Capstone Project: Seattle Collision Data

APPLIED DATA SCIENCE CAPSTONE ON COURSERA

BY JARED BEST

Introduction / Business Problem

- The goal of this capstone project is to predict:
  - **Injury collisions** (SEVERITYCODE = 2) and
  - **Property damage collisions** (SEVERITYCODE = 1).
- This project and its outcomes are relevant to those who are interested in preventing or reducing injury collisions and property damage.

Data Set Summary

Title	Collisions—All Years
Abstract	All collisions provided by SPD and recorded by Traffic Records.
Description	This includes all types of collisions. Collisions will display at the intersection or mid-block of a segment.
Timeframe	2004 to Present
Update Frequency	Weekly
Keyword(s)	SDOT, Seattle, Transportation, Accidents, Bicycle, Car, Collisions, Pedestrian, Traffic, Vehicle

Data Set Head

SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	...	ROADCOND	LIGHTCOND
0	2	-122.323148	47.703140	1	1307	1307	3502005	Matched	Intersection	37475.0	...	Wet
1	1	-122.347294	47.647172	2	52200	52200	2607959	Matched	Block	NaN	...	Wet
2	1	-122.334540	47.607871	3	26700	26700	1482393	Matched	Block	NaN	...	Dry
3	1	-122.334803	47.604803	4	1144	1144	3503937	Matched	Block	NaN	...	Dry
4	2	-122.306426	47.545739	5	17700	17700	1807429	Matched	Intersection	34387.0	...	Wet

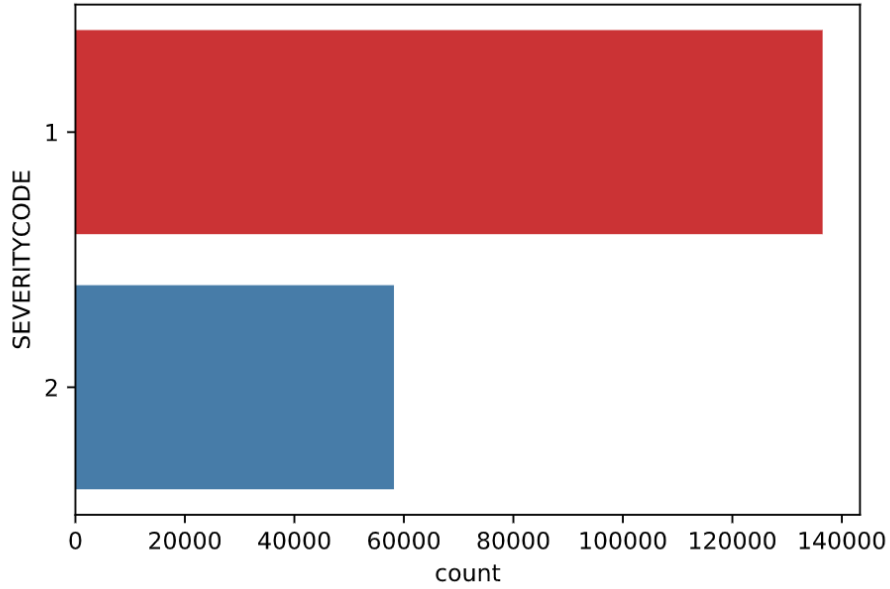
PEDROWNOTGRNT	SDOTCOLNUM	SPEEDING	ST_COLCODE	ST_COLDESC	SEGLANEKEY	CROSSWALKKEY	HITPARKEDCAR	
Daylight	NaN	NaN	NaN	10	Entering at angle	0	0	N
Dark - Street Lights On	NaN	6354039.0	NaN	11	From same direction - both going straight - bo...	0	0	N
Daylight	NaN	4323031.0	NaN	32	One parked--one moving	0	0	N
Daylight	NaN	NaN	NaN	23	From same direction - all others	0	0	N
Daylight	NaN	4028032.0	NaN	10	Entering at angle	0	0	N

Targets

Severity code [SEVERITYCODE]:

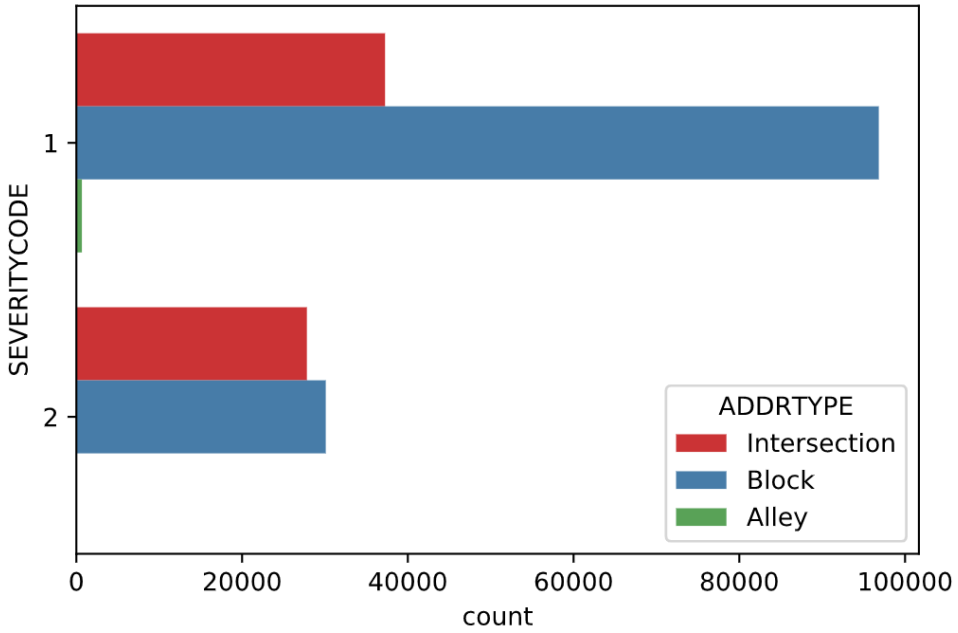
- 1 = Property damage collision
- 2 = Injury collision

Plot of collisions by severity code:



Address Type [ADDRTYPE]

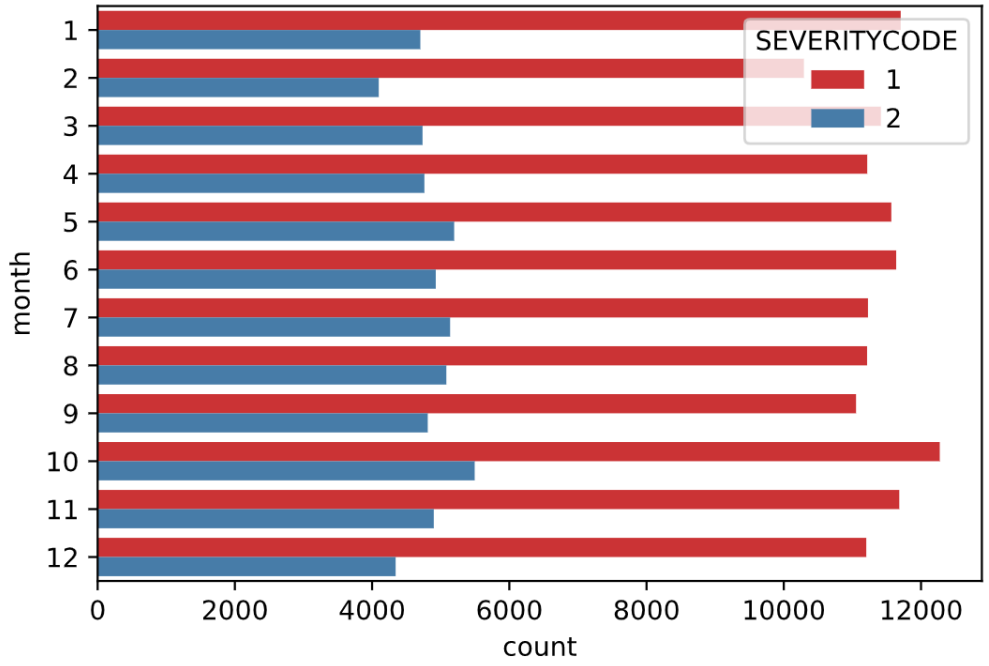
- The plot shows the count of collisions by severity code and grouped by address type
- The table shows the mean of all severity codes per address type
  - Indicates that intersections have more injury collisions than other address types



