# labb3

Emilio Otero

2024-11-26

## Exploratory Data Analysis

### The IRIS dataset

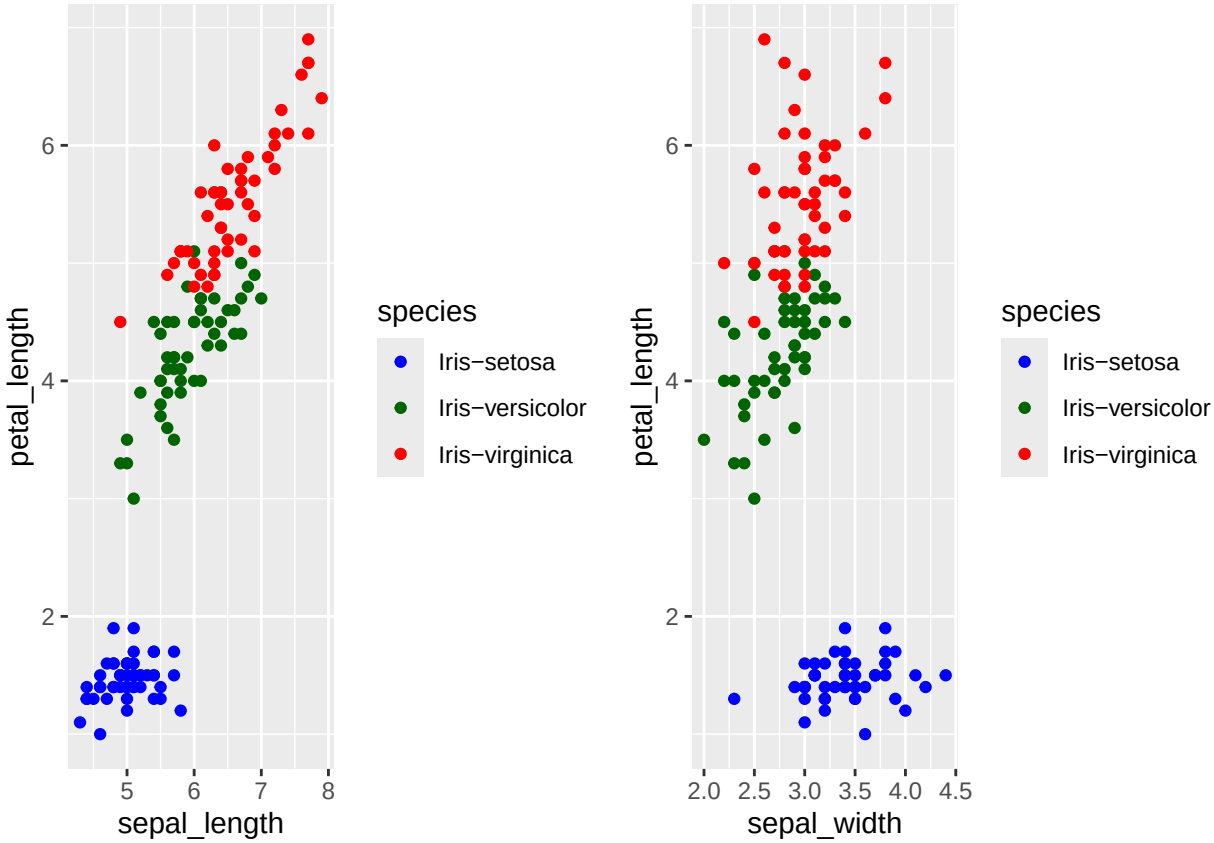
#### Scatterplots

We can observe from the scatterplot Fig 1 that sepal-length and petal length is closely positvly ocrrelated for Iris-versicolor and Iris-virginica and a slight positive correlation when looking at sepal-width and sepal-length. This is not the case however for Iris-setosa.

```
p1 <- ggplot(iris) +
  geom_point(aes(x = sepal_length, y = petal_length, color = species)) +
  scale_color_manual(values = c("Iris-setosa" =  "blue", "Iris-versicolor" = "darkgreen", "Iris-virginic

p2 <- ggplot(iris) +
  geom_point(aes(x = sepal_width, y = petal_length, color = species)) +
  scale_color_manual(values = c("Iris-setosa" =  "blue", "Iris-versicolor" = "darkgreen", "Iris-virginic

grid.arrange(p1, p2, ncol = 2)
```




### Boxplots

The boxplots in Fig 2 shows us the spread of the data for each species and attribute (sepal-width, -length, petal-width, -length). We se that for sepal-width Iris-setosa stands out by being wider but also having more variability. For sepal-length, petal-width, petal-length we observe a hierarchical order from smallest to largest Iris-setosa, -versicolor, -virginica.

