Final Report

Expenditure on capital investment in € per capita // Child poverty

Ausgaben für Sachinvestitionen in € je Einwohner // Kinderarmut

Expenditure on property, plant and equipment in € per capita // Child poverty

Correlation Coefficient

| cor(Var1, Var2) | > 0,60)

cor(df$A_Sachinvestitionen, df$Kinderarmut) = **-0.7012575**

Correlation coefficients determine the intensity of the linear relationship of two variables. A maximum value a variable can consist of is from -1 to +1. A positive relationship is demonstrated by a linear correlation coefficient greater than zero and negative relationship. If the correlation coefficient is less than zero, it is considered a negative correlation.Since the correlation coefficient value that we get here is in negative, that means both the variables here we get are heading in the same direction.

The correlation value in our case is **-0.7012575** which shows a significant correlation between child poverty and capital investment in an area. Although these two variables do not seem to correlate, the fact that higher capital investment signifies greater wealth of a family and thus reduces the chance of child poverty, explains the phenomenon well. The negative correlation says that higher the amount of capital investment, lower the chance of child poverty. This also is well explainable by real world observation.

2. Histogramme und Boxplots in ggplot2! →

As we create a histogram for each variable, we notice that the variables are linearly ordered i.e. values lie within an interval and do not rise exponentially from lower to higher or vice-versa. Due to this, further transformation of these variables is not deemed necessary. We can however divide e.g. Investment by 100 to make it the same scale as child poverty, but since it's already noticeable in histogram (scale of 100, 200, etc.) we won't transform our datasets.

Binwidth determines how narrow or wide intervals we want to take for an individual bar in histogram. e.g. binwidth of 10 gives no. of samples present in

intervals of 10 i.e. 0-10 or 10-20. It should be adjusted in such a way that the intervals reflect the general tendency of data. We choose a binwidth of 20 since the data is ranging from 0 to 1300 and taking each interval of 20 shows the general trend in data without being too general or too high resolution.

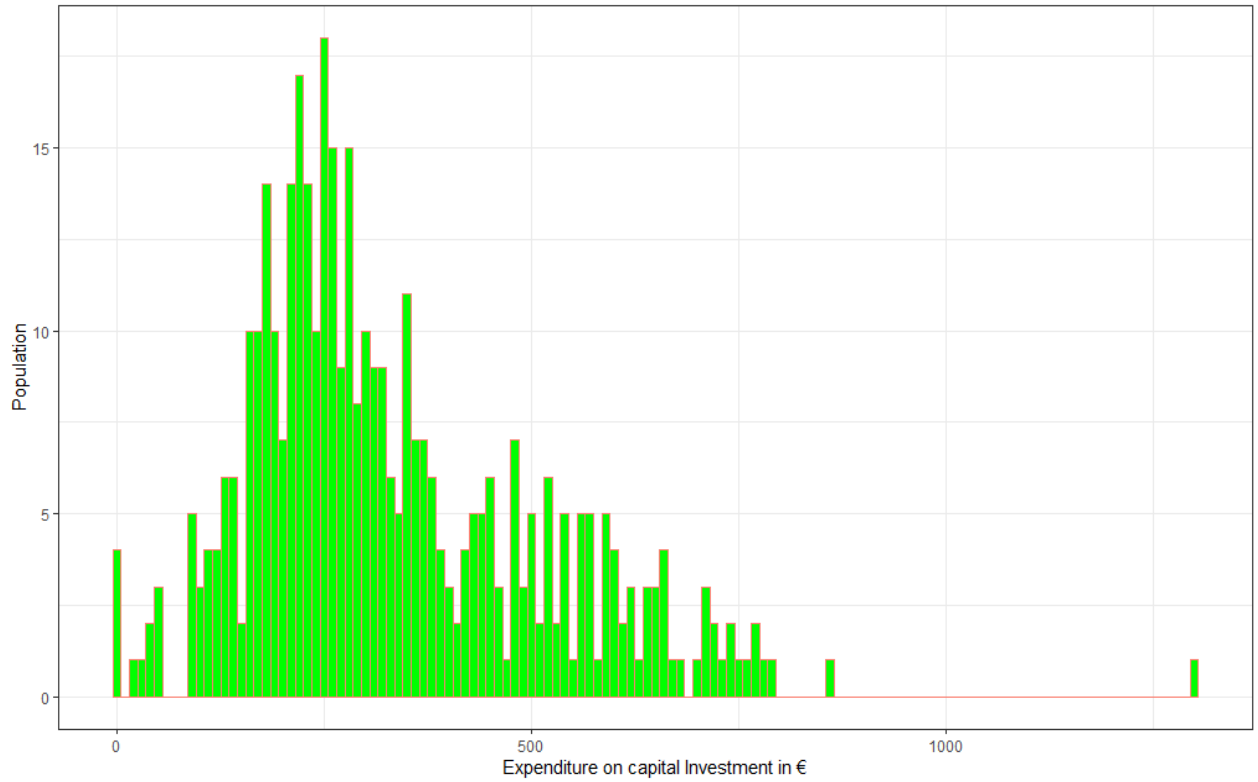

Fig 1 The Outliers remains away from the other datasets.

