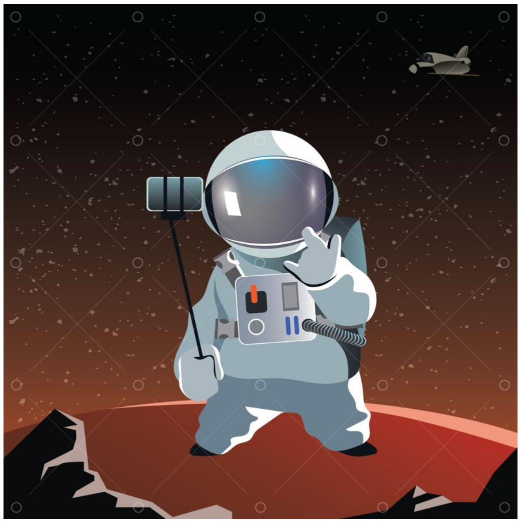
Data Science
Association
Munich

When is the prime time to be selected to fly to space?

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. So why not try to find under major simplifications of course at roughly what age I should be drafted into the elite team of astronauts and how much time I have left to become worthy ;).



First of all the R packages used:

```
library(tidyverse)
library(viridis)
library(patchwork)
library(ggtext)
theme_set(
  theme_light() +
    theme(plot.title = ggtext::element_markdown(size = 11),
          plot.subtitle = ggtext::element_markdown(size = 8))
)
```

(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(.))))
```

original_name	selection	mission_title	ascend_shuttle	descend_shuttle
NA	Air Gorce Group 6 - USSR	Soyuz 28	Soyuz 28	Soyuz 28
Farkas Bertalan	1978 Intercosmos Group	Soyuz 36/35	NA	NA
NA	Saudi-Arabia	STS-51G	STS-51G	STS-51G
NA	1985 NASA Group	STS-61-B	STS-61-B	STS-61-B
NA	Syria	1	Soyuz TM-3	Soyuz TM-2
NA	Afghanistan	3	Soyuz TM-6	Soyuz TM-5
Olsen, Gregory Hammond	NA	Soyuz TMA-7	Soyuz TMA-7	Soyuz TMA-7
Parmitano, Luca	2009 ESA Group	NA	Soyuz MS-13	not completed yet

```
### 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)
```

name	n_birth_years
Aleksandrov, Aleksandr	2

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

id	name	nationality	year_of_birth
248	Aleksandrov, Aleksandr	U.S.S.R/Russia	1943
249	Aleksandrov, Aleksandr	U.S.S.R/Russia	1943
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. As the only missing values occur in the variables `original_name`, once in the `selection`, `mission_title` and 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)
```

variable	unique_classes	n_obs
military_civilian	military	769
military_civilian	civilian	508
occupation	pilot	196
occupation	PSP	59
occupation	Pilot	1
occupation	commander	315
occupation	MSP	498
occupation	flight engineer	192
occupation	Other (Journalist)	1
occupation	Flight engineer	4
occupation	Other (space tourist)	5
occupation	Other (Space tourist)	3
occupation	Space tourist	2
occupation	spaceflight participant	1
sex	male	1134
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` ≤ `year_of_selection`
3. `year_of_selection` ≤ `year_of_mission`
4. Check some suspiciously high values from the summary above. (not the ones covered in 5-6.)
5. `hours_mission` ≤ `total_hrs_sum` and the sum should be the total + the max is suspiciously high.
6. `eva_hrs_mission` ≤ `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` ≤ `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)
```

id	name	year_of_selection	year_of_mission
648	Franco Malerba	1998	1992
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)
```

id	name	year_of_selection	year_of_mission
862	Thomas, Andrew S. W.	1992	1983
863	Thomas, Andrew S. W.	1992	1998
864	Thomas, Andrew S. W.	1992	2001

id	name	year_of_selection	year_of_mission
865	Thomas, Andrew S. W.	1992	2005

```
# 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)
```

id	name	nationwide_number	number
1271	Hague, Tyler	433	559
1276	Meir, Jessica	348	564
1275	Morgan, Andrew	347	563

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

id	name	nationwide_number	number
1270	Hague, Tyler	344	559
1271	Hague, Tyler	433	559

```
# 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)
```

id	name	year_of_selection	year_of_mission
1277	Al Mansoori, Hazzaa	2018	2019
1239	Zhang, Xiaoguang	2013	2013
1240	Wang, Yapi	2013	2013

```
# 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)
```

id	name	hours_mission
449	Polyakov, Valeri	10505.00
1163	Kimbrough, Robert S.	10383.25
1222	Borisenko, Andrei	10383.25
1262	Ryzhikov, Sergey	10383.25
638	Avdeyev, Sergei	9110.00

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

id	name	total_hrs_sum
942	Padalka, Gennady	21083.52
451	Krikalev, Sergei	19281.65
609	Kaleri, Aleksandr	18462.62

```r
# 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)
```

name	actual_total_hrs	total_hrs_sum	diff_total_hrs
Pesquet, Thomas	4721.83	15105.08	10383.25
Arnold, Richard R., II	5029.00	307.00	4722.00
Feustel, Andrew J.	5409.00	687.00	4722.00
Crouch, Roger Keith	472.55	4576.07	4103.52
Linteris, Gregory Thomas	472.55	4576.07	4103.52
Coleman, Catherine G.	4324.00	500.00	3824.00
Arnaldo Tamayo Mendez	1887.71	188.71	1699.00

name	actual_total_hrs	total_hrs_sum	diff_total_hrs
Henricks, Terence T.	1024.00	190.75	833.25
Weitz, Paul J.	387.00	793.22	406.22
Cassidy, Christopher J.	4367.00	4744.00	377.00

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)
```

id	name	eva_hrs_mission	total_eva_hrs
446	Solovyev, Anatoly	89.13	78.80
1275	Morgan, Andrew	39.52	39.52
1033	Whitson, Peggy A.	35.35	60.31
1015	Tani, Daniel M.	34.98	39.18
1021	Walheim, Rex J.	34.60	56.73

```
# 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)
```

name	actual_total_eva	total_eva_hrs	diff_total_eva
Hague, Tyler	39.86	19.93	19.93
Leestma, David C.	1.00	3.50	2.50

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

id	name	eva_hrs_mission	total_eva_hrs	in_orbit
1270	Hague, Tyler	19.93	19.93	aborted
1271	Hague, Tyler	19.93	19.93	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)
```

id	name	eva_hrs_mission	total_eva_hrs	in_orbit
316	Leestma, David C.	1	3.5	STS-41-G
317	Leestma, David C.	0	3.5	STS-28
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)
```

name	valid_mission_number
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)
```