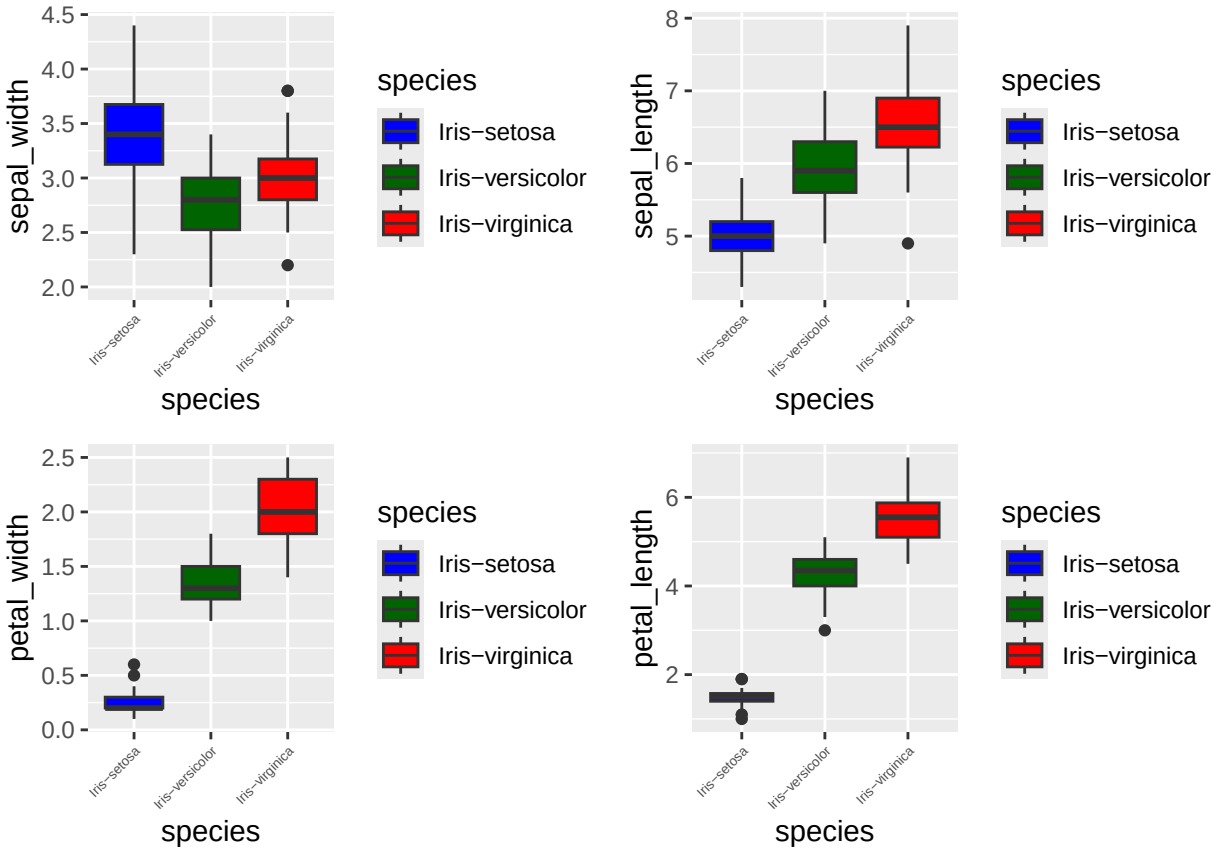
```
p1 <- ggplot(iris) +
  geom_boxplot(aes(x = species, y = sepal_width, fill = species)) +
  scale_fill_manual(values = c("Iris-setosa" =  "blue", "Iris-versicolor" = "darkgreen", "Iris-virginica
  theme(axis.text.x = element_text(size = 5, angle = 45, hjust = 1))

p2 <- ggplot(iris) +
  geom_boxplot(aes(x = species, y = sepal_length, fill = species)) +
  scale_fill_manual(values = c("Iris-setosa" =  "blue", "Iris-versicolor" = "darkgreen", "Iris-virginica
  theme(axis.text.x = element_text(size = 5, angle = 45, hjust = 1))

p3 <- ggplot(iris) +
  geom_boxplot(aes(x = species, y = petal_width, fill = species)) +
  scale_fill_manual(values = c("Iris-setosa" =  "blue", "Iris-versicolor" = "darkgreen", "Iris-virginica
  theme(axis.text.x = element_text(size = 5, angle = 45, hjust = 1))

p4 <- ggplot(iris) +
  geom_boxplot(aes(x = species, y = petal_length, fill = species)) +
  scale_fill_manual(values = c("Iris-setosa" =  "blue", "Iris-versicolor" = "darkgreen", "Iris-virginica
  theme(axis.text.x = element_text(size = 5, angle = 45, hjust = 1))
```

```
grid.arrange(p1, p2, p3, p4, nrow = 2, ncol = 2)
```

