# project

## Table of contents

```r
library(tidyverse)
library(ggplot2)
library(reshape2)
library(dplyr)
library(tidyr)
library(MASS)  # For Box-Cox
library(moments)  # For skewness/kurtosis
library(ggfortify)
library(viridis)  # For a professional color palette
library(ggpubr)   # For boxplots    # For normality tests and transformations
library(ggthemes) # For better themes
library(kableExtra)
library(GGally)
```

# 1 Introduction

# 2 Setting up our relevante Dataset:

```r
data_loc = "mc.csv"
data <- read.csv(data_loc)
#-3 values represent null values so we assing NA
data[data == -3] <- NA
#subset = styria -> NUTS2 = AT22
data <- data |> filter(xnuts2 == 22)
#filtering only for the relevant Predictors
data <- data |>dplyr::select(werr, dseitz, dstd, kjahr, xanzkind, xminalt,
            balt5, bsex,bfst, xbstaat, xbgeblan, xhatlevel, xeinw, xlfi,xpatch)
head(data)
```

```
  werr dseitz dstd kjahr xanzkind xminalt balt5 bsex bfst xbstaat xbgeblan
1    3     17   30    26       NA      NA     7    1    4       1        1
2    6     NA   NA    50        0      25    11    1    2       1        1
3    6     NA   NA    47        0      25    10    2    2       1        1
4    8     NA   NA    69        1      24    14    2    3       1        1
5    8     NA   NA    45        1      24    10    2    1       1        1
6    4     NA   NA    49       NA      NA    14    1    2       1        1
  xhatlevel xeinw xlfi xpatch
1        32     4    1     NA
```

```
2         32      1     3      NA
3         32      1     3      NA
4         21      2     3      NA
5         21      2     3      NA
6         51      4     3      NA
```

```
data <- data %>%
    mutate(
    werr = factor(werr, levels = 1:8,
                  labels = c("before 1919", "1919-1944", "1945-1960",
                             "1961-1970", "1971-1980", "1981-1990",
                             "1991-2000", "after 2000")),

    balt5 = factor(balt5, levels = 0:15,
                   labels = c("0-14", "15-19", "20-24", "25-29", "30-34", "35-39",
                              "40-44", "45-49", "50-54", "55-59", "60-64", "65-69",
                              "70-74", "75-79", "80-84", "85+")),

    bsex = factor(bsex, levels = c(1, 2), labels = c("Male", "Female")),

    bfst = factor(bfst, levels = 1:4,
                  labels = c("Single", "Married", "Widowed", "Divorced")),

    xbstaat = factor(xbstaat, levels = 1:7,
                     labels = c("Austria", "EU15 without Austria", "EU15 10 new members",
                                "Former Yugoslavia", "Turkey", "Other countries", "Bulgaria/

    xbgeblan = factor(xbgeblan, levels = 1:7,
                      labels = c("Austria", "EU15 without Austria", "EU15 10 new members",
                                 "Former Yugoslavia", "Turkey", "Other countries", "Bulgaria

    xhatlevel = factor(xhatlevel, levels = c(0, 11, 21, 22, 30, 31, 32, 41, 42, 43, 51, 52,
                       labels = c("ISCED 0/1", "ISCED 1", "ISCED 2", "ISCED 3c <2 years",
                                  "ISCED 3", "ISCED 3c 2+ years", "ISCED 3a, b", "ISCED 4a,
                                  "ISCED 4c", "ISCED 4", "ISCED 5b", "ISCED 5a", "ISCED 6",

    xeinw = factor(xeinw, levels = 1:4,
                   labels = c("up to 2000", "2001-10000", "10001-100000", "100001+")),

    xlfi = factor(xlfi, levels = 1:3,
                  labels = c("Employed", "Unemployed", "Not in labor force")),
```

```r
    xpatch = factor(xpatch, levels = c(1, 2), labels = c("Yes", "No"))
  )
data <- na.omit(data)

head(data)
```

```
          werr dseitz dstd kjahr xanzkind xminalt balt5   bsex    bfst xbstaat
10 after 2000      4   30    30        2      13 50-54   Male Married Austria
11 after 2000    156   70    15        2      13 45-49 Female Married Austria
14  1991-2000     74   38    31        2      14 45-49   Male  Single Austria
15  1991-2000    216   34    24        2      14 40-44 Female  Single Austria
16  1991-2000     11   40     3        2      14 20-24 Female  Single Austria
21  1981-1990     12   39     3        1      20 20-24   Male  Single Austria
   xbgeblan   xhatlevel      xeinw     xlfi xpatch
10  Austria ISCED 3a, b    100001+ Employed     No
11  Austria    ISCED 5a    100001+ Employed     No
14  Austria ISCED 3a, b up to 2000 Employed     No
15  Austria ISCED 3a, b up to 2000 Employed     No
16  Austria ISCED 3a, b up to 2000 Employed     No
21  Austria ISCED 3a, b 2001-10000 Employed     No
```

```r
data.numeric <-c("dseitz","dstd","kjahr","xanzkind")
data.polytomous <- c("balt5","bfst","xbstaat","xbgeblan","xhatlevel","xeinw","xlfi")
data.categorical <- c("balt5","bsex","bfst","xbstaat","xbgeblan","xhatlevel","xeinw","xlfi","
```

