Assignment R-Bootcamp

Lukas Niederhaeuser & Daniel Podolecki

25 February 2022

Analysis of Road Accidents in Switzerland

For this analysis a data set is provided by the *"Bundesamt für Strassen ASTRA"* in Switzerland. This data set contains accidents from 2011 to 2020, including for example the severity of the accident, whether a motorcycle, pedestrian or a bicycle was included.

This analysis is focused on accidents involving bicycles as well as animals. Furthermore the data set is getting extended with the population data from the 26 cantons of Switzerland and the LV03 Coordinates of the Cantons for a geographic map of Switzerland. Lastly, we added a fictitious data set to work with missing values. We assume that AXA gave us an excel file that records the amount of damage per accident event.

Reference to the data sources

1. **Road Accidents**
2. **Population of Switzerland**
3. For the map: **G1K09.shp**
4. For the map: **LV03 coordinates** for an exemplary city within a canton.
5. AXA damage reports per road accident (fictitious dataset).

Use Case

In our use case we both work for the *"Bundesamt für Strassen ASTRA"* as Data Analysts. We receive two different requests about the influence of traffic accidents in Switzerland.

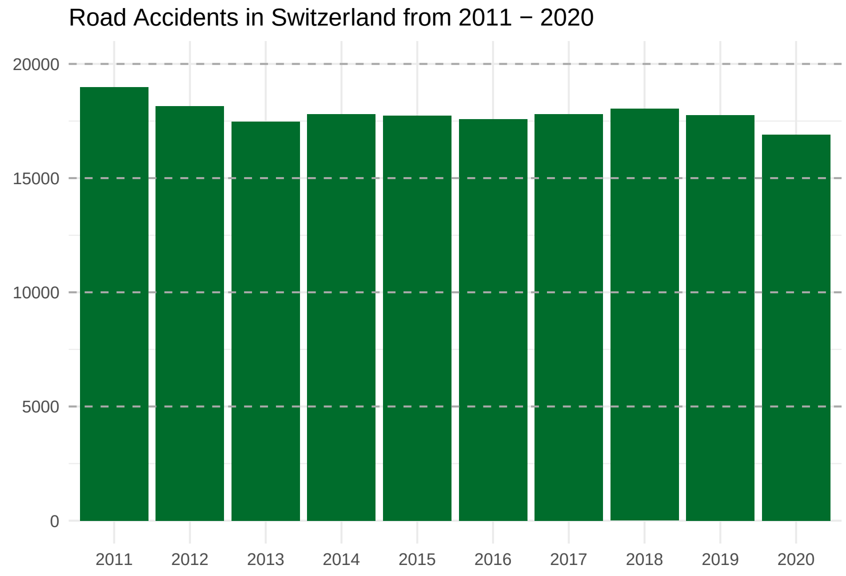
1. **The first request** comes from the nature conservation organization *"Pro Natura"*, which would like to work with the "Bundesamt für Strassen ASTRA". Pro Natura needs an analysis of all wildlife accidents in Switzerland. They ask which cantons are particularly affected by wildlife accidents, at which time of the day and in which months there is the highest danger. Based on this, they want to push for stronger measures in the respective cantons to better protect wildlife. Such could include, for example: Building wildlife bridges or expanding those, introducing speed limits, or putting up traffic signs. They are also interested in whether their past measures have been successful and whether the number of wildlife accidents has already decreased.

2. **The second request** deals with bicycle accidents in Switzerland. ASTRA would like to have more information on where accidents involving bicycles are occuring, specifically in which canton. Further more they would like to have a brief overview of the accidents which involve bicycles. With this information ASTRA will then be able to reach out to the cantons and ask them to further analyse the bicycle-accidents and the possible indicators of these accidents.

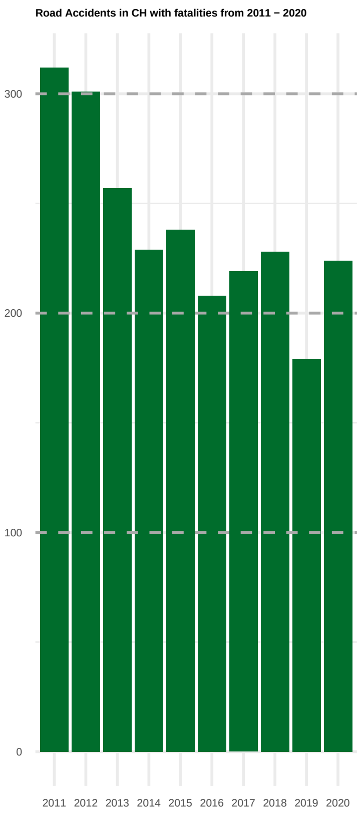
Figure 1: A wildlife bridge to reduce accidents with animals.

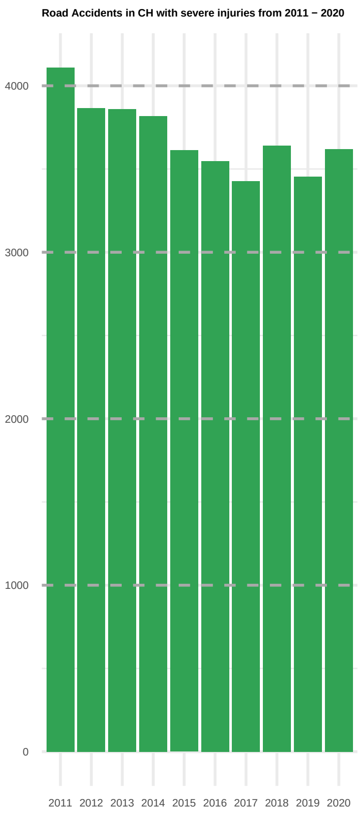
1. Analysis of Total Road Accidents in Switzerland 2011 - 2020

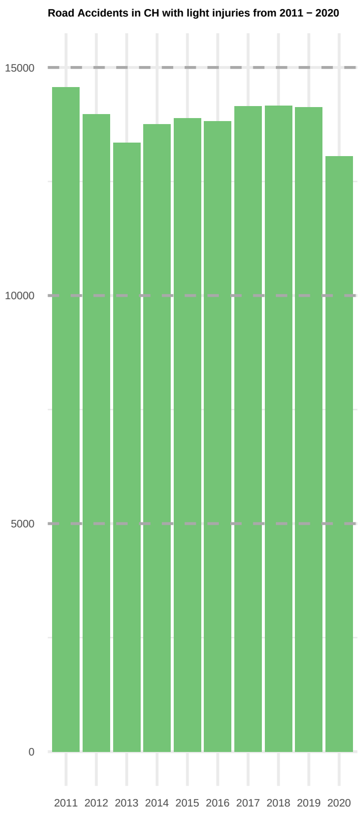
First the total accidents in Switzerland between 2011 - 2020 are getting examined. It can be seen that the total amount of road accidents is more or less stable with a slight decreasing trend. However the total number for 2020 is lower compared to previous years. This is most likely affected to the Covid-19 Pandemic with let to a lockdown in the second quarter of 2020 with led to less traffic on the road. Details on Bundesamt für Statistik.



The total road accidents are also categorised by their severity. The categorisation has three variables: Accident with fatalities, Accident with light injuries and Accidents with severe injuries. Below the three different severity types are plotted with their corresponding occurence per year. It can be seen that the overall number of accidents is decreasing as well as all the corresponding severity types are decreasing as well. However the accidents with fatalities have been increasing significantly in 2020 compared to 2019.







2. Analysis of Road Accidents with Animal Involvement

After looking broadly on the data we now dig deeper to find answers to the two use-cases. First we examine the accidents where animals were involved.

Pre-Processing as Chapter of Choice

In a first step we load the data from "Bundesamt für Strassen" and enrich it with population data for the different cantons in Switzerland. For this, we wanted to learn how to do a loop combined with if-else branching and used this to import and preprocess the data in the Excel File "PopulationCH". The main part of the code deals with the fact that each sheet in the Excel file creates a new object (data frame).

