[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

1. To access the gene expression data, there are two ways 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 protal

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

```
# library(tidyverse)

# 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_luad0)
#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"
```

```
identical(colnames(d_luad),colnames(d_lusc))
```

```
## [1] FALSE
```

```
# not same
```

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)
```

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

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
```

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

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

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

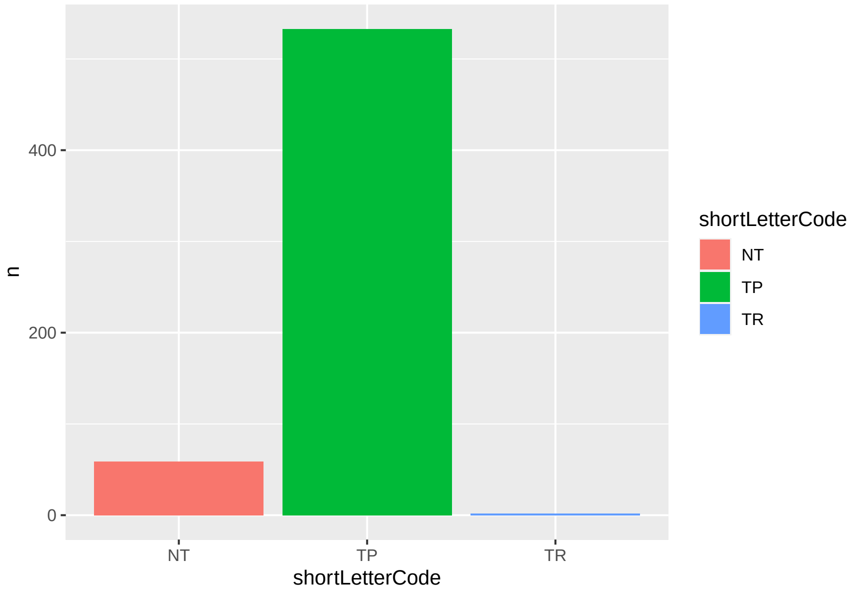
Check which samples are included in the LUSC dataset.

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

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

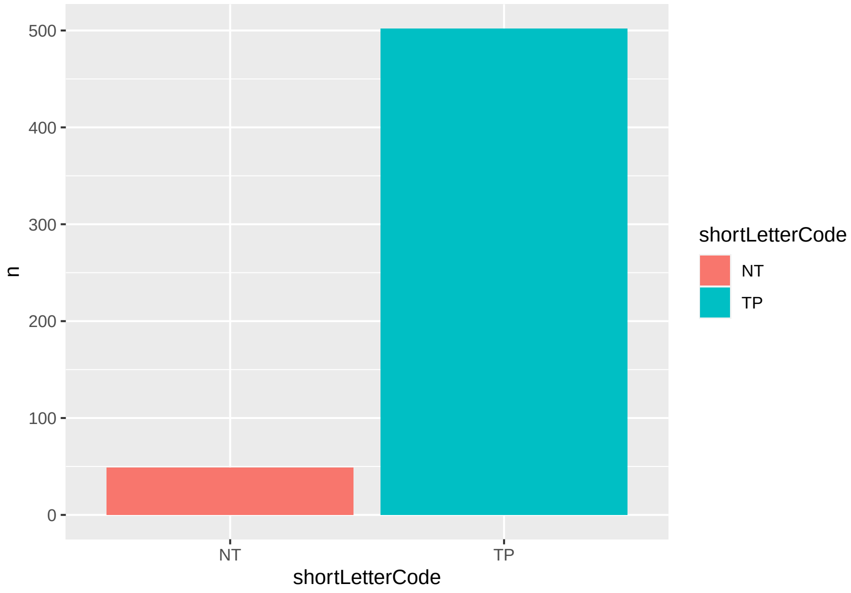
Let's make a simple bar plot to compare the number of tissues between two tissue types. We will look at the LUAD samples first.

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



Now we can plot the LUSC samples.

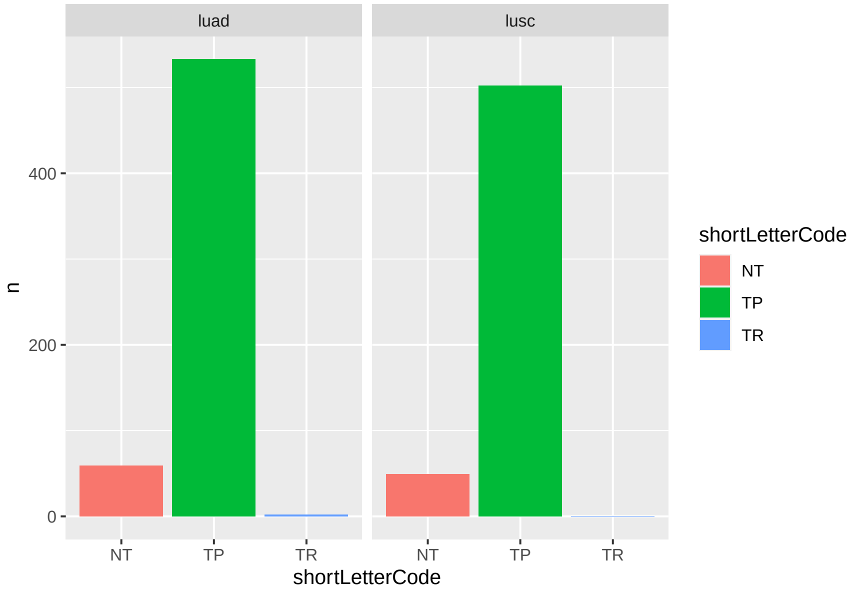
```
d_lusc %>%
  count(shortLetterCode) %>%
  ggplot(., aes(shortLetterCode, n, fill=shortLetterCode)) +
  geom_bar(stat="identity", position=position_dodge())
```