# 3 Descriptice Analysis

## 3.1 Numeric Variables

### 3.1.1 dseitz: working in the current job since. . . (in months)

```
summary(data$dseitz)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    0.0    29.5    95.5   127.6   187.5   469.0 
```

```
sd(data$dseitz)
```

```
[1] 117.1715
```

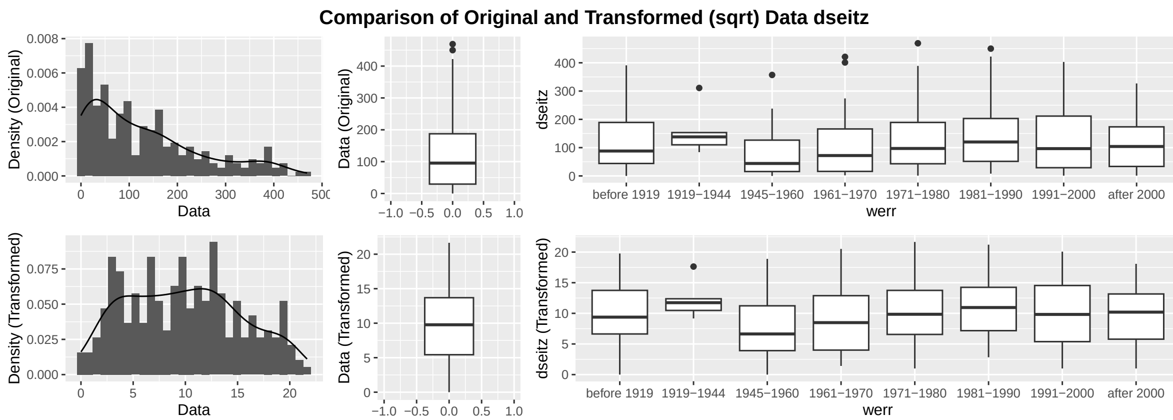
Spread & Variability: Ranges from 0 to 469, with a high standard deviation (117.17), indicating significant dispersion. Central Tendency: Median = 95.5, Mean = 127.6 (higher than the median), suggesting right-skewness. Quartiles: Q1 = 29.5, Q3 = 187.5, with a large IQR (158), showing a wide spread. Shape: The high max (469) and right-skewed distribution suggest potential outliers.

```
#|warning: false
plot_numeric_variable(data,
  "dseitz","werr","dseitz: working in the current job since. . . (in months)")
```


The histogram shows that the variable follows a right-skewed distribution and has a high spread. Through the boxplot we see that most values are between 0 and 200 with some outliers above 400. The box plot categorized by the buildings-Year shows that the distribution of dseitz differs across the categories. Additionally we see some outliers, but none of the seem to strongly influence the mean of the category, except 1919-1944 which shows a mean skewed towards the outlier.