Besides, we introduce a new column **CantonCode**, which we will need later for joining with the other dataset. An example sheet reduced to its most essential looks like this:

Canton	**CantonCode**	**Total**	**Age: 0-19**	**Age: 20-64**	**Age: 65 and older**	**Urban**	**Suburban**	**Countryside**
Aargau	AG	694,072	140,443	427,461	126,168	380,756	209,519	103,797

We also introduced new relative variables that measure age in three categories: "0-19", "20-64", and "65 and older" relative to the population of the canton. The same applies to the choice of living in Urban, Suburban and Countryside regions relative to the population.

The calculation for one of these relative variables is rounded by 2 digits:

$$Urban_r = \frac{Urban}{Total}$$

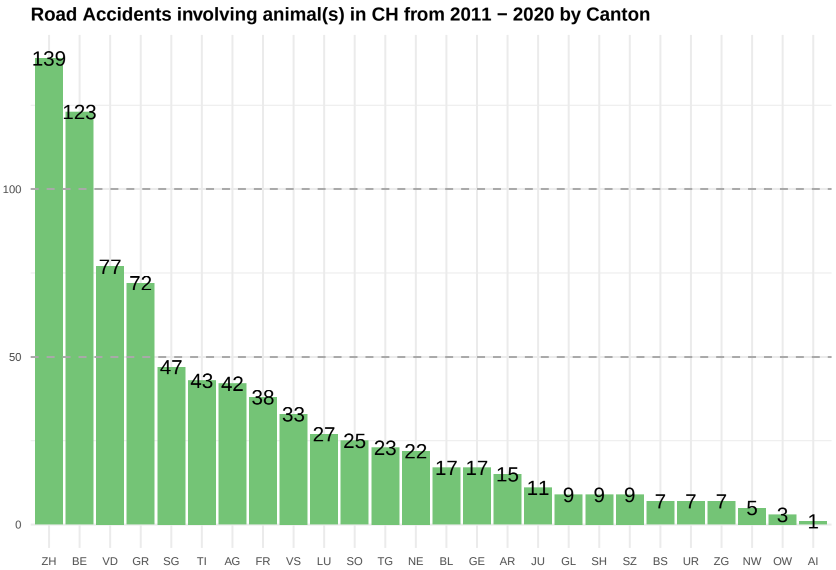
The relative proportion of the population in the canton of Aargau for the year 2020 living in Urban Areas has the value: 0.55 The extended dataset with the new variables look like this:

Canton	**Share of Urban Population (in %)**	**Share of Suburban Population (in %)**	**Share of Countryside Population (in %)**
Aargau	0.55	0.30	0.15
Appenzell A. Rh.	0.28	0.48	0.23
Appenzell I. Rh.	0.00	0.00	1.00
Basel-Landschaft	0.68	0.30	0.02
Basel-Stadt	1.00	0.00	0.00

In a second step the data is divided into subsets and joined so that they are available in an aggregated form for later analysis. After this pre-processing is done, the analysis is started.

Road Accidents with Animal-Involvement by Canton

The first plot shows an aggregation for the number of wildlife accidents grouped by canton. We see that most wildlife accidents occur in the cantons of *Zurich*, *Bern* and *Vaud*.



However, these are also the most populous cantons. It can be seen here:

Canton	Total
Zürich	1,553,423
Bern	1,043,132
Waadt	814,762
Aargau	694,072
St. Gallen	514,504
Genf	506,343
Luzern	416,347
Tessin	350,986
Wallis	348,503
Freiburg	325,496

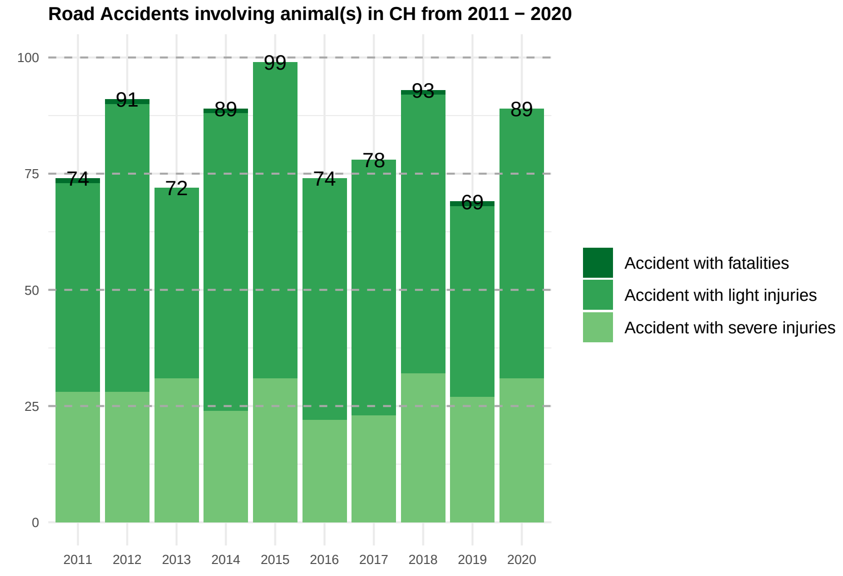
We can conclude that *Zurich*, *Bern* and *Vaud* has the highest population in total. Therefore, it is not surprising that the number of wildlife accidents is highest there.

Later on we will use the population and the calculated relative share of the population for further analysis as well. But first lets have a look at a few graphs to answer the other questions of Pro Natura.

Road Accidents with Animal-Involvement by Year

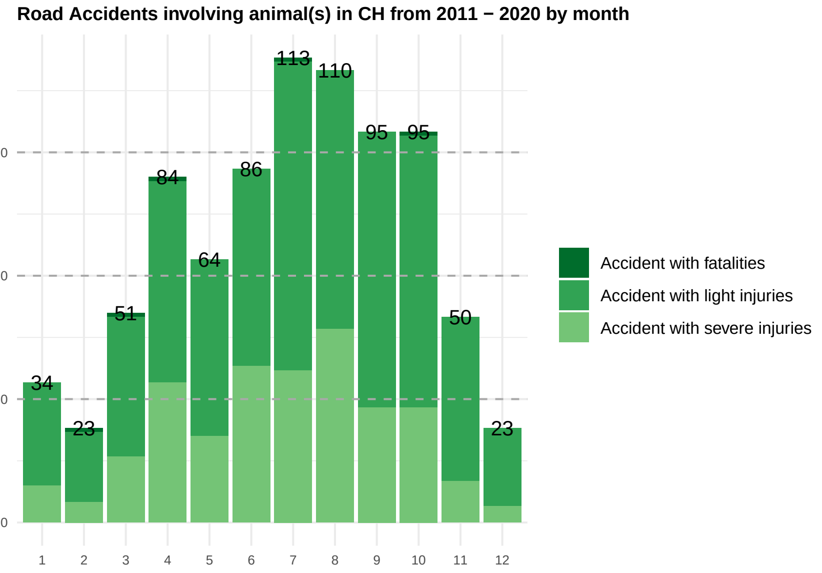
Pro Natura was interested in whether its wildlife conservation measures from previous years had been successful.

In this graph the amount of road accidents with animal involvement by year is plotted for the time period of 2011-2020. *No clear pattern* is visible, it seems that the accidents are in some year higher than in others. This means that Pro Natura's measures have not yet led to a significant reduction in wildlife accidents.

