Final Project

This project is focused on spatial and statistical data analysis of the Digital vector geographic database of the Czech Republic *ArcCR500* created in scale details 1:500 000 (km). First, data from the ESRI geodatabase in the form of feature classes was converted to three shapefiles - for the state, for regions, and for districts. The most interesting one is probably the district shapefile consisting of statistical indicators for each district in the Czech Republic. These three shapefiles were then loaded to RStudio. Stastical analysis were performed for districts shapefile as that divides the Czech area into much smaller parts, therefore, statistical analysis are even more interesting.

The first Markdown tab introduces those data and plots several maps capturing the population values across the area. Those maps were created using ggplot package. As statistical indicators are called in Czech language, the table explaining all indicators was drawn up.

The other four tabs are focused on Explanatory Data Analysis (EDA). First I have adjusted these statistics creating new columns that will show individual statistics with respect to inhabitants. This will make the results much more accurate. We can see histograms of all columns. First 18 columns are related mostly to population and are very similar. We can see very similar curves for numbers of wedding as well as for population. Last five columns related to unempoyment rate and life expectancy are i bit different. That can be mainly noticed on boxplots. Boxplots for those indicators are very regular. There are not too big differences in districts in terms of unemployment rate and life expectancy. Besides histograms and boxplots, also other statistical tools were used. I also checked probability density functions and empirical cumulative distribution functions using ggplot. Furthermore, the quantile-quantile (q-q) plots were examined which are a graphical technique for determining if a data set comes from normal distribution. Lastly, I have looked at dependacies within columns using correlation matrices.

As the aim of this project was to perform both Explanatory Data Analysis (EDA) of selected statistical indicators and the plot of these indicators in the form of maps. Therefore, after exploring the data I created several cartograms of these statistical indicators using *tmap* package that are inherent in the last tab.