Car accident severity

A Data Science Analysis of Causes and Correlations

1. Introduction

Traffic accidents, caused by a wide variety of factors, claim thousands of victims worldwide every year. Particularly in urban traffic in large cities, serious accidents occur again and again. In countries with a well-functioning infrastructure and administration, almost all these accidents are recorded by the authorities, so that comprehensive accident statistics are available. While these lists today contain several hundred thousand entries that are no longer manageable with normal manual methods, the modern and today accessible machine learning enables an efficient analysis of enormously large data sets. Using the example of an accident statistic of the U.S. city of Seattle such an analysis shall be carried out.

The analysis is based on accident statistics for the city of Seattle for the years 2004 to 2020. At first glance, the almost 200,000 entries make the recorded accidents seem unmanageable, so that it is not possible with conventional manual evaluation methods to uncover possible connections between different accidents or causes of accidents. Such an analysis of the extensive data set could, however, help to precisely identify these correlations and causes and in this way develop solutions to prevent or at least reduce the number of accidents. Such an analysis would enable the city of Seattle and its citizens to avoid dangerous situations or constellations of various factors. Furthermore, a targeted evaluation helps insurance companies to assess accident figures and their causes.

2. Data

2.1. Data source

The present raw data set is based on a survey of the city of Seattle itself and comprises about 200,000 entries/accidents. The data set includes criteria such as the severity of the accident, the number of passers-by involved, the road and lighting conditions. In the first step, an extensive analysis of the data set itself is to be carried out. In this way, only entries with sufficient significance should remain to be able to make valid statements about factors and constellations that are favorable to accidents. It must also be checked whether the number of remaining entries is sufficient to enable targeted analyses of the data using state-of-the-art machine learning tools. If this is the case, an extensive evaluation of the data will be carried out with the aim of formulating a problem-solving oriented model for reducing traffic accidents in Seattle.

2.2. Data cleaning

The first step is a comprehensive analysis and correction of the available data. A correlation analysis should provide information about which values have a high influence on the "Serevity_Code". In order to make valid statements and set up machine learning models, columns with very low correlation were removed. Likewise, the following values or rows were removed from the data set according to the following list:

- df.dropna(subset = ["X","Weather","Road_cond","Light_cond"], inplace=True)
- df = df[df.Weather != 'Unknown']
- df = df[df.Weather != 'Other']
- df = df[df.Road_cond != 'Other']
- df = df[df.Road_cond != 'Unknown']

- df = df[df.Light_cond != 'Unknown']
- df = df[df.Light_cond != 'Other']
- df = df[df.Light_cond != 'Dark - Unknown Lighting']

A subsequent evaluation of the data obtained using "df.isna().sum()" did not reveal any other missing values, so that data analysis and evaluation could begin from this point. Another adjustment of the data types from "float64" to "int64" followed. Excluded from this were the longitude and latitude of the accident positions. These were left at "float64". Furthermore, the column caption was adjusted and "to_datetime" conversion was carried out, as well as a further decomposition of the data into month, week, weekday and hour.

2.3. Feature selection

In addition to the correlation analysis, there was also an evaluation of the individual possible feature values, the results of which can be summarised as follows:

Severity_code	
Code	Number
1	111498
2	54711

Adress_type	
Code	Number
1 – Alley	0
2 – Block	105243
3 - Intersection	60966

Ped_count	
Code	Number
0	159619
1	6342
2	220
3	22
4	4
5	1
6	1

Ped_cycl_count	
Code	Number
0	160953
1	5214
2	42

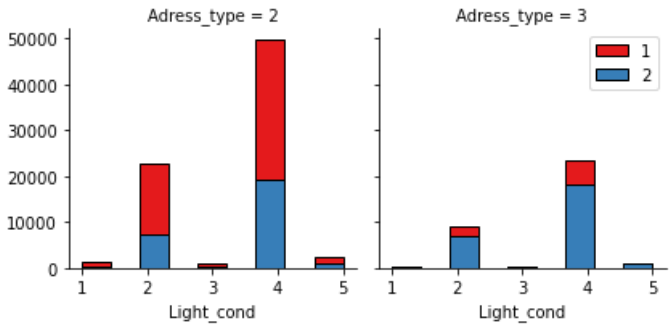
Due to the high impact, the following data evaluations and visualisations were carried out depending on the "Serevity_codes" and “Adress_type.

3. Exploratory Data Analysis

In order to be able to see initial trends and correlations in the selected feature data, "value_counts()" and plots were created for the feature values.

