**Introduction to Machine Learning**

**Major HW 1**

Submitted by: Sofia Blyufer 321128019 and Dor Rozen 318965365

13. The correlation table to the features left is presented below.   
Chart

Description automatically generated

To understand the correlations, we plotted histograms of highly correlated features. The pairs we decided to check were those with correlation 0.8 and above.

For two pairs we’ve received perfect correlation: steps per year vs. age group and number of cousins vs. age group. The plots below confirm these numbers.

Chart, scatter chart

Description automatically generated

Therefore, by keeping the feature of age group, number of cousins and steps per year do not add any new data. From the plots we can also see that there is no need of manipulation and taking a combination of the data (look the same).

Hence, we decided to remove ‘NrCousins’ and ‘StepsPerYear’, while keeping ‘AgeGroup’.

Before showing the next features, it should be noted that as previously explained, missing data was filled with mean. As a result, peaks in the mean are observed. Consequently, those effects slightly “ruin” correlations by creating a “cross”. For example, below is a plot of household expenses of parking tickets per year vs. studying per day. On the left is the original data, and on the right is the data normalized to z-score values, after filling missing data and outlier removal. For that reason, we also look at the plots before filling the data. Nevertheless, the correlation is worse after adding the “cross”, and therefore we are only being more strict. The correlation of the pair after filling the missing data is 0.91.

Hence, we decided to remove “StudingPerDay”, as it doesn’t add much more value to “HouseholdExpenseParkingTicketsPerYear”.

Chart, line chart

Description automatically generated

The other