Level of motorization in Zurich

What factors influence the number of cars ona a municipality level?

Philipp Eble, Nils Hass

2 February 2022

Introduction

Cars are one of the most popular means of transport in Switzerland. But unlike the public transport, walking or cycling, they lead to more traffic, more air pollution and take away valuable space for green areas and other environmentally friendly means of transport.

The large and, since its invention, increasing number of cars and other motorized vehicles plays a decisive role in this. Just about every municipality in Switzerland must deal with these problems. Therefore, we want to model the level of motorization within the municipalities across the canton of Zurich and investigate the factors affecting it.

Figure 1: Source: https://bit.ly/3BssiKR

1. Data Collection

To answer our research question, we first needed data to calculate the level of motorization for each of the municipalities in Zurich (number of motor vehicles per 1000 inhabitants). We then identified a number of possible influencing factors, which comprises traffic, public transport infrastructure, commercial activity, demographic data and many more. We have tried to find all available data on the structure of the municipality and the infrastructure for cars. Interesting factors, which we unfortunately could not find, would have been: Distance to motorways, number of parking spaces, proximity to shopping facilities and schools.

This data was directly sourced form the cantonal statistics office website via url (www.web.statistik.zh.ch). Due to the identically structured data sources of the statistics office, further factors can be added without problems with the help of the built-in loop function. Finally all data sets are merged together into a single data frame and temporarily saved.

2. Data Preperation

2.1 Initiation

In a first step of the data preparation we shortened the column names and gave them English notations. This gives us a better readability whereby the column order and rows are preserved. Let's have a look at the first entries of the data frame.

Table 1: Columns 2 - 7

No	BFS	Municipality	Year	Restaurant	Puplic_Transport_Volume	Privat_Transport_Volume
1	1	Aeugst a.A.	2011	3	NA	NA
2	1	Aeugst a.A.	2012	3	NA	NA
3	1	Aeugst a.A.	2013	3	970	5130
4	1	Aeugst a.A.	2014	3	NA	NA
5	1	Aeugst a.A.	2015	3	NA	NA

Table 2: Columns 9 - 13

No	PUT_Access_B	PUT_Access_C	PUT_Access_D	PUT_Access_E	PUT_Access_F
1	NA	NA	NA	NA	NA
2	NA	NA	NA	NA	NA
3	NA	NA	NA	NA	NA
4	NA	NA	NA	NA	NA
5	0	21.2	9.7	66.2	0

Table 3: Columns 14 - 17

No	PUT_Access_X	Road_Investments	Total_Tax_Rate	Expenses_Environmental_Protection
1	NA	19	NA	91
2	NA	4	96	95
3	NA	4	96	138
4	NA	0	98	97
5	0	0	98	100

Table 4: Columns 18 - 21

No	Expenses_Traffic	Contribution_to_ZVV	Share_Traffic_Area	Median_Taxable_Income
1	230	156	NA	59200
2	159	153	NA	60300
3	252	147	NA	58600
4	252	141	NA	59700
5	292	143	NA	57700

Table 5: Columns 22 - 26

No	Average_Age	Workplaces	Households	Motorcycles	Cars
1	41.8	137	NA	195	1100
2	41.9	137	797	200	1134
3	42.2	136	805	203	1157
4	42.6	136	796	215	1162
5	42.4	141	809	206	1173

2.2 Transform Variables

In the next data preparation step, we created two categorical variables based on average age statistics of the municipality (Average_Age: "<38","38-42","43-46",">48") and the public transport indicators (PUT_Access_: "A", "B", "C", "D", "E" "F", "X").

These variables allow us to segment data points by seniority of the citizens and by the degree of connectivity to public transport. Additionally, we created a normalized variable for the number of vehicles on the road per 1000 inhabitants (DOM: degree/level of motorization) and a date variable using a ISOdate format.

Table 6: New Features

No	PUT_Access_Category	DOM	Date
5	E	694.3605	2015-12-31
14	E	579.5639	2015-12-31
23	D	581.1725	2015-12-31
32	D	666.2831	2015-12-31
41	F	599.7264	2015-12-31

2.3 Missing Values

To assess the data quality of our data set, we created a NA-value table counting the number of missing values per variable. We discovered that missing values emerge exclusively from infrequent (non-annual basis) reporting or from missing municipalities.

Hence, our first step was to gather the missing municipality identifiers and create a vector containing the corresponding IDs. Subsequently, we filtered our data set to exclude these municipalities based on informative variables which had a high enough data quality.