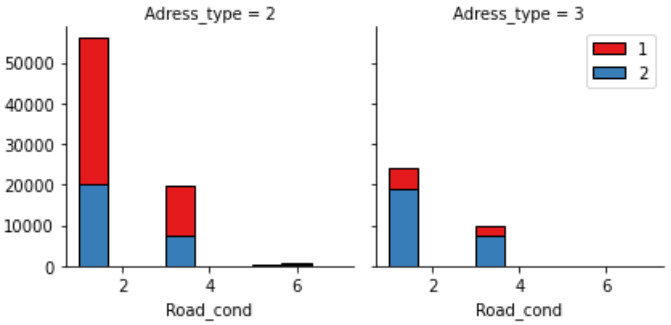
3.1. Cases considering Light Conditions (Light_cond) only

```
1 - Dark - No Street Lights / Street Lights Off
2 - Dark - Street Lights On
3 - Dawn
4 - Daylight
5 - Dusk
-------------------------------

4    110315
2     45621
5      5532
1      2404
3      2337
Name: Light_cond, dtype: int64
```



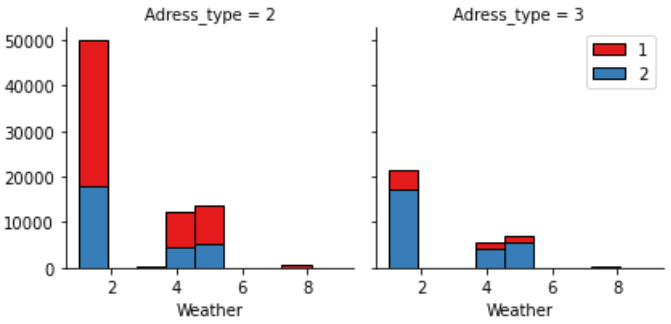
3.2. Cases considering Road Conditions (Road_cond) only

```
1 - Dry
2 - Sand/Mud/Dirt
3 - Wet
4 - Standing Water
5 - Snow/Slush
6 - Ice
7 - Oil
-------------------------------

1    119164
3     44970
6      1055
5       821
4        94
2        56
7        49
Name: Road_cond, dtype: int64
```



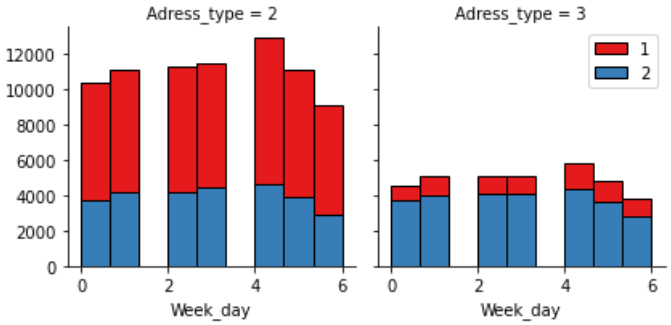
3.3. Cases considering Weather Conditions (Weather) only

```
1 - Clear
2 - Partly Cloudy
3 - Fog/Smog/Smoke
4 - Overcast
5 - Raining
6 - Severe Crosswind
7 - Sleet/Hail/Freezing Rain
8 - Snowing
9 - Blowing Sand/Dirt
-------------------------------

1    106749
5     31555
4     26378
8       813
3       537
7       106
9        42
6        24
2         5
Name: Weather, dtype: int64
```

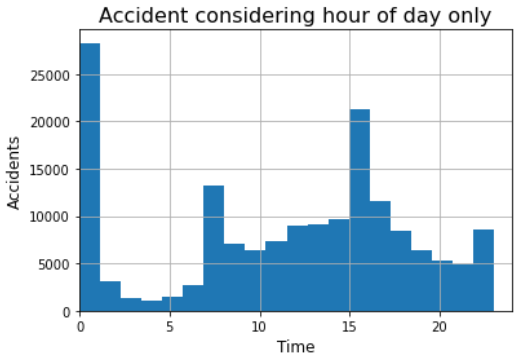


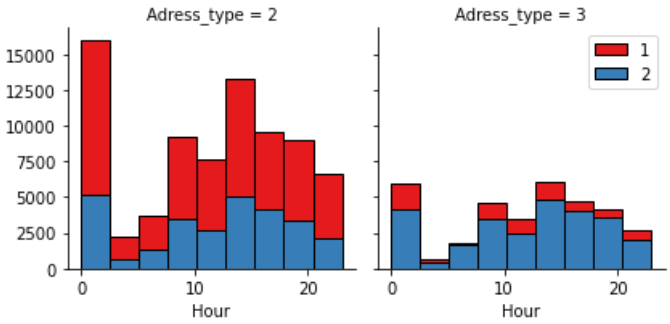
3.4. Cases considering Weather Conditions (Weather) only

```
0 - Sunday
1 - Monday
2 - Tuesday
3 - Wednesday
4 - Thursday
5 - Friday
6 - Saturday
------------------------------

4    27640
3    25127
2    24624
1    24392
5    23477
0    22341
6    18608
Name: Week_day, dtype: int64
```