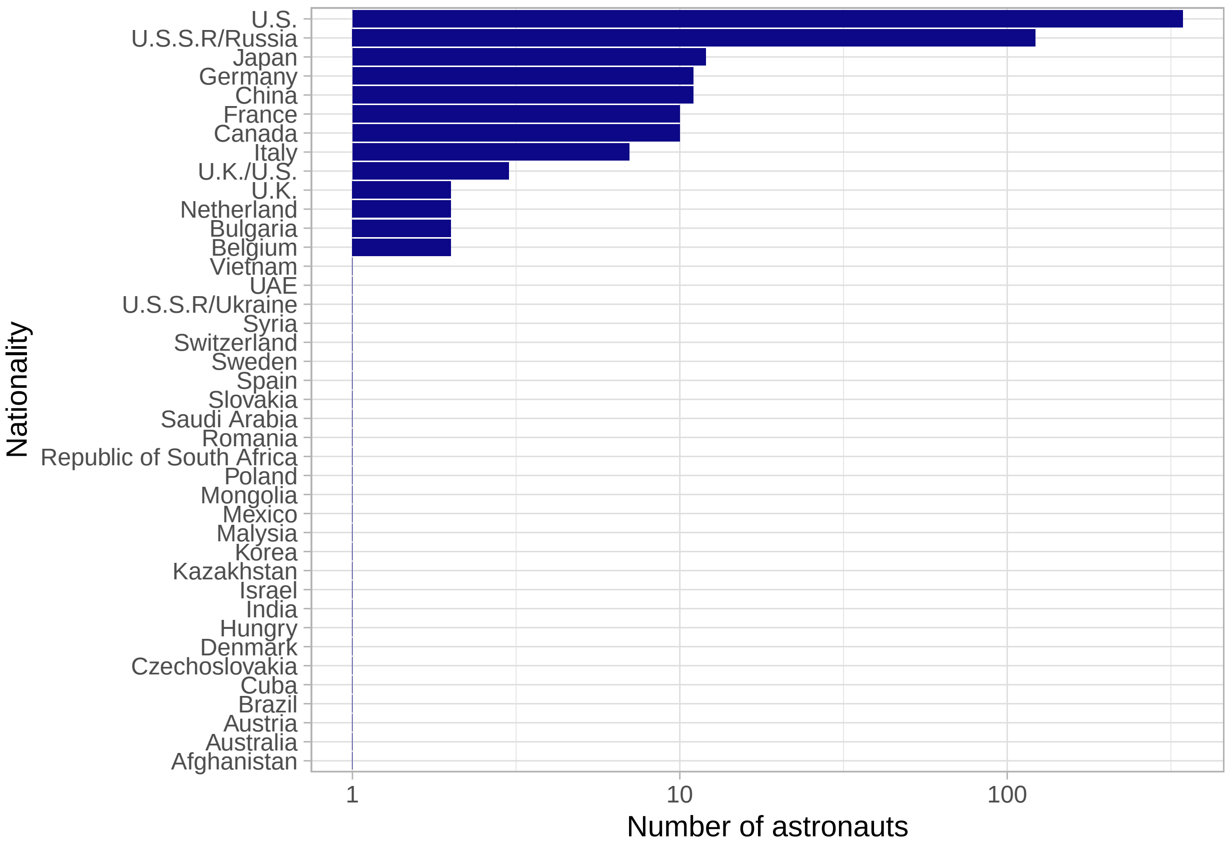
id	name	mission_number	total_number_of_missions	mission_title	year_of_mission
117	Shepard, Alan B., Jr.	2	2	Apollo 14	1971

```
# 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()
```



```
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.

```
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

So in order to take my selfie in space I have to be selected and to be selected I probably have to put in some work to become worthy. Thus the new variable `age_selected` will be at the core of the further analysis in order to find out a reasonable or even a perfect(?) selection age for me. On the way I might also stumble upon a nice hypothesis to check:). To accomplish this a close look at the univariate distribution and key statistical features is a good first step.

```
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)
```

name	age_selected	occupation	year_of_selection	nationality
Klimuk, Pyotr	23	Commander	1965	U.S.S.R/Russia
Sarafanov, Gennadi	23	Commander	1965	U.S.S.R/Russia
Zudov, Vyacheslav	23	Commander	1965	U.S.S.R/Russia
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)
```

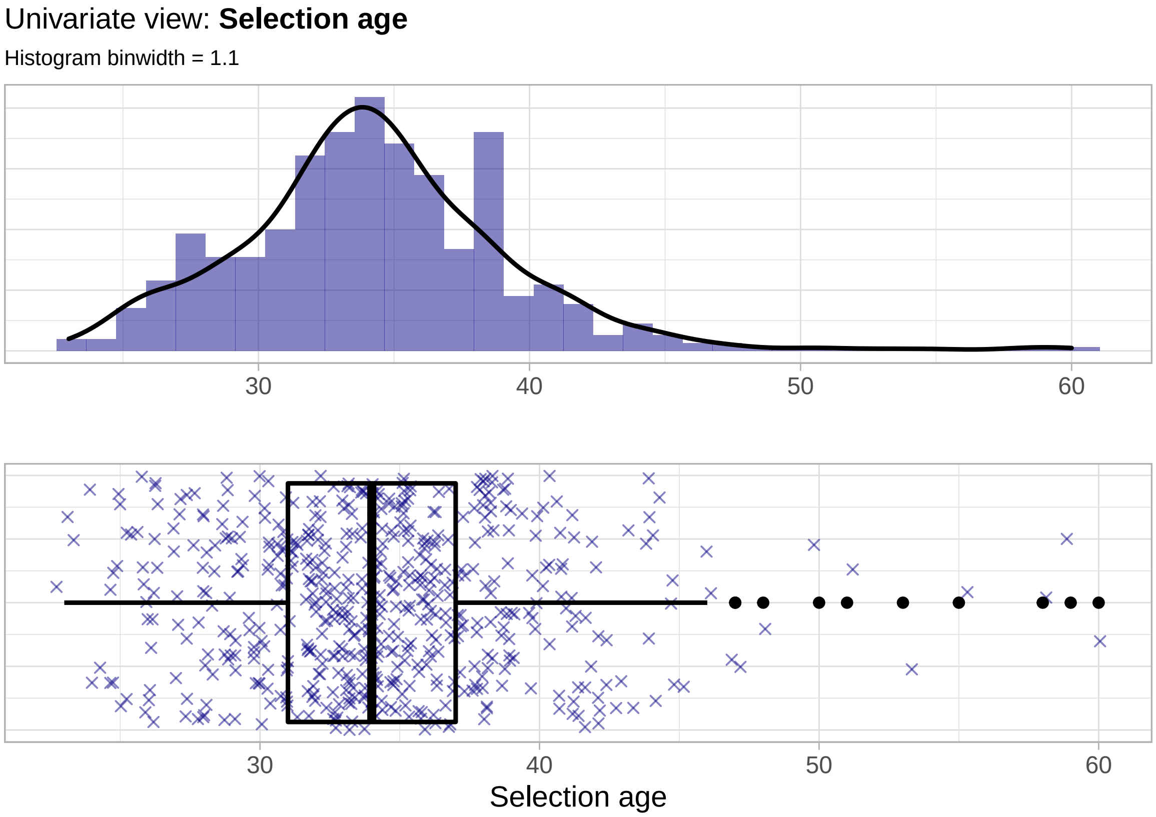
name	age_selected	occupation	year_of_selection	nationality
Crouch, Roger Keith	55	Payload Specialist	1995	U.S.
Simonyi, Charles (Karoly)	58	Space tourist	2006	U.S.
Olsen, Gregory Hammond	59	Space tourist	2004	U.S.
Tito, Dennis Anthony	60	Space tourist	2000	U.S.

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.

```r
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(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(axis.ticks.y = element_blank(),
        axis.text.y = element_blank())
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:** The selection age has changed over time and there are certain subgroups of astrounauts e.g. w.r.t. sex or occupation that show different patterns regarding the selection age. Were the selection ages of the astronauts of the space race lower and then grew over time. What were the selection ages of very successfull astronauts?

