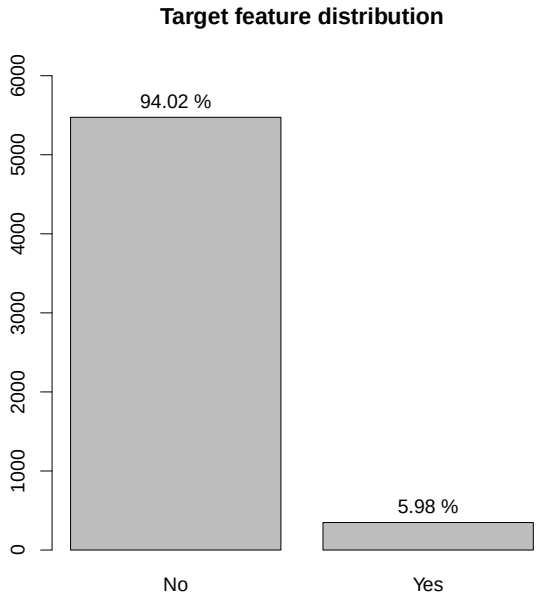
Final Homework Report

NPFL054 Introduction to Machine Learning

Milan Wikarski

Task 1 – Data analysis

Dataset consists of 5822 examples. Each example has 86 attributes. The 86-th attribute *Purchase* with values *Yes (1)* and *No (0)* is the targte attribute.



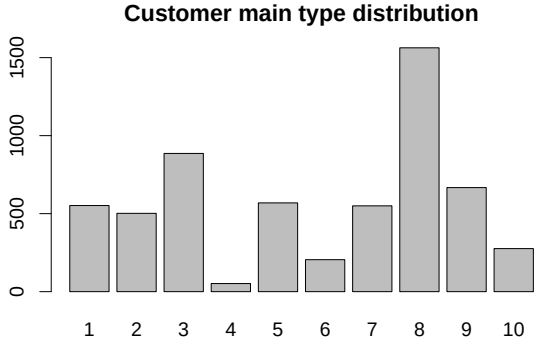
The distribution of the target attribute *Purchase* is heavily skewed towards *No*. The frequency of *Yes* is just 5.98%. This means that the expected precision when randomly selecting 100 examples whould be 5.98%.