3.5. Cases considering time/hour only





3.6. Summary

The following findings have so far emerged from the analyses carried out, some of them in graphic form:

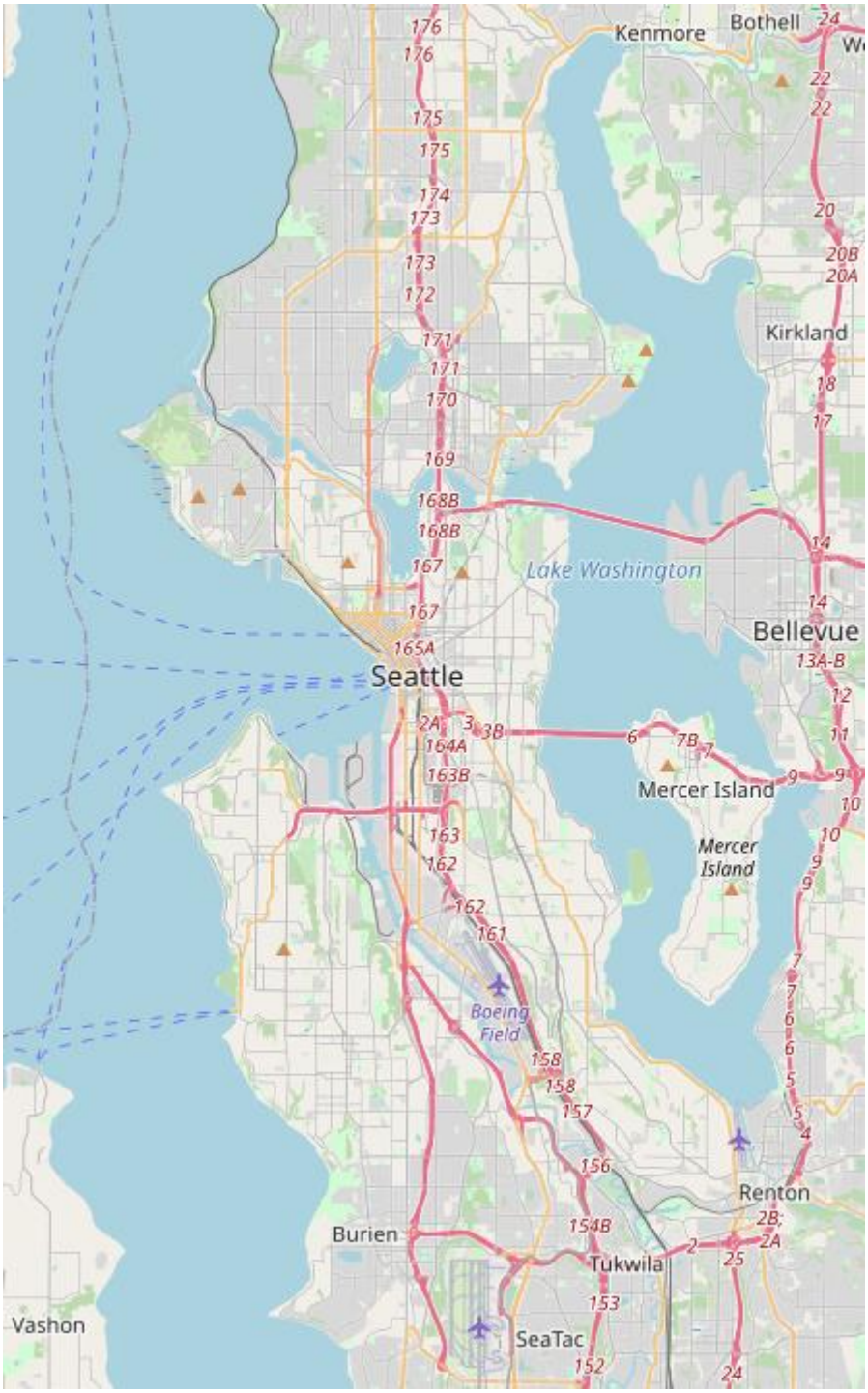
- two thirds of the accidents involved damage to property (1)
- One-third of the accidents is personal injury (2)
- No accidents on roads
- about two thirds of accidents occur in the city
- Approximately one third of accidents occur at road junctions
- Very few accidents involving passers-by (< 7,000)
- very few accidents involving cyclists (< 5,500)
- One-third of accidents occur in the dark with illuminated roads
- Two thirds of accidents happen during the day
- Three quarters of accidents occur on dry roads,
- One quarter of accidents occur on wet roads
- Three-fifths of accidents happen on a clear day
- Two-fifths of accidents occur under cloudy skies and rain
- The number of accidents increases slightly and continuously, starting with Sunday. The peak is reached on Thursday. Starting with Friday, the number of accidents falls drastically.
- In terms of individual days, most accidents occur at midnight, followed by 7 and 15 o'clock. After that, the number decreases again.

