Script 2a

Descriptive statistics at enrolment

Peter Kamerman

07 December 2019

Contents

Objective

To describe the demographic characteristics and disease status of the analysis cohort at study enrolment (week 0, baseline).

Analysis notes

Definitions of missingness

Data were regarded as **missing** when *pain in the last week* data were not present for one or more of weeks 0, 12, 24, 36, 48. Data also were classified as **missing** when there were inconsistencies in the data across the variables collected within a week.

Definition of data inconsistencies

Pain was defined as *pain in the last week* being 'Yes', and *pain at its worst* being > 0. These two measurements were then the "gatekeeper" measurements, such that the two measurements both had to be positive ('Yes' and '> 0', respectively) in order for there to be any entries for *site of pain* and *site of worst pain*. Were the data were inconsistent (e.g., when there was no *pain in the last week* and *pain at its worst* $= 0$, but there were entries for *site of pain* and *site of worst pain*), then the *site of pain* and *site of worst pain* entries were marked as **inconsistent**.

Data also were considered **inconsistent** when *pain in the last week* = 'Yes', but *site of worst pain* = 'None'.

Lastly, data were considered **inconsistent** when *site of worst pain* was not listed as one of the pain locations for a given measurement week.

For analysis purposes, missing data in the *site of pain* columns were changed to **'No'** (pain not present in the site). This approach was conservative, but we believed that the approach would have the least effect on the outcome, while still retaining as many participants as possible.

Import data

```
df <- read_rds('data-cleaned/data-ADVANCE.rds')
```

Quick look

```
head(df)
```

```
## # A tibble: 6 x 35
##   ranid interval_name site_name date_of_visit       pain_in_the_las~
##   <chr> <ord>         <chr>     <dttm>              <chr>
## 1 01-0~ 0 weeks       Wits RHI~ 2017-02-01 00:00:00 No
## 2 01-0~ 12 weeks      Wits RHI~ 2017-04-25 00:00:00 No
## 3 01-0~ 24 weeks      Wits RHI~ 2017-07-19 00:00:00 No
## 4 01-0~ 36 weeks      Wits RHI~ 2017-10-11 00:00:00 No
## 5 01-0~ 48 weeks      Wits RHI~ 2018-01-03 00:00:00 No
## 6 01-0~ 0 weeks       Wits RHI~ 2017-02-02 00:00:00 No
## # ... with 30 more variables: where_does_it_hurt_most <chr>,
## #   pain_worst <dbl>, pain_now <dbl>, head_pain <chr>,
## #   cervical_pain <chr>, shoulder_pain <chr>, arm_pain <chr>,
## #   hand_pain <chr>, chest_pain <chr>, abdominal_pain <chr>,
## #   low_back_pain <chr>, buttock_pain <chr>, hip_groin_pain <chr>,
## #   leg_pain <chr>, genital_pain <chr>, foot_pain <chr>, site_worst <chr>,
## #   age <dbl>, sex <chr>, ancestry <chr>, education <chr>,
## #   employment_status <chr>, cd4_cells.ul <dbl>, viral_load_cp.ml <dbl>,
## #   group <chr>, tb_screen <chr>, general_health <dbl>, mms_total <dbl>,
## #   interval_numeric <dbl>, any_missing <chr>
```

```
glimpse(df)
```

```
## Observations: 5,265
## Variables: 35
## $ ranid                   <chr> "01-0001", "01-0001", "01-0001", "01-0...
## $ interval_name           <ord> 0 weeks, 12 weeks, 24 weeks, 36 weeks,...
## $ site_name               <chr> "Wits RHI Yeoville Research Centre", "...
## $ date_of_visit           <dttm> 2017-02-01, 2017-04-25, 2017-07-19, 2...
## $ pain_in_the_last_week   <chr> "No", "No", "No", "No", "No", "No", "Y...
## $ where_does_it_hurt_most <chr> NA, NA, NA, NA, NA, NA, "Hip/groin lef...
## $ pain_worst              <dbl> 0, 0, 0, 0, 0, 0, 3, 3, 5, 0, 0, 0, 0,...
## $ pain_now                <dbl> NA, 0, NA, 0, NA, NA, 0, 2, 4, NA, NA,...
## $ head_pain               <chr> "No", "No", "No", "No", "No", "No", "N...
## $ cervical_pain           <chr> "No", "No", "No", "No", "No", "No", "N...
## $ shoulder_pain           <chr> "No", "No", "No", "No", "No", "No", "N...
## $ arm_pain                <chr> "No", "No", "No", "No", "No", "No", "N...
## $ hand_pain               <chr> "No", "No", "No", "No", "No", "No", "N...
## $ chest_pain              <chr> "No", "No", "No", "No", "No", "No", "N...
## $ abdominal_pain          <chr> "No", "No", "No", "No", "No", "No", "N...
## $ low_back_pain           <chr> "No", "No", "No", "No", "No", "No", "N...
## $ buttock_pain            <chr> "No", "No", "No", "No", "No", "No", "N...
## $ hip_groin_pain          <chr> "No", "No", "No", "No", "No", "No", "Y...
## $ leg_pain                <chr> "No", "No", "No", "No", "No", "No", "N...
## $ genital_pain            <chr> "No", "No", "No", "No", "No", "No", "N...
## $ foot_pain               <chr> "No", "No", "No", "No", "No", "No", "N...
## $ site_worst              <chr> "None", "None", "None", "None", "None"...
## $ age                     <dbl> 30, 30, 30, 30, 30, 34, 34, 34, 34, 34...
## $ sex                     <chr> "Male", "Male", "Male", "Male", "Male"...
## $ ancestry                <chr> "Black", "Black", "Black", "Black", "B...
## $ education               <chr> "Secondary", "Secondary", "Secondary",...
## $ employment_status       <chr> "Employed", "Employed", "Employed", "E...
## $ cd4_cells.ul            <dbl> 642, NA, 525, NA, 668, 241, NA, 364, N...
## $ viral_load_cp.ml        <dbl> 641, 50, 50, 50, 50, 3851, 50, 50, 50,...
## $ group                   <chr> "GROUP 1 (DTG + TAF + FTC)", "GROUP 1 ...
## $ tb_screen               <chr> "Negative", "Negative", "Negative", "N...
## $ general_health          <dbl> 4, 4, 5, 5, 4, 3, 5, 3, 3, 3, 4, 5, 5,...
## $ mms_total               <dbl> 0, 0, 0, 0, 0, 0, 7, 0, 3, 1, 0, 0, 0,...
## $ interval_numeric        <dbl> 0, 12, 24, 36, 48, 0, 12, 24, 36, 48, ...
## $ any_missing             <chr> "No", "No", "No", "No", "No", "No", "N...
```

Basic clean

```
# Remove missing data
df %<>%
    filter(any_missing == 'No')

# Extract enrolment data
df %<>%
    filter(interval_name == '0 weeks')
```

Quick tabulation

Analysis data set for the period 0 to 48 weeks

```
# Tabulate data
xtabs(~interval_name, data = df)
```

```
## interval_name
##  0 weeks 12 weeks 24 weeks 36 weeks 48 weeks
##      787        0        0        0        0
```

Basic summary

```
skim(df) %>%
    skimr::kable(caption = 'Quick summary')
```

Skim summary statistics
n obs: 787
n variables: 35

Variable type: character

variable	missing	complete	n	n_unique
abdominal_pain	0	787	787	2
ancestry	0	787	787	2
any_missing	0	787	787	1
arm_pain	0	787	787	2
buttock_pain	0	787	787	2
cervical_pain	0	787	787	2
chest_pain	0	787	787	2
education	4	783	787	4
employment_status	10	777	787	4
foot_pain	0	787	787	2
genital_pain	0	787	787	2
group	0	787	787	3
hand_pain	0	787	787	2
head_pain	0	787	787	2
hip_groin_pain	0	787	787	2
leg_pain	0	787	787	2
low_back_pain	0	787	787	2
pain_in_the_last_week	0	787	787	2
ranid	0	787	787	787
sex	0	787	787	2
shoulder_pain	0	787	787	2
site_name	0	787	787	2
site_worst	0	787	787	14
tb_screen	0	787	787	2
where_does_it_hurt_most	636	151	787	25

Variable type: factor

variable	missing	complete	n	n_unique	top_counts	ordered
interval_name	0	787	787	1	0 w: 787, 12 : 0, 24 : 0, 36 : 0	TRUE

Variable type: numeric

variable	missing	complete	n	mean	sd	p0	p25	p50	p75	p100
age	0	787	787	32.77	7.65	14	27	32	38	62
cd4_cells.ul	0	787	787	333.25	224.05	1	173.5	290	441.5	1721
general_health	4	783	787	3.45	0.82	1	3	3	4	5
interval_numeric	0	787	787	0	0	0	0	0	0	0
mms_total	5	782	787	1.08	2.13	0	0	0	1	17
pain_now	629	158	787	2	2.13	0	0	2	3	9
pain_worst	0	787	787	0.88	2.08	0	0	0	0	10
viral_load_cp.ml	0	787	787	98611.6	386719.99	50	5704.5	25853	85574	9475772