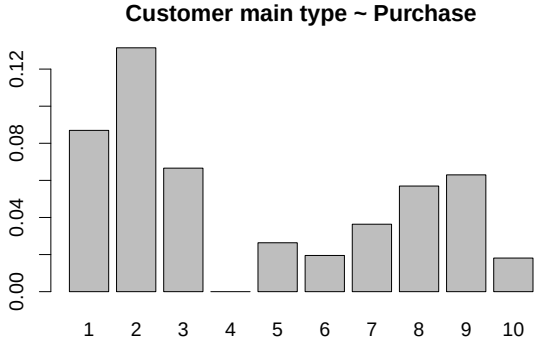
Task 1a

MOSHOODF

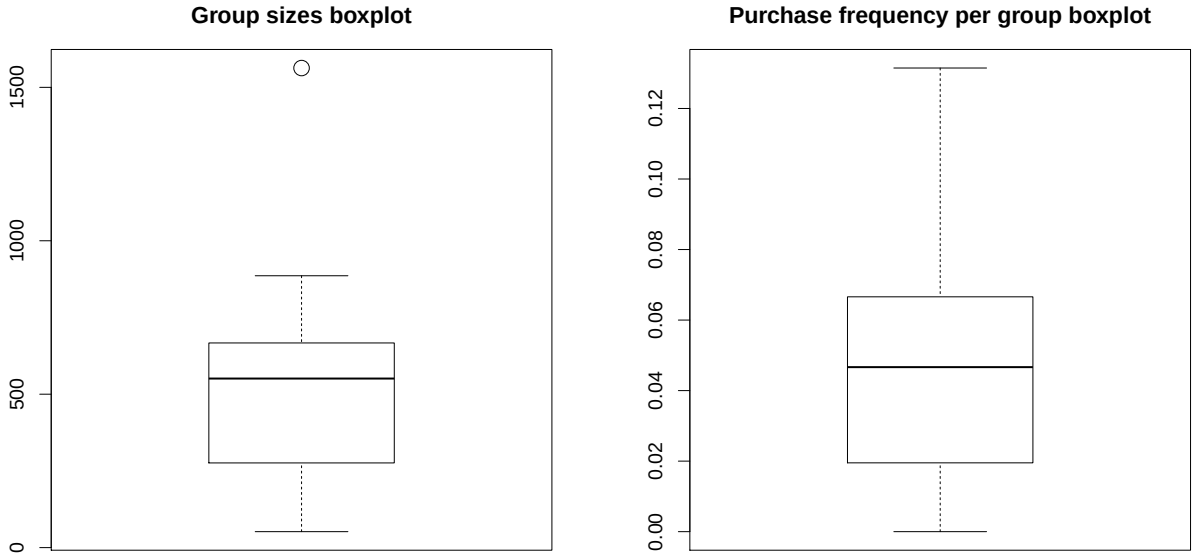
Attribute *MOSHOODF* represents the main customer type. This attribute divides customers into 10 groups described in *L2*. The table below lists the number of customers per group, as well as the percentage of examples in this group with target attribute *Purchase* value *Yes (1)* (ie. the percentage of customers who have purchased the caravan insurance policy):

This data can also be plotted into two barcharts:





We can further analyze irregularities in groups by plotting the data into two boxplots and looking for outliers:

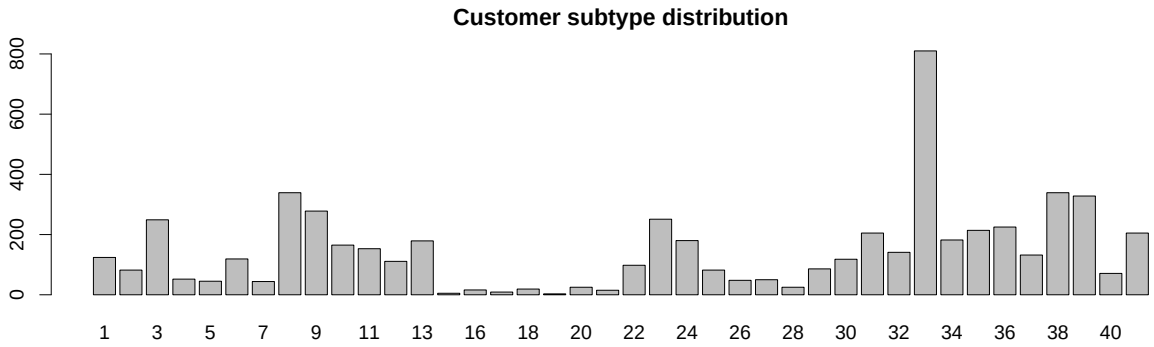


Group *4 (Career Loners)* is really small (only 52 members) and contains no positive examples but cannot be classified as outlier. Group *8 Family with grown ups* is the largest (1563 members), thus can be classified as outlier base on size, but the *Purchase* frequency of this group is slightly below average (5.69%). When it comes to *Purchase* frequncy in groups, there are no outliers.

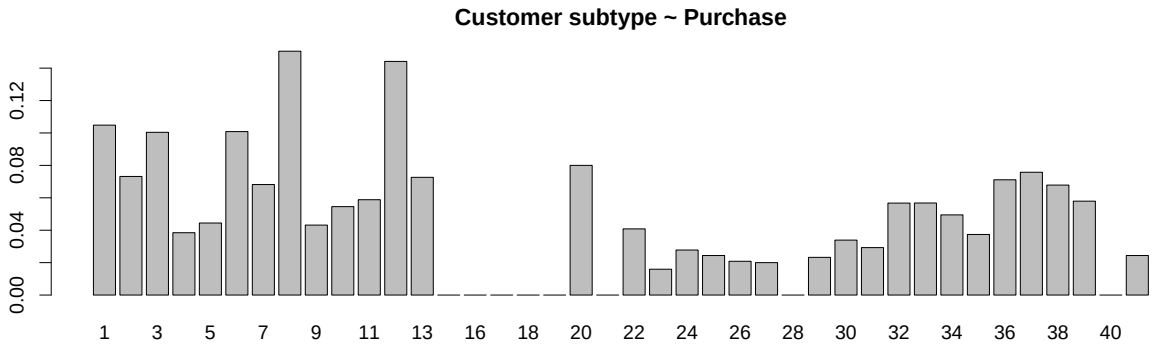
MOSTYPE

Attribute *MOSTYPE* represents the customer subtype. This attribute divides customers into 41 subgroups described in *L0*. The table below lists the number of customers per subgroup, as well as the percentage of examples in this subgroup with target attribute *Purchase* value *Yes (1)*:

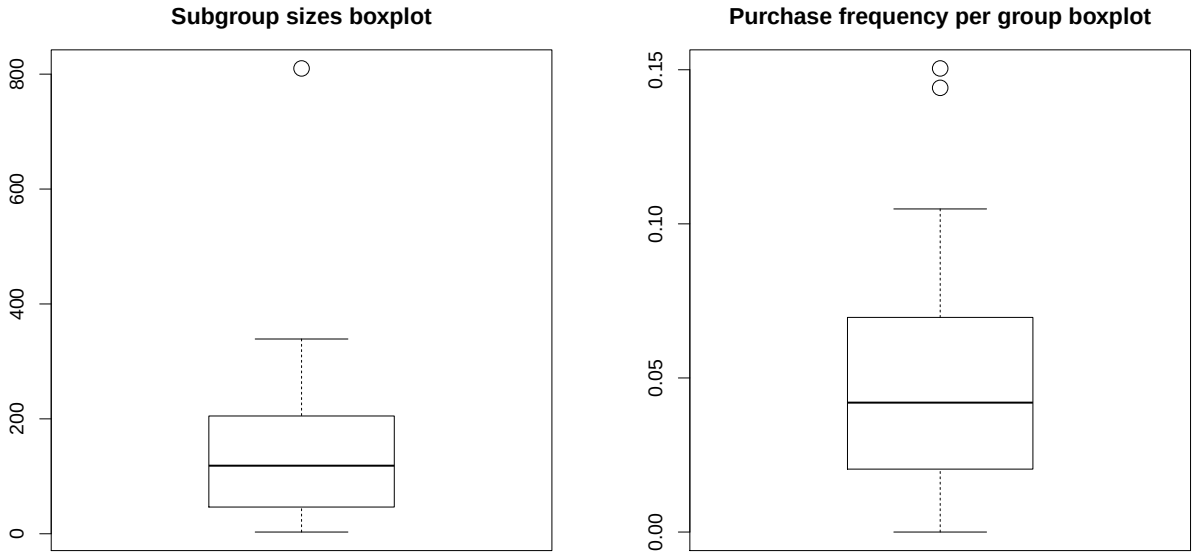
The subgroup sizes can also be visualized using a barchart:



The percentage of people in each subgroup who have purchased the caravan insurance policy can also be plotted to a barchart:



It should be noted that among 5822 examples, there is no representative of subgroup *14 (Junior cosmopolitan)*. To further see, which subgroups might be interessting, we can plot boxplots for size and *Purchase* frequency:



Subgroups *15 – 19* are very small and contain no positive examples but cannot be classified as outliers (the reason for this will be explained in the next chapter). Other subgroups with no positive examples are subgroup *21 Young urban have*-nots, subgroup *28 Residential elderly* and subgroup *40 Residential elderly*. Subgroups *21* and *28* are too small to draw any conclusions but there seems to be a correlation between being a member of subgroup *40* and not purchasing the caravan insurance policy.

The largest subgroup *33 Lower class large families* is composed of almost 14% of all examples but the *Purchase* frequency in this subgroup is slightly below average (5.68%). This subgroup is an outlier based on size.

The most promising subgroups are subgroup *12 Affluent young families* with 111 members and 14.41% Purchase frequency, and subgroup *8 Middle class families* with 339 members and *Purchase* frequency 15.04%, which is the greatest among all subgroups. These two subgroups can be classified as outliers based on *Purchase* frequency.

