Multivariate statistics - Microcensus report

Lukas Aichhorn, Johannes Gölles

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)
library(vcd)
```

1 Introduction

In this report, we focus on analyzing the factors influencing the construction year of buildings across different time periods. The dataset used for this analysis consists exclusively of households from Styria, Austria's 2nd largest state behind Lower Austria.

The goal of this report is to explore the relationship between various socio-economic factors and the construction year of buildings (werr). In order to achieve this, we selected four predictors: **job tenure (dseitz)**, **age of the youngest child (xminalt)**, **age category (balt5)**, and **family status (bfst)**.

1.1 Motivation behind predictor selection

The four predictors were chosen to reflect key aspects of household dynamics that could potentially influence housing preferences and the time of construction.

Job tenure captures the stability of the household's employment situation, which may be linked to the decision-making process around housing.

The age of the youngest child provides insights into the family's stage in the life cycle, which could affect housing preferences, particularly regarding the age of the home.

Age category reflects the demographic profile of household members, while **family status** provides an indication of the household's familial structure, which may impact both the choice of residence and the likelihood of living in newer or older buildings.

By analyzing these predictors, we aim to uncover patterns that shed light on how various demographic and economic factors influence housing choices in Styria.

1.2 Research questions and hypotheses

This section formulates the research questions and hypotheses that guide our analysis of how job tenure, family structure, and age-related characteristics influence the likelihood of living in buildings from different construction periods. Given that our data is limited to households in Styria, we specifically examine the individual and combined effects of job tenure (dseitz), the age of the youngest child (xminalt), age category (balt5), and family status (bfst) on the construction period of residential buildings (werr).

1.2.1 1st Research question

'How do job tenure (dseitz), the age of the youngest child (xminalt), age category (balt5), and family status (bfst) collectively influence the likelihood of Styrians living in buildings from different construction periods (werr)?'

This question examines how all predictors contribute to determining the construction year of the building, considering potential interactions between job tenure, family structure, and age.

1.2.2 Hypotheses 1st research question

Null Hypothesis (H): The predictors dseitz, xminalt, balt5, and bfst, both individually and in combination, have no significant effect on the construction period of the building (werr).

Alternative Hypothesis (H): At least one of the predictors or their interactions has a significant effect on the construction period of the building (werr).

1.2.3 2nd Research question

'Does the interaction between family status (bfst), age category (balt5), the age of the youngest child (xminalt), and family status (bfst) influence the likelihood of Styrian households living in buildings constructed before or after 2000 (werr)?'

This question focuses on how our predictors interact to shape housing choices, particularly around the boundary of the year 2000, which could be a key cutoff in housing trends.

1.2.4 Hypotheses 2nd research question

Null Hypothesis (H): The predictors dseitz, xminalt, balt5, and bfst, both individually and in combination, do not significantly influence the likelihood of living in houses older or newer than 25 years (werr = 2000).

Alternative Hypothesis (H): At least one of the predictors or their interactions significantly affect the likelihood of living in a house older or younger than 25 years (werr = 2000).

1.3 Starting point

Understanding the factors that influence residential choices and housing conditions is essential for assessing demographic and socioeconomic trends. This study investigates whether job tenure, family structure, and age-related factors play a role in determining the construction period of residential buildings in Styria. By analyzing these relationships, we aim to uncover potential patterns that could inform housing policy and urban development strategies.

1.4 Regression method

For this analysis, we will use ordinal logistic regression since the response variable, werr (construction period of the building), is ordinal in nature. The categories represent sequential time periods, making an ordinal model more appropriate than a standard multinomial logistic regression, which would disregard the inherent ordering.

Ordinal logistic regression allows us to estimate how the predictors—job tenure (dseitz), age of the youngest child (xminalt), age category (balt5), and family status (bfst)—affect the likelihood of living in buildings from different construction periods while preserving the ordinal structure of the dependent variable. This method also enables the inclusion of interactions to examine whether combined demographic factors influence housing conditions differently.

Alternative approaches, such as binary logistic regression, would require collapsing the categories into two groups, leading to information loss, while Poisson regression is not suitable since the response variable is categorical rather than a count.

2 Data collection

2.1 Type of survey

The data originates from a structured household survey conducted in Styria, focusing on demographic, employment, and housing characteristics. The survey follows a standardized questionnaire format, ensuring consistency across responses. Participants provided information

about their family composition, employment history, and housing conditions, allowing for a comprehensive analysis of potential influences on residential buildings' construction periods. According to **'Statistics Austria'**, every quarter, 22 500 households are selected for the survey. They are randomly drawn from the Central Residence Register (ZMR). Within ten years a private household can be surveyed in up to five consecutive calendar quarters.

(https://www.statistik.at/en/about-us/surveys/individual-and-household-surveys/microcensus#:~:text=The%

2.2 Description of data set

```
data_loc = "mc.csv"
data <- read.csv(data_loc)

str(data)
```

```
'data.frame':   9287 obs. of  27 variables:
 $ werr     : int  3 4 4 3 3 3 7 7 7 7 ...
 $ wanzw    : int  3 1 1 3 3 3 2 2 2 2 ...
 $ wm2      : int  2 4 4 2 2 2 5 5 5 5 ...
 $ wanzr    : int  3 8 8 3 3 3 7 7 7 7 ...
 $ wheiz    : int  1 2 2 2 2 2 3 3 3 3 ...
 $ wrecht   : int  3 1 1 6 6 6 1 1 1 1 ...
 $ wkges    : int  2 1 1 1 1 1 1 1 1 1 ...
 $ wvertr   : int  3 NA NA NA NA NA NA NA NA NA ...
 $ wfrist   : int  1 -3 -3 -3 -3 -3 -3 -3 -3 -3 ...
 $ bhhgr    : int  1 2 2 3 3 3 5 5 5 5 ...
 $ dseitz   : int  17 4 21 308 35 -3 -3 303 -3 -3 ...
 $ dstd     : num  30 40 45 40 20 -3 -3 40 -3 -3 ...
 $ kjahr    : int  26 40 12 28 30 1 28 27 4 2 ...
 $ xanzkind : int  -3 1 1 1 1 1 2 2 2 2 ...
 $ xminalt  : int  -3 21 21 15 15 15 16 16 16 16 ...
 $ balt5    : int  7 10 3 7 6 1 7 7 1 1 ...
 $ bsex     : int  1 2 2 1 2 2 2 1 1 1 ...
 $ bfst     : int  4 4 1 2 2 1 2 2 1 1 ...
 $ xbstaat  : int  1 1 1 1 6 1 1 1 1 1 ...
 $ xbgeblan : int  1 1 1 1 6 1 1 1 1 1 ...
 $ xhatlevel: int  32 51 32 32 21 21 32 32 21 21 ...
 $ xeinw    : int  4 4 4 2 2 2 2 2 2 2 ...
 $ xlfi     : int  1 1 1 1 1 3 3 1 3 3 ...
 $ xpatch   : int  -3 -3 -3 2 2 2 2 2 2 2 ...
 $ xurb     : int  1 1 1 2 2 2 3 3 3 3 ...
```

