

# Code together: From messy data to insightful visualizations –part 2

R-Ladies Frankfurt Meetup #5

8<sup>th</sup> of August 2019

# Artificial HIV-dataset

- Source: Github

<https://github.com/NFilmann/RLadiesFRA>

– Datasets basedata.csv, labdata.csv

Use `readr::read_delim` or `readr::read_csv2` for import

```
library(tidyverse)
labdata <-
read_delim("https://raw.githubusercontent.com/NFilmann/RLadies
FRA/master/labdata.csv", delim=";")
```

# Which factors are associated with therapeutic success?

- Therapeutic success in HIV-positive individuals, defined as
  - Primary goal is virologic response:, i.e. reduction of the viral load to an undetectable level below <20 copies per ml) by 24 weeks after start of treatment
  - CD4 cell counts: Key measure of immune status; they should rise 50 to 100 cells per ml in the first year of therapy (a CD4 count < 200 is defined as AIDS).

Source:

[https://en.wikipedia.org/wiki/Management\\_of\\_HIV/AIDS#Response\\_to\\_therapy](https://en.wikipedia.org/wiki/Management_of_HIV/AIDS#Response_to_therapy)

# basedata

```
> glimpse(basedata)
```

```
Observations: 1,299
```

```
Variables: 15
```

```
$ PatientID      <dbl> 1, 2, 3,  
4, 5, 6, 7, 8, 9, 10, 11,...
```

```
$ DateOfBirth    <chr>  
"17.02.1944", "02.06.1966", ...
```

```
$ Start_therapy  <chr>  
"09.09.2015", "24.07.2014",  
"13.06.2..."
```

```
$ Gender          <chr> "male",  
"male", "male", "male", ...
```

```
$ StudyCenter     <dbl> 1, 1, 1,  
1, 1, 1, 1, 1, 1, 1...
```

```
$ DateOfDiagnosis <chr>  
"01.01.1985", "01.01.1989", ...
```

*continued*

```
$ DateOfDeath     <chr> NA, NA,  
NA, NA, NA, NA, NA, ...
```

```
$ PreMedication   <chr> "N", "N",  
"N", "N", "Y", "N",...
```

```
$ HBVpos          <chr> NA, NA,  
NA, "Y", NA, NA, "..."
```

```
$ HCVpos          <chr> NA, NA,  
NA, NA, NA, NA, NA, ...
```

```
$ MedID1           <dbl> 1, 1, 2,  
4, 2, 1, 2, 2, 4, 2, 2...
```

```
$ MedID2           <dbl> 3, 4, 1,  
7, 1, 2, 1, 1, 2, 1, 1...
```

*(MedID3, MedID4, MedID5 accordingly)*

# Variables in basedata

PatientID: patient ID

DateOfBirth: date of birth (dd.mm.yyyy)

Start\_therapy: begin of therapy (dd.mm.yyyy)

Gender: gender

StudyCenter: study center (the hospital or medical practice) where the patient was treated

DateOfDiagnosis: date when HIV was diagnosed (dd.mm.yyyy)

DateOfDeath: if date available (dd.mm.yyyy), it depicts the date of death. NA indicates the patient is still alive.

PreMedication: Y indicates that patient has received HIV-specific treatment before, N that not.

HBVpos: Y indicates patient is infected with hepatitis B as well. NA indicates that no hepatitis B infection was diagnosed.

HCVpos: Y indicates patient is infected with hepatitis C as well. NA indicates that no hepatitis C infection was diagnosed.

MedID1, MedID2, ..., MedID5: HIV specific medications coded as numbers (1–14). Note that combination treatment with three up to five medications is common.

# labdata

```
> glimpse(labdata)
Observations: 7,794
Variables: 7
$ PatientID      <dbl> 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3,
3, 4, 4, 4, 4...
$ Time_weeks     <dbl> 0, 4, 8, 12, 16, 24, 0, 4, 8, 12, 16, 24, 0, 4, 8,
12, 16, 24, 0...
$ Test           <chr> "CD4", "CD4", "CD4", "CD4", "CD4", "CD4", "CD4",
"CD4", "CD4", "...
$ value          <dbl> 354, 595, 427, 699, 606, 660, 64, 102, 152, 112,
141, 172, 146, ...
$ Test_1         <chr> "HIVPCR", "HIVPCR", "HIVPCR", "HIVPCR", "HIVPCR",
"HIVPCR", "HIV...
$ TErgNumOperator <chr> NA, NA, NA, NA, NA, NA, NA, NA, "<", NA, "<", "<",
NA, NA, NA, "...
$ TErgNum        <dbl> 500, 49, 97, 31, 49, 23, 2840000, 3670, 20, 23, 20,
20, 3136, 38...
```

# Variables in labdata

PatientID: patient ID

Time\_weeks: measuring time (in weeks), start of therapy = 0

Test: Type of lab value (here CD4), corresponds to Value

Value: CD4 value (cells per ml )

Test\_1: Type of lab value (here HIVPCR), corresponds to TErgNumOperator and TErgNum

TErgNumOperator: < indicates if corresponding viral load in TErgNum is below a certain value, e.g. <20 indicates the viral load is below 20 copies per ml, the limit of quantification

TErgNum: viral load copies per ml (HIV)

# Steps to tidy the data

1. Join both datasets
2. Make sure that each observation corresponds to one row
3. Calculate the patient age at start of therapy
4. Calculate a new variable „ViralResponse“ indicating if therapeutic success (i.e. viral load <20 copies/ml) is reached 24 weeks after start of therapy  
(How do you deal with patients that died before end of treatment?)
5. Recode HBVpos and HCVpos in a meaningful way

See Github <https://github.com/NFilmann/RLadiesFRA> for the sample solution (Tidy.R) to have the dataset ready for plotting!!