Task 1b

After some analysis, it can be seen that *MOSHOODF* divides the customers into 10 groups and *MOSTYPE* further divides these customers into subgroups. This information can be gained by calling:

```
table(MOSTYPE, MOSHOOFD)
```

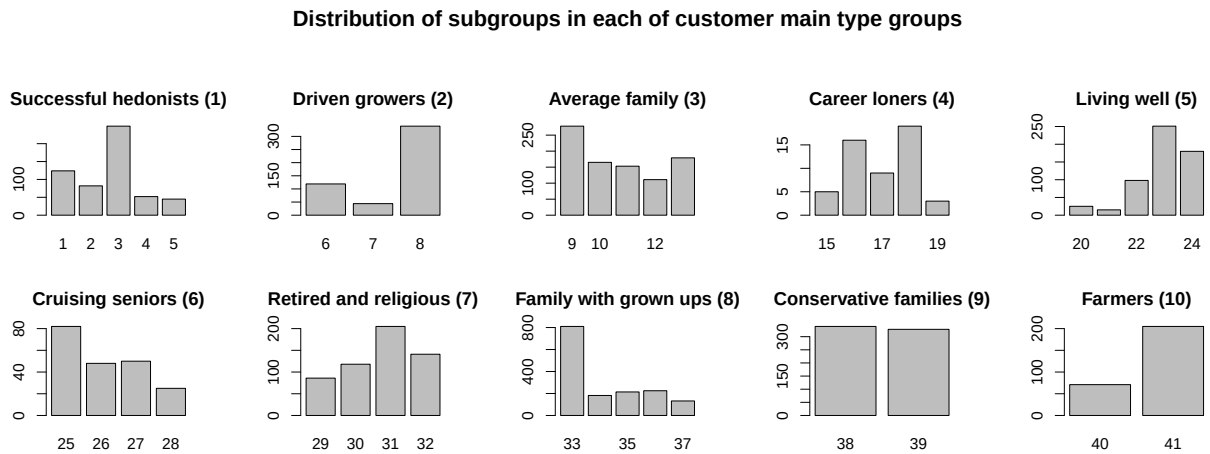
and analyzing the output:

```
        MOSHOOFD
MOSTYPE   1   2   3   4   5   6   7   8   9  10
     1  124   0   0   0   0   0   0   0   0   0
     2   82   0   0   0   0   0   0   0   0   0
     3  249   0   0   0   0   0   0   0   0   0
     4   52   0   0   0   0   0   0   0   0   0
     5   45   0   0   0   0   0   0   0   0   0
     6    0 119   0   0   0   0   0   0   0   0
     7    0  44   0   0   0   0   0   0   0   0
     8    0 339   0   0   0   0   0   0   0   0
     9    0   0 278   0   0   0   0   0   0   0
     10   0   0 165   0   0   0   0   0   0   0
    ...
```

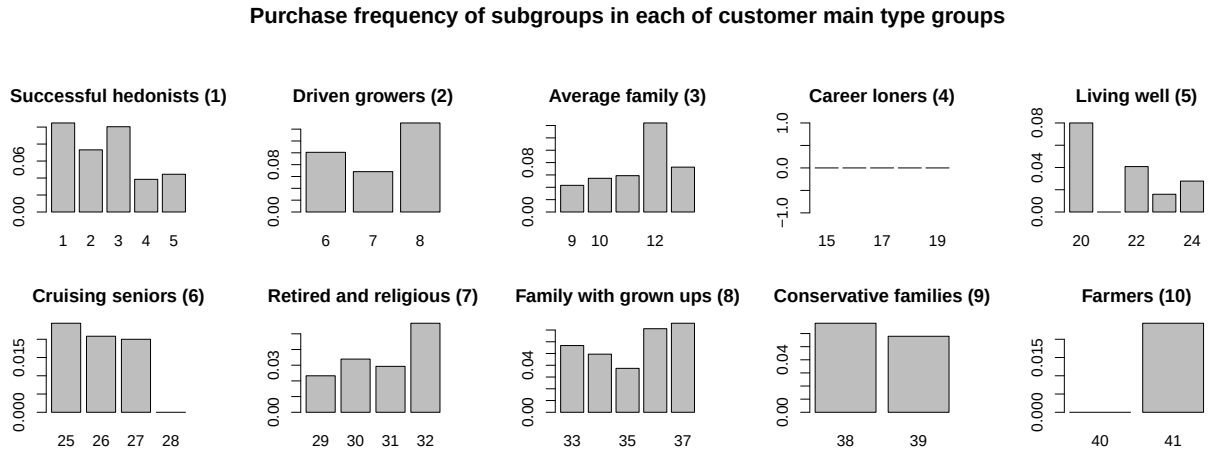
It is clear that all customers of certain subgroup *(MOSTYPE)* belong to just one group (*MOSHOOFD)* and that every customer from each group (*MOSHOOFD)* is assigned one subgroup (*MOSTYPE)*.