4. Visualization

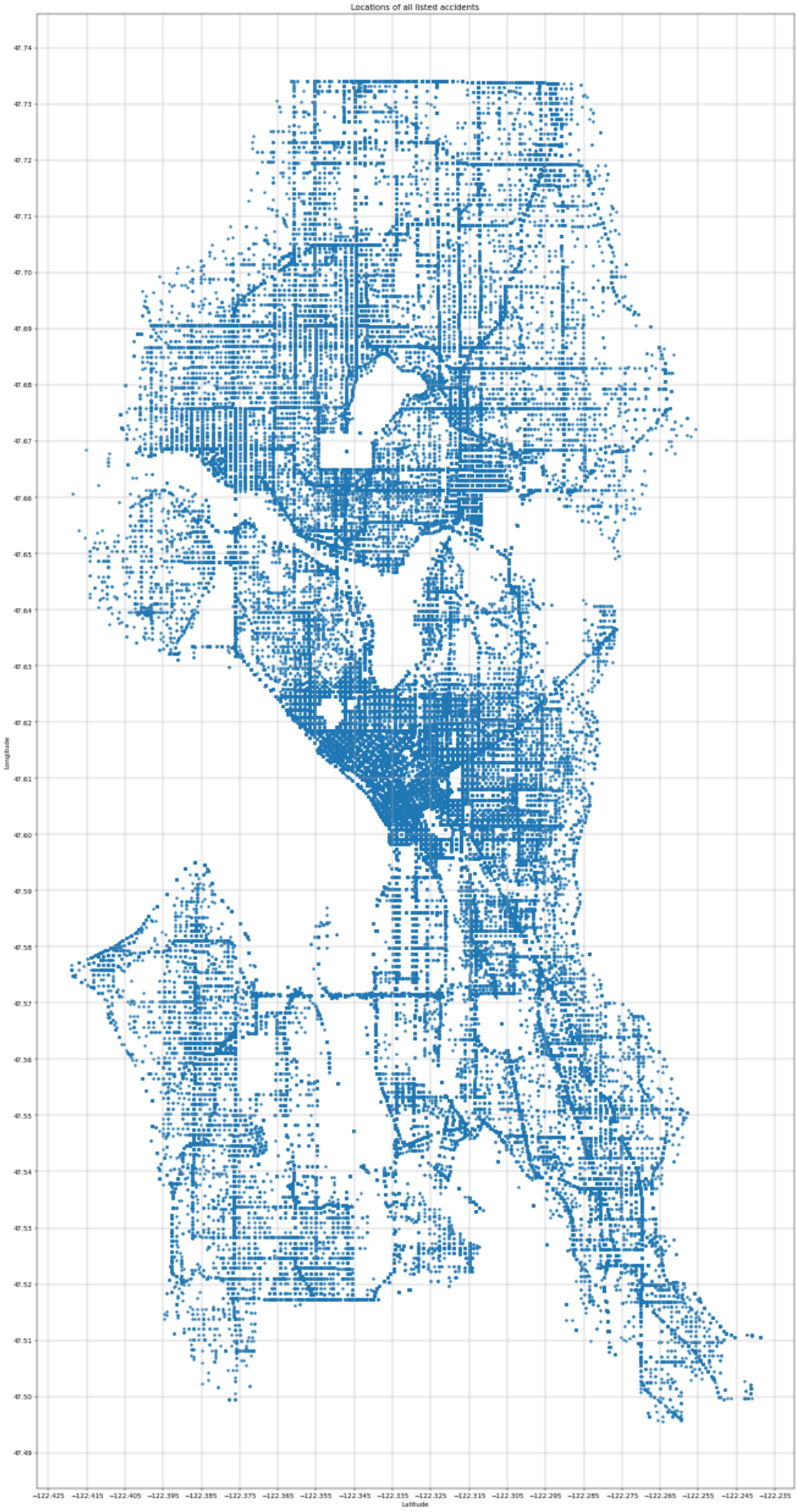
Since no geojson file was available and no adequate file for Seattle could be found during the research, a "scatter plot" was used, in which the positions of the accidents were entered for different variants according to the coordinates (X, Y). → next page.

