Analysis of Car Crashes in the Portland Metro Area

Thomas Sato, Miles DeVaney, Chris Olivia
CS-475 Machine Learning With Python

Project Abstract

This project will analyze a dataset of car crashes in the Portland Metro area released by Oregon Metro to algorithmically identify the areas of the city which are actually more prone to accidents.

We are using the *Oregon Statewide Crash Data System* data posted by the Oregon Department of Transportation

Main Goal: To come up with solutions or recommendations to decrease the car crashes rate and severity in a specific area of Portland

Context From State/City Programs

All Roads Transportation Safety (ARTS) Program

- Data driven approach to decreasing fatalities/serious injuries
- Roadway departure, intersection, bicycle and pedestrian
- Identification of hotspots of crashes and systemic targeting of issues

PBOT Vision Zero Program

- Identifies a High-Crash Network of streets/intersections
- Setting and enforcing safer speed limits
- Redesigning of streets and lane structure

What is the dataset

The Crash Analysis and Reporting (CAR) Unit of ODOT compiles data for reported motor vehicle traffic crashes occurring on city streets, county roads and state highways.

The data consist of crashes that happened between 2007 - 2021.

Description

Motor vehicle traffic crash points for 2007-2021 throughout the Metro region provided by ODOT. Includes injuries by severity. Additional attributes have been derived. ODOT's methodology can be found here: https://www.oregon.gov/odot/Data/documents/CDS_Code_Manual.pdf

Date of last data update: 2023-08-11

This is official RLIS data.

Data Cleaning

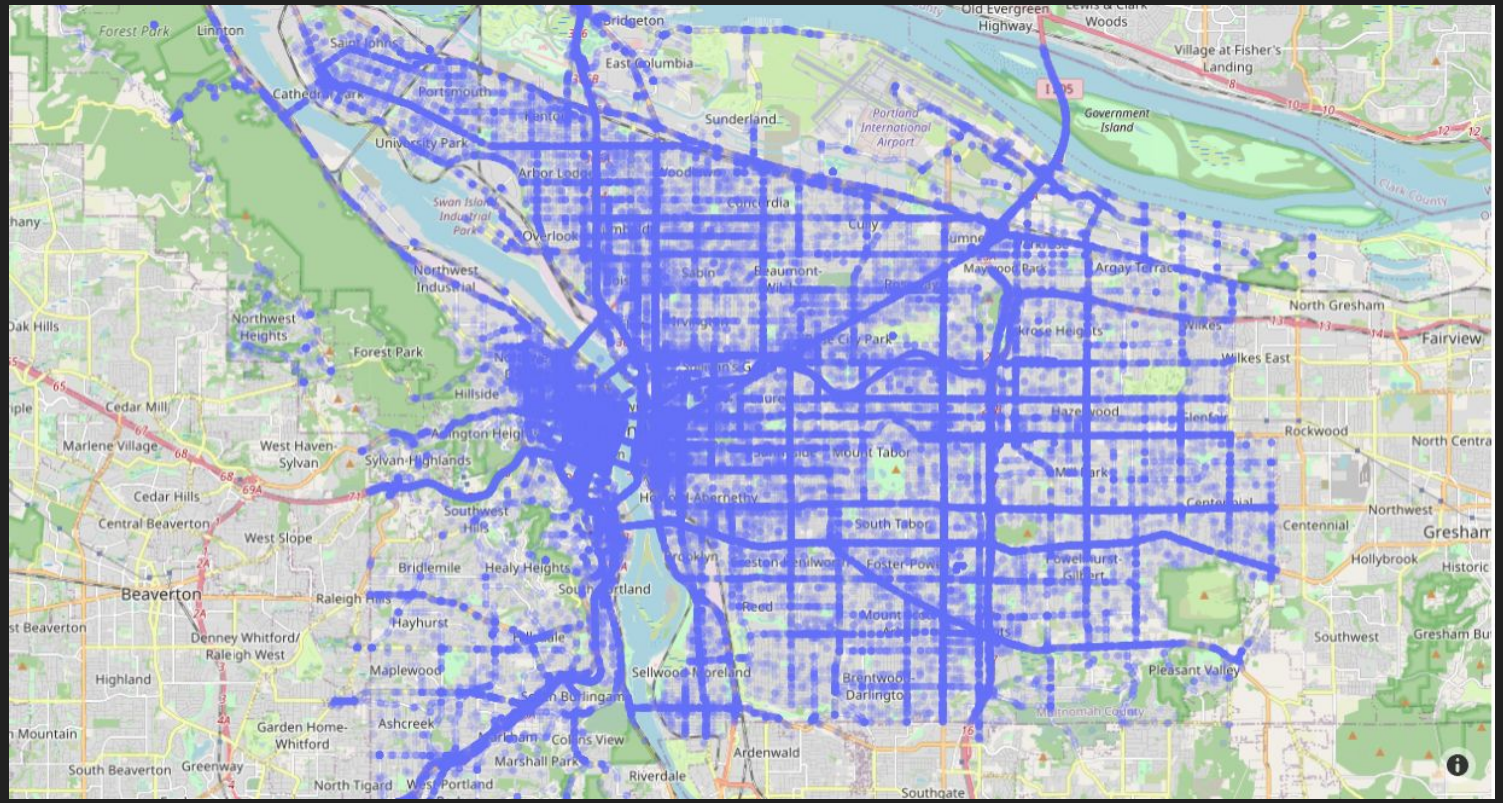
- Initial dataset had 151 columns and ~340,000 rows
- Many columns contained information which was duplicated elsewhere, or entirely contained the same value
- The dataset had crashes from the broader Portland area, we decided to focus only on crashes actually within the city of Portland
- Final dataset has 77 columns and ~150,000 rows