Variable type: POSIXct

variable	missing	complete	n	min	max	median	n_unique
date_of_visit	0	787	787	2017-02-01	2018-05-17	2017-09-12	247

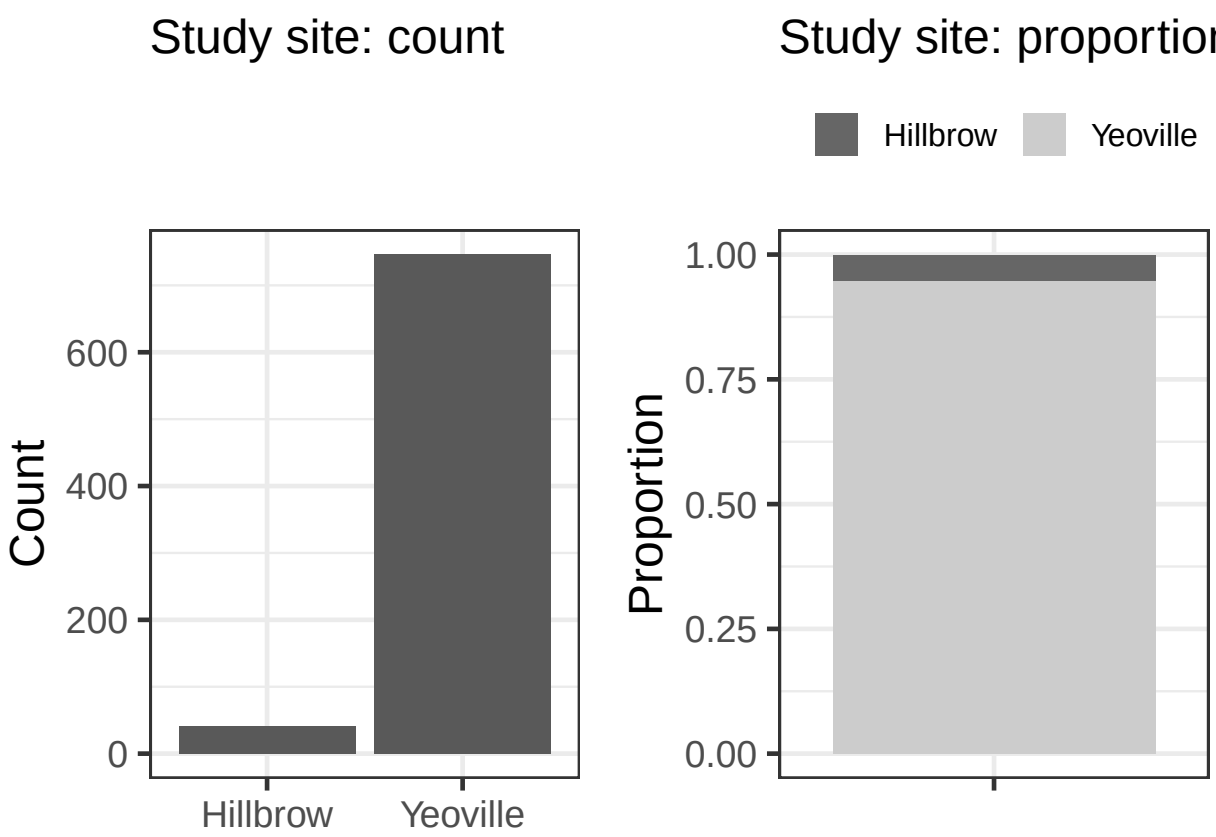
Study characteristics

Study site

```
# Plot
site_count <- ggplot(data = df) +
    aes(x = site_name) +
    geom_bar() +
    labs(subtitle = 'Study site: count',
         y = 'Count') +
    scale_x_discrete(labels = c('Hillbrow', 'Yeoville')) +
    theme(axis.title.x = element_blank())

site_prop <- df %>%
    group_by(site_name) %>%
    summarise(count = n()) %>%
    ggplot(data = .) +
    aes(x = '',
        y = count,
        fill = site_name) +
    geom_col(position = position_fill()) +
    labs(subtitle = 'Study site: proportion',
         y = 'Proportion') +
    scale_fill_manual(values = c('#666666', '#CCCCCC'),
                      labels = c('Hillbrow', 'Yeoville')) +
    theme(legend.title = element_blank(),
          legend.text = element_text(size = 12),
          legend.position = 'top',
          axis.title.x = element_blank())

site_count + site_prop
```




```
# Numeric summary
df %>%
    mutate(site_name = factor(site_name,
                              labels = c('Hillbrow', 'Yeoville'))) %>%
    group_by(site_name) %>%
    summarise(count = n()) %>%
    mutate(n = sum(count),
           missing = sum(is.na(df$site_name))) %>%
    mutate(proportion = round(count / n, 3)) %>%
    select(site_name, count, n, missing, proportion) %>%
    knitr::kable(caption = 'Study site: summary statistics')
```

Table 5: Study site: summary statistics

site_name	count	n	missing	proportion
Hillbrow	41	787	0	0.052
Yeoville	746	787	0	0.948

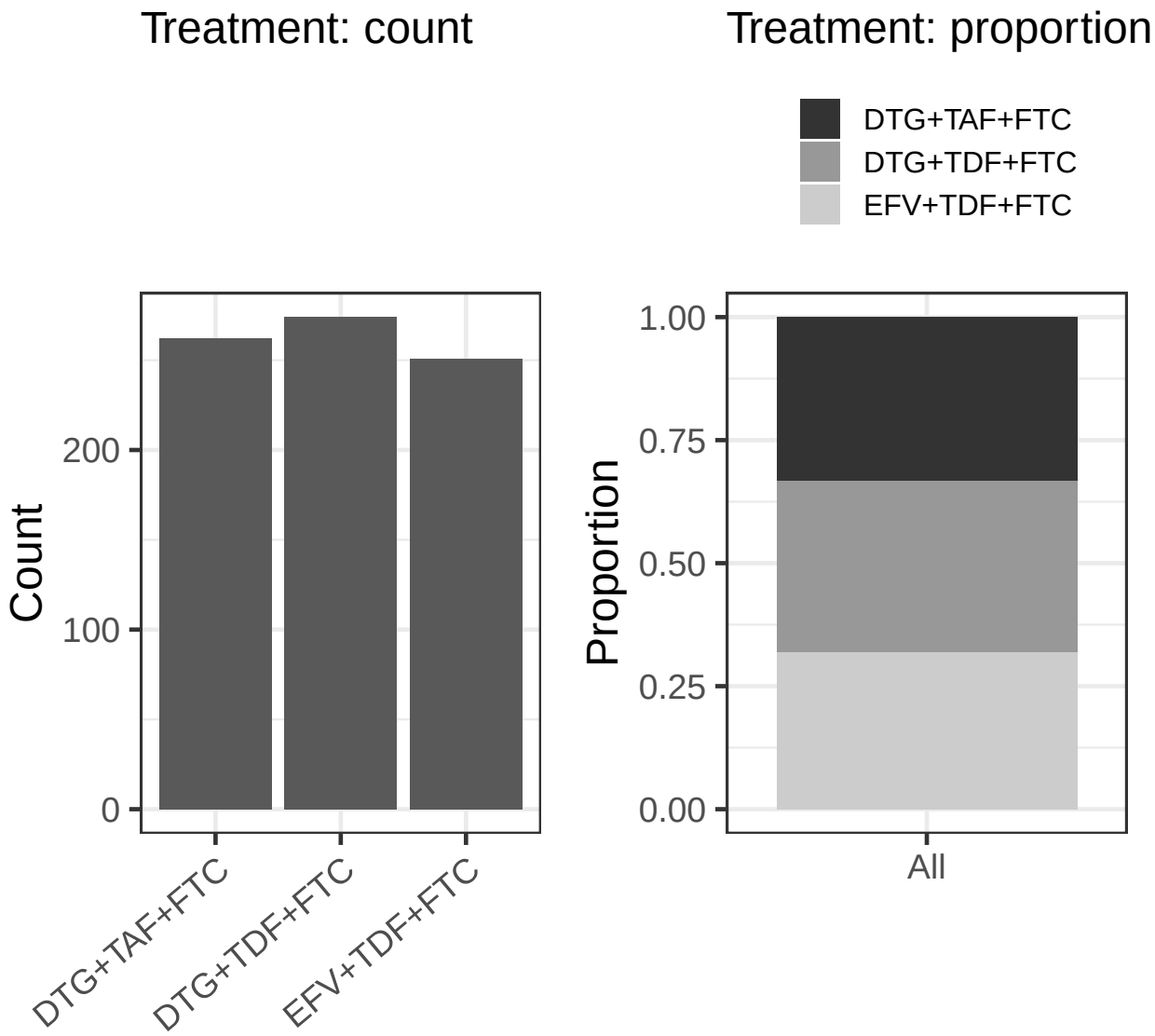
Treatment group allocation

```
# Plot
group_count <- ggplot(data = df) +
    aes(x = group) +
    geom_bar() +
    labs(subtitle = 'Treatment: count',
         y = 'Count') +
    scale_x_discrete(labels = c('DTG+TAF+FTC',
                                'DTG+TDF+FTC',
                                'EFV+TDF+FTC')) +
    theme(axis.title.x = element_blank(),
```

```
          axis.text.x = element_text(angle = 40, hjust = 1))

group_prop <- df %>%
    group_by(group) %>%
    summarise(count = n()) %>%
    ggplot(data = .) +
    aes(x = 'All',
        y = count,
        fill = group) +
    geom_col(position = position_fill()) +
    labs(subtitle = 'Treatment: proportion',
         y = 'Proportion') +
    scale_fill_grey(guide = guide_legend(ncol = 1),
                    labels = c('DTG+TAF+FTC',
                               'DTG+TDF+FTC',
                               'EFV+TDF+FTC')) +
    theme(legend.title = element_blank(),
          legend.text = element_text(size = 12),
          legend.position = 'top',
          axis.title.x = element_blank())

group_count + group_prop
```



```r
# Numeric summary
df %>%
    select(group) %>%
    group_by(group) %>%
    summarise(count = n(),
              missing = sum(is.na(group))) %>%
    mutate(n = sum(count),
           proportion = round(count / n, 3)) %>%
    select(group, count, proportion, missing, n) %>%
    knitr::kable(caption = 'Study group allocation: summary statistics')
```