# Sample code for tidying the data

```
library(tidyverse)
#Import the data
labdata <-
read_delim("https://raw.githubusercontent.com/NFilmann/RLadiesFRA/master/labdata.csv", delim=";")
basedata <-
read_delim("https://raw.githubusercontent.com/NFilmann/RLadiesFRA/master/basedata.csv", delim=";")

#Have a look at the data
str(labdata)
str(basedata)

# 1. Join both datasets
# The primary key in each table is the variable PatientID

# At first we assure, that we have no missing values in PatientID
summary(unique(labdata$PatientID))
summary(basedata$PatientID)
#-> no missings, and the IDs match

HIVdata <- full_join(labdata, basedata)
str(HIVdata)

# 2. Make sure that each observation corresponds to one row

HIVdata=split(labdata, labdata$Time_weeks) %>%
  reduce(left_join, by = "PatientID") %>%
  full_join(basedata, .)
str(HIVdata)

HIVdata <- HIVdata %>% rename_all(funs(str_replace(., ".y.y.y",
".24"))) %>%
  rename_all(funs(str_replace(., ".x.x.x", ".16"))) %>%
  rename_all(funs(str_replace(., ".y.y", ".12"))) %>%
  rename_all(funs(str_replace(., ".x.x", ".8"))) %>%
  rename_all(funs(str_replace(., ".y", ".4"))) %>%
  rename_all(funs(str_replace(., ".x", ".0")))

HIVdata <- HIVdata %>% rename_all(funs(str_replace(., "Start_thera.4",
"Start_therapy"))) %>%
  rename_all(funs(str_replace(., "Stu.4Center", "StudyCenter")))

str(HIVdata)
```

```
# 3. Calculate the patient age at start of therapy
HIVdata <-HIVdata %>% mutate(HIVdata,
Age=((difftime(as.Date(HIVdata$Start_therapy, format="%d.%m.%Y"),
as.Date(HIVdata$DateOfBirth, format="%d.%m.%Y")))) %>%
  mutate(Age=as.numeric(round(Age/365.25,0)))
str(HIVdata)

# 4&5. Calculate a new variable „VirResponse“ indicating if therapeutic
success (i.e. viral load <20 copies/ml) is reached 24 weeks after
start of therapy
# Recode HBVpos and HCVpos in a meaningful way

#For VirResponse you could use simply the variable
"TErgNumOperator.24", where an "<" indicates that the
#treatment was succesful
HIVdata <-HIVdata %>% mutate(
  VirResponse =
  as.numeric(!is.na(TErgNumOperator.24)),
  HBVpos = as.numeric(!is.na(HBVpos)),
  HCVpos = as.numeric(!is.na(HCVpos)))

#labeling the variables in a meaningful way
HIVdata <-HIVdata %>%
  mutate(VirResponse = recode(VirResponse, "0" = "no", "1" = "yes"),
  HBVpos = recode(HBVpos, "0" = "neg", "1" = "pos"),
  HCVpos = recode(HCVpos, "0" = "neg", "1" = "pos"))

# #Labeling the variable names
HIVdata <-HIVdata %>%
  mutate(VirResponse = structure(VirResponse, label = "Virologic
Response"),
  TErgNum.0 = structure(TErgNum.0, label = "viral load at baseline"),
  Value.0 = structure(Value.0, label = "CD4 at baseline"),
  HBVpos = structure(HBVpos, label = "HBV pos"),
  HCVpos = structure(HCVpos, label = "HCV pos"))

str(HIVdata)
#(How do you deal with patients that died before end of treatment?)
#They don't have any lab values at week 24, so they are already taken
into account.

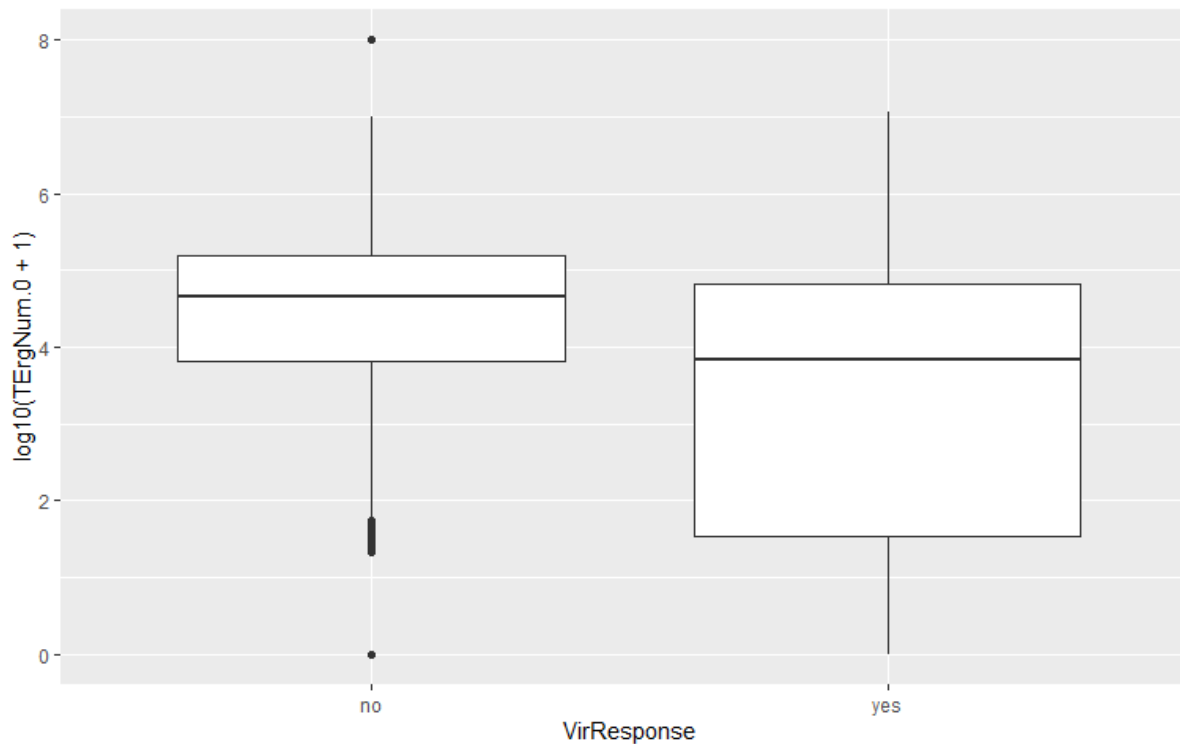
#Of course we could go on here, for example the variables, that
contain the medication are a mess,
#but we rather look at some plots now.
```