```
 $ xnuts1   : int  2 1 1 3 3 3 2 2 2 2 ...
 $ xnuts2   : int  22 13 13 33 33 33 21 21 21 21 ...
```

As stated in the introduction, this report is based on Microcensus data, which originally consists of 27 variables stored as integers and 9,287 rows. While all variables are initially represented numerically, closer examination reveals that many are categorical in nature (e.g., bsex, xbstaat). In the next section, all variables will be converted to their appropriate data types.

2.3 Data preparation

The first step of this report was the data preparation. Since NA values in the microcensus dataset were represented as '-3' we replaced them with NA in order to be able to later remove them using na.omit().

Additionally we filtered the data to only contain entries from our selected region Styria. Furthermore the categorical variables were transformed to factors. Our target variable((werr) and one of our predictors (balt5) were modified to incorporate the ordinality of their respective values.

```
#-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)

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, 6
                       labels = c("ISCED 0/1", "ISCED 1", "ISCED 2", "ISCED 3c <2 years",
                                  "ISCED 3", "ISCED 3c 2+ years", "ISCED 3a, b", "ISCED 4a, b
                                  "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")),

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

summary(data)
```

```
         werr        dseitz           dstd            kjahr
 1991-2000 :68   Min.   :  0.0   Min.   :  4.00   Min.   : 0.00
 1971-1980 :41   1st Qu.: 29.5   1st Qu.: 30.00   1st Qu.: 9.00
 1961-1970 :40   Median : 95.5   Median : 40.00   Median :20.00
 after 2000:39   Mean   :127.6   Mean   : 38.27   Mean   :18.56
 1981-1990 :26   3rd Qu.:187.5   3rd Qu.: 40.00   3rd Qu.:27.00
 1945-1960 :19   Max.   :469.0   Max.   :100.00   Max.   :45.00
 (Other)   :23
    xanzkind        xminalt          balt5        bsex           bfst
 Min.   :1.000   Min.   : 0.00   40-44  :44   Male  :133   Single  : 93
 1st Qu.:1.000   1st Qu.: 9.00   45-49  :39   Female:123   Married :142
 Median :2.000   Median :15.00   50-54  :38                Widowed :  2
 Mean   :1.656   Mean   :13.24   35-39  :34                Divorced: 19
```

```
3rd Qu.:2.000   3rd Qu.:20.00   20-24  :28
Max.   :4.000   Max.   :21.00   25-29  :25
                                (Other):48
                xbstaat                     xbgeblan         xhatlevel
Austria             :245   Austria             :232   ISCED 3a, b:149
EU15 without Austria:  3   EU15 without Austria:  3   ISCED 2    : 31
EU15 10 new members :  2   EU15 10 new members :  2   ISCED 4a, b: 26
Former Yugoslavia   :  3   Former Yugoslavia   :  8   ISCED 5a   : 20
Turkey              :  1   Turkey              :  1   ISCED 5b   : 17
Other countries     :  1   Other countries     :  7   ISCED 6    :  7
Bulgaria/Romania    :  1   Bulgaria/Romania    :  3   (Other)    :  6
         xeinw                      xlfi     xpatch
up to 2000  :117   Employed          :256   Yes: 29
2001-10000  :114   Unemployed        :  0   No :227
10001-100000:  6   Not in labor force:  0
100001+     : 19
```

This shows a short summary across all variables. Later on we will focus on the chosen predictors and their relationship with the target variable.

3 Descriptice Analysis

In this chapter, we will conduct a detailed descriptive analysis of the chosen predictors, focusing on understanding their individual characteristics and the relationships between them. The analysis will include visualizations and numerical summaries to describe the data and highlight any distinctive features, such as trends, outliers, or group differences.

We will start by presenting univariate visualizations for each variable, providing insights into their distribution and central tendencies. Additionally, bivariate relationships between the predictors and the response variable will be explored to understand how the former influence the latter. In particular, we will examine the interaction effects between pairs of predictors and the response variable, allowing us to identify any potential dependencies.

The visualizations will be complemented by detailed commentaries on all relevant statistics, focusing on distributions, and any note worthy findings from the plots.

Finally, we will summarize the key insights from the descriptive analysis. This will serve as the foundation for the more formal inferential analyses to follow.

3.1 Numeric Variables

This helper function was added to be able to create side by side plots for univariate and bivariate analysis of the two numerical predictors **xminalt** and **dseitz**.

```
# Define the function
plot_numeric_variable <- function(data, column_name, target_variable,plot_title) {

  # Histogram with density line
  hist_plot <- ggplot(data, aes_string(x = column_name)) +
    geom_histogram(aes(y = ..density..), bins = 30) +
    geom_density() +
    labs(x = "Data", y = "Density") +
    ggtitle("Histogram") +
    theme_grey() +
    scale_colour_grey()

  # Boxplot
  boxplot <- ggplot(data, aes_string(y = column_name)) +
    geom_boxplot() +
    xlim(-1, 1) +
    labs(y = "Data") +
    ggtitle("Boxplot") +
    theme_grey() +
```

```
    scale_colour_grey() +
    scale_fill_grey()

  # Relationship between the numeric variable and the categorical target variable
  relationship_plot <- ggplot(data, aes_string(x = target_variable, y = column_name)) +
    geom_boxplot() +
    labs(x = target_variable, y = column_name) +
    ggtitle("relationship with target werr") +
    theme_grey() +
    scale_colour_grey() +
    scale_fill_grey()

  # Arrange all three plots in one row
  plot <- ggarrange(hist_plot, boxplot, relationship_plot, ncol = 3, nrow = 1, widths = c(0.5

  # Add title to the combined plot
  plot_with_title <- annotate_figure(plot,
                                     top = text_grob(paste(plot_title, target_variable),
                                                     face = "bold", size = 14))

  # Return the plot with title
  return(plot_with_title)
}
```

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

The first variable we will examine is 'dseitz', which represents job tenure, measured in months. This variable provides insight into the length of time individuals have spent in their current employment.