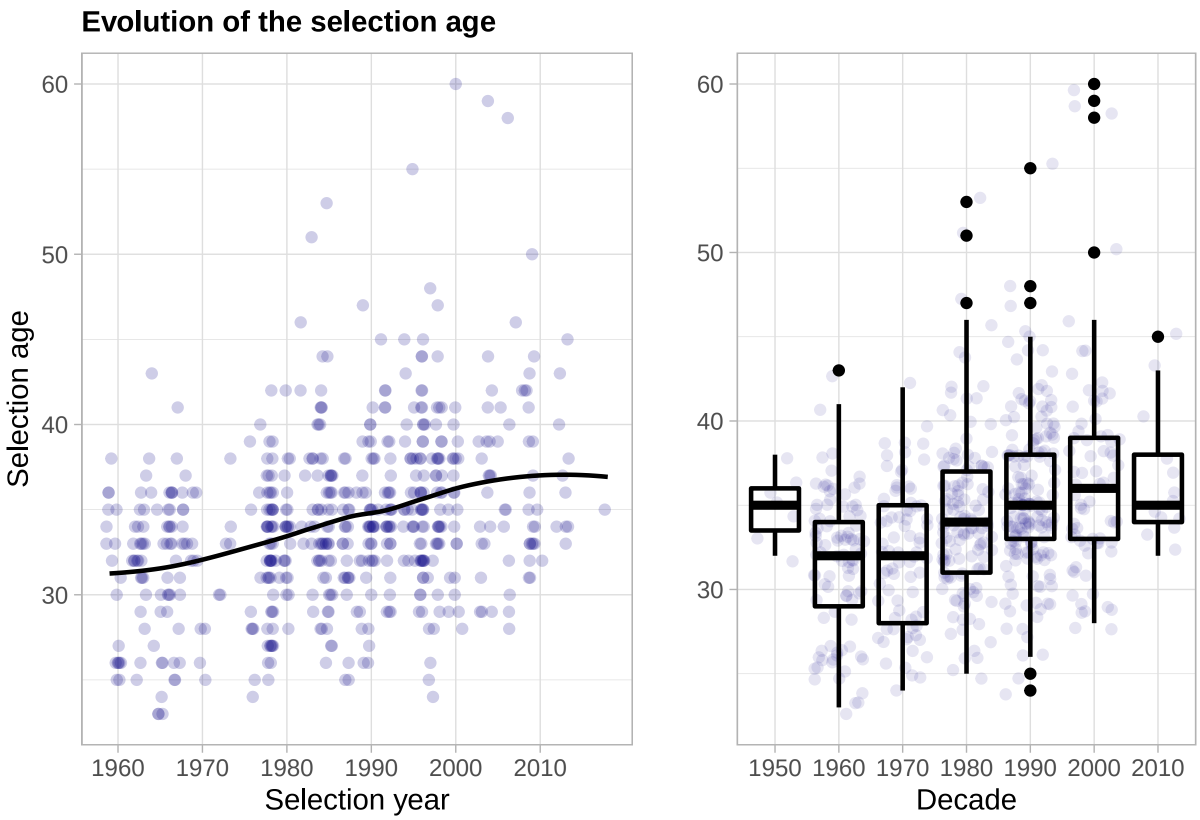
In order to find indications for or against this hypothesis the multivariate dependencies with other variables have to be examined in the next part.

–>

(3) Apogee & splashdown or visualization & conclusion

Selection age over time

```
# jittered scatterplot of selection age vs time
ggplot(astro_data_ind, aes(x = year_of_selection, y = age_selected)) +
  geom_jitter(alpha = 0.2, col = plasma(1), height = 0) +
  labs(x = "Selection year", y = "Selection age",
       title = "**Evolution of the selection age**") +
  geom_smooth(method = "loess", se = F, formula = 'y ~ x', size = 0.8,
              col = "black") +
  scale_x_continuous(breaks = seq(1960, 2010, 10)) +
  theme(panel.grid.minor.x = element_blank()) +
# grouped boxplots by decade
ggplot(astro_data_ind, aes(x = factor(decade_sel), y = age_selected)) +
  geom_jitter(col = plasma(1), alpha = 0.1) +
  geom_boxplot(col = "black", size = .8, fill = NA) +
  labs(x = "Decade", y = "") +
# get the desired layout
plot_layout(widths = c(6, 5))
```



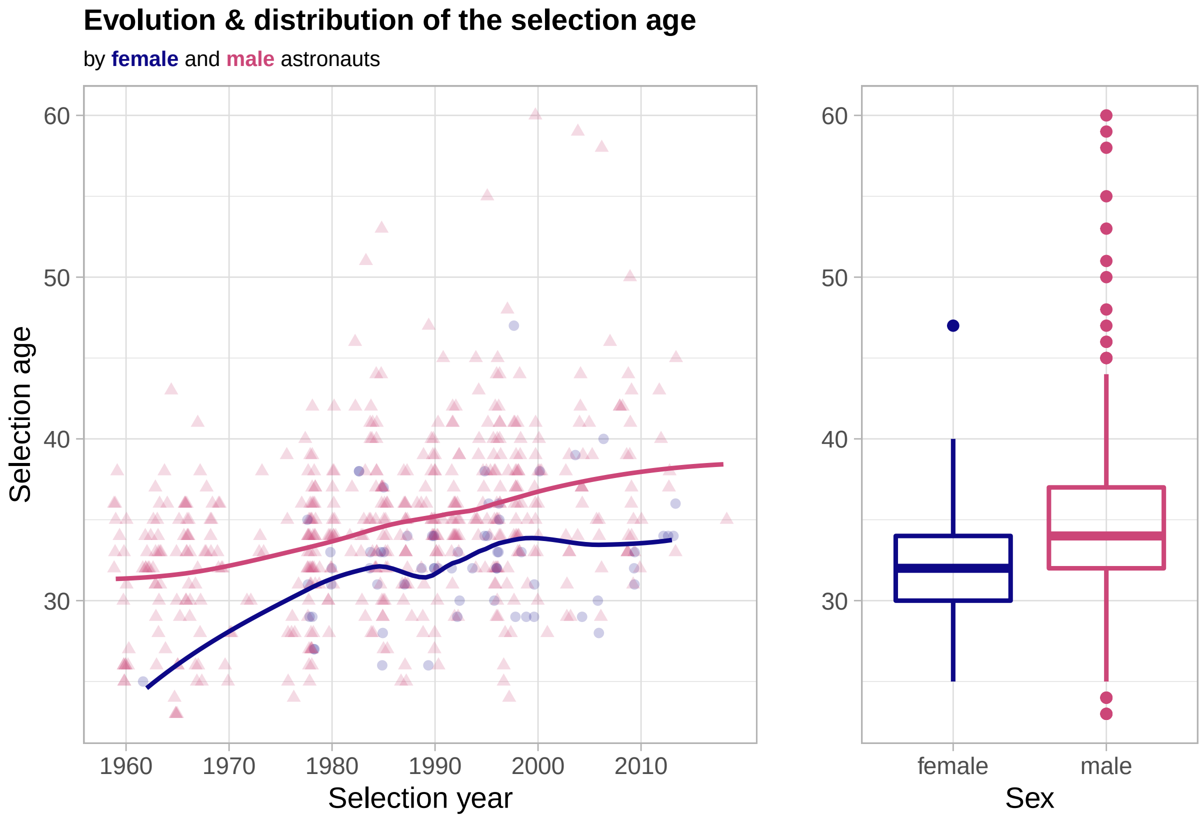
So indeed the data shows an overall positive trend that indicates an average selection age of roughly in the early thirties during the space race in the 1960ies that rises steadily until reaching a level of roughly 37 years in the 2010s. In the right plot that shows grouped boxplots w.r.t. the decade interesting enough the first an last boxplot contradict the overall trend somehow, in the first case this is probably due to the really small sample and in the latter case this could be due to the fact that the U.S. has lost it's capability to launch people into space due to the last space shuttle mission being in 2011. Only in the last years the U.S. regained the ability to put people into space with the Crew Dragon spacecraft. This could explain the reduced number of astronauts selected in this decade and gives rise to the question whether different nations select astronauts at differing ages. The rate of growth of the selection age overall also seems to dampen. So the average selection age will probably reach a stationary point somewhere below the average age of a working person for example 40. But before making such forecasts one should have a look at the other variables w.r.t. selection age.

Selection age w.r.t. categories

First of all one can look at the categorical variables and their influence on the evolution of the selection age.

