Exercise 3

for Data Analysis

Valentino Lazarevic - 1223211

09.04.2024

Setup

```r
library(nycflights13)
data("flights")
attach(flights)

library(laeken)
data(eusilc)
attach(eusilc)
library(DescTools)
```

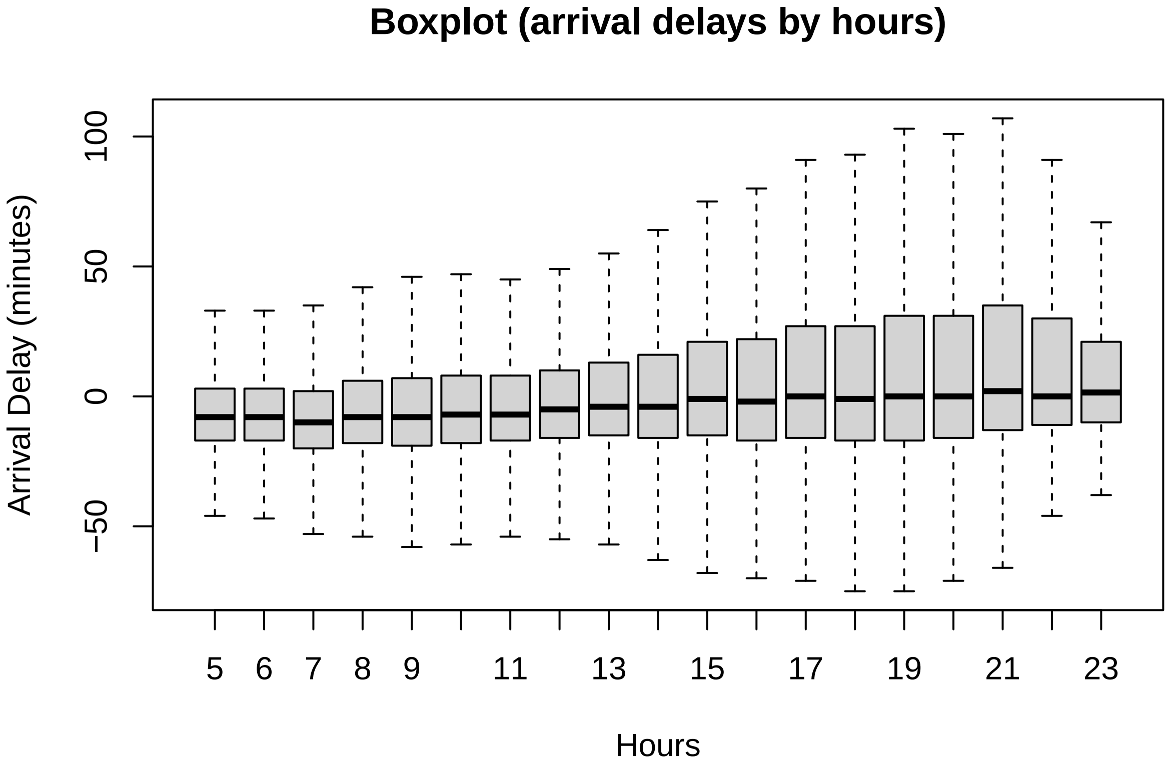
Exercise 1

Investigate by parallel boxplots if you can identify any tendency in the arrival delays concerning the hour of the day. What can you conclude?

Hint: The plots are heavily dominated by the outliers, and thus you might not want to show them.

Throughout the day, flight patterns reveal intriguing shifts. Departures commence at 7 am, reaching a peak at 21 pm. Between 5 and 11 am, consistency in interquartile ranges (IQRs) suggests stable performance. However, post-11 am, the IQR lengthens, especially in upper quartiles, indicating increased positive arrival delays. This trend peaks at 21 pm, aligning with the longest IQR and the largest median. Notably, whiskers extend significantly after 11 am, reflecting heightened variability in delays. Surprisingly, lower quartiles exhibit consistent departure delays, regardless of the hour, ranging from -10 to -20 minutes.