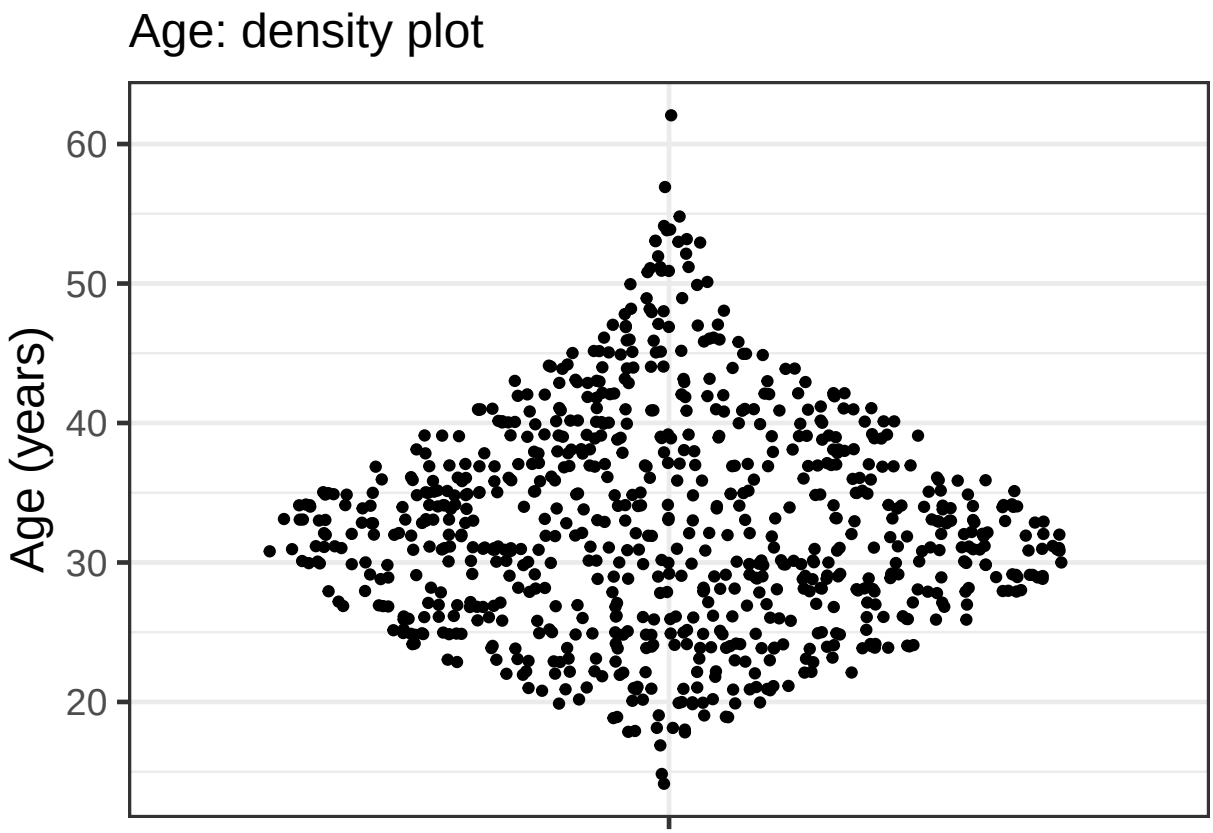
Table 6: Study group allocation: summary statistics

group	count	proportion	missing	n
GROUP 1 (DTG + TAF + FTC)	262	0.333	0	787
GROUP 2 (DTG + TDF + FTC)	274	0.348	0	787
GROUP 3 (EFV + TDF + FTC)	251	0.319	0	787

Demographics

Age

```r
# Plot
ggplot(data = df) +
    aes(x = 'Data',
        y = age) +
    geom_sina() +
    labs(subtitle = 'Age: density plot',
         y = 'Age (years)') +
    theme(axis.title.x = element_blank(),
          axis.text.x = element_blank())
```




```
# Numeric summary
df %>%
    select(age) %>%
    skim() %>%
    skimr::kable(caption = 'Age: summary statistics')
```

Skim summary statistics
n obs: 787
n variables: 1

Variable type: numeric

variable	missing	complete	n	mean	sd	p0	p25	p50	p75	p100
age	0	787	787	32.77	7.65	14	27	32	38	62

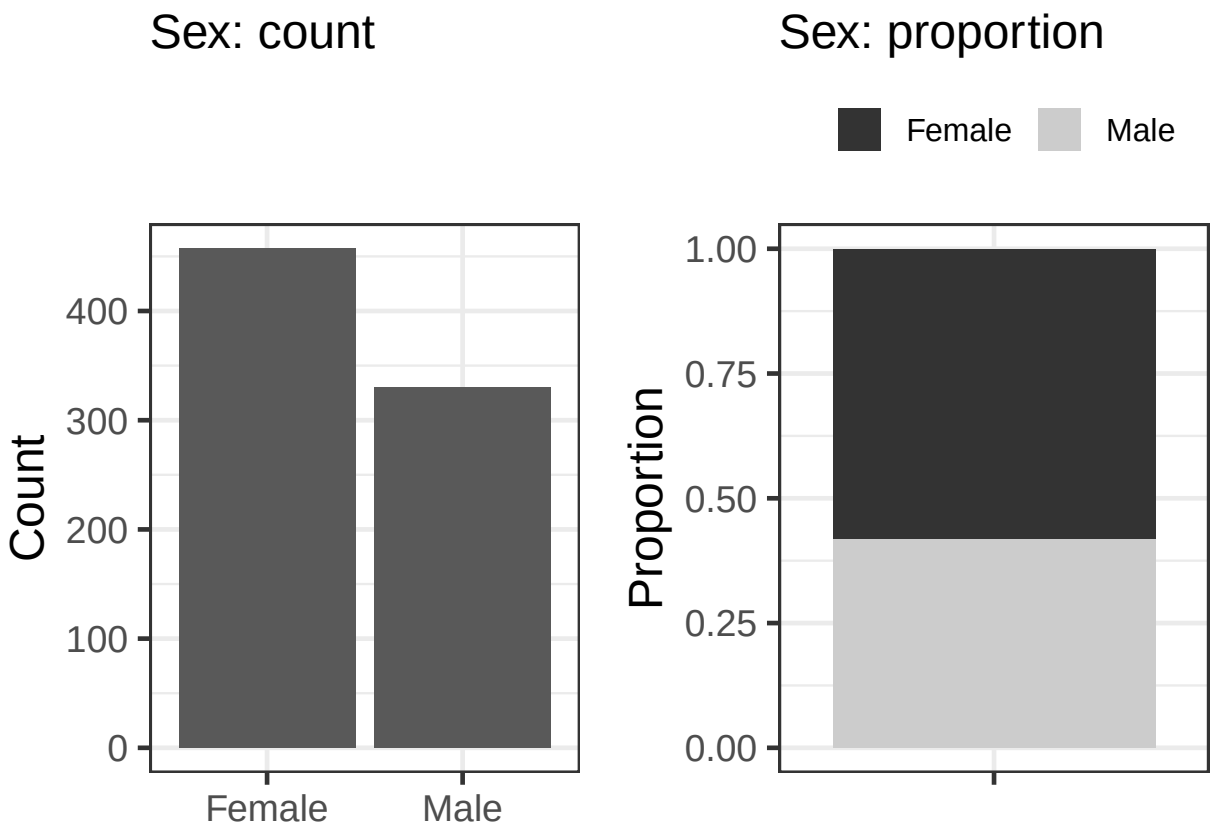
Sex

```
# Plot
sex_count <- ggplot(data = df) +
    aes(x = sex) +
    geom_bar() +
    labs(subtitle = 'Sex: count',
         y = 'Count') +
    theme(axis.title.x = element_blank())

sex_prop <- df %>%
    group_by(sex) %>%
    summarise(count = n()) %>%
    ggplot(data = .) +
    aes(x = 'Data',
        y = count,
```

```
        fill = sex) +
    geom_col(position = position_fill()) +
    labs(subtitle = 'Sex: proportion',
         y = 'Proportion') +
    scale_fill_grey() +
    theme(legend.title = element_blank(),
          legend.text = element_text(size = 12),
          legend.position = 'top',
          axis.title.x = element_blank(),
          axis.text.x = element_blank())

sex_count + sex_prop
```




```
# Numeric summary
df %>%
    select(sex) %>%
    mutate(sex = factor(sex)) %>%
    skim() %>%
    skimr::kable(caption = 'Sex: summary statistics')
```