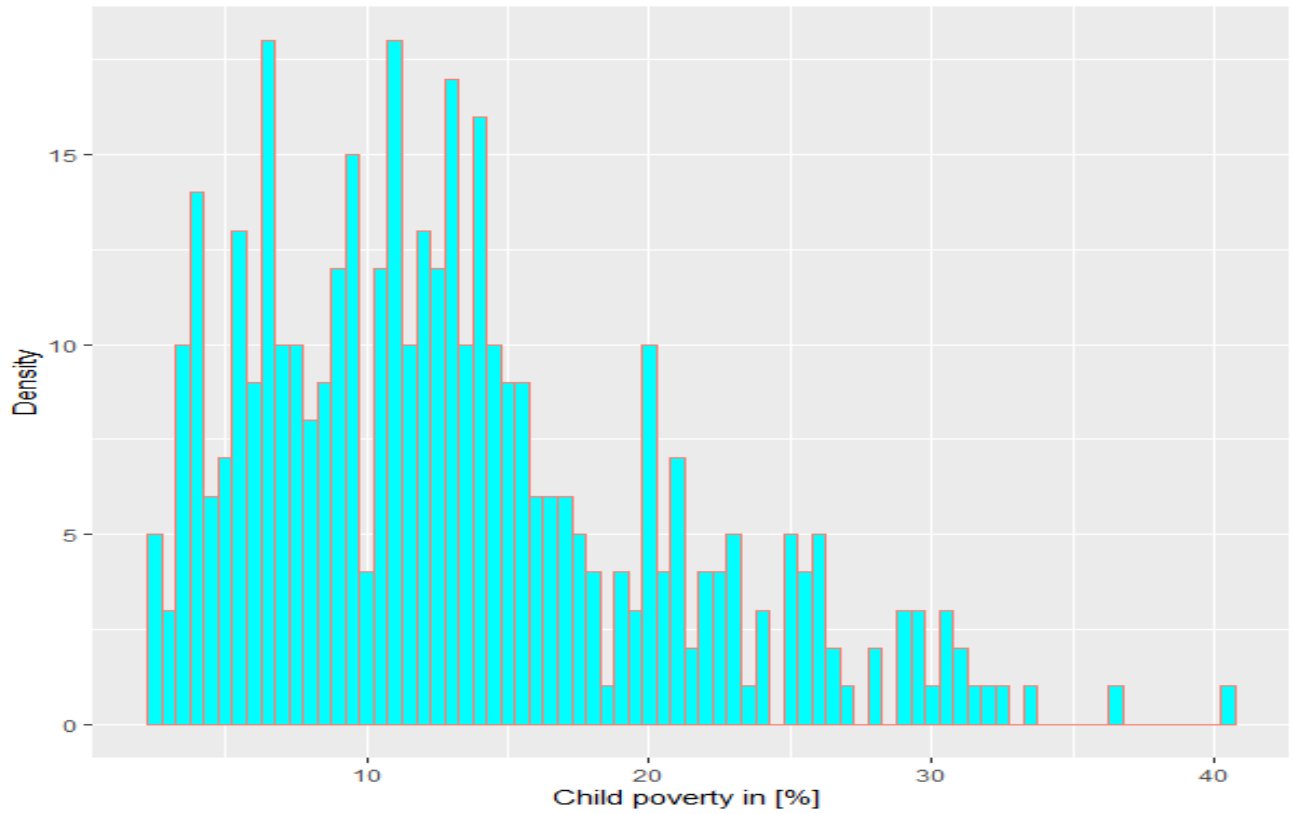

Fig. 2 Histogram of Child poverty

After rounding up the binwidth value between 0.3 -1.0, the histogram plot above with binwidth of 0.5 . We can relate the amount of money spent on capital investment to child poverty. The areas where the capital investment is higher, the chances of child poverty is very limited.
What we can assume from our data is that the areas where the capital investment is zero, a reason could be that even though they are well enough to expense on capital investment, they are reluctant to do it. And also for another reason some families do get the state aid "Staatshilfe or Kinderhilfe" which also signifies in reducing the child poverty and plays a direct role in not investing on capital expenditure.