As the day progresses, both median departure times and IQRs expand, signifying escalating delays. This trend likely results from mounting air and airport traffic, leading to congestion-induced delays peaking at 21:00. Subsequent reduction in flight volume post-peak corresponds with decreased congestion and lower delays.

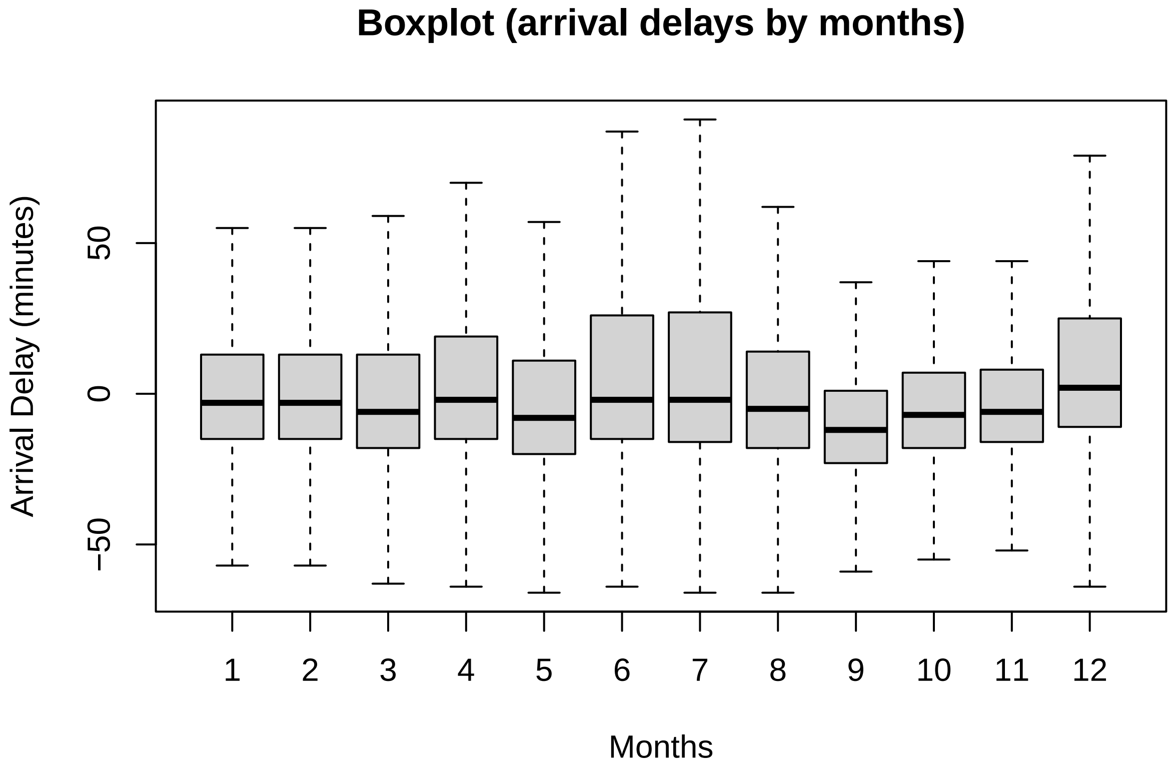


Exercise 2

Same as in 1., but for the different months.

Examining monthly flight data unveils intriguing variations. Initially, January and February exhibit uniform boxplot characteristics, suggesting consistency. However, March stands out with a smaller median arrival delay yet longer interquartile range and upper quartile length. June and July, despite similar median values to winter months, feature notably extended upper quartiles and whiskers, indicating significant variability in arrival times. Conversely, September boasts the smallest median and interquartile range, hinting at predictably smooth operations. December, paradoxically, experiences higher arrival delays compared to other winter months.

Interestingly, despite comparable data volumes across months, April, June, July, and December emerge as distinctive outliers. Weather conditions emerge as a potential explanation, with December's snow, June and July's heat, and April's rain likely influencing arrival delays. However, further analysis correlating weather data with flight delays is necessary to validate these hypotheses.