```
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
```

The distribution of the data exhibits considerable spread and variability, with values ranging from 0 to 469 and a high standard deviation of 117.17, indicating significant dispersion. The central tendency is characterized by a median of 95.5 and a mean of 127.6, with the mean being higher than the median, suggesting a right-skewed distribution. The interquartile range (IQR) is quite large, spanning from 29.5 (Q1) to 187.5 (Q3), with an IQR of 158, further emphasizing the wide spread of the data. The presence of a high maximum value of 469 also indicates potential outliers, reinforcing the right-skewness of the distribution.

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

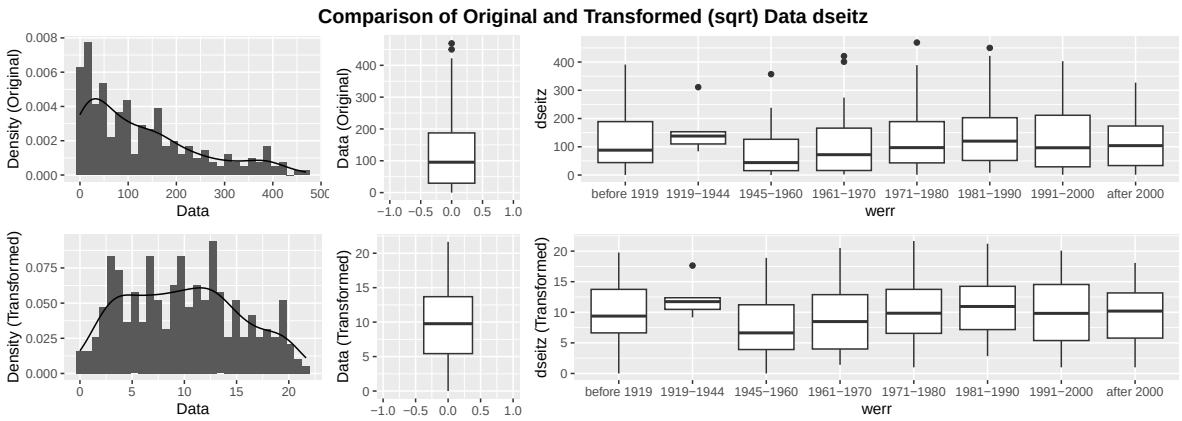
The histogram indicates a right-skewed distribution of the variable, accompanied by a broad spread. The boxplot reveals that most values of 'dseitz' lie between 0 and 200 months, with a few outliers extending beyond 400 months. When categorized by building construction year, the boxplot highlights that the distribution of 'dseitz' varies across the different categories. While some outliers are present, they do not significantly affect the mean in most categories. However, the 1919-1944 category shows a mean that is skewed due to an outlier. The observed skewness and differing distributions across categories suggest that a transformation might be beneficial to better approximate a normal distribution.

3.1.2 dseitz transformed (sqrt)

Out of curiosity and because it might later be useful in the part for creating the statistical model, we transformed the variable using a simple square root transformation to see the effect on the distribution.

Another helper function was written to be able to easily compare the plots before and after transformation.

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



After applying the square root transformation to the 'dseitz' variable, we observe a noticeable improvement in the distribution, as shown in the updated boxplot. The transformed data is now more concentrated around the median, with a tighter range of values. In comparison to the original boxplot, the transformation has reduced the high variability and the extreme skewness, bringing the data closer to a more symmetrical distribution.

When examining the medians across the categories of the target variable 'werr', the transformed data shows clearer and more consistent differences between categories, suggesting a more uniform distribution within each group. The individual medians appear more aligned, and the spread of values within each category has been reduced.

However, despite these improvements, there is still an outlier present in the 1919-1944 category, which stands apart from the rest of the data. This outlier may still influence the overall distribution and could be worth considering for removal, as it is likely distorting the mean and contributing to the skewness observed in this category even after the transformation. Overall, the transformation has helped normalize the data, but addressing the outlier might further improve the analysis.

```
# 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.

For the original data, the W-statistic is 0.88551, with a p-value of 0, which suggests that the original data deviates significantly from a normal distribution. The low W-statistic and extremely small p-value (essentially 0) indicate that the data is highly non-normal.

After applying the square root transformation, the W-statistic increases to 0.97086, and the p-value becomes 4e-05. Although the p-value remains small, indicating some deviation from normality, the W-statistic is considerably higher, reflecting an improvement in the distribution's symmetry and reduced skewness. The transformation has resulted in a distribution that is closer to normal, but still exhibits slight deviations, as indicated by the p-value. Therefore, while the square root transformation improves the normality of the data, the distribution may still not fully conform to a normal distribution.

3.1.3 xminalt age of youngest child in the family (in years)

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

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

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

```
[1] 6.682862
```

The variable 'xminalt', representing the age of the youngest child in the household, exhibits the following characteristics:

Range: The values span from 0 to 21 years, indicating that some households have infants while others have adult children still living at home.

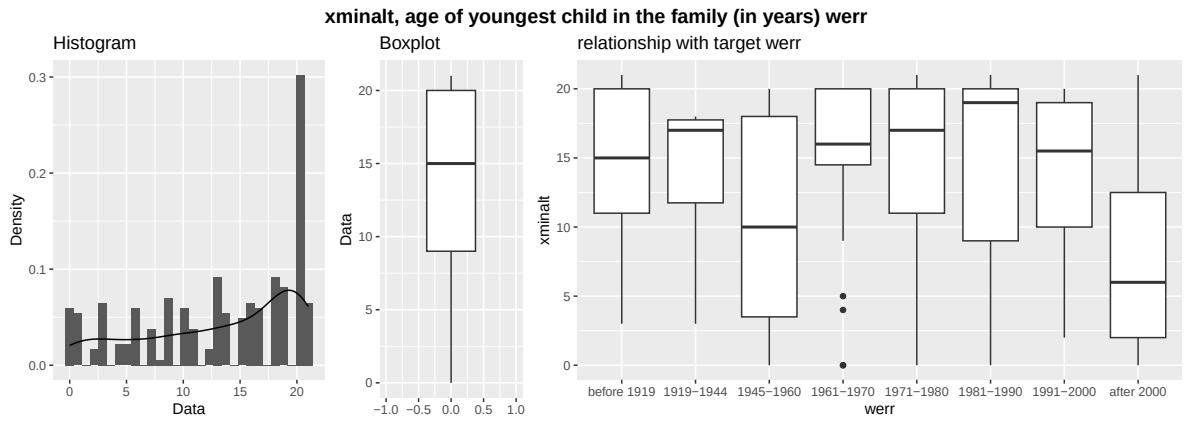
Central Tendency: The median age is 15 years, slightly higher than the mean of 13.24, suggesting a slight skew toward younger ages.

Quartiles: The first quartile (Q1) is 9, meaning that 25% of the households have their youngest child aged 9 or younger, while the third quartile (Q3) is 20, indicating that 75% of the youngest children are aged 20 or younger.

Spread: The standard deviation is 6.68, reflecting moderate variability in the distribution of ages.

The distribution suggests that most households have younger children, but there are some cases where the youngest child is already an adult.