Skim summary statistics
n obs: 787
n variables: 1

Variable type: factor

variable	missing	complete	n	n_unique	top_counts	ordered
sex	0	787	787	2	Fem: 457, Mal: 330, NA: 0	FALSE

```
df %>%
    group_by(sex) %>%
    summarise(count = n()) %>%
```

```
    mutate(n = sum(count),
           missing = sum(is.na(df$sex))) %>%
    mutate(proportion = round(count / n, 3)) %>%
    select(sex, count, n, missing, proportion) %>%
    knitr::kable(caption = 'Sex: summary statistics 2')
```

Table 9: Sex: summary statistics 2

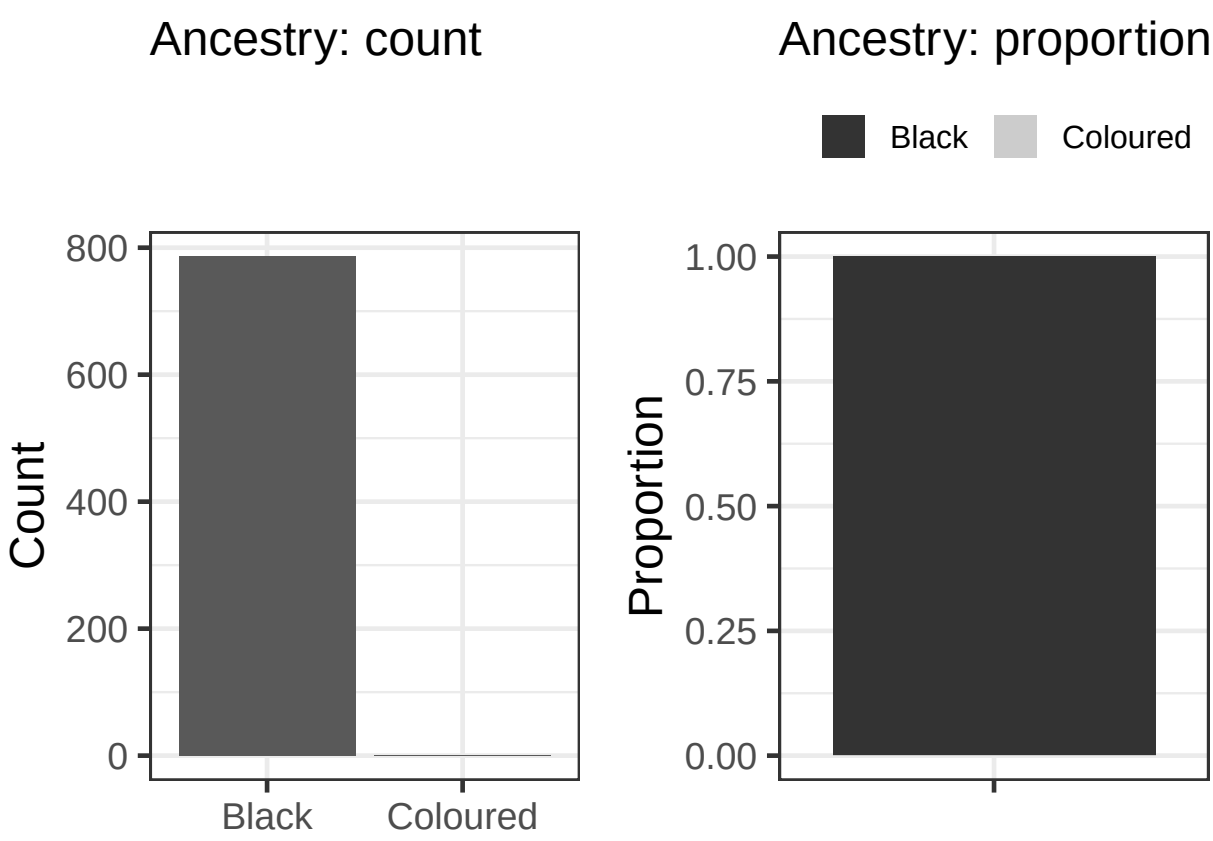
sex	count	n	missing	proportion
Female	457	787	0	0.581
Male	330	787	0	0.419

Ancestry

```
# Plot
anc_count <- ggplot(data = df) +
    aes(x = ancestry) +
    geom_bar() +
    labs(subtitle = 'Ancestry: count',
         y = 'Count') +
    theme(axis.title.x = element_blank())

anc_prop <- df %>%
    group_by(ancestry) %>%
    summarise(count = n()) %>%
    ggplot(data = .) +
    aes(x = 'Data',
        y = count,
        fill = ancestry) +
    geom_col(position = position_fill()) +
    labs(subtitle = 'Ancestry: proportion',
         y = 'Proportion') +
    scale_fill_grey() +
    theme(legend.title = element_blank(),
          legend.text = element_text(size = 12),
          legend.position = 'top',
          axis.title.x = element_blank(),
          axis.text.x = element_blank())

anc_count + anc_prop
```




```
# Numeric summary
df %>%
    select(ancestry) %>%
    mutate(ancestry= factor(ancestry)) %>%
    skim() %>%
    skimr::kable(caption = 'Ancestry: summary statistics')
```

Skim summary statistics
n obs: 787
n variables: 1

Variable type: factor

variable	missing	complete	n	n_unique	top_counts	ordered
ancestry	0	787	787	2	Bla: 786, Col: 1, NA: 0	FALSE

```
df %>%
    group_by(ancestry) %>%
    summarise(count = n()) %>%
    mutate(n = sum(count),
           missing = sum(is.na(df$ancestry))) %>%
    mutate(proportion = round(count / n, 3)) %>%
    select(ancestry, count, n, missing, proportion) %>%
    knitr::kable(caption = 'Ancestry: summary statistics 2')
```