EDA (Mapping)

All Data

2007-2021

Highest density downtown and along highways

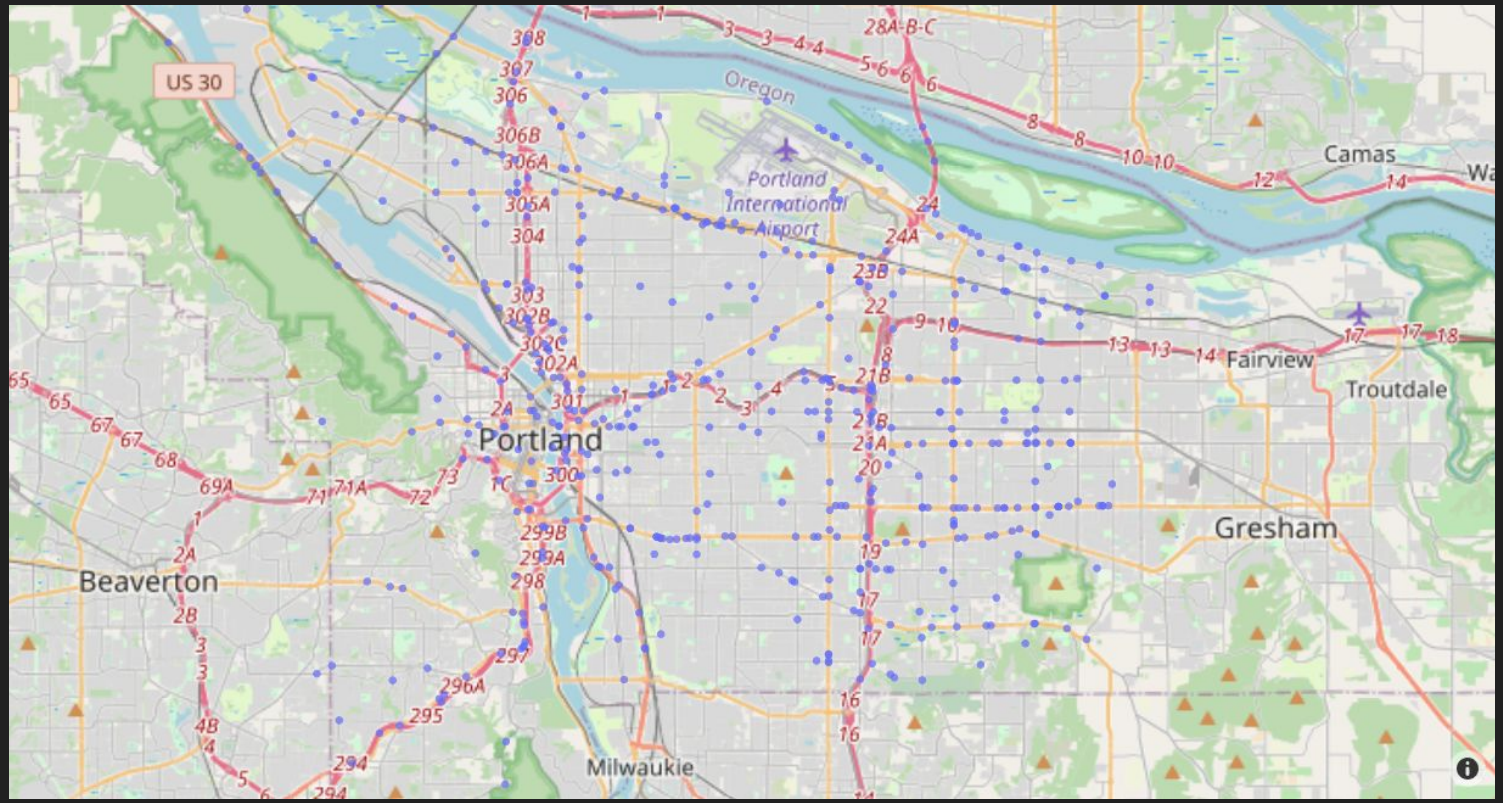


EDA (Mapping cont.)

