Quality of life in Swiss Cities

Pierre Erbacher
Anthony Piquet
Forough Habibollahi Saatlou

***Abstract*—In this project, we try to design a framework to predict specified indicators of quality of space in Swiss cities. We use insurance data as the input to our model to understand if they can represent or explain quality of urban environments. We analyze the customers data of la Mobilière in our model to efficiently predict different aspects of quality of space in a city. This understanding can help the improvement of social awareness and promote community development.**

I. Introduction

Spatial quality is about strategies, policies, design and effective creation and use of spaces. An effective design of buildings, landscapes and infrastructure in a city will lead to a better functioning society in the future where rising needs of people and organizations are met more effectively. The quality of space is an important factor which shall not be missed even in designing a learning environment which aims to be differentiated from the norm.
Especially during the last decade, in parallel with the technological and scientific developments in the world, empirical researches have been conducted on the livability in urban spaces, people's quality of life and the development of the urban space quality [1]. A model that can predict some key indicators of quality of space such as incomes and jobs, housing conditions, health, mobility, and etc. can be utilized to lead to better future planning. In this effort, we use la Mobilière insurance data from year 2018 for such a prediction. After a detailed exploration of the data, performing various steps of feature engineering and extracting a set of features which could efficiently represent the latent information of our target cities, our goal is to design a model fitted to this data which gives us high accuracy predictions of the specific indicators.

II. Dataset

For this study, we have been provided access, under confidentiality agreements for all members of the group, to la Mobilière insurance data from year 2018. The main research project is a collaboration between HERUS Lab and la Mobilière started from September 2017. The data includes information about costumers' personal information such as gender, age, language, nation, and job state and also information about their belongings which are insured such as houses, cars, the number of claims for each of them, etc. Next, from the Swiss Statistical Office (BFO), 90 indicators for 170 Swiss municipalities (more than 10'000 inhabitants) were collected which include indicators of transportation, work and workplace, housing, fiances, etc.
The aim would be to use the insurance data and try to develop a model that could predict a target set of these indicator with a high precision.

III. Models and Methods

In this section, we will now go into more detail about the various stages of the study and the performed methods towards reaching the goal of the research.

A. Data Preparation

Generally, in data science related topics, the data are far too biased to be used blindly by researchers in particular in the Social Science domain. Hence, we first need to become fully aware of the characteristics of our dataset and get a good understanding of what they represent. We also need to clean our data and remove the possible faulty or unknown values due to some errors while data collection. Eventually, we will be able to assess the cleaned dataset and select proper features for our model.

1) ***Exploratory data analysis****:* As the first step, we manually investigate the dataset and the information it holds about the insurance customers. Next, we select the relevant columns (features) which could potentially reflect some aspects of the data points that embed important information about the target indicators we are aiming to predict. Hence we select the following features [1]:

- *JobState*: Status of employment
- *Civil*: Civil status
- *YearOfBirth*: Year of birth
- *Gender*: Gender
- *Own/Rent*: If own or rent an house
- *Lang*: Speaking language
- *Nation*: Nation of origin
- *Children_0-26*: How many children
- *Car1_Price*: Price of the first car
- *Car1_ClaimsCt5Y*: Number of claims for the first car
- *Car1_ClaimsSum5Y*: Sum of money of claims for the first car
- *Car2_Price*: Price of the second car
- *Car2_ClaimsCt5Y*: Number of claims for the second car
- *Car2_ClaimsSum5Y*: Sum of money of claims for the second car
- *Car_Premium*: Premium class
- *HH_Ins_Sum*: Insured Sum
- *Stand_of_furn*: Standard of furniture

[1]The descriptions are retrieved from the original data description file

- *Rooms*: Number of rooms
- *Build_Ins_Sum*: Insured sum of the building
- *Year_of_constr*: Year of constructions
- *HHaB_ClaimsCt5Y*: Number of claims
- *HHaB_ClaimsSum5Y*: Sum of money of claims
- *HH_and_Bld_Prem*: Premium class
- *Zip*: Zip code of residence
- *BFS*: BFS number
- *City*: The city

In the next step, we simply translate the non-quantified values such as JobState or Gender into simple numerical values representing the different groups. We continue with replacing all the white spaces or *unknown* values with *NA* and then explore through the whole dataset to look into the number and percentage of the data points with *NA* or *0* values for each of the above numerical or categorical features.