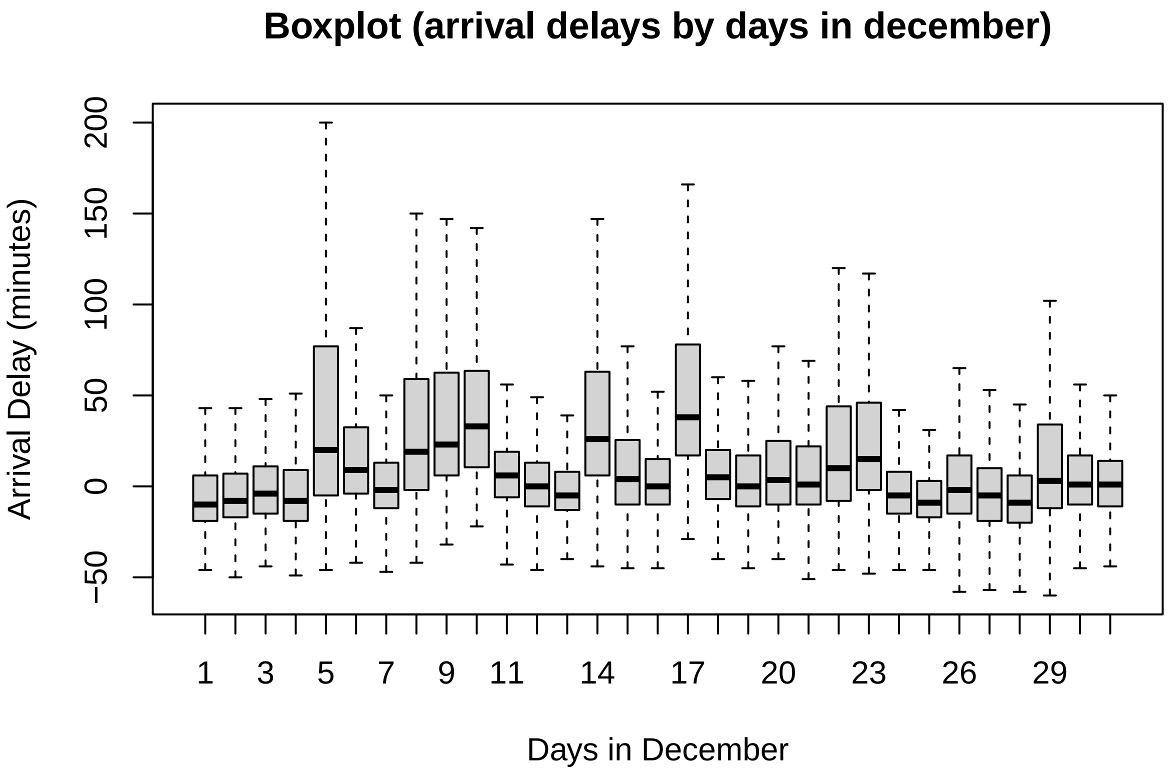
Exercise 3

Same as in 1., but for the different days of the month December.

Analyzing December flight data reveals intriguing patterns, particularly around the holiday season. Higher arrival delays on key dates such as the 22nd, 23rd, and 29th align with expectations, possibly due to increased baggage from holiday travel. Conversely, lighter and faster travel is anticipated on workdays throughout the month.

However, understanding delays in the first half of December poses challenges, as they do not seem to correlate with typical traffic patterns inferred from the histogram. The heightened delays on the 5th may stem from Thanksgiving holiday returns.

Notably, certain dates—5th, 8th, 9th, 10th, 14th, 17th, 22nd, and 23rd—stand out with significantly higher median arrival delays and longer interquartile ranges, indicating a wider range of delays. These outliers suggest an expectation of more extreme arrival delays on these dates.

