Data Analysis and Visualization in R (IN2339)

Case Study

Robin Mittas

2022-12-08

Motivation

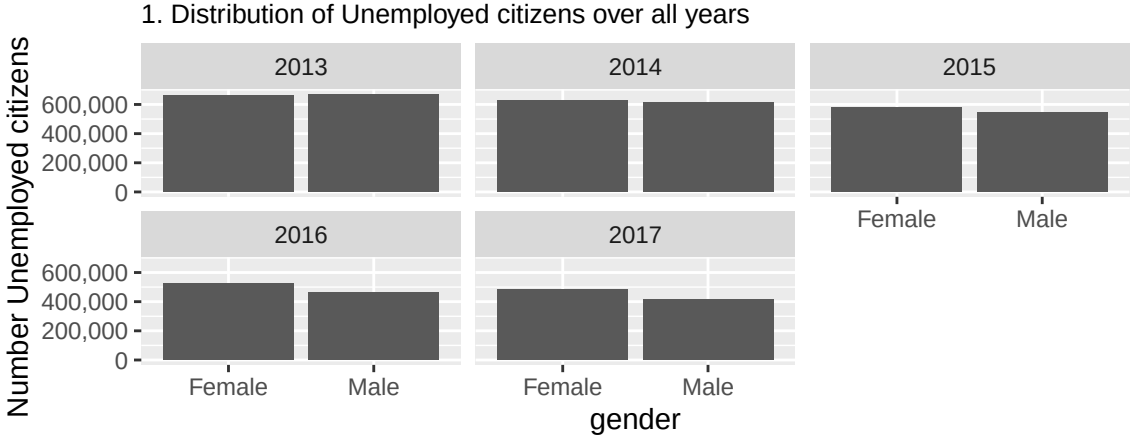
Since the financial crisis of 2007-2008, high unemployment rate amongst younger generations is one of the highest concerns. With only 15% of the people under 25 unemployed in 2008, this number rose to 23,6% by 2013 especially disadvantaged countries like Greece and Spain, which were strongly hit by the economic recession. This trend was strongly surpassed by Catalonian regions. This is why we decided to gain a deeper insight on this topic and analyze the given data sets and possibly better understand and detect some interesting factors on the youth employment in Barcelona between 2013-2017.

Data Preparation

In this section, we perform the needed data preparation steps for the analysis. Unnecessary chunk codes are omitted in the compiled pdf-file. After loading the files, some column names were renamed for easier joining and further processing.

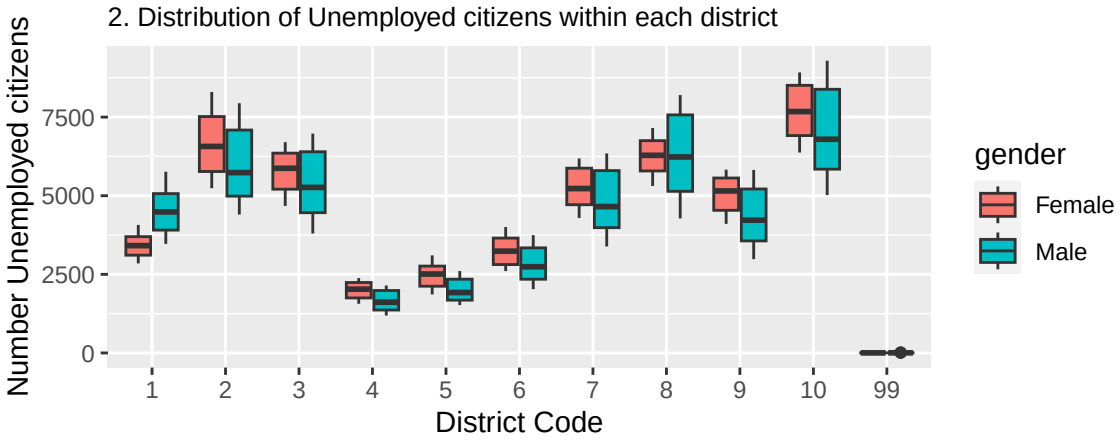
Data Analysis

To get a first overview of the number of unemployed citizens in each gender group we have created the following descriptive plot.




```
##    year female_unemployed male_unemployed diff_gender_groups_female_minus_male
## 1: 2013            660552          670661                                -10109
## 2: 2014            625114          616477                                  8637
## 3: 2015            581641          542940                                 38701
## 4: 2016            525817          465359                                 60458
## 5: 2017            485116          412780                                 72336
```

