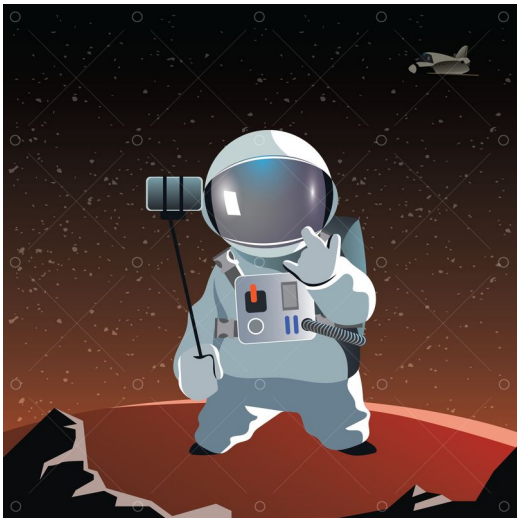
How many years do I have left to become an astronaut?

Report of Emanuel Sommer for the Nasa Hackaton of the DSA Munich

Hi, my name is Emanuel and I am a 23 year old master student of mathematics. Besides of being into data and R programming I am absolutely inspired and fascinated by space travel. I am not that guy who is into the astrophysics and aliens part but projects like *Artemis* (Moon → Mars) are so inspiring for me as they display how humanity can achieve tremendous goals through teamwork, ingenuity and innovation. And for real who hasn't dreamed about a selfie in space ;D.



First of all the R packages used:

```
library(tidyverse)
library(viridis)
library(patchwork)
library(ggtext)
```

(1) Get ready for takeoff or data importing & cleaning

Read in the data and get a first glimpse at the variables.

```
astro_data <- read_csv("astronauts.csv")
```

The next step is to detect outliers and missing data. Having this done one can decide on a strategy to deal with those.

```
# First the numeric features (get an overview of statistical key numbers here
# just for outlier detection)
astro_data %>%
  select(where(is.numeric)) %>%
  summary()
```

```
##        id           number      nationwide_number year_of_birth
##  Min.   :   1   Min.   :  1.0   Min.   :  1.0     Min.   :1921
##  1st Qu.: 320   1st Qu.:153.0   1st Qu.: 47.0     1st Qu.:1944
##  Median : 639   Median :278.0   Median :110.0     Median :1952
##  Mean   : 639   Mean   :274.2   Mean   :128.8     Mean   :1952
##  3rd Qu.: 958   3rd Qu.:390.0   3rd Qu.:204.0     3rd Qu.:1959
##  Max.   :1277   Max.   :565.0   Max.   :433.0     Max.   :1983
##  year_of_selection mission_number  total_number_of_missions year_of_mission
##  Min.   :1959      Min.   :1.000   Min.   :1.000            Min.   :1961
##  1st Qu.:1978      1st Qu.:1.000   1st Qu.:2.000            1st Qu.:1986
##  Median :1987      Median :2.000   Median :3.000            Median :1995
##  Mean   :1986      Mean   :1.992   Mean   :2.983            Mean   :1995
##  3rd Qu.:1995      3rd Qu.:3.000   3rd Qu.:4.000            3rd Qu.:2003
##  Max.   :2018      Max.   :7.000   Max.   :7.000            Max.   :2019
##  hours_mission   total_hrs_sum        field21       eva_hrs_mission
##  Min.   :    0   Min.   :    0.61   Min.   :0.0000   Min.   : 0.000
##  1st Qu.:  190   1st Qu.:  482.00   1st Qu.:0.0000   1st Qu.: 0.000
##  Median :  261   Median :  932.00   Median :0.0000   Median : 0.000
##  Mean   : 1051   Mean   : 2968.34   Mean   :0.6288   Mean   : 3.661
##  3rd Qu.:  382   3rd Qu.: 4264.00   3rd Qu.:1.0000   3rd Qu.: 4.720
##  Max.   :10505   Max.   :21083.52   Max.   :7.0000   Max.   :89.130
##  total_eva_hrs
##  Min.   : 0.00
##  1st Qu.: 0.00
##  Median : 0.00
##  Mean   :10.76
##  3rd Qu.:19.52
##  Max.   :78.80
```

First of all there are no observations with missing values present. Sanity checks for valid values are performed after one has checked the categorical features for missingness. Moreover note that the most recent mission covered here is from 2019.

```
# Check categorical data for NA's at least for the obvious ones
# (character NAs could be still in there)
astro_data %>%
  select(where(is.character)) %>%
  summarise(across(everything(), ~ sum(is.na(.)))) %>%
  as_vector()
```

```
##              name     original_name               sex       nationality
##                 0                 5                 0                 0
## military_civilian         selection        occupation     mission_title
##                 0                 1                 0                 1
##    ascend_shuttle          in_orbit   descend_shuttle
##                 1                 0                 1
```

So there are definitely some missing values. Let's have a closer look.

```
# display the missing data
astro_data %>%
  filter(if_any(everything(), ~ is.na(.))) %>%
  select(where(~ any(is.na(.))))
```

```
## # A tibble: 8 x 5
##   original_name          selection  mission_title ascend_shuttle descend_shuttle
##   <chr>                  <chr>      <chr>         <chr>          <chr>
## 1 <NA>                   Air Gorce~ Soyuz 28      Soyuz 28       Soyuz 28
## 2 Farkas Bertalan        1978 Inte~ Soyuz 36/35   <NA>           <NA>
## 3 <NA>                   Saudi-Ara~ STS-51G       STS-51G        STS-51G
## 4 <NA>                   1985 NASA~ STS-61-B      STS-61-B       STS-61-B
## 5 <NA>                   Syria      1             Soyuz TM-3     Soyuz TM-2
## 6 <NA>                   Afghanist~ 3             Soyuz TM-6     Soyuz TM-5
## 7 Olsen, Gregory Hammond <NA>       Soyuz TMA-7   Soyuz TMA-7    Soyuz TMA-7
## 8 Parmitano, Luca        2009 ESA ~ <NA>          Soyuz MS-13    not completed ~
```

```
### have a look whether name as the primary key is sufficient
# uniquely identified by name, nationality and year of birth:
astro_data %>%
  group_by(name, nationality, year_of_birth) %>%
  summarise(1, .groups = "drop") %>% nrow()
```

```
## [1] 565
```

```
# uniquely identified just by name:
length(unique(astro_data$name))
```

```
## [1] 564
```

```
# so one name is double

astro_data %>%
  group_by(name) %>%
  summarise(n_birth_years = length(unique(year_of_birth))) %>%
  filter(n_birth_years > 1)
```

```
## # A tibble: 1 x 2
##   name                   n_birth_years
##   <chr>                          <int>
## 1 Aleksandrov, Aleksandr             2
```

```
astro_data %>%
  filter(name == "Aleksandrov, Aleksandr") %>%
  select(id, name, nationality, year_of_birth)
```

```
## # A tibble: 3 x 4
##      id name                   nationality    year_of_birth
##   <dbl> <chr>                  <chr>                  <dbl>
## 1   248 Aleksandrov, Aleksandr U.S.S.R/Russia          1943
## 2   249 Aleksandrov, Aleksandr U.S.S.R/Russia          1943
## 3   447 Aleksandrov, Aleksandr Bulgaria                1951
```

```
# this is actually once the Bulgarian and once the Russian, as Wikipedia
# writes the Bulgarian one as Alexandar we can change this to have name as
# a unique key
astro_data$name[astro_data$id == 447] <- "Aleksandrov, Alexandar"
```

From this one can conclude that now one can use the `name` variable as the main key for a single astronaut as there are no missing values. Down below I actually showed that two astronauts had the same name As the only missing values occur in the variables `original_name`, once in the `selection` and once in the `ascent_shuttle` and `descent_shuttle` for the same mission in the latter cases one does not have to remove them in this particular case. This is because these variables will not play a crucial role in the further analysis.