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,'free_x', ncol=5) # Q2: What is this?
```

```
## Warning: Coercing 'nrow' to be an integer.

## Warning in sanitise_dim(nrow): NAs introduced by coercion

## Warning: 'nrow' is missing or less than 1 and will be treated as NULL.
```



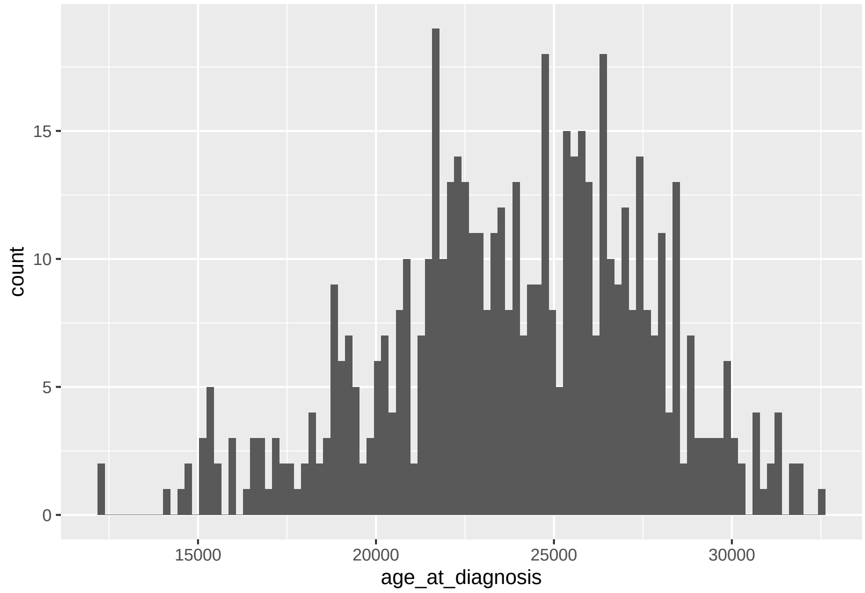
```
#Q1 tidyr::complete()
#Turns implicit missing values into explicit missing values
#In this case, lusc,TR -> n=0
# fill=list(n=0), this argument fill n=0 instead of NA.

#Q2 facet_wrap()
#similar to facet-grid but better sometimes..
```

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 fowlloing 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)
```

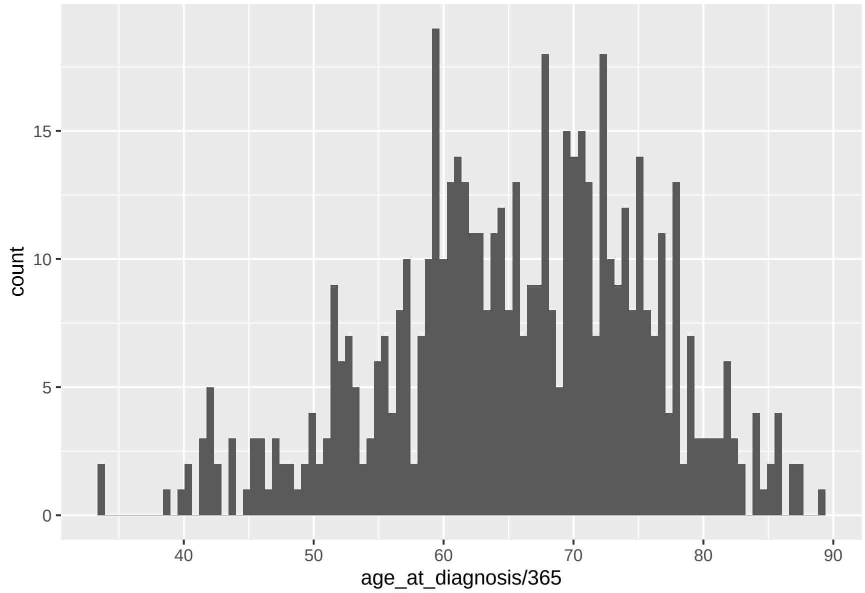
```
## Warning: Removed 37 rows containing non-finite values (stat_bin).
```