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

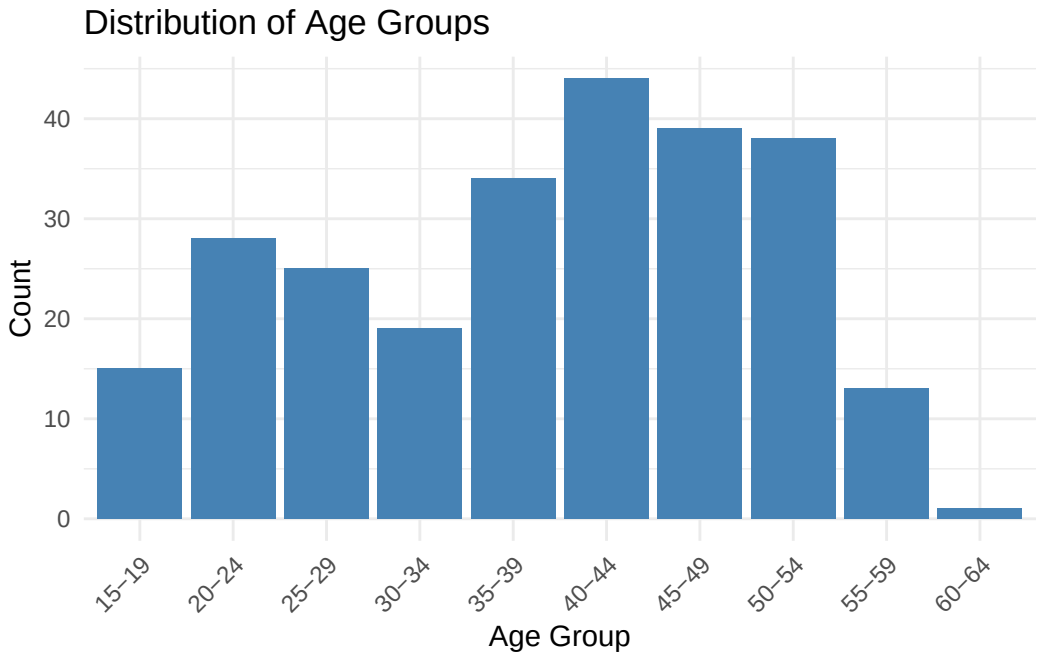


The histogram reveals a left-skewed distribution, with most values clustered near 20 years and a sharp spike at this upper limit, possibly due to rounding. The boxplot confirms this skewness, with the median near the upper quartile and whiskers extending further toward lower values, where some outliers (0–5 years) appear.

Boxplots by werr categories highlight differences in child age distribution, with some groups (e.g., 1961-1970) showing more lower-end outliers, while others (before 1919, 1981-1990) have higher medians. This suggests a potential relationship between xminalt and the construction period of the building.

3.2 Categorical Variables

```
# Bar plot of age group distribution
ggplot(data, aes(x = balt5)) +
  geom_bar(fill = "steelblue") +
  labs(title = "Distribution of Age Groups",
       x = "Age Group",
       y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate labels for readability
```



The bar plot displays the frequency distribution of individuals across age groups. The majority fall within the 40-44 and 45-49 age brackets, with 44 and 39 individuals, respectively. The distribution shows a peak in middle-aged groups, gradually declining in older categories. Notably, there are no individuals aged 70 and above, and only one person in the 60-64 range, indicating a strong skew toward younger and middle-aged populations. The lack of representation in the youngest (0-14) and oldest (70+) categories may suggest a dataset focused on working-age individuals.

```
summary(data$balt5)
```

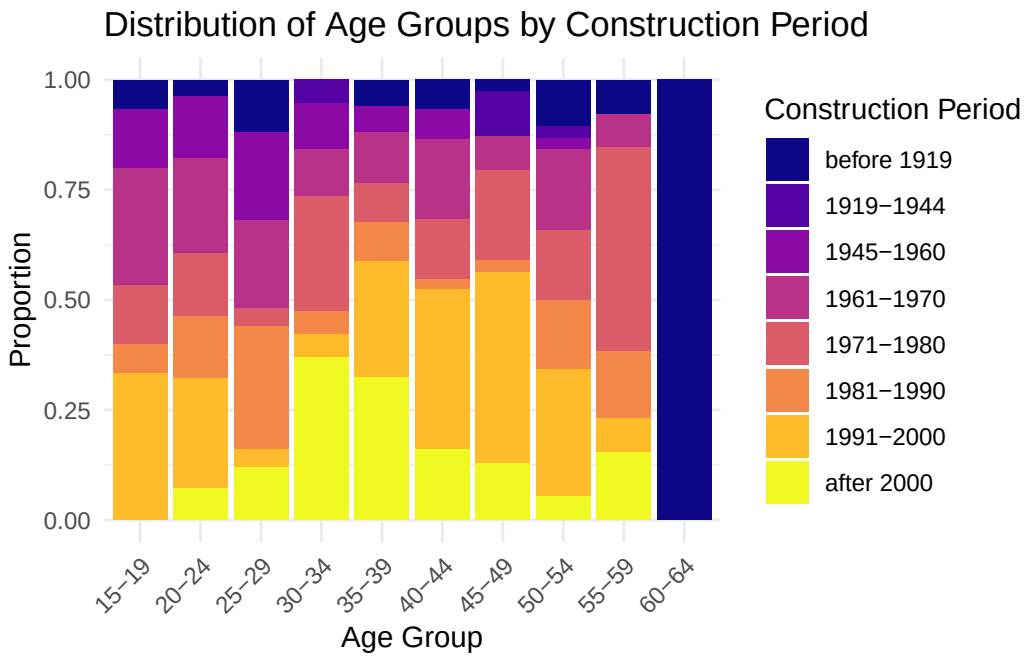
```
0-14 15-19 20-24 25-29 30-34 35-39 40-44 45-49 50-54 55-59 60-64 65-69 70-74
```

```
    0     15     28     25     19     34     44     39     38     13      1      0      0
75-79 80-84    85+
    0      0      0
```

```r
#dropping levels without any entries from the dataset
data$balt5 <- droplevels(data$balt5)
summary(data$balt5)
```

```
15-19 20-24 25-29 30-34 35-39 40-44 45-49 50-54 55-59 60-64
   15    28    25    19    34    44    39    38    13     1
```

```r
ggplot(data, aes(x = balt5, fill = werr)) +
  geom_bar(position = "fill") +  # Stacked proportionally
  scale_fill_viridis_d(option = "plasma") +  # Gradual color transition
  labs(title = "Distribution of Age Groups by Construction Period",
       x = "Age Group",
       y = "Proportion",
       fill = "Construction Period") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate labels for readability
```



```r
prop.table(table(data$balt5, data$werr), margin = 1)
```