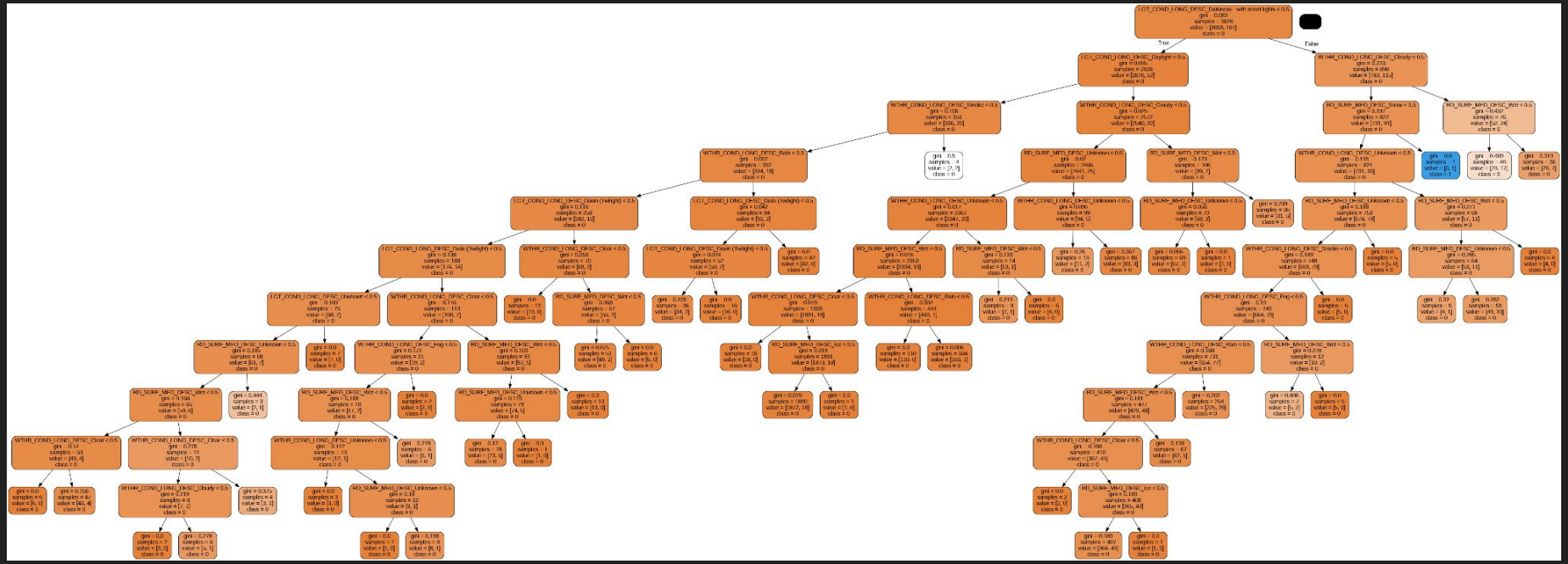
Fatal Only

2007-2021

Harder to find obvious clusters



Classification Tree



Feature Selection Methodology

- Chi-squared tests of independence
- Discretizing numeric features, encoding of categorical features

```
# Defining target and feature data
X = chi2_crashes.drop(["FATAL", "CRASH_SVRTY_LONG_DESC"], axis = 1)
y = chi2_crashes['FATAL']

preprocessor = ColumnTransformer(
    transformers=[
        ('numeric', KBinsDiscretizer(n_bins = 200, encode = 'ordinal'), numeric_columns),
        ('categorical', OneHotEncoder(), categorical_columns)
    ])

pipeline = Pipeline([
    ('preprocessor', preprocessor),
    ('feature_selection', SelectKBest(chi2, k = 15))  # You can specify the desired number of features (k)
])

pipeline.fit(X, y)
X_transformed = pipeline.transform(X)

selected_feature_indices = pipeline.named_steps['feature_selection'].get_support(indices = True)
transformed_feature_names = pipeline.named_steps['preprocessor'].get_feature_names_out()
```