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)
```

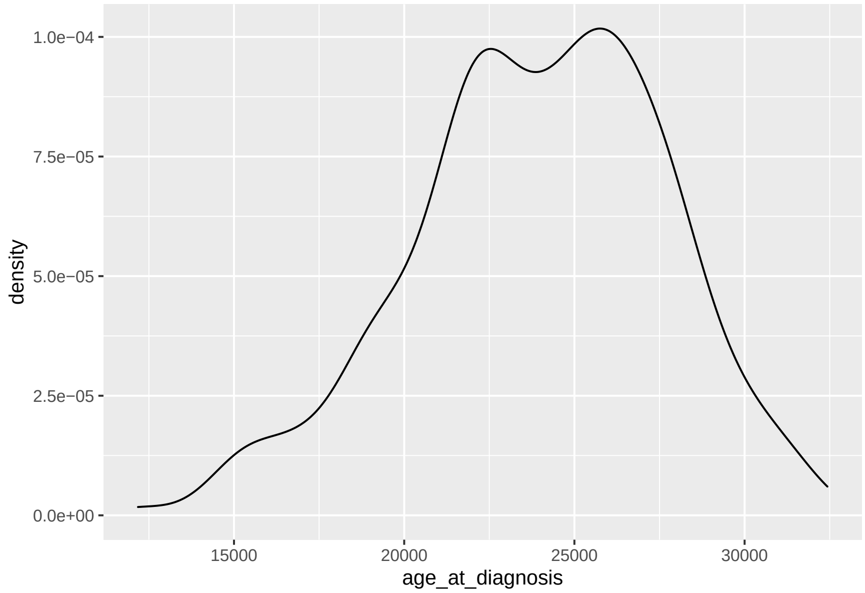
```
## Warning: Removed 37 rows containing non-finite values (stat_bin).
```



To creat 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).
```

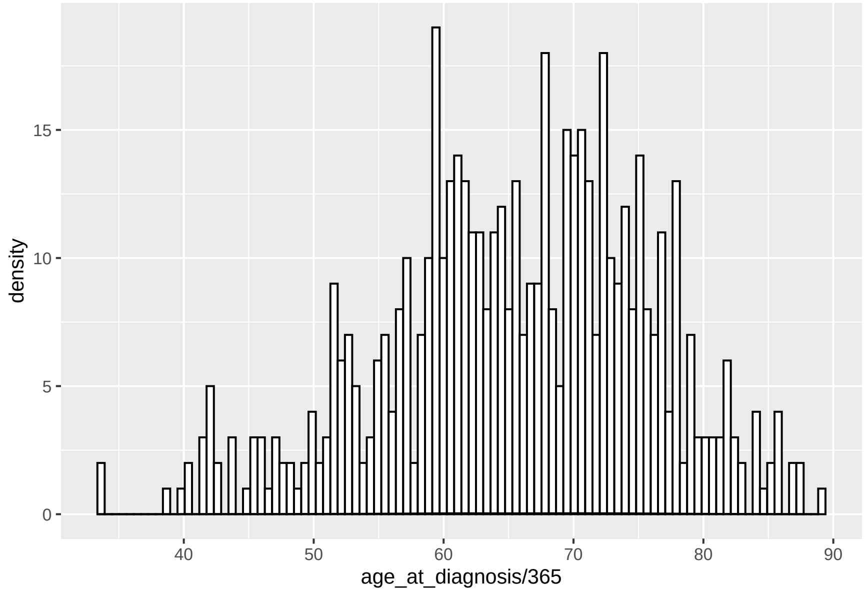


More plot types on distribution. First we can try the plot for both histogram and density.

```
# 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?

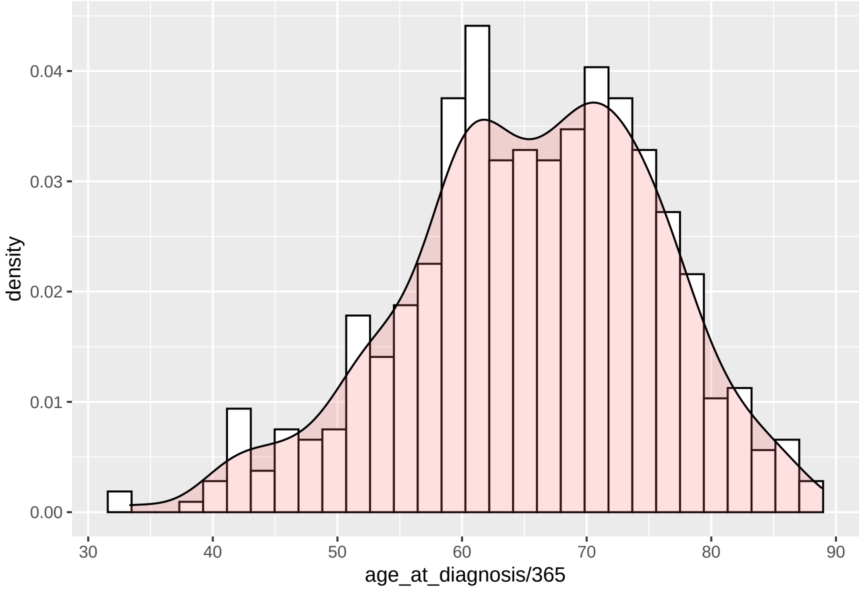
Let's plot a bit different version. We will replace the y-axis of the histogram with the y-axis of the density plot.