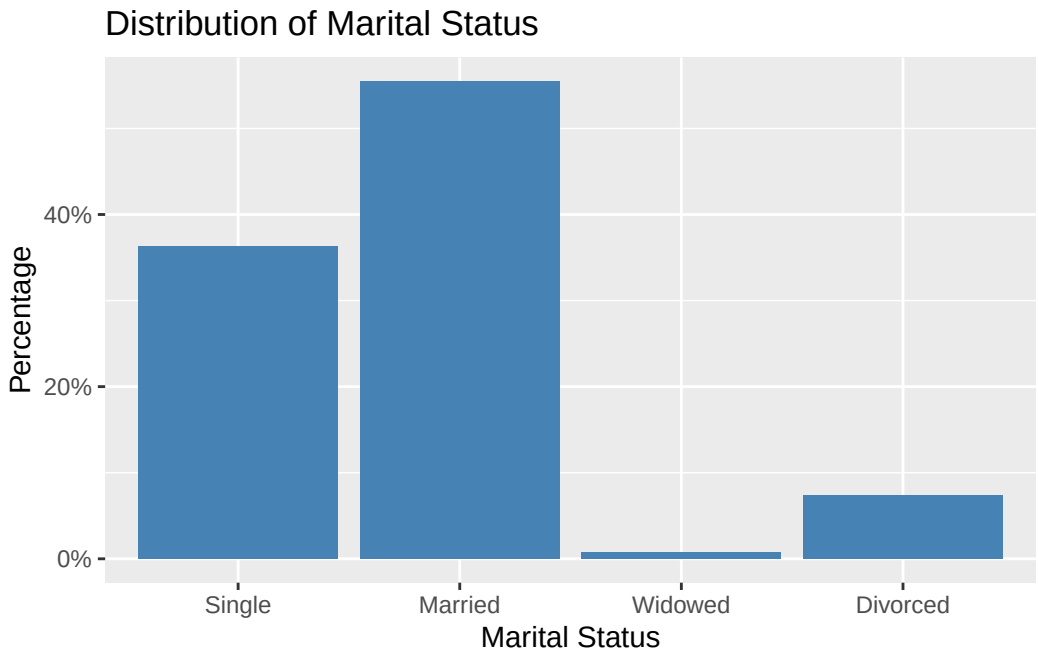
```
       before 1919  1919-1944  1945-1960  1961-1970  1971-1980  1981-1990
  15-19  0.06666667 0.00000000 0.13333333 0.26666667 0.13333333 0.06666667
  20-24  0.03571429 0.00000000 0.14285714 0.21428571 0.14285714 0.14285714
  25-29  0.12000000 0.00000000 0.20000000 0.20000000 0.04000000 0.28000000
  30-34  0.00000000 0.05263158 0.10526316 0.10526316 0.26315789 0.05263158
  35-39  0.05882353 0.00000000 0.05882353 0.11764706 0.08823529 0.08823529
  40-44  0.06818182 0.00000000 0.06818182 0.18181818 0.13636364 0.02272727
  45-49  0.02564103 0.10256410 0.00000000 0.07692308 0.20512821 0.02564103
  50-54  0.10526316 0.02631579 0.02631579 0.18421053 0.15789474 0.15789474
  55-59  0.07692308 0.00000000 0.00000000 0.07692308 0.46153846 0.15384615
  60-64  1.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000

       1991-2000 after 2000
  15-19 0.33333333 0.00000000
  20-24 0.25000000 0.07142857
  25-29 0.04000000 0.12000000
  30-34 0.05263158 0.36842105
  35-39 0.26470588 0.32352941
  40-44 0.36363636 0.15909091
  45-49 0.43589744 0.12820513
  50-54 0.28947368 0.05263158
  55-59 0.07692308 0.15384615
  60-64 0.00000000 0.00000000
```

This stacked barplot, along with the contingency table displaying proportions, illustrates the distribution of construction periods across the age groups with non-zero counts. It is evident that middle-aged individuals (between 30 and 40) disproportionately reside in newer houses, with around 35% living in houses built after 2000. Meanwhile, individuals over 40 are more likely to live in houses built in the 1990s or earlier. In contrast, people under thirty are spread across various construction periods, with no single group dominating—except among teenagers, where 33% reside in houses built in the 1990s. There is one single value in the group of 60-64 year old people which lives in a property built before 1919 which accounts for the 100% in this category.

```r
ggplot(data, aes(x = bfst)) +
  geom_bar(aes(y = (..count..) / sum(..count..) * 100), fill = "steelblue") +
  labs(title = "Distribution of Marital Status", x = "Marital Status", y = "Percentage") +
  theme_minimal() +
```

```
  theme_grey() +
  scale_y_continuous(labels = scales::percent_format(scale = 1))
```



As visible in this barplot the majority of surveyants were either single or married and only a rather small percentage of the people in the survey are widowed or divorced (below 10% combined). This could lead to false assumptions in the model because the two major groups are so dominant in the dataset.

3.2.1 Relationship between numerical and Categorical variables

3.2.1.1 dseitz ~ balt5 Relationship 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. . .
  face = "bold", size = 14))
```

```
plot_with_title
```

The median job tenure (dseitz) increases with age, as older employees have had more time to build tenure. Younger employees show lower and less variable tenures, likely due to frequent job changes or being new to the workforce. Middle-aged employees display the greatest variation, possibly reflecting career shifts or promotions, while older employees tend to have longer, more stable tenures. This pattern suggests that regrouping the age groups into three broader categories may be beneficial.

3.2.1.2 dseitz ~ bfst Relation ship between dseitz: working in the current job since. . . (in months) and family status")

```
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

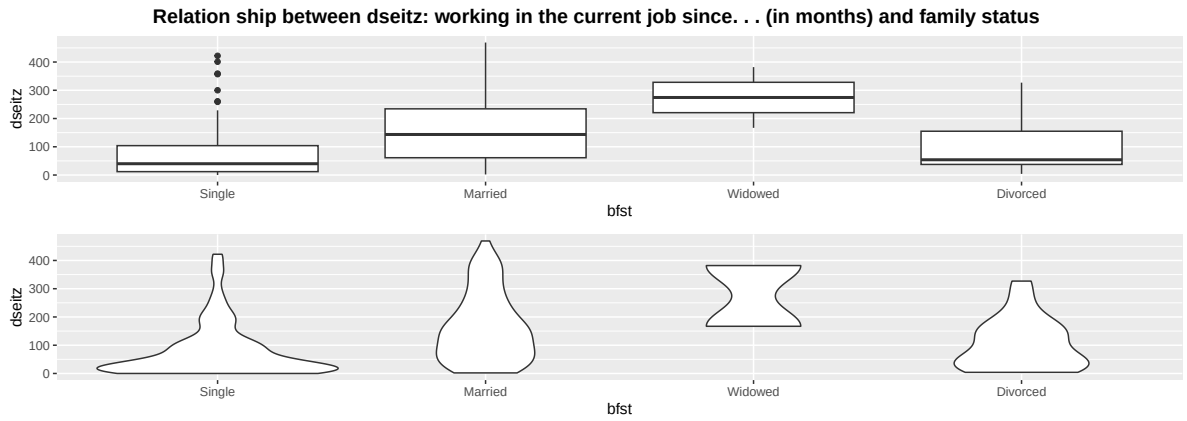
This table again shows the very skewed distribution for family status in this dataset with less than one percent being widowed (2 individuals) and the majority married or single.

```
# 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
```



This combination of boxplot and violin plot illustrates the relationship between job tenure (in months) and family status. It shows that single and divorced individuals generally have lower job tenure, while married—and to a lesser extent, widowed—individuals tend to have longer tenures. The plots reveal that married individuals exhibit a higher median tenure and a broader range, suggesting both long-term stability and recent entry into their positions, whereas the distributions for single and divorced groups are more concentrated at lower values with fewer long-term outliers.