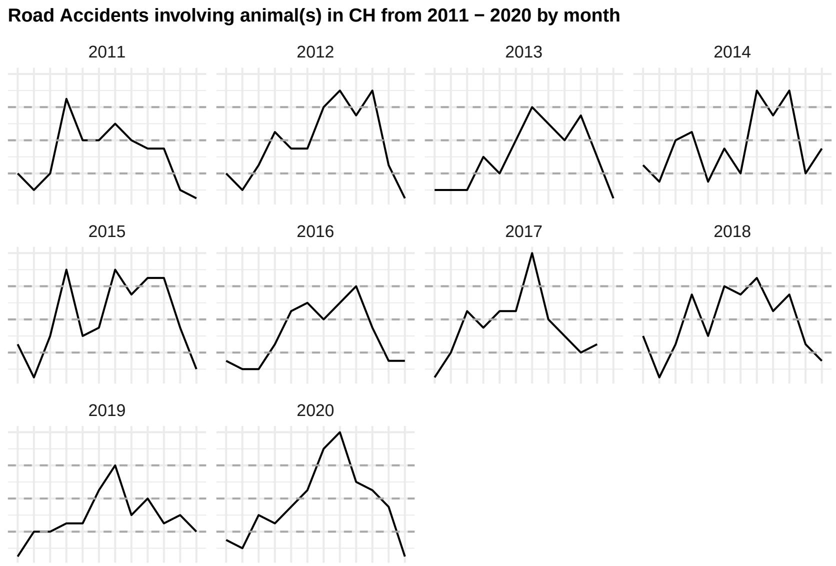


Road Accidents with Animal-Involvement by Month

Furthermore Pro Natura asked, if there is any pattern in the monthly distribution. It seems that between *July* and *October* the most accidents involving animals are happening. This also makes sense since the hunting season for some animals like deer is starting in autumn. Therefore it can be that deer as an example are fleeing from hunters and then crossing a street.

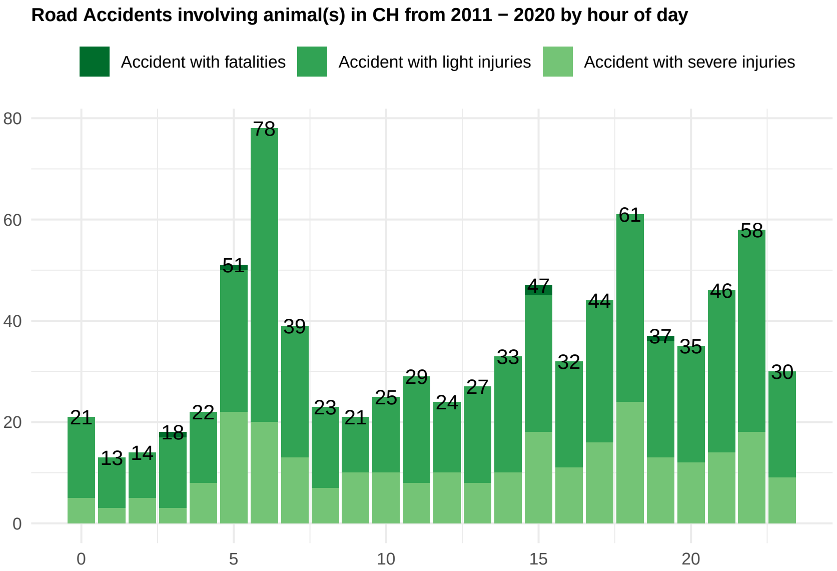


This monthly distribution will below be investigated further to see the distribution per year. The plots below are splitted by year and month. It can be seen that there is *no clear distribution / pattern*. The only observation which we can make is that it seems some years have more accidents with animal involvement in autumn as stated above. This is true for the years 2012, 2014, 2017 and partially 2020.



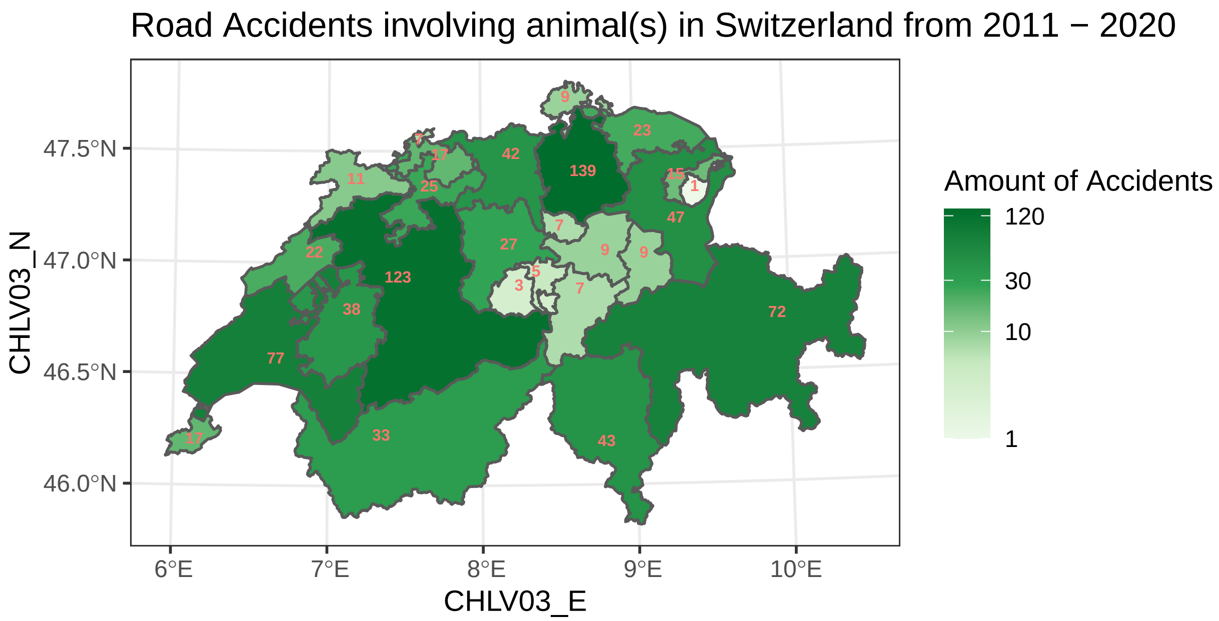
Road Accidents with Animal-Involvement by Hour of Day

Recently, Pro Natura has also asked for an analysis of the time of day at which a particularly high number of accidents occur. When investigating the occurrence of accidents with animal by hour of the day a clear pattern is visible. A peak in the morning at around *07:00* is visible and then at *18:00* in the evening as well as *22:00*. Since most animals are active at night this makes sense, since at the same time of the day most of the road traffic is taking place.


Road Accidents involving animal(s) in CH from 2011 – 2020 by hour of day

Map of Road Accidents with Animal-Involvement

As we have already seen, most wildlife accidents occur in the cantons of *Zurich*, *Bern* and *Vaud*, as they have the highest population and therefore the probability of colliding with a wild animal is higher. We have visualized this in a map of Switzerland:



However, since the absolute number of traffic accidents is not a meaningful key figure, we have introduced a relative key figure called "Accidents Per Inhabitants". With this key figure we can analyse the accidents in the map of Switzerland based on its population. For this we join the Population data with the Road Accidents data using the Primary Keys Canton Code.

