ANALYSIS OF CAR ACCIDENTS IN SEATTLE

Python Course Capstone

Introduction: Business Problem

- Safety of Seattle Roads analyzing hours, locations, weather conditions and people involve in each accident
- The main objective is to predict injury
- Use of maps to see the most stacked areas
- Easy to predict using people involve
- Use of plots to compare variables between light accidents and dangerous accidents

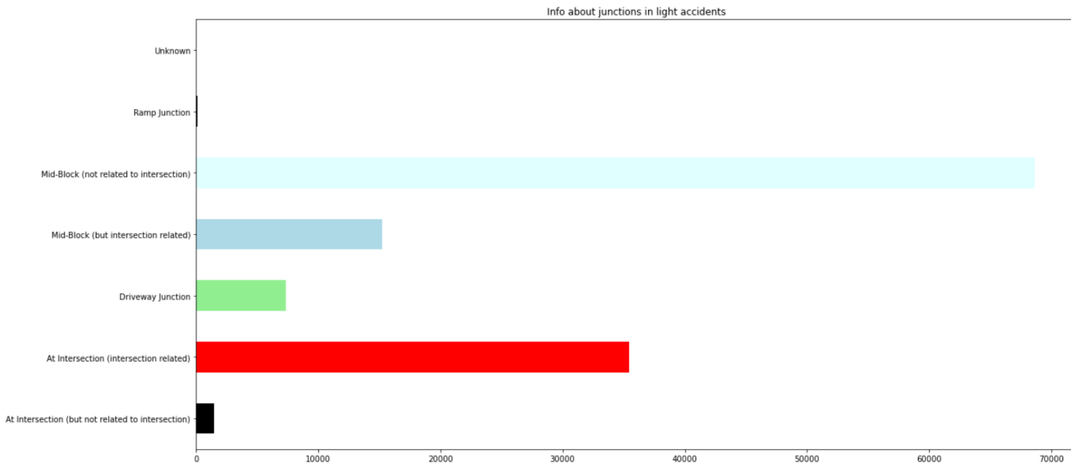
Data Preparing and Cleaning

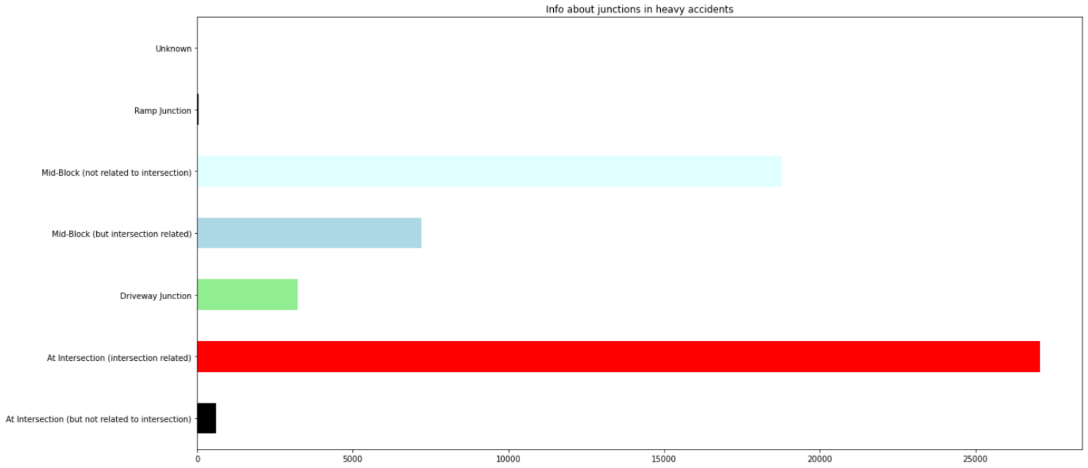
- The final data to use

	SEVERITYCODE	X	Y	OBJECTID	ADDRTYPE	LOCATION	SEVERITYDESC	COLLISIONTYPE	PERSONCOUNT	PEDCOUNT	...	JUNCTIONTYPE	INATTENTIONIND	UNDERINFL	WEATHER	ROADCOND	LIGHTCOND	PEDROWNOTGRNT	SPEEDING	ST_COLCODE	HITPARKEDCAR
0	2	-122.323148	47.703140	1	Intersection	5TH AVE NE AND NE 103RD ST	Injury Collision	Angles	2	0	...	At Intersection (intersection related)	NaN	N	Overcast	Wet	Daylight	NaN	NaN	10	N
1	1	-122.347294	47.647172	2	Block	AURORA BR BETWEEN RAYE ST AND BRIDGE WAY N	Property Damage Only Collision	Sideswipe	2	0	...	Mid-Block (not related to intersection)	NaN	0	Raining	Wet	Dark - Street Lights On	NaN	NaN	11	N
2	1	-122.334540	47.607871	3	Block	4TH AVE BETWEEN SENECA ST AND UNIVERSITY ST	Property Damage Only Collision	Parked Car	4	0	...	Mid-Block (not related to intersection)	NaN	0	Overcast	Dry	Daylight	NaN	NaN	32	N
3	1	-122.334803	47.604803	4	Block	2ND AVE BETWEEN MARION ST AND MADISON ST	Property Damage Only Collision	Other	3	0	...	Mid-Block (not related to intersection)	NaN	N	Clear	Dry	Daylight	NaN	NaN	23	N
