4153-9a90-8b7fb8d4ba0a
## TCGA-49-6745-01A-11R-1858-07 357f3efe-e2c9-432a-a2b6-75a5fdbae3be
## TCGA-49-6767-01A-11R-1858-07 708a8742-f6d6-411d-9340-2c6c25885d22
## TCGA-05-4382-01A-01R-1206-07 cce6d71f-369e-467f-bd7e-03d20e97b7f3
## TCGA-05-5429-01A-01R-1628-07 fbc90b42-7d26-456b-a166-41993a1d3f0d
## TCGA-38-A44F-01A-11R-A24H-07 45c17476-0525-46d2-8448-48d4de6713f1
## TCGA-44-8117-01A-11R-2241-07 d84e1b10-b7c6-4916-afbc-b311d2ccf19d
## TCGA-71-6725-01A-11R-1858-07 83973755-fa6a-460a-a042-385b34698c39
## TCGA-78-7542-01A-21R-2066-07 43b96e20-df08-4c56-915f-5d626f90c415
##                                       sample_type days_to_collection    state
## TCGA-55-6982-11A-01R-1949-07 Solid Tissue Normal                 NA released
## TCGA-99-8033-01A-11R-2241-07       Primary Tumor                 NA released
## TCGA-49-6745-01A-11R-1858-07       Primary Tumor                 NA released
## TCGA-49-6767-01A-11R-1858-07       Primary Tumor                 NA released
## TCGA-05-4382-01A-01R-1206-07       Primary Tumor                 NA released
## TCGA-05-5429-01A-01R-1628-07       Primary Tumor                 NA released
## TCGA-38-A44F-01A-11R-A24H-07       Primary Tumor                127 released
## TCGA-44-8117-01A-11R-2241-07       Primary Tumor                 NA released
## TCGA-71-6725-01A-11R-1858-07       Primary Tumor                 NA released
## TCGA-78-7542-01A-21R-2066-07       Primary Tumor                 NA released
##                                  initial_weight intermediate_dimension
## TCGA-55-6982-11A-01R-1949-07             NA                    0.7
## TCGA-99-8033-01A-11R-2241-07             NA                    0.6
## TCGA-49-6745-01A-11R-1858-07             NA                    0.7
## TCGA-49-6767-01A-11R-1858-07             NA                    0.5
## TCGA-05-4382-01A-01R-1206-07             NA                    0.8
## TCGA-05-5429-01A-01R-1628-07             NA                    0.9
## TCGA-38-A44F-01A-11R-A24H-07            580                     NA
## TCGA-44-8117-01A-11R-2241-07             NA                    0.7
```

```
## TCGA-71-6725-01A-11R-1858-07             NA                    1.0
## TCGA-78-7542-01A-21R-2066-07             NA                    0.8
##                                                pathology_report_uuid submitter_id
## TCGA-55-6982-11A-01R-1949-07                                    <NA> TCGA-55-6982
## TCGA-99-8033-01A-11R-2241-07 a7f10fd8-ab32-4adb-94a9-31632e286b11 TCGA-99-8033
## TCGA-49-6745-01A-11R-1858-07 e291ad97-8658-435d-ae71-f0eb80bbb816 TCGA-49-6745
## TCGA-49-6767-01A-11R-1858-07 8e8171dc-1baa-4b81-8e05-e15714135c1f TCGA-49-6767
## TCGA-05-4382-01A-01R-1206-07 952c0f32-1472-49e1-8334-b0f1de4ac921 TCGA-05-4382
## TCGA-05-5429-01A-01R-1628-07 087C433B-9B20-439A-932C-4C852341A052 TCGA-05-5429
## TCGA-38-A44F-01A-11R-A24H-07 FC3ACA4F-AA49-47F4-9BA3-7B18C56F64AD TCGA-38-A44F
## TCGA-44-8117-01A-11R-2241-07 af49cbe4-eb8a-46df-bd70-4023acee6ae1 TCGA-44-8117
## TCGA-71-6725-01A-11R-1858-07 0d31265b-d292-4550-81cb-509a2ab9a4ce TCGA-71-6725
## TCGA-78-7542-01A-21R-2066-07 07a2fac9-0532-442e-ae75-5a7cc5d298d4 TCGA-78-7542
##                              shortest_dimension oct_embedded longest_dimension
## TCGA-55-6982-11A-01R-1949-07                0.5         <NA>                0.8
## TCGA-99-8033-01A-11R-2241-07                0.5         <NA>                0.6
## TCGA-49-6745-01A-11R-1858-07                0.3         <NA>                1.1
## TCGA-49-6767-01A-11R-1858-07                0.4         <NA>                1.5
## TCGA-05-4382-01A-01R-1206-07                0.3         <NA>                0.9
## TCGA-05-5429-01A-01R-1628-07                0.3         <NA>                1.0
## TCGA-38-A44F-01A-11R-A24H-07                 NA        false                 NA
## TCGA-44-8117-01A-11R-2241-07                0.4         <NA>                1.1
## TCGA-71-6725-01A-11R-1858-07                0.8         <NA>                2.0
## TCGA-78-7542-01A-21R-2066-07                0.6         <NA>                1.0
##                              is_ffpe  tissue_type synchronous_malignancy
## TCGA-55-6982-11A-01R-1949-07   FALSE Not Reported                    No
## TCGA-99-8033-01A-11R-2241-07   FALSE Not Reported                    No
## TCGA-49-6745-01A-11R-1858-07   FALSE Not Reported          Not Reported
## TCGA-49-6767-01A-11R-1858-07   FALSE Not Reported                    No
## TCGA-05-4382-01A-01R-1206-07   FALSE Not Reported          Not Reported
## TCGA-05-5429-01A-01R-1628-07   FALSE Not Reported                    No
## TCGA-38-A44F-01A-11R-A24H-07   FALSE Not Reported                    No
## TCGA-44-8117-01A-11R-2241-07   FALSE Not Reported                    No
## TCGA-71-6725-01A-11R-1858-07   FALSE Not Reported                    No
## TCGA-78-7542-01A-21R-2066-07   FALSE Not Reported                    No
##                              ajcc_pathologic_stage tumor_stage
## TCGA-55-6982-11A-01R-1949-07             Stage IIB   stage iib
## TCGA-99-8033-01A-11R-2241-07              Stage IV    stage iv
## TCGA-49-6745-01A-11R-1858-07            Stage IIIA  stage iiia
## TCGA-49-6767-01A-11R-1858-07             Stage IIB   stage iib
## TCGA-05-4382-01A-01R-1206-07              Stage IB    stage ib
## TCGA-05-5429-01A-01R-1628-07            Stage IIIA  stage iiia
## TCGA-38-A44F-01A-11R-A24H-07              Stage IB    stage ib
## TCGA-44-8117-01A-11R-2241-07              Stage IB    stage ib
## TCGA-71-6725-01A-11R-1858-07              Stage IB    stage ib
## TCGA-78-7542-01A-21R-2066-07              Stage IB    stage ib
##                              days_to_diagnosis   treatments
## TCGA-55-6982-11A-01R-1949-07                 0 c(NA, NA....
## TCGA-99-8033-01A-11R-2241-07                 0 c(NA, NA....
## TCGA-49-6745-01A-11R-1858-07                 0 c(NA, NA....
## TCGA-49-6767-01A-11R-1858-07                 0 c(NA, NA....
## TCGA-05-4382-01A-01R-1206-07                 0 c(NA, NA....
## TCGA-05-5429-01A-01R-1628-07                 0 c(NA, NA....
## TCGA-38-A44F-01A-11R-A24H-07                 0 c(NA, NA....
```