The groups and subgroups can be visualized as follows:

With this information, we can explore the groups (*MOSHOOFD)* in more detail by looking at the size of its subgroups:

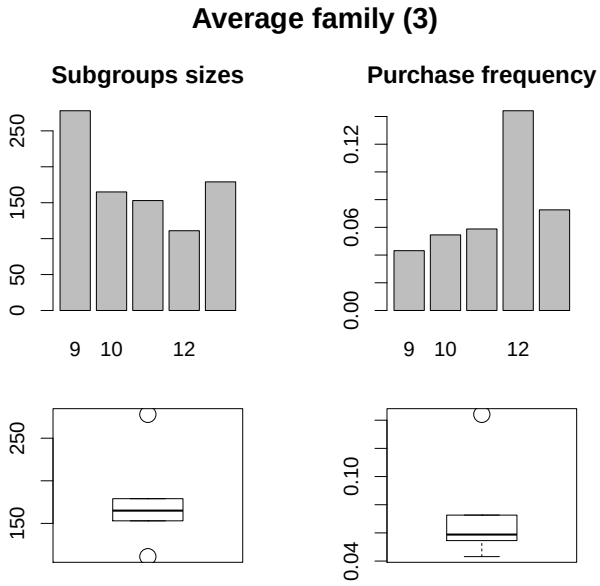


and the *Purchase* frequency of its subgroups:

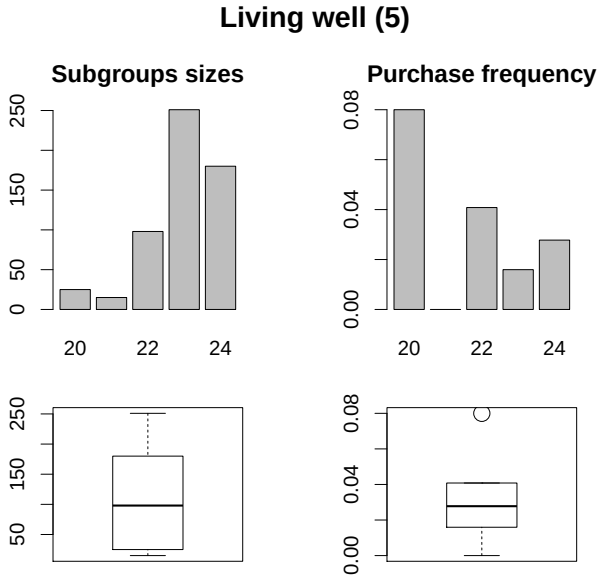


This visuallization is really useful when exlaining why subgroups *15-19* contain no examples. These subgroups are part of group *4 Career loners*, which as was shown in **Task 1a** contains no positive examples.

Groups which are interesting are the ones with irregularities in the *Purchase* distribution. Most notably, the likelihood of the members of group *3 Average family* to purchase the caravan insurance policy is average (6.66%), unless they are part of subgroup *12 Affluent young families*, in which case it is 14.41%.



The same can be said about members of group *5 Living well* and subgroup *20 Ethnically diverse*. Although this subgroup can be classified as an outlier compared to other subgrous in the parent group (based on *Purchase* frequency), it contains only 25 members which drastically decreases its importance.



Charts like these two (group details) were created for each group and can be seen in */out/group-<group number>-detail.pdf*.

Task 2 – Model fitting, optimization and selection

Task 3 – Model interpretation and feature selection

Task 4 – Final prediction on the blind test set