2) ***Data cleaning****:* In this step, after analyzing all the features, we need to clean our dataset from unwanted or meaningless values.

Towards this direction, first, we remove the features with 50% or more *unknown* or meaningless *0* values. To give a few examples, we saw that almost 88% of values in *Year_of_constr* feature is Zeros (which is not a valid value). We also explored more realized that the rest of non-zero values still contain a lot of invalid numbers. So, overall we decided to drop this feature. In a similar case for *Car2_ClaimsSum5Y* and *Car2_ClaimsCt5Y* features, we saw that almost 99.2% of the values are equal to zero. This means that there is almost no information in these features for prediction. So, we decided to drop those columns.

Also, in features that had only a small number of *unknown* or meaningless *0* values (less than 1%), we simply dropped the data points corresponding to that values.

At the end of this procedure, we remove *Year_of_constr, Car2_ClaimsSum5Y, Car2_ClaimsCt5Y, and Rooms*.

3) ***Feature Extraction****:* After having the finalized set of chosen features to work with, we still need to handle the missing values in the remaining data points and for the remaining features.

For handling missing values, we apply two different pipelines:

1) Considering a new category 'unknown' when creating dummy variables and also replacing the missing numerical values with the mean value of the whole dataset.
2) Using predictive models (e.g. Linear Regression) to predict the missing values by utilizing the rest of the features for that data point.

In this work, we use only the first approach since after a detailed exploration through the dataset, it can be seen than when the value of a specific feature is *unknown*, all the rest of the selected features are *unknown* as well, meaning that there has been some faulty measurement and that specific data point holds no information for us. For this reason, it is now clear that the second mentioned method for handling the missing variables will not be applicable to this dataset.

4) ***Feature Engineering****:* Eventually, before feeding the features into our model for prediction, we first standardize our features. So we do this by removing the mean and scaling to unit variance. Next, we use PCA as a dimensionality reduction method before training our model.

B. Prediction Model

In the final steps of our project, before training a model on the data, we needed to take into account the fact that we have many customers (data points) for each of the 170 cities in the dataset and we are trying to predict a single value for each indicator corresponding to each city. Therefore, we cannot train our model simply on the whole dataset and we need to group all the data points corresponding to each city and then perform our classification. Next, we also realized that the very different nature of various Swiss cities and hence the varying range of indicators among these cities could make the model unnecessarily complicated. After a few trial tuns and trying to train a model based on all the data points (customers) we had and without considering the probable categories among these cities, we reached very low results indicating that not all cities fit in the same model. To solve this issue, we first classify the 170 cities in accordance to their population and train the model separately in each class of cities and predict the indicators for all the cities in that specific class. This process can act as a similarity measure between cities based on their populations and help to improve the prediction procedure. So we put the cities into 85 classes of 2 according to the similarity between their populations and train the network for each class. Eventually, we will report the mean R^2 score for the predictions of each indicator among all the city classes.

In this project, we will focus on predicting two groups of indicators which are related to **1. Population:** Number of Inhabitants, male, female, age distribution, foreigners, etc. and **2. Work and Workplace:** Unemployment rate, kind of work, etc.