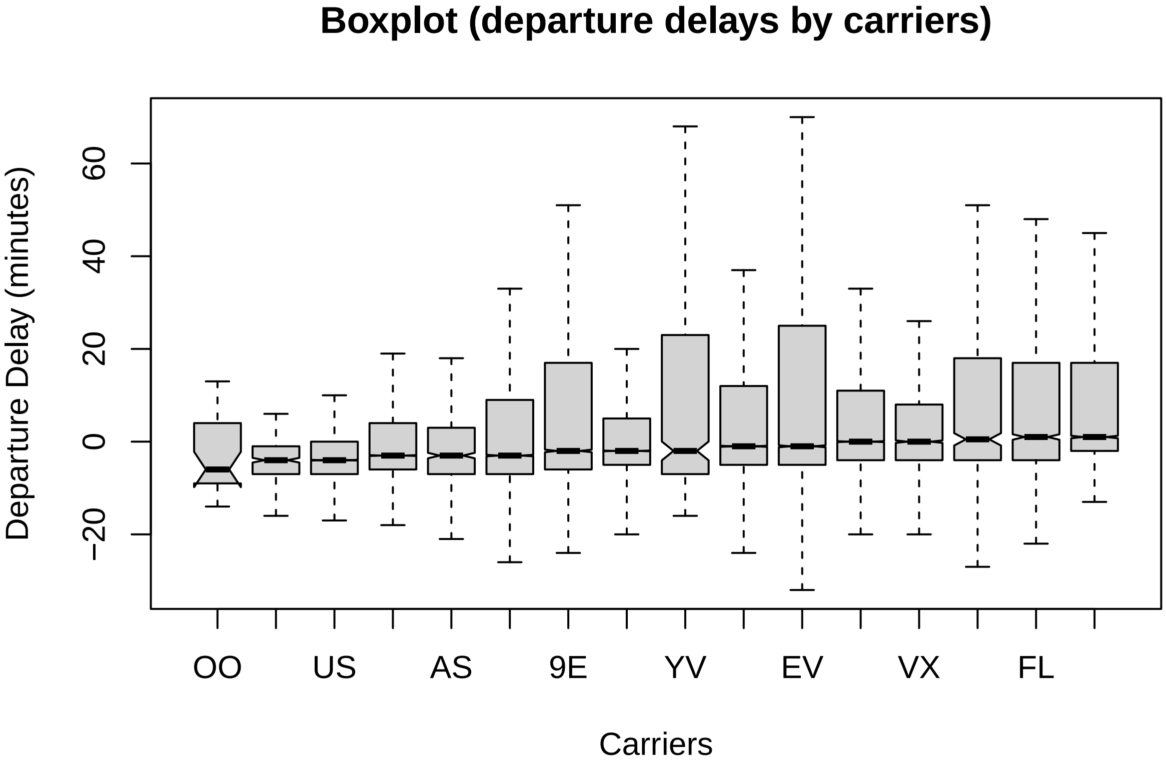


Exercise 4

Show parallel boxplots with notches for the departure delays of the different carriers. Sort the boxplots according to increasing values of the medians. What can you conclude? Is there a substantial ("significant"?) difference in the median departure delays of the carriers?

Analyzing carrier performance via boxplots reveals intriguing insights. Notably, some carriers' notches extend beyond hinges, potentially skewing median-based confidence intervals. For instance, "OO" exhibits the lowest median departure delay, around -6 minutes, with fewer outliers indicated by smaller whiskers. Conversely, "WN" shows the highest median at approximately -2 minutes. While median delays show little difference among carriers, variations emerge in upper whisker lengths. For example, "YV" displays longer upper whiskers, suggesting more extreme departure delays compared to "WN" or "VX."

