Data understanding:

In this phase, you need to collect or extract the dataset from various sources such as csv file or SQL database. Then, you need to determine the attributes (columns) that you will use to train your machine learning model. Also, you will assess the condition of chosen attributes by looking for trends, certain patterns, skewed information, correlations, and so on.

Dataset: Seattle Collision Data

Datashape:
```
(194673, 38)
```

Columns :

```
['SEVERITYCODE', 'X', 'Y', 'OBJECTID', 'INCKEY', 'COLDETKEY', 'REPORTNO',
       'STATUS', 'ADDRTYPE', 'INTKEY', 'LOCATION', 'EXCEPTRSNCODE',
       'EXCEPTRSNDESC', 'SEVERITYCODE.1', 'SEVERITYDESC', 'COLLISIONTYPE',
       'PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT', 'INCDATE',
       'INCDTTM', 'JUNCTIONTYPE', 'SDOT_COLCODE', 'SDOT_COLDESC',
       'INATTENTIONIND', 'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND',
       'PEDROWNOTGRNT', 'SDOTCOLNUM', 'SPEEDING', 'ST_COLCODE',
       'ST_COLDESC','SEGLANEKEY', 'CROSSWALKKEY', 'HITPARKEDCAR']
```

Groups: Locations or hazardous areas
Conditions caused by Nature
Human Failure
Count accident participants
Time of the Accident
Unnecessary columns (to be dropped)
To be determined

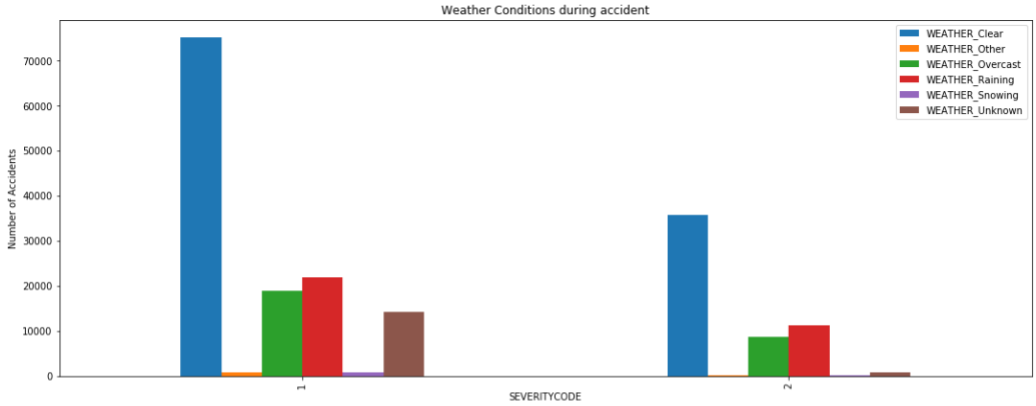
194673 accidents
→ Large dataset.

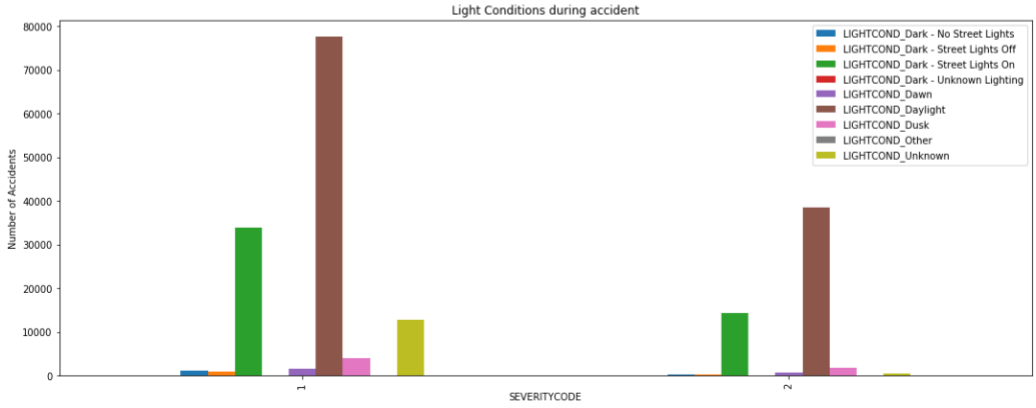
Dataset is not balanced
1 136485 (not severe)
2 58188 (severe)

