Factors

0.1 Libraries

```
library(tidyverse)
```

0.2 Defining factors

```
# usual to convert strings to factors
ch <- c("plus","minus")
ch
```

```
[1] "plus"  "minus"
```

```
# note: levels are alphabetically placed
fch <- factor(ch)
fch
```

```
[1] plus  minus
Levels: minus plus
```

```
# or using `as.factor` from BaseR
as.factor(ch)
```

```
[1] plus  minus
Levels: minus plus
```

0.3 Ordering of `factors`

0.3.1 baseR:: `as.factor` or `factor`

0.3.1.1 (default to alphabetical)

```
# factor levels normally arranged alphabetically
bloodtypes <- c("O","A","B","AB")
b1 <- factor(bloodtypes)
levels(b1)
```

```
[1] "A"  "AB" "B"  "O"
```

```
b2 <- as.factor(bloodtypes)
levels(b2)
```

```
[1] "A"  "AB" "B"  "O"
```

- the `Levels` default to an alphabetical sequence, but
- if I want to list factors Levels by the order in which they first appear

    - you can define levels explicitly with factor() '
    - you can use as_factor from tidyverse which will retain the order

- hence, good idea to use `as_factor`

0.3.1.2 (specify factor order)

```
# method 1
# specified levels
bt1 <- factor(bloodtypes,
        levels=c("O","A","B","AB"))
levels(bt1)
```

```
[1] "O"  "A"  "B"  "AB"
```

```
# method 2
# using as_factor (instead of as.factor from baseR)
# will retain the original factor order.
# better than using relevel or fct_reorder
bt2 <- as_factor(bloodtypes)
levels(bt2)
```

```
[1] "O"  "A"  "B"  "AB"
```