Boxplots

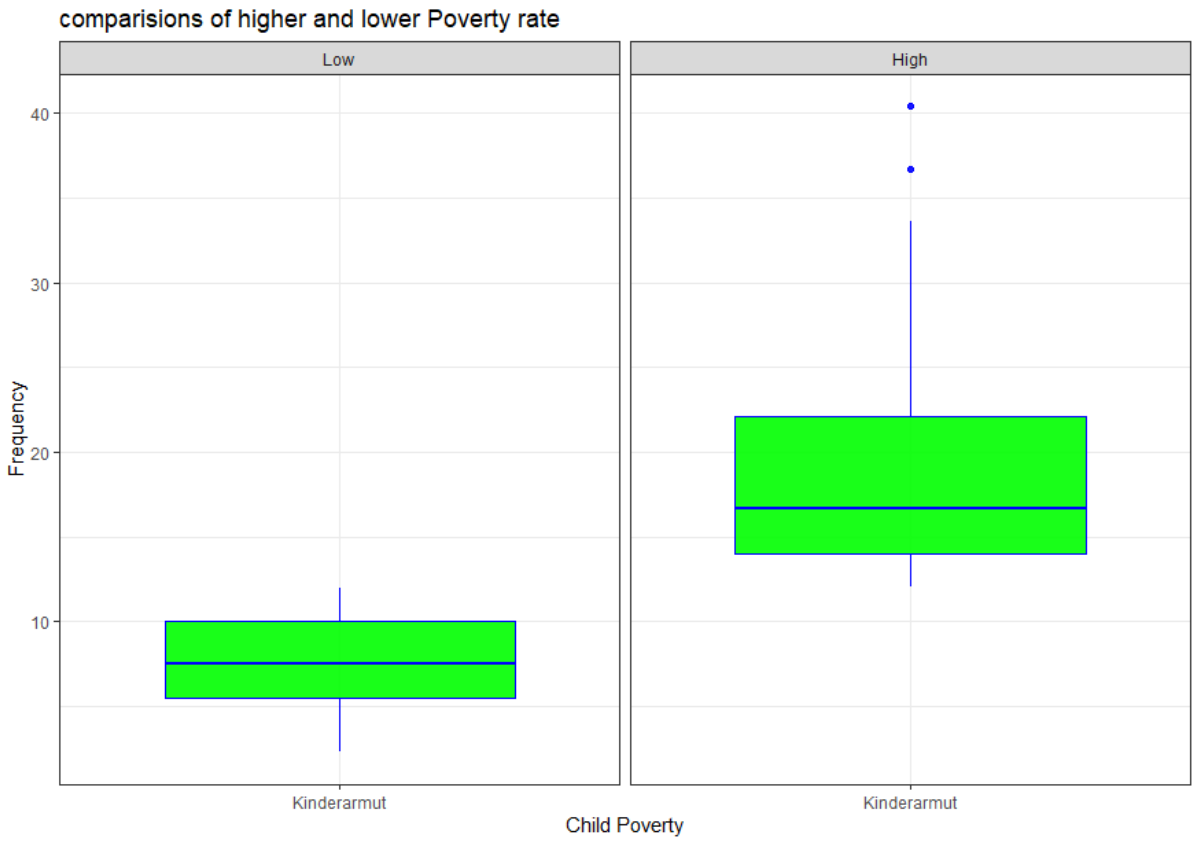

Fig.3 Boxplots of two data sets

In fig 3, the comparisons between two features of a continuous (mean area) data is examined.
On the left boxplot, we can see the data are almost normally distributed where the median of the dataset is centrally located but upper 25% data are slightly lower than lower 25%.

The median is closer to the lower quartile, so we can say the skewness of the data is positively skewed(Positive skew).

There are some data in outliers which are proportional to child poverty.

3. **Removing NA- value and Outliers**

Since we don't have NA values in data, we don't need to filter out these data but there are outliers and to remove them we set a limit of 900, this roots out one outlier. we build a subset of the data with above conditions met.

new_df <- **filter**(df, A_Sachinvestitionen<900)

With the help of filter function in R-studio we have filtered out the **Outlier** data and limited it to 900, of the least possible value.

4. Scatterplot

As indicated by the correlation coefficient, it can be seen that the data has a fairly linear structure. A linear fit is shown in the diagram (red line), which shows the linear trend of the distribution of the data. The data is sloped from top left to bottom right which signifies the negative nature of correlation.