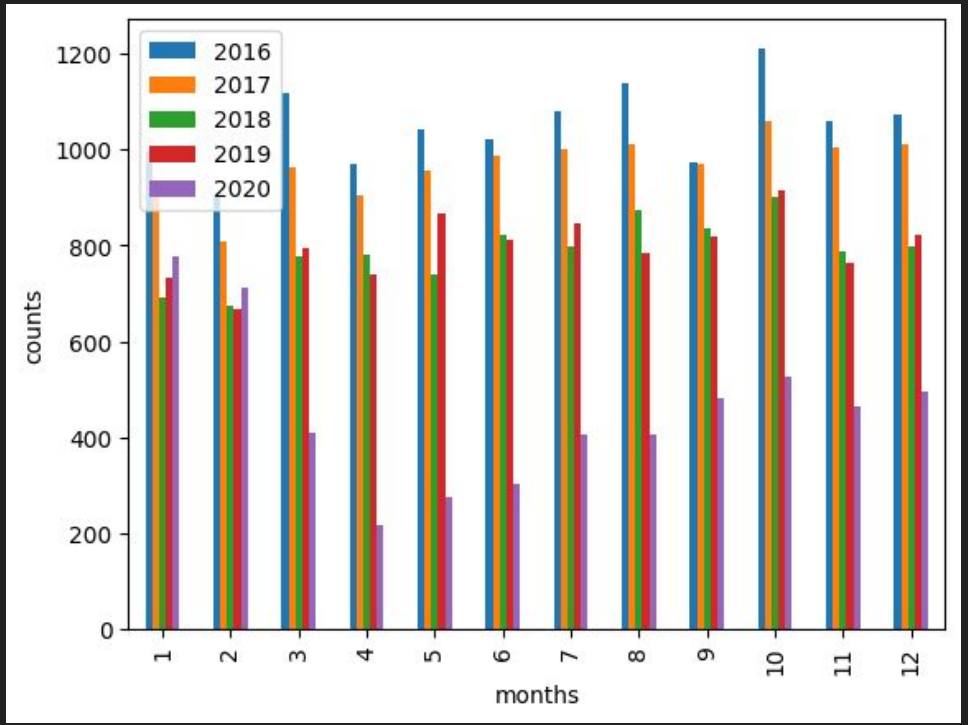
Feature Selection Results

Most important features that indicate dependence on whether or not a car crash is fatal. Goal is to identify features that are *controllable*

1. Drug Involvement
2. Driving in Excess of Posted Speed
3. Number of Pedestrians Involved
4. Alcohol Involvement
5. School Zone
6. Work Zone
7. Report Received by City Police
8. Cloudy Weather Condition
9. Total Persons Not Using Safety Equipment
10. Number of Motorcyclists Involved

Analysis of Features (Year 2016 - 2020)

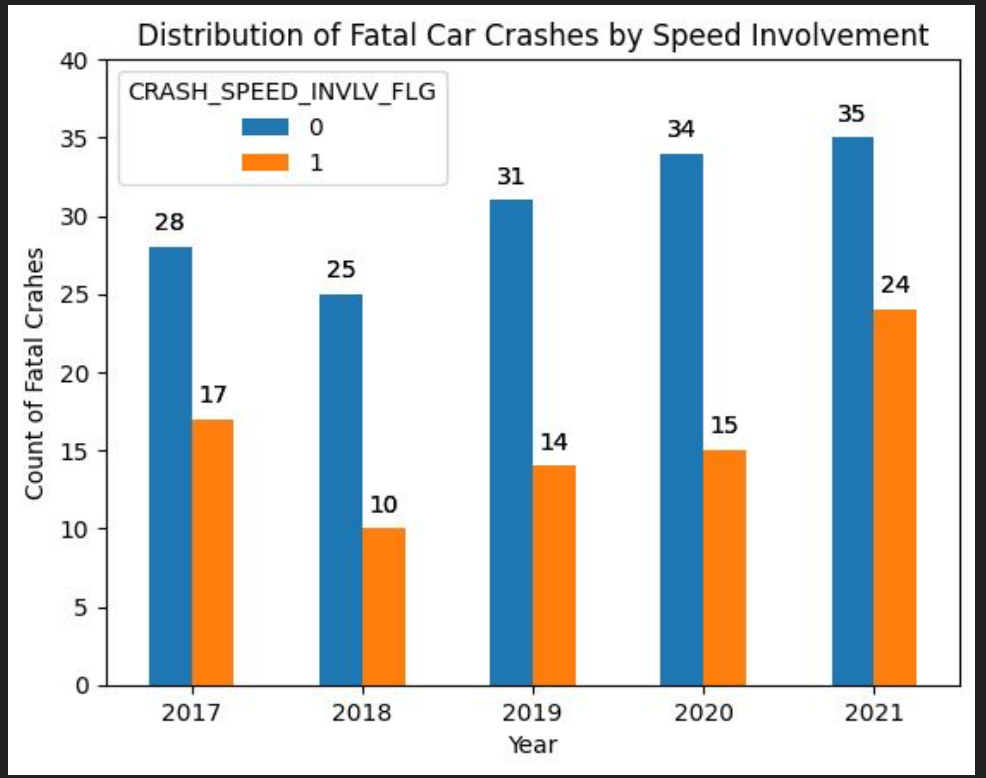
- 2016 has the highest crashes rate
- Crashes count for 2018 and 2019 are significantly less than 2016
- Crashes rate significantly decreased during the start of covid starting from March 2020



Analysis of Features (Speed Involvement)

- Around ⅓ of fatal car crashes every year involve driving at excessive speed
- The ratio of speed involvement in fatal crashes seems to be rising