Table 11: Ancestry: summary statistics 2

ancestry	count	n	missing	proportion
Black	786	787	0	0.999
Coloured	1	787	0	0.001

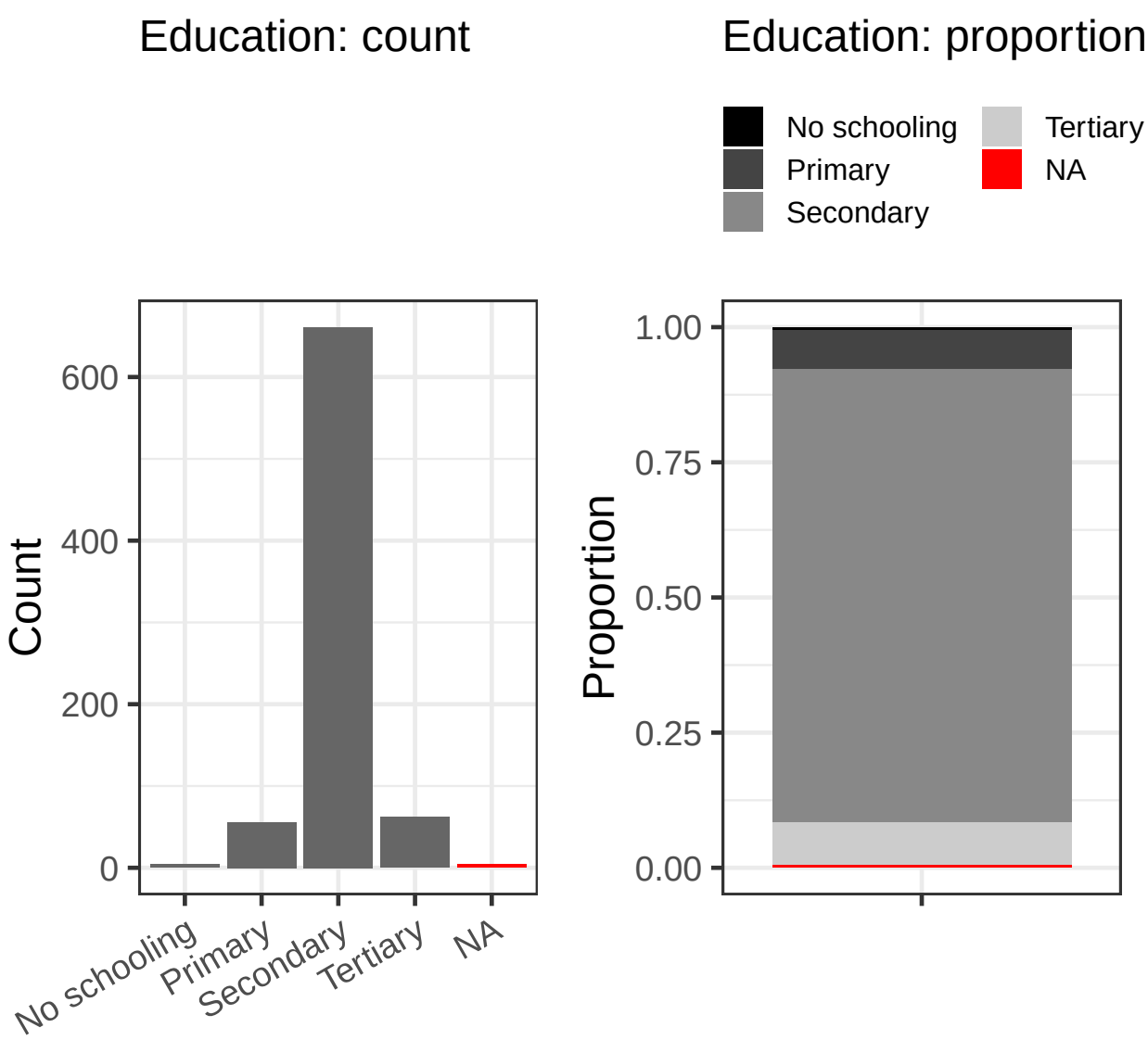
Education

```
# Plot
edu_count <- df %>%
    mutate(education = str_replace_na(education)) %>%
    mutate(education = factor(education,
                              levels = c('No schooling', 'Primary',
                                         'Secondary', 'Tertiary',
                                         'NA'),
                              ordered = TRUE)) %>%
    ggplot(data = .) +
    aes(x = education,
        fill = education) +
    geom_bar() +
    labs(subtitle = 'Education: count',
         y = 'Count') +
        scale_fill_manual(values = c(rep('#666666', 4), '#FF0000')) +
    theme(legend.position = 'none',
          axis.title.x = element_blank(),
          axis.text.x = element_text(angle = 30, hjust = 1))

grey_pal <- colorRampPalette(colors = c('#CCCCCC', '#000000'),
                             interpolate = 'linear')
grey_red <- c(rev(grey_pal(4)), '#FF0000')

edu_prop <- df %>%
    mutate(education = str_replace_na(education)) %>%
    mutate(education = factor(education,
                              levels = c('No schooling', 'Primary',
                                         'Secondary', 'Tertiary',
                                         'NA'),
                              ordered = TRUE)) %>%
    group_by(education) %>%
    summarise(count = n()) %>%
    ggplot(data = .) +
    aes(x = 'Data',
        y = count,
        fill = education) +
    geom_col(position = position_fill()) +
    labs(subtitle = 'Education: proportion',
         y = 'Proportion') +
    scale_fill_manual(values = grey_red,
                      guide = guide_legend(ncol = 2)) +
    theme(legend.title = element_blank(),
          legend.text = element_text(size = 12),
          legend.position = 'top',
          axis.title.x = element_blank(),
          axis.text.x = element_blank())

edu_count + edu_prop
```




```
# Numeric summary
df %>%
    select(education) %>%
    mutate(education = factor(education)) %>%
    skim() %>%
    skimr::kable(caption = 'Education: summary statistics')
```

Skim summary statistics
n obs: 787
n variables: 1

Variable type: factor

variable	missing	complete	n	n_unique	top_counts	ordered
education	4	783	787	4	Sec: 661, Ter: 62, Pri: 56, No : 4	FALSE

```
df %>%
    group_by(education) %>%
    summarise(count = n()) %>%
    mutate(n = sum(count),
           missing = sum(is.na(df$education)),
           n = n - missing) %>%
    mutate(proportion = round(count / n, 3)) %>%
```

```
    select(education, count, n, missing, proportion) %>%
    filter(education != 'NA') %>%
    knitr::kable(caption = 'Education: summary statistics 2')
```

Table 13: Education: summary statistics 2

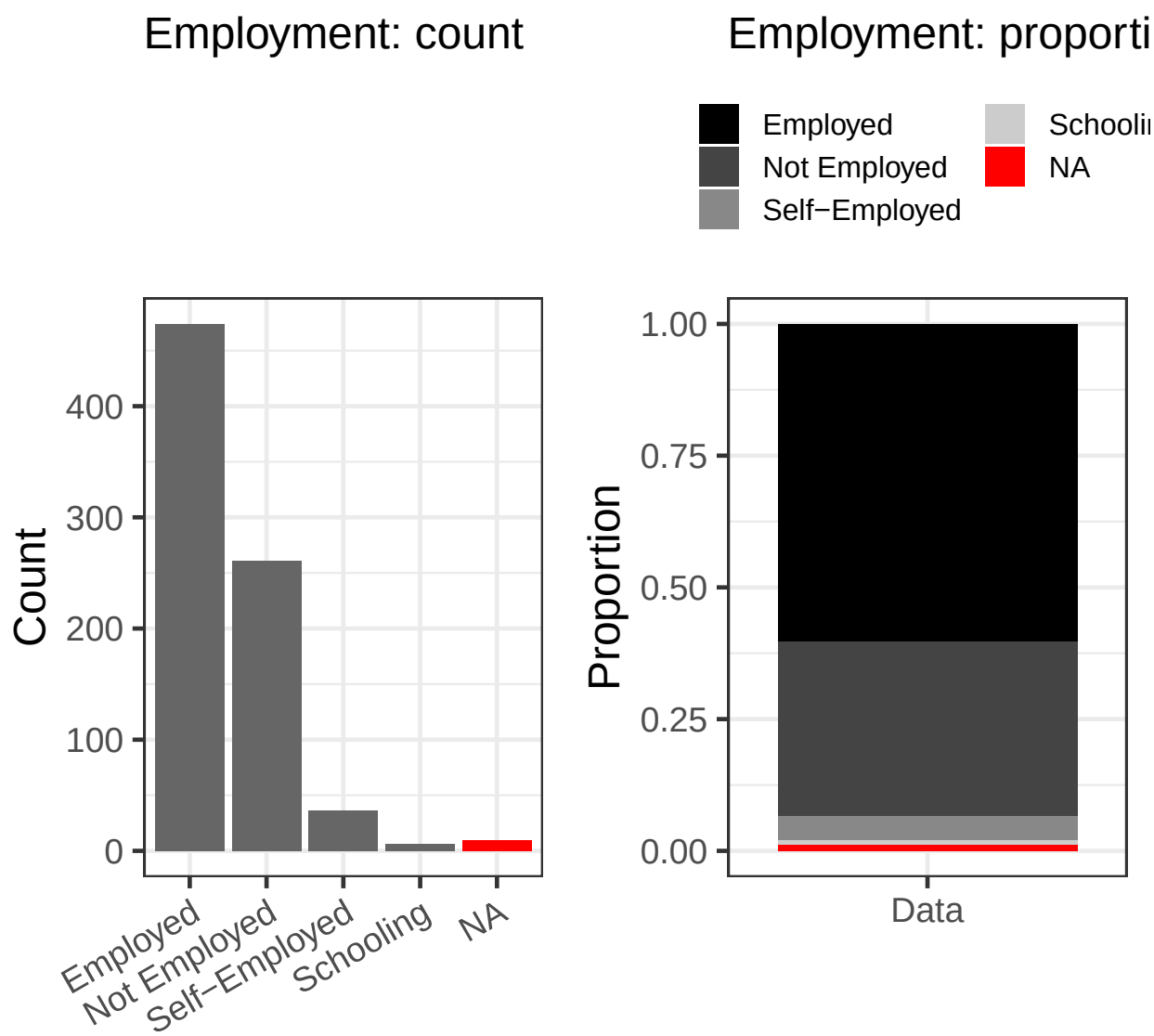
education	count	n	missing	proportion
No schooling	4	783	4	0.005
Primary	56	783	4	0.072
Secondary	661	783	4	0.844
Tertiary	62	783	4	0.079