```
# 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).
```



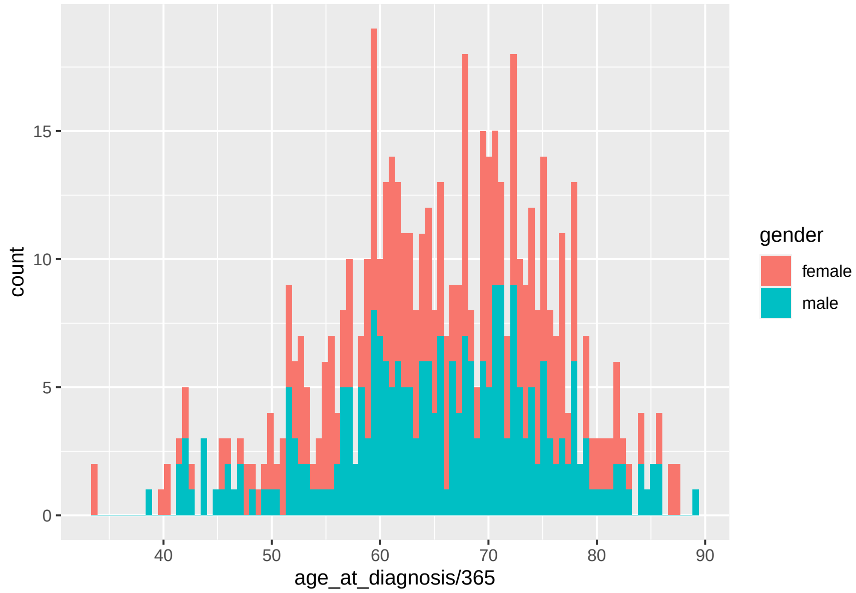
Q: Do you think whether it is good visualization for the distribution?

```
# not bod I think.
```

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)
```

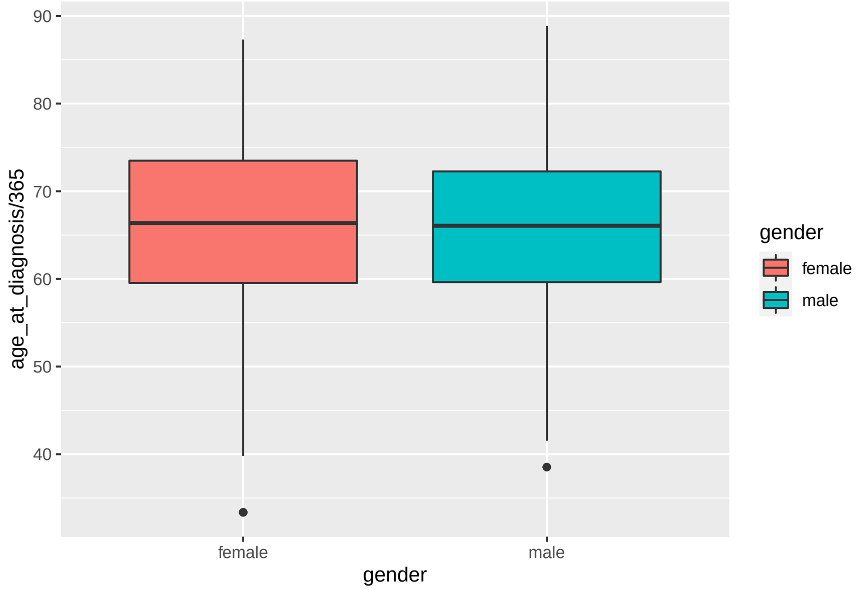
```
## Warning: Removed 37 rows containing non-finite values (stat_bin).
```



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()
```

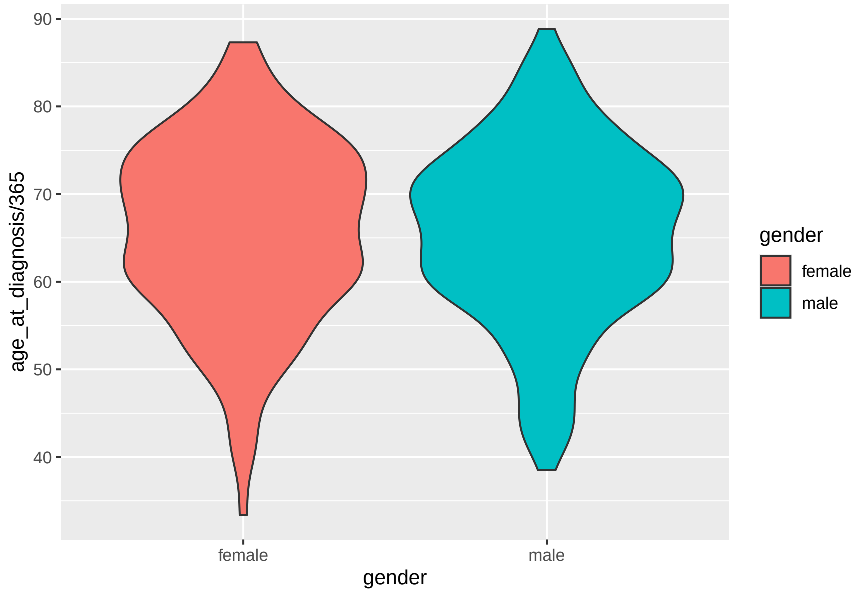
```
## Warning: Removed 37 rows containing non-finite values (stat_boxplot).
```



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).
```