# Now it's your turn – plot the data

- 1. Make a boxplot between baseline viral load and Virologic response,
  - b) stratify also for gender

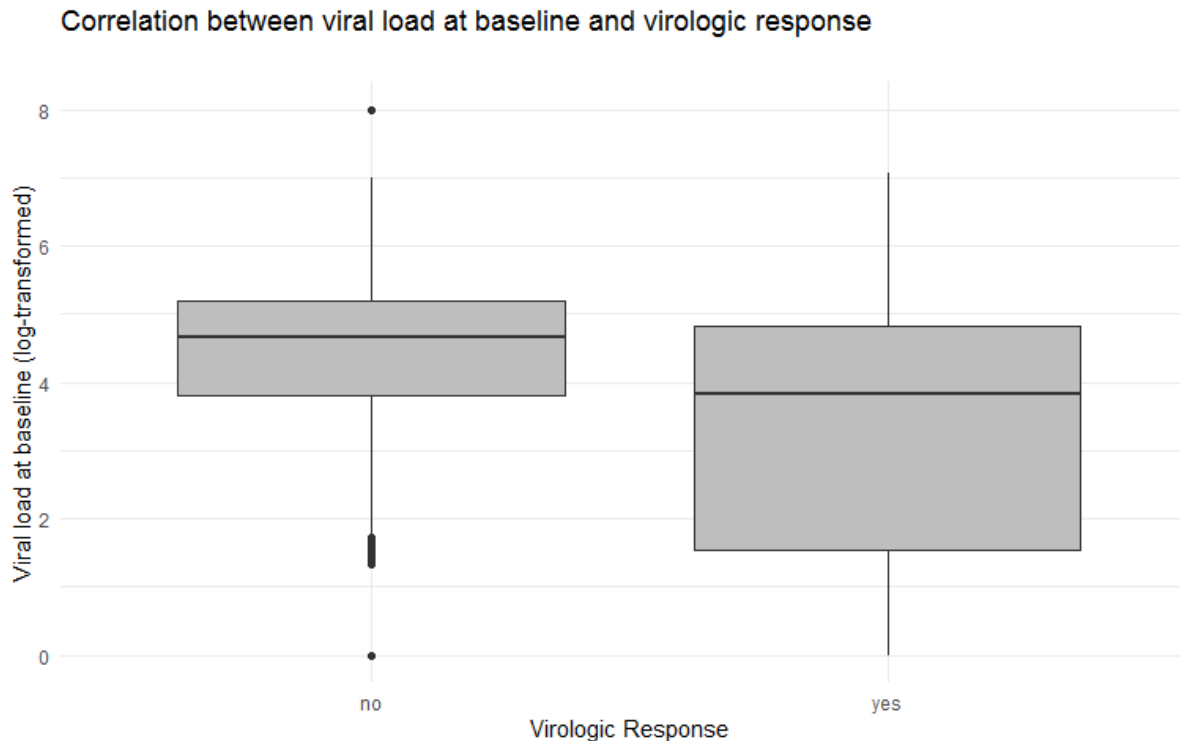
# Basic boxplot

```
ggplot( data = HIVdata ) +  
  aes(x = VirResponse, y = log10(TErgNum.0+1))+  
  geom_boxplot()
```