From the calculated dataset (see under the first plot) we see that in every year, apart of 2013, there are more female citizens registered unemployed. The second plot outlines the distribution of the number of registered unemployed citizens within each district per gender over all available years. We have one data point per district, month, year and gender-group.



From the plot we can see that the most unemployed citizens are living in District 2 (Eixample), 8 (Nou Barris) and 10 (Sant Marta). Let us especially remember district 2.

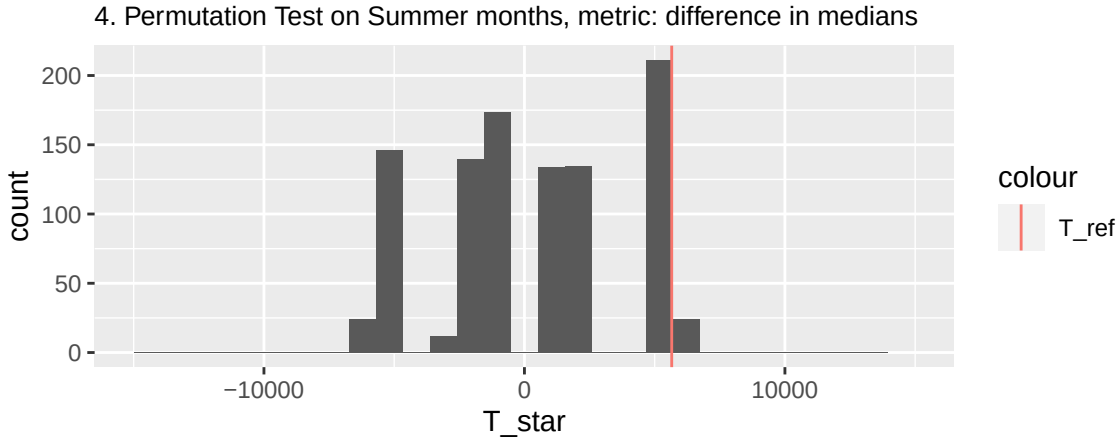
In order to check if there are seasonal effects due to e.g. tourism, we plot the monthly data (for this we created a plot which is not shown in the respective document - see here R Studio). From the plot we can conclude that there is almost no seasonal effect BUT: the data points are per district, let us see how it is for complete Barcelona and what kind of effect it might have. We will consider each year independently as the number of unemployed citizens should also decrease over the years.

We see that the number of unemployed people decreases over the summer months in each year. This is probably because of seasonal effects e.g.: construction industry, tourism.

To underline the result we can now test with a t test (one binary, one continuous variable) → p_value of 3% - 2017. The p-value of 2017 is definitely underlying our hypothesis. → p.value # 21% p_value - 2016. The p-value of 2016 might be an outlier year.

We see that when we aggregate the number of unemployed citizens among all districts/ neighborhoods, seasonal effects can be seen. From the plot we might conclude that the differences from the median are quite high. But as we have learned in the lecture, we need to be careful before making any conclusions. The pattern might be an artifact of random variation and might disappear if we had more data. For that we are doing a statistical permutation test, meaning we consider the possibility that the difference in medians could