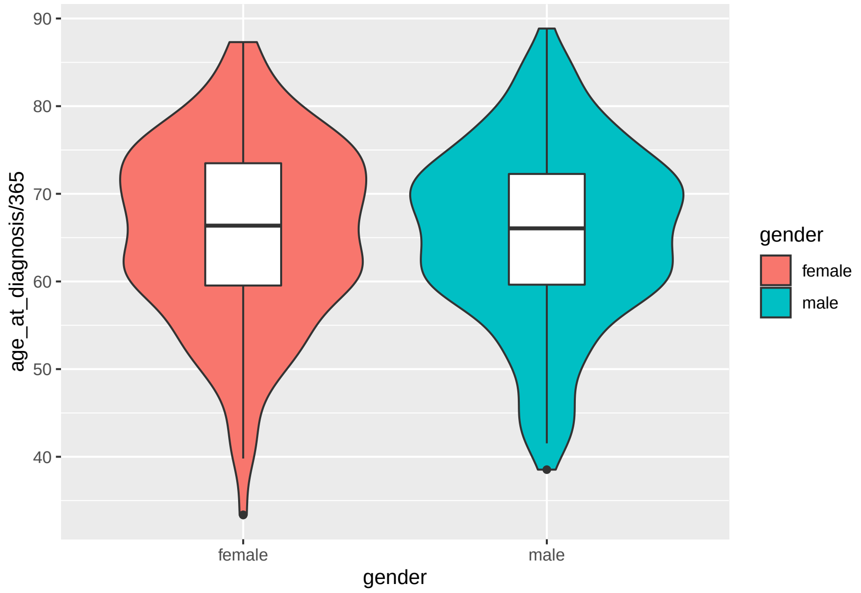


Can you tell the difference between boxplot and violin plot?

```
# 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)
```

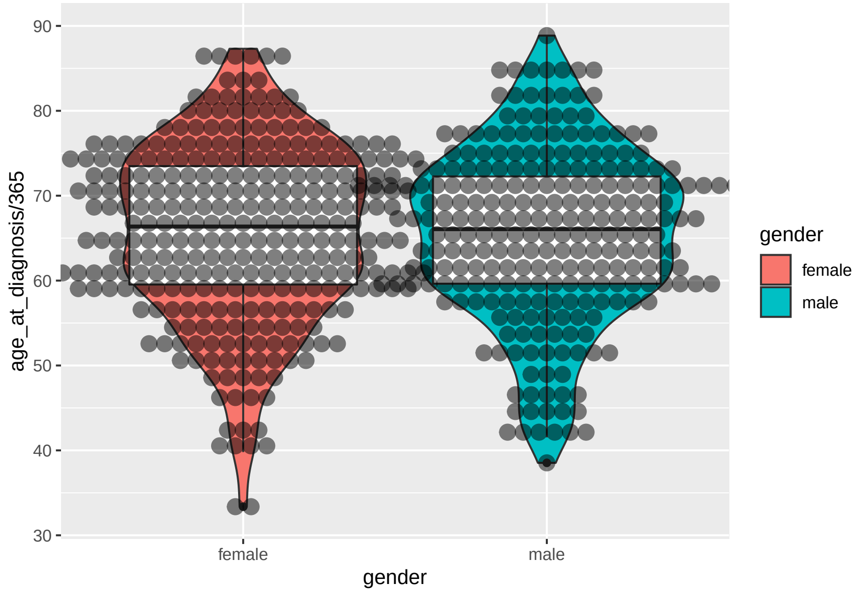
```
## Warning: Removed 37 rows containing non-finite values (stat_ydensity).
```

```
## 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).
```

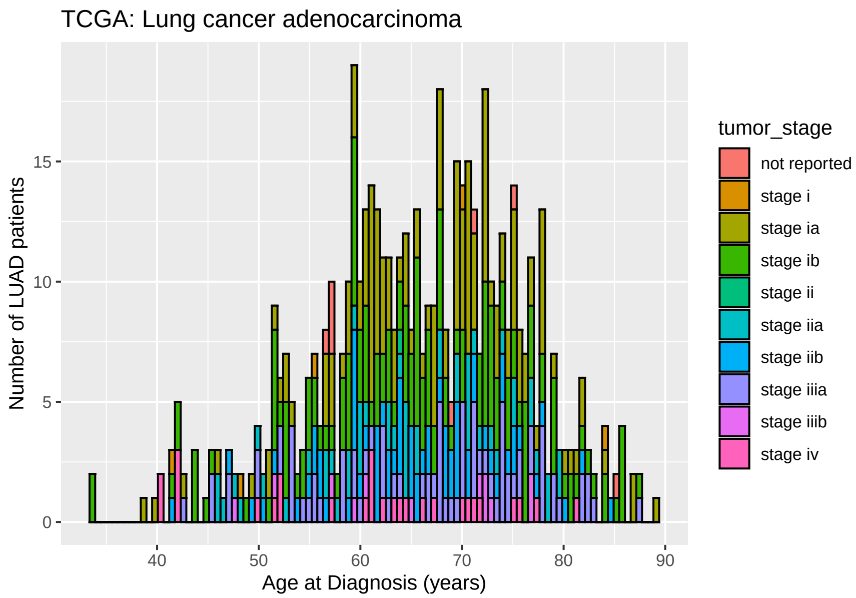


```
# Try different options from the manual
# http://www.sthda.com/english/wiki/ggplot2-violin-plot-quick-start-guide-r-software-and-data-visualiza
```