As shown below mostly the categorical features have quite many classes and thus are not easily checkable manually, for variables with less then 20 classes this is manageable.

```
# Have a look at how many unique values the variables have
astro_data %>%
  select(where(is.character)) %>%
  mutate(across(everything(), as_factor)) %>%
  summarise(across(everything(), ~ length(unique(.)))) %>%
  as_vector()
```

```
##              name     original_name               sex       nationality
##               565               561                 2                40
## military_civilian         selection        occupation     mission_title
##                 2               230                12               362
##    ascend_shuttle          in_orbit   descend_shuttle
##               437               289               433
```

```
# the classes of vars with less than 20 unique classes
astro_data %>%
  select(where(is.character)) %>%
  mutate(across(everything(), as_factor)) %>%
  select(where(~ length(unique(.)) < 20)) %>%
  pivot_longer(everything(), names_to = "variable",
               values_to = "unique_classes") %>%
  group_by(variable, unique_classes) %>%
  summarise(n_obs = n(), .groups = "keep") %>%
  arrange(variable)
```

```
## # A tibble: 16 x 3
## # Groups:   variable, unique_classes [16]
```

```
##    variable          unique_classes          n_obs
##    <chr>             <fct>                   <int>
##  1 military_civilian military                  769
##  2 military_civilian civilian                  508
##  3 occupation        pilot                     196
##  4 occupation        PSP                        59
##  5 occupation        Pilot                       1
##  6 occupation        commander                 315
##  7 occupation        MSP                       498
##  8 occupation        flight engineer           192
##  9 occupation        Other (Journalist)          1
## 10 occupation        Flight engineer             4
## 11 occupation        Other (space tourist)       5
## 12 occupation        Other (Space tourist)       3
## 13 occupation        Space tourist               2
## 14 occupation        spaceflight participant     1
## 15 sex               male                     1134
## 16 sex               female                    143
```

For the variables `military_civilian` and `sex` the results are not too surprising, but the `occupation` variable shows some data quality issues! For example Flight engineer is covered twice with the only difference being the case of the first letter, same goes for pilot. The occupations PSP and MSP were not clear to me and after a bit of searching I found out that these are the Payload Specialist and the Mission Specialist roles. Regarding all classes that cover somewhat space tourists I will group them all into one category called 'Space tourist'. Funnily enough on Wikipedia (https://en.wikipedia.org/wiki/Astronaut) it says that as of 2020 nobody has even qualified for the status of space tourist and one could read the full taxonomy in order to correctly assign each of the space travelers the actual correct class but I will omit this here for simplicity reasons.

```r
# clean the data according to results
astro_data <- astro_data %>%
  mutate(occupation = case_when(
    occupation == "pilot" ~ "Pilot",
    occupation == "flight engineer" ~ "Flight engineer",
    occupation == "PSP" ~ "Payload Specialist",
    occupation == "MSP" ~ "Mission Specialist",
    occupation == "Other (Journalist)" ~ "Space tourist",
    occupation == "Other (space tourist)" ~ "Space tourist",
    occupation == "Other (Space tourist)" ~ "Space tourist",
    occupation == "spaceflight participant" ~ "Space tourist",
    # for consistent upper case:
    occupation == "commander" ~ "Commander",
    TRUE ~ occupation
  ))

unique(astro_data$occupation)
```

```
## [1] "Pilot"              "Payload Specialist" "Commander"
## [4] "Mission Specialist" "Flight engineer"    "Space tourist"
```

Now one can formulate and perform some basic **sanity checks**:

1. `id` should be unique.

2. `year_of_birth` $\leq$ `year_of_selection`
3. `year_of_selection` $\leq$ `year_of_mission`
4. Check some suspiciously high values from the summary above. (not the ones covered in 5-6.)
5. `hours_mission` $\leq$ `total_hrs_sum` and the sum should be the total + the max is suspiciously high.
6. `eva_hrs_mission` $\leq$ `total_eva_hrs` and the sum should be the total + this is definitely violated as can be seen from the max. So there are definitely data quality issues.
7. All astronauts should have the same number, year of selection and birth in each row.
8. `mission_number` $\leq$ `total_number_of_missions`
9. Every astronaut has to have a mission number 1.

```
### 1. ----------------
length(unique(astro_data$id)) == nrow(astro_data)
```

```
## [1] TRUE
```

```
### 2.  ----------------
all(astro_data$year_of_birth <= astro_data$year_of_selection)
```

```
## [1] TRUE
```

```
### 3.  ----------------
all(astro_data$year_of_selection <= astro_data$year_of_mission)
```

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

```
astro_data %>%
  filter(year_of_selection > year_of_mission) %>%
  select(id, name, year_of_selection, year_of_mission)
```

```
## # A tibble: 2 x 4
##      id name                 year_of_selection year_of_mission
##   <dbl> <chr>                            <dbl>           <dbl>
## 1   648 Franco Malerba                    1998            1992
## 2   862 Thomas, Andrew S. W.              1992            1983
```

```
# Franco Malerba was actually selected 1977 and not 1998 year of the
# mission is correct
# Thomas, Andrew S. W. actually had his first spaceflight in 1996
# (STS-77 as correctly written)

# the rest of Andrew is ok
astro_data %>%
  filter(name == "Thomas, Andrew S. W.") %>%
  select(id, name, year_of_selection, year_of_mission)
```

```
## # A tibble: 4 x 4
##      id name                 year_of_selection year_of_mission
##   <dbl> <chr>                            <dbl>           <dbl>
## 1   862 Thomas, Andrew S. W.              1992            1983
## 2   863 Thomas, Andrew S. W.              1992            1998
## 3   864 Thomas, Andrew S. W.              1992            2001
## 4   865 Thomas, Andrew S. W.              1992            2005
```

```r
# correct the errors
astro_data$year_of_selection[astro_data$id == 648] <- 1977
astro_data$year_of_mission[astro_data$id == 862] <- 1996

### 4. -----------------
# high max nationwide number
astro_data %>%
  arrange(desc(nationwide_number)) %>%
  select(id, name, nationwide_number, number) %>%
  head(3)
```

```
## # A tibble: 3 x 4
##      id name           nationwide_number number
##   <dbl> <chr>                      <dbl>  <dbl>
## 1  1271 Hague, Tyler                 433    559
## 2  1276 Meir, Jessica                348    564
## 3  1275 Morgan, Andrew               347    563
```

```r
astro_data %>%
  filter(name == "Hague, Tyler") %>%
  select(id, name, nationwide_number, number)
```

```
## # A tibble: 2 x 4
##      id name         nationwide_number number
##   <dbl> <chr>                    <dbl>  <dbl>
## 1  1270 Hague, Tyler               344    559
## 2  1271 Hague, Tyler               433    559
```

```r
# the nationwide number of Hague, Tyler is indeed not correct and thus
# will be adjusted
astro_data$nationwide_number[astro_data$id == 1271] <- 344

# high year of selection
astro_data %>%
  arrange(desc(year_of_selection)) %>%
  select(id, name, year_of_selection, year_of_mission) %>%
  head(3)
```

```
## # A tibble: 3 x 4
##      id name                year_of_selection year_of_mission
##   <dbl> <chr>                           <dbl>           <dbl>
## 1  1277 Al Mansoori, Hazzaa              2018            2019
## 2  1239 Zhang, Xiaoguang                 2013            2013
## 3  1240 Wang, Yapi                       2013            2013
```

```r
# this is correct
```

```r
### 5. -----------------
# first a look at the biggest values
astro_data %>%
  arrange(desc(hours_mission)) %>%
  select(id, name, hours_mission) %>%
  head(5)
```

```
## # A tibble: 5 x 3
##      id name                 hours_mission
##   <dbl> <chr>                        <dbl>
## 1   449 Polyakov, Valeri            10505
## 2  1163 Kimbrough, Robert S.        10383.
## 3  1222 Borisenko, Andrei           10383.
## 4  1262 Ryzhikov, Sergey            10383.
## 5   638 Avdeyev, Sergei              9110
```

```
astro_data %>%
  arrange(desc(total_hrs_sum)) %>%
  distinct(name, .keep_all = TRUE) %>%
  select(id, name, total_hrs_sum) %>%
  head(3)
```

```
## # A tibble: 3 x 3
##      id name              total_hrs_sum
##   <dbl> <chr>                     <dbl>
## 1   942 Padalka, Gennady         21084.
## 2   451 Krikalev, Sergei         19282.
## 3   609 Kaleri, Aleksandr        18463.
```

```
# the max values are correct

# now check the sum over missions (just roughly)
astro_data %>%
  group_by(name) %>%
  summarise(actual_total_hrs = sum(hours_mission),
            total_hrs_sum = total_hrs_sum,
            .groups = "drop") %>%
  # roughly the same +- 2 hour the actual
  mutate(diff_total_hrs = abs(actual_total_hrs - total_hrs_sum)) %>%
  filter(diff_total_hrs > 2) %>%
  arrange(desc(diff_total_hrs)) %>%
  distinct(name, .keep_all = TRUE) %>%
  head(10)
```

```
## # A tibble: 10 x 4
##    name                     actual_total_hrs total_hrs_sum diff_total_hrs
##    <chr>                               <dbl>         <dbl>          <dbl>
##  1 Pesquet, Thomas                    4722.         15105.         10383.
##  2 Arnold, Richard R., II             5029            307            4722
##  3 Feustel, Andrew J.                 5409            687            4722
##  4 Crouch, Roger Keith                 473.          4576.          4104.
##  5 Linteris, Gregory Thomas            473.          4576.          4104.
##  6 Coleman, Catherine G.              4324            500            3824
##  7 Arnaldo Tamayo Mendez              1888.           189.          1699
##  8 Henricks, Terence T.               1024            191.           833.
##  9 Weitz, Paul J.                      387            793.           406.
## 10 Cassidy, Christopher J.            4367           4744            377
```