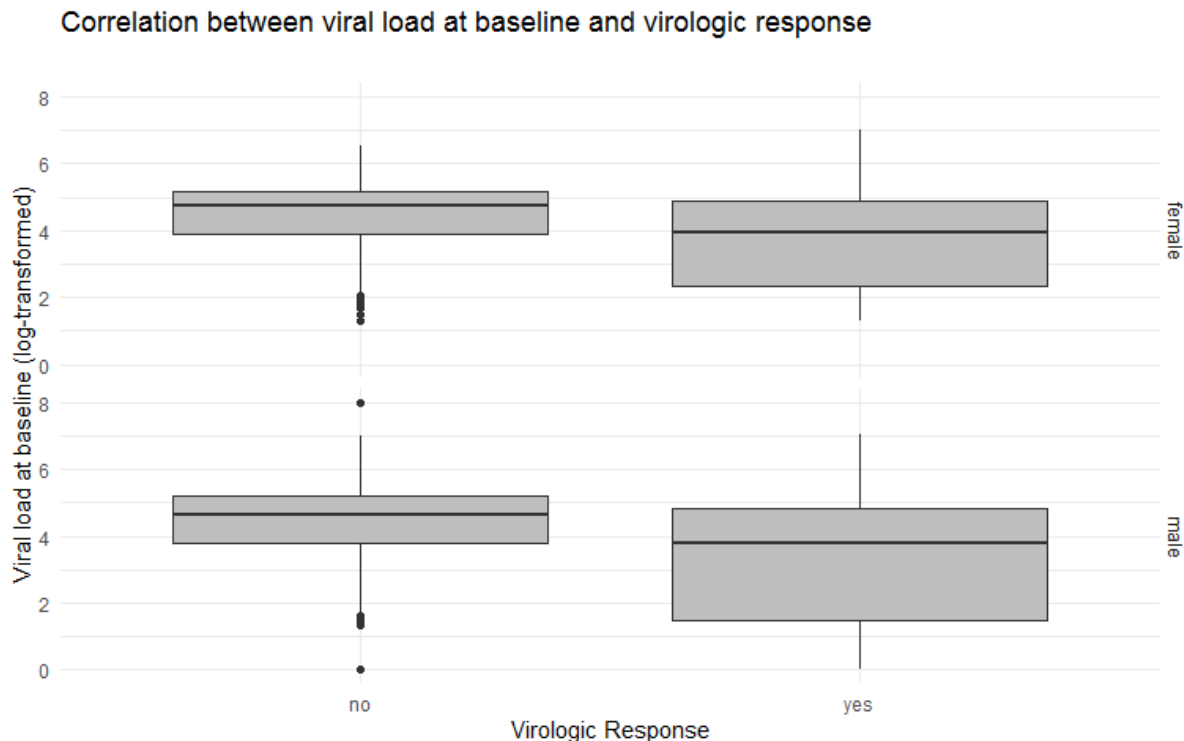
# Boxplot with more features

```
ggplot( data = HIVdata ) +  
  aes(x = virResponse, y = log10(TErgNum.0+1))+  
  geom_boxplot(fill = "grey")+  
  theme_minimal()+  
  ggtitle("Correlation between viral load at baseline and virologic response") +  
  xlab(attributes(HIVdata$virResponse)$label) +  
  ylab(paste(attributes(HIVdata$TErgNum.0)$label, "(log-transformed)"))
```



# Boxplot – stratified for gender

```
ggplot( data = HIVdata ) +  
  aes(x = virResponse, y = log10(TErgNum.0+1))+  
  facet_grid(rows = vars(Gender))+  
  geom_boxplot(fill = "grey")+  
  theme_minimal()+  
  ggtitle("Correlation between viral load at baseline and virologic response") +  
  xlab(attributes(HIVdata$VirResponse)$label) +  
  ylab(paste(attributes(HIVdata$TErgNum.0)$label, "(log-transformed)"))
```