As the predictor, we use several models to eventually be able to choose the most efficient one. In all the methods, we perform a supervised learning procedure with 80% of the data as *train set*. Here is the list of utilized models:

- Multilayer Perceptron (MLPRegressor): Hiddel layer size: 100; Activation Function: *logistic*, the logistic sigmoid function; Solver: *adam*, a stochastic gradient-based optimizer; Learning rate: *adaptive*, keeps the learning rate constant to the initial value chosen as 0.01 as long as training loss keeps decreasing. Each time two consecutive epochs fail to decrease training loss by at least *tol* (tolerance defined as 10e-4), or fail to increase validation score by at least tol. Regularization term: chosen as 0.01.
- Ridge Regression
- Linear Regression

The results were achieved from utilizing all three methods and they are reported and compared in the next section. The two latter methods (ridge and linear regression) produced very similar results and we will only report the scores from the 2 first methods here.

IV. Results

In this section we will report the mean R^2 score for the predictions of indicators in all city classes. These results are produced by using 3 different models and using the indicators in group *Population* which include pop22, po23,...,pop30 and group *Work & Workplace* including trv1, trv2, ... trv11. More detailed descriptions on all of the indicators can be found in the file *swissdatadescription.pdf*[2] provided as part of our project. We should now mention that, after the training, the results had a very interesting and curious characteristic. They showed that among all the 170 cities, we basically have 2 major groups of cities that show very high and rather low results in the process of prediction. These cities are not related by their population index at all. Histograms below demonstrate this statement better for 1 sample indicator from each of two groups. The results showed the similar traits for all other indicators using all the predictor models

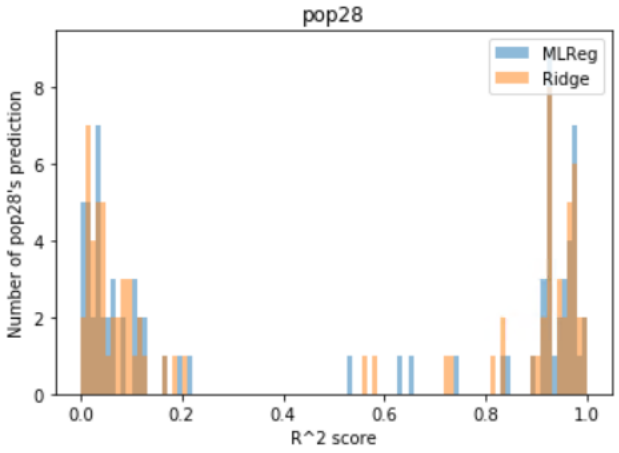


Figure 1: Histogram of prediction scores for pop28 indicator

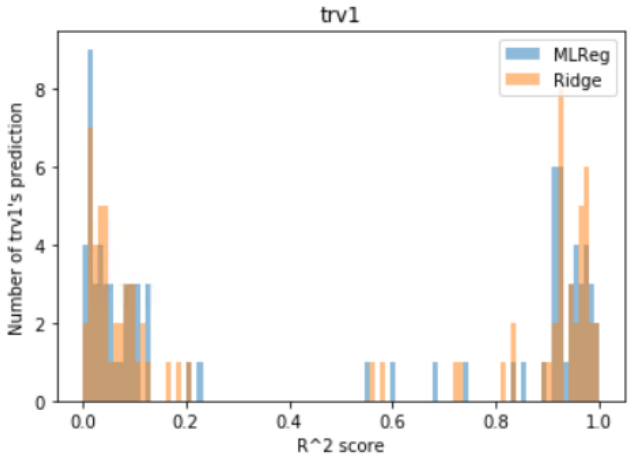


Figure 2: Histogram of prediction scores for trv1 indicator

[2]Due to the confidentiality of the dataset, we are not authorized to provide the original dataset including the features and indicators.

This could suggest that there are some other latent similarities among different cities (other than their population) which could affect the indicators by far. Finding these similarities and categorizing cities in accordance to these new similarity measures between them, could help the training model to fit to the data better.