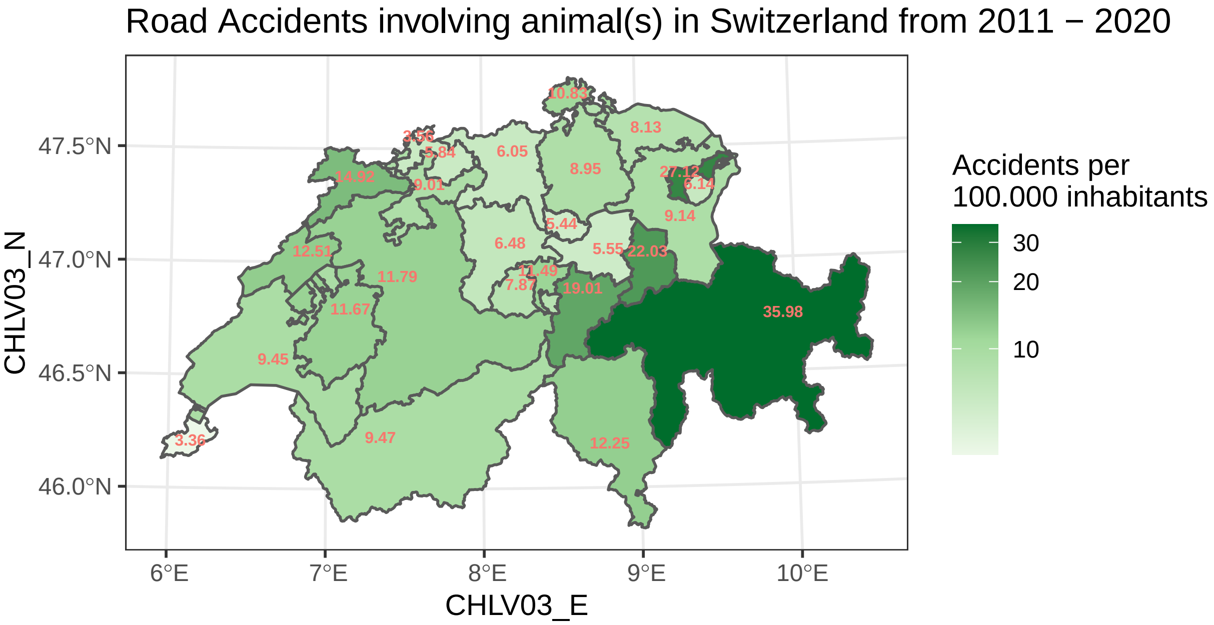
The key figure Accidents per Inhabitants is calculated as follows:

$$AccidentsPerInhabitants = \frac{Accidents}{PopulationTotal/100000} = \frac{Accidents}{PopulationTotal} * 100000$$

Now we have a good indicator to compare between the cantons.

Map of Road Accidents per 100.000 inhabitants with Animal-Involvement

The map shows that the cantons of *Graubünden*, *Appenzell Ausserrhoden* and *Glarus* have proportionally the most wildlife accidents. *Geneva* and *Basel City*, on the other hand, have the fewest cases. The top 3 cantons *Zurich*, *Bern* and *Vaud* from the previous map with the absolute cases are now in the middle range.



While researching on the Internet, we came across a very similar graph from Axa, which, however, shows a slightly different distribution. It is true that *Graubünden* is still in the top 5. But this is followed by cantons such as *Jura* and *Fribourg*. Details on: **Wild beim Wild.com**.

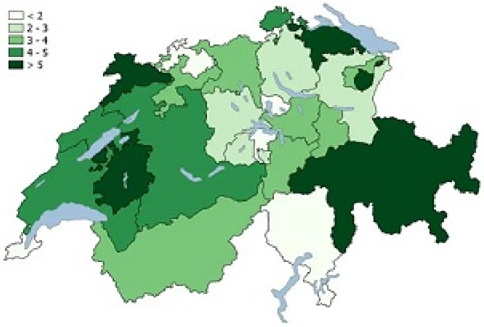


Figure 2: All wildlife accidents reported to Axa per canton in 2017.

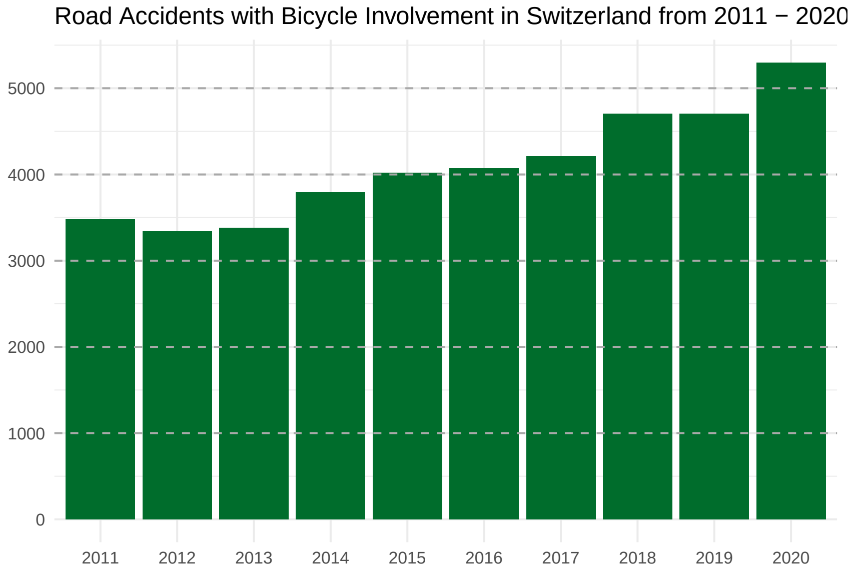
However, a closer look shows that the data basis is different, as Axa considers all wildlife accidents reported to Axa. We, on the other hand, look at the cases reported to the *"Bundesamt für Strassen ASTRA"*.

We will now turn our attention to the second use case.

3. Analysis of Road Accidents with Bicycle Involvement

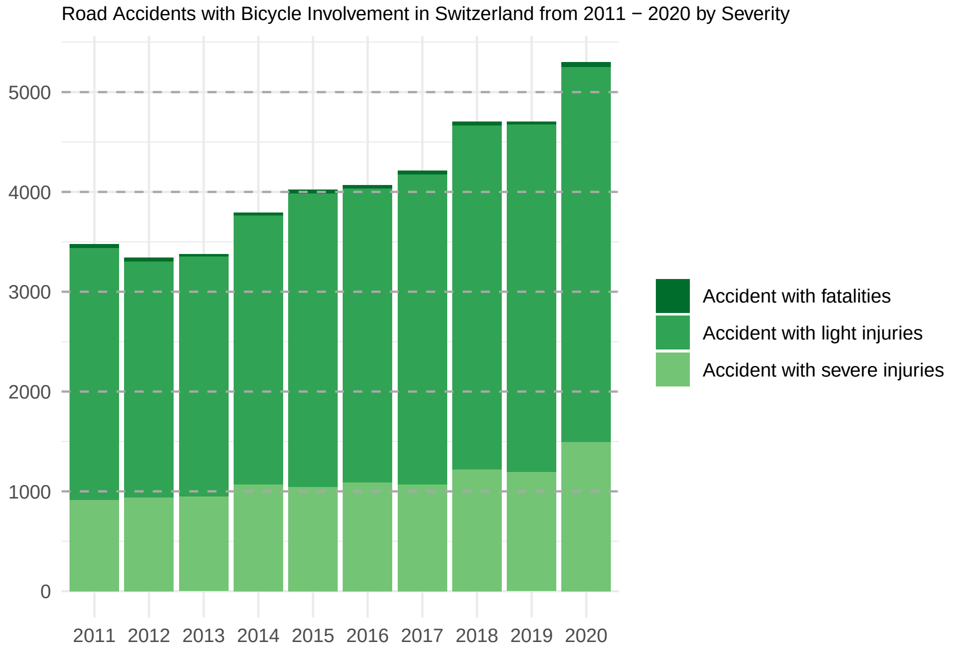
Road Accidents with Bicycle-Involvement by Year

Now the road accidents with bicycle involvement are analyzed. Therefore the dataset has been subsetted to only include the accidents with bicycle involvement. Below the plot of the development from 2011-2020 of these accidents is shown. It can be seen that the accidents with bicycles is increasing, giving reason for further and detailed analysis.