Sanity check 5. showed that actually the column `total_hrs_sum` has serious quality issues! I checked for the first 4 largest differences between calculated total time spend in space and the given variable

`total_hrs_sum` and every time the `total_hrs_sum` variable was wrong and the newly calculated column was spot on right. Thus as there at least ~80 rows with at least some issue I will proceed by removing `total_hrs_sum` from the data set and add the calculated `calc_total_hrs` column instead, which corresponds to the `actual_total_hrs` above.

```
astro_data <- astro_data %>%
  select(-total_hrs_sum) %>%
  group_by(name) %>%
  mutate(calc_total_hrs = sum(hours_mission)) %>%
  ungroup()
```

```
### 6. ----------------
# the highest eva_hrs_mission is for sure wrong so have a look
astro_data %>%
  arrange(desc(eva_hrs_mission)) %>%
  select(id, name, eva_hrs_mission, total_eva_hrs) %>%
  head(5)
```

```
## # A tibble: 5 x 4
##      id name              eva_hrs_mission total_eva_hrs
##   <dbl> <chr>                       <dbl>         <dbl>
## 1   446 Solovyev, Anatoly            89.1          78.8
## 2  1275 Morgan, Andrew               39.5          39.5
## 3  1033 Whitson, Peggy A.            35.4          60.3
## 4  1015 Tani, Daniel M.              35.0          39.2
## 5  1021 Walheim, Rex J.              34.6          56.7
```

```
# Solovyev, Anatoly indeed has the biggest total eva time of correctly 78 hours
# but thus the 89.13 is not one of his 16 spacewalks but instead ...
solovey_446_spacewalk <- astro_data %>%
  filter(name == "Solovyev, Anatoly") %>%
  filter(eva_hrs_mission < 80) %>%
  summarise(total_eva_hrs - sum(eva_hrs_mission)) %>%
  distinct() %>%
  as.numeric()
solovey_446_spacewalk
```

```
## [1] 35.14
```

```
# correct it
astro_data$eva_hrs_mission[astro_data$id == 446] <- solovey_446_spacewalk

# now again check the sum over missions (just roughly)
astro_data %>%
  group_by(name) %>%
  summarise(actual_total_eva = sum(eva_hrs_mission),
            total_eva_hrs = total_eva_hrs,
            .groups = "drop") %>%
  # roughly the same +- 2 hour the actual
  mutate(diff_total_eva = abs(actual_total_eva - total_eva_hrs)) %>%
  filter(diff_total_eva > 2) %>%
  arrange(desc(diff_total_eva)) %>%
  distinct(.keep_all = TRUE)
```

```
## # A tibble: 2 x 4
##   name              actual_total_eva total_eva_hrs diff_total_eva
##   <chr>                        <dbl>         <dbl>          <dbl>
## 1 Hague, Tyler                  39.9          19.9           19.9
## 2 Leestma, David C.              1             3.5            2.5
```

```
astro_data %>%
  filter(name == "Hague, Tyler") %>%
  select(id, name, eva_hrs_mission, total_eva_hrs, in_orbit)
```

```
## # A tibble: 2 x 5
##      id name         eva_hrs_mission total_eva_hrs in_orbit
##   <dbl> <chr>                  <dbl>         <dbl> <chr>
## 1  1270 Hague, Tyler            19.9          19.9 aborted
## 2  1271 Hague, Tyler            19.9          19.9 ISS
```

```
# so wrongly Hague, Tyler got his eva hours also for the aborted flight
# this is of course wrong and should be corrected

astro_data$eva_hrs_mission[astro_data$id == 1270] <- 0


astro_data %>%
  filter(name == "Leestma, David C.") %>%
  select(id, name, eva_hrs_mission, total_eva_hrs, in_orbit)
```

```
## # A tibble: 3 x 5
##      id name              eva_hrs_mission total_eva_hrs in_orbit
##   <dbl> <chr>                       <dbl>         <dbl> <chr>
## 1   316 Leestma, David C.               1           3.5 STS-41-G
## 2   317 Leestma, David C.               0           3.5 STS-28
## 3   318 Leestma, David C.               0           3.5 STS-45
```

```
# here the eva_hrs_mission of the sts-41-G is wrongly set to 1 but it were
# 3.5 hours as correctly stated in the total_eva_hrs column
astro_data$eva_hrs_mission[astro_data$id == 316] <- 3.5
```

```
### 7. ----------------
astro_data %>%
  group_by(name) %>%
  summarise(n_number = length(unique(number)),
            n_nat_number = length(unique(nationwide_number)),
            n_selection = length(unique(year_of_selection)),
            n_birth = length(unique(year_of_birth))) %>%
  filter(n_number > 1 | n_nat_number > 1 |
           n_selection > 1 | n_birth > 1) %>%
  nrow() == 0
```

```
## [1] TRUE
```

```
### 8. ----------------
all(astro_data$mission_number <= astro_data$total_number_of_missions)
```

```
## [1] TRUE
```

```
### 9. ----------------
astro_data %>%
  group_by(name) %>%
  summarise(valid_mission_number = all(sort(mission_number) == seq(n()))) %>%
  filter(!valid_mission_number)
```

```
## # A tibble: 1 x 2
##   name                  valid_mission_number
##   <chr>                 <lgl>
## 1 Shepard, Alan B., Jr. FALSE
```

```
astro_data %>%
  filter(str_detect(name, "Shepard")) %>%
  select(id, name, mission_number, total_number_of_missions, mission_title, year_of_mission)
```

```
## # A tibble: 1 x 6
##      id name      mission_number total_number_of_~ mission_title year_of_mission
##   <dbl> <chr>              <dbl>             <dbl> <chr>                   <dbl>
## 1   117 Shepard,~              2                 2 Apollo 14                1971
```