Variables with too many missing values were subsequently removed from the data set resulting in 0 remaining NA values for our analysis.

Table 7: NA-Values per Year/Column

Year	2015	2016	2017	2018	2019
No_NA	0	0	0	0	0
BFS_NA	0	0	0	0	0
Municipality_NA	0	0	0	0	0
Restaurant_NA	0	0	0	3	0
Puplic_Transport_Volume_NA	174	4	172	9	162
Privat_Transport_Volume_NA	174	4	172	9	162
PUT_Access_A_NA	12	10	10	9	0
PUT_Access_B_NA	12	10	10	9	0
PUT_Access_C_NA	12	10	10	9	0
PUT_Access_D_NA	12	10	10	9	0
PUT_Access_E_NA	12	10	10	9	0
PUT_Access_F_NA	12	10	10	9	0
PUT_Access_X_NA	12	10	10	9	0
Road_Investments_NA	0	0	0	3	0
Total_Tax_Rate_NA	5	4	4	5	0
Expenses_Environmental_Protection_NA	0	0	0	3	0
Expenses_Traffic_NA	0	0	0	3	0
Contribution_to_ZVV_NA	4	4	2	0	0
Share_Traffic_Area_NA	174	172	172	3	162
Median_Taxable_Income_NA	0	0	0	3	162
Average_Age_NA	0	0	0	3	0
Workplaces_NA	0	0	0	3	0
Households_NA	0	0	0	3	0
Motorcycles_NA	0	0	0	3	0
Cars_NA	0	0	0	3	0
Population_NA	0	0	0	3	0
PUT_Access_Category_NA	12	10	10	9	0
DOM_NA	0	0	0	3	0
Date_NA	0	0	0	0	0