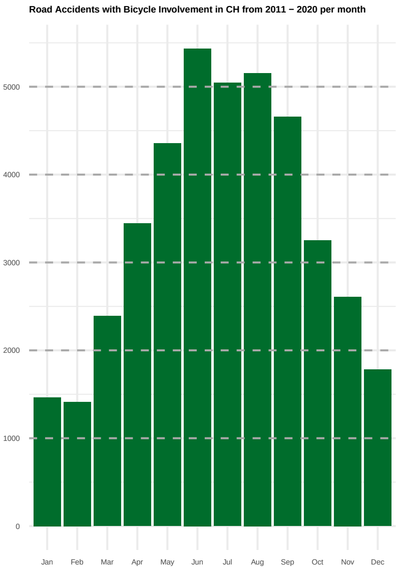
Road Accidents with Bicycle-Involvement by Severity

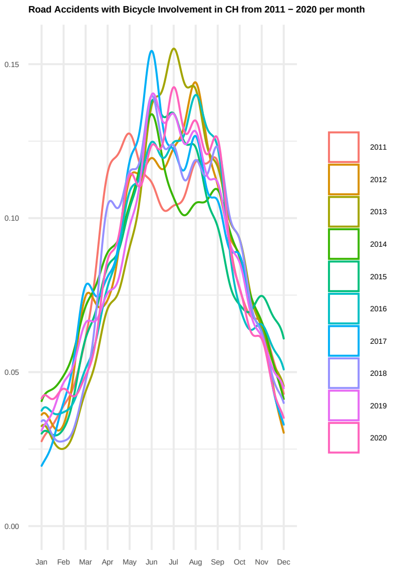
As the total bicycle accidents are increasing it is also analyzed whether the severity of these accidents is also following a pattern. It can be seen that the accidents with fatalities are not increasing drastically. But the accidents with severe injuries and the ones with light injuries have been increasing slightly year by year reaching a peak in 2020.



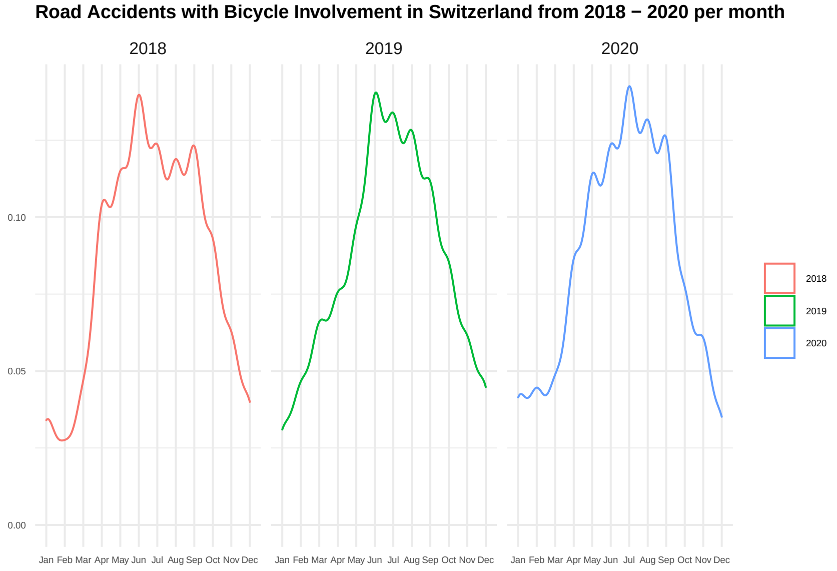
Road Accidents with Bicycle-Involvement by Month

It is also of interest whether the accidents involving bicycles is following a pattern throughout the year. Therefore the accidents are plotted with their occurrence over the months. It is clearly visible that during the summer (May - September) the most bicycle accidents have happened between 2011-2020. In this period of the year a lot of people are using the bicycle to commute or to travel from A to B, whereas during the winter a lot of people a leaving their bicycle unused due to cold temperatures. In contrast also the total accidents including all observations are following a same pattern but with less deviation throughout the year.



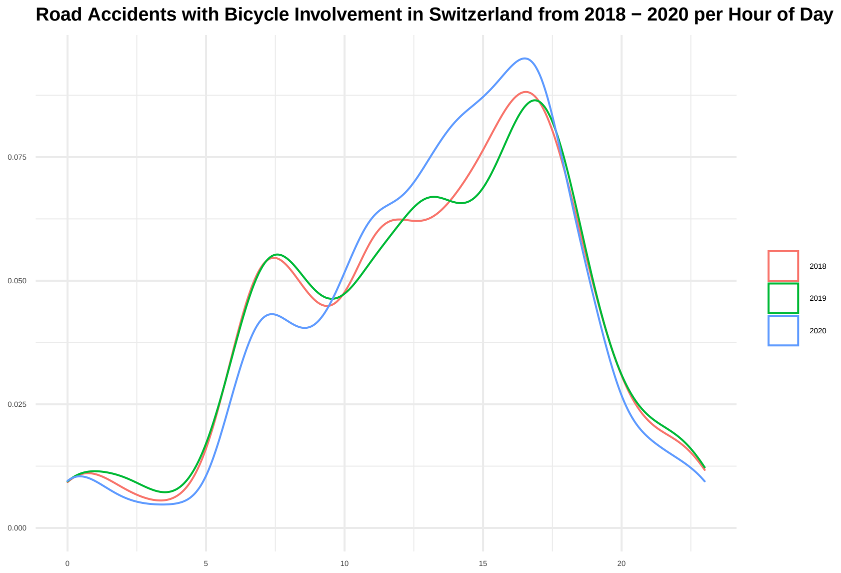


In the above charts the years 2011-2020 have been analyzed. To check whether there are any differences between the overall observations from these 9 years and the last three years (2018-2020) we also plot the occurrences of bicycle accidents per month only from the last three years. I can be seen that the distribution is following more or less the same pattern with the most observations between May and September.



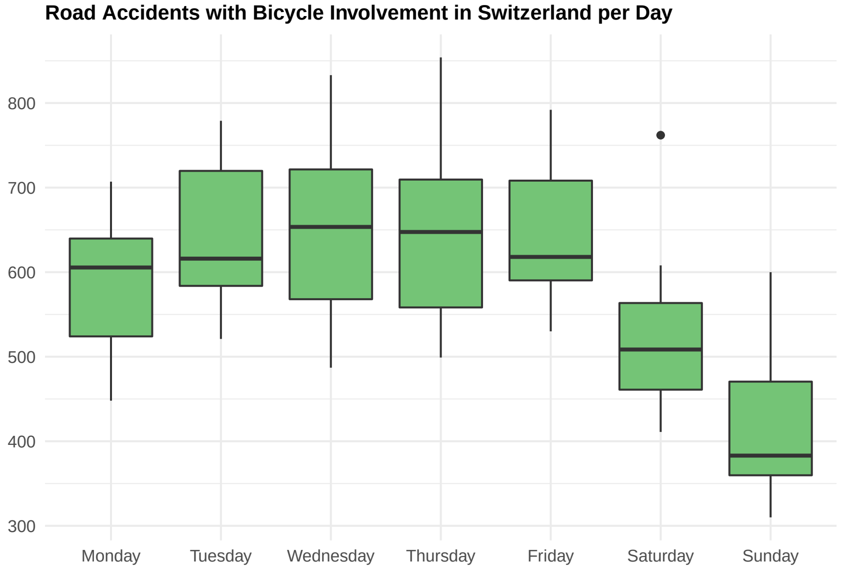
Road Accidents with Bicycle-Involvement by Hour of Day

Further investigations are made to check for patterns in the occurence of accidents throughout the day. Therefore a density-plot is made with the accidents per hour between 2018-2020. The bicycle accidents are also in this perspective following a pattern. Most of the accidents are happening during the afternoon - peak at around 17:00 to 18:00. Furthermore there also seems to be a peak in the morning at around 07:00. This indicates that a lot of people are using the bicycle to commute to work for example. In the rush-hour the bicycle traffic is then on the road with the cars and other users of the road. This would maybe indicate that there is a need for separate bicycle lanes where the different road users would not interfere with each other.