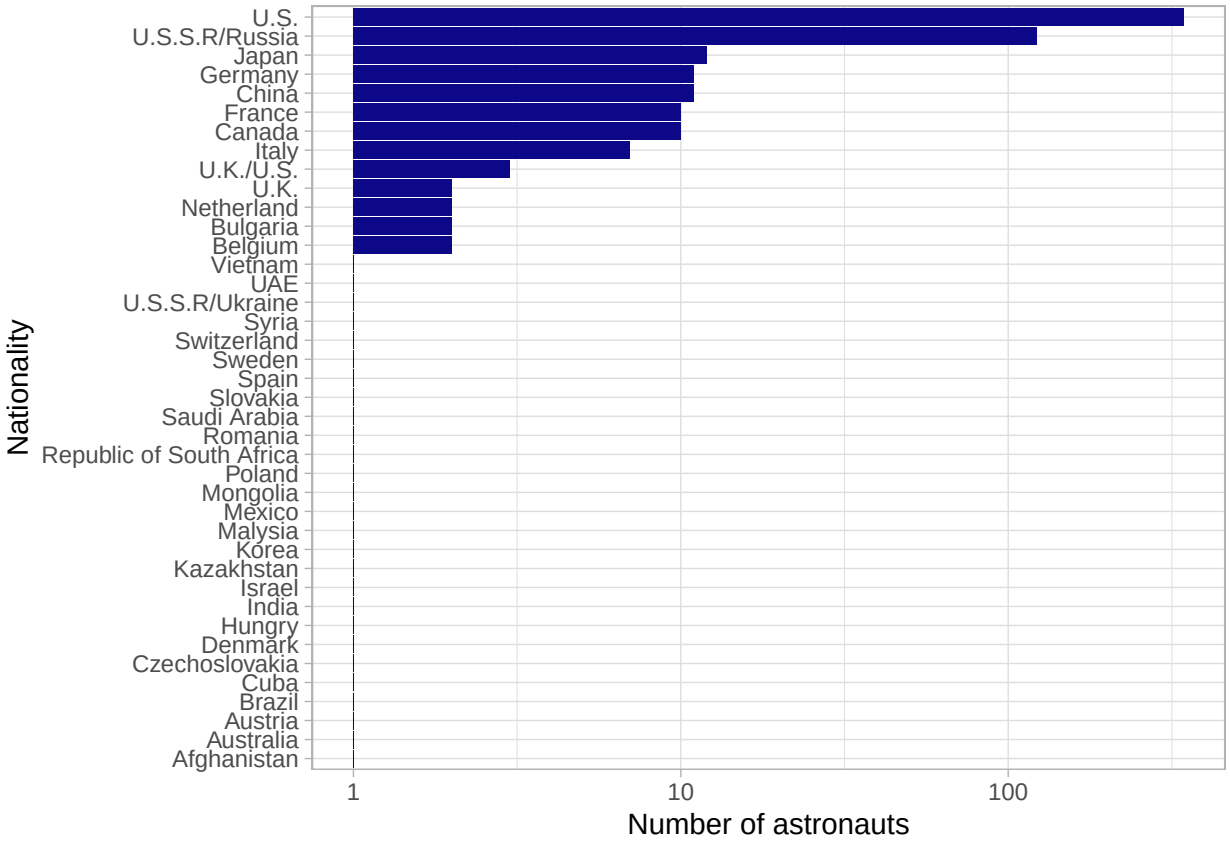
```
# So actually the first mission of Alan Shepard with the Mercury-Redstone 3
# is not in the data set maybe as it was only a suborbital flight but then
# the mission number should be 1. From the two ways of proceeding i.e.
# setting the mission number to 1 or adding an additional row for the MR-3
# launch I will pick the first option. Either way this decision won't impact the
# further analysis a lot
astro_data$mission_number[astro_data$id == 117] <- 1
astro_data$total_number_of_missions[astro_data$id == 117] <- 1
```

Now as the data set can pass the specified sanity checks it is time for some feature engineering i.e. add interesting additional variables for further analysis to the data set.

1. The age of the astronaut when selected: `age_selected`
2. The duration from selection to the first mission: `train_time`
3. The decade an astronaut was selected: `decade_sel`
4. As can be seen in the below visualization Russia and the US have by far the most space-travelers and thus the new column `nationality_red` covers just the three levels 'U.S.', 'U.S.S.R./Russia' and 'Rest of the world'.

```
astro_data %>%
  filter(mission_number == 1) %>%
  group_by(nationality) %>%
  summarise(n = n()) %>%
  ungroup() %>%
  mutate(nationality = as_factor(nationality),
```

```
        nationality = fct_reorder(nationality, n)) %>%
  ggplot(aes(x = nationality, y = n)) +
  geom_col(fill = plasma(1)) +
  coord_flip() +
  labs(y = "Number of astronauts", x = "Nationality") +
  scale_y_log10() +
  theme_light()
```




```
astro_data <- astro_data %>%
  mutate(age_selected = year_of_selection - year_of_birth) %>%
  group_by(name) %>%
  mutate(train_time = min(year_of_mission) - year_of_selection) %>%
  ungroup() %>%
  mutate(decade_sel = 10 * (year_of_selection %/% 10),
         nationality_red = case_when(
           nationality %in% c("U.S.", "U.S.S.R/Russia") ~ nationality,
           TRUE ~ "Rest of the world"
         ))
```

As the current data set has possibly multiple rows corresponding to the missions of an astronaut I create also a data set with just one row per astronaut that contains the most important facts around the individual. That means of course that no mission specific data is contained in the new `astro_data_ind` data set. As the further analysis will focus on the individual level and not the mission level one saves a lot of `group_by()` calls.

```r
astro_data_ind <- astro_data %>%
  filter(mission_number == 1) %>%
  select(-mission_number, -year_of_mission, -mission_title,
         -ascend_shuttle, -in_orbit, -descend_shuttle,
         -hours_mission, -field21, -eva_hrs_mission)
```

(2) Takeoff or statistical key numbers

As I find the new variables `age_selected` and `train_time` very interesting they will be at the core of the further analysis. Thus a close look at the univariate distributions and key statistical features is performed now.

Selection age

```
summary(astro_data_ind$age_selected)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   23.00   31.00   34.00   34.29   37.00   60.00
```

One can see that overall the IQR of 6 is quite small. So most of the astronauts are selected in their early to mid thirties. The fact that the mean and median are quite close to each other indicates that the empirical distribution could be symmetric. The youngest astronaut to be selected was 23 years and the oldest 60 years old. Let's get to know the youngest and oldest to be selected:

```
# the youngest
astro_data_ind %>%
  arrange(age_selected) %>%
  select(name, age_selected, occupation, year_of_selection, nationality) %>%
  head(4)
```

```
## # A tibble: 4 x 5
##   name               age_selected occupation year_of_selection nationality
##   <chr>                     <dbl> <chr>                   <dbl> <chr>
## 1 Klimuk, Pyotr                23 Commander                1965 U.S.S.R/Russia
## 2 Sarafanov, Gennadi           23 Commander                1965 U.S.S.R/Russia
## 3 Zudov, Vyacheslav            23 Commander                1965 U.S.S.R/Russia
## 4 Kizim, Leonid                24 Commander                1965 U.S.S.R/Russia
```

Clearly the youngest selected astronauts were all selected in the midst of the space race in the former U.S.S.R.

```
# the oldest
astro_data_ind %>%
  arrange(age_selected) %>%
  select(name, age_selected, occupation, year_of_selection, nationality) %>%
  tail(4)
```

```
## # A tibble: 4 x 5
##   name                        age_selected occupation year_of_selecti~ nationality
##   <chr>                              <dbl> <chr>                 <dbl> <chr>
## 1 Crouch, Roger Keith                   55 Payload S~             1995 U.S.
## 2 Simonyi, Charles (Karoly)             58 Space tou~             2006 U.S.
## 3 Olsen, Gregory Hammond                59 Space tou~             2004 U.S.
## 4 Tito, Dennis Anthony                  60 Space tou~             2000 U.S.
```

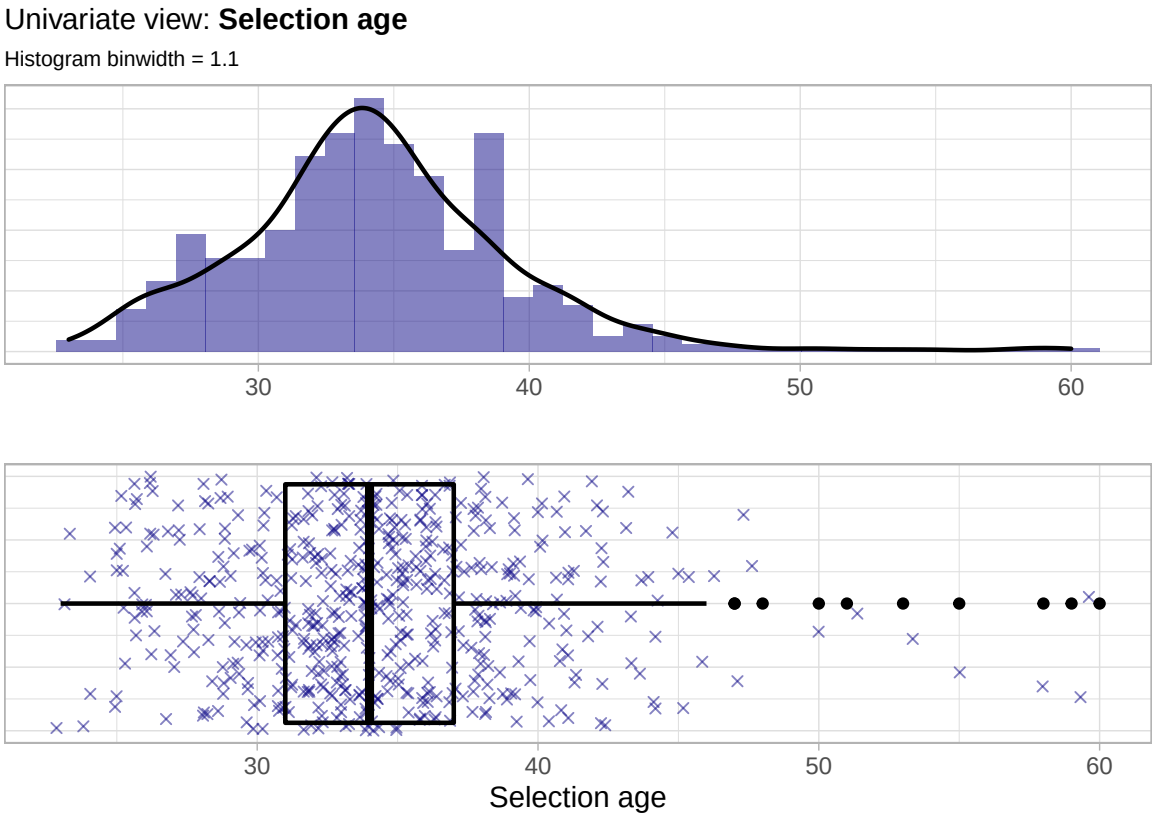
The oldest selected astronauts are all space tourist but actually followed quite closely by a professional astronaut. The fact that space tourists are generally quite old when selected comes probably from the hefty