```
## TCGA-44-8117-01A-11R-2241-07                0 c(NA, NA....
## TCGA-71-6725-01A-11R-1858-07                0 c(NA, NA....
## TCGA-78-7542-01A-21R-2066-07                0 c(NA, NA....
##                              last_known_disease_status
## TCGA-55-6982-11A-01R-1949-07              not reported
## TCGA-99-8033-01A-11R-2241-07              not reported
## TCGA-49-6745-01A-11R-1858-07              not reported
## TCGA-49-6767-01A-11R-1858-07              not reported
## TCGA-05-4382-01A-01R-1206-07              not reported
## TCGA-05-5429-01A-01R-1628-07              not reported
## TCGA-38-A44F-01A-11R-A24H-07              not reported
## TCGA-44-8117-01A-11R-2241-07              not reported
## TCGA-71-6725-01A-11R-1858-07              not reported
## TCGA-78-7542-01A-21R-2066-07              not reported
##                              tissue_or_organ_of_origin days_to_last_follow_up
## TCGA-55-6982-11A-01R-1949-07          Lower lobe, lung                     NA
## TCGA-99-8033-01A-11R-2241-07          Upper lobe, lung                    170
## TCGA-49-6745-01A-11R-1858-07          Upper lobe, lung                    522
## TCGA-49-6767-01A-11R-1858-07         Middle lobe, lung                    677
## TCGA-05-4382-01A-01R-1206-07          Upper lobe, lung                    607
## TCGA-05-5429-01A-01R-1628-07                 Lung, NOS                     30
## TCGA-38-A44F-01A-11R-A24H-07          Upper lobe, lung                    133
## TCGA-44-8117-01A-11R-2241-07          Upper lobe, lung                    385
## TCGA-71-6725-01A-11R-1858-07          Upper lobe, lung                    256
## TCGA-78-7542-01A-21R-2066-07          Upper lobe, lung                     NA
##                              age_at_diagnosis
## TCGA-55-6982-11A-01R-1949-07               NA
## TCGA-99-8033-01A-11R-2241-07            27342
## TCGA-49-6745-01A-11R-1858-07            30133
## TCGA-49-6767-01A-11R-1858-07            17108
## TCGA-05-4382-01A-01R-1206-07            24868
## TCGA-05-5429-01A-01R-1628-07            22066
## TCGA-38-A44F-01A-11R-A24H-07            29534
## TCGA-44-8117-01A-11R-2241-07            19855
## TCGA-71-6725-01A-11R-1858-07            17793
## TCGA-78-7542-01A-21R-2066-07            20763
##                                          primary_diagnosis
## TCGA-55-6982-11A-01R-1949-07                Adenocarcinoma, NOS
## TCGA-99-8033-01A-11R-2241-07                Adenocarcinoma, NOS
## TCGA-49-6745-01A-11R-1858-07      Papillary adenocarcinoma, NOS
## TCGA-49-6767-01A-11R-1858-07                Adenocarcinoma, NOS
## TCGA-05-4382-01A-01R-1206-07 Adenocarcinoma with mixed subtypes
## TCGA-05-5429-01A-01R-1628-07                Adenocarcinoma, NOS
## TCGA-38-A44F-01A-11R-A24H-07               Acinar cell carcinoma
## TCGA-44-8117-01A-11R-2241-07                Adenocarcinoma, NOS
## TCGA-71-6725-01A-11R-1858-07      Papillary adenocarcinoma, NOS
## TCGA-78-7542-01A-21R-2066-07                Adenocarcinoma, NOS
##                              prior_malignancy year_of_diagnosis prior_treatment
## TCGA-55-6982-11A-01R-1949-07               no              2004              No
## TCGA-99-8033-01A-11R-2241-07               no              2011              No
## TCGA-49-6745-01A-11R-1858-07              yes              2011              No
## TCGA-49-6767-01A-11R-1858-07               no              2010              No
## TCGA-05-4382-01A-01R-1206-07              yes              2009              No
## TCGA-05-5429-01A-01R-1628-07               no              2010              No
```

```
## TCGA-38-A44F-01A-11R-A24H-07              no             2012             No
## TCGA-44-8117-01A-11R-2241-07              no             2011             No
## TCGA-71-6725-01A-11R-1858-07              no             2011             No
## TCGA-78-7542-01A-21R-2066-07              no             1993             No
##                              ajcc_staging_system_edition ajcc_pathologic_t
## TCGA-55-6982-11A-01R-1949-07                         6th               T2
## TCGA-99-8033-01A-11R-2241-07                         7th               TX
## TCGA-49-6745-01A-11R-1858-07                         7th              T2a
## TCGA-49-6767-01A-11R-1858-07                         7th               T3
## TCGA-05-4382-01A-01R-1206-07                         6th               T2
## TCGA-05-5429-01A-01R-1628-07                         7th               T3
## TCGA-38-A44F-01A-11R-A24H-07                         7th              T2a
## TCGA-44-8117-01A-11R-2241-07                         7th              T2a
## TCGA-71-6725-01A-11R-1858-07                         6th               T2
## TCGA-78-7542-01A-21R-2066-07                         6th               T2
##                              morphology ajcc_pathologic_n ajcc_pathologic_m
## TCGA-55-6982-11A-01R-1949-07    8140/3               N1               M0
## TCGA-99-8033-01A-11R-2241-07    8140/3               NX               M1
## TCGA-49-6745-01A-11R-1858-07    8260/3               N2               M0
## TCGA-49-6767-01A-11R-1858-07    8140/3               N0               MX
## TCGA-05-4382-01A-01R-1206-07    8255/3               N0               M0
## TCGA-05-5429-01A-01R-1628-07    8140/3               N2               M0
## TCGA-38-A44F-01A-11R-A24H-07    8550/3               N0               M0
## TCGA-44-8117-01A-11R-2241-07    8140/3               N0               M0
## TCGA-71-6725-01A-11R-1858-07    8260/3               N0               M0
## TCGA-78-7542-01A-21R-2066-07    8140/3               N0               M0
##                              classification_of_tumor
## TCGA-55-6982-11A-01R-1949-07            not reported
## TCGA-99-8033-01A-11R-2241-07            not reported
## TCGA-49-6745-01A-11R-1858-07            not reported
## TCGA-49-6767-01A-11R-1858-07            not reported
## TCGA-05-4382-01A-01R-1206-07            not reported
## TCGA-05-5429-01A-01R-1628-07            not reported
## TCGA-38-A44F-01A-11R-A24H-07            not reported
## TCGA-44-8117-01A-11R-2241-07            not reported
## TCGA-71-6725-01A-11R-1858-07            not reported
## TCGA-78-7542-01A-21R-2066-07            not reported
##                                                 diagnosis_id icd_10_code
## TCGA-55-6982-11A-01R-1949-07 78008708-c100-5080-8d97-4666f0840a4b       C34.3
## TCGA-99-8033-01A-11R-2241-07 0fe120c3-7b42-5c53-ad08-bf9e7a1ca154       C34.1
## TCGA-49-6745-01A-11R-1858-07 f17af2ab-6e23-55d4-946b-99873bc2e707       C34.1
## TCGA-49-6767-01A-11R-1858-07 7266a073-3136-5348-85d6-04625138a28a       C34.2
## TCGA-05-4382-01A-01R-1206-07 eece3e4c-7790-5d2a-bfa5-68bfb23ab769       C34.1
## TCGA-05-5429-01A-01R-1628-07 06eed07a-e385-5f1c-a0e6-c57bbe8a5aae       C34.1
## TCGA-38-A44F-01A-11R-A24H-07 da0e718e-d622-5e8a-874f-b7f3d74832cd       C34.1
## TCGA-44-8117-01A-11R-2241-07 f86b9b90-0009-56d6-afa2-b57635c9f168       C34.1
## TCGA-71-6725-01A-11R-1858-07 71b3c9c2-e113-506c-ab80-1b5c13dd0ff6       C34.1
## TCGA-78-7542-01A-21R-2066-07 d1d15f1e-57d9-5b74-b37a-32228cf63cf4       C34.1
##                              site_of_resection_or_biopsy  tumor_grade
## TCGA-55-6982-11A-01R-1949-07           Lower lobe, lung not reported
## TCGA-99-8033-01A-11R-2241-07           Upper lobe, lung not reported
## TCGA-49-6745-01A-11R-1858-07           Upper lobe, lung not reported
## TCGA-49-6767-01A-11R-1858-07          Middle lobe, lung not reported
## TCGA-05-4382-01A-01R-1206-07           Upper lobe, lung not reported
```