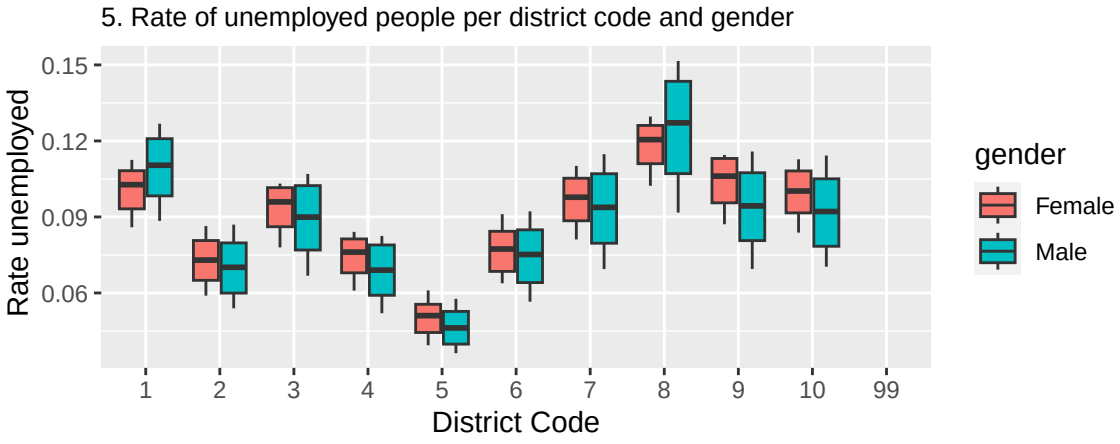
often arise by chance. We are randomly assigning months to a row, grouped by the year to infer from the plot that the number of unemployed citizens is slowly going down over the years (only one exception in 2015 can be observed).



There is a possibility that the observed "event" occured by chance (also indicated by the p value). However, the result underlines that there could exist some kind of seasonal effects. We have just observed a few differences in the medians bigger than the original one (T_ref).

Data Preparation 2

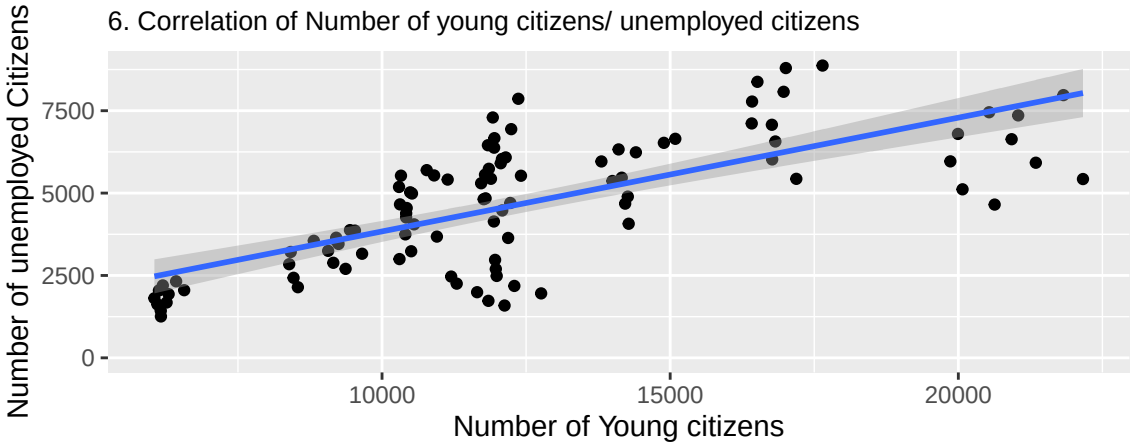
As a preparation for our next analysis, we aggregated the population by year, district and gender to gain direct information on the distribution between the individual age groups. To achieve this data basis we have first calculated the total number of immigrants/youth via the total number of citizens. We define all citizens in the age groups 15 - 64 as potential workers and citizens between 15- 29 as youth citizens. The total number of immigrants is aggregated by all immigrants except Spanish, as they may have come for studying etc. By merging all data we now have added some additional information to the unemployment data set with the rate_unemployed (unemployed/ total workers), the share_youth (youth citizens/total workers) and share_immigrants (immigrants/number citizens). Let us now dig deeper into the different information we have now gained and investigate if there are co-founding factors for the unemployment. In comparison to the plots in the beginning, let us put the number of unemployed people per district/ gender into proportion of number of citizens per district/ gender.