The Scatterplot expresses a highly negative linear correlation(~0.71) between the capital investment and the child poverty. As the investment on capital increases the child poverty decreases and vice versa which is a perfect negative correlation.
There are several reasons causing child poverty and from the Scatterplot we can express that investment on capital is also one of them.

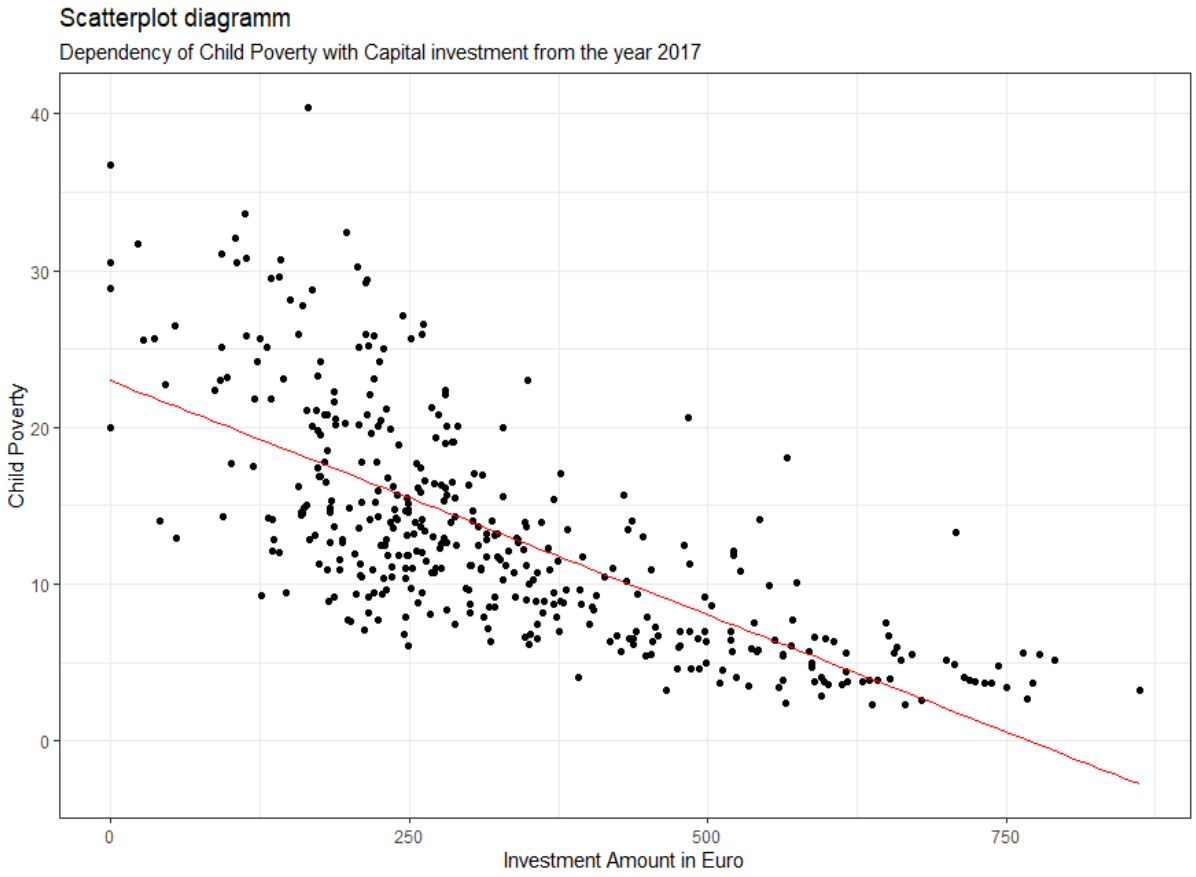

Fig. 5 Scatterplot

5) Linear Regression Models

→

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	22.969510	0.546932	42.0	<2e-16 ***
A_Sachinvestitionen	-0.029856	0.001478	-20.2	<2e-16 ***

A linear regression helps in modelling the relationship between two variables from an observed data. The equation is in the form of $\boldsymbol{Y = a + bX}$, *where* $\boldsymbol{X}$ *is the capital investment in our case and* $\boldsymbol{Y}$ *is the child poverty.* *As per our calculated regression output, the regression model will be as follows:*

Y = 22.9695 -0.0298X
e.g. Y = 22.9695 -0.0298 * 1300 = - 15.7705 → Negative

Y = Child poverty
a = intercept value
b = slope
X = independent variable

Code in R:

```
regressionmodel<-lm(Kinderarmut~A_Sachinvestitionen,data =new_df)
```

Here child poverty is used as a predictor and the capital investment is used as a response factor.