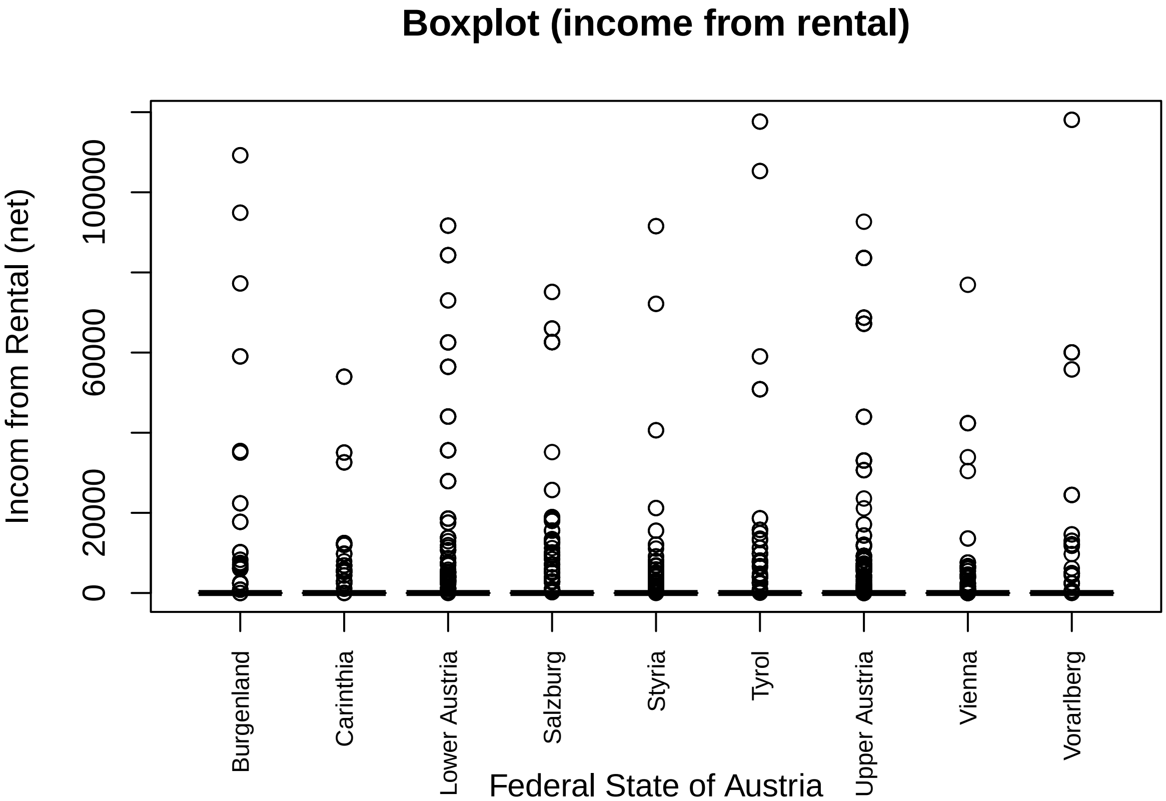
Interestingly, neighboring carriers often share similar median values, as indicated by non-overlapping confidence intervals. However, significant differences emerge when comparing carriers positioned at different ends of the plot. Additionally, utilizing a barplot to visualize the number of flights per carrier illustrates how tighter notches, representing confidence intervals, correspond with increasing data points.



Exercise 5

We are interested in the variable hy040n (income from rental of a property or land). Can you see income differences for the different Austrian federal states (variable db040)? Try to find an answer by using parallel boxplots, as well as by location/scale estimators.

When examining the boxplots, it becomes apparent that the majority of households do not earn any income from renting property or land. Across all federal states, the median income from this source is 0. Employing q0.95 as an estimator, we find that only a minority, approximately 5 percent of households in certain federal states, derive income from rental properties or land. However, analyzing mean household incomes by federal state reveals that households in "Burgenland" have the highest income from rentals.