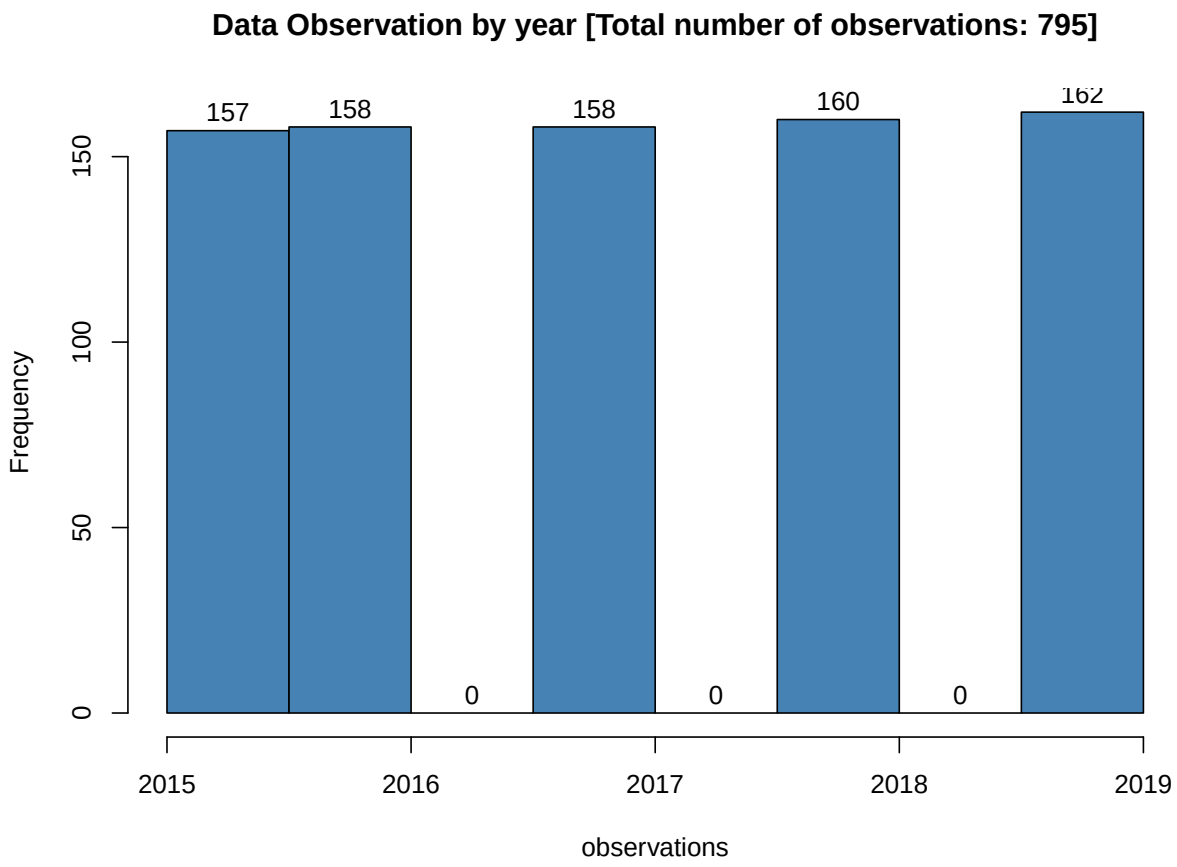
3. Exploratory Data Analysis

3.1 Structure

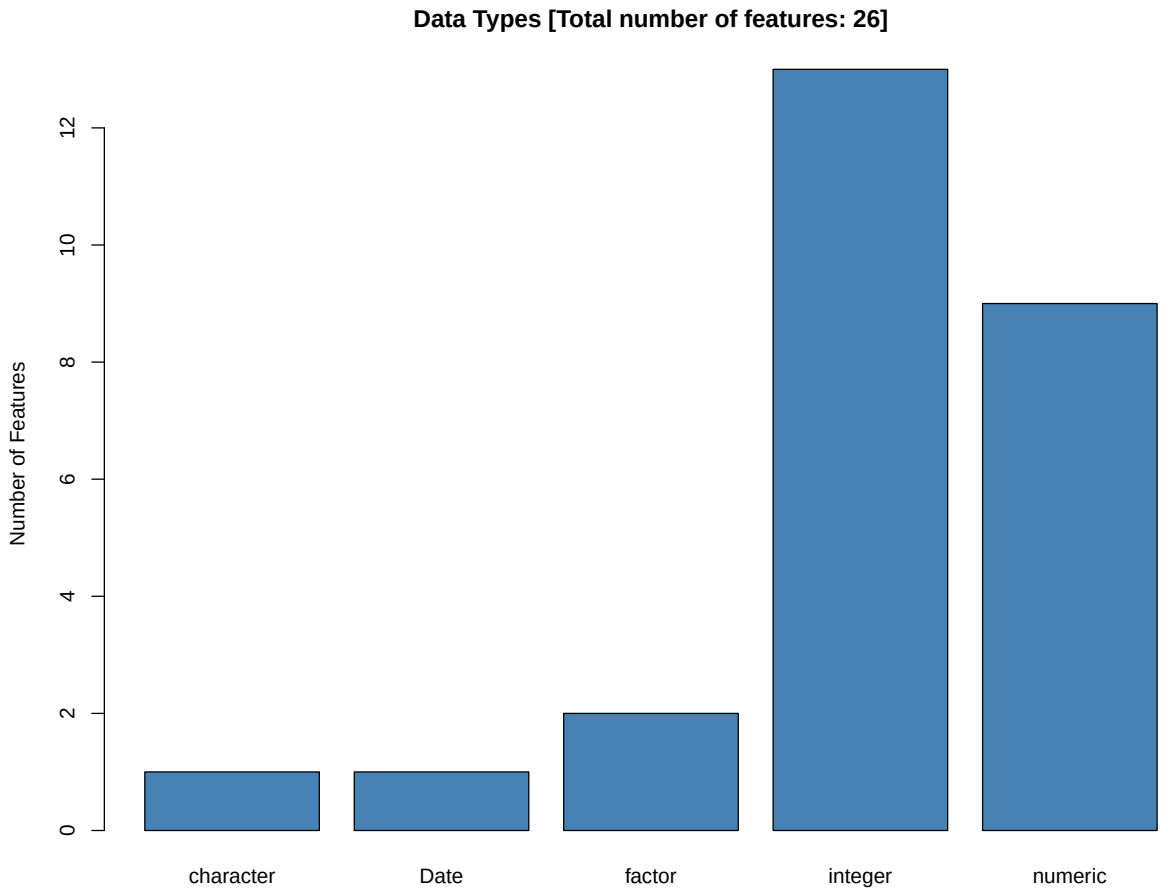
After the data preparation and cleaning, we can take a look at the structure of our data frame using different plots and the summary function. For example, let's take a look at the measure of location and variation for the variable 'Population'.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     336    1834    4341    9170    7734  419012
```

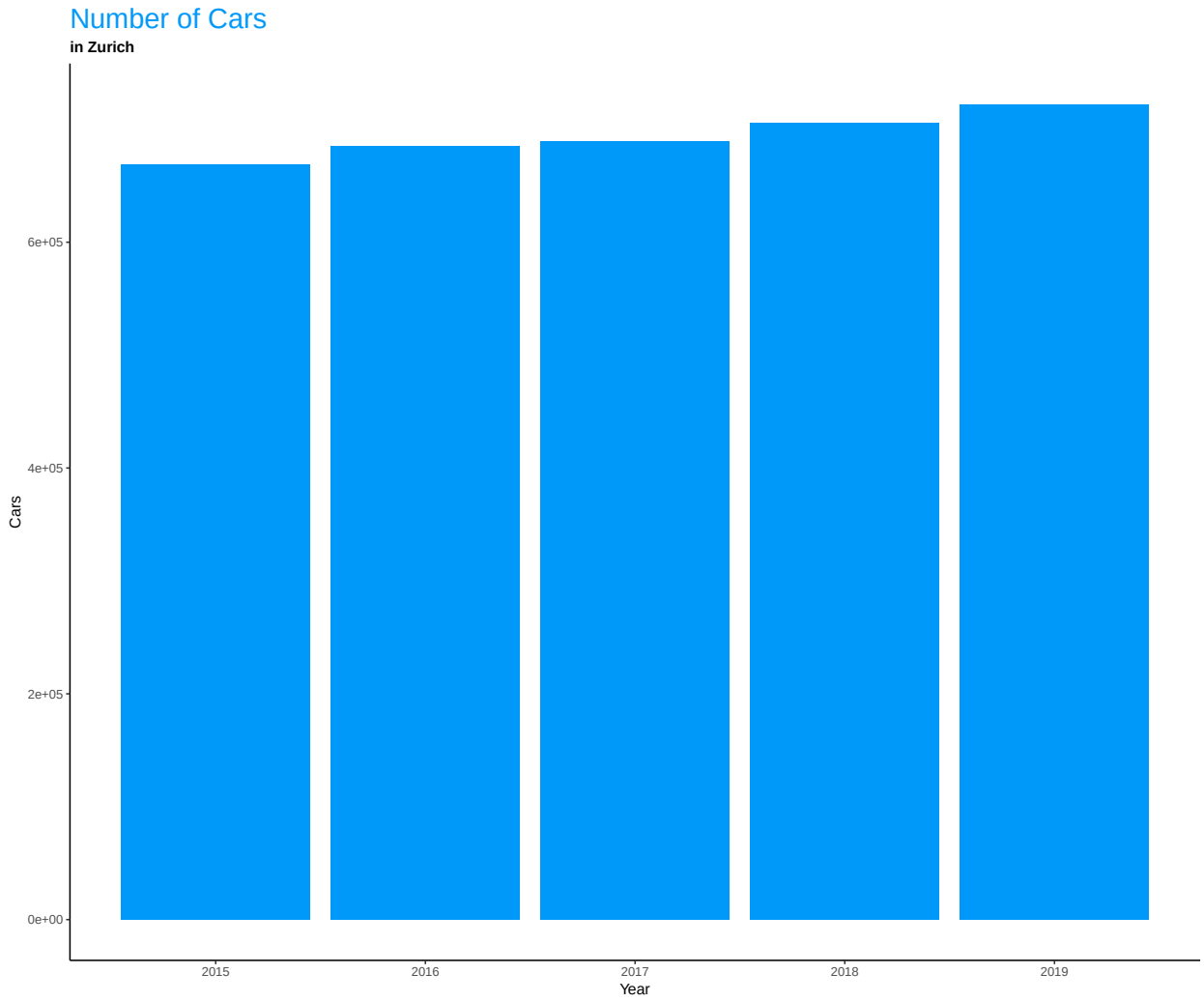
Let's inspect the data further and move on to the visualizations:



As we can see from the first histogram, our data frame contains 795 observations, which are evenly distributed over the years.



Besides integer, numeric and character features, the data set now also contains date & factor features.



With the last graph, we show the addressed, rising trend in the number of cars in Zurich.

3.2 Correlations & Density

To get a sense of the data variables and their correlations, we created a shiny application that allows us to select two variables, view their correlation, their density plots and mean values. This was done to get a preliminary overview of the relationships in our data and identify outliers.

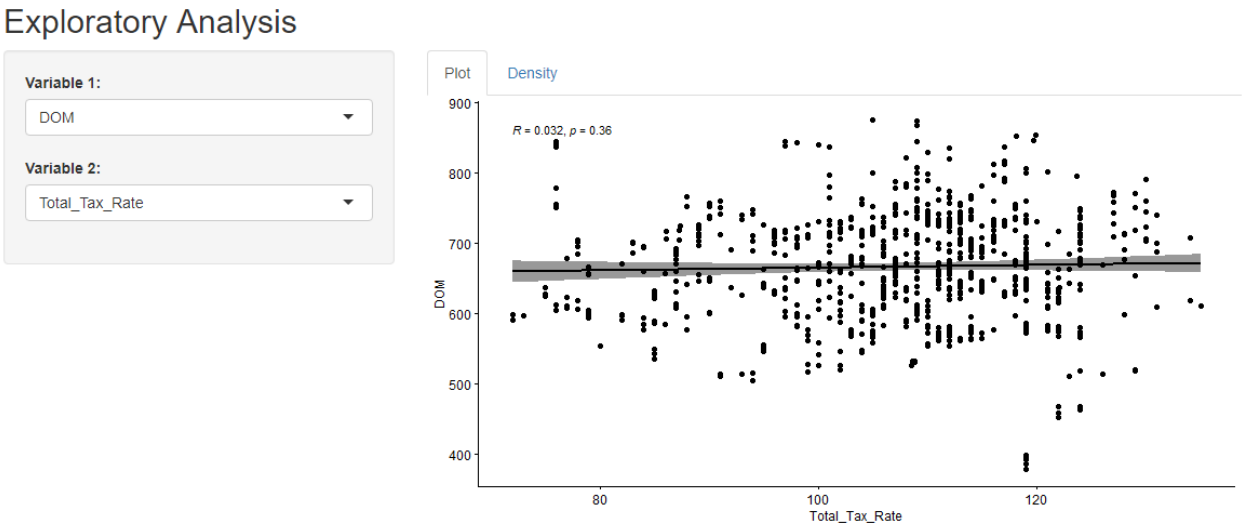


Figure 2: Screenshot ShinyApp: Correlation Plots

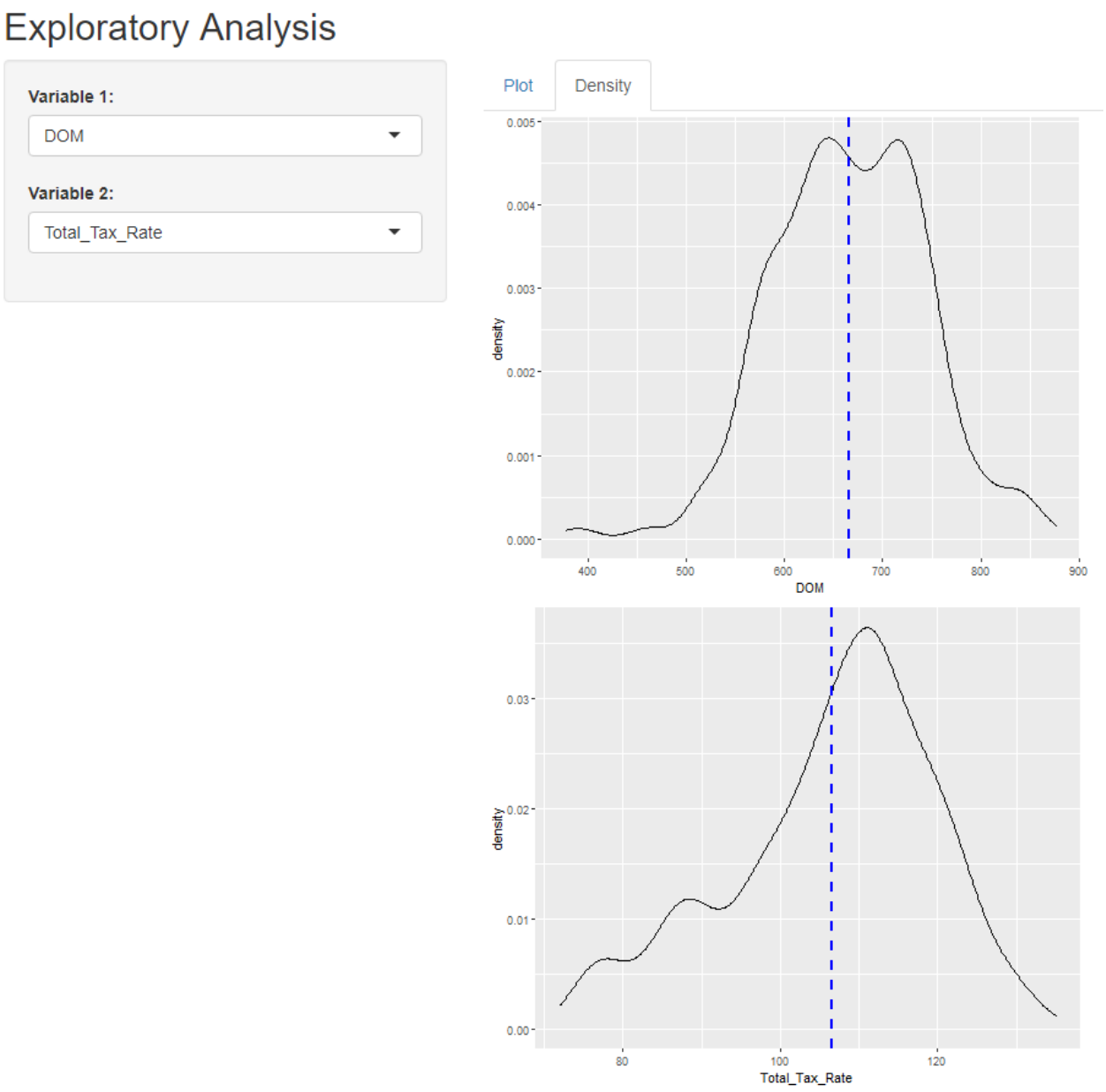


Figure 3: Screenshot ShinyApp: Density Plots

3.3 Outliers

Using the shiny app we discovered that our results were strongly distorted by some of the more populous and resourceful municipalities in Zurich which relatively had far larger values.

We therefore proceeded to remove these values, thereby attaining more balanced variables better suited for linear modelling of the underlying effects in the data.