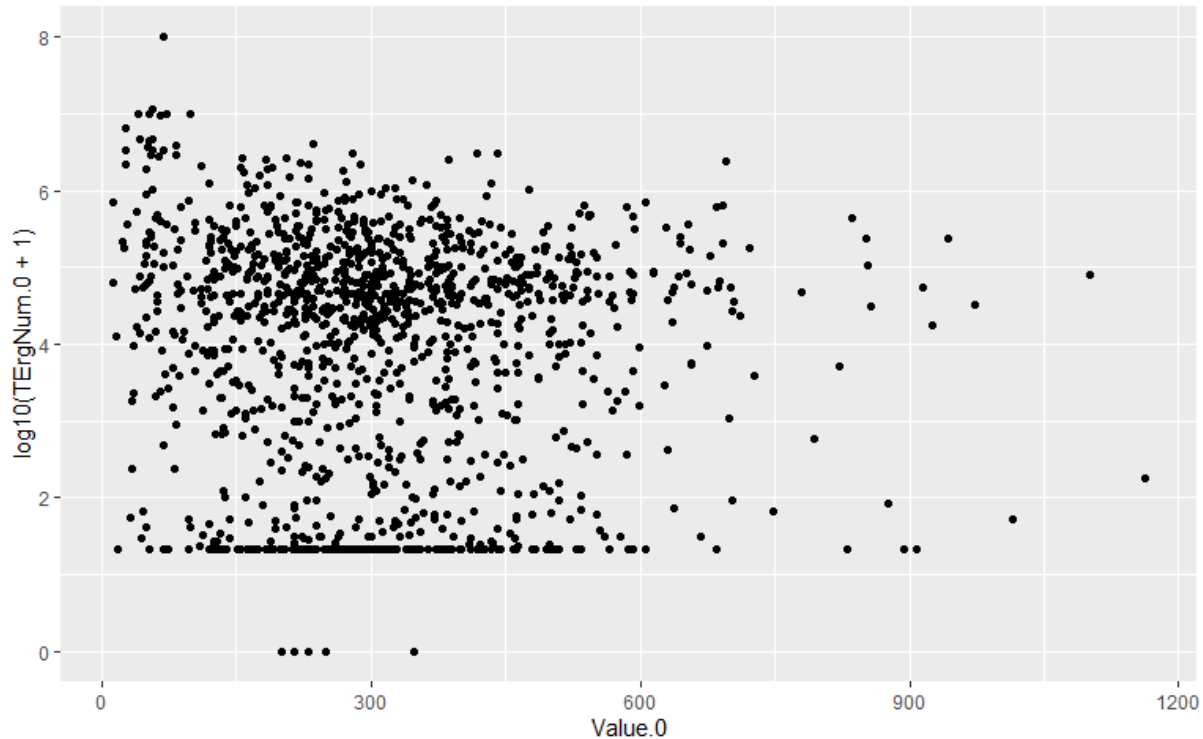


# Plot the data

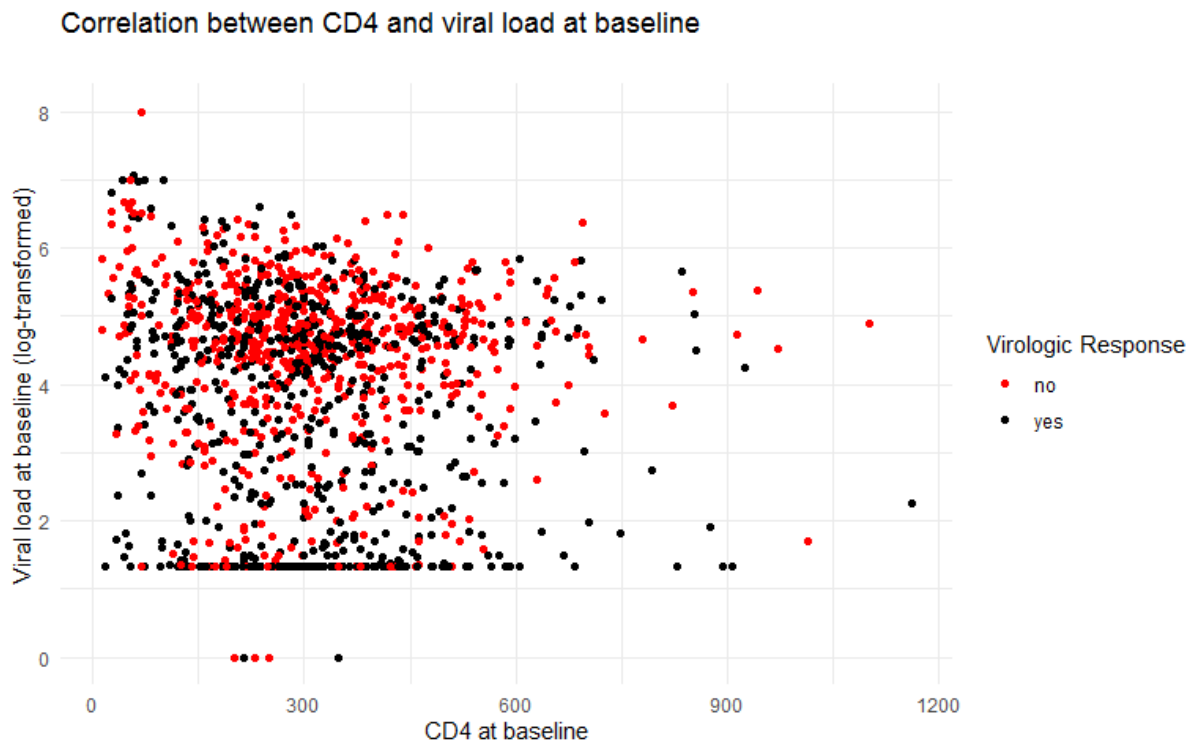
- 2. Use a scatterplot to visualize the correlation between baseline CD4 and baseline viral load

# Basic scatterplot

```
ggplot( data = HIVdata ) +  
  aes(x = value.0, y = log10(TErgNum.0+1))+  
  geom_point()
```



```
ggplot( data = HIVdata ) +
  aes(x = value.0, y = log10(TERGNum.0+1), color=VirResponse)+
  geom_point()+
  theme_minimal()+
  ggtitle("Correlation between CD4 and viral load at baseline", subtitle = " ") +
  xlab(attributes(HIVdata$value.0)$label) +
  ylab(paste(attributes(HIVdata$TERGNum.0)$label, "(log-transformed)"))+
  scale_color_manual(values=c("red", "black"))+
  labs(color= xlab(attributes(HIVdata$VirResponse)$label))
#geom_smooth(span = 1.5)
```