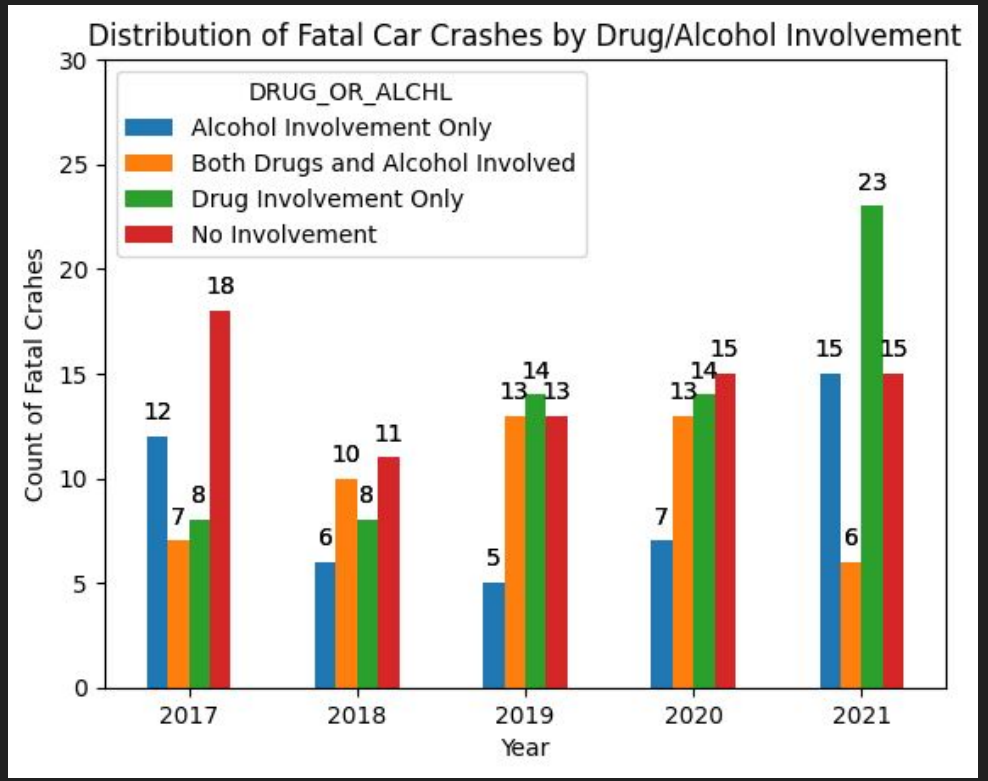
Fatal Crashes	Speed Not Involved	Speed Involved	Involvement Ratio
2017	28	17	0.39
2018	25	10	0.30
2019	31	14	0.32
2020	34	15	0.32
2021	35	24	0.41



Analysis of Features (Drug/Alcohol Involvement)

- Over 60% of fatal car crashes involved drugs or alcohol every year from 2017 - 2021
- Drug usage seems to be more prevalent than alcohol usage in terms of fatal crashes

Fatal Crashes	Alcohol Involvement Only	Both Drugs and Alcohol Involved	Drug Involvement Only	No Involvement	Involvement Ratio
2017	12	7	8	18	0.60
2018	6	10	8	11	0.69
2019	5	13	14	13	0.71
2020	7	13	14	15	0.69
2021	15	6	23	15	0.75



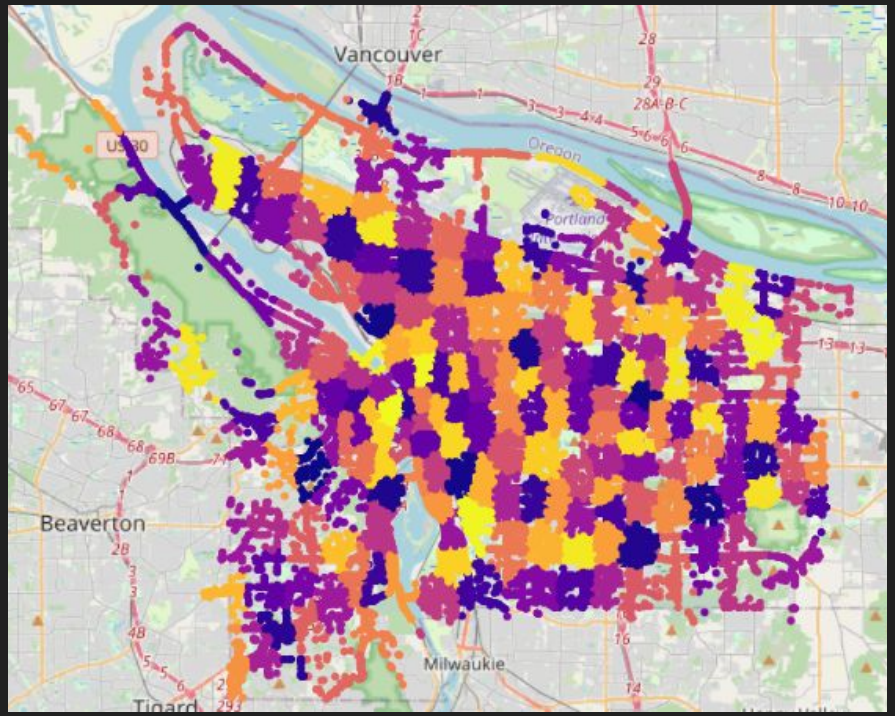
Clustering Methods

- K-Means clustering algorithm
- Works well with geographic data
- Examined clustered areas for patterns of drug/alcohol involvement and excessive speeding