4	2	-122.306426	47.545739	5	Intersection	SWIFT AVE S AND SWIFT AV OFF RP	Injury Collision	Angles	2	0	...	At Intersection (intersection related)	NaN	0	Raining	Wet	Daylight	NaN	NaN	10	N
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
194668	2	-122.290826	47.565408	219543	Block	34TH AVE S BETWEEN S DAKOTA ST AND S GENESEE ST	Injury Collision	Head On	3	0	...	Mid-Block (not related to intersection)	NaN	N	Clear	Dry	Daylight	NaN	NaN	24	N
194669	1	-122.344526	47.690924	219544	Block	AURORA AVE N BETWEEN N 85TH ST AND N 86TH ST	Property Damage Only Collision	Rear Ended	2	0	...	Mid-Block (not related to intersection)	Y	N	Raining	Wet	Daylight	NaN	NaN	13	N
194670	2	-122.306689	47.683047	219545	Intersection	20TH AVE NE AND NE 75TH ST	Injury Collision	Left Turn	3	0	...	At Intersection (intersection related)	NaN	N	Clear	Dry	Daylight	NaN	NaN	28	N
194671	2	-122.355317	47.678734	219546	Intersection	GREENWOOD AVE N AND N 68TH ST	Injury Collision	Cycles	2	0	...	At Intersection (intersection related)	NaN	N	Clear	Dry	Dusk	NaN	NaN	5	N
194672	1	-122.289360	47.611017	219547	Block	34TH AVE BETWEEN E MARION ST AND E SPRING ST	Property Damage Only Collision	Rear Ended	2	0	...	Mid-Block (not related to intersection)	NaN	N	Clear	Wet	Daylight	NaN	NaN	14	N

194673 rows × 24 columns

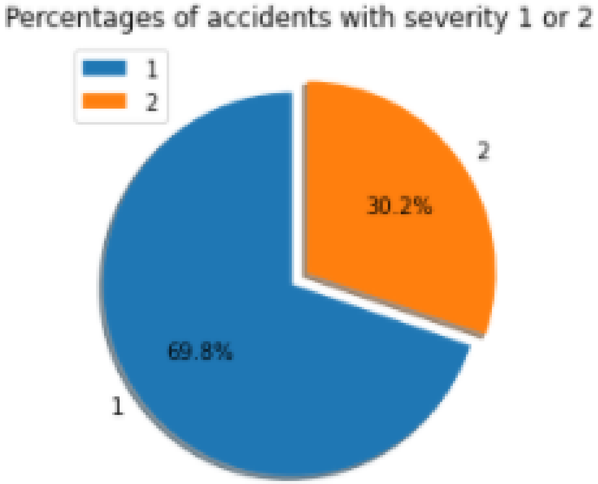
Analysis of Severity by groups





This parameter is very visual the change between the two groups of severity

```
SEVERITYCODE
1    69.83
2    30.17
Name: X, dtype: float64
```



Model without weather

$$Severity\ code = 1.11854 + Pedestrians \cdot 0.5963 + bicylce \cdot 0.6248 + vehicles \cdot 0.0129 + people \cdot 0.0473$$