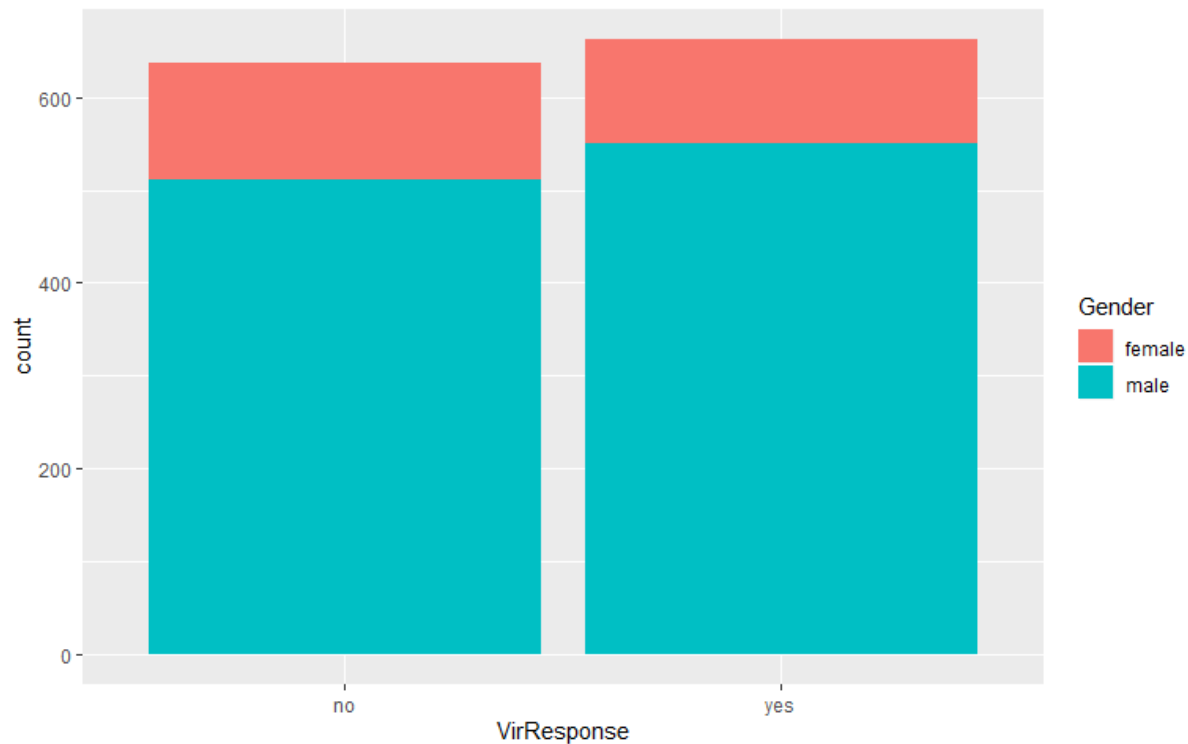


# Plot the data

- 3. Visualize with barplots the association between gender and therapeutic response,
  - b) and between HCV- + HBV-status and therapeutic response

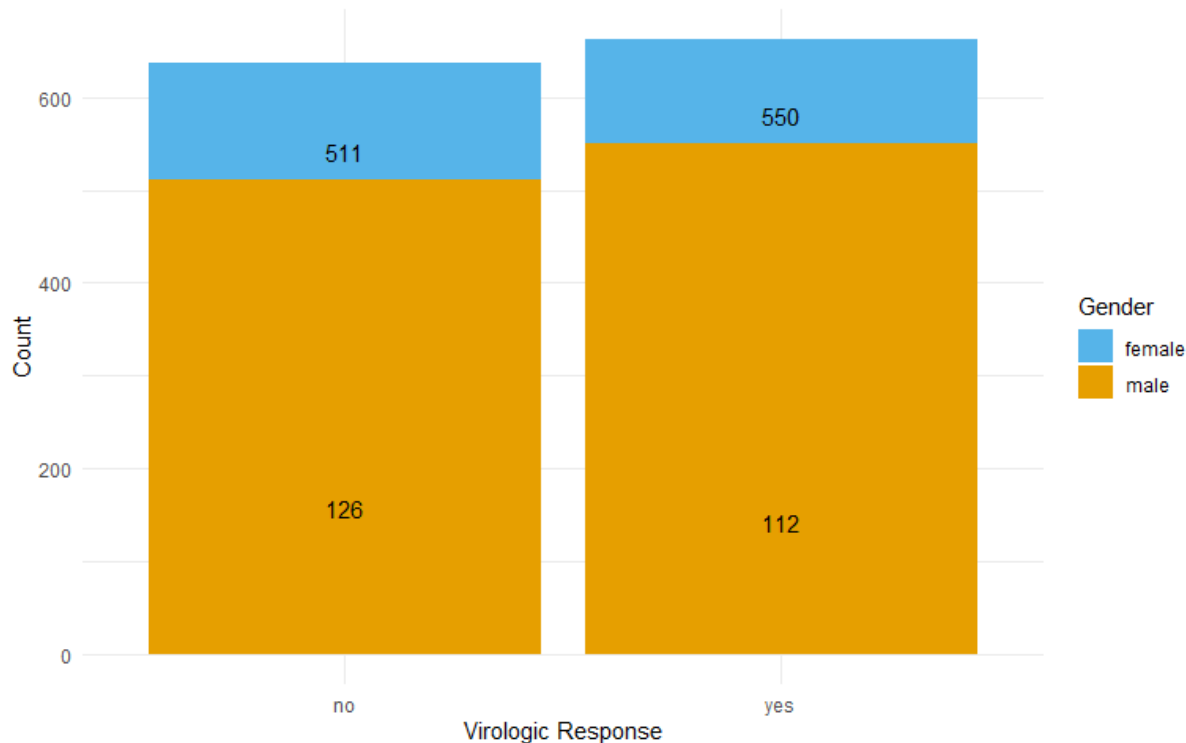
# Basic barplot

```
ggplot( data = HIVdata ) +  
  aes( x = VirResponse, fill = Gender)+  
  geom_bar()
```