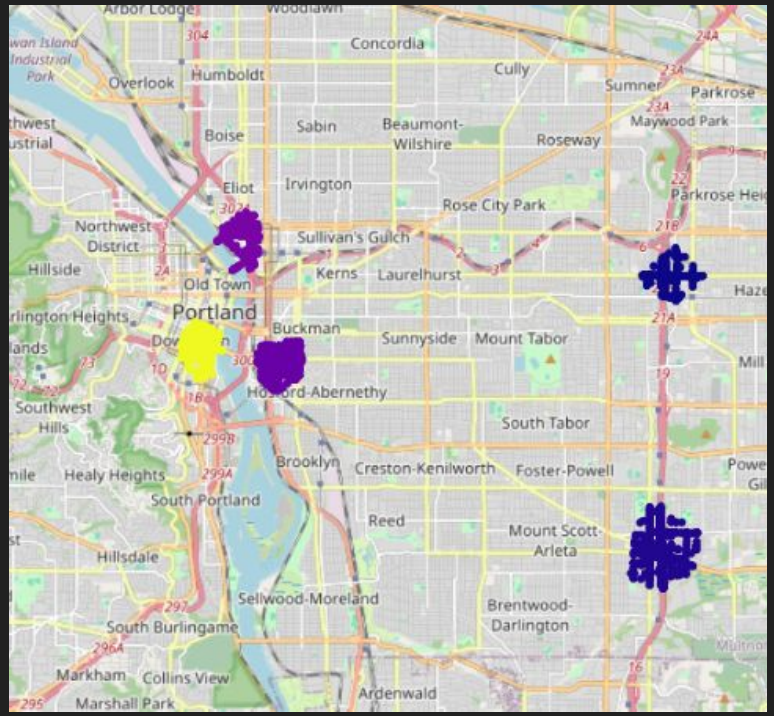
```
from sklearn.cluster import KMeans

yrs = crashes_portland[crashes_portland['CRASH_YR_NO'].isin(range(2017, 2022))]
features = yrs[['LAT_DD', 'LONGTD_DD', 'DRUG_INVLV_FLG', 'ALCHL_INVLV_FLG', 'CRASH_SPEED_INVLV_FLG']]

kmeans = KMeans(n_clusters = 150, random_state = 42)
y_pred = kmeans.fit_predict(features[['LAT_DD', 'LONGTD_DD']])
features['cluster'] = y_pred
```

Overview of Clusters Found

All clusters 2007-2021:

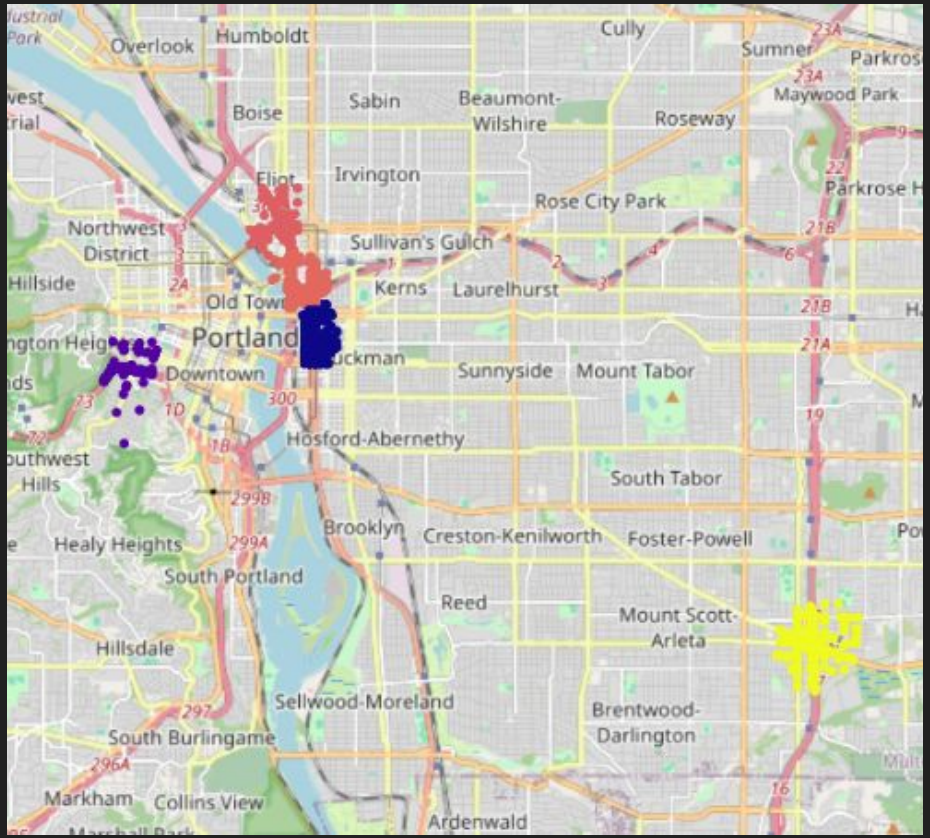


5 Largest clusters 2007-2021:



Clusters Overview (cont.)