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(aes(fill=tumor_stage),col="black",bins=100)
```

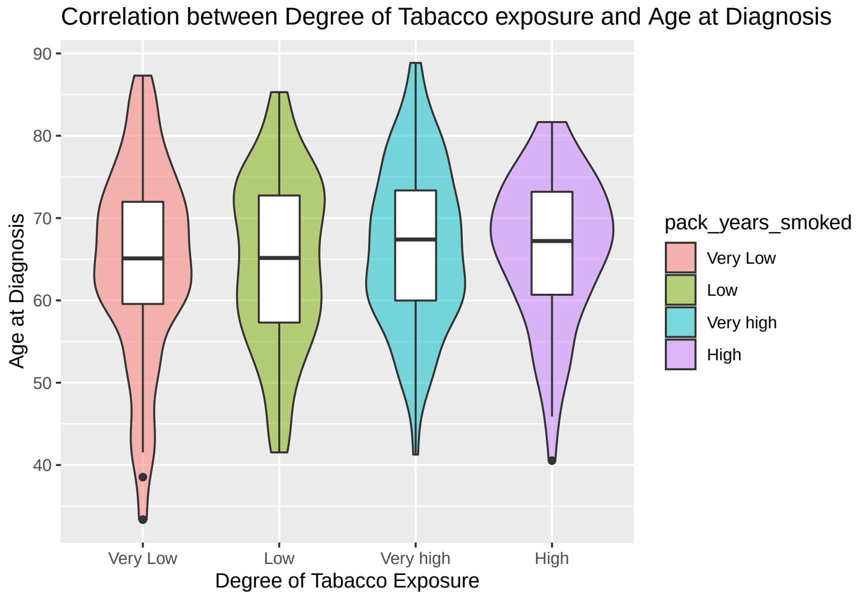
```
## Warning: Removed 37 rows containing non-finite values (stat_bin).
```



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.

```
# assign p for plot
library(tidyverse)
library(ggrepel)
library(ggplot2)
library(ggridges)



p<-d_luad%>%
  filter(!is.na(pack_years_smoked))%>%
  select(patient,age_at_diagnosis,shortLetterCode,pack_years_smoked)%>%
  mutate(pack_years_smoked=case_when(scale(pack_years_smoked)<qnorm(0.25)~"Very Low",
                                     scale(pack_years_smoked)<qnorm(0.5)~"Low",
                                     scale(pack_years_smoked)<qnorm(0.75)~"High",
                                     TRUE~"Very high"))%>%
  mutate(pack_years_smoked=factor(pack_years_smoked,levels=unique(pack_years_smoked)))%>%
  ggplot(aes(pack_years_smoked,age_at_diagnosis/365))+
  geom_violin(aes(fill=pack_years_smoked),alpha=0.5)+
  geom_boxplot(width=0.3)+
  xlab("Degree of Tabacco Exposure")+
  ylab("Age at Diagnosis")+
  ggtitle("Correlation between Degree of Tabacco exposure and Age at Diagnosis")

p
```

```
## Warning: Removed 15 rows containing non-finite values (stat_ydensity).