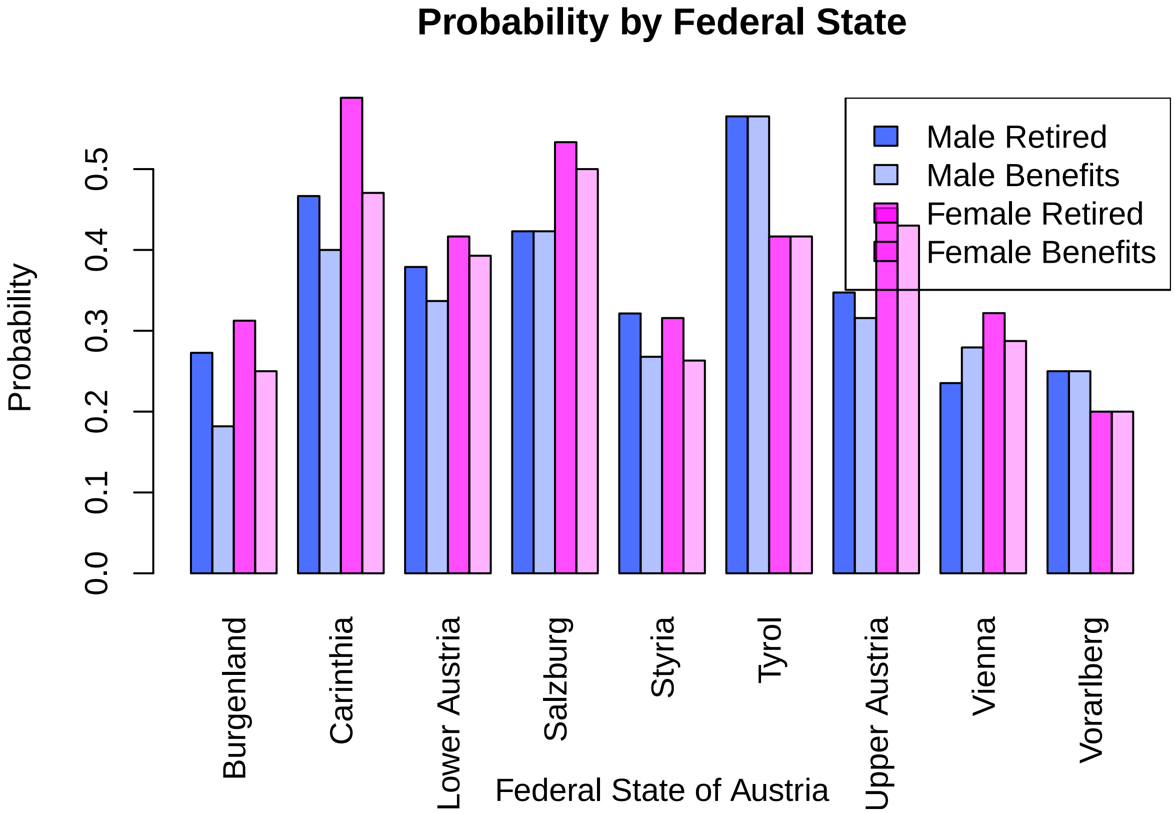
Exercise 6

Compare in a simple plot again the Austrian federal states for males in the age between 55 and 59 years the proportion of retired persons (pl030==5) with the proportion of persons receiving old-age benetits (py100n>0). Which federal states deviate? Show the results for the females for the same age group in the same plot.

Interpreting the data requires caution due to various constraints, notably the narrow age range of 55-59, resulting in a limited dataset, particularly when stratified by federal state. For instance, in Burgenland, all metrics exhibit overlap, likely due to a small sample size.

Across most regions, the number of individuals receiving old-age benefits aligns closely with the number of retirees. However, disparities emerge, such as in Tyrol, where fewer men receive benefits, possibly due to exceeding eligibility thresholds based on wealth. Additionally, more women tend to be retired than men in many states, with Vorarlberg being the only exception, where more women receive benefits than are retired.

Examining males separately reveals that, except for Vienna, the number of retirees equals or exceeds those receiving benefits. This pattern holds true for females across all regions. Notably, Lower and Upper Austria stand out, with twice the number of retirees or benefit recipients compared to other federal states. However, to gain a clearer understanding, it's essential to normalize these numbers based on the total population of older individuals.