Employment

```
# Plot
emp_count <- df %>%
    mutate(employment_status = str_replace_na(employment_status)) %>%
    mutate(employment_status = factor(employment_status,
                                      levels = c('Employed', 'Not Employed',
                                                 'Self-Employed', 'Schooling',
                                                 'NA'),
                                      ordered = TRUE)) %>%
    ggplot(data = .) +
    aes(x = employment_status,
        fill = employment_status) +
    geom_bar() +
    labs(subtitle = 'Employment: count',
         y = 'Count') +
    scale_fill_manual(values = c(rep('#666666', 4), '#FF0000')) +
    theme(legend.position = 'none',
          axis.title.x = element_blank(),
          axis.text.x = element_text(angle = 30, hjust = 1))

emp_prop <- df %>%
    mutate(employment_status = str_replace_na(employment_status)) %>%
    mutate(employment_status = factor(employment_status,
                                      levels = c('Employed', 'Not Employed',
                                                 'Self-Employed', 'Schooling',
                                                 'NA'),
                                      ordered = TRUE)) %>%
    group_by(employment_status) %>%
    summarise(count = n()) %>%
    ggplot(data = .) +
    aes(x = 'Data',
        y = count,
        fill = employment_status) +
    geom_col(position = position_fill()) +
    labs(subtitle = 'Employment: proportion',
         y = 'Proportion') +
    scale_fill_manual(values = grey_red,
                      guide = guide_legend(ncol = 2)) +
    theme(legend.title = element_blank(),
          legend.text = element_text(size = 12),
          legend.position = 'top',
          axis.title.x = element_blank())
```

```
emp_count + emp_prop
```




```
# Numeric summary
df %>%
    select(employment_status) %>%
    mutate(employment_status = factor(employment_status)) %>%
    skim() %>%
    skimr::kable(caption = 'Employment status: summary statistics')
```

Skim summary statistics
n obs: 787
n variables: 1

Variable type: factor

variable	missing	complete	n	n_unique	top_counts	ordered
employment_status	10	777	787	4	Emp: 474, Not: 261, Sel: 36, NA: 10	FALSE

```
df %>%
    group_by(employment_status) %>%
    summarise(count = n()) %>%
    mutate(n = sum(count),
           missing = sum(is.na(df$employment_status)),
```

```
          n = n - missing) %>%
    mutate(proportion = round(count / n, 3)) %>%
    select(employment_status, count, n, missing, proportion) %>%
    filter(employment_status != 'NA') %>%
    knitr::kable(caption = 'Employment status: summary statistics 2')
```

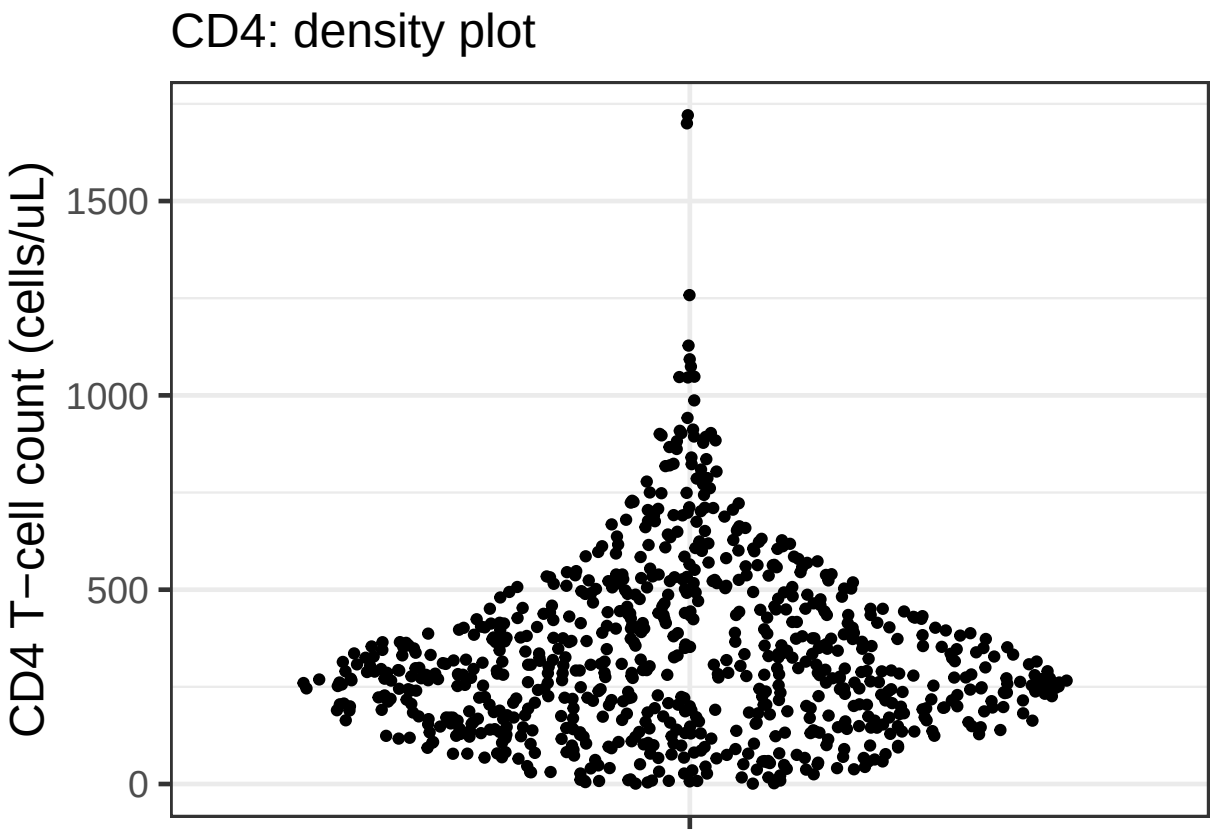
Table 15: Employment status: summary statistics 2

employment_status	count	n	missing	proportion
Employed	474	777	10	0.610
Not Employed	261	777	10	0.336
Schooling	6	777	10	0.008
Self-Employed	36	777	10	0.046

Clinical

CD4 T-cell count

```
# Plot
ggplot(data = df) +
    aes(x = 'Data',
        y = cd4_cells.ul) +
    geom_sina() +
    labs(subtitle = 'CD4: density plot',
         y = 'CD4 T-cell count (cells/uL)') +
    theme(axis.title.x = element_blank(),
          axis.text.x = element_blank())
```



```r
# Numeric summary
df %>%
    select(cd4_cells.ul) %>%
    skim() %>%
    skimr::kable(caption = 'CD4 T-cell count: summary statistics')
```

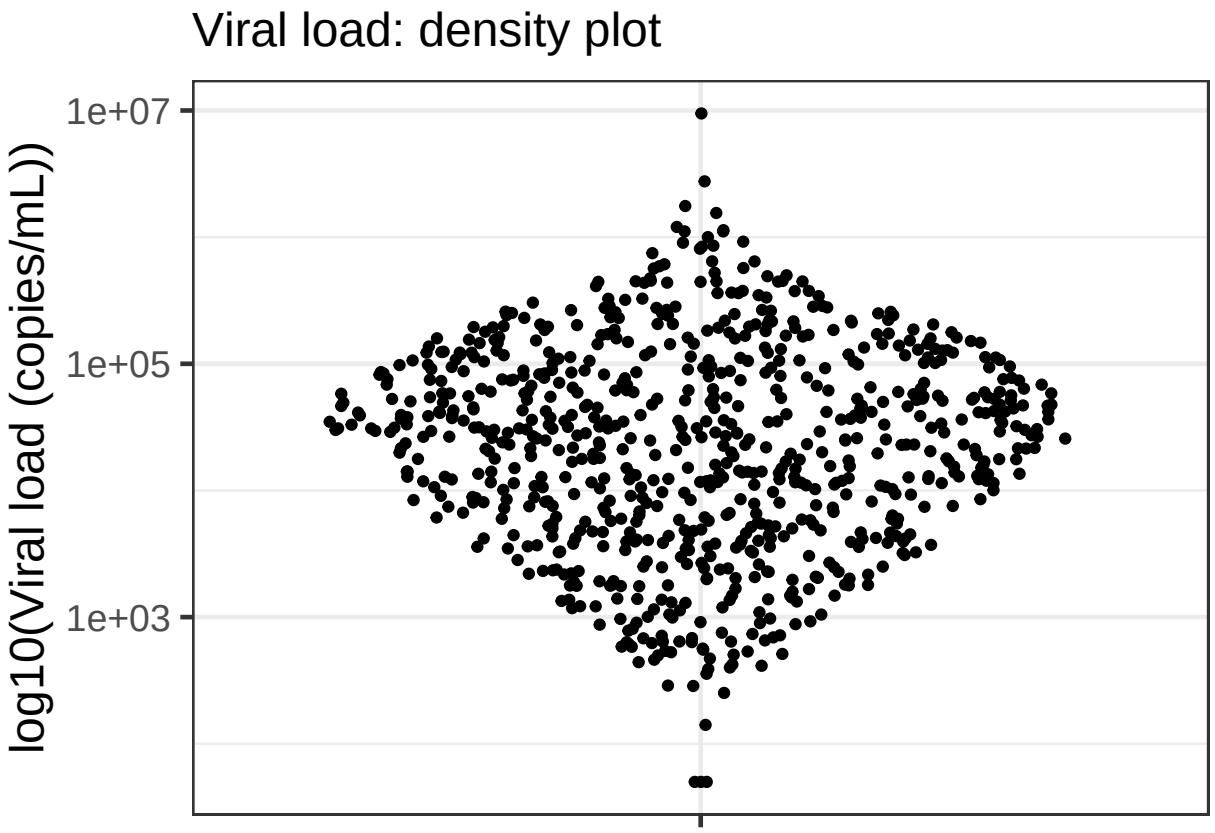
Skim summary statistics
n obs: 787
n variables: 1

