**Introduction to Machine Learning**

**Major HW 1**

Submitted by: Sofia Blyufer 321128019 and Dor Rozen 318965365

1. See code.
2. In the table below, we can see for each attribute a proposed meaning and a proposed type of attribute.
3. See in the table above.
4. See code.
5. See 2.
6. The strings discussed are “Job”, “CurrentLocation” and “Address”.   
   Job – we have decided to not extract any information from this feature for two reasons. The first, there is more than 25% missing data for this attribute. The second, there are too many unique values (600 out of 3000) and is not informative enough.  
   Current Location – location coordinates can be important to identify risk area. As the data already numerical, which is easiest for further analysis, we have decided to split the data into two attributes – X and Y (x and y coordinates separately).   
   Address – we have seen fit to drop this attribute, as the location information is already numerically convenient in “CurrentLocation”. Address is an ID feature and therefore might only harm our modelling. Using the same reason, we have decided to drop “ID” now.
7. Chart, histogram

   Description automatically generatedAs previously mentioned, job, address and ID were deleted altogether. Additionally, we have dropped “PCR\_11” and “PCR\_15” as they both have above 80% missing values, and imputation would be too inaccurate in such a case. The rest of the features were chosen to be imputed as there is enough data to draw conclusions based on the its distribution.   
   For example, the histogram below is of the attribute “MedicalCarePerYear”. It can be approximated well to a normal distribution. Therefore, for such features we think imputation is relatively straightforward.  
   For other features that had enough data we also decided to impute, using other methods, as will be described below.
8. We have handled our features according to 3 types:
9. See code.
10. Let show BMI as an example:  
    Chart, scatter chart

    Description automatically generated  
    cvb

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

Description automatically generated with low confidence

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 (not normalized) 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 stricter. 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

Another case of correlation is