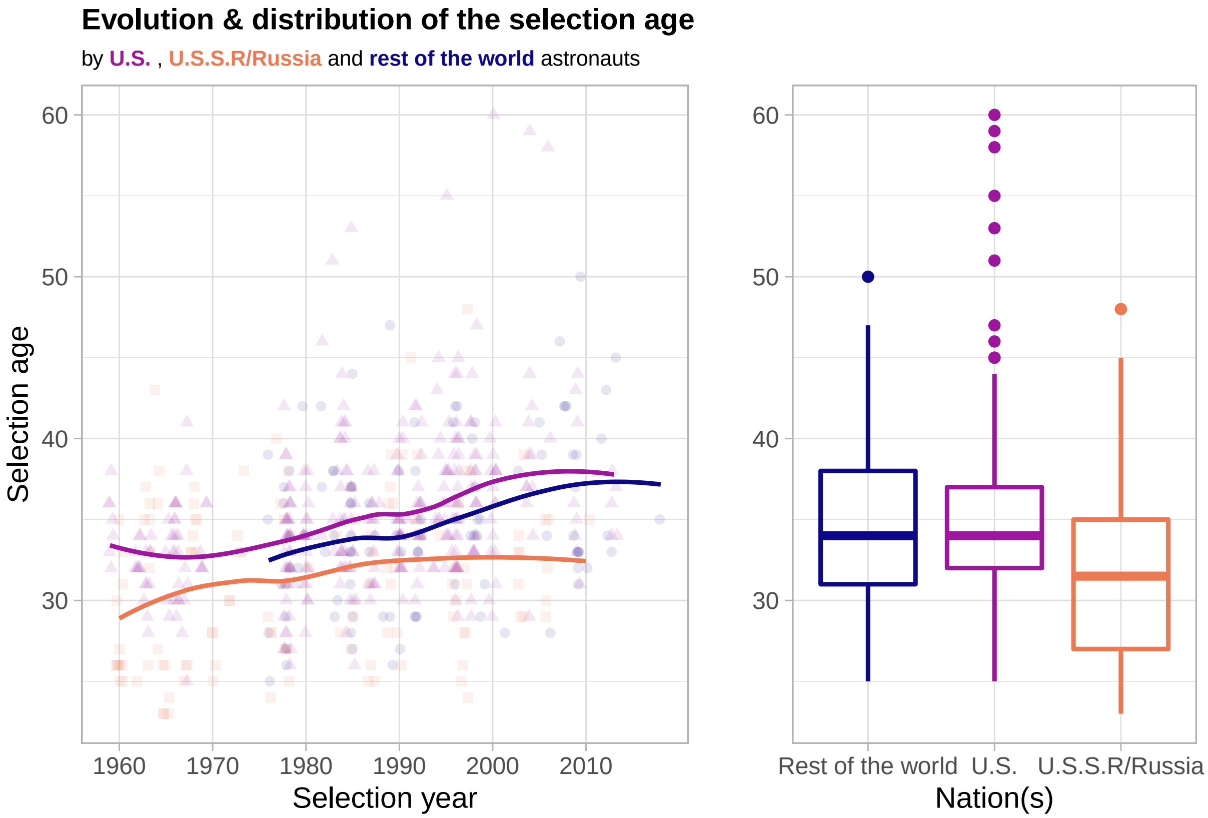
- `sex`
- `nationality_red` just compare the two biggest space nations so far and the rest of the world.
- `military_civilian`
- `occupation`

```
# conditional scatterplot -wrt sex- of the selection age over time
ggplot(astro_data_ind, aes(x = year_of_selection, y = age_selected,
                           col = factor(sex))) +
  geom_jitter(aes(shape = factor(sex)), alpha = 0.2, height = 0) +
  labs(x = "Selection year", y = "Selection age",
       title = "**Evolution & distribution of the selection age**",
       subtitle = paste("by",
                        "<span style='color:",
                        plasma(3)[1],
                        "'>**female**</span>",
                        " and ",
                        "<span style='color:",
                        plasma(3)[2],
                        "'>**male**</span>",
                        " astronauts")) +
  geom_smooth(method = "loess", se = F, formula = 'y ~ x', size = 0.8) +
  scale_color_manual(values = plasma(3)[1:2]) +
  scale_x_continuous(breaks = seq(1960, 2010, 10)) +
  guides(col = "none", shape = "none") +
  theme(panel.grid.minor.x = element_blank()) +
# compare the conditional distributions with side to side boxplots
ggplot(astro_data_ind, aes(x = factor(sex), y = age_selected,
                           col = factor(sex))) +
  geom_boxplot(size = .8, fill = NA) +
  scale_color_manual(values = plasma(3)[1:2]) +
  guides(col = "none") +
  labs(x = "Sex", y = "") +
# get the desired layout
plot_layout(widths = c(2, 1))
```

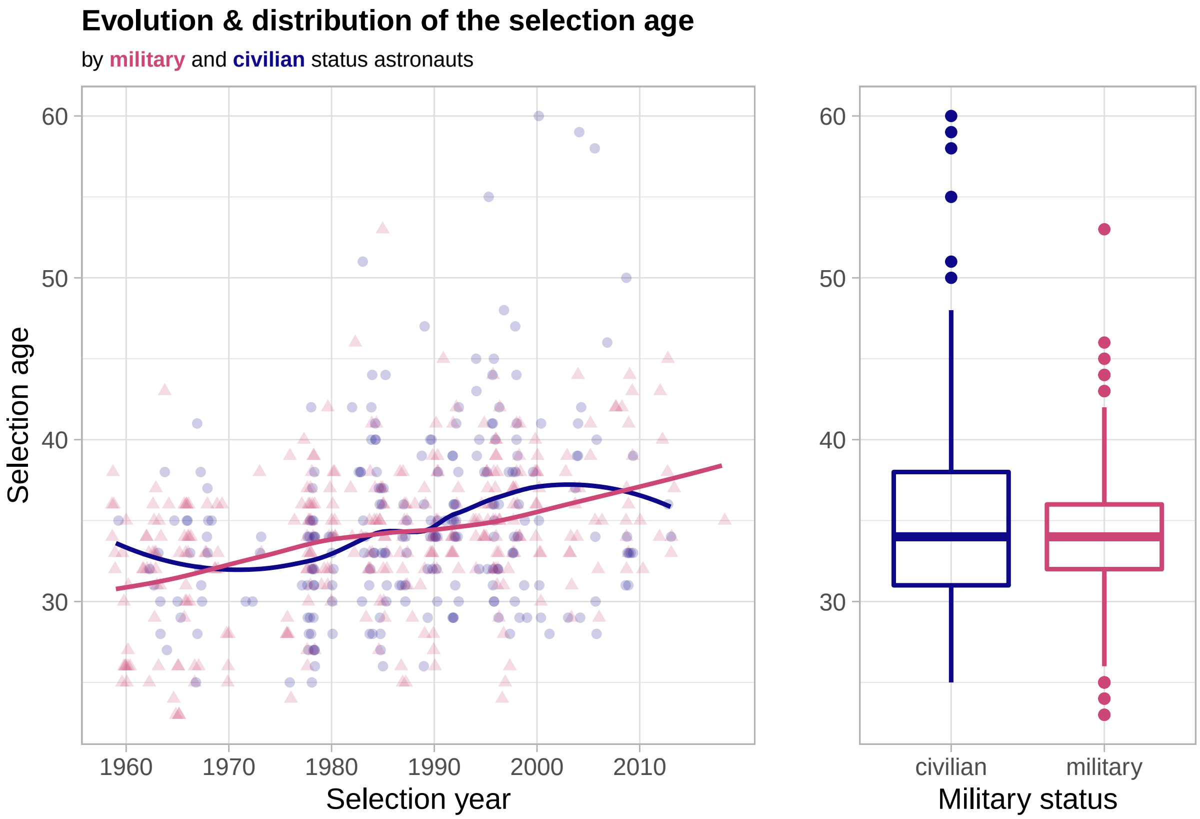


```
# conditional scatterplot -wrt sex- of the selection age over time
ggplot(astro_data_ind, aes(x = year_of_selection, y = age_selected,
                           col = factor(nationality_red))) +
  geom_jitter(aes(shape = factor(nationality_red)),
              alpha = 0.1, height = 0) +
  labs(x = "Selection year", y = "Selection age",
       title = "**Evolution & distribution of the selection age**",
       subtitle = paste("by",
                        "<span style='color:",
                        plasma(4)[2],
                        "'>**U.S.**</span>",
                        ", ",
                        "<span style='color:",
                        plasma(4)[3],
                        "'>**U.S.S.R/Russia**</span>",
                        " and ",
                        "<span style='color:",
                        plasma(4)[1],
                        "'>**rest of the world**</span>",
                        " astronauts")) +
  geom_smooth(method = "loess", se = F, formula = 'y ~ x', size = 0.8) +
  scale_color_manual(values = plasma(4)[1:3]) +
  scale_x_continuous(breaks = seq(1960, 2010, 10)) +
  guides(col = "none", shape = "none") +
  theme(panel.grid.minor.x = element_blank()) +
# compare the conditional distributions with side to side boxplots
```