```
## TCGA-05-5429-01A-01R-1628-07                  Lung, NOS not reported
## TCGA-38-A44F-01A-11R-A24H-07           Upper lobe, lung not reported
## TCGA-44-8117-01A-11R-2241-07           Upper lobe, lung not reported
## TCGA-71-6725-01A-11R-1858-07           Upper lobe, lung not reported
## TCGA-78-7542-01A-21R-2066-07           Upper lobe, lung not reported
##                               progression_or_recurrence cigarettes_per_day
## TCGA-55-6982-11A-01R-1949-07              not reported                NA
## TCGA-99-8033-01A-11R-2241-07              not reported                NA
## TCGA-49-6745-01A-11R-1858-07              not reported         1.0958904
## TCGA-49-6767-01A-11R-1858-07              not reported         1.6438356
## TCGA-05-4382-01A-01R-1206-07              not reported         3.3972603
## TCGA-05-5429-01A-01R-1628-07              not reported                NA
## TCGA-38-A44F-01A-11R-A24H-07              not reported         0.6575342
## TCGA-44-8117-01A-11R-2241-07              not reported         2.9589041
## TCGA-71-6725-01A-11R-1858-07              not reported                NA
## TCGA-78-7542-01A-21R-2066-07              not reported         3.2876712
##                               alcohol_history
## TCGA-55-6982-11A-01R-1949-07    Not Reported
## TCGA-99-8033-01A-11R-2241-07    Not Reported
## TCGA-49-6745-01A-11R-1858-07    Not Reported
## TCGA-49-6767-01A-11R-1858-07    Not Reported
## TCGA-05-4382-01A-01R-1206-07    Not Reported
## TCGA-05-5429-01A-01R-1628-07    Not Reported
## TCGA-38-A44F-01A-11R-A24H-07    Not Reported
## TCGA-44-8117-01A-11R-2241-07    Not Reported
## TCGA-71-6725-01A-11R-1858-07    Not Reported
## TCGA-78-7542-01A-21R-2066-07    Not Reported
##                                                      exposure_id years_smoked
## TCGA-55-6982-11A-01R-1949-07 1d9b9b61-89f3-5b30-8ee2-1ecaf1498ea8           NA
## TCGA-99-8033-01A-11R-2241-07 144ee914-ba8a-5d4d-bf91-b257c12eb162           NA
## TCGA-49-6745-01A-11R-1858-07 b6e58fb2-f1ec-539f-a467-c8394af02460           NA
## TCGA-49-6767-01A-11R-1858-07 41270366-24bb-5315-a254-22cd7c74fa2c           NA
## TCGA-05-4382-01A-01R-1206-07 ef57bd0c-1455-550d-a1eb-06f2d5620c8e           NA
## TCGA-05-5429-01A-01R-1628-07 9b1bed7c-7b43-5fa5-9cfc-630d620435d0           NA
## TCGA-38-A44F-01A-11R-A24H-07 6e5a723e-37c2-5f59-a502-8335b30af5e6           12
## TCGA-44-8117-01A-11R-2241-07 cc92b3af-cbe1-5a17-8eab-93a6cd6a6b93           36
## TCGA-71-6725-01A-11R-1858-07 0668841b-8d9c-5fbe-b377-781326034024           NA
## TCGA-78-7542-01A-21R-2066-07 c24ecf41-e2b6-5c40-adf6-7dacfa6519e2           NA
##                               pack_years_smoked gender              ethnicity
## TCGA-55-6982-11A-01R-1949-07                NA female           not reported
## TCGA-99-8033-01A-11R-2241-07                NA female not hispanic or latino
## TCGA-49-6745-01A-11R-1858-07                20   male           not reported
## TCGA-49-6767-01A-11R-1858-07                30 female           not reported
## TCGA-05-4382-01A-01R-1206-07                62   male           not reported
## TCGA-05-5429-01A-01R-1628-07                NA   male           not reported
## TCGA-38-A44F-01A-11R-A24H-07                12   male not hispanic or latino
## TCGA-44-8117-01A-11R-2241-07                54 female not hispanic or latino
## TCGA-71-6725-01A-11R-1858-07                NA female not hispanic or latino
## TCGA-78-7542-01A-21R-2066-07                60   male           not reported
##                                race vital_status age_at_index
## TCGA-55-6982-11A-01R-1949-07  white         Dead           79
## TCGA-99-8033-01A-11R-2241-07  white         Dead           74
## TCGA-49-6745-01A-11R-1858-07  white        Alive           82
## TCGA-49-6767-01A-11R-1858-07  white        Alive           46
```

```
## TCGA-05-4382-01A-01R-1206-07 not reported        Alive           68
## TCGA-05-5429-01A-01R-1628-07 not reported         Dead           60
## TCGA-38-A44F-01A-11R-A24H-07        white        Alive           80
## TCGA-44-8117-01A-11R-2241-07        white        Alive           54
## TCGA-71-6725-01A-11R-1858-07        asian        Alive           48
## TCGA-78-7542-01A-21R-2066-07        white         Dead           56
##                               days_to_birth year_of_birth
## TCGA-55-6982-11A-01R-1949-07            NA          1925
## TCGA-99-8033-01A-11R-2241-07        -27342          1937
## TCGA-49-6745-01A-11R-1858-07        -30133          1929
## TCGA-49-6767-01A-11R-1858-07        -17108          1964
## TCGA-05-4382-01A-01R-1206-07        -24868          1941
## TCGA-05-5429-01A-01R-1628-07        -22066          1950
## TCGA-38-A44F-01A-11R-A24H-07        -29534          1932
## TCGA-44-8117-01A-11R-2241-07        -19855          1957
## TCGA-71-6725-01A-11R-1858-07        -17793          1963
## TCGA-78-7542-01A-21R-2066-07        -20763          1937
##                                                 demographic_id days_to_death
## TCGA-55-6982-11A-01R-1949-07 0fe268b0-fcba-5ee9-a593-d484d7e9324d           995
## TCGA-99-8033-01A-11R-2241-07 be35b84c-8c75-5fd5-ad4d-6e7d55ae36c4           656
## TCGA-49-6745-01A-11R-1858-07 96208e4a-c1c5-5e28-b8e5-c6e6e4e096c8            NA
## TCGA-49-6767-01A-11R-1858-07 fcadd633-ccce-577d-a1a3-be171e9128df            NA
## TCGA-05-4382-01A-01R-1206-07 c2f0c511-8275-502d-b819-2405c10d0fbd            NA
## TCGA-05-5429-01A-01R-1628-07 97dbd192-260d-5a7b-a942-2f5117d24934           275
## TCGA-38-A44F-01A-11R-A24H-07 0d71918d-d17d-5726-a5ce-c9e65d7ae4dc            NA
## TCGA-44-8117-01A-11R-2241-07 37435a10-39c6-5b05-85dc-4511664352f2            NA
## TCGA-71-6725-01A-11R-1858-07 926be778-dafa-583b-ab44-194c2a63cef3            NA
## TCGA-78-7542-01A-21R-2066-07 c0ae101c-b643-5333-86bf-6c2484db38ac           321
##                               year_of_death bcr_patient_barcode primary_site
## TCGA-55-6982-11A-01R-1949-07          2006    TCGA-55-6982-11A Bronchus....
## TCGA-99-8033-01A-11R-2241-07            NA    TCGA-99-8033-01A Bronchus....
## TCGA-49-6745-01A-11R-1858-07            NA    TCGA-49-6745-01A Bronchus....
## TCGA-49-6767-01A-11R-1858-07            NA    TCGA-49-6767-01A Bronchus....
## TCGA-05-4382-01A-01R-1206-07            NA    TCGA-05-4382-01A Bronchus....
## TCGA-05-5429-01A-01R-1628-07            NA    TCGA-05-5429-01A Bronchus....
## TCGA-38-A44F-01A-11R-A24H-07            NA    TCGA-38-A44F-01A Bronchus....
## TCGA-44-8117-01A-11R-2241-07            NA    TCGA-44-8117-01A Bronchus....
## TCGA-71-6725-01A-11R-1858-07            NA    TCGA-71-6725-01A Bronchus....
## TCGA-78-7542-01A-21R-2066-07          1993    TCGA-78-7542-01A Bronchus....
##                               disease_type project_id releasable
## TCGA-55-6982-11A-01R-1949-07 Acinar C....  TCGA-LUAD       TRUE
## TCGA-99-8033-01A-11R-2241-07 Acinar C....  TCGA-LUAD       TRUE
## TCGA-49-6745-01A-11R-1858-07 Acinar C....  TCGA-LUAD       TRUE
## TCGA-49-6767-01A-11R-1858-07 Acinar C....  TCGA-LUAD       TRUE
## TCGA-05-4382-01A-01R-1206-07 Acinar C....  TCGA-LUAD       TRUE
## TCGA-05-5429-01A-01R-1628-07 Acinar C....  TCGA-LUAD       TRUE
## TCGA-38-A44F-01A-11R-A24H-07 Acinar C....  TCGA-LUAD       TRUE
## TCGA-44-8117-01A-11R-2241-07 Acinar C....  TCGA-LUAD       TRUE
## TCGA-71-6725-01A-11R-1858-07 Acinar C....  TCGA-LUAD       TRUE
## TCGA-78-7542-01A-21R-2066-07 Acinar C....  TCGA-LUAD       TRUE
##                                                 name released paper_patient
## TCGA-55-6982-11A-01R-1949-07 Lung Adenocarcinoma     TRUE          <NA>
## TCGA-99-8033-01A-11R-2241-07 Lung Adenocarcinoma     TRUE          <NA>
## TCGA-49-6745-01A-11R-1858-07 Lung Adenocarcinoma     TRUE  TCGA-49-6745
```