	SEVERITYCODE
ADDRTYPE	
Intersection	1.427524
Block	1.237115
Alley	1.109188

Collision Breakdown per Month

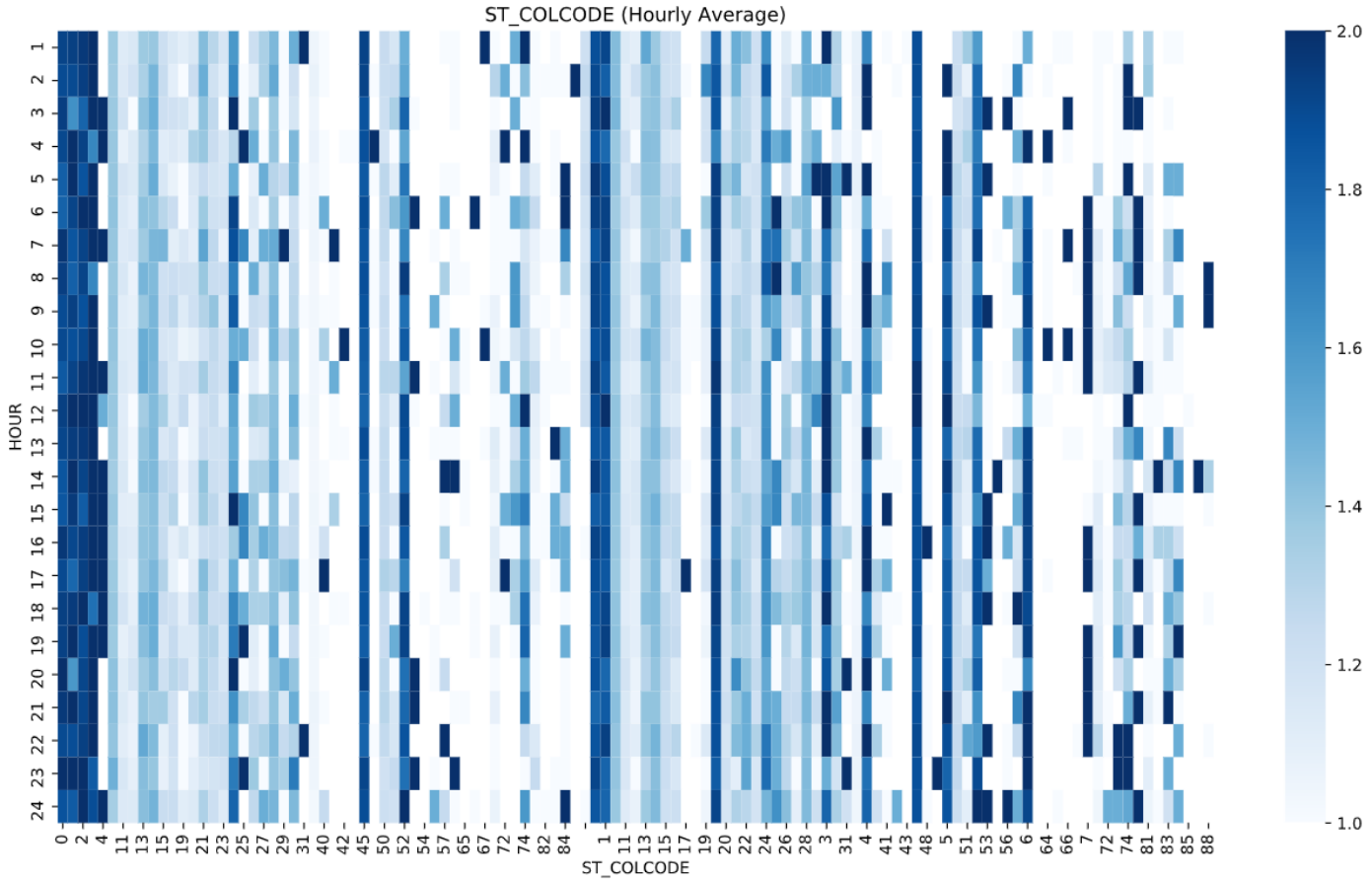
- The following chart shows the number of collisions by severity code for each month