Table I: R^2 Mean scores for two sides of the histogram for Population indicators

Indicator	MLP		Ridge	
$R^2(\%)$	Mean_1	Mean_2	Mean_1	Mean_2
pop22	5.4295	91.1513	4.5607	90.9148
pop23	5.6918	92.0803	4.4593	91.1974
pop24	5.2008	91.0791	4.5607	90.9148
pop25	4.8745	91.1901	4.5607	90.9148
pop26	5.1462	90.9595	4.5607	90.9148
pop27	5.2494	91.1912	4.5968	91.1420
pop28	4.7693	91.2432	4.5607	90.9148
pop29	5.3975	91.1558	4.5607	90.9148
pop30	5.3470	91.1508	4.5607	90.9148

Table II: R^2 Mean scores for two sides of the histogram for Work & Workplace indicators

Indicator	MLP		Ridge	
$R^2(\%)$	Mean_1	Mean_2	Mean_1	Mean_2
trv1	5.8053	91.2661	4.5607	90.9149
trv2	5.8096	91.1289	4.5607	90.9149
trv3	5.4161	91.0392	4.5607	90.9149
trv4	5.5872	91.1237	4.5607	90.9149
trv5	5.6903	91.1773	4.5607	90.9149
trv6	5.6358	90.9781	4.5607	90.9149
trv7	3.3918	91.2009	5.0499	94.5783
trv8	5.8162	91.0164	4.6312	91.1420
trv9	5.9589	91.1751	4.5607	90.9149
trv10	5.4596	91.0270	4.5607	90.9149
trv11	5.0460	91.1901	4.5607	90.9149

Tables I and II, show the mean prediction scores over all city categories for all the indicators. We have reported two different mean values for each case that correspond to the mean value of the first and second peaks in Figures 1 and 2 and 2 respectively. The similar predicted values in some entries of Table I and II for the predictors is due to the very high correlation between different indicators in each group. The actual values are also highly correlated and similar.

V. Discussions

The reported results in the previous section as well as the results of less efficient methods produced in the procedure by our group members, clearly indicate that much more can be done for a more precise analysis. The high number of indicators and the variety of methods that can be applied give us a very wide research question that can be answered from different perspectives which will require much more time. A framework which predicts indicators one by one (instead of predicting each group of indicators all together) can improve the results while, of course, increasing the cost and time of the analysis. Other training models could be applied and compared

to the ones tested in this work. Also the classification of cities by their population in order to train the model on different classes could be improved or more generalized. This step can also be done considering other possible similarities among cities in accordance to the type of target indicators one wants to predict. Some options are the geographical distance between cities or the age of the city's population.
Furthermore, after the establishment of a final efficient framework, a temporal analysis on data from multiple years could be done in order to achieve a robust model over different years. A comparison between results could then be performed in order to look for meaningful trends in the paths that indicators have changed over the years.

VI. Acknowledgments

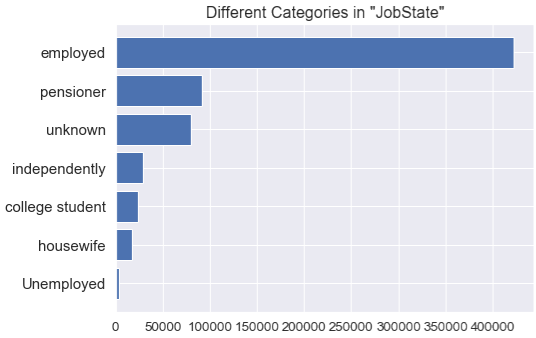
The authors would like to thank Dr. Emanuele Massaro from the LABORATORY FOR HUMAN-ENVIRONMENT RELATIONS IN URBAN SYSTEMS(HERUS) who provided us access to the dataset and supervised us through the project.