Variable type: numeric

variable	missing	complete	n	mean	sd	p0	p25	p50	p75	p100
cd4_cells.ul	0	787	787	333.25	224.05	1	173.5	290	441.5	1721

Viral load

```r
# Plot
ggplot(data = df) +
    aes(x = 'Data',
        y = viral_load_cp.ml) +
    geom_sina() +
    scale_y_log10() +
    labs(subtitle = 'Viral load: density plot',
         y = 'log10(Viral load (copies/mL))') +
    theme(axis.title.x = element_blank(),
          axis.text.x = element_blank())
```




```r
# Numeric summary
df %>%
    select(viral_load_cp.ml) %>%
```

```
    skim() %>%
    skimr::kable(caption = 'Viral load: summary statistics')
```

Skim summary statistics
n obs: 787
n variables: 1

Variable type: numeric

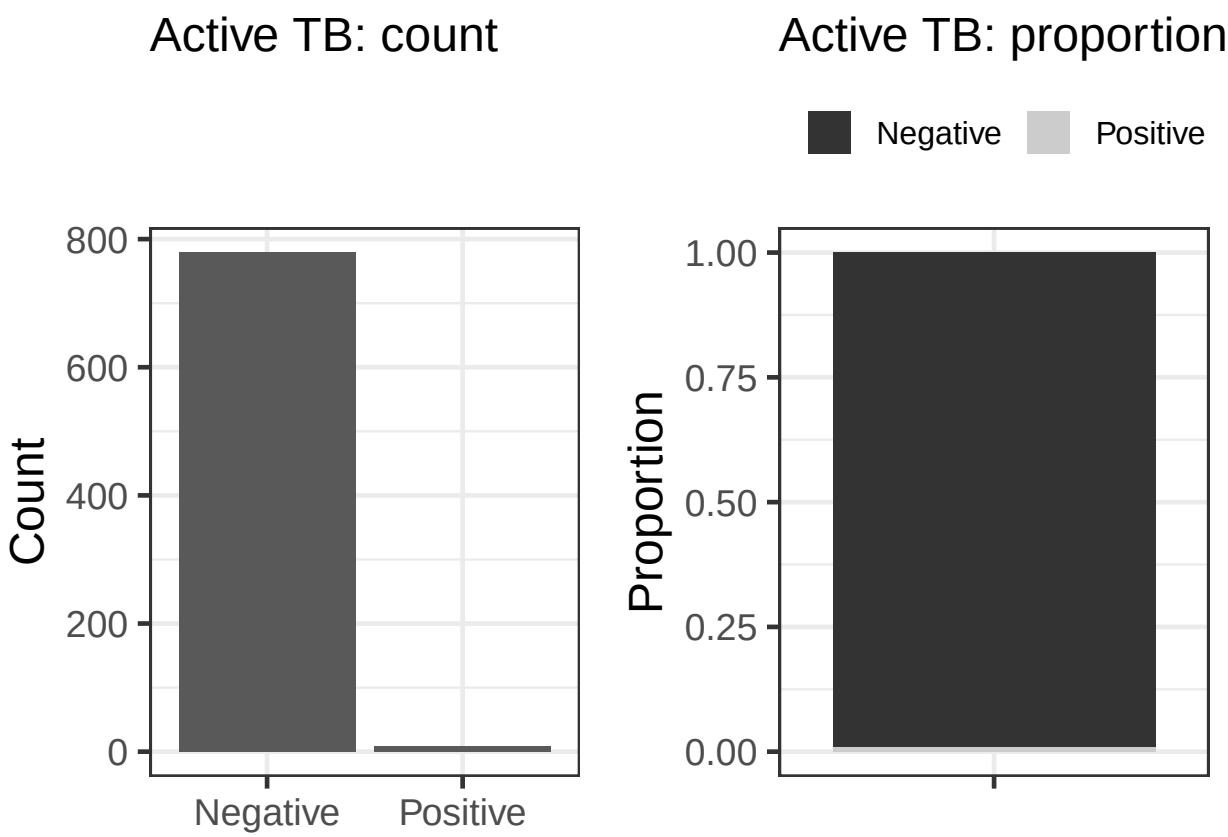
variable	missing	complete	n	mean	sd	p0	p25	p50	p75	p100
viral_load_cp.ml	0	787	787	98611.6	386719.99	50	5704.5	25853	85574	9475772

Active TB

```
# Plot
tb_count <- ggplot(data = df) +
    aes(x = tb_screen) +
    geom_bar() +
    labs(subtitle = 'Active TB: count',
         y = 'Count') +
    theme(axis.title.x = element_blank())

tb_prop <- df %>%
    group_by(tb_screen) %>%
    summarise(count = n()) %>%
    ggplot(data = .) +
    aes(x = 'Data',
        y = count,
        fill = tb_screen) +
    geom_col(position = position_fill()) +
    labs(subtitle = 'Active TB: proportion',
         y = 'Proportion') +
    scale_fill_grey() +
    theme(legend.title = element_blank(),
          legend.text = element_text(size = 12),
          legend.position = 'top',
          axis.title.x = element_blank(),
          axis.text.x = element_blank())

tb_count + tb_prop
```




```
# Numeric summary
df %>%
    select(tb_screen) %>%
    mutate(tb_screen = factor(tb_screen)) %>%
    skim() %>%
    skimr::kable(caption = 'Active TB: summary statistics')
```

Skim summary statistics
n obs: 787
n variables: 1

Variable type: factor

variable	missing	complete	n	n_unique	top_counts	ordered
tb_screen	0	787	787	2	Neg: 779, Pos: 8, NA: 0	FALSE

```
df %>%
    group_by(tb_screen) %>%
    summarise(count = n()) %>%
    mutate(n = sum(count),
           missing = sum(is.na(df$tb_screen))) %>%
    mutate(proportion = round(count / n, 3)) %>%
    select(tb_screen, count, n, missing, proportion) %>%
    knitr::kable(caption = 'Sex: summary statistics 2')
```

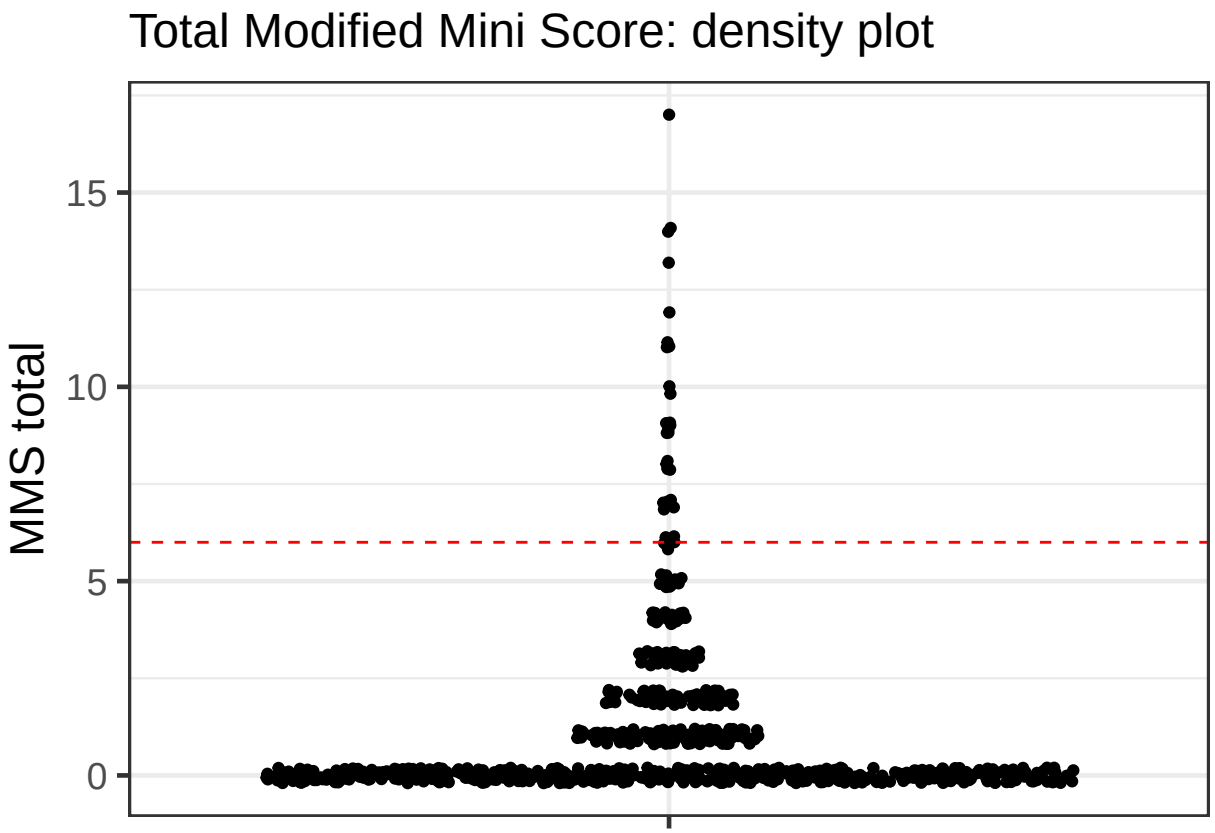
Table 19: Sex: summary statistics 2

tb_screen	count	n	missing	proportion
Negative	779	787	0	0.99
Positive	8	787	0	0.01

