Accident Number Forecasting

Introduction (Challenge Mission 1)

Forecasting the accidents counts is an important topic to prepare aids, ambulances for victims. Also, it helps to expect the hospitals preparations. Also, it helps to avoid it. This dataset has been prepared, cleaned, analyzed, and visualized as all in the All_Accidents_Data_Analysis.ipynb. and in From this Analysis, I found the following properties:

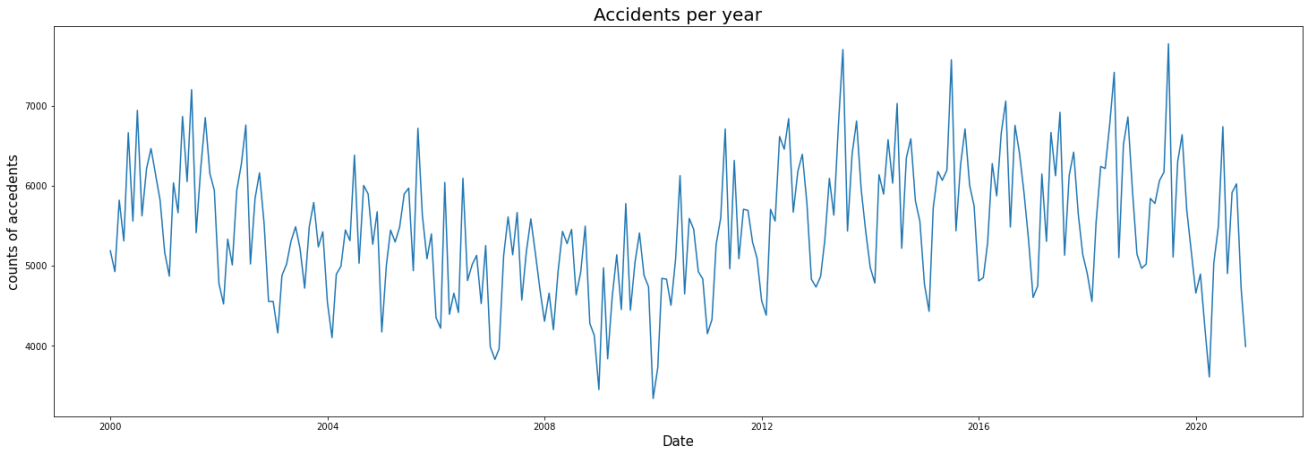
- 1764 rows of data.
- This data is from 2000 to 2020
- It has three types of accidents as follows 'insgesamt', 'Verletzte und Getötete', and 'mit P ersonenschäden'
- These accidents have the following categories 'Alkoholunfälle', 'Fluchtunfälle', and 'Verke hrsunfälle

Analyzing all data Notebook Parts (All_Accidents_Data_Analysis.ipynb)

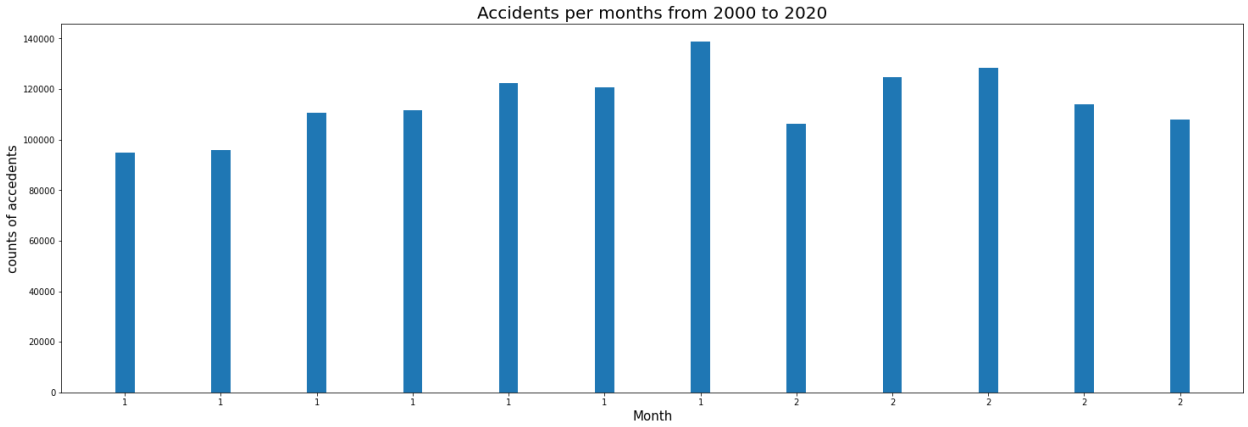
In the cleaning part, I applied the following to clean it before analysis or visualization.

In the Analysis data part, I analyzed and visualized the data to explore some insights and relationsip between data columns such as,

- Accidents per year from different types and categories.