```
## TCGA-49-6767-01A-11R-1858-07 Lung Adenocarcinoma     TRUE  TCGA-49-6767
## TCGA-05-4382-01A-01R-1206-07 Lung Adenocarcinoma     TRUE  TCGA-05-4382
## TCGA-05-5429-01A-01R-1628-07 Lung Adenocarcinoma     TRUE  TCGA-05-5429
## TCGA-38-A44F-01A-11R-A24H-07 Lung Adenocarcinoma     TRUE          <NA>
## TCGA-44-8117-01A-11R-2241-07 Lung Adenocarcinoma     TRUE          <NA>
## TCGA-71-6725-01A-11R-1858-07 Lung Adenocarcinoma     TRUE  TCGA-71-6725
## TCGA-78-7542-01A-21R-2066-07 Lung Adenocarcinoma     TRUE  TCGA-78-7542
##                              paper_Sex paper_Age.at.diagnosis paper_T.stage
## TCGA-55-6982-11A-01R-1949-07      <NA>                  <NA>          <NA>
## TCGA-99-8033-01A-11R-2241-07      <NA>                  <NA>          <NA>
## TCGA-49-6745-01A-11R-1858-07      MALE                    82           T2a
## TCGA-49-6767-01A-11R-1858-07    FEMALE                    46            T3
## TCGA-05-4382-01A-01R-1206-07      MALE                    68            T2
## TCGA-05-5429-01A-01R-1628-07      MALE                    60            T3
## TCGA-38-A44F-01A-11R-A24H-07      <NA>                  <NA>          <NA>
## TCGA-44-8117-01A-11R-2241-07      <NA>                  <NA>          <NA>
## TCGA-71-6725-01A-11R-1858-07    FEMALE                    48            T1
## TCGA-78-7542-01A-21R-2066-07      MALE                    56            T2
##                              paper_N.stage paper_Tumor.stage
## TCGA-55-6982-11A-01R-1949-07          <NA>              <NA>
## TCGA-99-8033-01A-11R-2241-07          <NA>              <NA>
## TCGA-49-6745-01A-11R-1858-07            N2        Stage IIIA
## TCGA-49-6767-01A-11R-1858-07            N0         Stage IIB
## TCGA-05-4382-01A-01R-1206-07            N0          Stage IB
## TCGA-05-5429-01A-01R-1628-07            N2        Stage IIIA
## TCGA-38-A44F-01A-11R-A24H-07          <NA>              <NA>
## TCGA-44-8117-01A-11R-2241-07          <NA>              <NA>
## TCGA-71-6725-01A-11R-1858-07            N0          Stage IA
## TCGA-78-7542-01A-21R-2066-07            N0          Stage IB
##                                                paper_Smoking.Status
## TCGA-55-6982-11A-01R-1949-07                                   <NA>
## TCGA-99-8033-01A-11R-2241-07                                   <NA>
## TCGA-49-6745-01A-11R-1858-07 Current reformed smoker for < or = 15 years
## TCGA-49-6767-01A-11R-1858-07                          Current smoker
## TCGA-05-4382-01A-01R-1206-07 Current reformed smoker for < or = 15 years
## TCGA-05-5429-01A-01R-1628-07                       Lifelong Non-smoker
## TCGA-38-A44F-01A-11R-A24H-07                                   <NA>
## TCGA-44-8117-01A-11R-2241-07                                   <NA>
## TCGA-71-6725-01A-11R-1858-07                          Current smoker
## TCGA-78-7542-01A-21R-2066-07                          Current smoker
##                              paper_Survival paper_Transversion.High.Low
## TCGA-55-6982-11A-01R-1949-07          <NA>                        <NA>
## TCGA-99-8033-01A-11R-2241-07          <NA>                        <NA>
## TCGA-49-6745-01A-11R-1858-07        LIVING                         Low
## TCGA-49-6767-01A-11R-1858-07        LIVING                        High
## TCGA-05-4382-01A-01R-1206-07        LIVING                        High
## TCGA-05-5429-01A-01R-1628-07        LIVING                         Low
## TCGA-38-A44F-01A-11R-A24H-07          <NA>                        <NA>
## TCGA-44-8117-01A-11R-2241-07          <NA>                        <NA>
## TCGA-71-6725-01A-11R-1858-07        LIVING                         Low
## TCGA-78-7542-01A-21R-2066-07      DECEASED                        High
##                              paper_Nonsilent.Mutations
## TCGA-55-6982-11A-01R-1949-07                        NA
## TCGA-99-8033-01A-11R-2241-07                        NA
```

```
## TCGA-49-6745-01A-11R-1858-07                      121
## TCGA-49-6767-01A-11R-1858-07                      410
## TCGA-05-4382-01A-01R-1206-07                     1316
## TCGA-05-5429-01A-01R-1628-07                       37
## TCGA-38-A44F-01A-11R-A24H-07                       NA
## TCGA-44-8117-01A-11R-2241-07                       NA
## TCGA-71-6725-01A-11R-1858-07                       51
## TCGA-78-7542-01A-21R-2066-07                      313
##                              paper_Nonsilent.Mutations.per.Mb
## TCGA-55-6982-11A-01R-1949-07                             <NA>
## TCGA-99-8033-01A-11R-2241-07                             <NA>
## TCGA-49-6745-01A-11R-1858-07                             3,47
## TCGA-49-6767-01A-11R-1858-07                            11,77
## TCGA-05-4382-01A-01R-1206-07                            37,78
## TCGA-05-5429-01A-01R-1628-07                             1,06
## TCGA-38-A44F-01A-11R-A24H-07                             <NA>
## TCGA-44-8117-01A-11R-2241-07                             <NA>
## TCGA-71-6725-01A-11R-1858-07                             1,46
## TCGA-78-7542-01A-21R-2066-07                             8,99
##                              paper_Oncogene.Negative.or.Positive.Groups
## TCGA-55-6982-11A-01R-1949-07                                       <NA>
## TCGA-99-8033-01A-11R-2241-07                                       <NA>
## TCGA-49-6745-01A-11R-1858-07                          Oncogene Positive
## TCGA-49-6767-01A-11R-1858-07                          Oncogene Negative
## TCGA-05-4382-01A-01R-1206-07                          Oncogene Negative
## TCGA-05-5429-01A-01R-1628-07                          Oncogene Negative
## TCGA-38-A44F-01A-11R-A24H-07                                       <NA>
## TCGA-44-8117-01A-11R-2241-07                                       <NA>
## TCGA-71-6725-01A-11R-1858-07                          Oncogene Negative
## TCGA-78-7542-01A-21R-2066-07                          Oncogene Negative
##                              paper_Fusions paper_expression_subtype
## TCGA-55-6982-11A-01R-1949-07          <NA>                     <NA>
## TCGA-99-8033-01A-11R-2241-07          <NA>                     <NA>
## TCGA-49-6745-01A-11R-1858-07                            prox.-inflam
## TCGA-49-6767-01A-11R-1858-07                            prox.-inflam
## TCGA-05-4382-01A-01R-1206-07                            prox.-inflam
## TCGA-05-5429-01A-01R-1628-07                           prox.-prolif.
## TCGA-38-A44F-01A-11R-A24H-07          <NA>                     <NA>
## TCGA-44-8117-01A-11R-2241-07          <NA>                     <NA>
## TCGA-71-6725-01A-11R-1858-07                                     TRU
## TCGA-78-7542-01A-21R-2066-07                            prox.-inflam
##                              paper_chromosome.affected.by.chromothripsis
## TCGA-55-6982-11A-01R-1949-07                                        <NA>
## TCGA-99-8033-01A-11R-2241-07                                        <NA>
## TCGA-49-6745-01A-11R-1858-07                                        <NA>
## TCGA-49-6767-01A-11R-1858-07                                        <NA>
## TCGA-05-4382-01A-01R-1206-07                                        <NA>
## TCGA-05-5429-01A-01R-1628-07                                        <NA>
## TCGA-38-A44F-01A-11R-A24H-07                                        <NA>
## TCGA-44-8117-01A-11R-2241-07                                        <NA>
## TCGA-71-6725-01A-11R-1858-07                                        <NA>
## TCGA-78-7542-01A-21R-2066-07                                        <NA>
##                              paper_iCluster.Group
## TCGA-55-6982-11A-01R-1949-07                   NA
```