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

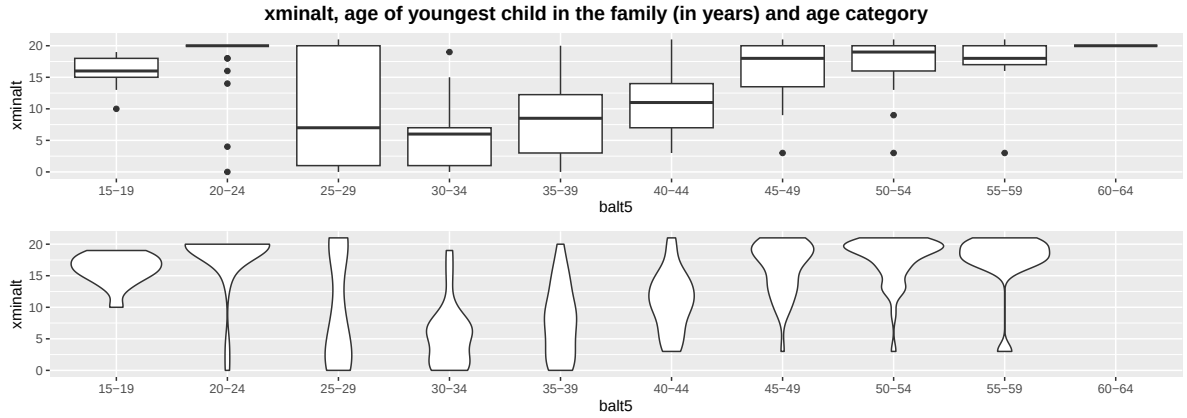
```
# Histogram with density line
relationship_plot <- ggplot(data, aes_string(y = "xminalt", x = "balt5")) +
    geom_boxplot() +
    labs(y = "xminalt" , x = "balt5") +
   # ggtitle("relationship between dseitz and balt5") +
    theme_grey() +
    scale_colour_grey() +
    scale_fill_grey()

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
```



This analysis explores the relationship between xminalt (age of the youngest child) and balt5 (age group of the parent). Younger parents (15-29 years) tend to have a wider spread of youngest child ages, with some having older children. Middle-aged parents (30-44 years) are more likely to have younger children, as indicated by lower medians. Older parents (50+ years) predominantly have older youngest children, with few cases of very young children. The violin plots suggest bimodal distributions in certain age groups (20-24, 50-54), indicating two distinct patterns.

Overall, the data aligns with expected parental age-child age trends, with younger parents having more variation and older parents having consistently older youngest children.

3.2.1.4 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
```

The violin plot displays the distribution of xminalt across different categories, excluding the "Widowed" group. The Single category shows a concentration of values primarily in the 15 to 20 range, indicating a peak in this region. In contrast, the Married category exhibits a high density of values at the upper end, followed by a gradual and even decline, suggesting a steady tapering of the distribution. The Widowed category has only values above 19, indicating a single-point distribution.

4 Joint influences

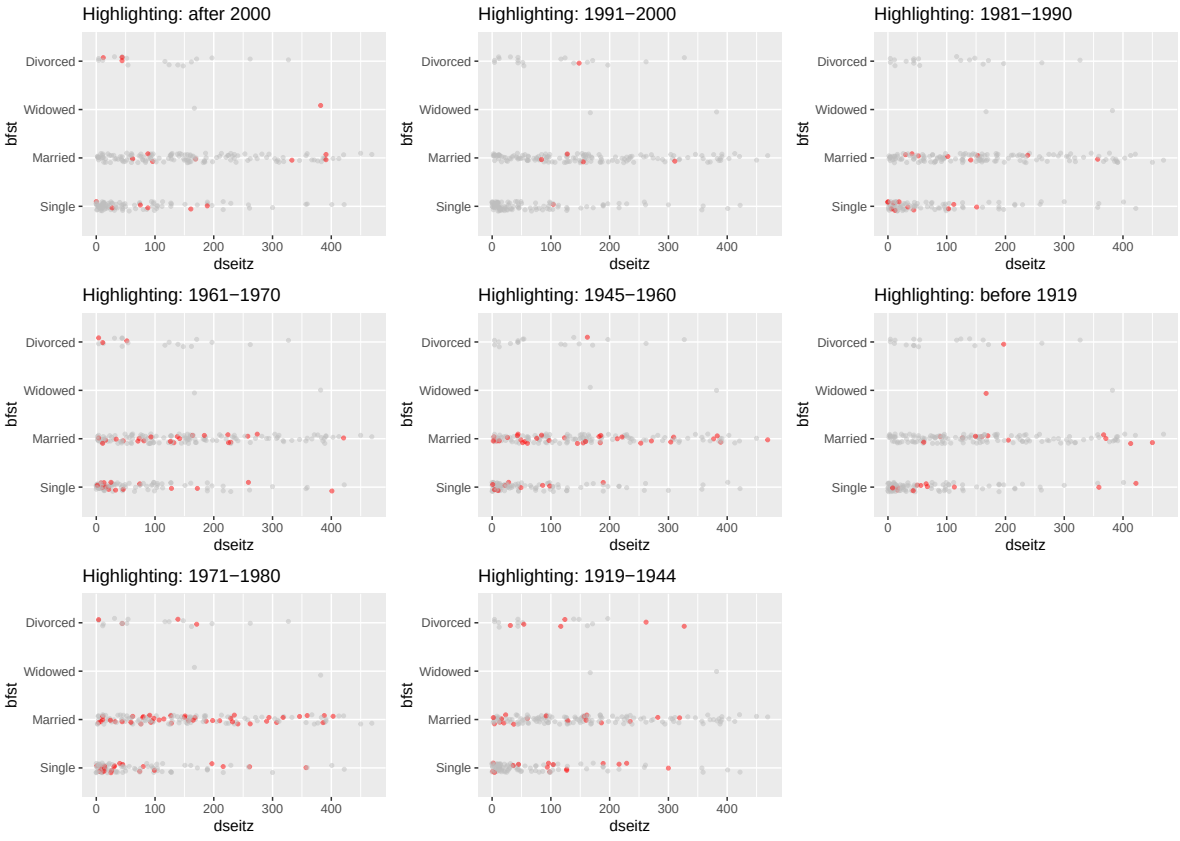
Each subplot highlights a different house construction period, showing how one categorical variable are distributed against a numeric one. The red points represent individuals whose houses were built in the corresponding period.

4.1 Joint influence of bfst ~ dseitz on the Target variable Werr