price tag a trip to space as a tourist still has. Basically right now it is billionaires only. If Virgin Galactic and Blue Origin can keep their promises to make space travel more affordable in the upcoming years the average age of at least the commercial astronauts could decrease.

In order to get a really good sense of an univariate distribution I like to combine some different visualization techniques that let me grab not only the overall silhouette of the distribution (e.g. with KDE), but also statistical key numbers (e.g. through a boxplot) and most importantly also subtle details like discrete clusters of data points (e.g. suitable histogram). Moreover a jittered pointcloud can mitigate the general problem of boxplots that they could be shaped strongly by small point clusters especially in low sample size scenarios.

```
age_selected_box <- ggplot(astro_data_ind, aes(x = 1, y = age_selected)) +
  geom_jitter(alpha = 0.5, col = plasma(1), shape = 4) +
  geom_boxplot(col = "black", size = .8, fill = NA) +
  labs(x = "", y = "Selection age") +
  coord_flip() +
  theme_light() +
  theme(axis.ticks.y = element_blank(),
        axis.text.y = element_blank())

age_selected_hist <- ggplot(astro_data_ind, aes(x = age_selected)) +
  geom_histogram(aes(y = ..density..),
                 fill = plasma(1), binwidth = 1.1,
                 alpha = 0.5) +
  geom_density(aes(y = ..density..),
                 col = "black", size = 0.8) +
  labs(x = "", y = "", subtitle = "Histogram binwidth = 1.1",
       title = "Univariate view:  **Selection age**") +
  theme_light() +
  theme(axis.ticks.y = element_blank(),
        axis.text.y = element_blank(),
        plot.title = ggtext::element_markdown(size = 11),
        plot.subtitle = ggtext::element_markdown(size = 8))
age_selected_hist / age_selected_box
```



Except for the outliers that produce an overall just slightly right skewed empirical distribution the empirical distribution is actually quite symmetric and bell shaped. As mentioned above most of the astronauts were selected in their early to mid thirties.

> **Hypothesis 1:** The selection age has changed over time and there are certain subgroups of astrounauts e.g. w.r.t. sex or occupation that show a different patterns regarding the selection age.

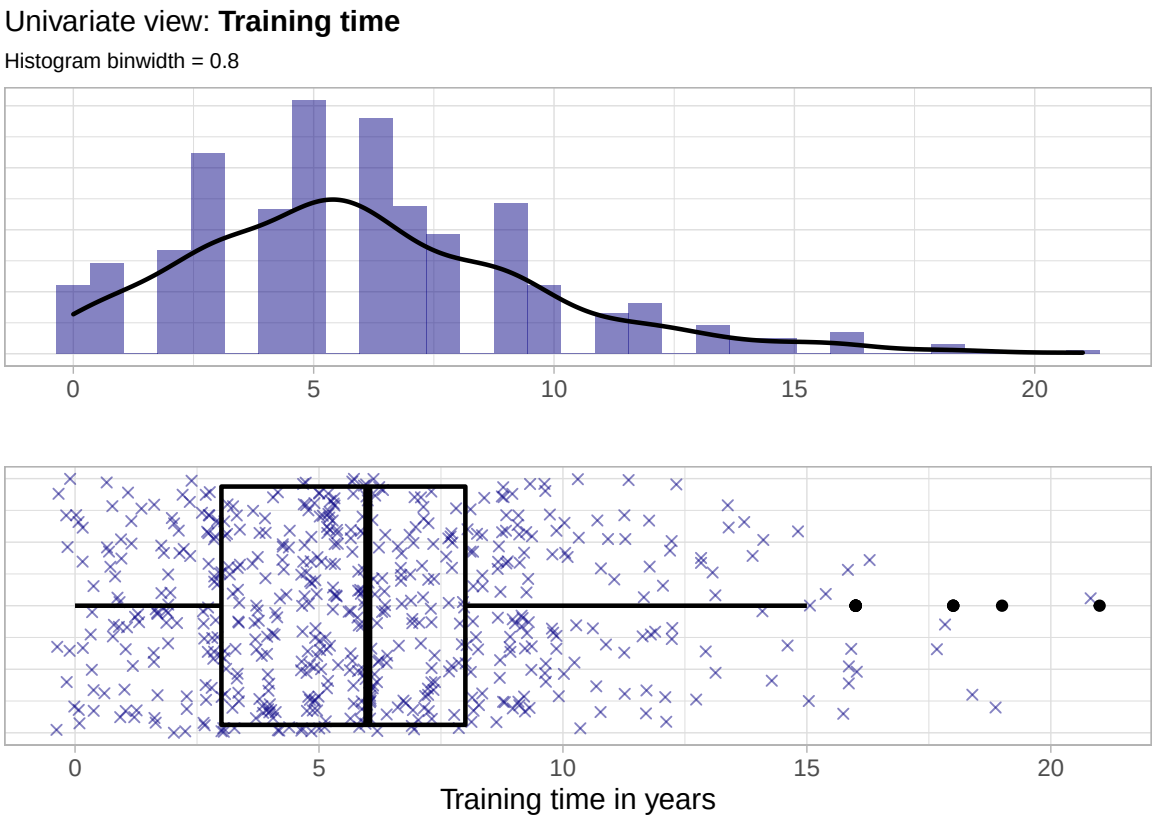
Training time

```
summary(astro_data_ind$train_time)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   3.000   6.000   6.042   8.000  21.000
```