The skeweness and the different distributions throughout the categories might indicate that a transformation would help to conform more to a normal-distribution.

```
plot_numeric_variable_with_transformation(data,"dseitz","werr")
```




After applying the Square root transformation we can see in the boxplot that the data is now more organised around the median value. Comparing the individual medians in the categories of the target variable shows some improvements in the distribution. There is still one outlier in the 1919-1944 category which might be good to remove.

```
# Perform Shapiro-Wilk test before transformation
shapiro_before <- shapiro.test(data$dseitz)
data_sqrt <- sqrt(data$dseitz)
shapiro_after <- shapiro.test(data_sqrt)

# Create a formatted table of test results
shapiro_results <- data.frame(
  Test = c("Original Data", "Square Root Transformed"),
  W_Statistic = c(shapiro_before$statistic, shapiro_after$statistic),
  P_Value = c(shapiro_before$p.value, shapiro_after$p.value)
)
kable(shapiro_results, caption = "Shapiro-Wilk Normality Test Results", digits = 5)
```

Table 1: Shapiro-Wilk Normality Test Results

| Test | W_Statistic | P_Value |
|---|---|---|
| Original Data | 0.88551 | 0e+00 |
| Square Root Transformed | 0.97086 | 4e-05 |

Using the Shapiro-Wilk Normality Test we can show that we improved the normality of the data by some degree by comparing the W-Value. By transforming we were able to increase the W-value closer to 1.

While building the model, we will try the transformend variable as well and see how it influences the performance.

### 3.1.2 xminalt age of youngest child in the family (in years)

```r
summary(data$xminalt)
```

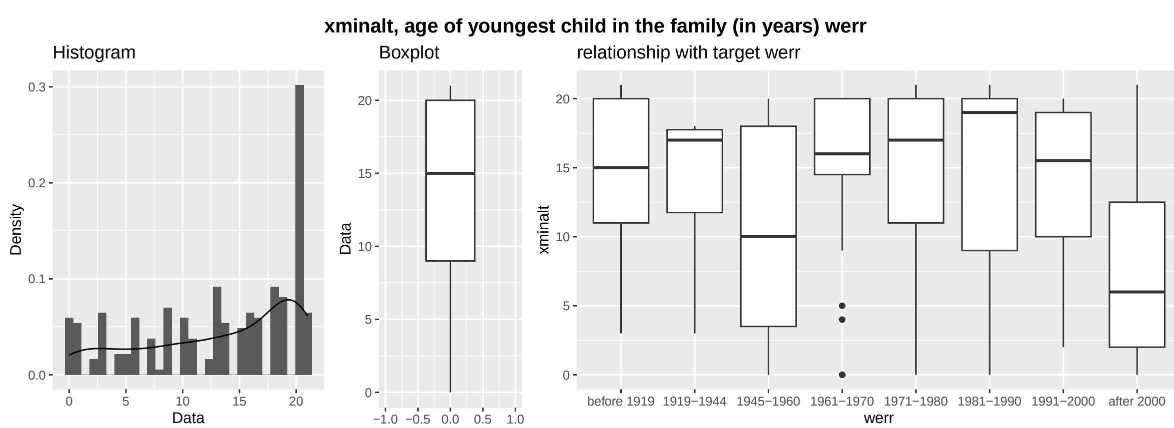
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    9.00   15.00   13.24   20.00   21.00
```

```r
sd(data$xminalt)
```

```
[1] 6.682862
```

The variable shows a wide spread, high variability, and a right-skewed distribution. The presence of a very high maximum value (469) compared to Q3 (187.5) suggests possible outliers.