0.4 Ordering factors for ggplots

```
# problems when you want to plot the categorical variable
# in a certain (original) order
tb <- read_csv("cancer, age, sex
              Colorectal, 66, M
              Lung, 29, M
              Lung, 21, F
              Nasal, 17, M
              Colorectal, 55, F
              Nasal, 31,F
              Breast, 37, F
              Liver, 55, M
              Liver, 63, F
              ")
tb1 <- tb
tb1$cancer <- factor(tb1$cancer)
levels(tb1$cancer)
```

```
[1] "Breast"     "Colorectal" "Liver"      "Lung"       "Nasal"
```

```
ggplot(tb1, aes(x=cancer))+
  geom_bar(fill="lightblue",color="blue")
```
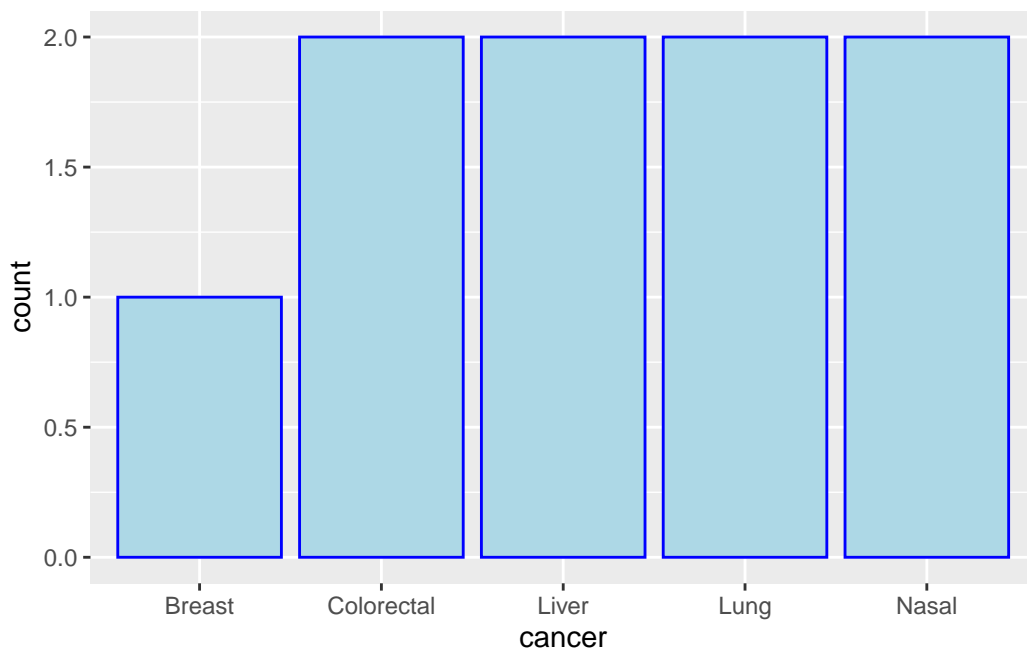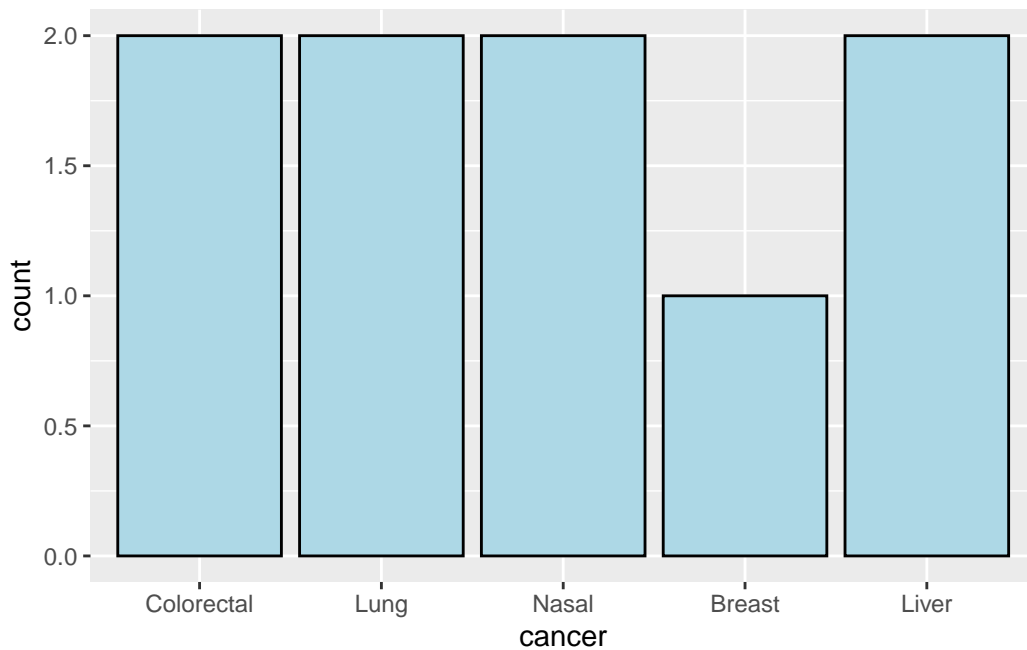
- or I can order the factor cancer in its original position

```
library(tidyverse)
tb2 <- tb
tb2$cancer <- as_factor(tb$cancer)
levels(tb2$cancer)
```

```
[1] "Colorectal" "Lung"       "Nasal"      "Breast"     "Liver"
```

```
ggplot(tb2, aes(x=cancer))+
  geom_bar(fill="lightblue",color="black")
```


0.5 Renaming Factors

```
tb3<-read_csv("
Dept, Views
National Cancer Center, 683
National Heart Center Singapore, 697
National Eye Center, 534
")
tb3$Dept <- as_factor(tb3$Dept)    # retains order
levels(tb3$Dept)
```

```
[1] "National Cancer Center"          "National Heart Center Singapore"
[3] "National Eye Center"
```

```r
tb3$Dept <- fct_recode(tb3$Dept,
          NCC="National Cancer Center",
          NHCS = "National Heart Center Singapore",
          NEC = "National Eye Center")
levels(tb3$Dept)
```

```
[1] "NCC"  "NHCS" "NEC"
```

```r
tb3
```

```
# A tibble: 3 x 2
  Dept  Views
  <fct> <dbl>
1 NCC     683
2 NHCS    697
3 NEC     534
```