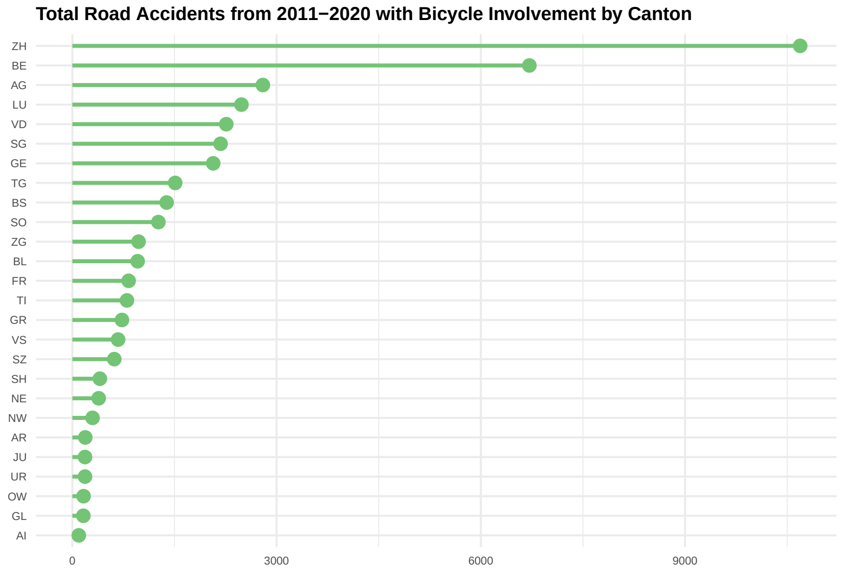
Road Accidents with Bicycle-Involvement by Weekday

Now the occurrence of accidents per month and per Hour are plotted. What is missing is the occurrence of bicycle accidents per weekday which is plotted below. It can clearly be seen that between 2011-2020 most of the accidents are happening between monday and friday, whereas in the weekend less accidents are happening. Since it can be assumed that a lot of people are using the bicycle to commute, they interfere with other road user such as cars mostly during the normal "working week". Whereas in the weekend there might be in general less traffic on the road or the usage of a bicycle / car during the day is more distributed. It can also be assumed that a lot of people are simply staying at home during the weekend and are not using their car or bicycle.

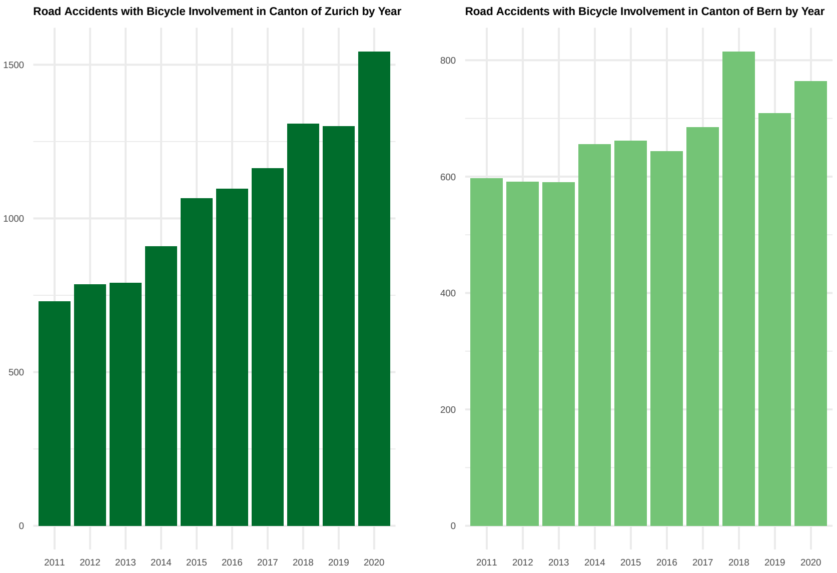


Road Accidents with Bicycle-Involvement by Canton

Below the cantons with the most bicycle accidents can be seen in a descending order. The canton of Zurich has had the most bicycle accidents from 2011-2020 followed by the canton of Bern. These two cantons have significantly more bicycle accidents than the other cantons.



Therefore Zurich and Bern are investigated further. In the two barcharts below the development of accidents in Zurich and Bern over the time-period 2011-2020 can be seen. While both cantons have seen a steady increase in bicycle accidents, the accidents in the canton of Zurich have roughly doubled between 2011-2020. In Bern the accidents have "only" increased by roughly 30% in the same period.



4. Analysis of the Demographic Development

Next, we would like to analyze the demographic development within a canton. We often read in the press that Switzerland as a whole is getting “older”. But are there also cantons that have become younger? And which canton has aged particularly strongly?

For this purpose, we use the well-known dataset about the population. The category “Age: 65 and older_r” provides information about the proportion of the population that is over 65 years old.

For this purpose, we have assigned the data of the Excel Sheets 2020 and 1999 to two different matrices. Now, to calculate the difference, we subtract the 2020 matrix from the 1999 matrix and we get the following result:

```
##    Age: 0-19 Age: 20-64 Age: 65 and older
## NW     -0.07      -0.01              0.09
## GR     -0.06      -0.01              0.07
## BL     -0.03      -0.04              0.06
## OW     -0.08       0.02              0.06
## TI     -0.02      -0.04              0.06
```

Compared with 1999, the proportion of the population aged over 65 has grown most in the cantons of Nidwalden with 9 %, Graubünden (7%), Basel Land (6%), Obwalden (6%) and Ticino (6%). Or to put it another way: these 5 cantons have grown older the most.

```
##    Age: 0-19 Age: 20-64 Age: 65 and older
## NE     -0.02       0.00              0.02
## VD     -0.01       0.01              0.01
## BS      0.00       0.01             -0.01
```

In contrast, we see Basel Stadt as the only canton where the share of the population over 65 has decreased by 1% percent.

5. Missing Values

Do we have any NA in the ID column? TRUE. In which row? Row 4.

ID	DamageTotal	AccidentType_en	CantonCode
1	30,000	Accident involving animal(s)	LE
2	20,000		BE
3	500	Accident when parking	GL
	50,000	Accident involving pedestrian(s)	AG
5	2,000	Accident with rear-end collision	GR

Now we create a missing value on our own, because the Canton Code LE is not a valid Canton Code.

ID	DamageTotal	AccidentType_en	CantonCode
1	30,000	Accident involving animal(s)	
2	20,000		BE
3	500	Accident when parking	GL
	50,000	Accident involving pedestrian(s)	AG
5	2,000	Accident with rear-end collision	GR

Axa wants us to calculate the mean of the damage total:

- With NA we get an error: NA
- Without NA: 9605.8823529

Axa wants at last a clean data set. We call this cleaned dataset "AxaDropNA". Lets check if we have still missing values there:

- Check: 0, 0, 0, 0
- All columns are 0. That means: No missing values left.

6. Fit models

Model 1: Road Accidents with Animal-Involvement depending on Hour of Day

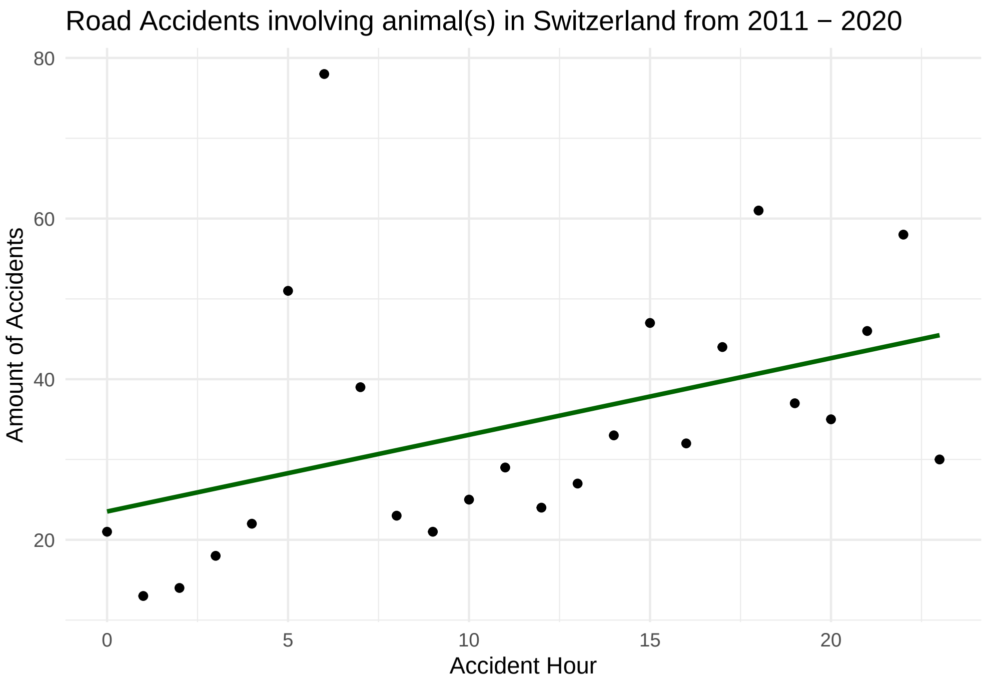
After analyzing the data, the question now is whether there is a correlation between individual variables. In chapter 2 we have already seen that there is no clear pattern for the years.