Pedestrians = 0
Bycicles = 0
Vehicles = 2
People = 0
Severity = 1

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           SEVERITYCODE   R-squared:                       0.130
Model:                            OLS   Adj. R-squared:                  0.130
Method:                 Least Squares   F-statistic:                     7258.
Date:                Sun, 20 Sep 2020   Prob (F-statistic):               0.00
Time:                        22:16:52   Log-Likelihood:            -1.1059e+05
No. Observations:              194673   AIC:                         2.212e+05
Df Residuals:                  194668   BIC:                         2.212e+05
Df Model:                           4
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
const           1.1185      0.003    323.685      0.000       1.112       1.125
PEDCOUNT        0.5963      0.005    116.883      0.000       0.586       0.606
PEDCYLCOUNT     0.6248      0.006    103.846      0.000       0.613       0.637
VEHCOUNT        0.0129      0.002      7.211      0.000       0.009       0.016
PERSONCOUNT     0.0473      0.001     60.457      0.000       0.046       0.049
==============================================================================
Omnibus:                    25605.630   Durbin-Watson:                   1.993
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            31499.516
Skew:                           0.949   Prob(JB):                         0.00
Kurtosis:                       2.473   Cond. No.                         22.7
==============================================================================
```

Model with weather

$$\textit{Severity code} = 1.33804 + \textit{Pedestrians} \cdot 0.5527 + \textit{bicylce} \cdot 0.5846 + \textit{vehicles} \cdot 0.0168 + \textit{people} \cdot 0.0399 + \textit{speed cause} \cdot 0.1218 + \textit{parked car} \cdot (-0.1218) + \textit{junction} \cdot (-0.0441) + \textit{road} \cdot (-0.0027) + \textit{light} \cdot (-0.0045)$$

Pedestrians = 0
Bycicles = 0
Vehicles = 2
People = 0
Speeding = 1
Road = 1
Junctions = 2
Light = 1
Parked car = 1
Severity = 1

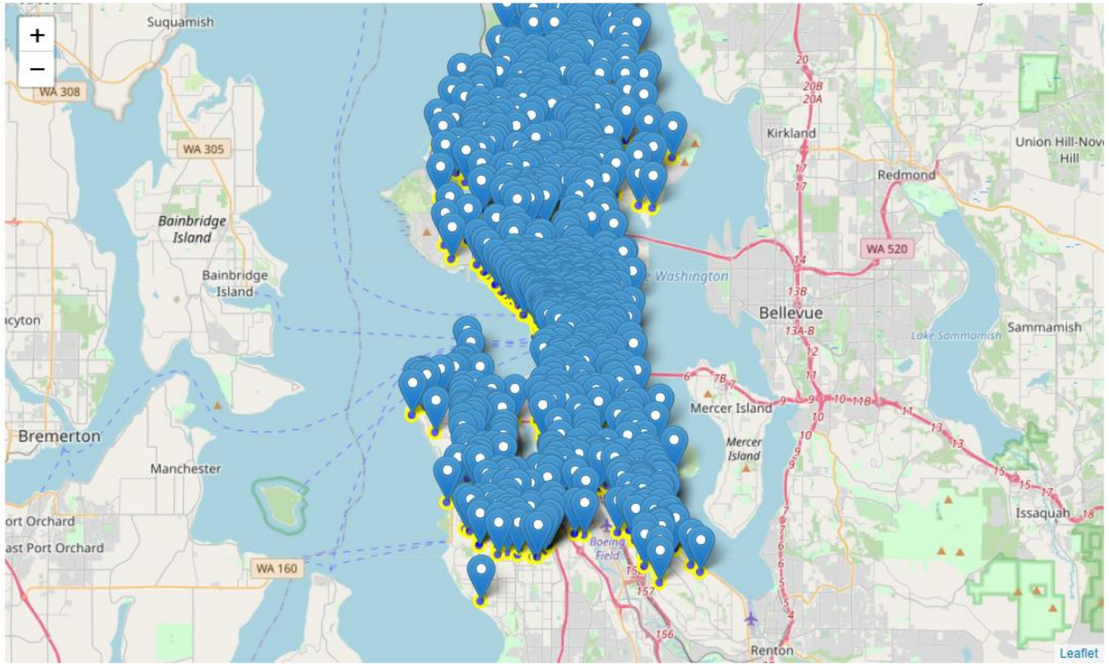
OLS Regression Results

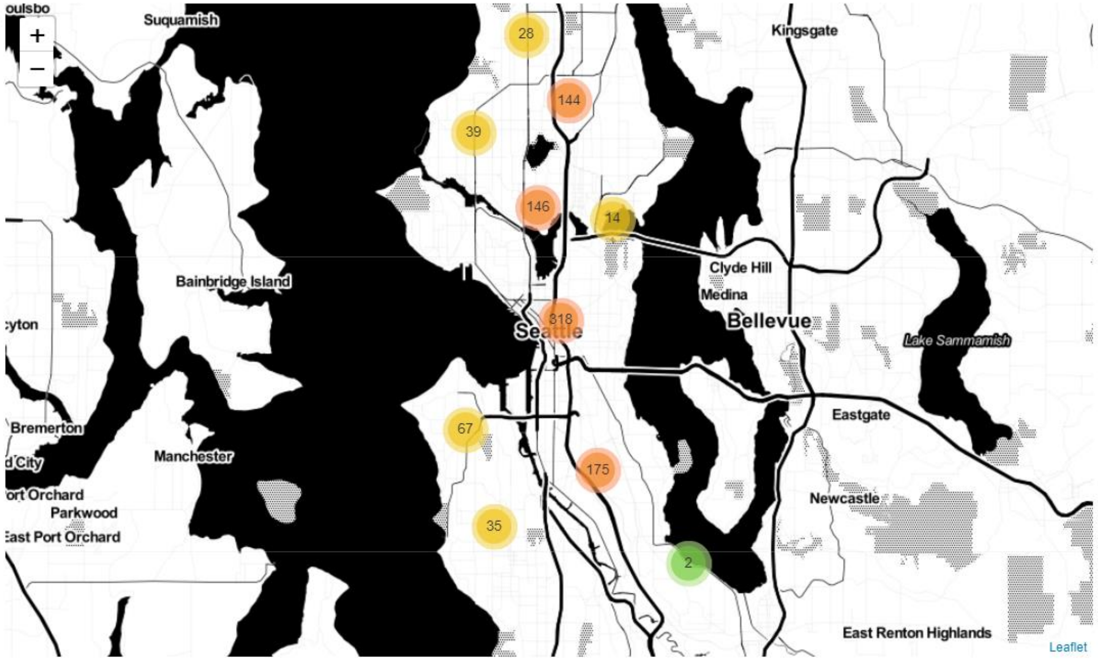
Dep. Variable:	SEVERITYCODE	R-squared:	0.150
Model:	OLS	Adj. R-squared:	0.150
Method:	Least Squares	F-statistic:	3824.
Date:	Sun, 20 Sep 2020	Prob (F-statistic):	0.00
Time:	23:10:04	Log-Likelihood:	-1.0827e+05
No. Observations:	194673	AIC:	2.166e+05
Df Residuals:	194663	BIC:	2.167e+05