```
## TCGA-99-8033-01A-11R-2241-07                   NA
## TCGA-49-6745-01A-11R-1858-07                    4
## TCGA-49-6767-01A-11R-1858-07                    3
## TCGA-05-4382-01A-01R-1206-07                    6
## TCGA-05-5429-01A-01R-1628-07                    4
## TCGA-38-A44F-01A-11R-A24H-07                   NA
## TCGA-44-8117-01A-11R-2241-07                   NA
## TCGA-71-6725-01A-11R-1858-07                    3
## TCGA-78-7542-01A-21R-2066-07                    3
##                              paper_CIMP.methylation.signature.
## TCGA-55-6982-11A-01R-1949-07                              <NA>
## TCGA-99-8033-01A-11R-2241-07                              <NA>
## TCGA-49-6745-01A-11R-1858-07                             high
## TCGA-49-6767-01A-11R-1858-07                     intermediate
## TCGA-05-4382-01A-01R-1206-07                              <NA>
## TCGA-05-5429-01A-01R-1628-07                             high
## TCGA-38-A44F-01A-11R-A24H-07                              <NA>
## TCGA-44-8117-01A-11R-2241-07                              <NA>
## TCGA-71-6725-01A-11R-1858-07                     intermediate
## TCGA-78-7542-01A-21R-2066-07                     intermediate
##                              paper_MTOR.mechanism.of.mTOR.pathway.activation
## TCGA-55-6982-11A-01R-1949-07                                            <NA>
## TCGA-99-8033-01A-11R-2241-07                                            <NA>
## TCGA-49-6745-01A-11R-1858-07                                       unaligned
## TCGA-49-6767-01A-11R-1858-07                                       unaligned
## TCGA-05-4382-01A-01R-1206-07                                            <NA>
## TCGA-05-5429-01A-01R-1628-07                                       unaligned
## TCGA-38-A44F-01A-11R-A24H-07                                            <NA>
## TCGA-44-8117-01A-11R-2241-07                                            <NA>
## TCGA-71-6725-01A-11R-1858-07                                       unaligned
## TCGA-78-7542-01A-21R-2066-07                                       unaligned
##                              paper_Ploidy.ABSOLUTE.calls
## TCGA-55-6982-11A-01R-1949-07                        <NA>
## TCGA-99-8033-01A-11R-2241-07                        <NA>
## TCGA-49-6745-01A-11R-1858-07                        2,15
## TCGA-49-6767-01A-11R-1858-07                         3,3
## TCGA-05-4382-01A-01R-1206-07                        2,95
## TCGA-05-5429-01A-01R-1628-07                        2,07
## TCGA-38-A44F-01A-11R-A24H-07                        <NA>
## TCGA-44-8117-01A-11R-2241-07                        <NA>
## TCGA-71-6725-01A-11R-1858-07                        2,69
## TCGA-78-7542-01A-21R-2066-07                         4,7
##                              paper_Purity.ABSOLUTE.calls
## TCGA-55-6982-11A-01R-1949-07                        <NA>
## TCGA-99-8033-01A-11R-2241-07                        <NA>
## TCGA-49-6745-01A-11R-1858-07                        0,37
## TCGA-49-6767-01A-11R-1858-07                        0,43
## TCGA-05-4382-01A-01R-1206-07                        0,23
## TCGA-05-5429-01A-01R-1628-07                        0,72
## TCGA-38-A44F-01A-11R-A24H-07                        <NA>
## TCGA-44-8117-01A-11R-2241-07                        <NA>
## TCGA-71-6725-01A-11R-1858-07                        0,63
## TCGA-78-7542-01A-21R-2066-07                        0,36
```

```r
d_luad %>% count(synchronous_malignancy)
```

```
##   synchronous_malignancy   n
## 1                     No 515
## 2           Not Reported  70
## 3                    Yes   9
```

```r
#paper_Smoking.Status
#longest_dimension, shortest_dimension
#cigarettes_per_day
#race
#paper_Nonsilent.Mutations
#paper_Nonsilent.Mutations.per.Mb
#paper_Oncogene.Negative.or.Positive.Groups
```

```r
library(RColorBrewer)

#assigning number of samples
  d_luad <- d_luad %>%
  group_by(paper_Smoking.Status) %>% mutate(n = n()) %>%
  mutate(label = paste0(paper_Smoking.Status,'\nN = ',n))


q <-  d_luad %>%
  filter(!is.na(paper_Smoking.Status)) %>%
  filter(!paper_Smoking.Status %in% "[Not Available]") %>%
  mutate(paper_Smoking.Status = reorder(paper_Smoking.Status,
                                        age_at_diagnosis, FUN = median)) %>%
  ggplot(aes(as.factor(label), age_at_diagnosis/365, fill = as.factor(label)))+
  geom_violin(trim = F, alpha = 0.3)+
  geom_boxplot(fill = "white", width = 0.2) +
  coord_flip() +
  labs(x = "",
       title = "Distribution of Age at Diagnosis by Smoking Status",
       y = "Age at Diagnosis") +
  theme(legend.position = "none") +
  scale_fill_brewer(palette="Dark2")

ggsave('plot.TCGA_LUAD.smokingstatus_AgeOfDX.20211020.pdf', q, width = 16, height = 9)
```

```
## Warning: Removed 24 rows containing non-finite values (stat_ydensity).
```