We could see there, however, that there is a pattern for the time of day. In this especially at 7 o’clock occur and at 18 o’clock. Times when there is a lot of traffic on the roads as people commute to work. But is this time of day really significant for the number of wildlife accidents? Let’s check it with a linear model.

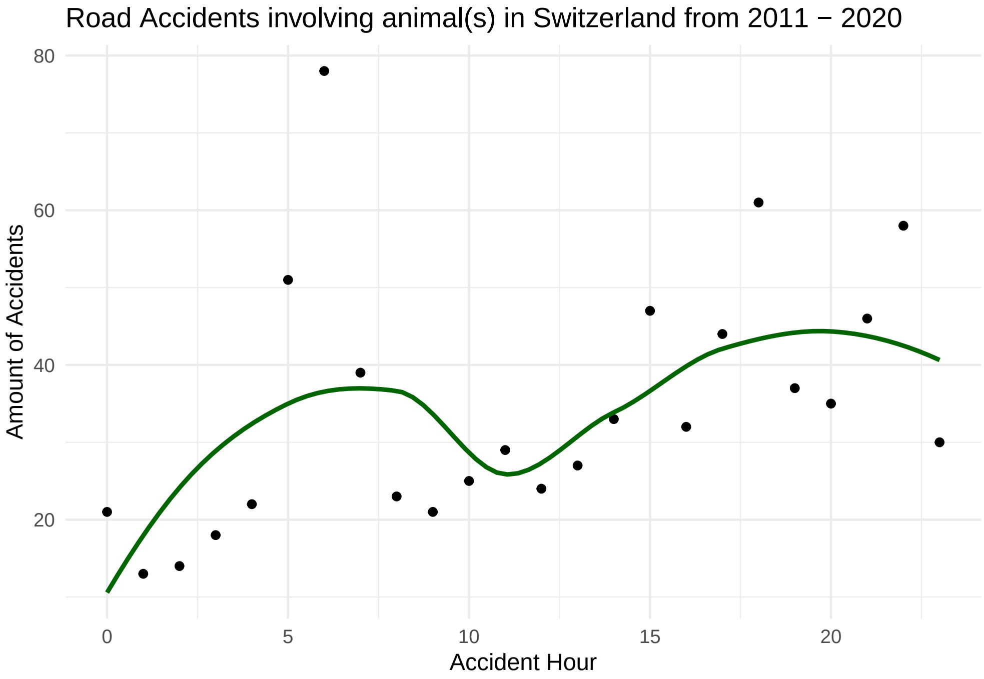
```
##
## Call:
## lm(formula = animals_accidents_per_hour$n ~ animals_accidents_per_hour$AccidentHour,
##     data = animals_accidents_per_hour)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.480  -8.521  -5.181   5.386  48.751
##
## Coefficients:
##                                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)                              23.5200     5.9069   3.982 0.000631
## animals_accidents_per_hour$AccidentHour   0.9548     0.4401   2.170 0.041113
##
## (Intercept)                             ***
## animals_accidents_per_hour$AccidentHour *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 14.92 on 22 degrees of freedom
## Multiple R-squared:  0.1763, Adjusted R-squared:  0.1388
## F-statistic: 4.707 on 1 and 22 DF,  p-value: 0.04111
```

We can see that the p-value is with a value of 0.04111 less than the alpha of 5 %. That means the independent variable “Hour” has a significant effect on the dependent variable “Amount of Accidents”. The Adjusted R Squared of 13.88% is very low. This means that only 13.88 % of the variance of the data is described by the regression line. The Intercept states that at 0 a.m./or 24 p.m., the number of wildlife accidents averages 23. With the increase of each hour of the day, the number of wildlife accidents increases by about 1 accident on average. This is what the slope value of 0.9548 tells us. We can also see the value of 23.52 at 0 AM expressed differently as a global minimum.

However the p value is very close to the significance level of 0.05. Therefore, although there is a linear relationship, this relationship is very weak.



GG Plot suggests a different line of best fit. It also shows us a trend that statistically wildlife accidents are more common at the end of the day than at the beginning. However, this trend is not strongly formed by a linear straight line, but the regression curve is more wavy like.



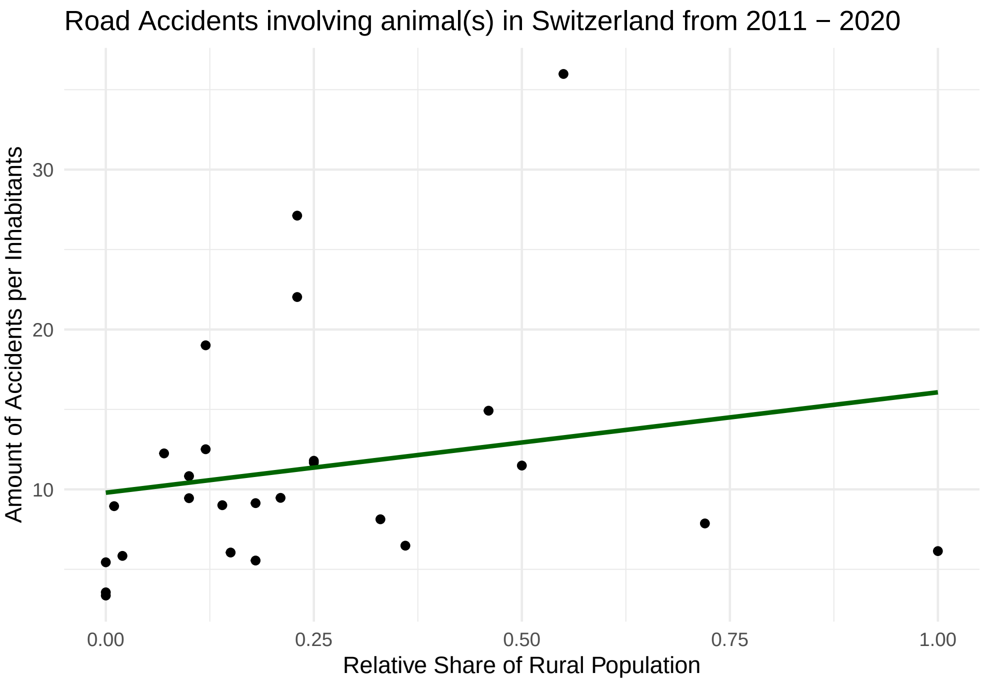
Model 2: Road Accidents with Animal-Involvement depending on Share of Countryside

Looking at the relative key figure **Accidents Per Inhabitants** of the cantons in the Swiss map, we found that *Graubünden*, *Appenzell Ausserrhoden* and *Glarus* have proportionally the most wildlife accidents.