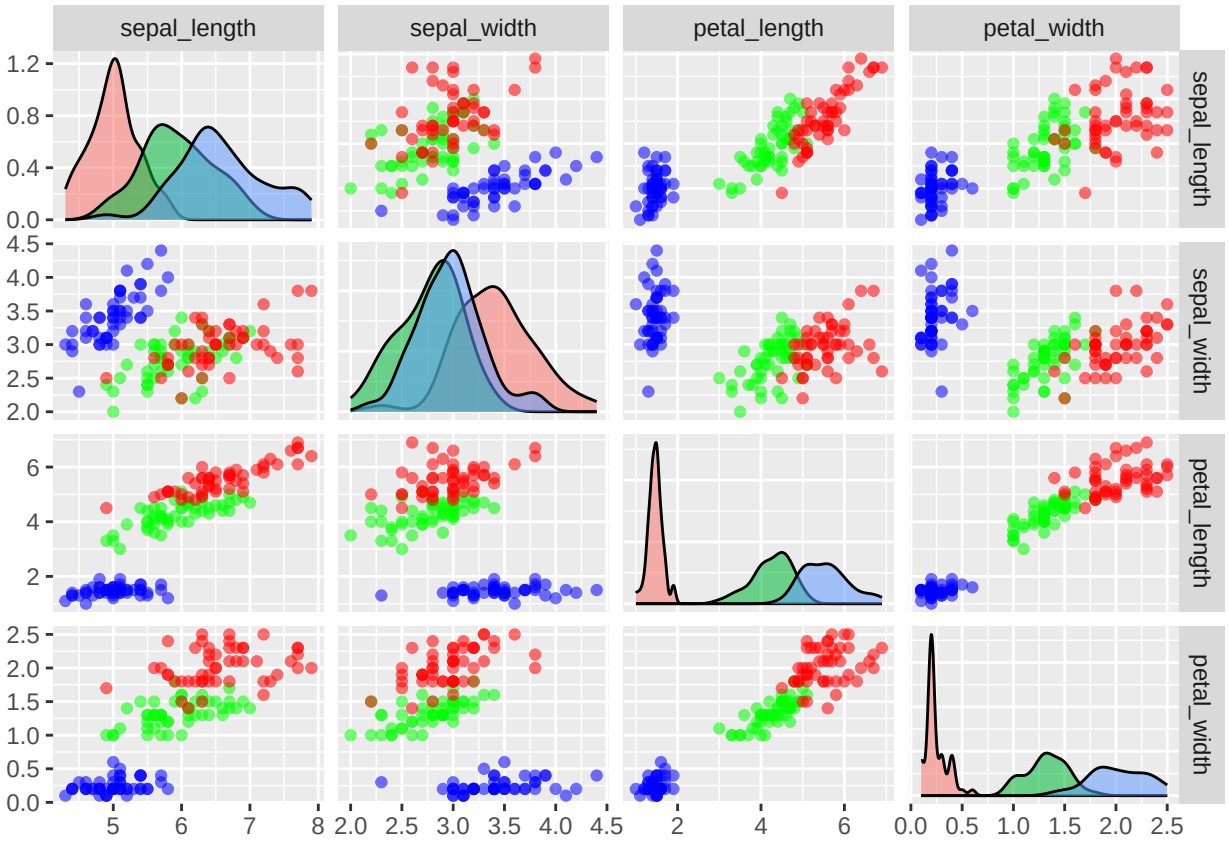



### Pairs plot

The pairs plot gives us an overview how the different variables of the data are related.

```
ggpairs(iris[, -5],
        aes(color = iris$species, alpha = 0.5),
        upper = list(continuous = "points")) +
          scale_color_manual(values = c("Iris-setosa" = "blue",
                                        "Iris-versicolor" = "green",
                                        "Iris-virginica" = "red"))
```

```
## Warning: No shared levels found between `names(values)` of the manual scale and the
## data's colour values.
## No shared levels found between `names(values)` of the manual scale and the
## data's colour values.
## No shared levels found between `names(values)` of the manual scale and the
## data's colour values.
## No shared levels found between `names(values)` of the manual scale and the
## data's colour values.
```




## Highest count species

```
summary(artportalen)
```

```
##        Id             Taxonsorteringsordning  Rödlistade
##  Min.   : 97785066   Min.   :53905          Length:21916
##  1st Qu.: 98957872   1st Qu.:54222          Class :character
##  Median : 99796691   Median :54944          Mode  :character
##  Mean   : 99812517   Mean   :54752
##  3rd Qu.:100764332   3rd Qu.:55213
##  Max.   :101507783   Max.   :55488
##    Artnamn           Vetenskapligt.namn    Auktor             Antal
##  Length:21916       Length:21916       Length:21916       Length:21916
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Ålder.stadium          Kön             Aktivitet          Lokalnamn