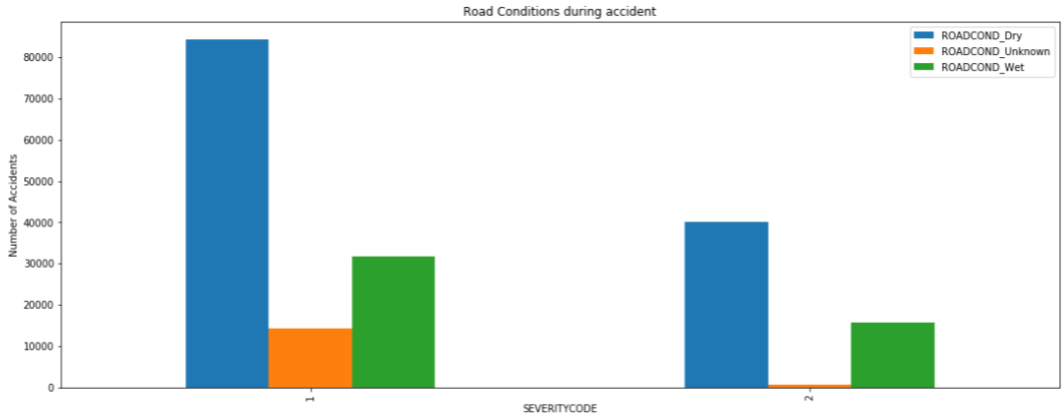
Conditions caused by Nature

'WEATHER', 'ROADCOND', 'LIGHTCOND'

Each of the data has 9 or 11 variables. So one creates dummies and plots and group them by severity.







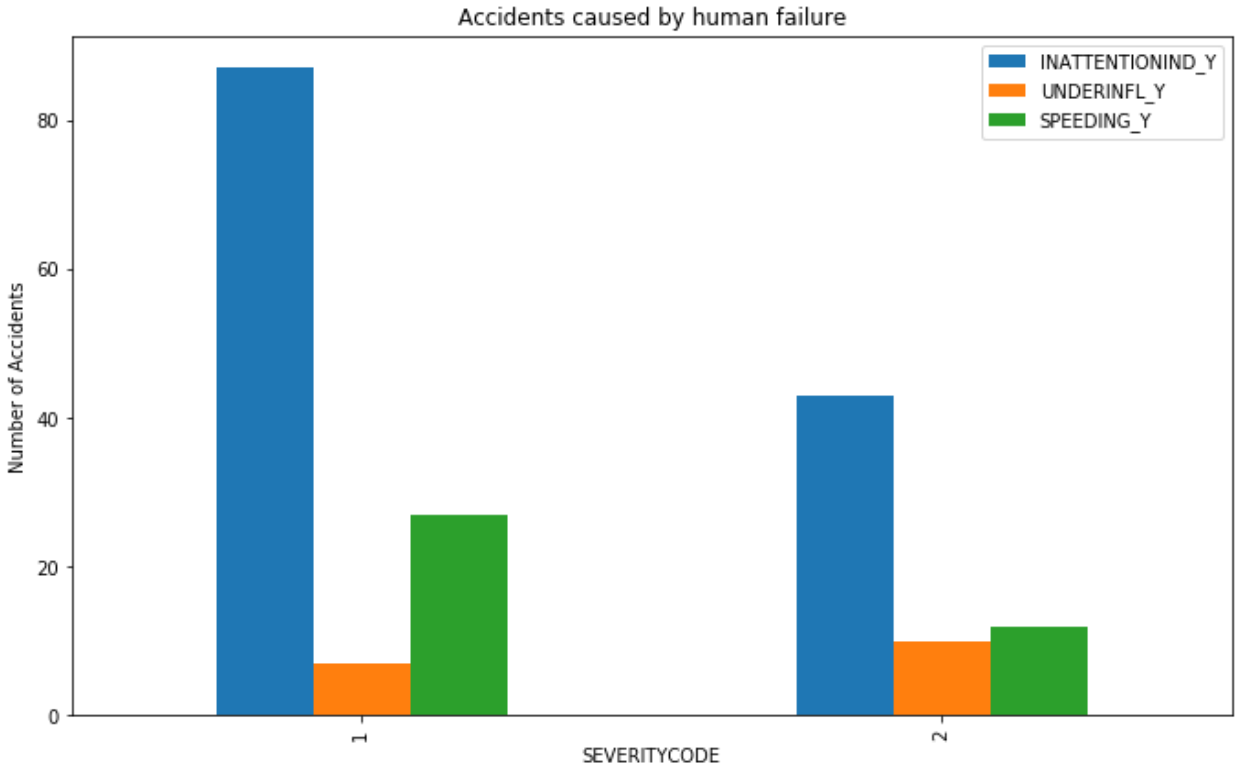
Trends

- Weather:
  Severe Accidents more likely when: ‘Rainy’, ‘Overcast’
- Light Conditions
  Severe Accidents more likely when: ‘Dark’
- Road Conditions:
  Severe Accidents more likely when: ‘Wet’

Human Failure
'INATTENTIONIND', 'UNDERINFL', ‘SPEEDING’

Data: UNDERINFL has Yes No and 0 1 data.