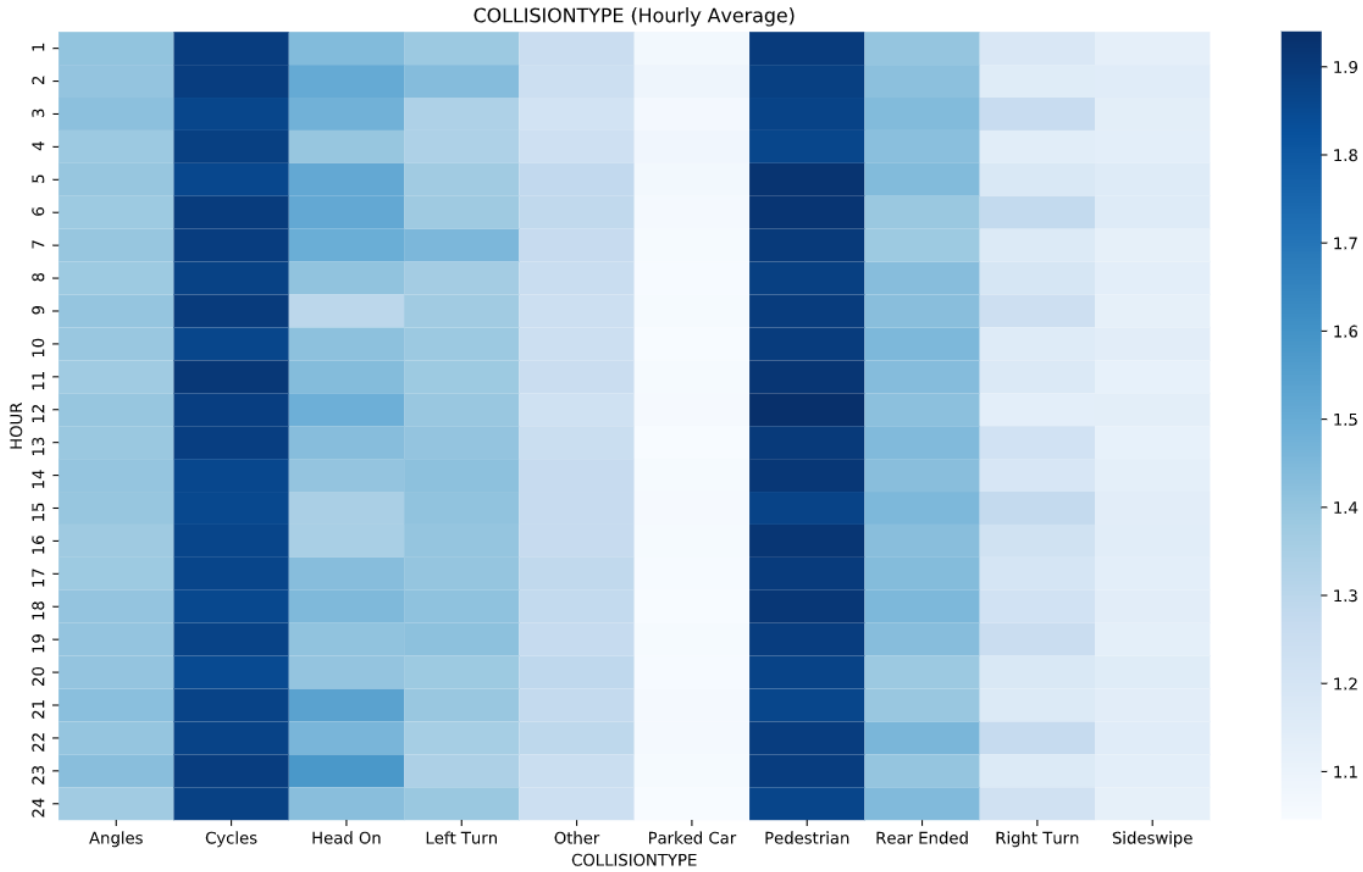
month	SEVERITYCODE
7	1.313921
8	1.311856
5	1.309968
10	1.309264
9	1.303265
4	1.298035
6	1.297477
11	1.295441
3	1.293189
1	1.286646
2	1.284712
12	1.279382