```r
plot_numeric_variable(data, "xminalt","werr","xminalt, age of youngest child in the family (i
```

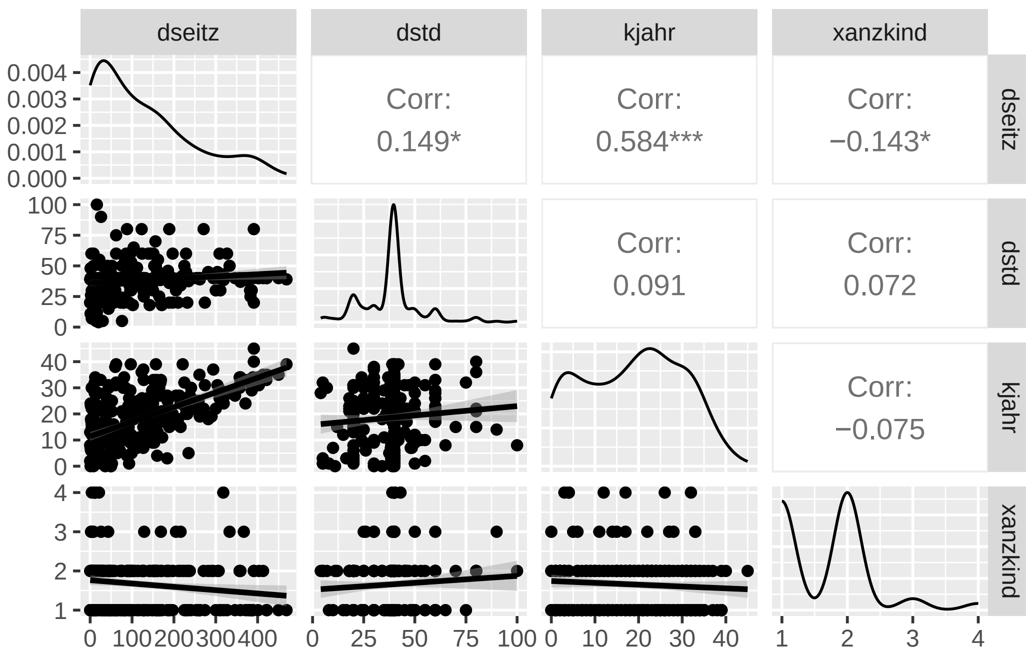



The histogram shows a left-skewed distribution, with many values concentrated at the higher end (close to 20 years). The boxplot confirms this skewness, as the median is closer to the upper quartile. A sharp spike at 20 years suggests a possible ceiling effect or rounding in data collection.

The whiskers extend further toward lower values, and some outliers appear at the lower end (0-5 years).

The boxplots by groups of werr indicate variations in child age distribution across different categories. Some groups (e.g., 1961-1970) show more lower-end outliers, while others (before 1919, 1981-1990) have higher medians. This suggests that the age of the youngest child may have a meaningful relationship with the target variable.

```r
pair_data <- data[,data.numeric]
ggpairs(pair_data,
        lower = list(continuous = "smooth"), # Smoothed scatterplots on the lower panel
        diag = list(continuous = "densityDiag"), # Density plots on the diagonal
        upper = list(continuous = "cor")) # Add correlation coefficients on the upper panel
```



### 3.1.3 Relationship between numerical and Categorical variables

#### 3.1.3.1 dseitz ~ balt5 Relation ship between dseitz: working in the current job since. . . (in months) and age category"

```
library(knitr)

# Compute count and percentage
freq_table <- data %>%
  group_by(balt5) %>%
  summarise(Count = n()) %>%
  mutate(Percentage = round((Count / sum(Count)) * 100, 1))  # Calculate percentage

freq_table_wide <- freq_table %>%
  pivot_longer(cols = c(Count, Percentage),
               names_to = "Metric",
               values_to = "Value") %>%
  pivot_wider(names_from = balt5, values_from = Value) %>%
  dplyr::select(Metric, everything())  # Ensure 'Metric' is first

# Print the table as a kable
kable(freq_table_wide, format = "html", caption = "Age Group Count and Percentage")
```

Table 2: Age Group Count and Percentage

| Metric | 15-19 | 20-24 | 25-29 | 30-34 | 35-39 | 40-44 | 45-49 | 50-54 | 55-59 | 60-64 |
|---|---|---|---|---|---|---|---|---|---|---|
| Count | 15.0 | 28.0 | 25.0 | 19.0 | 34.0 | 44.0 | 39.0 | 38.0 | 13.0 | 1.0 |
| Percentage | 5.9 | 10.9 | 9.8 | 7.4 | 13.3 | 17.2 | 15.2 | 14.8 | 5.1 | 0.4 |

counts <- data %>% group_by(balt5) %>% summarise(count = n()) # Histogram with density line relationship_plot <- ggplot(data, aes_string(x = "balt5", y = "dseitz")) + geom_boxplot() + labs( x= "",y ="dseitz") + # ggtitle("relationship between dseitz and balt5") + theme_grey() + scale_colour_grey() + scale_fill_grey()

violinplot <- ggplot(data, aes(x = balt5, y = dseitz)) + geom_violin()+ theme_grey() + theme(plot.margin = margin(b = 0))+ scale_colour_grey() + scale_fill_grey()

plot <- ggarrange(relationship_plot, violinplot, ncol = 1, nrow = 2, heights = c(0.5,0.5)) # Add title to the combined plot plot_with_title <- annotate_figure( plot, top = text_grob(paste("Relation ship between dseitz: working in the current job since. . . (in months) and age category"), face = "bold", size = 14))

plot_with_title