In comparison with the actual map of Seattle, accident hotspots can be identified in this way.

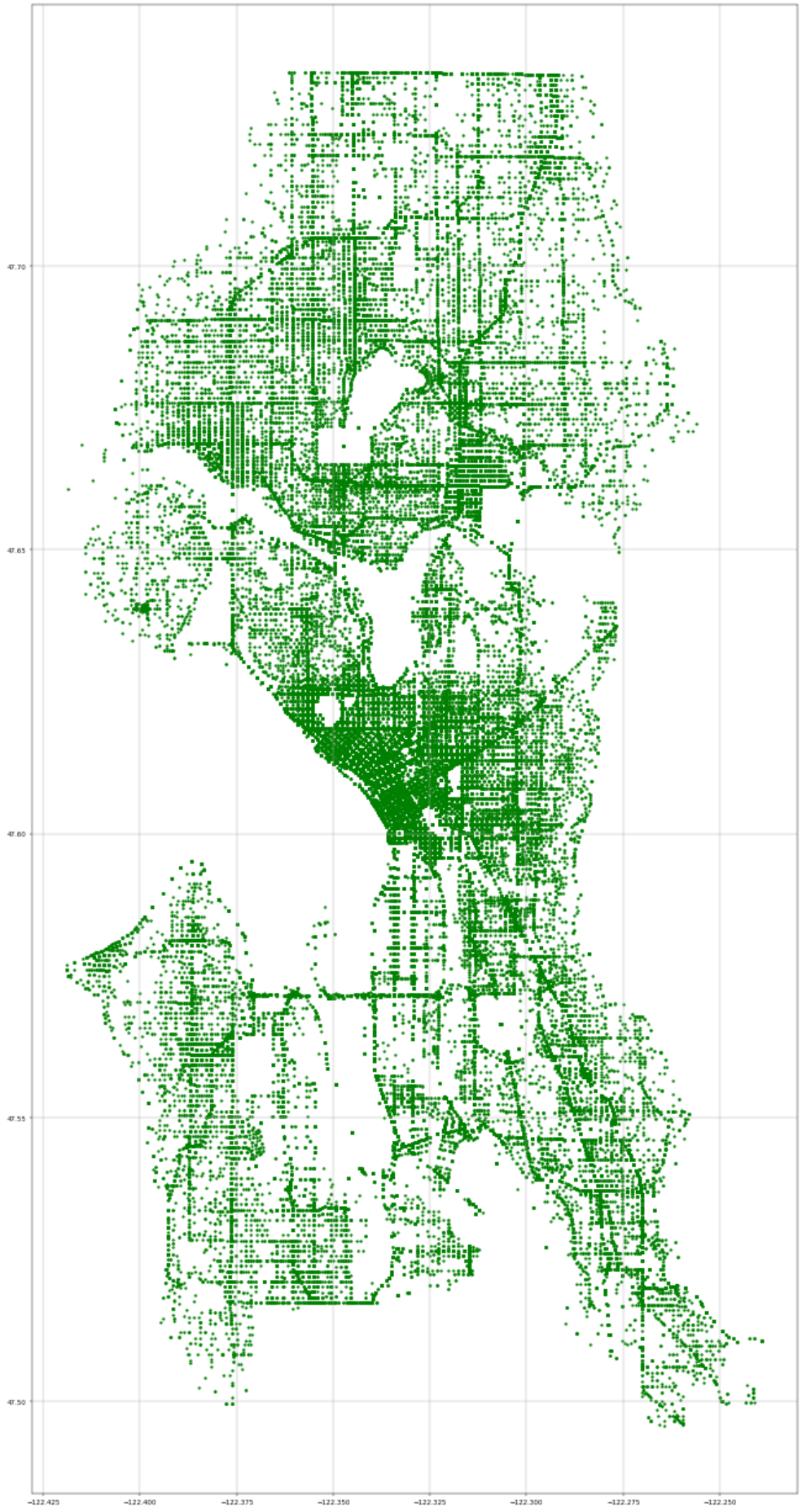
4.1. Map of Seattle



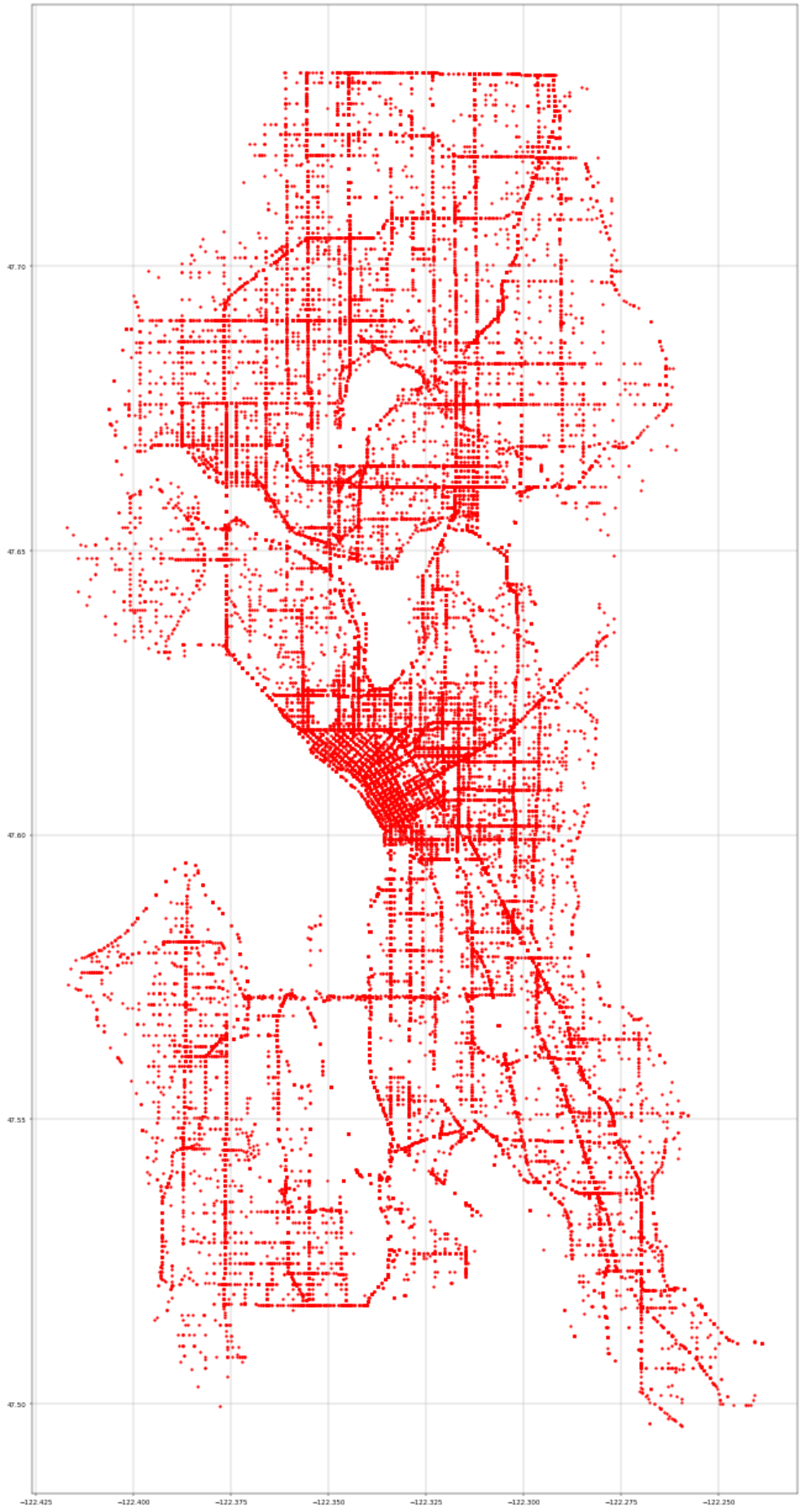
4.2. Locations of all