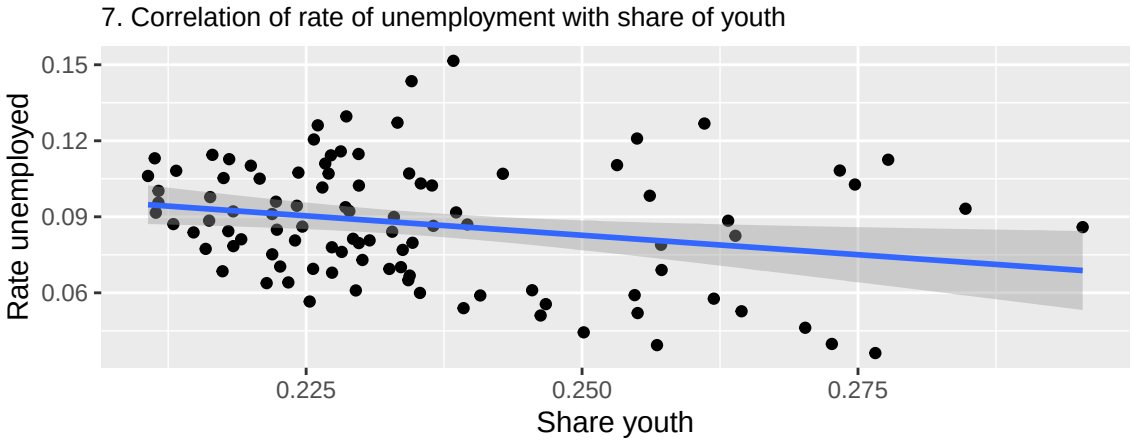


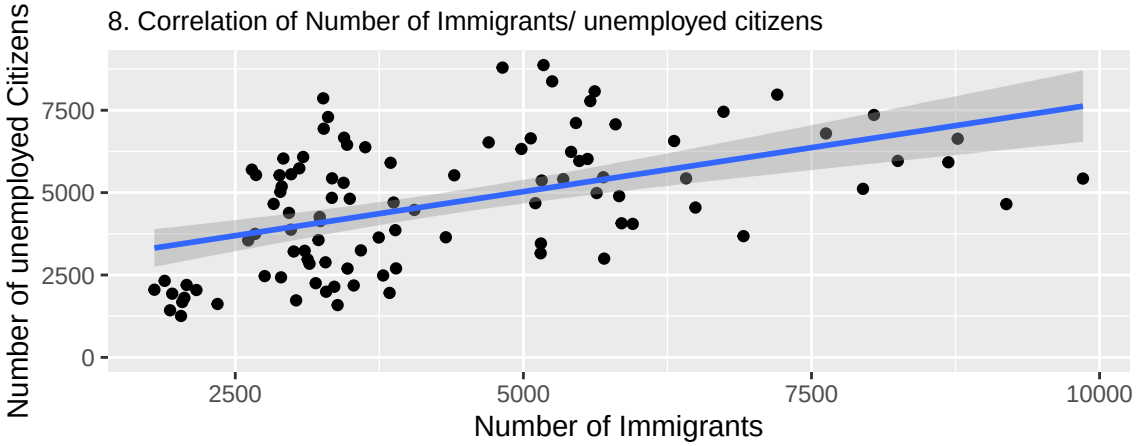
As we can now see e.g. district 2 doesn't has such a high unemployment rate, but from the first plot one could have falsley inferred that the rate might also be high due to the total number of unemployed citizens.