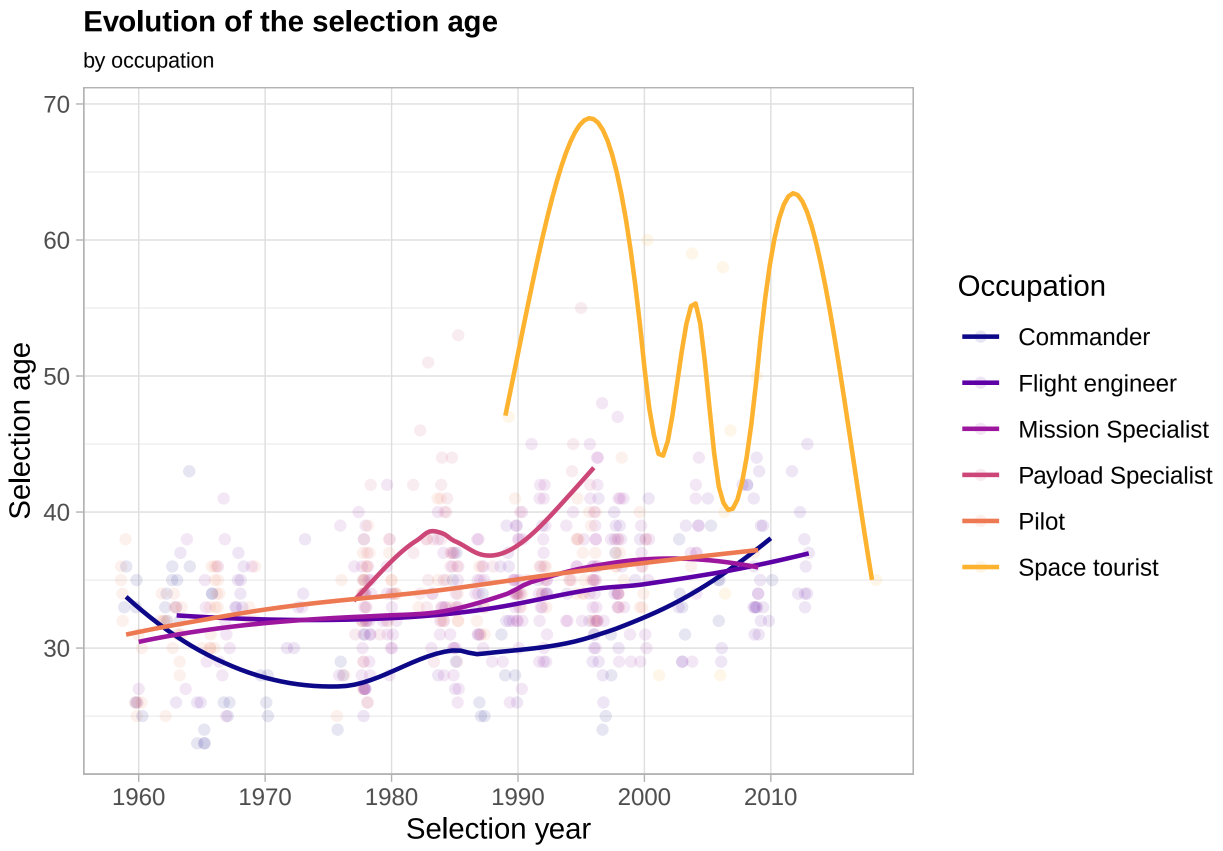
```
ggplot(astro_data_ind, aes(x = factor(nationality_red), y = age_selected,
                           col = factor(nationality_red))) +
  geom_boxplot(size = .8, fill = NA) +
  scale_color_manual(values = plasma(4)[1:3]) +
  guides(col = "none") +
  labs(x = "Nation(s)", y = "") +
# get the desired layout
plot_layout(widths = c(3, 2))
```



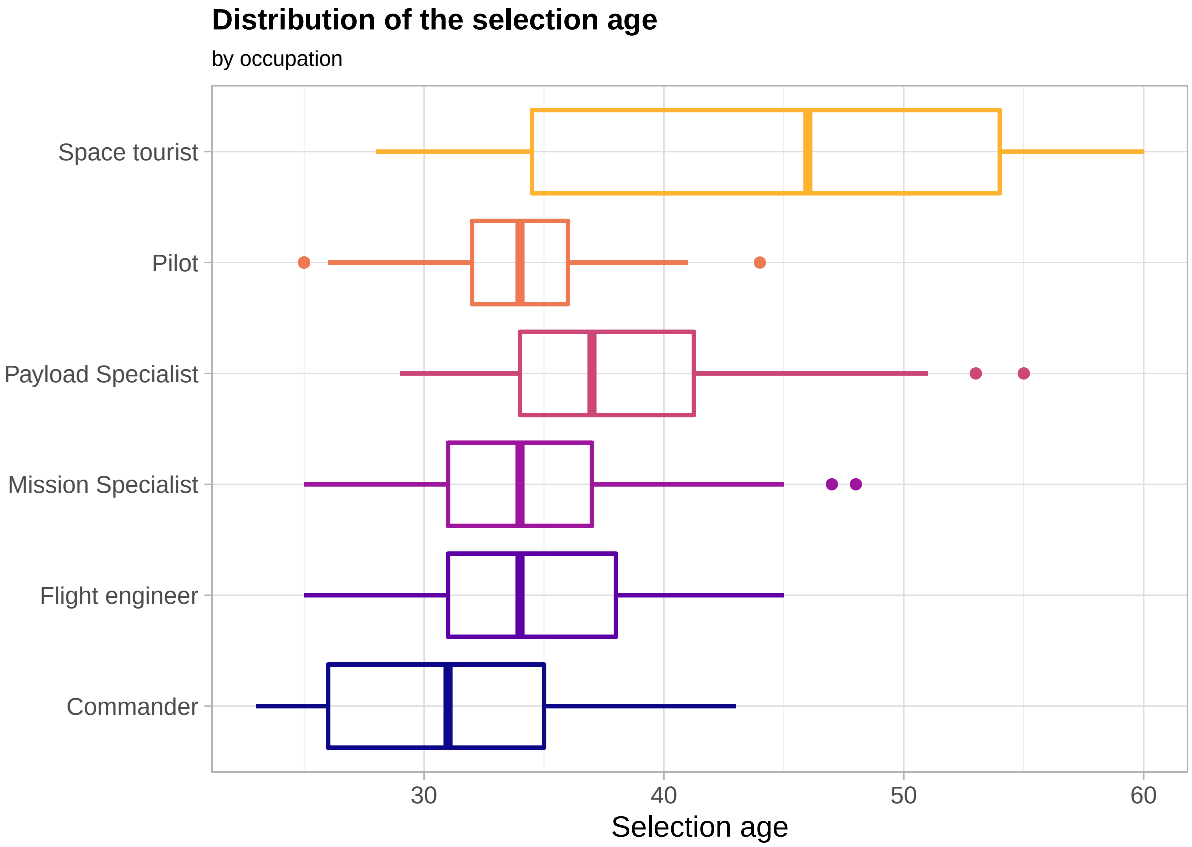
```
# conditional scatterplot -wrt sex- of the selection age over time
ggplot(astro_data_ind, aes(x = year_of_selection, y = age_selected,
                           col = factor(military_civilian))) +
  geom_jitter(aes(shape = factor(military_civilian)),
              alpha = 0.2, height = 0) +
  labs(x = "Selection year", y = "Selection age",
       title = "**Evolution & distribution of the selection age**",
       subtitle = paste("by",
                        "<span style='color:",
                        plasma(3)[2],
                        "'>**military**</span>",
                        " and ",
                        "<span style='color:",
                        plasma(3)[1],
                        "'>**civilian**</span>",
                        " status astronauts")) +
  geom_smooth(method = "loess", se = F, formula = 'y ~ x', size = 0.8) +
  scale_color_manual(values = plasma(3)[1:2]) +
  scale_x_continuous(breaks = seq(1960, 2010, 10)) +
  guides(col = "none", shape = "none") +
  theme(panel.grid.minor.x = element_blank()) +
# compare the conditional distributions with side to side boxplots
ggplot(astro_data_ind, aes(x = factor(military_civilian), y = age_selected,
                           col = factor(military_civilian))) +
  geom_boxplot(size = .8, fill = NA) +
  scale_color_manual(values = plasma(3)[1:2]) +
  guides(col = "none") +
  labs(x = "Military status", y = "") +
# get the desired layout
plot_layout(widths = c(2, 1))
```



```
# conditional scatterplot -wrt sex- of the selection age over time
ggplot(astro_data_ind, aes(x = year_of_selection, y = age_selected,
                           col = factor(occupation))) +
  geom_jitter(alpha = 0.1, height = 0) +
  labs(x = "Selection year", y = "Selection age",
       title = "**Evolution of the selection age**",
       subtitle = "by occupation") +
  geom_smooth(method = "loess", se = F, formula = 'y ~ x', size = 0.8) +
  scale_color_manual(values = plasma(7)[1:6], name = "Occupation") +
  scale_x_continuous(breaks = seq(1960, 2010, 10)) +
  # guides(col = "none", shape = "none") +
  theme(panel.grid.minor.x = element_blank())
```



```
# compare the conditional distributions with side to side boxplots
ggplot(astro_data_ind, aes(x = factor(occupation), y = age_selected,
                           col = factor(occupation))) +
  geom_boxplot(size = .8, fill = NA) +
  scale_color_manual(values = plasma(7)[1:6]) +
  guides(col = "none") +
  coord_flip() +
  labs(x = "", y = "Selection age",
       title = "**Distribution of the selection age**",
       subtitle = "by occupation")
```