Heatmap: ST_COLCODE (Hourly)

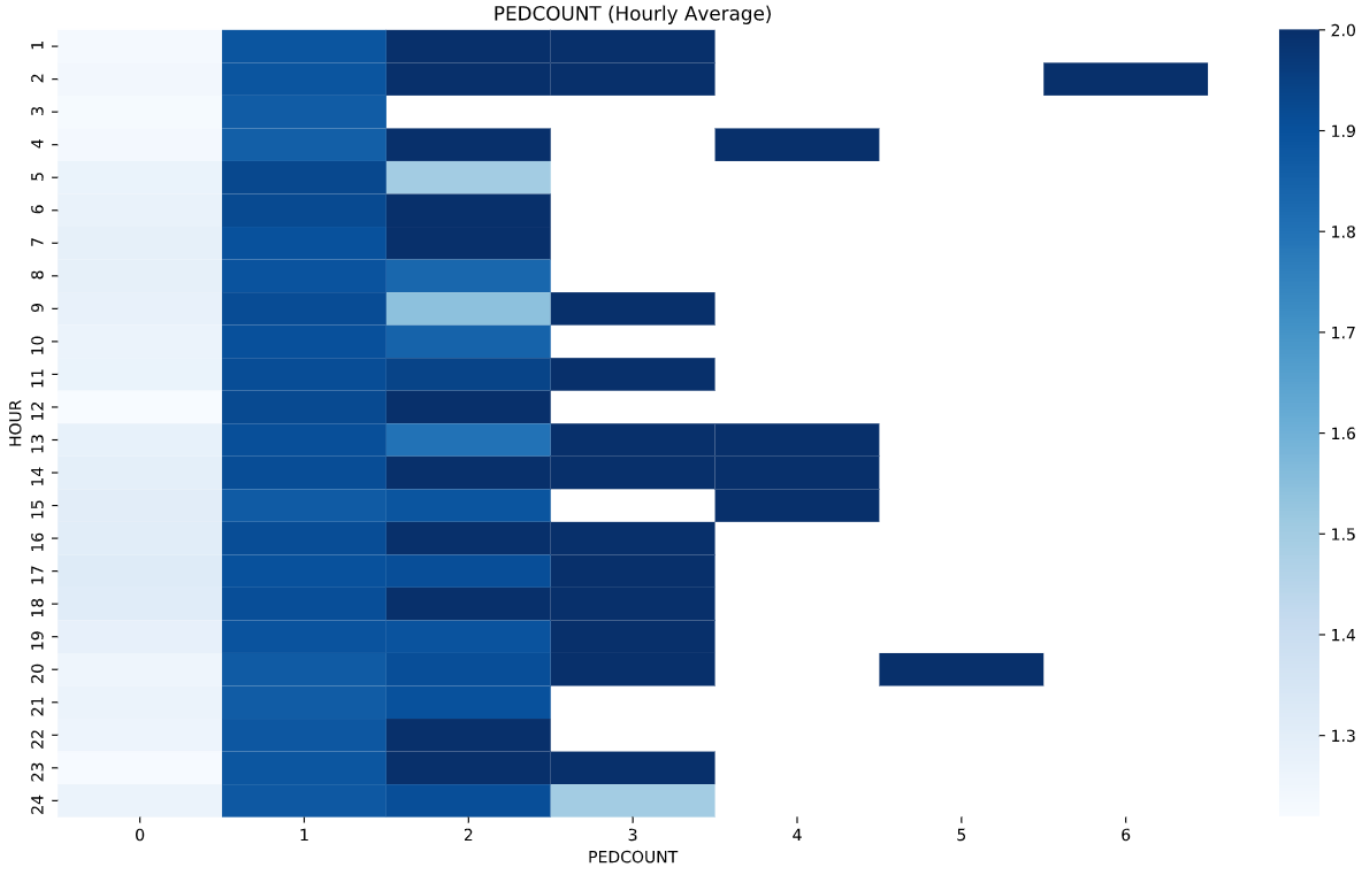
ST_COLCODE	SEVERITYCODE
49	2.000000
87	2.000000
87	2.000000
4	1.950000
7	1.944444
...	...
65	1.000000
82	1.000000
60	1.000000
85	1.000000
65	1.000000