Residual standard error: 5.047 on 398 degrees of freedom
Multiple R-squared: 0.5061, Adjusted R-squared: 0.5049
F-statistic: 407.9 on 1 and 398 DF, p-value: < 2.2e-16

→ As per above findings in the regression model, the slope of the independent variable is negative to the dependent variables which is -0.0298. The calculated output of the R^2 in the regression model is 0.5061; where the model represents 50% of the variations in response variables around its mean.

Only 50% of this data represents the relation between the counties of Germany which shows the negative linear regression model. Thus, with only one independent variable we cannot describe the output of the predictor, while many other independent variables could also lead to the output/child poverty.

6).

The region based shape data was downloaded from the ESRI Deutschland open Data portal and was loaded into R-Studio and is merged. These shapefiles represent the spatial vector data representing points, lines and polygons in a map.
There was some error while merging the data as the problem was on data due to leading zeros, which were added from 0 to 318.

Code to add a leading zeros.

```
df$Kennziffer[0:318] <- paste0("0", df$Kennziffer[0:318])
```

7)

The child poverty distribution, as shown in figure 5., shows that the poverty distribution is comparatively higher in the northern part of the country compared to the southern part. This

is in accordance to the lower investment in those areas, as shown in figure 6. Comparatively, the southern part of the country has higher investment.

Another interesting region is the NRW region where highly populated cities are located, the investment is relatively low, the child poverty rate is also higher.

In the eastern region, provinces like in Berlin, Neu-Brandenburg, the capital investment is very low which expresses that the child poverty percentage is very high in these areas.

Interpretations:

Northern and Eastern provinces are relatively less prosperous than the Southern and Western provinces. This is reflected in the plotted map of capital investment. As the child poverty is considerably negatively correlated with investment. The plotted map of child poverty shows inverse tendency to that of capital investment. i.e higher child poverty in northern and eastern regions. One peculiar observation in terms of child poverty is seen in the NRW regions, where despite the area being relatively prosperous the child poverty rate is higher than expected. The reason for this could not be easily discerned from the data at hand.

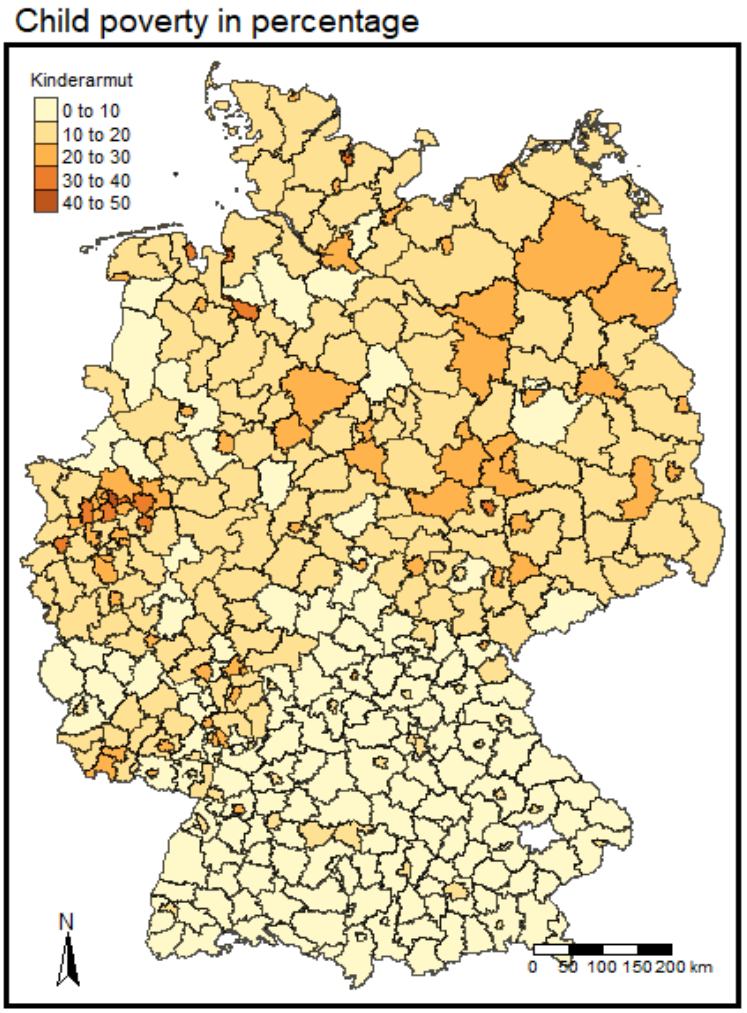

Fig. 5 -> Child poverty in percentage

Fig. 6 → Capital Investment in € per capita