You got selected but how long are you gonna be trained

Scatter train vs age selected

Selection age w.r.t. greatest achievements

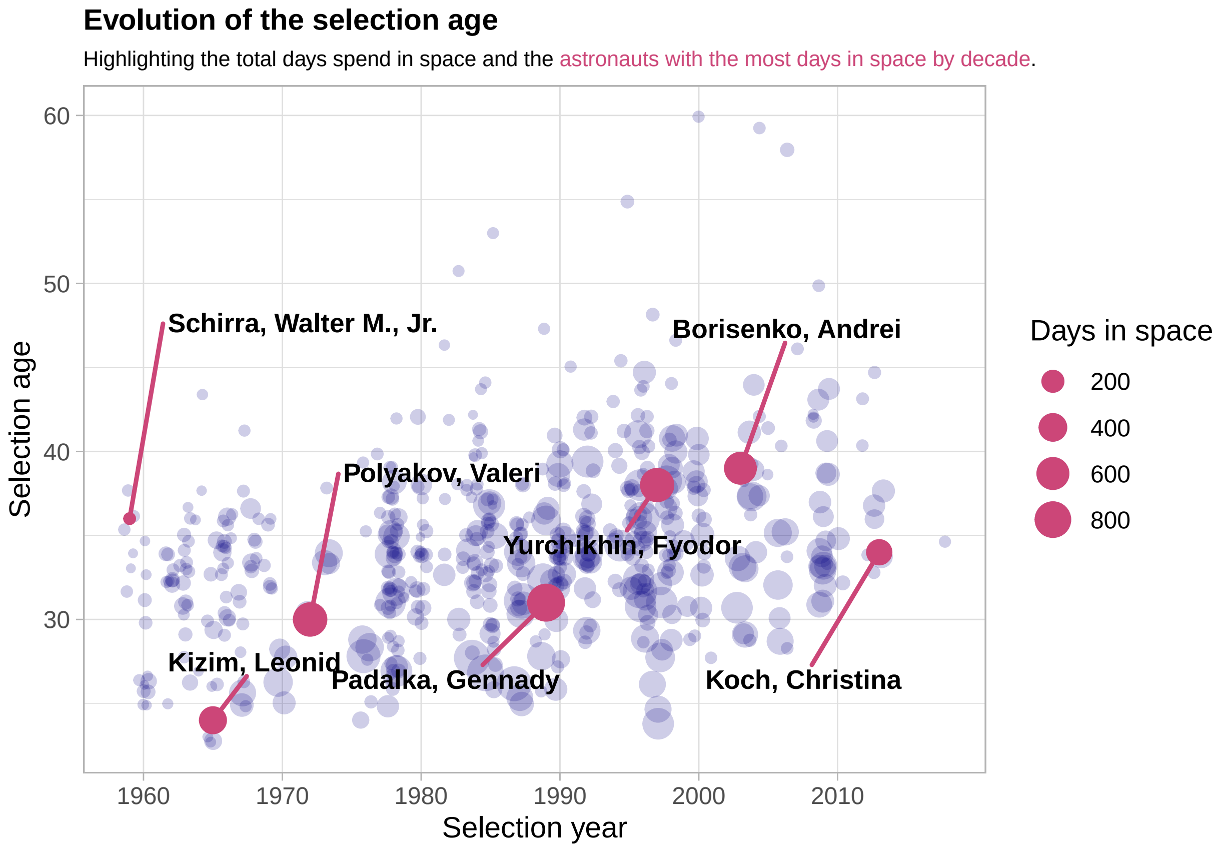
- total number of missions
- total eva hours
- calc total hours

```r
# extract the astronauts with the most time spend in space per decade selected
max_time_spend_per_decade <- astro_data_ind %>%
  mutate(calc_total_days = calc_total_hrs / 24) %>%
  group_by(decade_sel) %>%
  slice_max(order_by = calc_total_hrs, n = 1) %>%
  ungroup()
max_time_spend_per_decade %>%
  select(name, age_selected, year_of_selection, nationality, calc_total_days)
```

name	age_selected	year_of_selection	nationality	calc_total_days
Schirra, Walter M., Jr.	36	1959	U.S.	12.30083
Kizim, Leonid	24	1965	U.S.S.R/Russia	374.70833
Polyakov, Valeri	30	1972	U.S.S.R/Russia	678.62500

name	age_selected	year_of_selection	nationality	calc_total_days
Padalka, Gennady	31	1989	U.S.S.R/Russia	878.37500
Yurchikhin, Fyodor	38	1997	U.S.S.R/Russia	672.79167
Borisenko, Andrei	39	2003	U.S.S.R/Russia	596.84375
Koch, Christina	34	2013	U.S.	307.17917