```
# Unique categories in 'werr'
categories <- unique(data$werr)

# Create a plot for each category
plots <- lapply(categories, function(cat) {
  ggplot(data, aes(x = dseitz, y = bfst, color = as.factor(werr))) +
    geom_jitter(width = 0.1, height = 0.1, size = 1,,alpha = 0.5) +
    scale_color_manual(values = ifelse(categories == cat, "red", "grey")) +
    ggtitle(paste("Highlighting:", cat)) +
    theme_grey()+
  theme(legend.position = "none")
})

# Print plots
ggarrange(plotlist = plots, ncol = 3, nrow = 3)
```



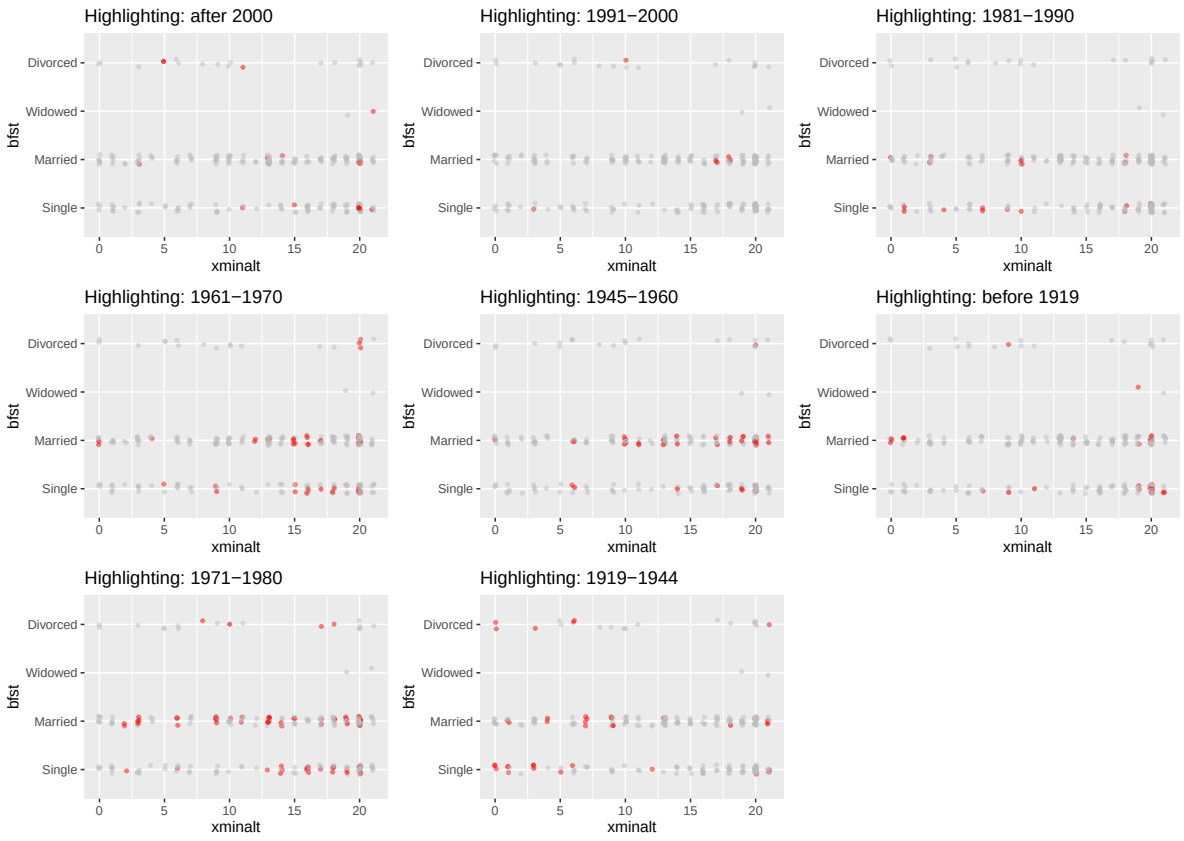
1971-1980 and 1945-1960 1961-1970 have a high concentration in being married or single with a wide spread especially in the Married category. 1981-1990 shows a value concentration in the lower left. All Other Target Categories show no distinct pattern of distribution.

4.2 Joint influence of bfst ~ xminalt on the Target variable Werr

```
# Ensure that the levels of werr are consistent with the facet variable
# Unique categories in 'werr'
categories <- unique(data$werr)

# Create a plot for each category
plots <- lapply(categories, function(cat) {
  ggplot(data, aes(x = xminalt, y = bfst, color = as.factor(werr))) +
    geom_jitter(width = 0.1, height = 0.1, size = 1,,alpha = 0.5) +
    scale_color_manual(values = ifelse(categories == cat, "red", "grey")) +
    ggtitle(paste("Highlighting:", cat)) +
    theme_grey()+
  theme(legend.position = "none")
})

# Print plots
ggarrange(plotlist = plots, ncol = 3, nrow = 3)
```



1961-1970, 1945-1960 and 1971-1980 show a light orientation to a higher min age, with more values being in the Married category. most observation in the 1919-1944 are in the lower range of xmin alt between 0 and 10.

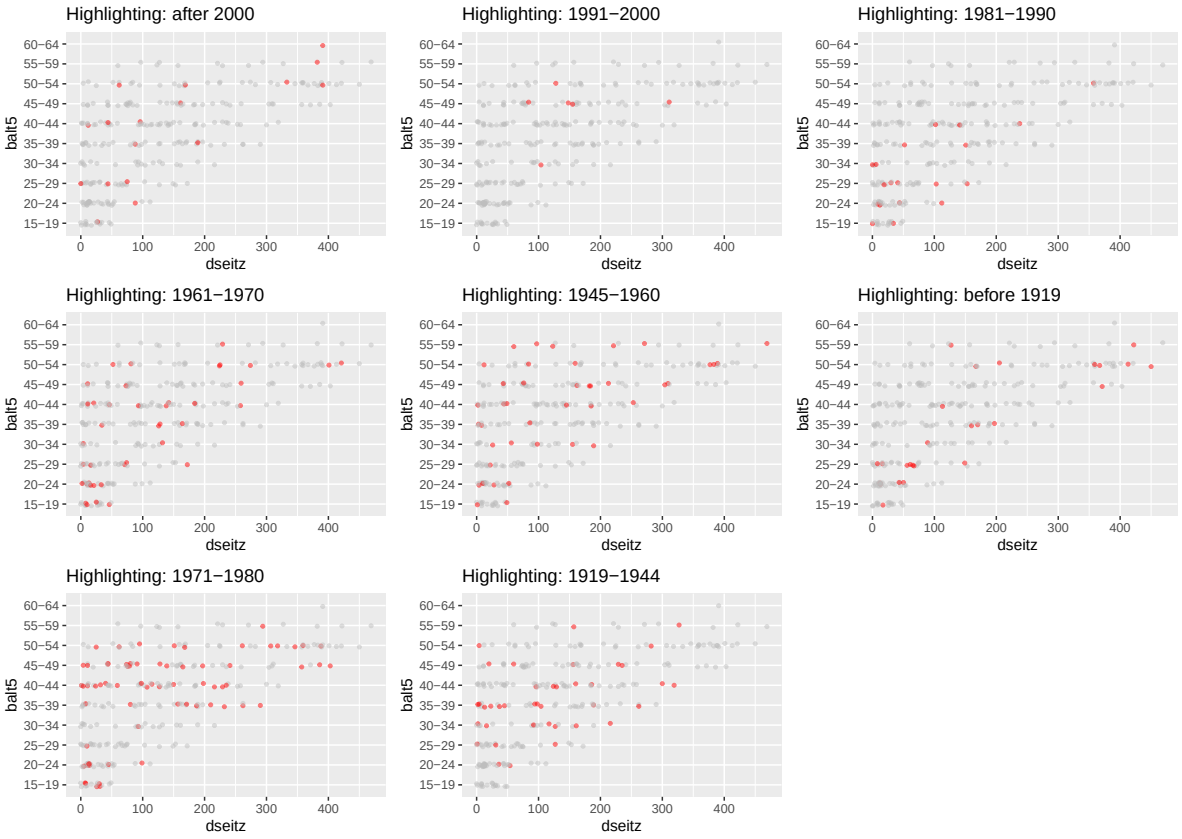
4.3 Joint influence of balt5 ~ dseitz on the Target variable Werr