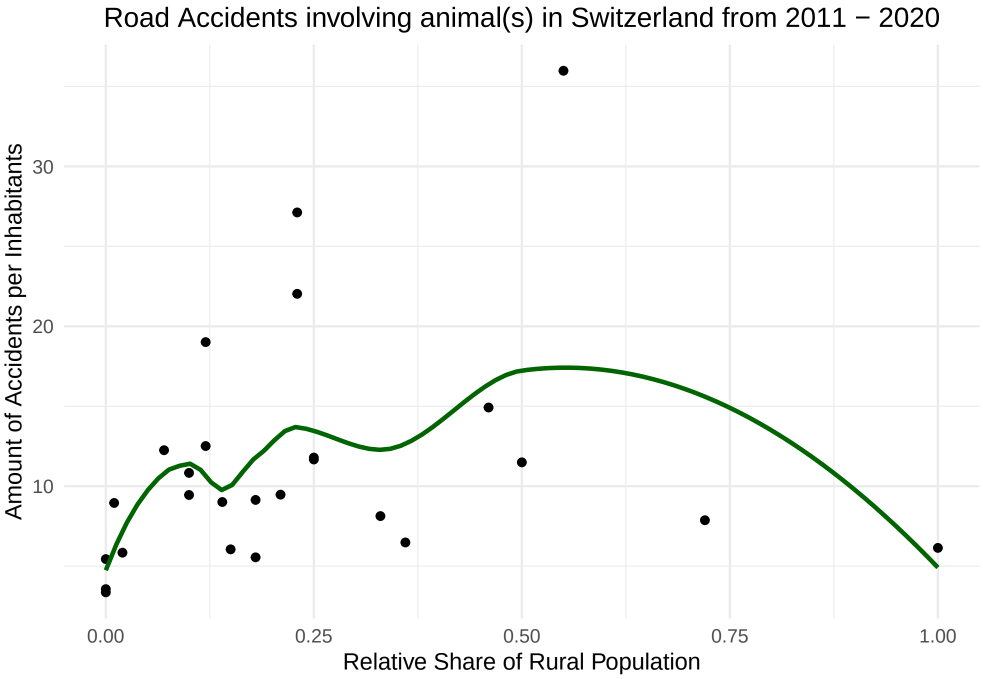
Since all three cantons are considered more rural, there could be a correlation with the urbanization of a canton and the number of wildlife accidents. Accordingly, our assertion would be that cantons whose population is predominantly rural have more frequent wildlife accidents. Is this assertion correct? We test it again with a linear model.

```
## 
## Call:
## lm(formula = AccidentsPerInhabitants ~ Countryside_r, data = animals_accidents_per_canton_pop)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -9.932 -4.601 -1.542  1.580 22.734 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      9.793      2.096   4.672 9.55e-05 ***
## Countryside_r    6.279      6.215   1.010    0.322    
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 7.458 on 24 degrees of freedom
## Multiple R-squared:  0.0408, Adjusted R-squared:  0.0008345 
## F-statistic: 1.021 on 1 and 24 DF,  p-value: 0.3224
```

We can see that the p-value is with a value of 0,3224 much larger than the alpha of 5 %. That means the independent variable “Countryside_r” does not have a significant effect on the dependend variable “AccidentsPerInhabitants”. Another look at Adjusted R-Squared also presents that the linear model is not a good fit for this context.



GG Plot recommend us the following regression curve. It seems much more suitable for this purpose. However it is even more wavier than in the previous model. We assume that this line of gg plot might be an overfit.

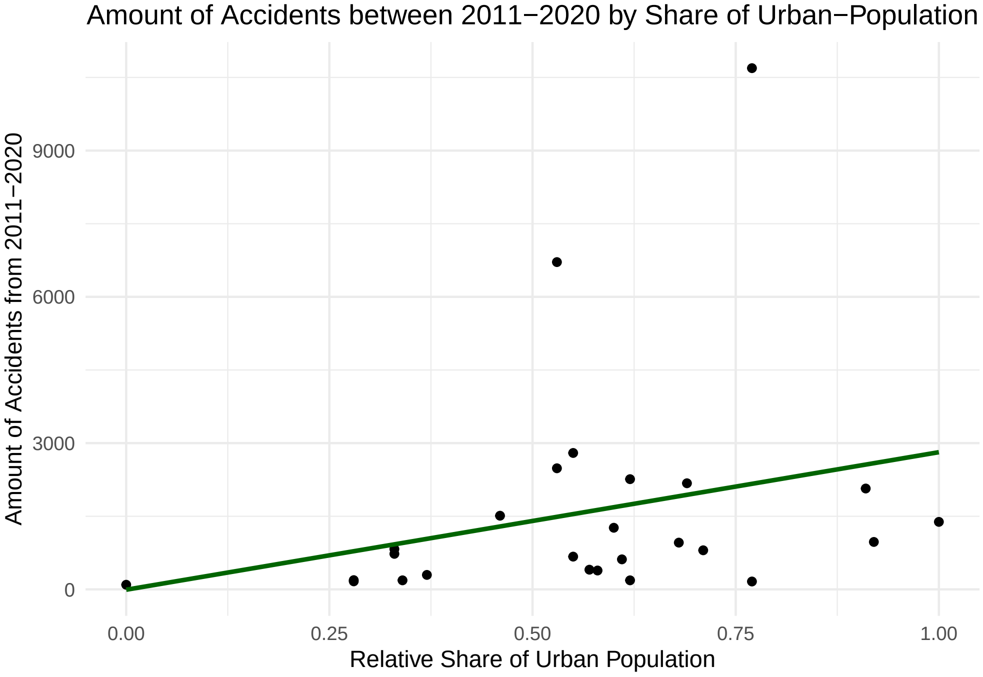


Model 3: Road Accidents with Animal-Involvement depending on City Country Distribution

A further model is made to examine whether a higher share of urban population within a canton is influencing the amount of bicycle accidents. First a table with the overall number of accidents between 2011-2020 per Canton is made. This gives 26 observations. Then the relative share of urban population is inserted to this table. The assumption is that the higher the share of urban population the more accidents there are. As a depending variable the total amount of bicycle accidents is chosen and as an independent variable the relative share of urban population.

```
##
## Call:
## lm(formula = accidents_bicycle_model1$n ~ accidents_bicycle_model1$Urban_r)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2003.1 -1169.8  -605.3   190.8  8525.9
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        -7.055   1214.483  -0.006    0.995
## accidents_bicycle_model1$Urban_r 2820.989   2012.285   1.402    0.174
##
## Residual standard error: 2270 on 24 degrees of freedom
## Multiple R-squared:  0.07569,    Adjusted R-squared:  0.03718
## F-statistic: 1.965 on 1 and 24 DF,  p-value: 0.1738
```

The summary of the linear model shows for the independent variable “Share of Urban Population” with a value of 0.1738 no significant value (significance level of 0.05). Also the Multiple R-squared of the model is with 0.07569 or 7.6% very low and has therefore no exlanatory power.



7. Summary

The results of the analysis for Pro Natura shows: The canton of *Graubünden*, *Appenzell Ausserrhoden* and *Glarus* have the highest relative share of road accidents involving animals. Most of these accidents happen either early in the morning or in the evening. A fully clear pattern for the occurence of accidents during the year can not be stated. However there is an indication that the accidents are occurring mostly with beginning of autumn. We would advice the nature conservation organization **Pro Natura** to investigate where these accidents happen. By doing so an evaluation can be done whether there is a need to support animal by making sure they are able to cross roads safely. This could be for example a wildlife bridge. Furthermore it may be useful to restrict hunting in certain areas where animals are crossing the road when fleeing from hunters.

The second request from ASTRA to reach out the cantons shows: Bicycle accidents mostly happen from Monday to Friday and are following an almost normal distribution during the year, with most of the accidents happening during summertime. It can also be stated that the most road accidents involving bicycles are happening during the rush-hour period. This means in the morning at *around 07:00* and in the evening at around 17:00 to 18:00. When analyzing the cantons and the amount of accidents it can be seen that *Zurich* and *Bern* are the cantons with the highest occurrences from 2011-2020. The accidents involving bicycles have significantly increased in the Canton of Zurich. We strongly advice the canton of Zurich to map the accidents to the exact locations where they occurred, to analyze whether there are locations where accidents occur a lot.

Figure 3: Deer on the way to the wildlife bridge.