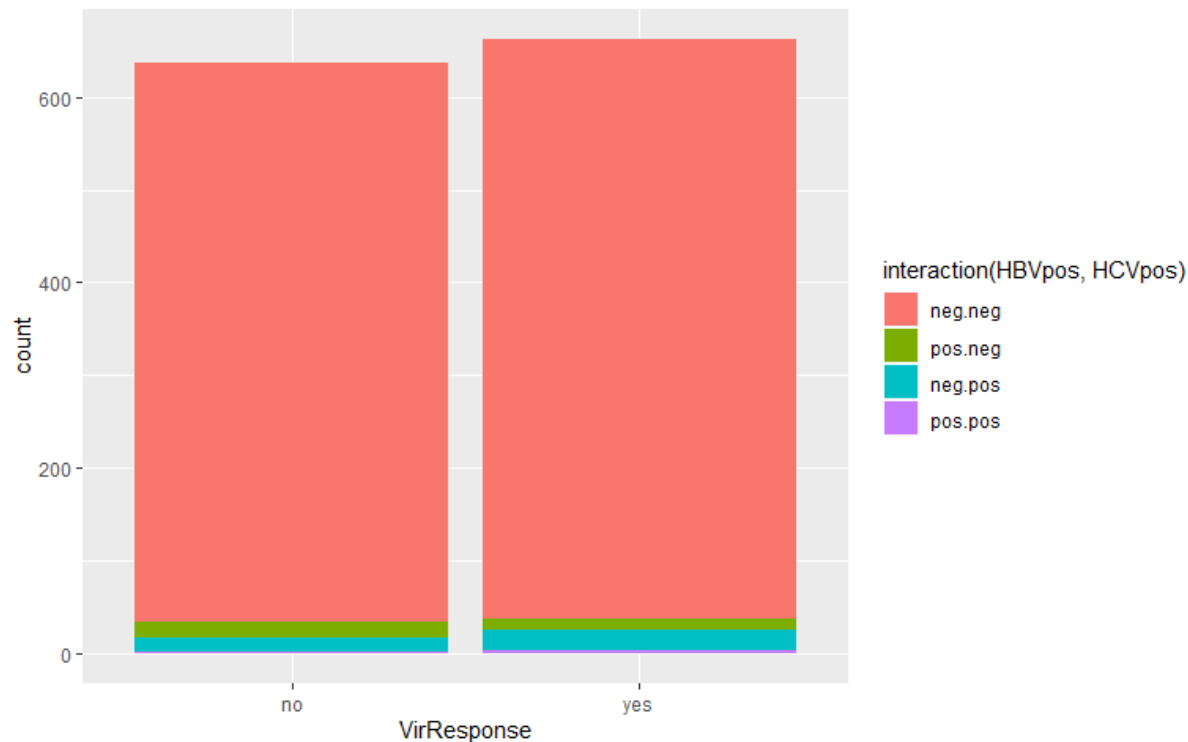
# Barplot with more features

```
ggplot( data = HIVdata ) +  
  aes( x = VirResponse, fill = Gender)+  
  geom_bar(show.legend = TRUE)+  
  theme_minimal()+  
  scale_fill_manual(values=c("#56B4E9", "#E69F00"))+  
  geom_text(stat='count', aes(label=..count..), vjust=-1)+  
  xlab(attributes(HIVdata$VirResponse)$label) +  
  ylab("Count")
```



# Barplot – association between HBV-HCV-status and treatment success

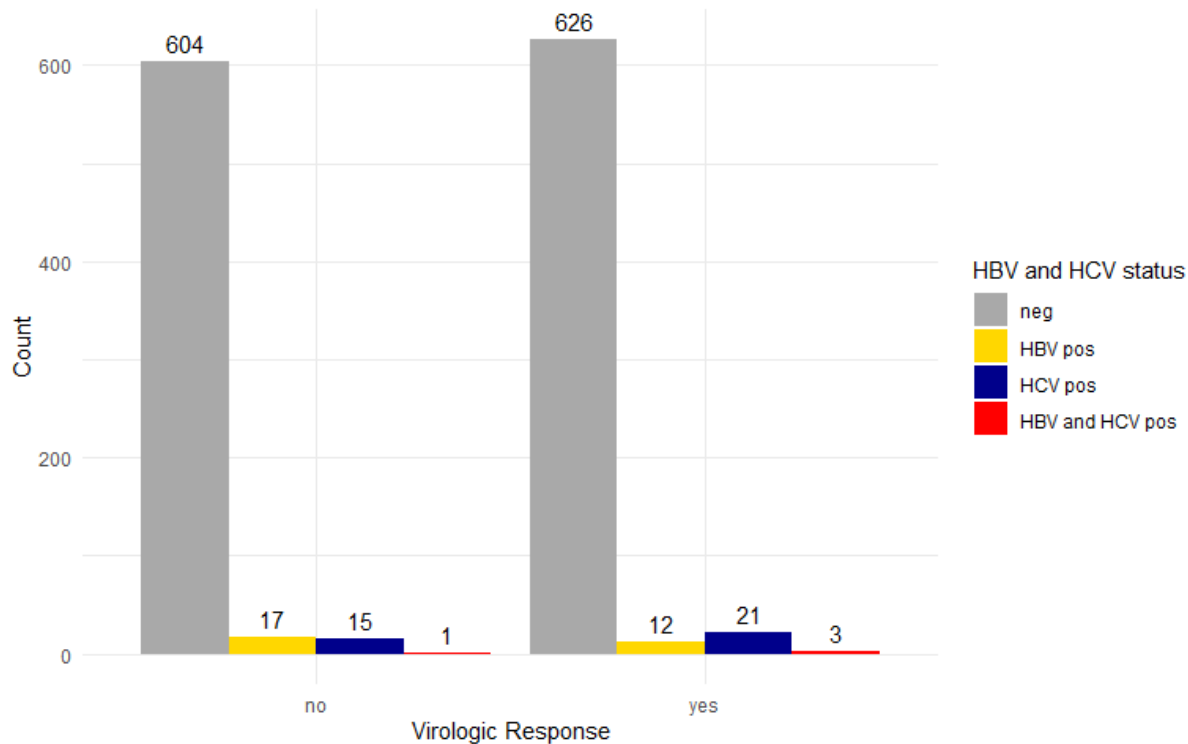
```
ggplot(data = HIVdata) +  
  aes( x = VirResponse, fill = interaction(HBVpos, HCVpos)) +  
  geom_bar(show.legend = TRUE)
```



```

ggplot(data = HIVdata) +
  aes( x = VirResponse, fill = interaction(HBVpos, HCVpos))+
  geom_bar(show.legend = TRUE, position ="dodge")+
  labs(fill="HBV and HCV status")+
  scale_fill_manual(labels = c("neg", "HBV pos", "HCV pos", "HBV and HCV pos"), values
= c("darkgrey", "gold","darkblue", "red"))+
  theme_minimal()+
  geom_text(stat = 'count', aes(label=..count..),
            position = position_dodge(.9),
            vjust = -0.5) +
  xlab(attributes(HIVdata$VirResponse)$label) +
  ylab("Count")