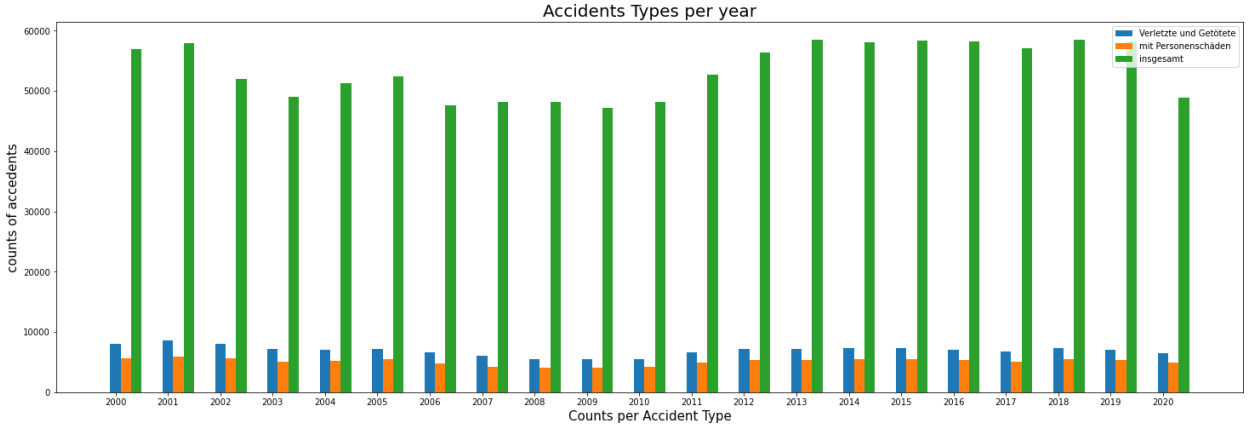


- Summation Accidents per month



We notice from above figure that July month has the maximum number of accidents.

- Accident's types per year



From here, we found that 'insgesamt' caused the highest count of accidents.

- Obtaining the relationship between accidents' types and categories. We notice from her that the 'Verkehrsunfälle ' has the highest number of accidents.

category	accident_type	
Alkoholunfälle	Verletzte und Getötete	5216
	insgesamt	11026
Fluchtunfälle	Verletzte und Getötete	11312
	insgesamt	221616
Verkehrsunfälle	Verletzte und Getötete	128906
	insgesamt	891374
	mit Personenschäden	106986

<u>In Saving data part</u>

After all cleaning steps, the data is save to be used in the Insgesamt_Accidents_Analysis_and_Modeling notebook to build the model.

Insgesamt Accidents Analysis and Modeling Notebook

Insgesamt_Accidents_Analysis_and_Modeling)

In this notebook, I will model the data of only 'insgesamt' from 'Alkoholunfälle' category. As far as I know in time-series, (It is my 1st time to work time series) the data should be sequential to build sequences. But the data in our hand has more than type of accidents happened in the same month. So, you can see more than row has the same date. As result of that I propose to build model for each category inside each type to model all data.

So, I decided to model 'insgesamt' type from the Alkoholunfälle category the 'to predict your example inside the challenge'. But in real world every type of each category should be modeled individually. In prediction time, we should call the most suitable model according to the data category. Lets go through notebook parts.

Data Preparing

Where the cleaned data is read and extract the data of 'Alkoholunfälle' category and 'insgesamt' accident type.

Preprocessing

Modelling

As mentioned above, I modelled the 'insgesamt' type from the Alkoholunfälle category.

Although, all rows in my hand are 252 only. I could tune the LSTM to model it ☺.

The model architecture is as follows:

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
rnn (RNN)                    (None, 64)                17408
_________________________________________________________________
dense (Dense)                (None, 1)                 65
=================================================================
Total params: 17,473
Trainable params: 17,473
Non-trainable params: 0
_________________________________________________________________
```

During modeling, I used the following configuration parameters:

- Epoch=10000 with early stoping, learning rate= 0.000001, LSTM size =64, batch_size=4, validation split = 0.25 of training data, testing split = 0.1 of data, loss is mean square error, patience parameter in early stopping is 2, and dense layer size=1.
- keras is used to build, train this model.

Inference (Inside the notebook)

I tried to forecast the value of the following example (inside the challenge):

Category: 'Alkoholunfälle'
Type: 'insgesamt
Year: '2021'
Month: '01'

The predicted value from the 1st model is 12.4 = 13, the 2nd model predict it as 13.006=13 (there is no float in number of accidents), the real value is 35

Note: there are two models, 1st is Insgesamt_Accidents_Analysis_and_Modeling, 2nd is Insgesamt_Accidents_Analysis_and_Modeling_copy1. The difference between both is the learning rate.