```
## Warning: Removed 24 rows containing non-finite values (stat_boxplot).
```

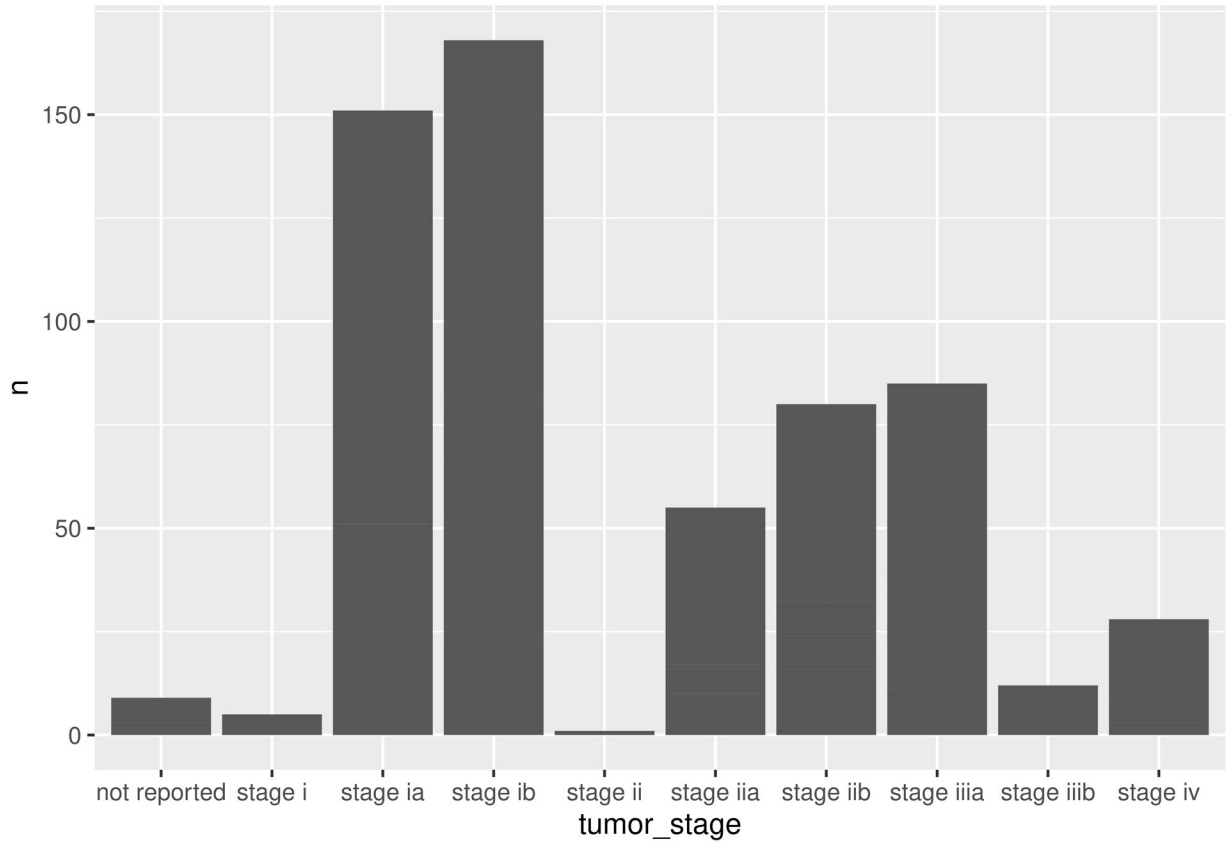
```r
# assign p for plot
p <- ggplot(d_luad, aes(age_at_diagnosis/365, fill = gender)) +
  labs(x ='Age at Diagnosis (years)', y = 'Number of LUAD patients',
       title = 'TCGA: Lung cancer adenocarcinoma') +
  geom_histogram(bins=100)


# save to file
ggsave('plot.TCGA_LUAD.histogram_AgeOfDX.20191002.pdf', p, width = 16, height = 9)
```

```
## Warning: Removed 37 rows containing non-finite values (stat_bin).
```

3.3. Tumor stages of lung cancers

First information you might want to explore is staging cancers. Different types of staging systems are used for different types of cancer. You can read further information in general staging rules, or specifics in lung cancers (e.g. stage IA).

```
# Count the number of tumors in the LUAD dataset
counts_tumor <- d_luad %>% count(tumor_stage)

# Try bar plot
ggplot(counts_tumor, aes( tumor_stage, n)) + geom_bar(stat="identity")
```

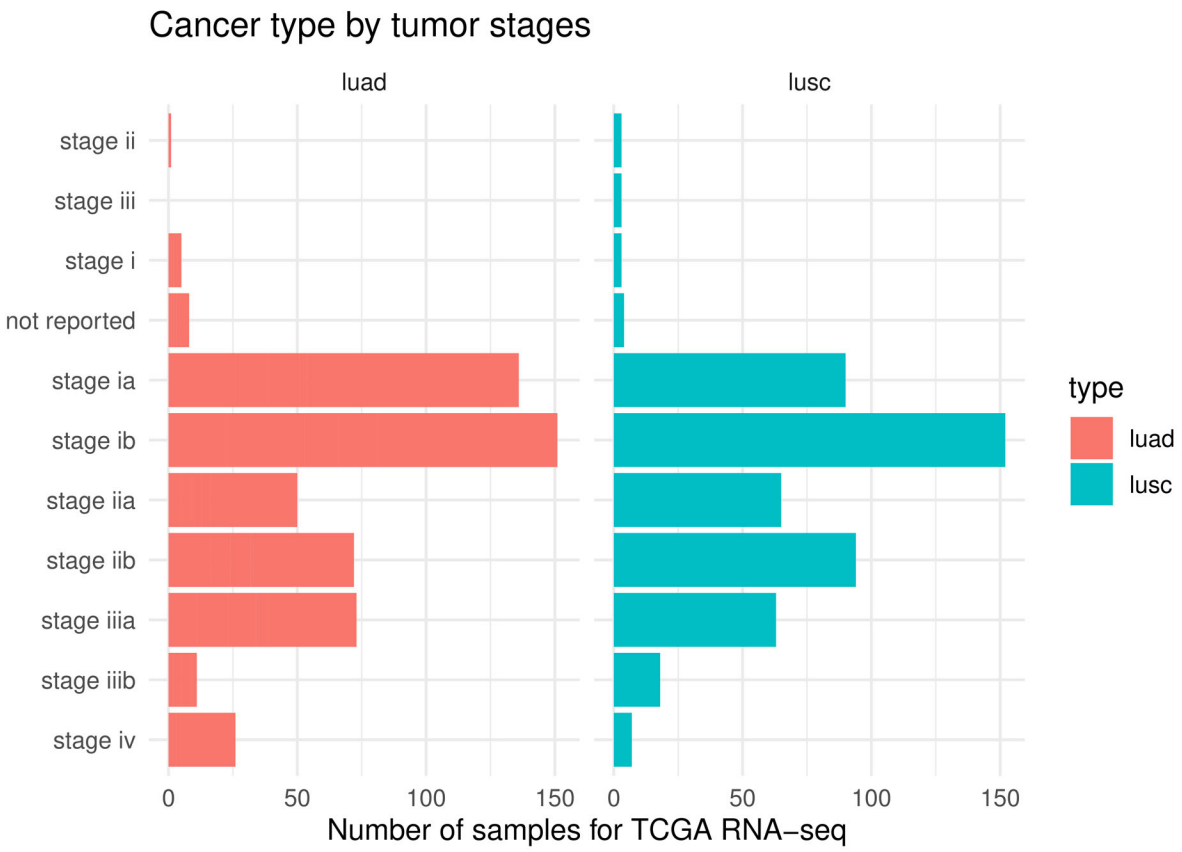


Now you can combine both for visualization.

```
# Cancer type by tumor stages
bind_rows(d_luad %>%
            filter(shortLetterCode == 'TP') %>%
            mutate(type='luad') %>%
            select(type, tumor_stage),
          d_lusc %>%
            filter(shortLetterCode == 'TP') %>%
            mutate(type='lusc') %>%
            select(type, tumor_stage)) %>%
```

```
  count(type, tumor_stage) %>%
  mutate(tumor_stage = factor(tumor_stage, levels=rev(unique(tumor_stage)))) %>%
  ggplot(aes( tumor_stage, n, fill=type)) +
  labs(title = 'Cancer type by tumor stages',
       x = '', y='Number of samples for TCGA RNA-seq') +
  theme_minimal() + geom_bar(stat="identity") +
  facet_wrap(~type) + coord_flip()
```