```



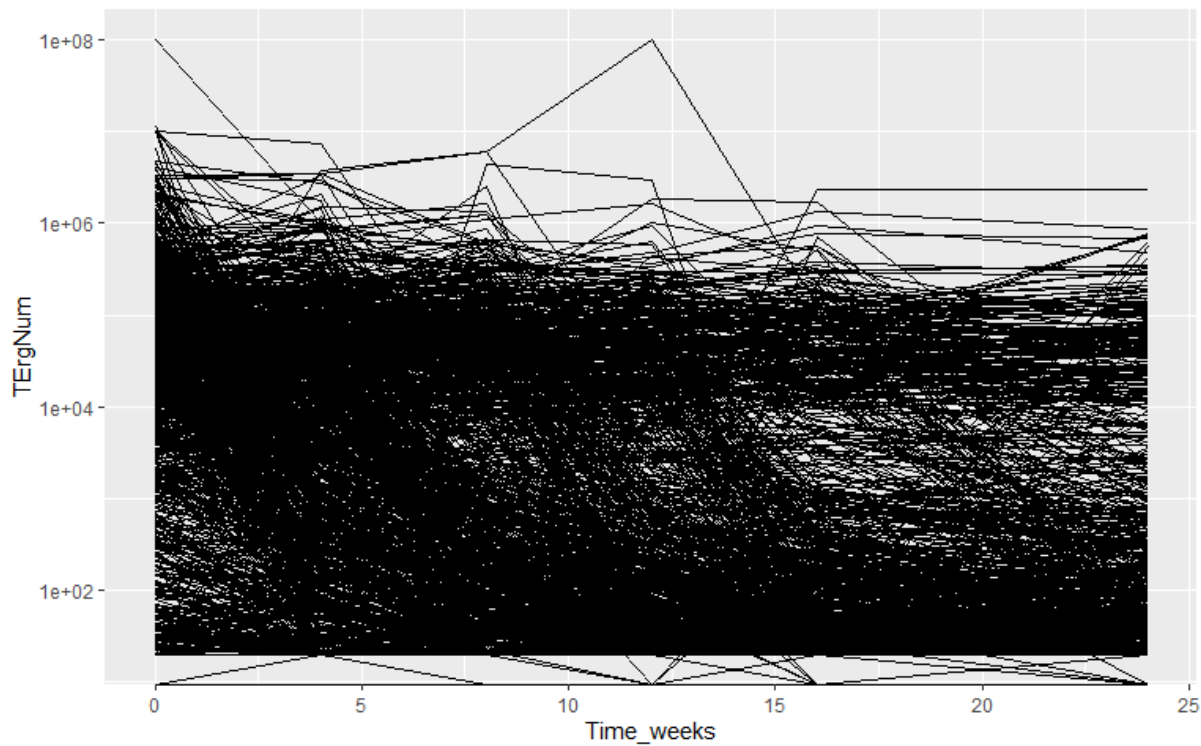
# Plot the data

- 4. Depict the viral kinetics with a spaghetti plot

```
# here the original data formate is more appropriate
HIVdata_long <- full_join(labdata, basedata) %>%
  mutate(Time_weeks = structure(Time_weeks, label = "Time (weeks)"),
  TErgNum = structure(TErgNum, label = "Viral load"))
```

# Basic spaghetti

```
ggplot(data = HIVdata_long)+  
  aes(x = Time_weeks, y = TErgNum, group = PatientID)+  
  scale_y_log10()+  
  geom_line()
```



```

ggplot(data = HIVdata_long, aes(x = Time_weeks, y = TErgNum+1, group = PatientID))+
  geom_line(col="grey")+
  theme_minimal()+
  #facet_grid(. ~ Gender)+
  scale_y_log10()+
  stat_summary(aes(group = 1), geom = "line", fun.y = median,col="red",size = 2) +
  xlab(attributes(HIVdata_long$Time_weeks)$label) +
  ylab(attributes(HIVdata_long$TErgNum)$label)+
  geom_hline(yintercept = 20) +
  annotate("text", 5, 40, label = "Limit of detection", size=3)

```