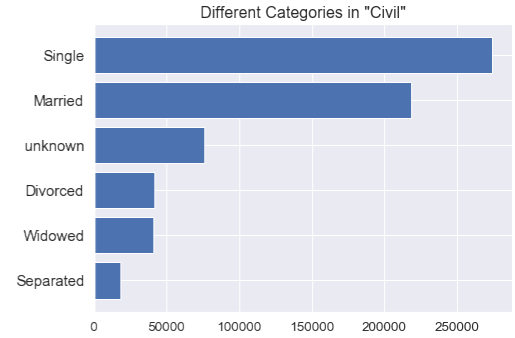
References

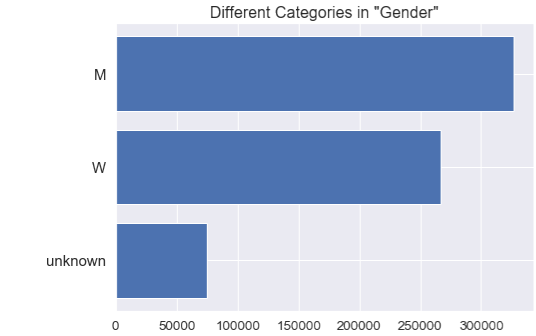
[1] İnceoğlu, Mehmet, and Ayfer Aytuğ. "Kentsel Mekânda Kalite Kavramı." Megaron 4.3 (2009).

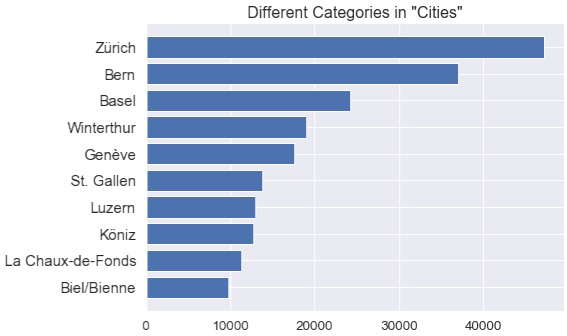
VII. Appendix

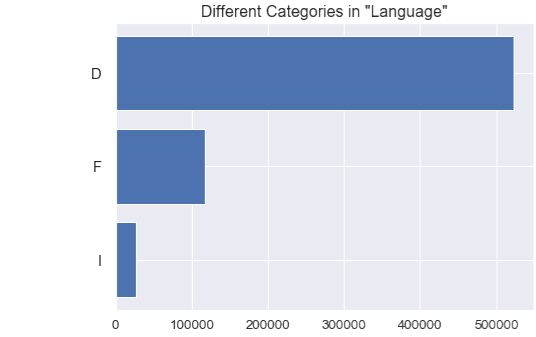
The figures below show some important steps of exploratory data analysis and feature processing which we performed.