## Warning: Removed 15 rows containing non-finite values (stat_boxplot).
```

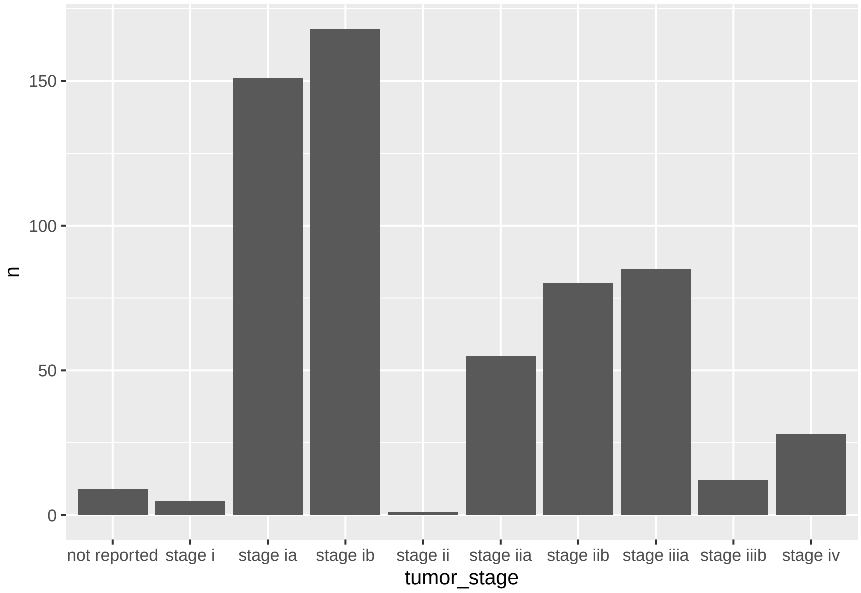


```
# ggsave('Degree_of_Tabacco_Exposure.pdf',p,width = 16,height=9)
# To save to PDF is done.
```

3.3 Tumor stages of lung cancers

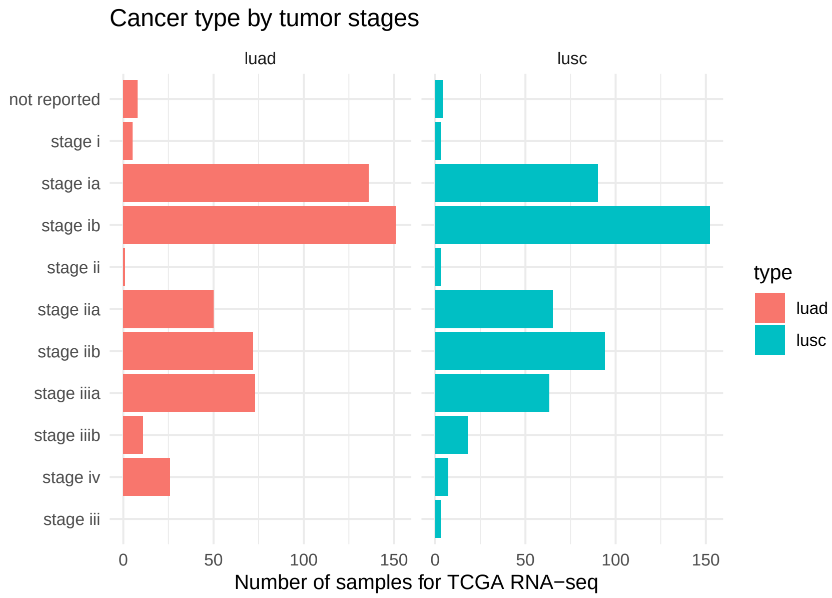
First information you might want to explore is staging cancers. Different types of staging systems are used for different types of cancer. You can read further information in general staging rules, or specifics in lung cancers (e.g. stage IA).

```
# 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.

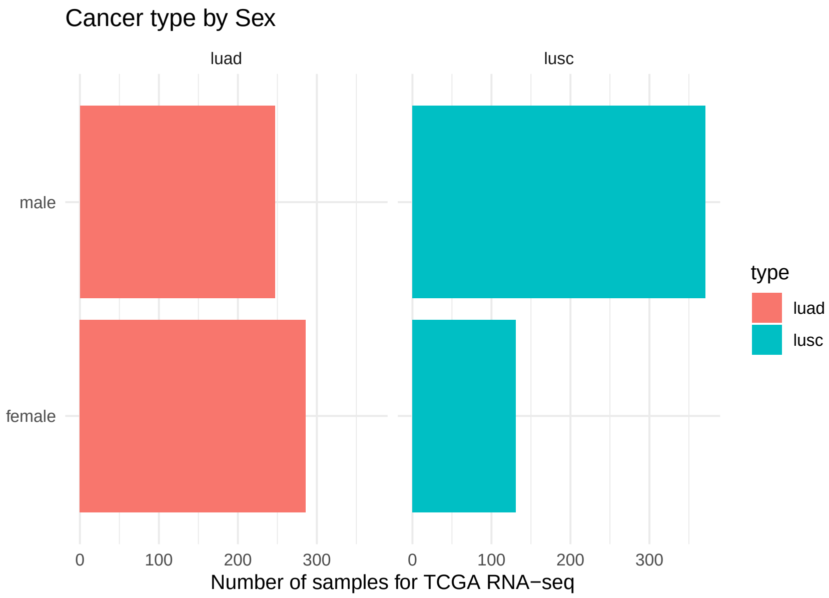
```
# Cancer type by tumor stages
bind_rows(d_luad %>%
            filter(shortLetterCode == 'TP') %>%
            mutate(type='luad') %>%
            select(type, tumor_stage),
          d_lusc %>%
            filter(shortLetterCode == 'TP') %>%
            mutate(type='lusc') %>%
            select(type,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



3.4 Cancer type by gender

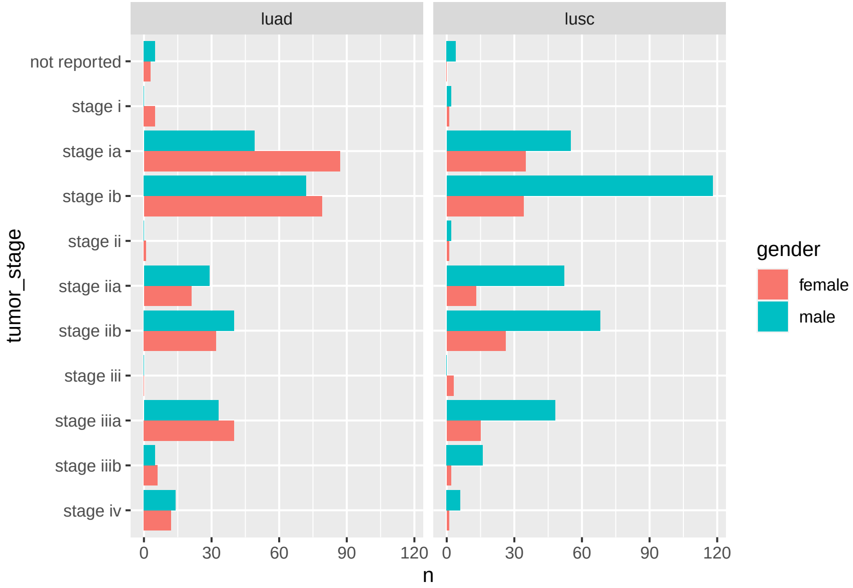
Let's try another information - gender. We will describe this by a bar plot like above. After plotting, which information you can read from this? NB: I am not pretty sure why TCGA put gender, not sex in the dataset, in addition to mixed use of female/male with gender.