listed accidents



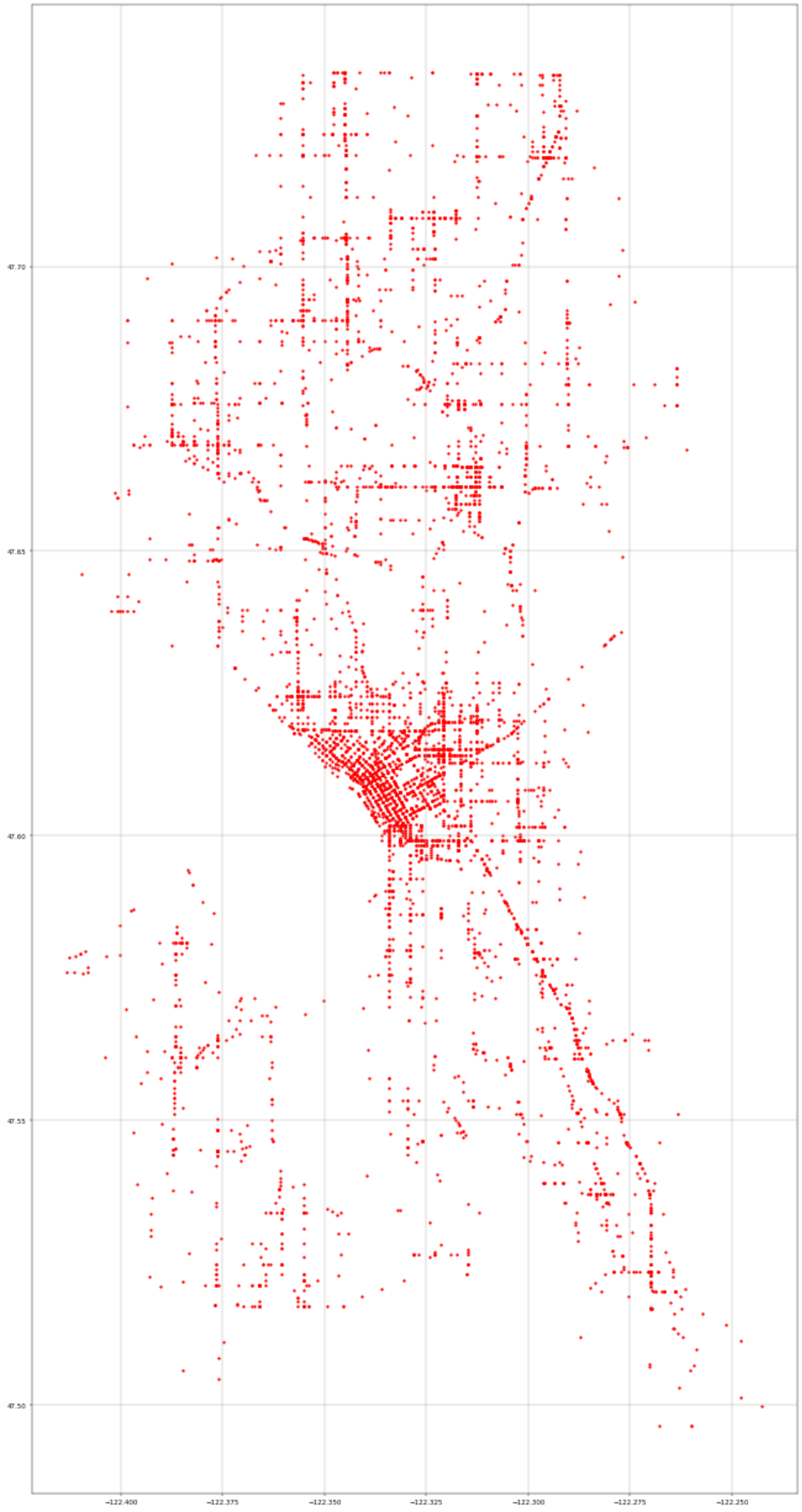
4.3. Locations of all accidents with prop damage