5 Largest clusters 2017-2021:

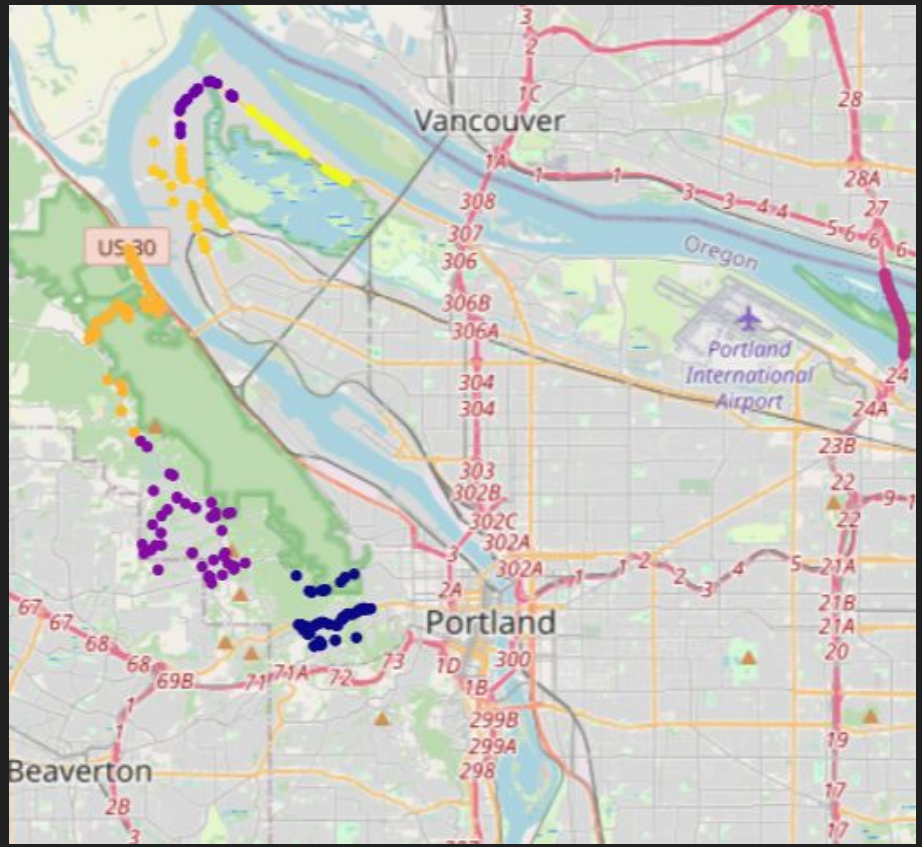


Analysis of Clusters (speed)

Clusters for “driving in excess of posted speed” from 2017-2022:

The highest speeding percentages were the yellow and purple clusters in the top left, with a posted speed limit of 40.

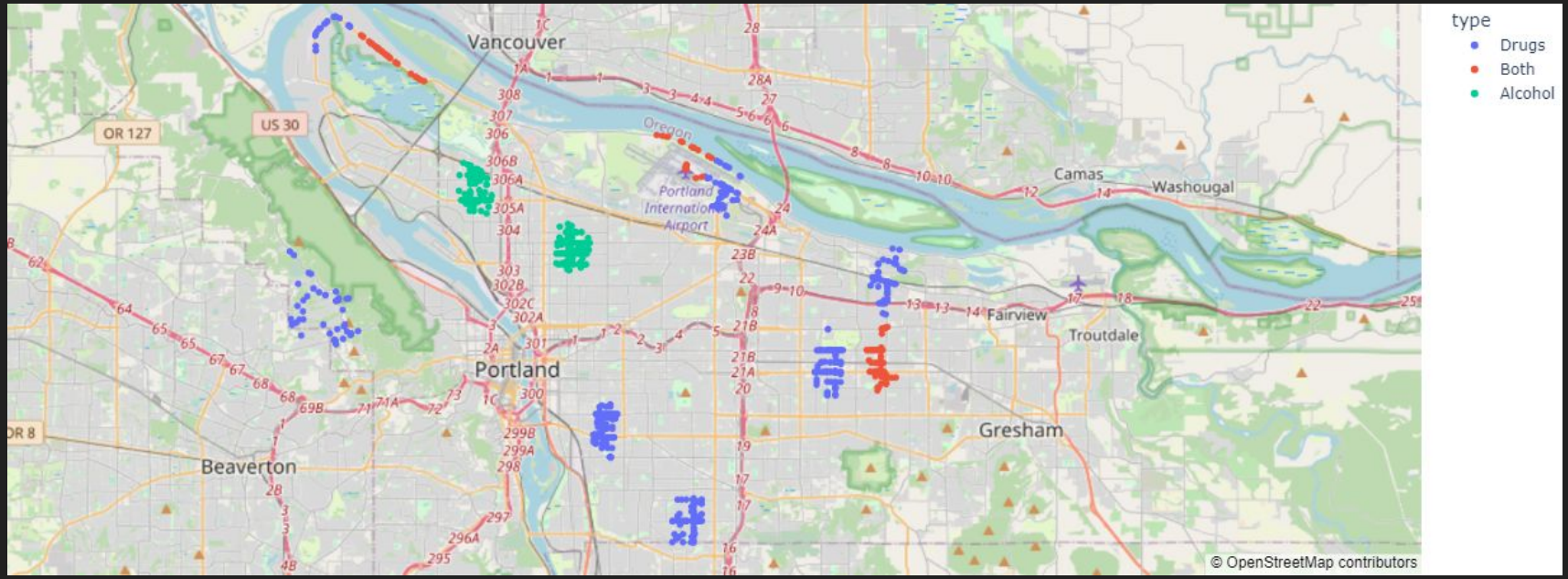
The other clusters had even lower posted speed limits, from 25-40.