##  Length:21916       Length:21916       Length:21916       Length:21916
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
```

```
##
##
##   Ostkoordinat     Nordkoordinat      Noggrannhet       Diffusion
##  Min.   :1622830   Min.   :6579964   Min.   :   0.0   Min.   :0
##  1st Qu.:1626851   1st Qu.:6580712   1st Qu.:  94.0   1st Qu.:0
##  Median :1630050   Median :6581857   Median : 187.0   Median :0
##  Mean   :1629865   Mean   :6583355   Mean   : 288.5   Mean   :0
##  3rd Qu.:1633107   3rd Qu.:6585409   3rd Qu.: 250.0   3rd Qu.:0
##  Max.   :1635000   Max.   :6590025   Max.   :4679.0   Max.   :0
##      Län               Kommun            Provins           Församling
##  Length:21916       Length:21916       Length:21916       Length:21916
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##   Startdatum          Starttid          Slutdatum            Sluttid
##  Length:21916       Length:21916       Length:21916       Length:21916
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##   Kommentar            Biotop           Rapportör         Observatörer
##  Length:21916       Length:21916       Length:21916       Length:21916
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
```

```
artportalen <- artportalen %>%
  mutate(Antal = as.numeric(Antal))
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `Antal = as.numeric(Antal)`.
## Caused by warning:
## ! NAs introduced by coercion
```

```
# 5 most prevalent species
artportalen %>%
  group_by(Vetenskapligt.namn) %>%
  summarize(total = sum(Antal, na.rm = TRUE)) %>%
  arrange(desc(total)) %>%
  slice(1:5) %>%
  kable()
```