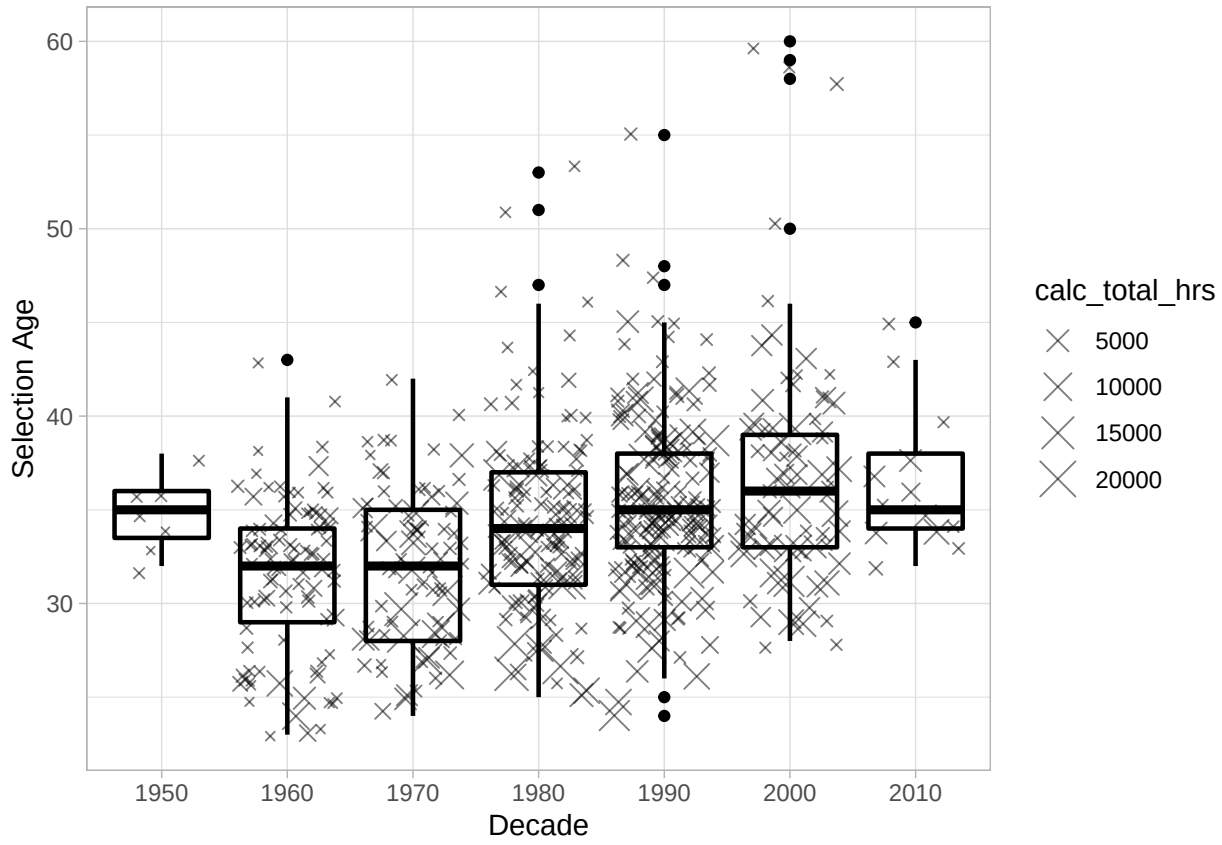
```
train_time_box <- ggplot(astro_data_ind, aes(x = 1, y = train_time)) +
  geom_jitter(alpha = 0.5, col = plasma(1), shape = 4) +
  geom_boxplot(col = "black", size = .8, fill = NA) +
  labs(x = "", y = "Training time in years") +
  coord_flip() +
  theme_light() +
  theme(axis.ticks.y = element_blank(),
        axis.text.y = element_blank())
```

```
train_time_hist <- ggplot(astro_data_ind, aes(x = train_time)) +
  geom_histogram(aes(y = ..density..),
                 fill = plasma(1), binwidth = 0.7,
                 alpha = 0.5) +
  geom_density(aes(y = ..density..),
                 col = "black", size = 0.8) +
  labs(x = "", y = "", subtitle = "Histogram binwidth = 0.8",
       title = "Univariate view:  **Training time**") +
  theme_light() +
  theme(axis.ticks.y = element_blank(),
        axis.text.y = element_blank(),
        plot.title = ggtext::element_markdown(size = 11),
        plot.subtitle = ggtext::element_markdown(size = 8))
train_time_hist / train_time_box
```

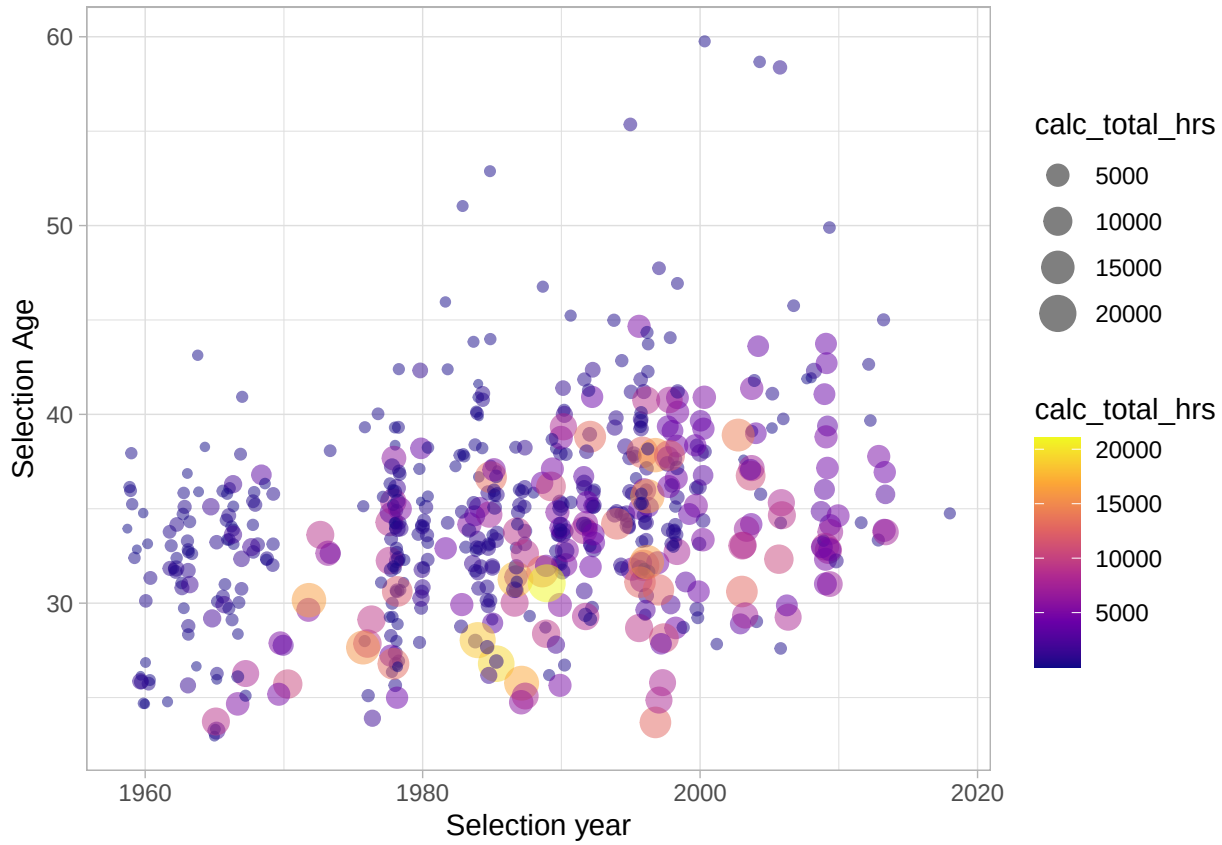


(3) Apogee & splashdown or visualization & conclusion

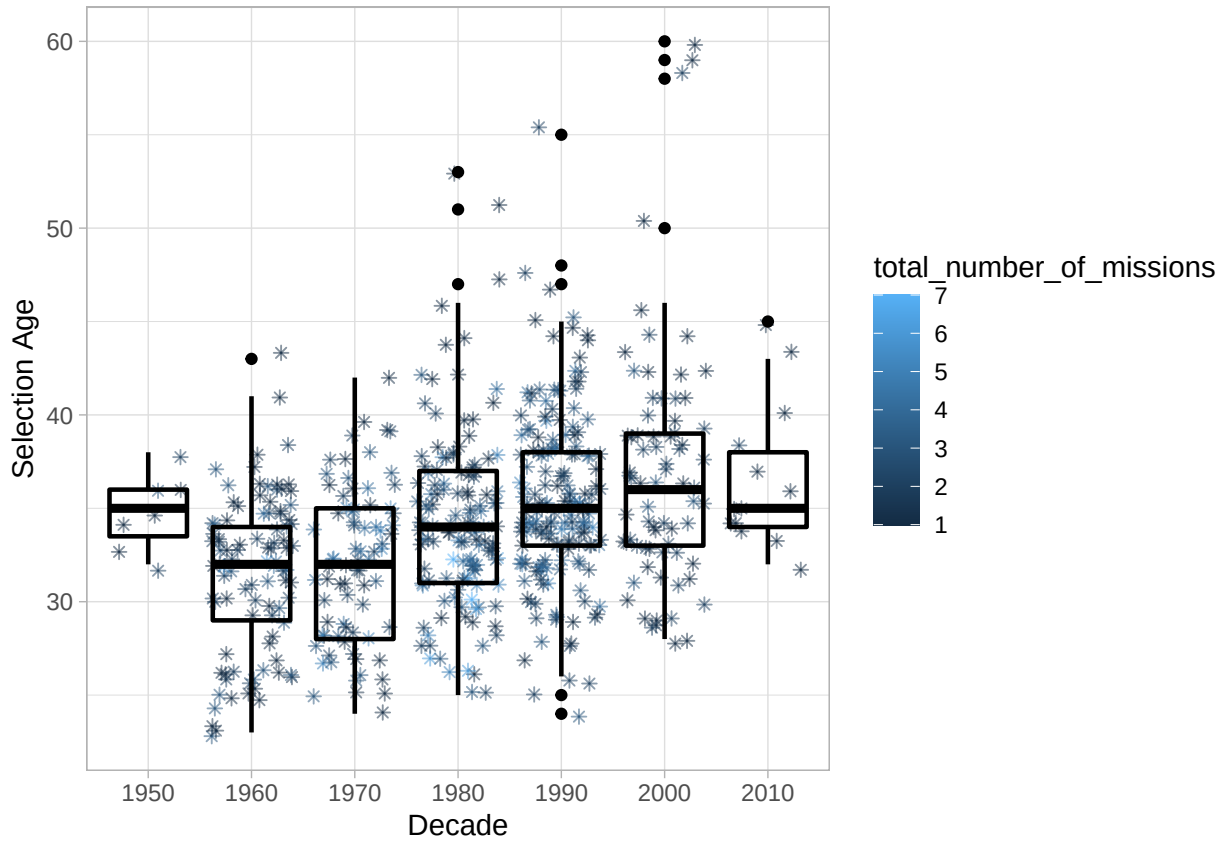
```
ggplot(astro_data_ind, aes(x = factor(decade_sel), y = age_selected)) +
  geom_jitter(aes(size = calc_total_hrs), alpha = 0.5, shape = 4) +
  geom_boxplot(col = "black", size = .8, fill = NA) +
  labs(x = "Decade", y = "Selection Age") +
  theme_light()
```