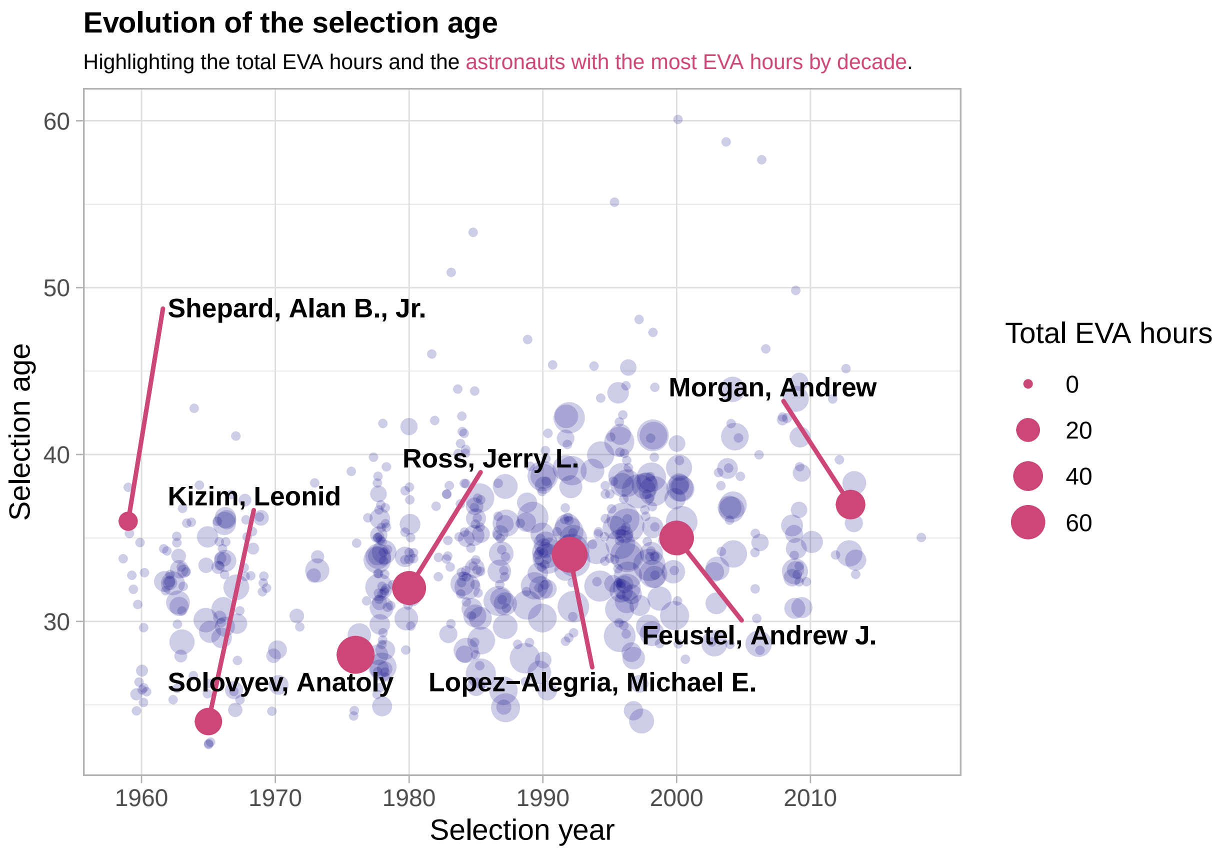
```
astro_data_ind %>%
  filter(!(name %in% max_time_spend_per_decade$name)) %>%
  mutate(calc_total_days = calc_total_hrs / 24) %>%
  ggplot(aes(x = year_of_selection, y = age_selected,
             size = calc_total_days, label = name)) +
    geom_jitter(alpha = 0.2, col = plasma(3)[1]) +
    geom_point(data = max_time_spend_per_decade,
      col = plasma(3)[2]) +
    labs(x = "Selection year", y = "Selection age",
         size = "Days in space",
         title = "**Evolution of the selection age**",
         subtitle = paste0(
           "Highlighting the total days spend in space and the",
           "<span style='color:",
           plasma(3)[2],
           "'> astronauts with the most days in space by decade</span>",
           ".")) +
    ggrepel::geom_text_repel(
      data = max_time_spend_per_decade,
      max.overlaps = Inf, box.padding = 2.2,
      size = 3.5, segment.color = plasma(3)[2],
      segment.size = .8,
      col = "black",
      fontface = "bold") +
    scale_x_continuous(breaks = seq(1960, 2010, 10)) +
    theme(panel.grid.minor.x = element_blank())
```



```
# extract the astronauts with the most time spend during eva per decade selected
max_eva_spend_per_decade <- astro_data_ind %>%
  group_by(decade_sel) %>%
  slice_max(order_by = total_eva_hrs, n = 1) %>%
  ungroup()
max_eva_spend_per_decade %>%
  select(name, age_selected, year_of_selection, nationality, total_eva_hrs)
```

name	age_selected	year_of_selection	nationality	total_eva_hrs
Shepard, Alan B., Jr.	36	1959	U.S.	9.25
Kizim, Leonid	24	1965	U.S.S.R/Russia	31.48
Solovyev, Anatoly	28	1976	U.S.S.R/Russia	78.80
Ross, Jerry L.	32	1980	U.S.	58.53
Lopez-Alegria, Michael E.	34	1992	U.S.	67.67
Feustel, Andrew J.	35	2000	U.S.	61.80
Morgan, Andrew	37	2013	U.S.	39.52

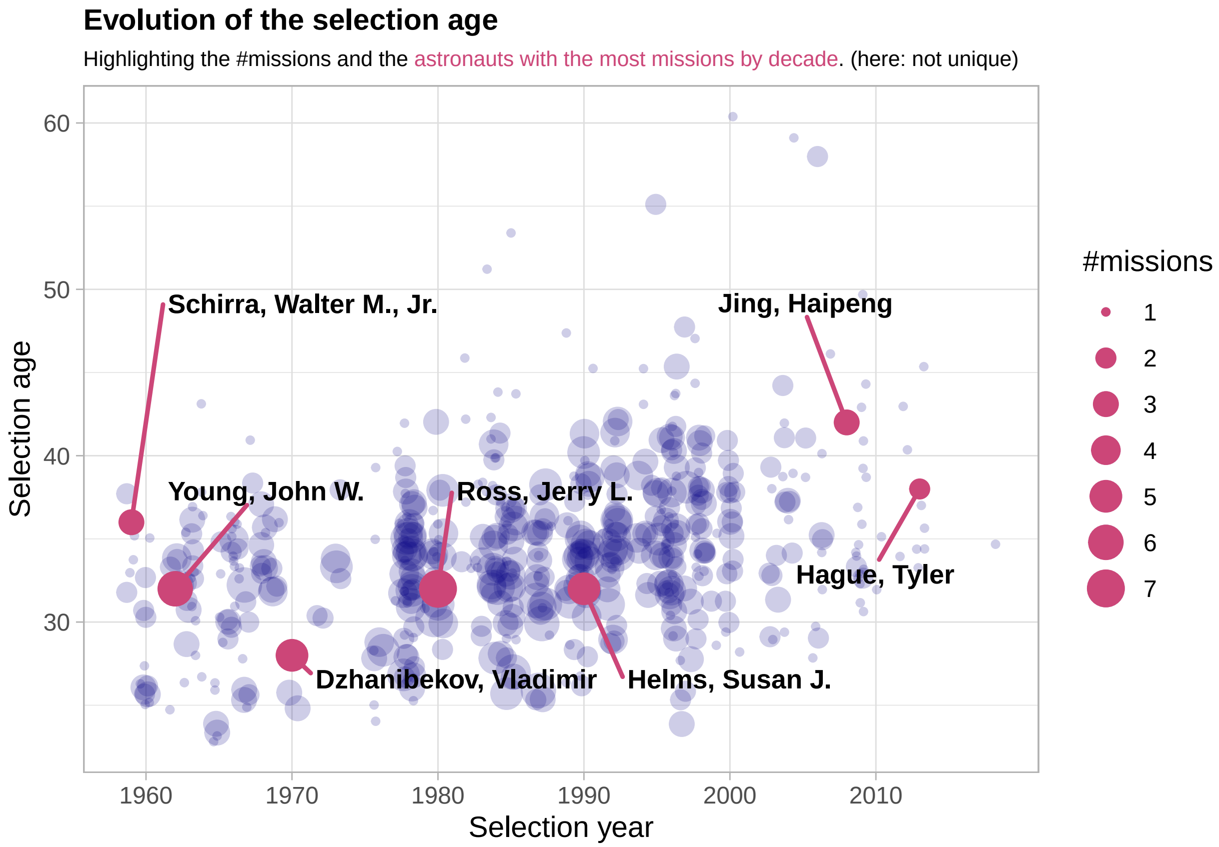
```
astro_data_ind %>%
  filter(!(name %in% max_eva_spend_per_decade$name)) %>%
  ggplot(aes(x = year_of_selection, y = age_selected,
             size = total_eva_hrs, label = name)) +
    geom_jitter(alpha = 0.2, col = plasma(3)[1]) +
    geom_point(data = max_eva_spend_per_decade,
      col = plasma(3)[2]) +
    labs(x = "Selection year", y = "Selection age",
         size = "Total EVA hours",
         title = "**Evolution of the selection age**",
         subtitle = paste0(
           "Highlighting the total EVA hours and the",
           "<span style='color:",
           plasma(3)[2],
           "'> astronauts with the most EVA hours by decade</span>",
           ".")) +
    ggrepel::geom_text_repel(
      data = max_eva_spend_per_decade,
      max.overlaps = Inf, box.padding = 2.2,
      size = 3.5, segment.color = plasma(3)[2],
      segment.size = .8,
      col = "black",
      fontface = "bold") +
    scale_x_continuous(breaks = seq(1960, 2010, 10)) +
    theme(panel.grid.minor.x = element_blank())
```



```
# extract the astronauts with the most missions per decade selected
# importantly without ties
max_missions_per_decade <- astro_data_ind %>%
  group_by(decade_sel) %>%
  slice_max(order_by = total_number_of_missions, n = 1,
            with_ties = FALSE) %>%
  ungroup()
max_missions_per_decade %>%
  select(name, age_selected, year_of_selection, nationality,
         total_number_of_missions)
```