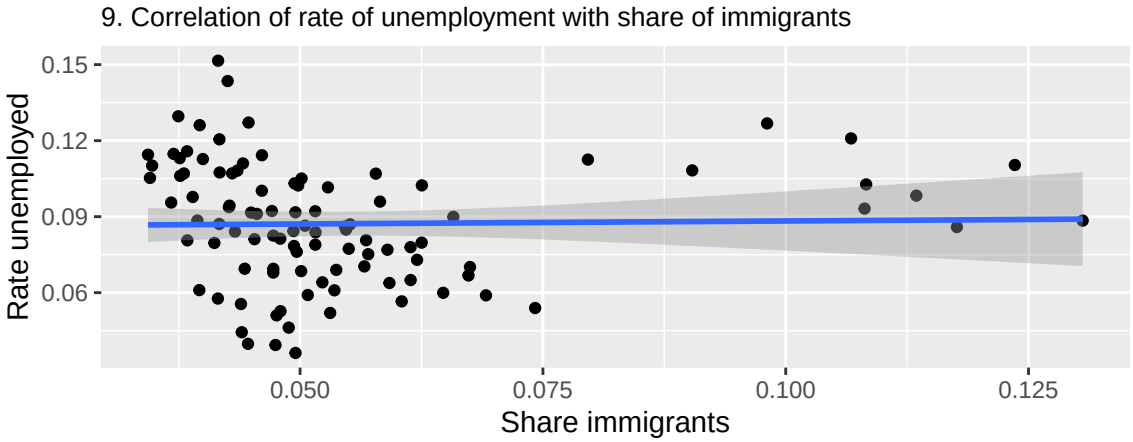
Next, we will check if the rate of unemployed citizens and the share of youth/immigrants correlate. We

achieve this by testing these statistics in relation to the total number of citizens within the separate districts. On the plot Nr. 6 we see that the number of young perople correlates strongly with the number of unemployed citizens. But as this is an obvious conclusion, we will now check for a possbible cofounding factor, the number of citizens. Let us rather take the rate of unemplyoment and the share of young citizens instead of the total numbers. The same holds for number of immigrants.









We see that the share of youth within all districts does actually not correlate with unemployment rate: Districts with a high share of immigrants do not necessarily have a high unemployment rate, there might be a weak correlation. Let us test this with a Spearman test to clarify if the two continuous variables correlate. The first rho is the statistic for the correlation between share of youth and rate of unemployment, the second one for the share of immigrants. Both are negative, thus we conclude they definitley do not correlate.

```
##        rho
## -0.2436964
```

```
##       rho
## -0.270339
```

We conclude that the share of immigrants as well as the share of young people within the population does not correlate with a high unemployment rate. For Gender we will use a t-test, as we have one boolean (Male=True, Female=False) and one continious variable (number of unemployed citizens):

```
## [1] 0.5567207
```

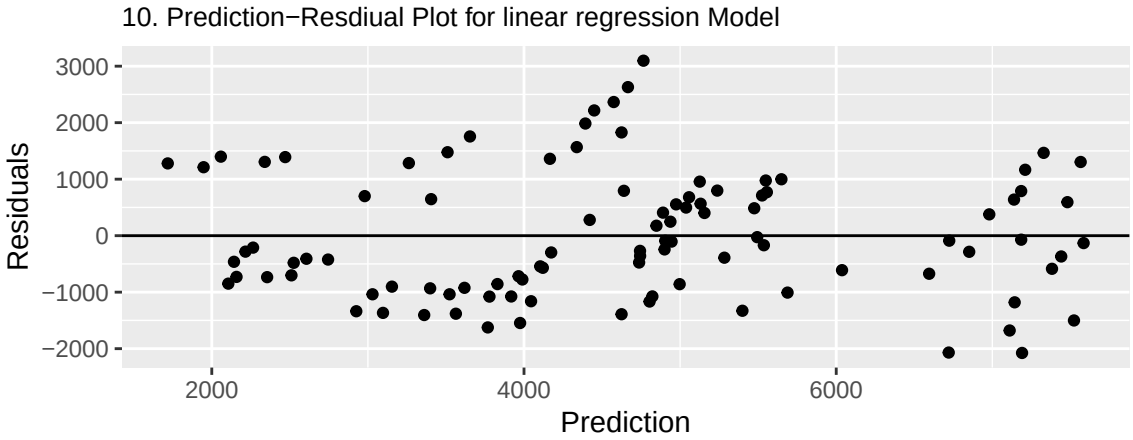
The p value is at 55%, which is quite high and indicates that gender is not really associated with unemployment. To summarize, summer months are a cofounding factor for unemployment but the share of immigrants as well as share of young people within district are not. We always also have to compare all numbers (such as number of young people, number of unemployed people) to the total numbers in a district to remove all cofounding factors.