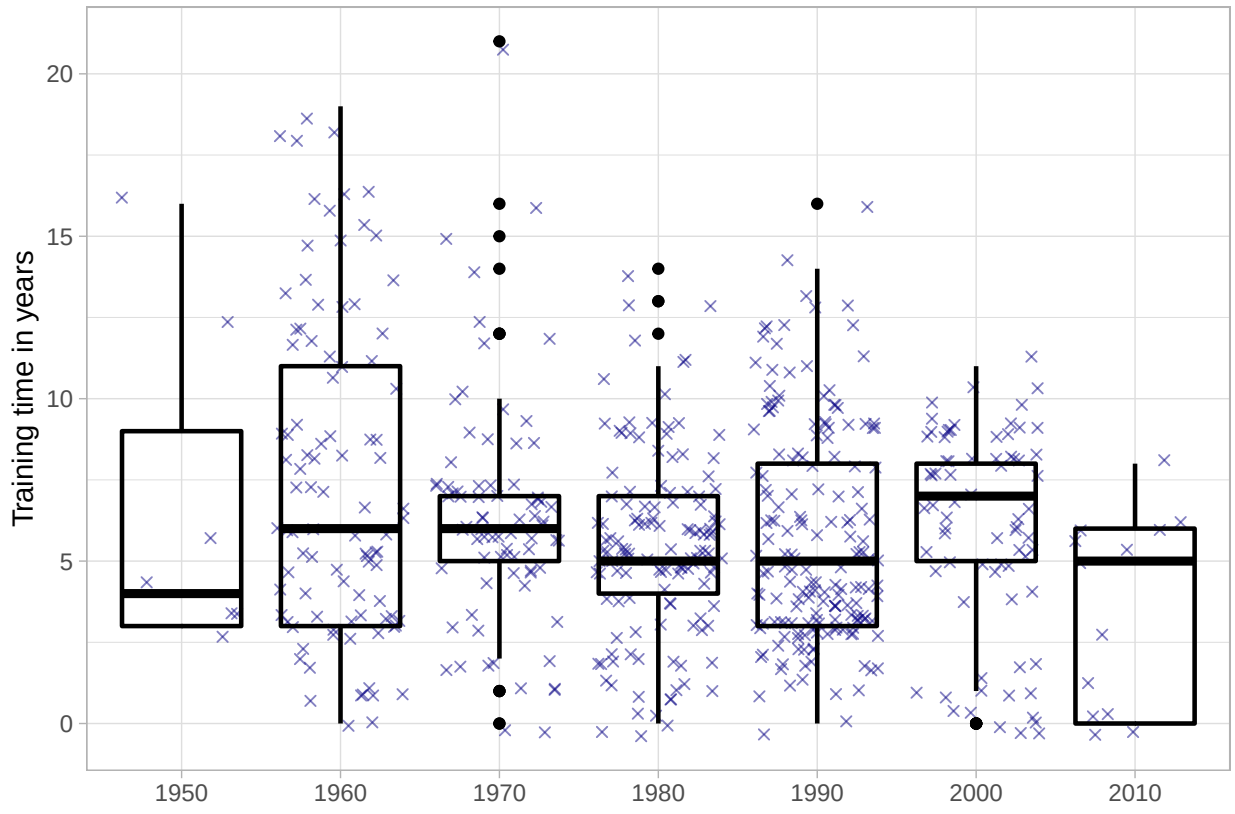

```
ggplot(astro_data_ind, aes(x = year_of_selection, y = age_selected)) +
  geom_jitter(aes(size = calc_total_hrs, col = calc_total_hrs), alpha = 0.5) +
  labs(x = "Selection year", y = "Selection Age") +
  scale_color_viridis_c(option = "C") +
  theme_light()
```



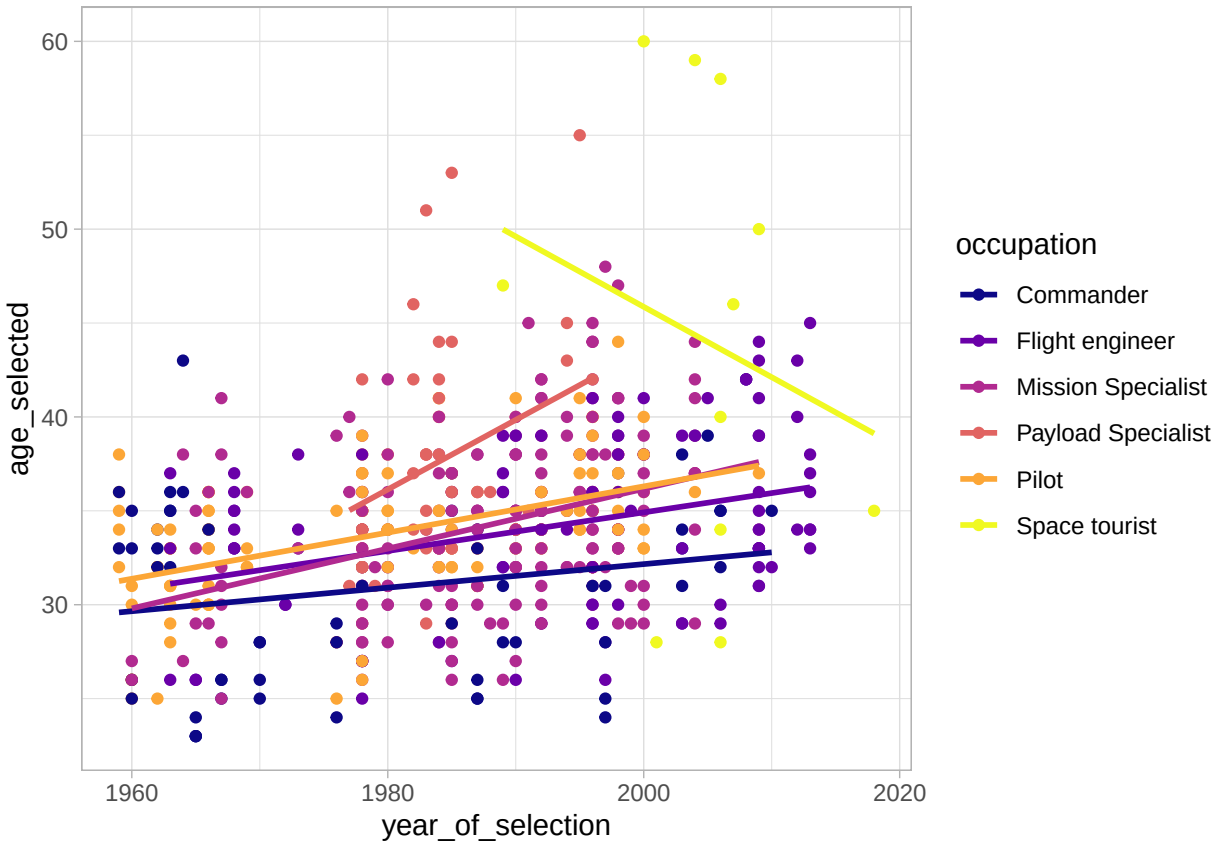

```
ggplot(astro_data_ind, aes(x = factor(decade_sel), y = age_selected)) +
  geom_jitter(aes(col = total_number_of_missions), alpha = 0.5, shape = 8) +
  geom_boxplot(col = "black", size = .8, fill = NA) +
  labs(x = "Decade", y = "Selection Age") +
  theme_light()
```