➔ Preprocess data to map Y -> 1 and N -> 0

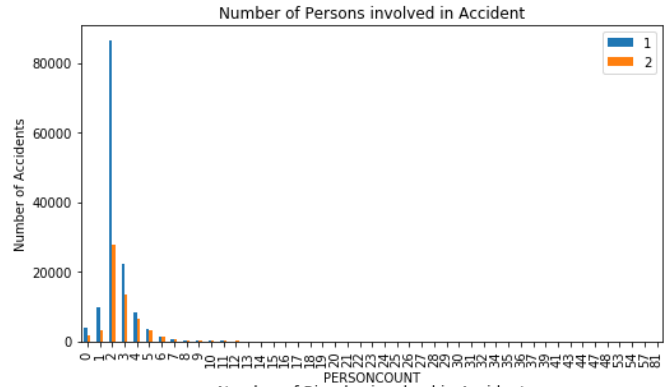


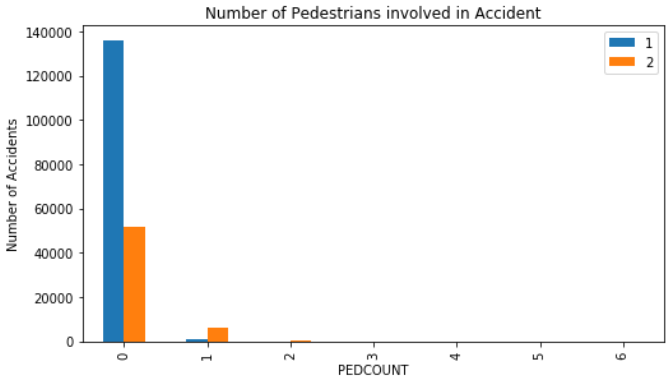
Trends:

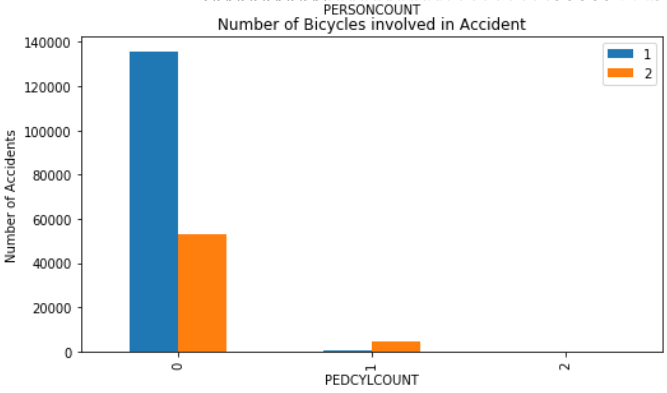
→ Severe accidents are more likely to happen when Speeding and under influence.

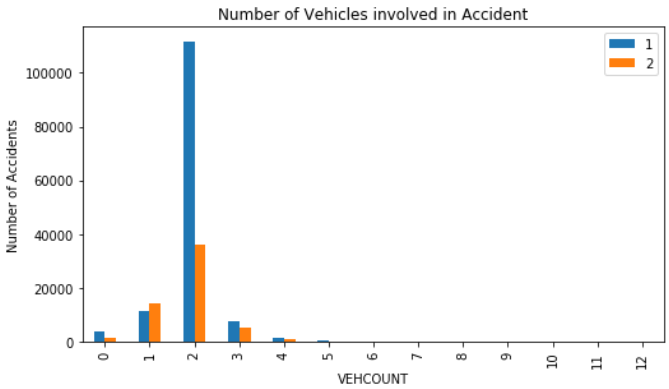
Count of Accident Particpants

'PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT'