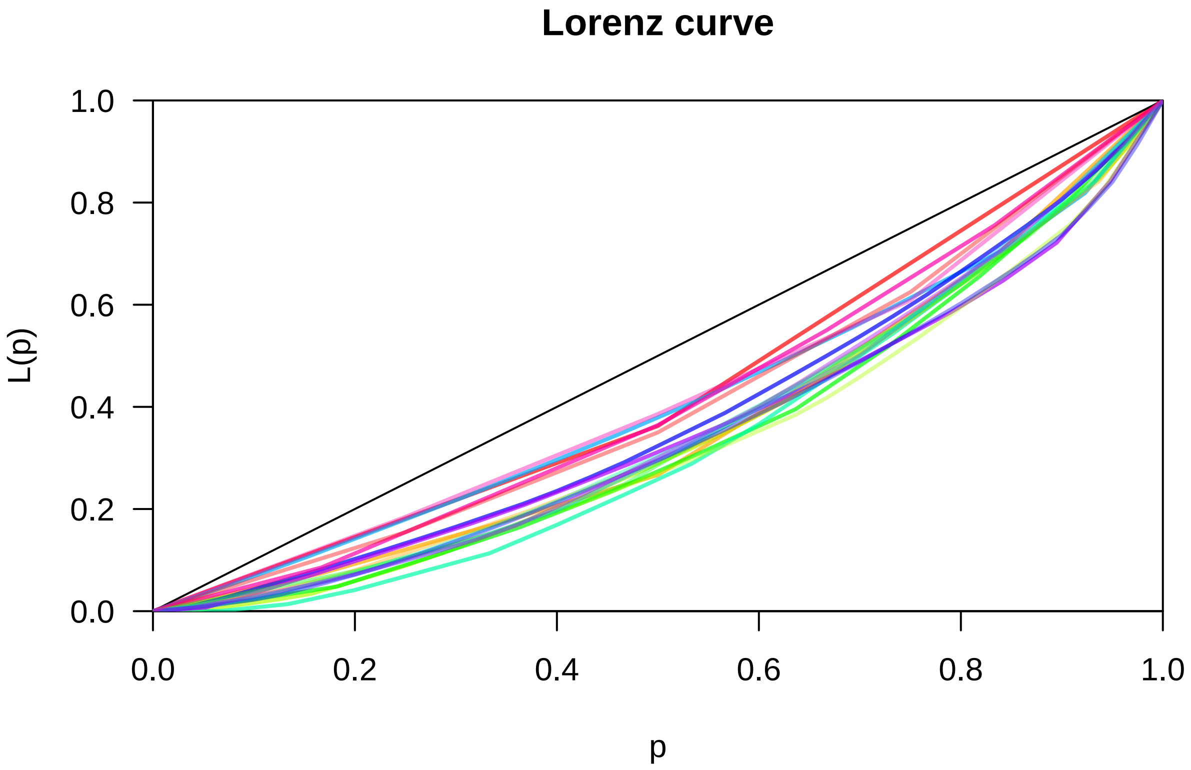
Exercise 7

Focus again on the age group from 55 to 59 years, and on all persons receiving old-age benefits. Present the Lorenz curve for males, separately for the different federal states, and do the same for the females. Compare the corresponding Gini values for males and females in the different federal states. What do the plots and values show? Are the income distributions balanced among the federal states and across the genders?

Hint: The R package DescTools provides the function Lc() to compute the Lorenz curve. The result can be plotted, and it also contains the Gini coefficient.

The Lorenz curve illustrates income distribution cumulatively across the population, compared to the line of perfect equality (y = x). The Gini coefficient measures inequality by comparing the area between the Lorenz curve and the line of perfect equality to the total area below the latter, with 0 indicating perfect equality and 1 perfect inequality.

Typically, male and female Lorenz curves closely resemble each other across states, with Gini values ranging between 0.2 and 0.4, indicating moderate inequality. However, two outliers exist: Burgenland, where male income inequality is notably higher at 0.6, likely due to a small sample size; and Vorarlberg, where female income is twice that of males, possibly also influenced by sample size discrepancies.