Modified Mini Score

```
# Plot
ggplot(data = df) +
    aes(x = 'Data',
        y = mms_total) +
    geom_sina() +
    geom_hline(yintercept = 6,
               linetype = 2,
               colour = '#FF0000') +
    labs(subtitle = 'Total Modified Mini Score: density plot',
         y = 'MMS total') +
    theme(axis.title.x = element_blank(),
          axis.text.x = element_blank())
```

```
## Warning: Removed 5 rows containing non-finite values (stat_sina).
```




```
# Numeric summary
df %>%
    select(mms_total) %>%
    skim() %>%
    skimr::kable(caption = 'Total Modified Mini Score: summary statistics')
```

Skim summary statistics
n obs: 787
n variables: 1

Variable type: numeric

variable	missing	complete	n	mean	sd	p0	p25	p50	p75	p100
mms_total	5	782	787	1.08	2.13	0	0	0	1	17

```r
# Mode
xtabs(~mms_total, data = df) %>%
    knitr::kable(caption = 'Total Modified Mini Score: modal distribution')
```

Table 21: Total Modified Mini Score: modal distribution

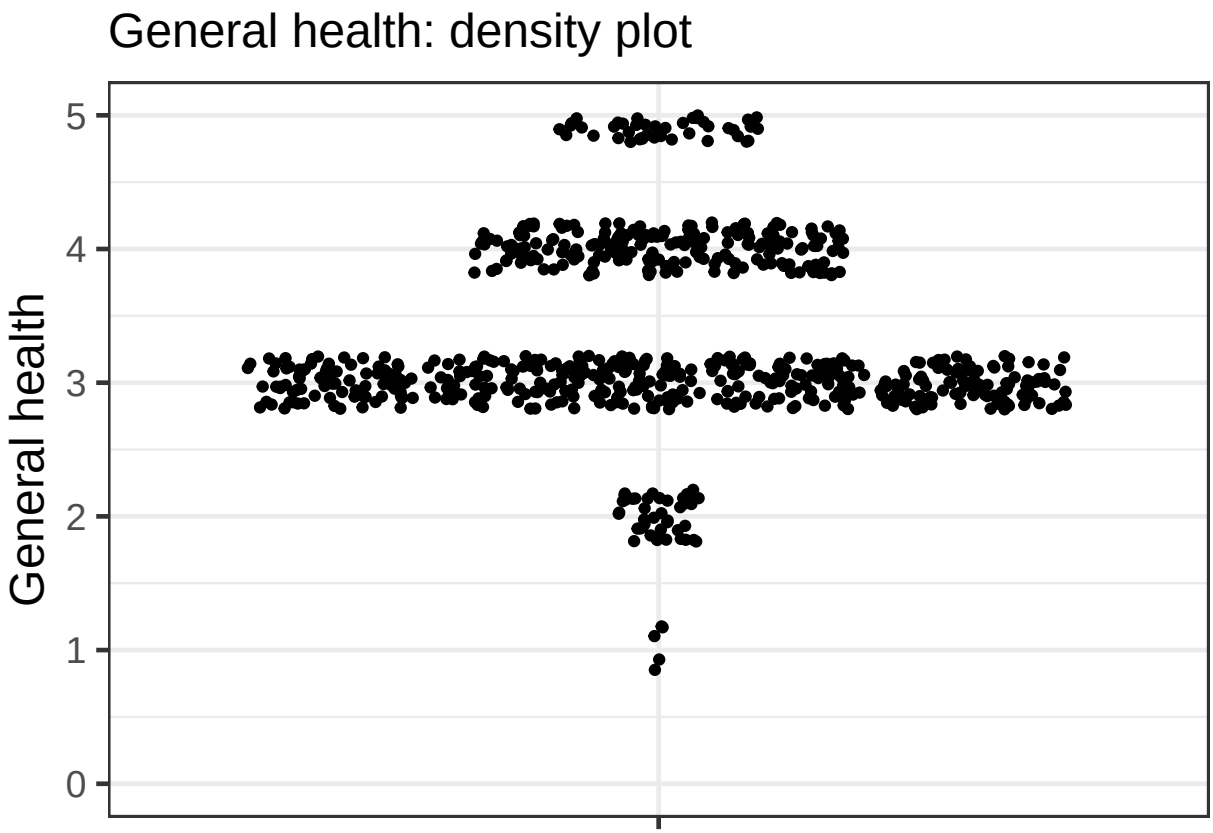
mms_total	Freq
0	487
1	111
2	77
3	36
4	20
5	15
6	8
7	7
8	4
9	7
10	2
11	3
12	1
13	1
14	2
17	1

Perception of health (baseline)

```r
# Plot
ggplot(data = df) +
    aes(x = 'Data',
        y = general_health) +
    geom_sina() +
    scale_y_continuous(limits = c(0, 5)) +
    labs(subtitle = 'General health: density plot',
         y = 'General health') +
    theme(axis.title.x = element_blank(),
          axis.text.x = element_blank())
```

```
## Warning: Removed 4 rows containing non-finite values (stat_sina).
```

```
## Warning: Removed 63 rows containing missing values (geom_point).
```




```
# Numeric summary
df %>%
    select(general_health) %>%
    skim() %>%
    skimr::kable(caption = 'General health: summary statistics')
```

Skim summary statistics
n obs: 787
n variables: 1

Variable type: numeric

variable	missing	complete	n	mean	sd	p0	p25	p50	p75	p100
general_health	4	783	787	3.45	0.82	1	3	3	4	5

```
# Mode
xtabs(~general_health, data = df) %>%
    knitr::kable(caption = 'General health: modal distribution')
```

Table 23: General health: modal distribution

general_health	Freq
1	5
2	43
3	435
4	195
5	105

Session information

```
sessionInfo()
```

```
## R version 3.6.1 (2019-07-05)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS Mojave 10.14.6
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] patchwork_0.0.1 skimr_1.0.7     ggforce_0.3.1   magrittr_1.5
##  [5] forcats_0.4.0   stringr_1.4.0   dplyr_0.8.3     purrr_0.3.3
##  [9] readr_1.3.1     tidyr_1.0.0     tibble_2.1.3    ggplot2_3.2.1
## [13] tidyverse_1.2.1
##
## loaded via a namespace (and not attached):
##  [1] tidyselect_0.2.5 xfun_0.10        haven_2.1.1      lattice_0.20-38
##  [5] colorspace_1.4-1 vctrs_0.2.0      generics_0.0.2   htmltools_0.4.0
##  [9] yaml_2.2.0       utf8_1.1.4       rlang_0.4.0      pillar_1.4.2
## [13] glue_1.3.1       withr_2.1.2      tweenr_1.0.1     modelr_0.1.5
## [17] readxl_1.3.1     lifecycle_0.1.0  munsell_0.5.0    gtable_0.3.0
## [21] cellranger_1.1.0 rvest_0.3.4      evaluate_0.14    labeling_0.3
## [25] knitr_1.25       fansi_0.4.0      highr_0.8        broom_0.5.2
## [29] Rcpp_1.0.2       scales_1.0.0     backports_1.1.5  jsonlite_1.6
## [33] farver_1.1.0     hms_0.5.1        digest_0.6.22    stringi_1.4.3
## [37] polyclip_1.10-0  grid_3.6.1       cli_1.1.0        tools_3.6.1
## [41] lazyeval_0.2.2   crayon_1.3.4     pkgconfig_2.0.3  zeallot_0.1.0
## [45] MASS_7.3-51.4    xml2_1.2.2       lubridate_1.7.4  assertthat_0.2.1
## [49] rmarkdown_1.16   httr_1.4.1       rstudioapi_0.10  R6_2.4.0
## [53] nlme_3.1-141     compiler_3.6.1
```