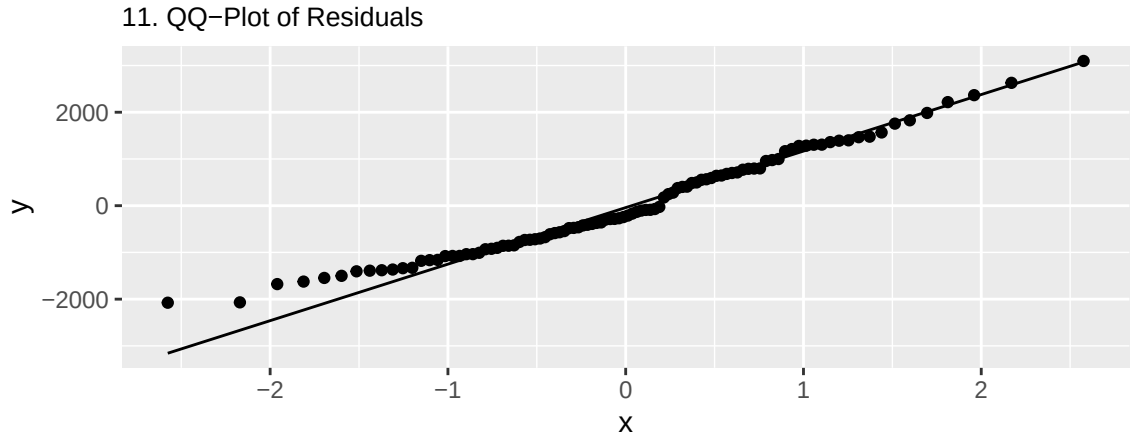
I will now quickly add some basic Linear Regression Model to predict the number of unemployed citizens given the total number of citizens. We are using following formula (explanatory and response variables):

```
## number_unemployed ~ number_citizens_total + number_workers_total +
##     number_citizens_youth + number_immigrants
```

From the Plot 10 we conclude that the average of the residuals doesn't seem to change across predicted values. Therefore the residual vs predicted values plot doesn't provide evidence against the linearity assumption, namely that the expected values of the response are a linear combinations of the explanatory variables.

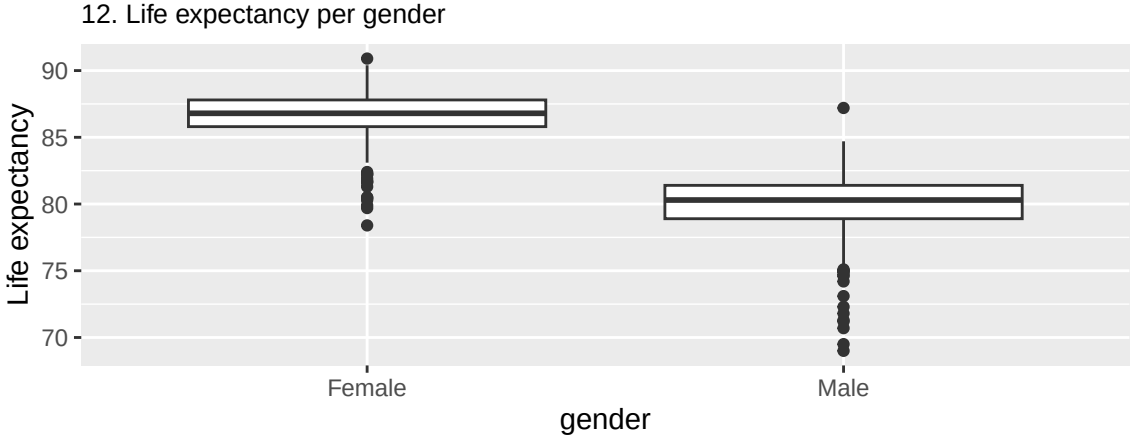
Also, the variance of the residuals seems to be constant across predicted values. Together with the previous observation, such supports that the errors are identically and independently distributed. An implication that the errors follow a Gaussian distribution is that the residuals also follow a Gaussian distribution. The Q-Q plot (Nr. 11) supports such distribution.