Legend of Exercise 7

Male/State

- ● Burgenland.male
- ● Carinthia.male
- ● Lower Austria.male
- ● Salzburg.male
- ● Styria.male
- ● Tyrol.male
- ● Upper Austria.male
- ● Vienna.male
- ● Vorarlberg.male

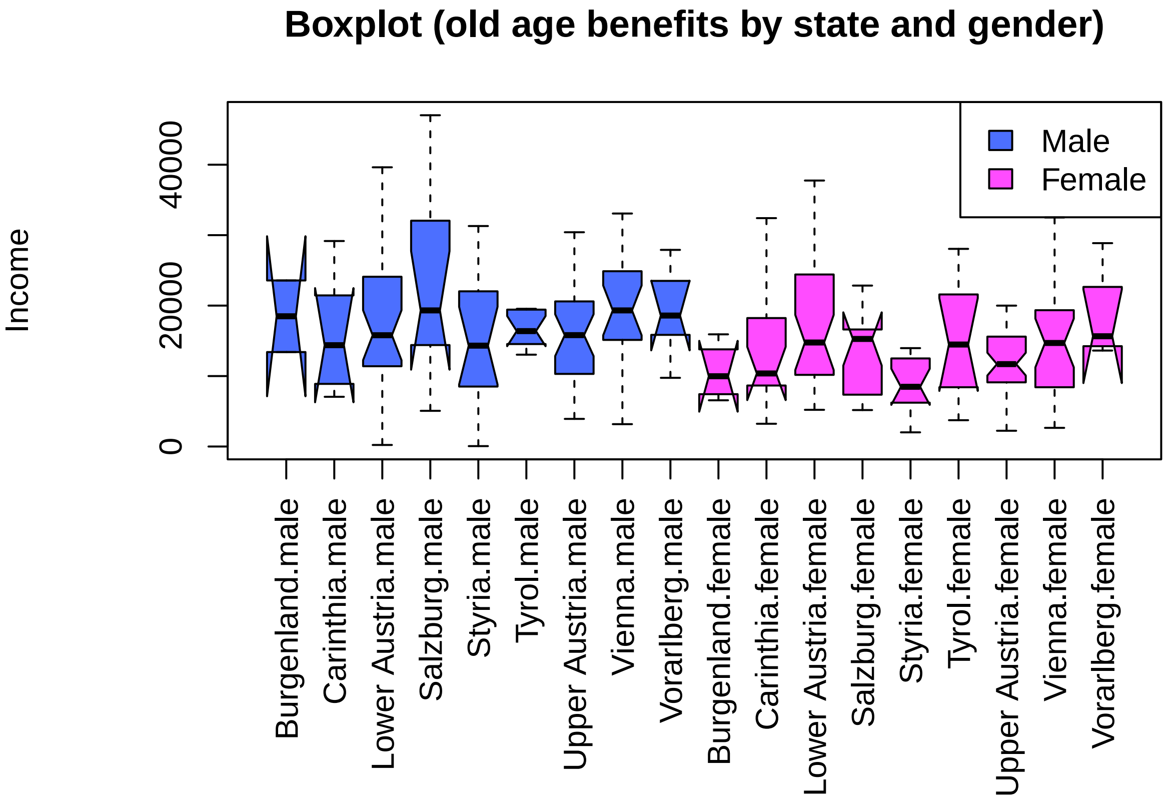
Female/State

- ● Burgenland.female
- ● Carinthia.female
- ● Lower Austria.female
- ● Salzburg.female
- ● Styria.female
- ● Tyrol.female
- ● Upper Austria.female
- ● Vienna.female
- ● Vorarlberg.female

Exercise 8

Use the same data as in 8., but show them with parallel notched boxplots for the different federal states and genders. What do you conclude?

The correlation between this boxplot and the previously discussed Lorenz curves isn't entirely clear. While certain observations, such as a large interquartile range (IQR), typically align with higher Gini values and greater inequality, as seen in males from Burgenland or females from Vorarlberg, a small IQR doesn't necessarily indicate a lower Gini value compared to other states. This discrepancy might stem from unaccounted outliers or the small sample size, influencing the interpretation.


Boxplot (old age benefits by state and gender)
Income
Male
Female
Burgenland.male
Carinthia.male
Lower Austria.male
Salzburg.male
Styria.male
Tyrol.male
Upper Austria.male
Vienna.male
Vorarlberg.male
Burgenland.female
Carinthia.female
Lower Austria.female
Salzburg.female
Styria.female
Tyrol.female
Upper Austria.female
Vienna.female
Vorarlberg.female