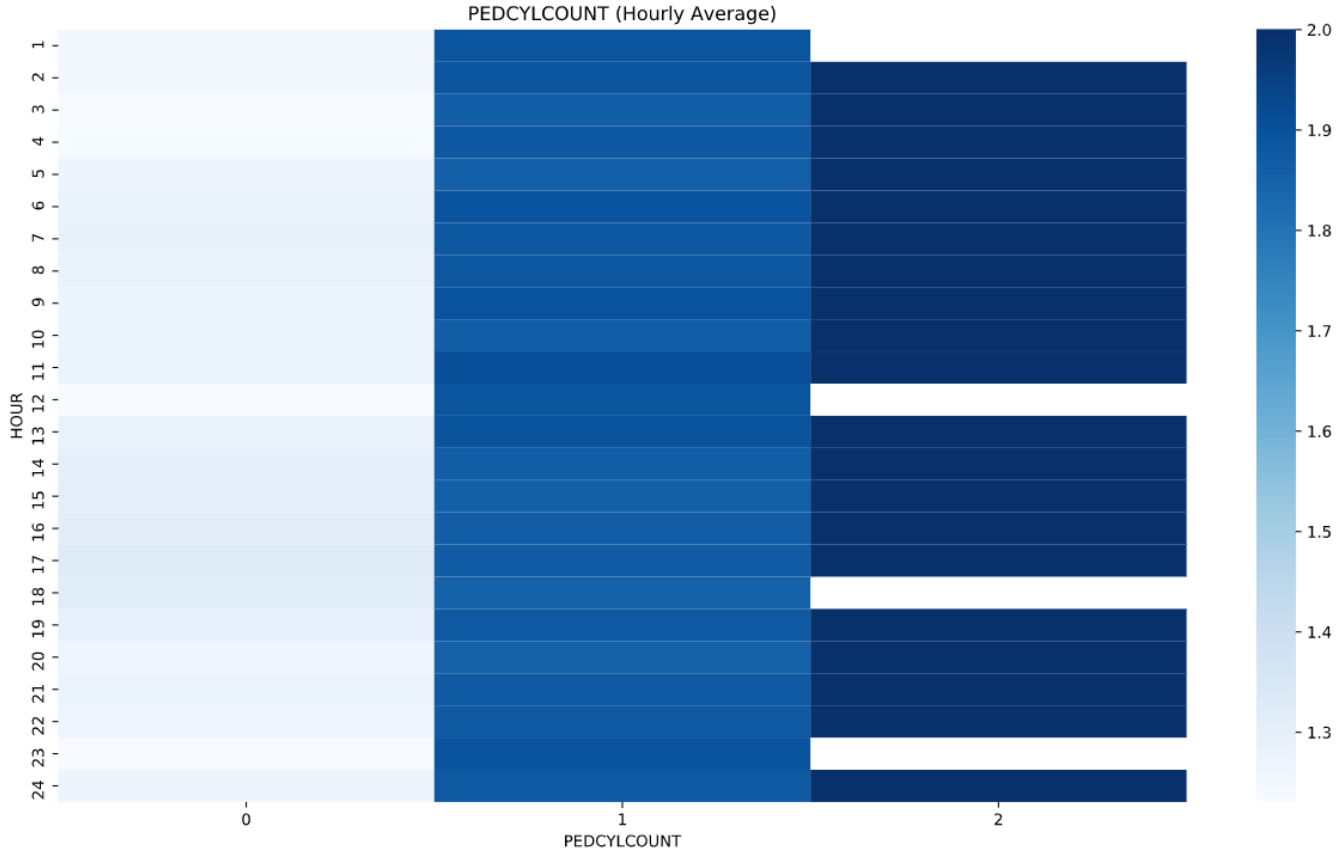
Heatmap: COLLISIONTYPE (Hourly)