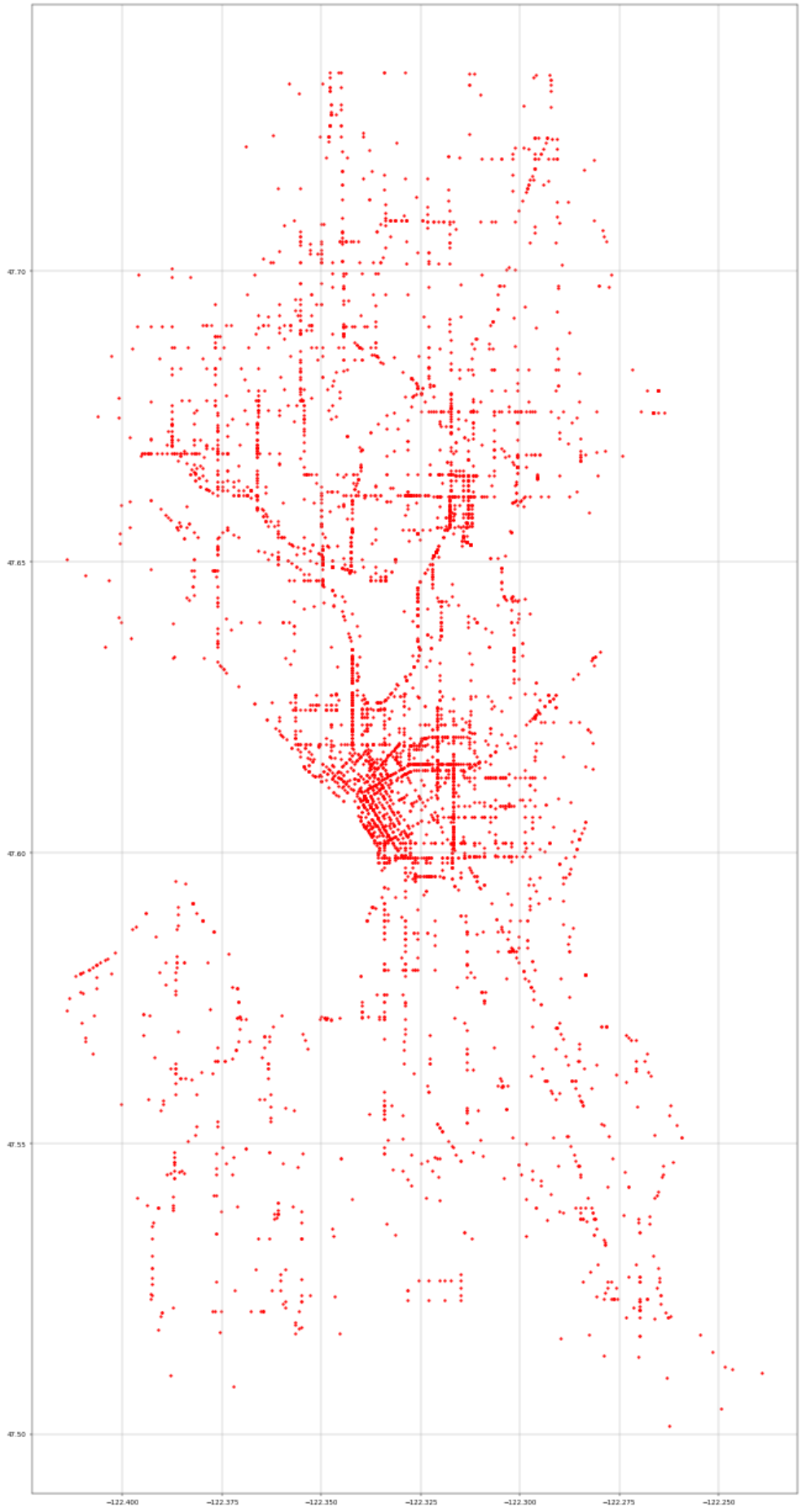
4.4. Locations of all accidents with injuries



4.5. Locations of all accidents with pedestrians



4.6. Locations of all accidents with bicycles



5. Machine Learning

To setup a predictive model, several machine learning methods were conducted. Starting with normalizing the date, K Nearest Neighbor (KNN), Decision Tree (DT), Support Vector Machine (SVM) and Logistic Regression were carried out with corresponding Jaccard index, F1-score and, if appropriate, LogLoss. The results are summarized in the following table and can be analyzed in detail in the ipynb.-file.

ML-Model	Jaccard-score	F1-score	LogLoss
KNN	0.64	0.59	---
DT	0.67	0.54	---
SVM	0.67	0.54	0.62

6. Discussion

The results obtained indicate that most accidents are due to stress and/or fatigue. Especially the cases occurring at midnight are evidence of this. Most accidents occur under normal weather and lighting conditions. In bad weather and light conditions, most people seem to avoid traffic, so there seem to be fewer accidents. Accident peaks occur at 7 clock in the morning (on the way to work) and at 15 clock (on the way home). Trends can also be seen over the course of the week: Starting with Sunday, a day with a comparatively low accident frequency, the accident frequency increases continuously from Monday to Thursday and only falls with Friday and the beginning of the weekend. A lightly local accident peak seems to be in the region within “West Edge”. With the help of Machine Learning, three models for prognosis could be established.