name	age_selected	year_of_selection	nationality	total_number_of_missions
Schirra, Walter M., Jr.	36	1959	U.S.	3
Young, John W.	32	1962	U.S.	6
Dzhanibekov, Vladimir	28	1970	U.S.S.R/Russia	5
Ross, Jerry L.	32	1980	U.S.	7
Helms, Susan J.	32	1990	U.S.	5
Jing, Haipeng	42	2008	China	3
Hague, Tyler	38	2013	U.S.	2

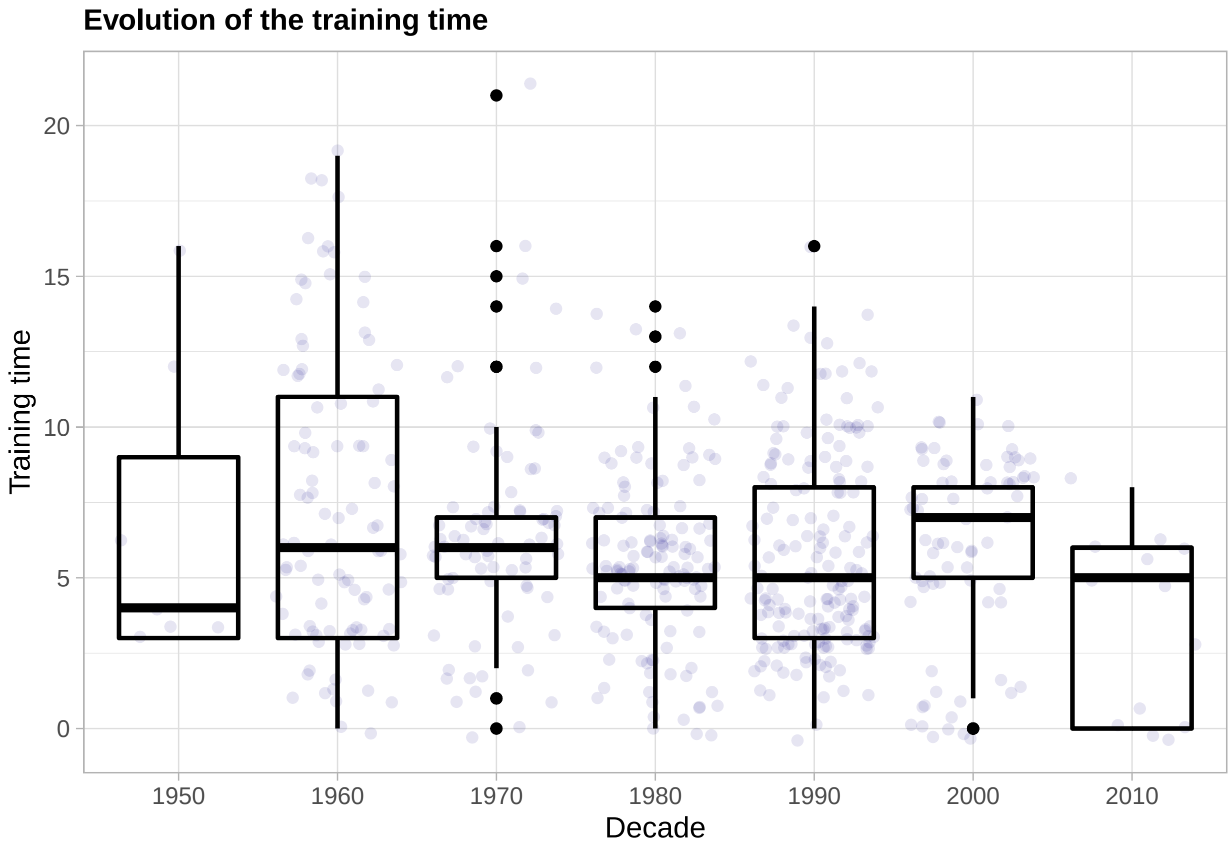
```
astro_data_ind %>%
  filter(!(name %in% max_missions_per_decade$name)) %>%
  ggplot(aes(x = year_of_selection, y = age_selected,
             size = total_number_of_missions, label = name)) +
    geom_jitter(alpha = 0.2, col = plasma(3)[1]) +
    geom_point(data = max_missions_per_decade,
      col = plasma(3)[2]) +
    labs(x = "Selection year", y = "Selection age",
         size = "#missions",
         title = "**Evolution of the selection age**",
         subtitle = paste0(
           "Highlighting the #missions and the",
           "<span style='color:",
           plasma(3)[2],
           "'> astronauts with the most missions by decade</span>",
           ". (here: not unique)")) +
    ggrepel::geom_text_repel(
      data = max_missions_per_decade,
      max.overlaps = Inf, box.padding = 2.2,
      size = 3.5, segment.color = plasma(3)[2],
      segment.size = .8,
      col = "black",
      fontface = "bold") +
    scale_x_continuous(breaks = seq(1960, 2010, 10)) +
    theme(panel.grid.minor.x = element_blank())
```


Evolution of the selection age

Highlighting the #missions and the astronauts with the most missions by decade. (here: not unique)

So when can I expect to make that space selfie?

```
ggplot(astro_data_ind, aes(x = factor(decade_sel), y = train_time)) +
  geom_jitter(col = plasma(1), alpha = 0.1) +
  geom_boxplot(col = "black", size = .8, fill = NA) +
  labs(x = "Decade", y = "Training time",
       title = "**Evolution of the training time**")
```


Evolution of the training time
Training time
0
5
10
15
20
1950
1960
1970
1980
1990
2000
2010
Decade