Data preparing

Data source

Data and metadata were downloaded from the Seattle Geo Data portal (link). Data include all types of collisions from 2004 to present.

Data construction

Data are provided by Seattle Police Department and recorded weekly by Traffic Records. The data contains 39 attributes (described in Tab. 1.) and 221 389 entries.

Tab. 1. Data attribute information.

Attribute	Description
X,Y	Geographic location.
OBJECTID	ESRI unique identifier.
INCKEY	A unique key for the incident.
COLDETKEY	Secondary key for the incident.
REPORTNO	Report number.
STATUS	N/A
ADDRTYPE	Collision address type: Alley, Block, Intersection.
INTKEY	Key that corresponds to the intersection associated with a collision.
LOCATION	Description of the general location of the collision.
EXCEPTRSNCODE	N/A
EXCEPTRSNDESC	N/A
SEVERITYCODE	A code that corresponds to the severity of the collision: 3 - fatality, 2b - serious injury, 2 - injury, 1 - property damage, 0 - unknown.
SEVERITYDESC	A detailed description of the severity of the collision.
COLLISIONTYPE	Collision type.
PERSONCOUNT	The total number of people involved in the collision.
PEDCOUNT	The number of pedestrians involved in the collision. This is entered by the state.
PEDCYLCOUNT	The number of bicycles involved in the collision. This is entered by the state.
VEHCOUNT	The number of vehicles involved in the collision. This is entered by the state.
INJURIES	The number of total injuries in the collision. This is entered by the state.
SERIOUSINJURIES	The number of serious injuries in the collision. This is entered by the state.
FATALITIES	The number of fatalities in the collision. This is entered by the state.
INCDATE	The date of the incident.
INCDTTM	The date and time of the incident.
JUNCTIONTYPE	Category of junction at which collision took place.
SDOT_COLCODE	A code given to the collision by SDOT.
SDOT_COLDESC	A description of the collision corresponding to the collision code.
INATTENTIONIND	Whether or not collision was due to inattention. (Y/N)
UNDERINFL	Whether or not a driver involved was under the influence of drugs or alcohol.
WEATHER	A description of the weather conditions during the time of the collision.
ROADCOND	The condition of the road during the collision.
LIGHTCOND	The light conditions during the collision.
PEDROWNOTGRNT	Whether or not the pedestrian right of way was not granted. (Y/N)
SDOTCOLNUM	A number given to the collision by SDOT.

SPEEDING	Whether or not speeding was a factor in the collision. (Y/N)
ST_COLCODE	A code provided by the state that describes the collision.
ST_COLDESC	A description that corresponds to the state's coding designation.
SEGLANEKEY	A key for the lane segment in which the collision occurred.
CROSSWALKKEY	A key for the crosswalk at which the collision occurred.
HITPARKEDCAR	Whether or not the collision involved hitting a parked car. (Y/N)

Data selection and initial cleaning

- The unnecessary columns were removed. For the next stages the following columns will be needed: ADDRTYPE, LOCATION, SEVERITYCODE, COLLISIONTYPE, FATALITIES, INCDATE, INATTENTIONIND, UNDERINFL, WEATHER, ROADCOND, LIGHTCOND, PEDROWNOTGRANTED, SPEEDING.
- Entries with attribute "PEDESTRIAN" were selected in order to obtain a dataset regarding only the collisions with pedestrians.
- After choosing the "PEDESTRIAN" attribute, the COLLISIONTYPE column wasn't needed anymore, so it was removed.
- New dataset has 12 columns and 7665 rows.
- There were a couple of problems with data:
    - In column INATTENTIONID there were only two types of entry: Y and NaN. I have assumed that NaN stands for no or unknown. I have replaced Y with 1 and NaN with 0. Similar situation was with columns PEDROWNOTGRNT and SPEEDING.
    - In column UNDERINFL there were four types of entry: Y, N, 0 and 1. I have assumed that 0 stands for No and 1 stands for Y Data were standardized to 0 and 1.
    - Column INCDATE was contaminated by time string containing only zeroes: "00:00:00+00". Unnecessary string was removed.

Data application

Obtained dataset will be used to point which places in Seattle city are the most dangerous for pedestrians and which factors are the most important regarding the severity of pedestrian's injuries. Data will be used to prepare appropriate histograms and decision tree machine learning algorithm.