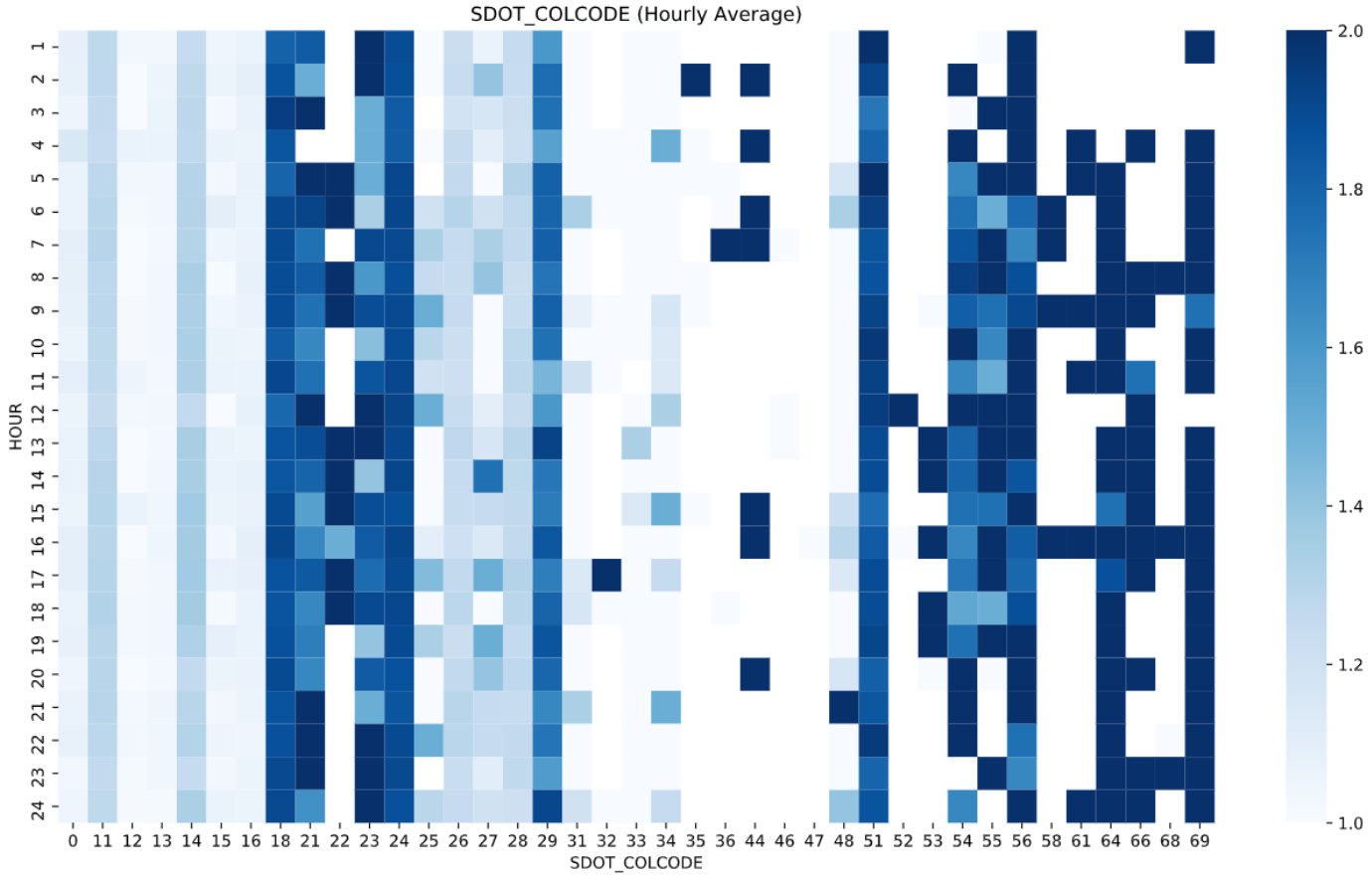
Heatmap: PEDCOUNT (Hourly)