```
## Adding missing grouping variables: `paper_Smoking.Status`
```



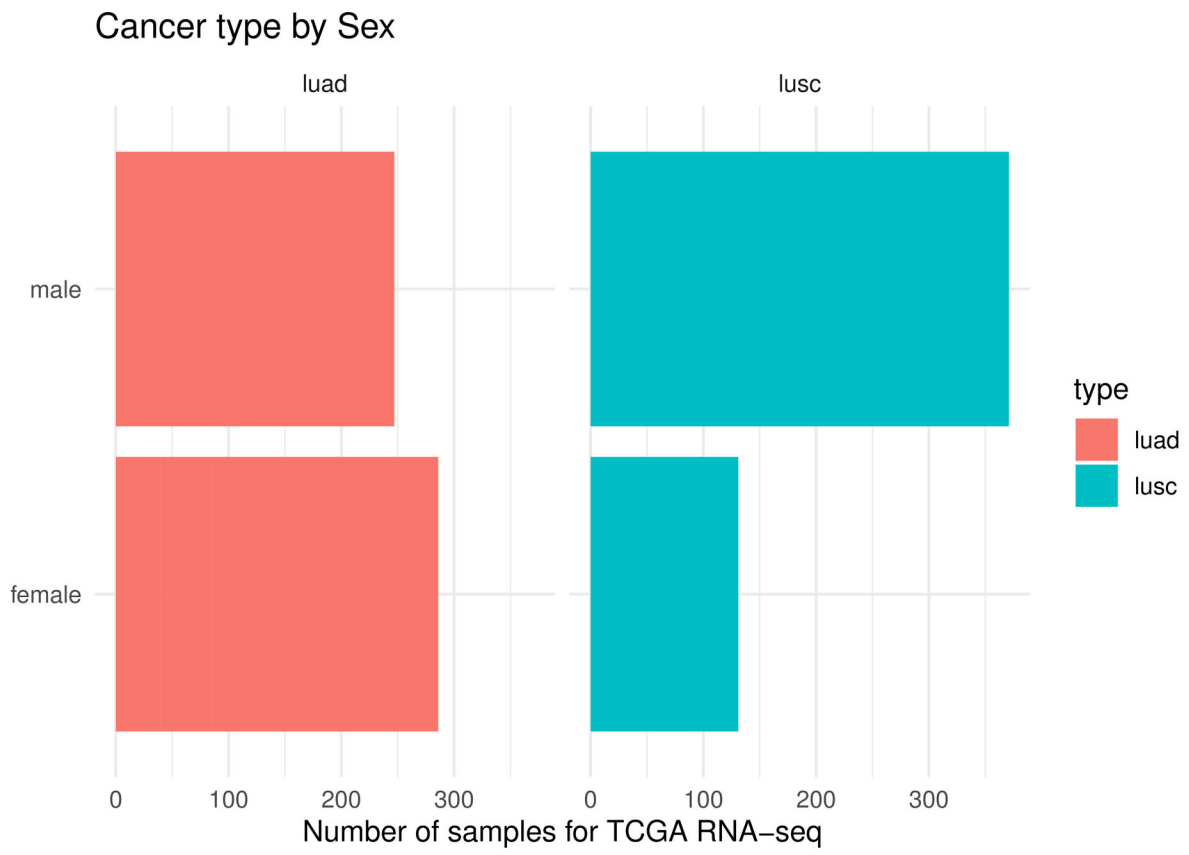
3.4. Cancer type by gender

Let's try another information - gender. We will describe this by a bar plot like above. After plotting, which information you can read from this? NB: I am not pretty sure why TCGA put gender, not sex in the dataset, in addition to mixed use of female/male with gender.

```
# Cancer type by Sex
bind_rows(d_luad %>%
            filter(shortLetterCode == 'TP') %>%
            mutate(type='luad') %>%
            select(type, gender),
          d_lusc %>%
            filter(shortLetterCode == 'TP') %>%
```

```
            mutate(type='lusc') %>%
            select(type, gender)) %>%
  count(type, gender) %>%
  ggplot(aes(gender, n, fill=type)) +
  labs(title = 'Cancer type by Sex',
       x = '', y='Number of samples for TCGA RNA-seq') +
  theme_minimal() + geom_bar(stat="identity") +
  facet_wrap(~type) + coord_flip()
```

```
## Adding missing grouping variables: `paper_Smoking.Status`
```

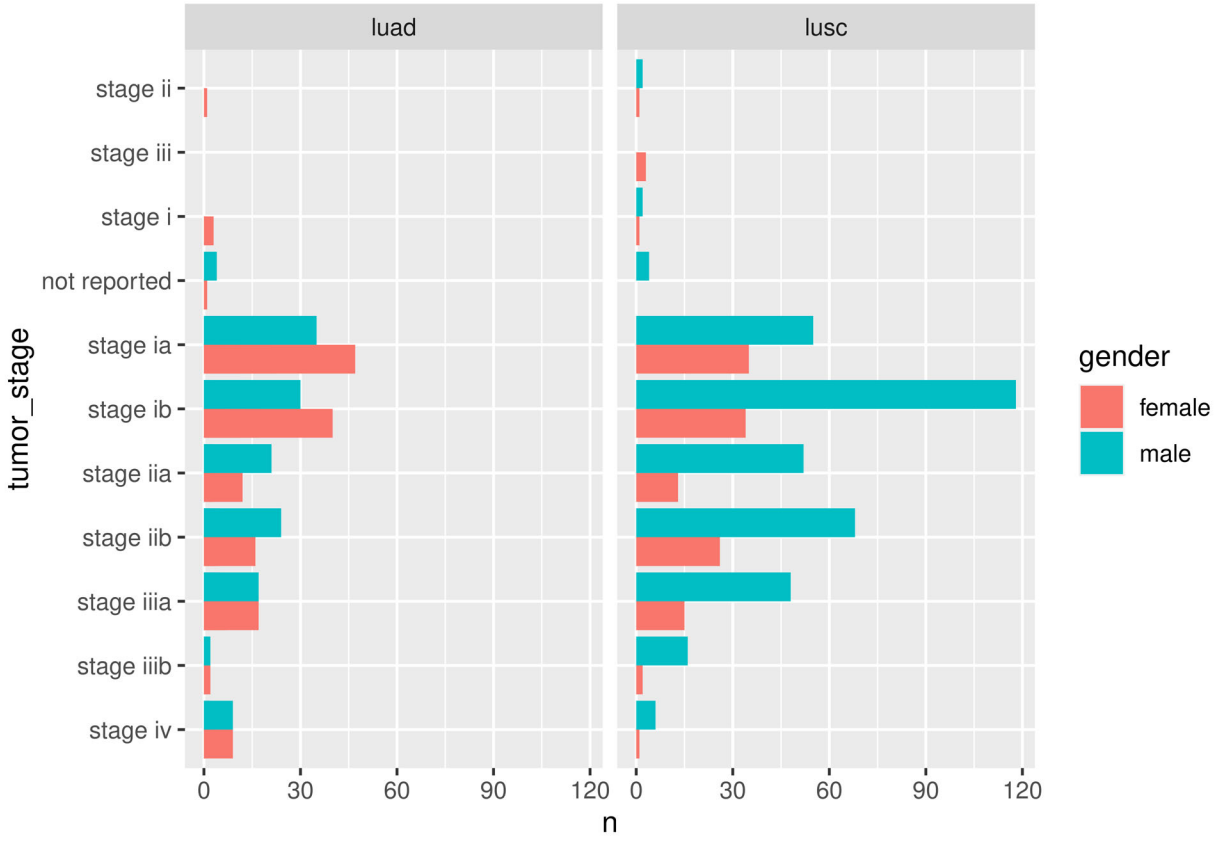


Is there difference in tumor stages by gender? Let's check with LUSC.

```
# Add gender with stages
bind_rows(d_luad %>%
            filter(shortLetterCode == 'TP') %>%
            mutate(type='luad') %>%
            select(type, tumor_stage, gender),
          d_lusc %>%
            filter(shortLetterCode == 'TP') %>%
            mutate(type='lusc') %>%
            select(type, tumor_stage, gender)) %>%
  count(type, tumor_stage, gender) %>%
  complete(gender, type, tumor_stage, fill = list(n = 0)) %>%
  mutate(tumor_stage = factor(tumor_stage, levels=rev(unique(tumor_stage))),
```

```
        gender = factor(gender)) %>%
  ggplot(., aes(tumor_stage, n, fill=gender)) +
  geom_bar(stat="identity", position=position_dodge()) +
  facet_wrap(~type) + coord_flip()
```