```
# Ensure that the levels of werr are consistent with the facet variable
# Unique categories in 'werr'
categories <- unique(data$werr)

# Create a plot for each category
plots <- lapply(categories, function(cat) {
  ggplot(data, aes(x = dseitz, y = balt5, color = as.factor(werr))) +
    geom_jitter(width = 0.1, height = 0.1, size = 1,,alpha = 0.5) +
    scale_color_manual(values = ifelse(categories == cat, "red", "grey")) +
    ggtitle(paste("Highlighting:", cat)) +
    theme_grey()+
```

```
  theme(legend.position = "none")
})

# Print plots
ggarrange(plotlist = plots, ncol = 3, nrow = 3)
```



Pre-1919 buildings demonstrate pronounced clustering of long tenures (300-450 months) among older residents (50-64), indicating potential non-linear relationship between building age and tenure length.

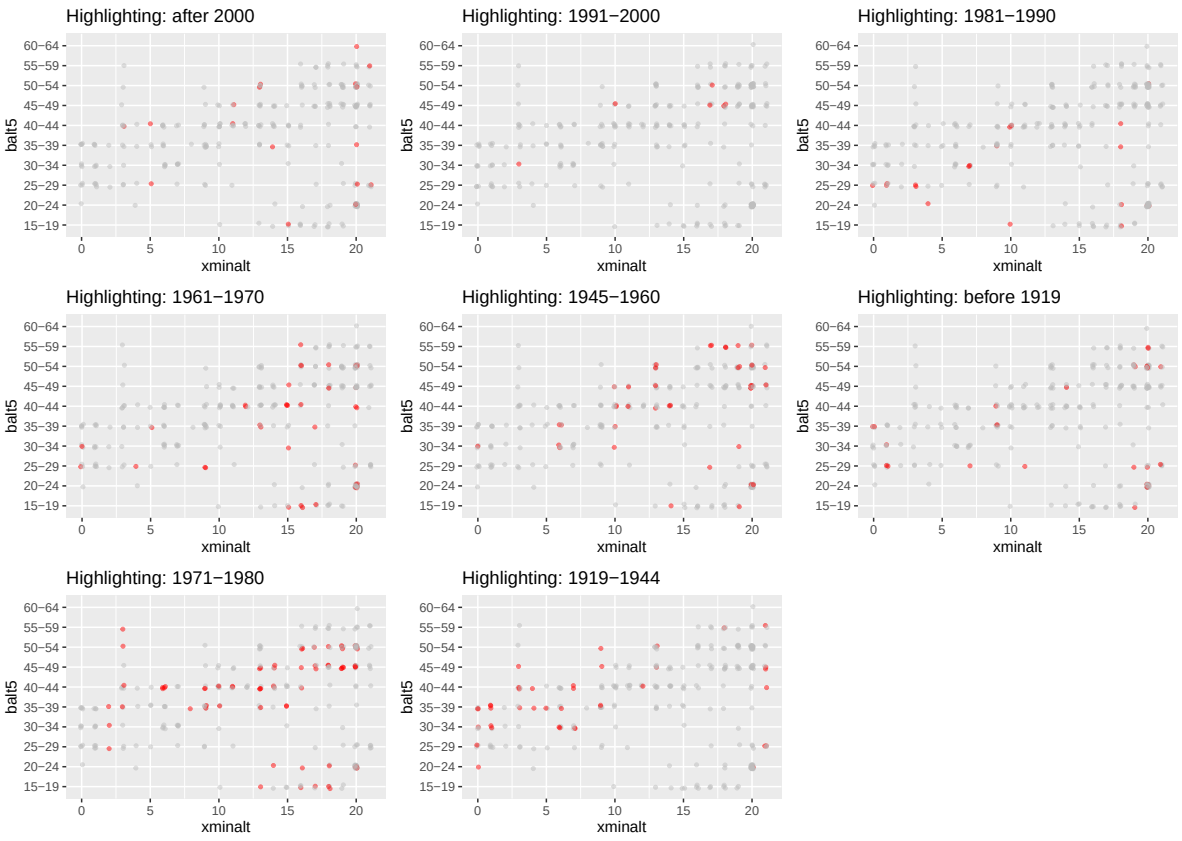
The 1981-1990 data reveals a distinct pattern for younger residents with shorter tenures, suggesting a potential interaction effect between younger age groups and this specific construction period.

4.4 Joint influence of balt5 ~ xminalt on the Target variable Werr

```
# Ensure that the levels of werr are consistent with the facet variable
# Unique categories in 'werr'
categories <- unique(data$werr)

# Create a plot for each category
plots <- lapply(categories, function(cat) {
  ggplot(data, aes(x = xminalt, y = balt5, color = as.factor(werr))) +
    geom_jitter(width = 0.1, height = 0.1, size = 1,,alpha = 0.5) +
    scale_color_manual(values = ifelse(categories == cat, "red", "grey")) +
    ggtitle(paste("Highlighting:", cat)) +
    theme_grey()+
  theme(legend.position = "none")
})

# Print plots
ggarrange(plotlist = plots, ncol = 3, nrow = 3)
```



Older buildings (pre-1919) appear to house a more diverse age range of parents with children of various ages The 1970s show the most widespread distribution of highlighted family structures. More recent periods (post-2000) show a trend toward older parents with teenage children The periods 1981-2000 show notably fewer highlighted points than other periods, suggesting different demographic patterns during these construction eras