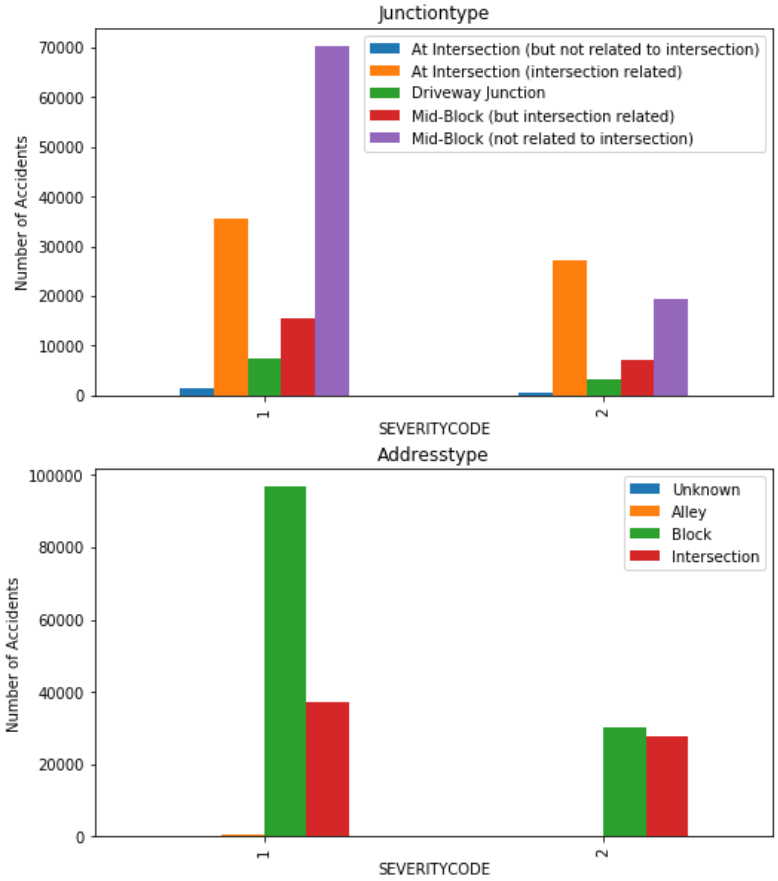
Trends:

- Persons:
  The more persons involved, the more likely it is for the accident to be severe

- Pedestrians and Bicycle riders:
  If pedestrians and bicycle riders are involved, the accident is likely to be severe.

- Vehicle:
  The accident is likely to be severe when more vehicles are involved

Address types



Trends:

Severe accidents are more likely to happen at

- Intersections

Locations
‘LOCATION’

In the dataset Locations are type string. Hence, one can filter the most common strings:

Format: (‘string’, count)

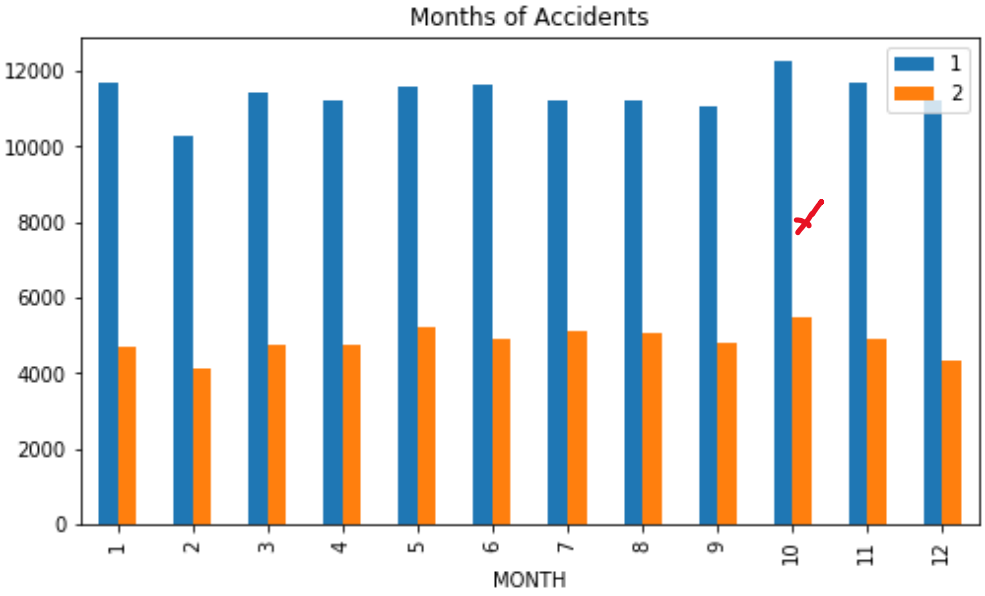
```
('ST', 28030), ('AVE', 27716), ('AND', 24104), ('BETWEEN', 16491), ('S', 1
5619), ('NE', 10026), ('N', 9381), ('SW', 8016), ('E', 6375), ('NW', 6121)
, ('W', 4429), ('WAY', 3504), ('PL', 1809), ('1ST', 858), ('DEAD', 835), (
'END', 835), ('15TH', 781), ('RP', 687), ('8TH', 670), ('DR', 665), ('5TH'
, 603), ('4TH', 581), ('3RD', 580), ('6TH', 574), ('KING', 547), ('35TH',
535), ('2ND', 532), ('LAKE', 522), ('12TH', 510), ('45TH', 494), ('17TH',
494), ('14TH', 491), ('24TH', 481), ('42ND', 479), ('20TH', 479), ('30TH',
464), ('39TH', 458), ('23RD', 441), ('JR', 440), ('M', 439), ('L', 439), (
'32ND', 428), ('16TH', 421), ('9TH', 407), ('25TH', 400), ('36TH', 391), (
'22ND', 389), ('11TH', 383), ('65TH', 383), ('26TH', 380), ('46TH', 371),
('34TH', 370), ('28TH', 366), ('41ST', 363), ('40TH', 363), ('RAINIER', 35
8), ('50TH', 357), ('AURORA', 356), ('38TH', 356), ('WR', 355), ('7TH', 35
2), ('31ST', 345), ('SPOKANE', 344), ('ER', 344), ('BLVD', 344), ('BEACON'
, 342), ('37TH', 342), ('47TH', 337), ('18TH', 335), ('44TH', 334), ('13TH
', 330), ('21ST', 318), ('OFF', 299), ('43RD', 297), ('ROOSEVELT', 296), (
'80TH', 295), ('19TH', 291), ('75TH', 289), ('10TH', 288), ('48TH', 283),
('29TH', 277), ('70TH', 276), ('EAST', 275), ('55TH', 274), ('85TH', 268),
('GREENWOOD', 264), ('RD', 264), ('FREMONT', 264), ('27TH', 262), ('PARK',
254), ('MERCER', 254), ('WASHINGTON', 253), ('LINDEN', 252), ('NB', 247),
('BR', 244), ('33RD', 241), ('AV', 239), ('DENNY', 239), ('ON', 232), ('62
ND', 232)
```