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.3225	0.008	155.646	0.000	1.306	1.339
PEDCOUNT	0.5871	0.005	115.628	0.000	0.577	0.597
PEDCYLCOUNT	0.6212	0.006	103.689	0.000	0.609	0.633
VEHCOUNT	0.0352	0.002	18.938	0.000	0.032	0.039
PERSONCOUNT	0.0405	0.001	51.769	0.000	0.039	0.042
SPEEDING	0.1248	0.005	27.564	0.000	0.116	0.134
HITPARKEDCAR	-0.1765	0.005	-34.651	0.000	-0.187	-0.167
JUNCTIONTYPE	-0.0289	0.001	-44.801	0.000	-0.030	-0.028
ROADCOND	-0.0048	0.000	-17.961	0.000	-0.005	-0.004
LIGHTCOND	-0.0108	0.001	-20.357	0.000	-0.012	-0.010

Omnibus:	24984.877	Durbin-Watson:	1.990
Prob(Omnibus):	0.000	Jarque-Bera (JB):	29117.008
Skew:	0.904	Prob(JB):	0.00
Kurtosis:	2.437	Cond. No.	84.6

Maps


Suquamish
WA 308
WA 305
Bainbridge Island
Bainbridge Island
Bremerton
Manchester
ort Orchard
ast Port Orchard
WA 160
Kirkland
Redmond
Union Hill-Novelty Hill
WA 520
Bellevue
Sammamish
Lake Sammamish
Mercer Island
Mercer Island
Issaquah
Boeing Field
Renton
Leaflet


Suquamish
Kingsgate
28
144
39
146
14
Bainbridge Island
Clyde Hill
Medina
Bellevue
318
Seattle
Lake Sammamish
Bremerton
Manchester
67
Eastgate
175
ort Orchard
Parkwood
East Port Orchard
35
Newcastle
2
East Renton Highlands
Leaflet

Results

- Weather doesnt have a bigger influence
- Easy to predict using people involve in the crash
- Difficul to predict trying to find the causes of accident
- Accidents are condensed in the road main city and city center
- 1 of each 3 accidents is dangerous

Discussions

- Careful with too many variables – Overfitting
- Too many variables – bar plot
- Use of subplots
- To represent maps is better use only a part of dataset
- Subtitute qualitive variables for number to add it to the model
- Use statsmodels.api to see the model efficient

Conclusions

- The spread of accidents is irregular. We only plot in the map a randomly part of 1000 accidents. In the city-center is higher.
- Carrying on with these, one reason could be that is most common accidents in interceptions (are principal in the city center).
- The model has a higher dependence of how many people and of what type are in the crash.
- The conditions of the enviroment are important too, but less, with them we have over fitting in our model.
- Weather conditions are not good enough for our model, cause most of accidents occurs days with good weather. With the bar plots we can see the influence, but the big amount of sunny days with accidents make that when we add this data to the model, we have overfitting.