4. Data Modeling

4.1 Simple linear regression

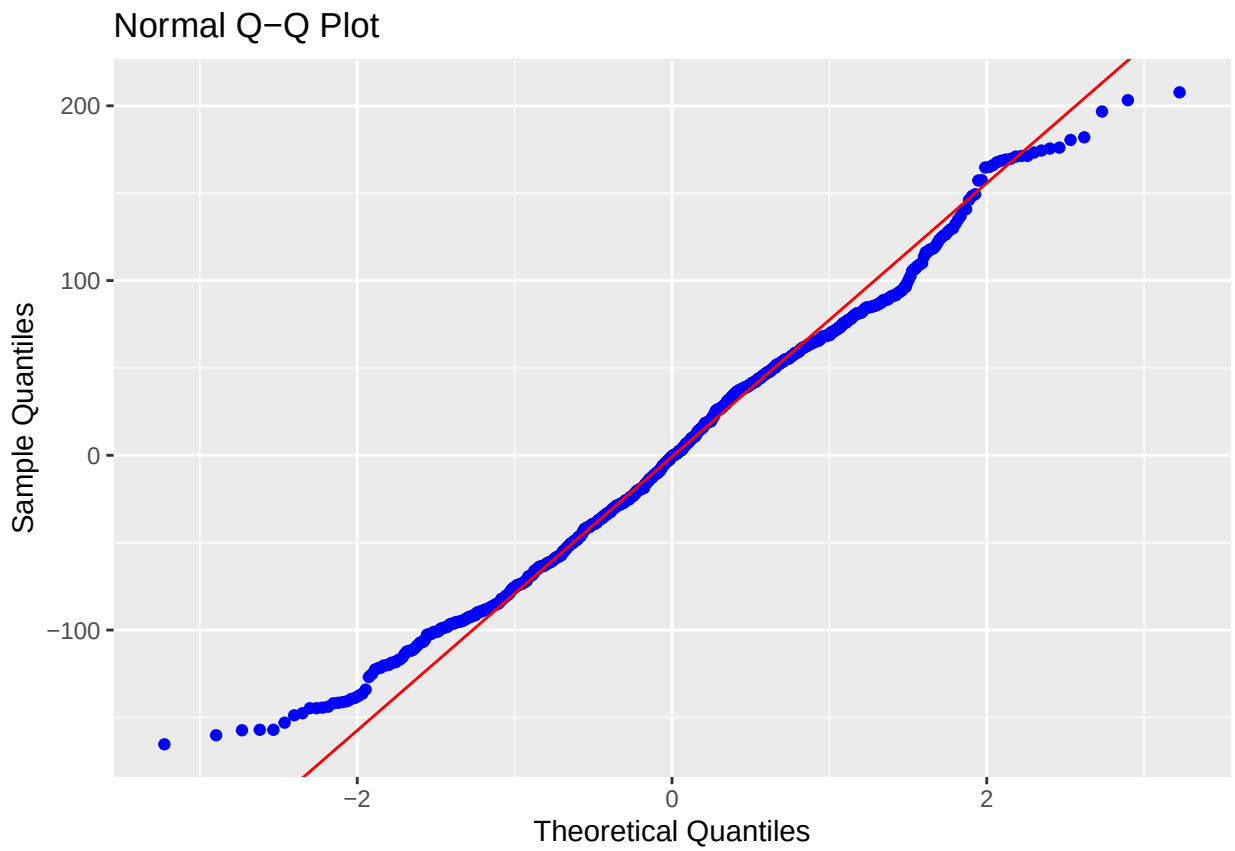
From this, we set out to develop a simple and multiple linear model of the motorization level.

Our first assumption is that the number of workplaces affects the level of motorization. To answer this, a simple linear regression fits pretty well. Our independent (predictor) variable, we are interested in, will be DOM (Degree/Level of motorization). We plot DOM as a function of Workplaces (dependent variable) and get our first regression model:

```
##
## Call:
## lm(formula = DOM ~ Workplaces, data = df.final)
##
## Coefficients:
## (Intercept)   Workplaces
##  672.004885    -0.008149
```

DOM = 672.290871 -0.008469 * Workplaces.

To be noted is that linear regression makes several assumptions about the data at hand. With the QQ plot and histogram we visually check the normality assumption. In our case, almost all the points fall approximately along this reference line, so we can assume normality. The correlation between our observed residuals and expected residuals shows a correlation of nearly 1.As we can see there is a negative correlation between the independent variable and the dependent variable. This suggests that as the number of Workplaces increases, the level of motorization tends to decrease. In other words, if there are more workplaces in a municipality, then there are also fewer vehicles there.



4.2 Multiple linear regression

We applied the AIC stepwise regression technique for feature selection of our multiple linear regression model. This technique estimates in-sample prediction error and penalizes the addition of features.

The technique yielded traffic expenses, restaurants, environmental protection expenses, road investments and the PUT Access categorical variable as the most significant features for our model.

While the effect of traffic expenses on motorization levels do not appear to be statistical significant, we see a significant positive relationships between motorization levels and environmental protection expenses as with each unit invested, motorization is roughly 0.43 units greater. Significant negative relationships are encountered for road investments and restaurants, -0.037 and 0.225, respectively. For the categorical variable Category B, C and D show significant positive relationships.

The interpretation of the categorical variable is different as we have proportional data ranging from 0 to 1. Hence, the coefficients represent percentage increases that describe positive significant relationships particularly between B and C with motorization levels.