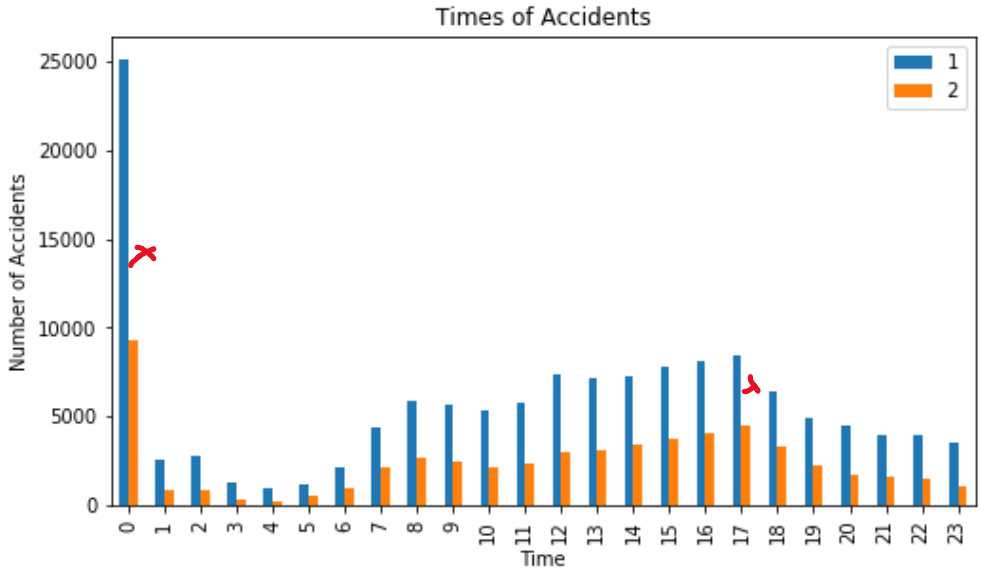
It seems like the most accidents occur in the South in ‘Dead Ends’ or near the lake.

More to be evaluated.

Time of Accident
‘INCDTTM’

Convert the INCDTTM into Panda datetime.





Trends:

- Months:
  (Severe) accidents are more likely to happen in October
- Time:
  Severe accidents are more likely to happen during rush-hour (4-5PM) and at mid night

Conclusion:

After determining groups containing obvious patterns, the next step will be to run those data groups through machine learning algorithms like Logistic Regression, etc. .