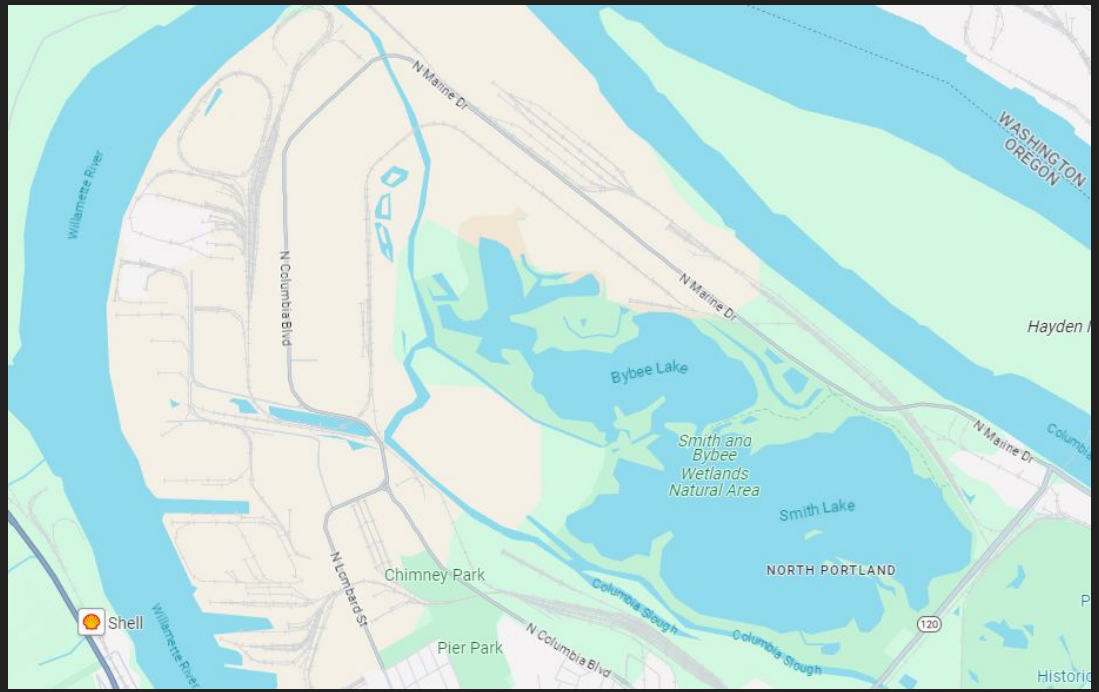
Analysis of Clusters (drugs/alcohol)

Clusters for drug and alcohol involvement from 2017-2022:

There is some noticeable overlap with the previous clusters, again specifically in the northwest corner



Our solution



- City measures to enforce speed limits on Marine Drive
- Implementation of speed cameras
- Increased police surveillance, specifically focused on catching DUI offences

Comparing City Initiates

- City focuses on pedestrian safety and structural changes
- We found school/work zones to be effective
- Focus on clusters of drug/alcohol usage
- More enforcement in lower-speed areas

Answering a question from the professor

- Is it generally safer to take I-5 or 205 through Portland?
- In terms of likelihood to be involved in a crash, no specific factors were significant

```
Percentage of I5 crashes that were fatal: 0.23859515174651652
Percentage of 205 crashes that were fatal: 0.133712184522814 66
Ratio 205 over I5: 0.5604145077720208

Percentage of rainy I5 crashes which were fatal: 0.1929570670525808
Percentage of rainy 205 crashes which were fatal: 0.07304601899196494
Ratio 205 over I5: 0.3785609934258583
```

- It seems that if you are involved in a crash, it is likely to be much less severe on 205 compared to I-5

Questions?

Sources

https://rlisdiscovery.oregonmetro.gov/datasets/drcMetro::crashes/about

https://www.oregon.gov/odot/Data/documents/CDS_Code_Manual.pdf

https://www.oregon.gov/odot/Engineering/ARTS/Key-Facts.pdf

https://www.portland.gov/transportation/vision-zero/documents/vision-zero-portland-2022-deadly-traffic-crash-report/download