```
The median dseitz (job tenure) increases as age increases, which is expected since older empl
Young employees tend to switch jobs more frequently or have just entered the workforce, leadi
Middle-aged employees exhibit the most variation in tenure, possibly due to career shifts, pr
Older employees generally have longer tenures and less variation, indicating job stability an

This suggests that it might be beneficial to regroup the age groups into three bigger catego


#### **dseitz ~ bfst** Relation ship between dseitz: working in the current job since. . . (
::: {.cell}

```{.r .cell-code}
library(knitr)

# Compute count and percentage
freq_table <- data %>%
  group_by(bfst) %>%
  summarise(Count = n()) %>%
  mutate(Percentage = round((Count / sum(Count)) * 100, 1))  # Calculate percentage

freq_table_wide <- freq_table %>%
  pivot_longer(cols = c(Count, Percentage),
               names_to = "Metric",
               values_to = "Value") %>%
  pivot_wider(names_from = bfst, values_from = Value) %>%
  dplyr::select(Metric, everything())  # Ensure 'Metric' is first

# Print the table as a kable
kable(freq_table_wide, format = "html", caption = "family status Count and Percentage")
```

Table 3: family status Count and Percentage

| Metric | Single | Married | Widowed | Divorced |
|---|---|---|---|---|
| Count | 93.0 | 142.0 | 2.0 | 19.0 |
| Percentage | 36.3 | 55.5 | 0.8 | 7.4 |

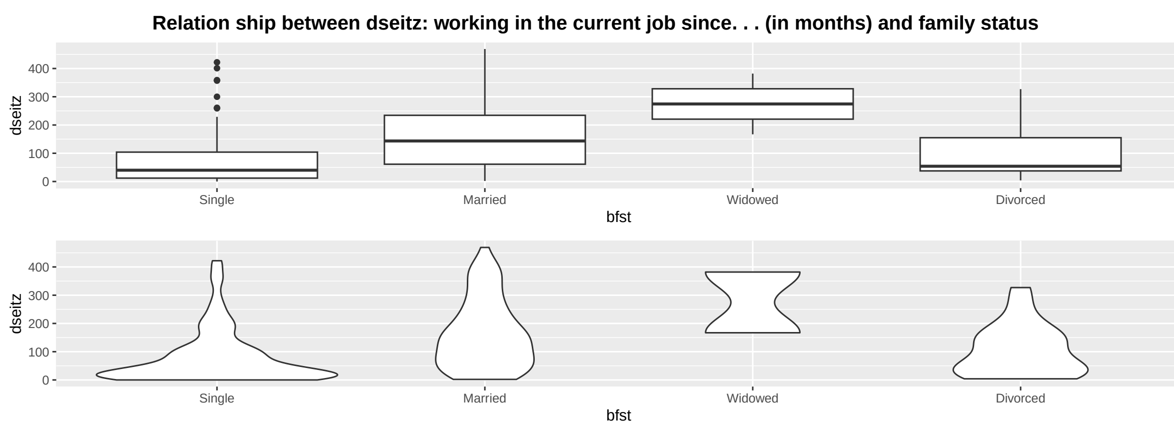
:::
```

```
# Histogram with density line
relationship_plot <- ggplot(data, aes_string(y = "dseitz", x = "bfst")) +
    geom_boxplot() +
    labs(y = "dseitz" , x = "bfst") +
   # ggtitle("relationship between dseitz and balt5") +
    theme_grey() +
    scale_colour_grey() +
    scale_fill_grey()