```
##
## Call:
## lm(formula = DOM ~ Expenses_Traffic + Restaurant + Expenses_Environmental_Protection +
##     Road_Investments + PUT_Access_Category, data = df.final)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -162.987  -45.238    0.127   41.231  194.997
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     536.20487   29.38583  18.247  < 2e-16 ***
## Expenses_Traffic                  0.00839    0.02660   0.315 0.752570
## Restaurant                       -0.22262    0.03009  -7.398 3.57e-13 ***
## Expenses_Environmental_Protection   0.43061    0.07089   6.074 1.94e-09 ***
## Road_Investments                 -0.03766    0.01618  -2.328 0.020182 *
## PUT_Access_CategoryB            162.35642   29.48109   5.507 4.95e-08 ***
## PUT_Access_CategoryC            134.91552   28.75923   4.691 3.20e-06 ***
## PUT_Access_CategoryD            110.38824   28.44019   3.881 0.000113 ***
## PUT_Access_CategoryE             87.73230   28.56092   3.072 0.002202 **
## PUT_Access_CategoryF             39.02610   29.72187   1.313 0.189553
## PUT_Access_CategoryX             65.29649   39.62230   1.648 0.099759 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62.83 on 784 degrees of freedom
## Multiple R-squared:  0.3422, Adjusted R-squared:  0.3338
## F-statistic: 40.78 on 10 and 784 DF,  p-value: < 2.2e-16
```

5 Chapter of choice

5.1 Municipality Map

In our chapter of choice we tried to display the level of motorization with the help of geographical data respectively a .shp-file. The level can be indicated with a slider and with the help of "shiny" the user is shown the municipalities which are above the indicated level of motorization.

The .shp-file was downloaded from https://www.stadt-zuerich.ch/portal/de/index/ogd/werkstatt/shp_shapefile.html.

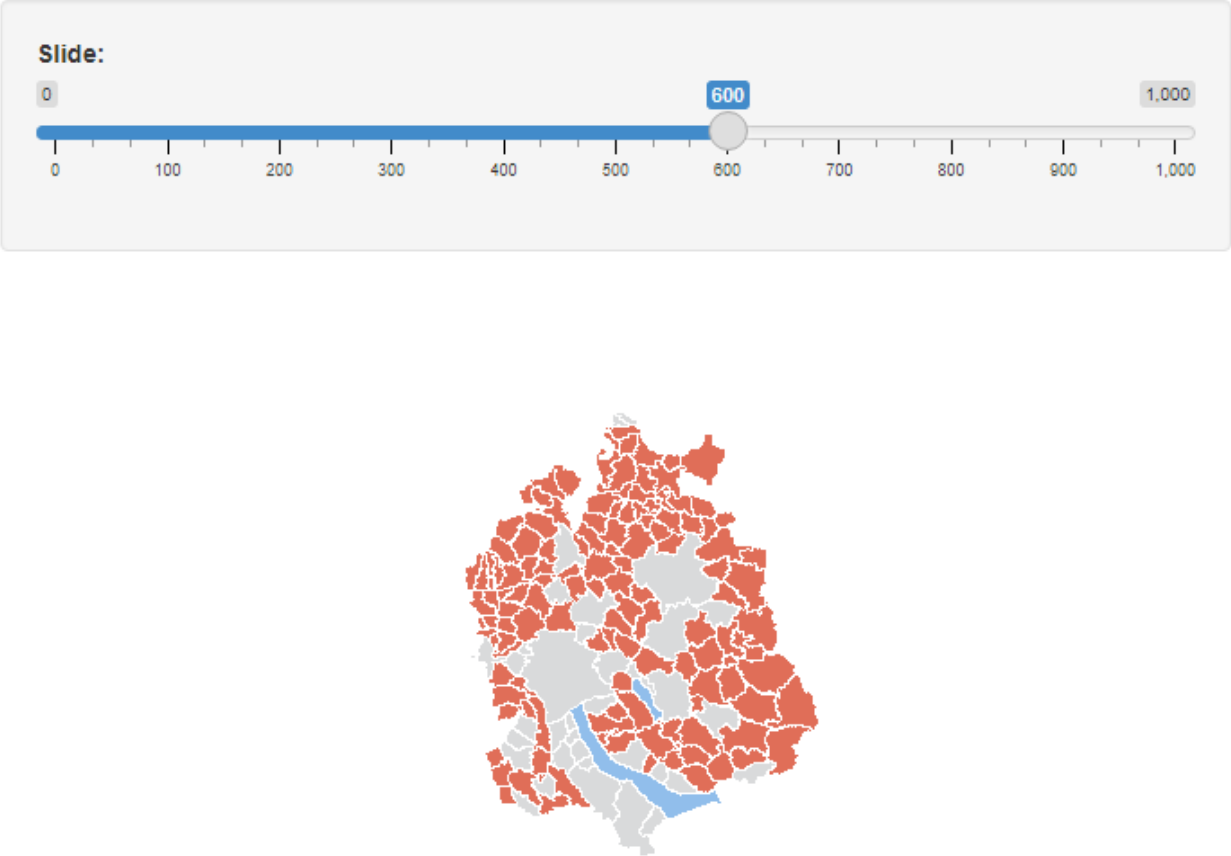


Figure 4: Screenshot ShinyApp: degree of motorization

Conclusion

Our research has shed light on some of the factors impacting the levels of motorization from data sets of the cantonal office of statistics in the canton of Zurich. We have demonstrated that public and commercial spaces such as restaurants and workplaces impact the level of motorization in communities. Additionally, municipal investments in traffic management and environmental protection appear to be related with the number of vehicles in the canton. Finally, also the class of remuneration for public transport particularly for municipalities with a high proportion of class B and C tend to have high levels of motorization.

The insights of our report can be used in policy discussions at the municipal and cantonal levels to address the expansions of individualized rather than shared public transport which is more economical and environmentally friendly.