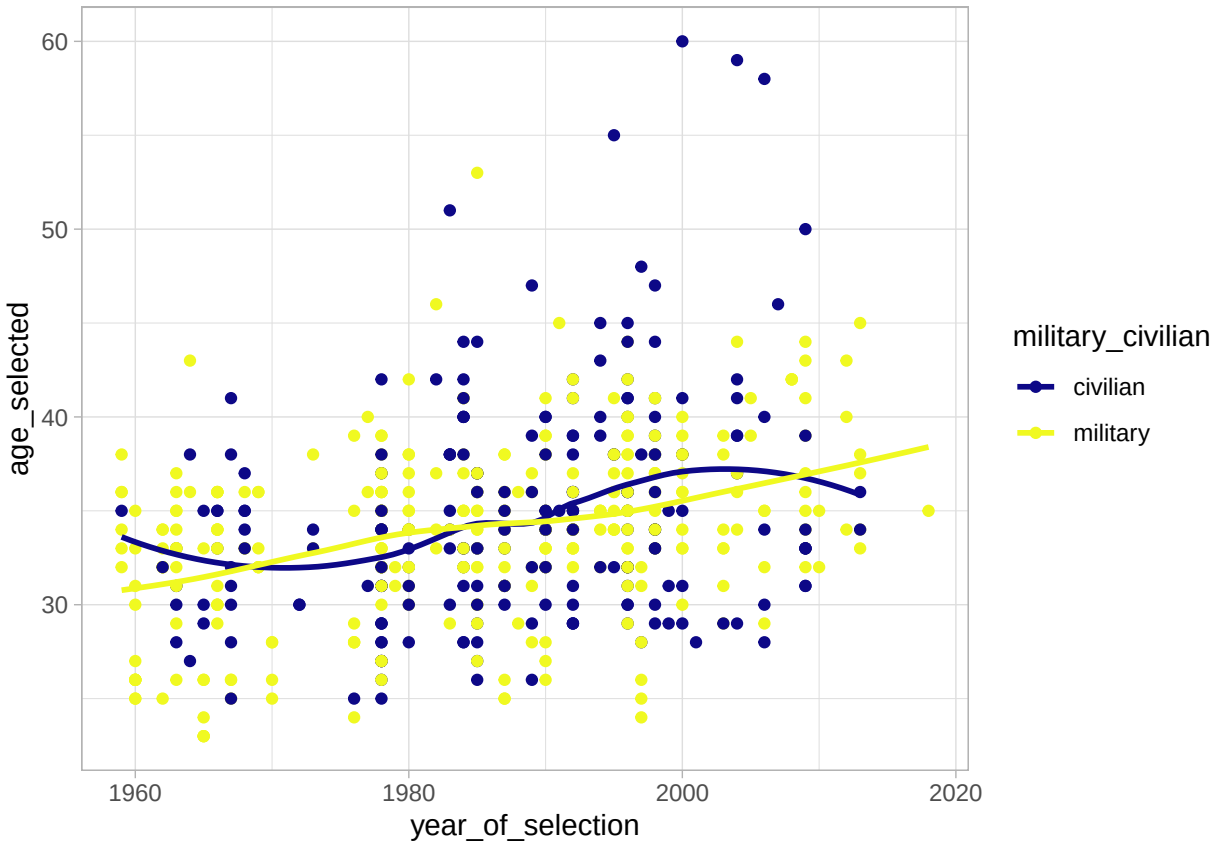

```
ggplot(astro_data_ind, aes(x = factor(decade_sel), y = train_time)) +
  geom_jitter(alpha = 0.5, col = plasma(1), shape = 4) +
  geom_boxplot(col = "black", size = .8, fill = NA) +
  labs(x = "", y = "Training time in years") +
  theme_light()
```



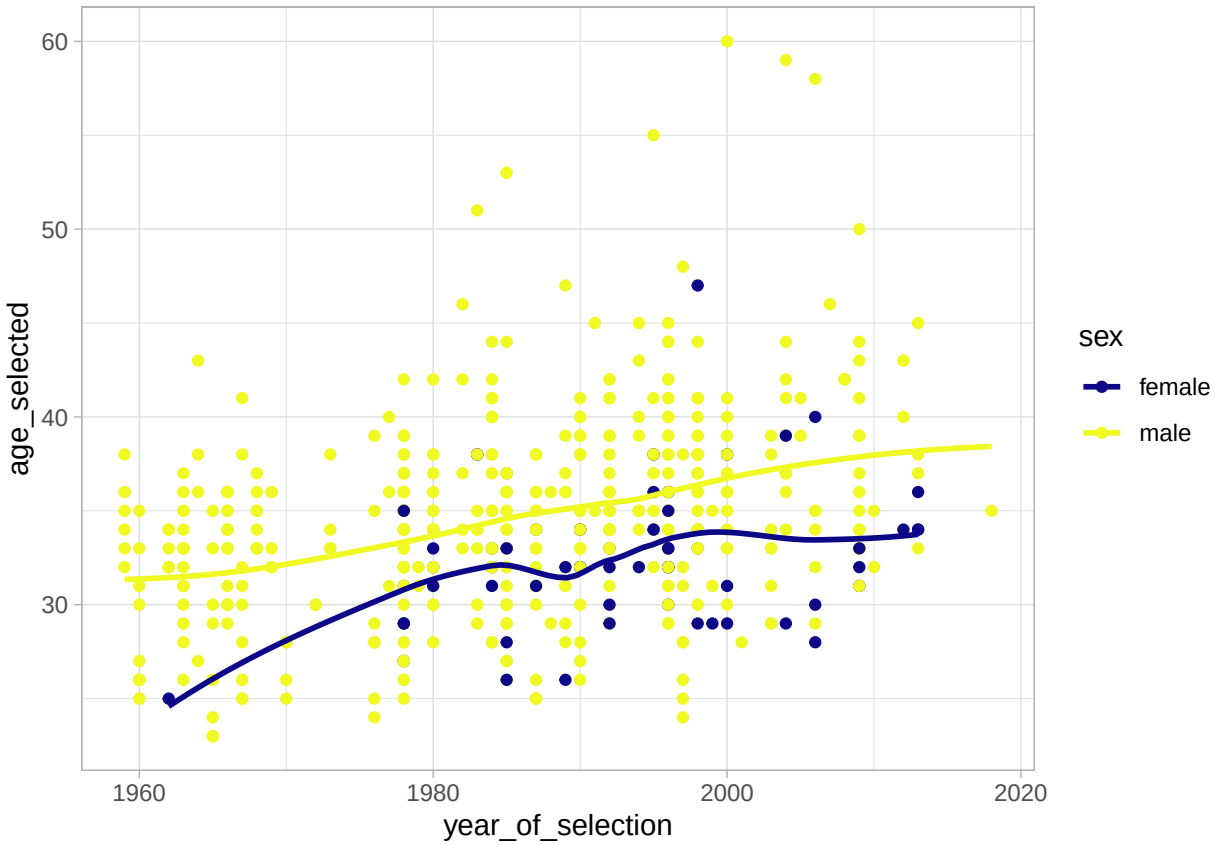

```
ggplot(astro_data_ind, aes(x = year_of_selection, y = age_selected, col = occupation)) +
  geom_point() +
  geom_smooth(method = "lm", se = F, formula = 'y ~ x') +
  scale_color_viridis_d(option = "C") +
  theme_light()
```




```
ggplot(astro_data_ind, aes(x = year_of_selection, y = age_selected, col = military_civilian)) +
  geom_point() +
  geom_smooth(method = "loess", se = F, formula = 'y ~ x') +
  scale_color_viridis_d(option = "C") +
  theme_light()
```




```
ggplot(astro_data_ind, aes(x = year_of_selection, y = age_selected, col = sex)) +
  geom_point() +
  geom_smooth(method = "loess", se = F, formula = 'y ~ x') +
  scale_color_viridis_d(option = "C") +
  theme_light()
```


60
50
40
30
age_selected
1960
1980
2000
2020
year_of_selection
sex
female
male