violinplot <- ggplot(data, aes(x = bfst, y = dseitz)) +
  geom_violin()+
  theme_grey() +
  scale_colour_grey() +
  scale_fill_grey()

plot <- ggarrange(relationship_plot, violinplot, ncol = 1, nrow = 2, widths = c(0.5,0.5))
# Add title to the combined plot
  plot_with_title <- annotate_figure(
  plot,
  top = text_grob(paste("Relation ship between dseitz: working in the current job since. . .
  face = "bold", size = 14))

plot_with_title
```



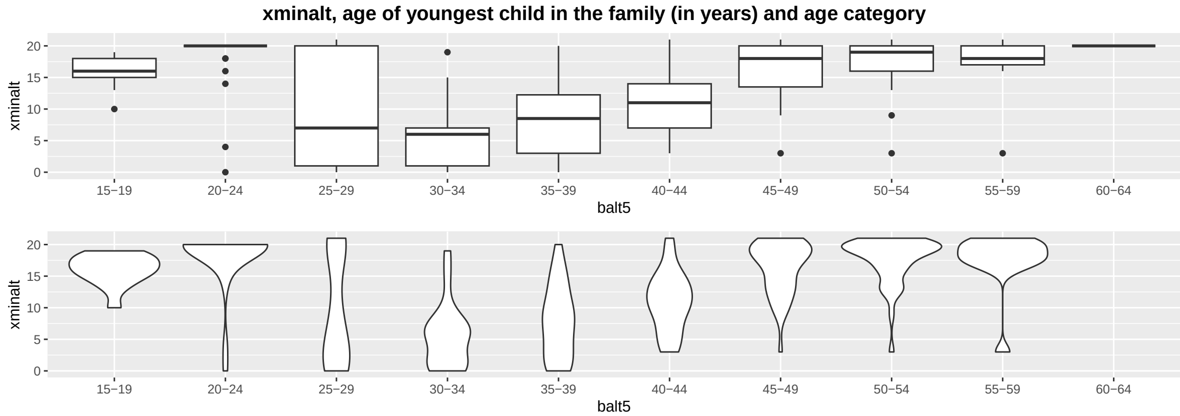

#### 3.1.3.2 'xminalt ~ bfst age of youngest child in the family (in years) and age category"

```r
# Histogram with density line
relationship_plot <- ggplot(data, aes_string(y = "xminalt", x = "balt5")) +
    geom_boxplot() +
    labs(y = "xminalt" , x = "balt5") +
   # ggtitle("relationship between dseitz and balt5") +
    theme_grey() +
    scale_colour_grey() +
    scale_fill_grey()

violinplot <- ggplot(data, aes(x = balt5, y = xminalt)) +
  geom_violin()+
  theme_grey() +
  scale_colour_grey() +
  scale_fill_grey()

plot <- ggarrange(relationship_plot, violinplot, ncol = 1, nrow = 2, widths = c(0.5,0.5))
# Add title to the combined plot
  plot_with_title <- annotate_figure(
  plot,
  top = text_grob(paste("xminalt, age of youngest child in the family (in years) and age cate
  face = "bold", size = 14))

plot_with_title
```




The higer the age group, the older the jungest child gets with some outlier. We see that it makes sence to combine some

#### 3.1.3.3 xminalt ~ bfsft

```
# Histogram with density line
relationship_plot <- ggplot(data, aes_string(y = "xminalt", x = "bfst")) +
    geom_boxplot() +
    labs(y = "xminalt" , x = "bfst") +
   # ggtitle("relationship between dseitz and balt5") +
    theme_grey() +
    scale_colour_grey() +
    scale_fill_grey()

violinplot <- ggplot(data, aes(x = bfst, y = xminalt)) +
  geom_violin()+
  theme_grey() +
  scale_colour_grey() +
  scale_fill_grey()

plot <- ggarrange(relationship_plot, violinplot, ncol = 1, nrow = 2, widths = c(0.5,0.5))
# Add title to the combined plot
  plot_with_title <- annotate_figure(
  plot,
  top = text_grob(paste("xminalt, age of youngest child in the family (in years) and family s
  face = "bold", size = 14))

plot_with_title
```