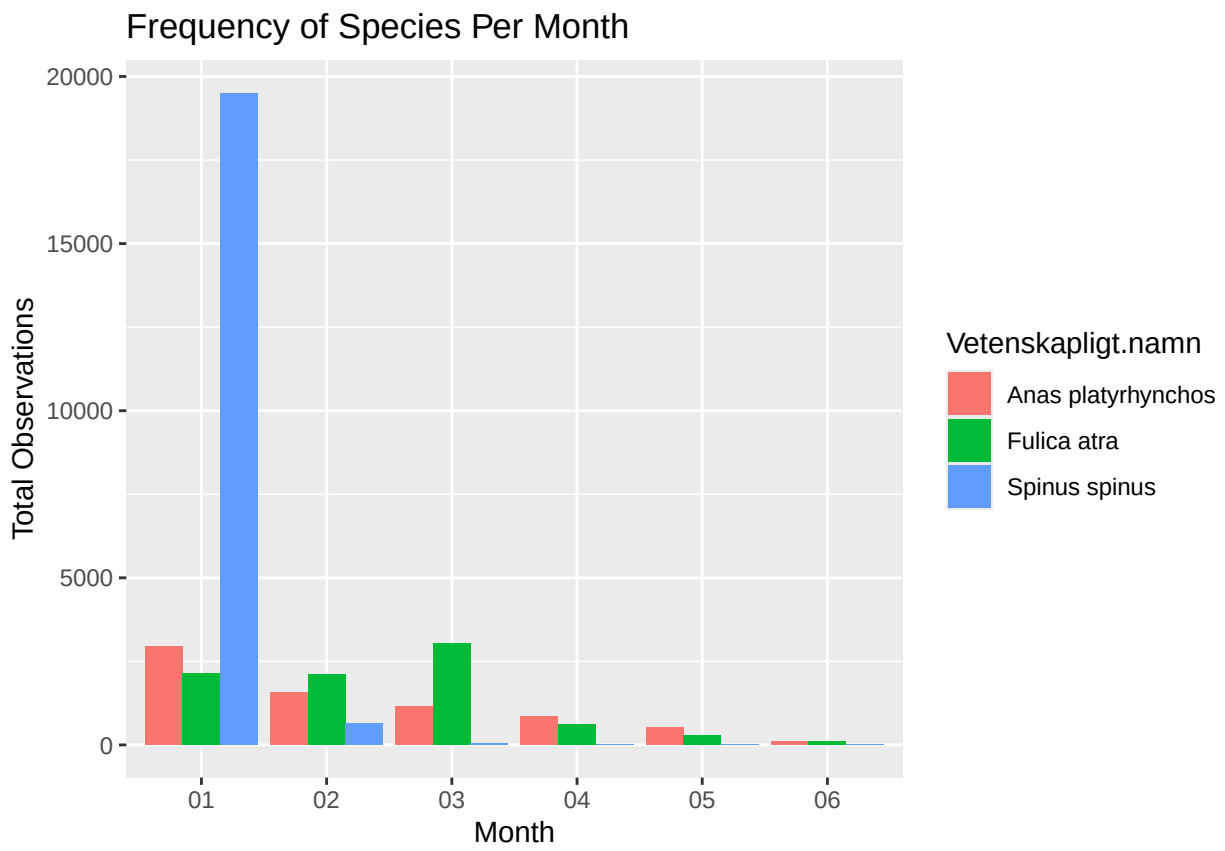
| Vetenskapligt.namn | total |
|---|---|
| Spinus spinus | 20211 |
| Fulica atra | 8308 |
| Anas platyrhynchos | 7167 |
| Mergus merganser | 6750 |

| Vetenskapligt.namn | total |
|---|---|
| Branta leucopsis | 6345 |

```
# Monthly distribution of the 3 most prevalent species
dist_art <- artportalen %>%
  filter(Vetenskapligt.namn %in% c("Spinus spinus", "Fulica atra", "Anas platyrhynchos")) %>%
  mutate(month = format(as.Date(Startdatum), "%m")) %>% # creates new variable month
  group_by(Vetenskapligt.namn, month) %>%
  summarize(total = sum(Antal, na.rm = TRUE))
```

```
## `summarise()` has grouped output by 'Vetenskapligt.namn'. You can override
## using the `.groups` argument.
```

```
ggplot(dist_art, aes(x = month, y = total, fill = Vetenskapligt.namn)) +
  geom_bar(stat = "identity", position = "dodge") +  # Use "dodge" to make bars side-by-side
  labs(
    title = "Frequency of Species Per Month",
    x = "Month",
    y = "Total Observations"
  )
```




```
# 5 rarest species
artportalen %>%
```

```
  group_by(Vetenskapligt.namn) %>%
  summarize(total = sum(Antal, na.rm = TRUE)) %>%
  filter(total == 1) %>%
  kable()
```