```
## Adding missing grouping variables: `paper_Smoking.Status`
```

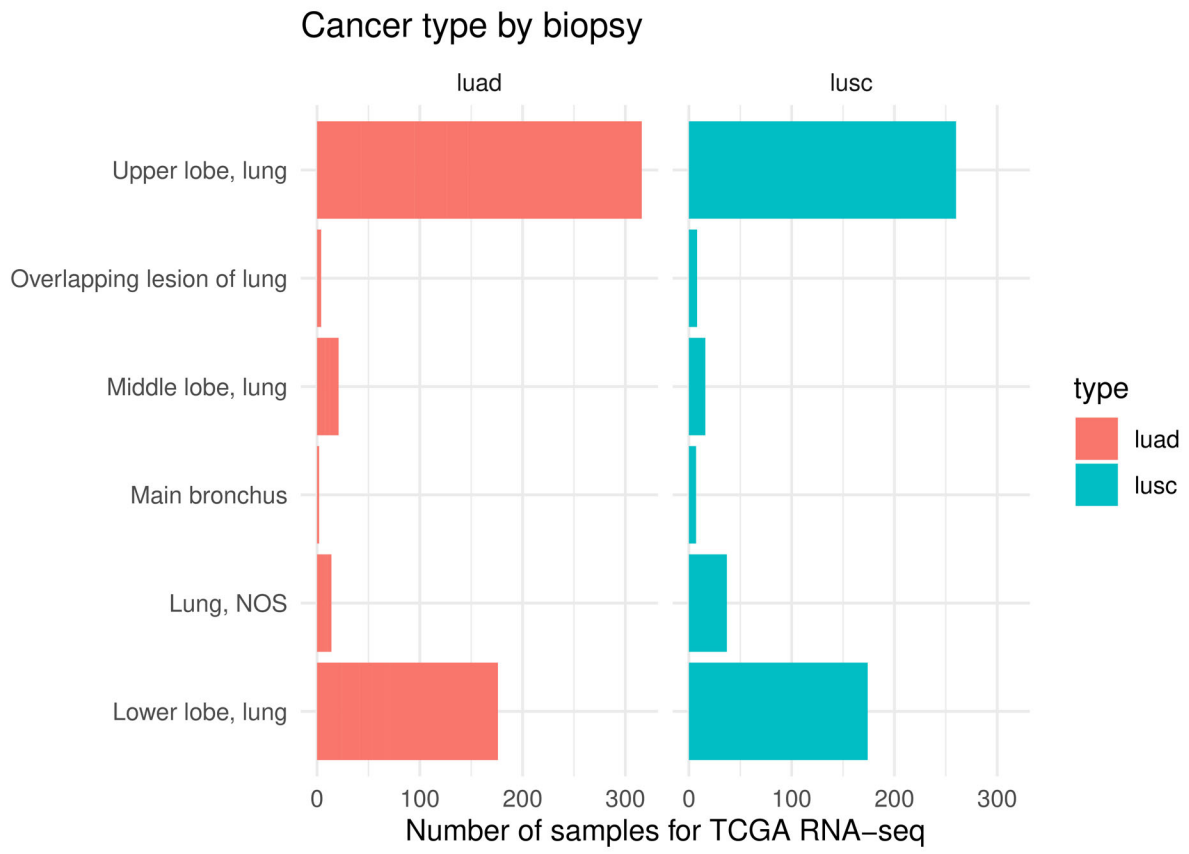


3.5. Site of biopsy

```
# Cancer type by biopsy
bind_rows(d_luad %>%
            filter(shortLetterCode == 'TP') %>%
            mutate(type='luad') %>%
            select(type, site_of_resection_or_biopsy),
          d_lusc %>%
            filter(shortLetterCode == 'TP') %>%
            mutate(type='lusc') %>%
            select(type, site_of_resection_or_biopsy)) %>%
  count(type, site_of_resection_or_biopsy) %>%
  ggplot(aes( site_of_resection_or_biopsy, n, fill=type)) +
  labs(title = 'Cancer type by biopsy',
       x = '', y='Number of samples for TCGA RNA-seq') +
```

```
  theme_minimal() + geom_bar(stat="identity") +
  facet_wrap(~type) + coord_flip()
```

```
## Adding missing grouping variables: ‘paper_Smoking.Status‘
```

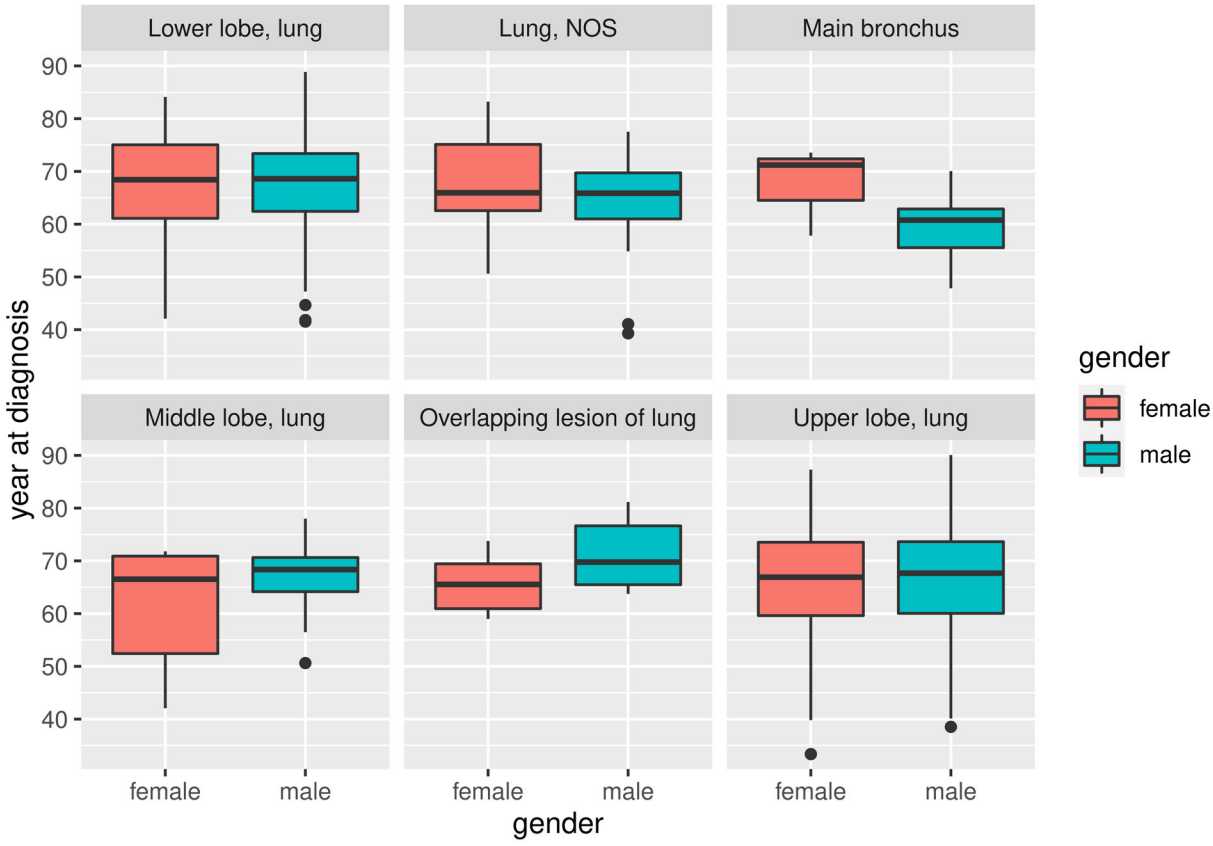


Plot the years at diagnosis by tissue or organ of origin. We would also include difference by gender.

```
bind_rows(d_luad %>%
            filter(shortLetterCode == 'TP') %>%
            mutate(type='luad') %>%
            select(type, gender, age_at_diagnosis, tissue_or_organ_of_origin),
          d_lusc %>%
            filter(shortLetterCode == 'TP') %>%
            mutate(type='lusc') %>%
            select(type, gender, age_at_diagnosis, tissue_or_organ_of_origin)) %>%
  ggplot(., aes(gender, age_at_diagnosis/365, fill=gender)) +
  geom_boxplot() + labs(y='year at diagnosis') +
  facet_wrap(~tissue_or_organ_of_origin)
```

```
## Adding missing grouping variables: ‘paper_Smoking.Status‘
```

```
## Warning: Removed 40 rows containing non-finite values (stat_boxplot).
```


Lower lobe, lung
Lung, NOS
Main bronchus
Middle lobe, lung
Overlapping lesion of lung
Upper lobe, lung
year at diagnosis
90
80
70
60
50
40
female
male
gender
gender
female
male