Heatmap: PEDCYLCOUNT (Hourly)



Heatmap: SDOT_COLCODE (Hourly)



Balancing Targets

- As the data set is unbalanced, we must perform sampling methods to avoid any potential unbalanced issues.
- Over sampling is selected in this case

Models

- This focus of this project is on predicting injury and property damage collisions; therefore, this is a classification problem.
- The following models are evaluated:
  - MLPClassifier
  - Naive Bayes
  - XGBoost
- These were selected due to their performance and popularity in the literature.

Model Comparison

 # MLPClassifier

Test_Confusion_Matrix:

[[16792 9765]

[2378 8933]]

Train_ROC_AUC_Score: 0.7115526574803149

Test_ROC_AUC_Score: 0.711031256769306

Test_Sensitivity : 0.789762178410397

Test_Specificity : 0.6323003351282148

Test_Accuracy_Score: 0.6793334741734446

	precision	recall	f1-score	support
1	0.88	0.63	0.73	26557
2	0.48	0.79	0.60	11311
accuracy			0.68	37868
macro avg	0.68	0.71	0.66	37868
weighted avg	0.76	0.68	0.69	37868

 # Naive Bayes

Test_Confusion_Matrix:

[[26273 284]

[9043 2268]]

Train_ROC_AUC_Score: 0.5936080405814658

Test_ROC_AUC_Score: 0.5949093980930091

Test_Sensitivity : 0.20051277517460878

Test_Specificity : 0.9893060210114094

Test_Accuracy_Score: 0.7536970529206718

	precision	recall	f1-score	support
1	0.74	0.99	0.85	26557
2	0.89	0.20	0.33	11311
accuracy			0.75	37868
macro avg	0.82	0.59	0.59	37868
weighted avg	0.79	0.75	0.69	37868

 # XGBoost

Test_Confusion_Matrix:

[[16798 9759]

[2400 8911]]

Train_ROC_AUC_Score: 0.7118129164142943

Test_ROC_AUC_Score: 0.7101717166945873

Test_Sensitivity : 0.7878171691273981

Test_Specificity : 0.6325262642617766

Test_Accuracy_Score: 0.6789109538396535

	precision	recall	f1-score	support
1	0.87	0.63	0.73	26557
2	0.48	0.79	0.59	11311
accuracy			0.68	37868
macro avg	0.68	0.71	0.66	37868
weighted avg	0.76	0.68	0.69	37868

Final Model

- AUC, ROC, Sensitivity, Specificity, Precision, and Accuracy are compared for model evaluation.
- In the three models used in the current study, Naive Bayes has the highest Accuracy of 0.75; however, it has the lowest AUC, ROC score. Furthermore, its sensitivity is also low (at 0.2). Despite its high specificity value, Naive Bayes will not be considered.
- XGBoost and MLPClassifier both perform well, but as MLPClassifier slightly outperforms XGBoost.
- MLPClassifier will be selected as a final model.

Conclusion

- The purpose of this project is to predict injury collisions and property damage collisions.
- The final model selected is the MLPClassifier, which provides AUC, ROC score of .71 and a detection rate of injury collisions equal to 79% and of property damage collisions equal to 63%.
- Its overall accuracy is 68%.
- This model will inform relevant stakeholders to predict injury and property damage collisions.