| Vetenskapligt.namn | total |
|---|---|
| Ardea alba | 1 |
| Buteo buteo buteo | 1 |
| Buteo lagopus | 1 |
| Calcarius lapponicus | 1 |
| Carpodacus erythrinus | 1 |
| Cinclus cinclus | 1 |
| Dryocopus martius | 1 |
| Falco tinnunculus | 1 |
| Gavia arctica | 1 |
| Lanius excubitor | 1 |
| Linaria flavirostris | 1 |
| Loxia bifasciata | 1 |
| Lullula arborea | 1 |
| Milvus milvus | 1 |
| Motacilla cinerea | 1 |
| Pernis apivorus | 1 |
| Somateria mollissima | 1 |
| Tadorna tadorna | 1 |
| Tringa totanus | 1 |

**Data**

```
cell_phone <- read.csv("cell_phones_total.csv")

cell_t <- cell_phone %>%
  # This section transforms k,M,B in to numerical values acordingly
  mutate(across(-1,
                ~ ifelse(grepl("k", .), as.numeric(gsub("k", "", .)) * 1e3, .))) %>%
  mutate(across(-1,
                ~ ifelse(grepl("M", .), as.numeric(gsub("M", "", .)) * 1e6, .))) %>%
  mutate(across(-1,
                ~ ifelse(grepl("B", .), as.numeric(gsub("B", "", .)) * 1e9, .))) %>%
  # Transforms the empty chr in to NA
  mutate(across(where(is.character), ~ na_if(., ""))) %>%
  # Transforms chr 0 in to dbl 0
  mutate(across(where(is.character), ~ ifelse(. == "0", 0, .))) %>%
  # Transforms logical values in to dbl.
  mutate(across(where(is.logical), ~ as.numeric(.)))
```

```
## Warning: There were 33 warnings in `mutate()`.
## The first warning was:
## i In argument: `across(...)`.
## Caused by warning in `ifelse()`:
## ! NAs introduced by coercion
```

```
## i Run `dplyr::last_dplyr_warnings()` to see the 32 remaining warnings.

## Warning: There were 8 warnings in `mutate()`.
## The first warning was:
## i In argument: `across(...)`.
## Caused by warning in `ifelse()`:
## ! NAs introduced by coercion
## i Run `dplyr::last_dplyr_warnings()` to see the 7 remaining warnings.
```

```
# View(cell_t)
# View(cell_phone)
cell_tibble <- as_tibble(cell_t)
cell_tibble
```

```
## # A tibble: 214 x 57
##    iso.3 X1960 X1965 X1966 X1967 X1968 X1969 X1970 X1971 X1972 X1973 X1974 X1975
##    <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1 ABW       0     0    NA    NA    NA    NA     0    NA    NA    NA    NA     0
##  2 AFG       0     0    NA    NA    NA    NA     0    NA    NA    NA    NA     0
##  3 AGO       0     0    NA    NA    NA    NA     0    NA    NA    NA    NA     0
##  4 ALB       0     0    NA    NA    NA    NA     0    NA    NA    NA    NA     0
##  5 AND       0     0    NA    NA    NA    NA     0    NA    NA    NA    NA     0
##  6 ARE      NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
##  7 ARG       0     0    NA    NA    NA    NA     0    NA    NA    NA    NA     0
##  8 ARM       0     0    NA    NA    NA    NA     0    NA    NA    NA    NA     0
##  9 ASM       0     0    NA    NA    NA    NA     0    NA    NA    NA    NA     0
## 10 ATG       0     0    NA    NA    NA    NA     0    NA    NA    NA    NA     0
## # i 204 more rows
## # i 44 more variables: X1976 <dbl>, X1977 <dbl>, X1978 <dbl>, X1979 <dbl>,
## #   X1980 <chr>, X1981 <chr>, X1982 <chr>, X1983 <chr>, X1984 <chr>,
## #   X1985 <chr>, X1986 <chr>, X1987 <chr>, X1988 <chr>, X1989 <chr>,
## #   X1990 <chr>, X1991 <chr>, X1992 <chr>, X1993 <chr>, X1994 <chr>,
## #   X1995 <chr>, X1996 <chr>, X1997 <chr>, X1998 <chr>, X1999 <chr>,
## #   X2000 <chr>, X2001 <chr>, X2002 <chr>, X2003 <chr>, X2004 <chr>, ...
```