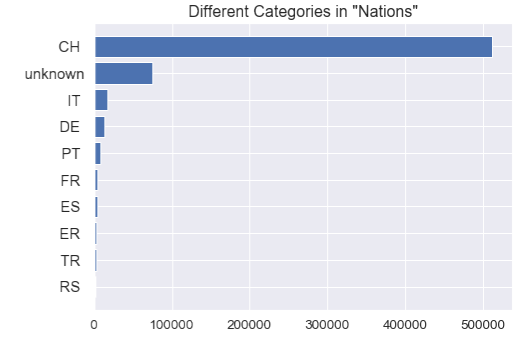




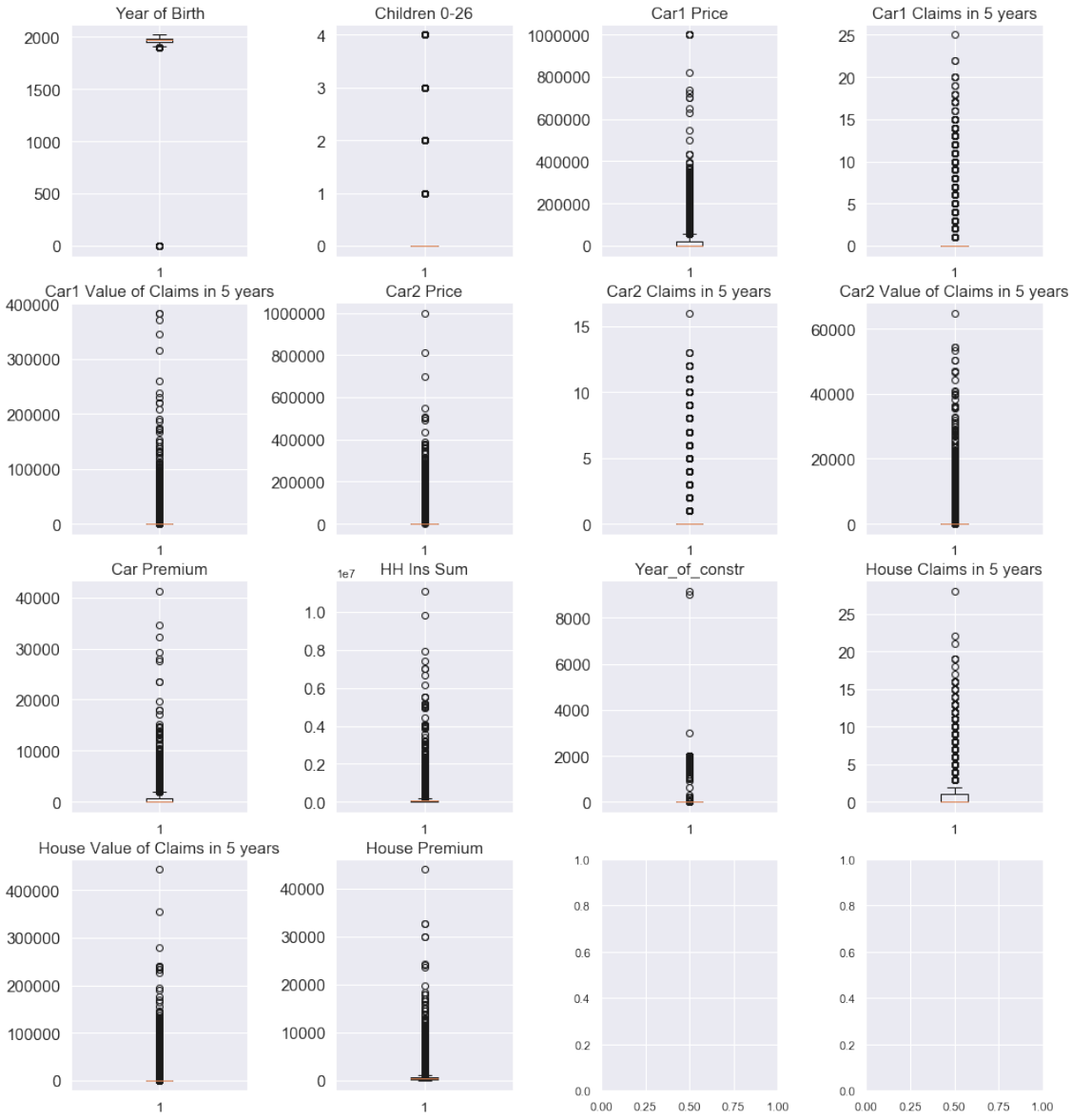




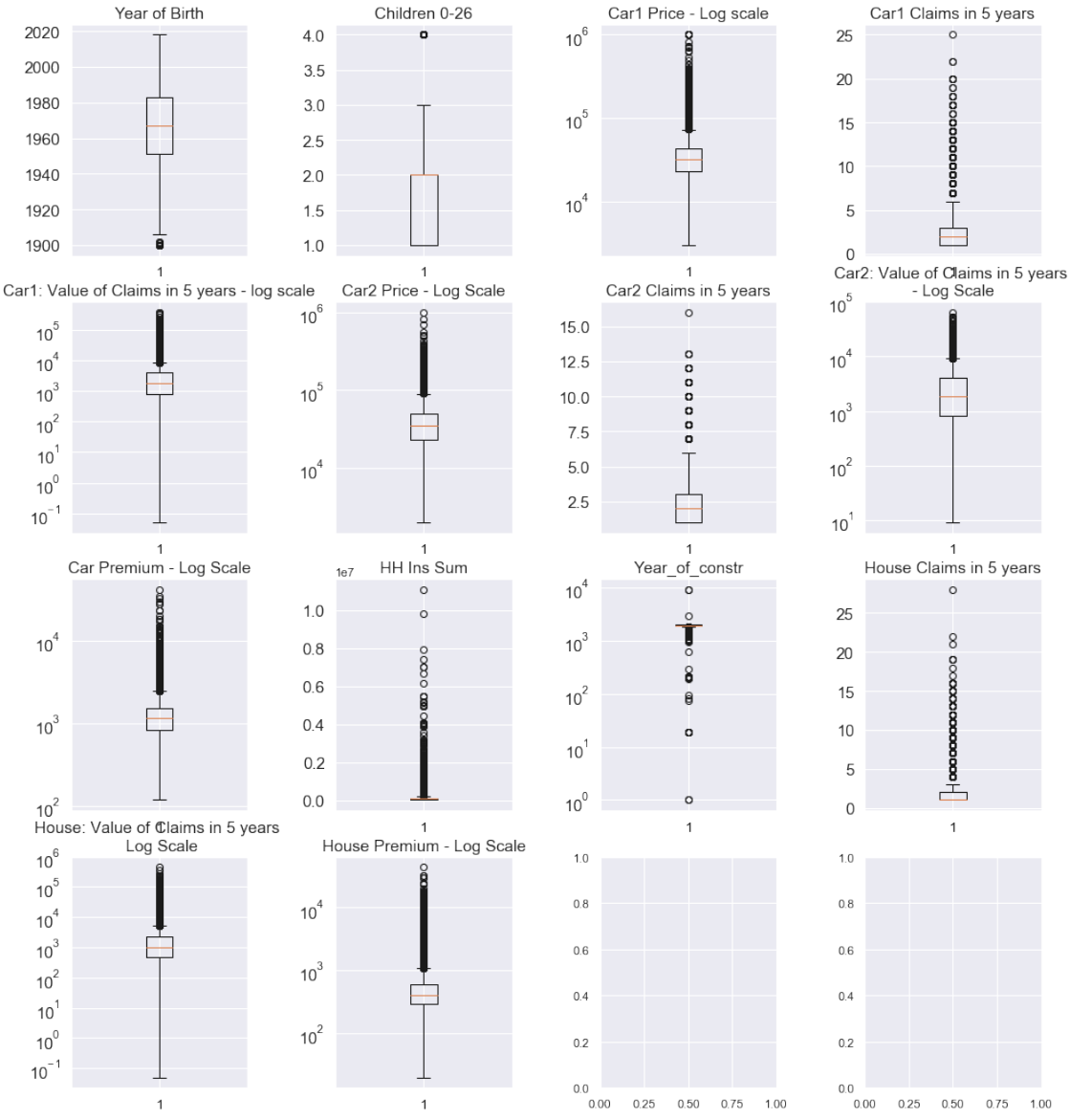


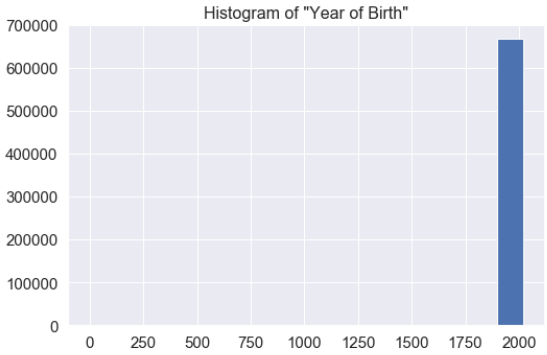


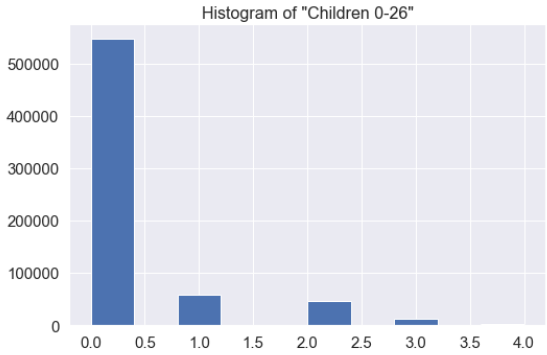
Boxplots for Original Data

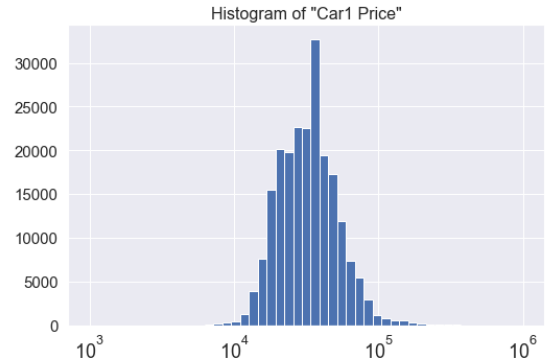


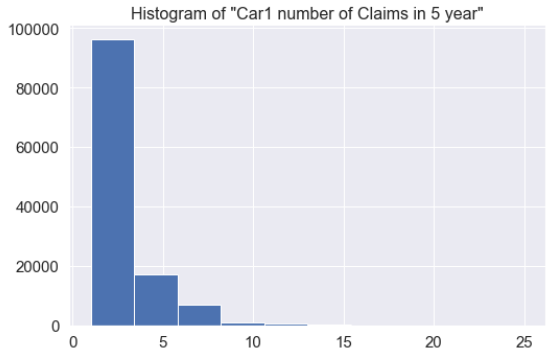