```
# 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()
```



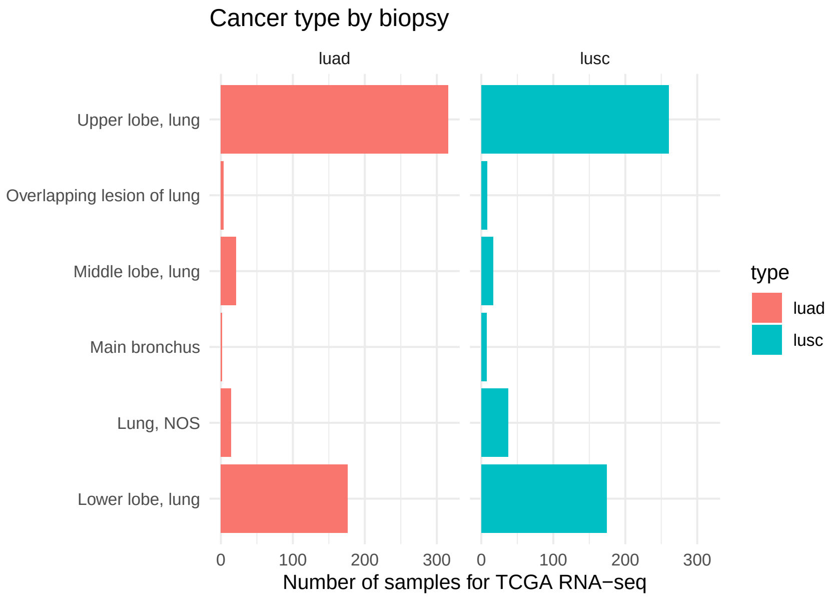
Is there difference in tumor stages by gender? Let's check with LUSC.

```
# 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()
```



3.5 Site of biopsy

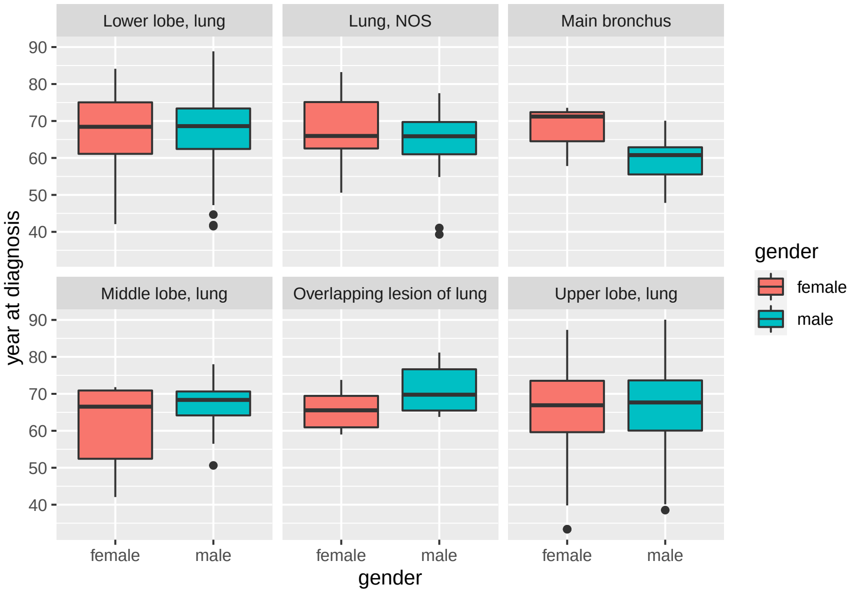
```
# Cancer type by biopsy
bind_rows(d_luad %>%
            filter(shortLetterCode == 'TP') %>%
            mutate(type='luad') %>%
            select(type, site_of_resection_or_biopsy),
          d_lusc %>%
            filter(shortLetterCode == 'TP') %>%
            mutate(type='lusc') %>%
       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



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)
```

```
## 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