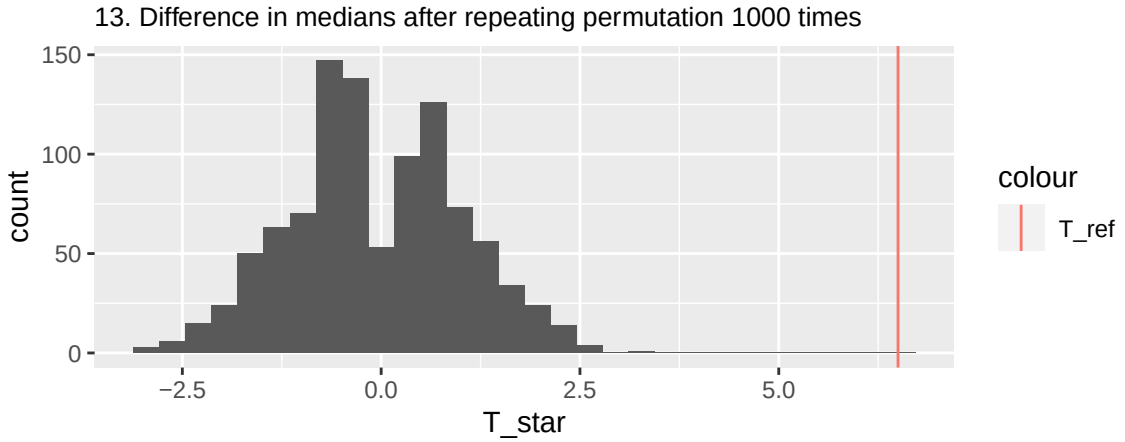




As we have focused a lot on differences in gender groups we want to finish the case study by a quick statistical test. The statistically supported hypothesis should be that women have a higher life expectancy than men in Barcelona.

The demonstrative plot 12 shows that a difference in the life expectancy of females and male exists. In general females live longer than men. But the question arises whether the difference arose by any chance. For this we run a Permutation test. Plot 13 shows the clear result underlying our hypothesis.





After conducting the permutation test we can conclude that the life expectancy of women is higher than of men as we have not observed a difference equal or larger than the original one among 1000 permutations.

Conclusion

The general ideas about issues surrounding increased unemployment among residents in Spain, and especially hotspots such as Barcelona, may well be revealed by analyzing the available data set. In addition, we statistically showed that there is an influence of the respective season, e.g. summer months, in the number of unemployed citizens which we have seen in the demonstrative plot number 3. In fact the season is also a co-founding factor of unemployment. Thus,the problem of unemployment intensifies especially in the winter months, because then probably short-term jobs from e.g. the tourism and construction industry are not available. The gender of the persons concerned is not a decisive factor for unemployment. Women and men are therefore at virtually the same risk of unemployment.
The potential stereotype of unemployed immigrants cannot be substantiated by our analysis either. Due to the data we could just show that in districts with an increasing share of immigrants, the unemployment rate is not increasing. Summarizing we have also investigated different co-founding factors for unemployment, namely the number of immigrants or the share of youth within a district and have also tested statistical whether they influence the unemployment rate. In addition, we conclude based on a statistical permutation test that the life expectancy of women is higher than that of men.