Boxplots with Zeros removed (Log scale when Necessary)

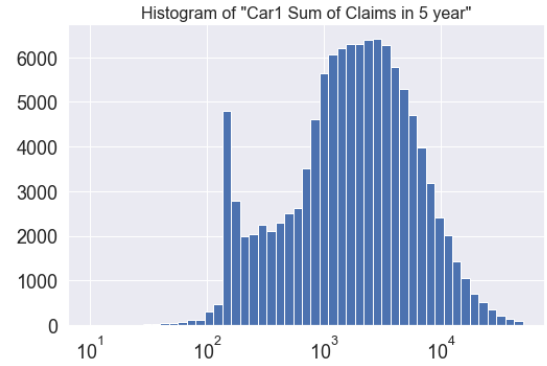


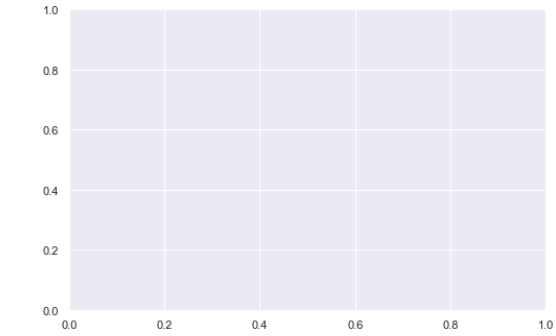

Histogram of "Year of Birth"
700000
600000
500000
400000
300000
200000
100000
0
0
250
500
750
1000
1250
1500
1750
2000


Histogram of "Children 0-26"
500000
400000
300000
200000
100000
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0


Histogram of "Car1 Price"
30000
25000
20000
15000
10000
5000
0
10^3
10^4
10^5
10^6


Histogram of "Car1 number of Claims in 5 year"
100000
80000
60000
40000
20000
0
0
5
10
15
20
25


Histogram of "Car1 Sum of Claims in 5 year"
6000
5000
4000
3000
2000
1000
0
10^1
10^2
10^3
10^4


1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0