UNIVERSITY OF MARYLAND

Beyond Good and Evil: Analyzing Washington Crash Files

Team IC23001
Members: Chaitanya Pohnerkar, Akhil Reddy, Jarrar Haider, Eeshan Agarwal
Organization Name: Washington Traffic Safety Commission (WTSC)
UMD Information Challenge 2023

We are first year graduate students at Robert H. Smith School of Business

Akhil Reddy

Chaitanya Pohnerkar

Jarrar Haider

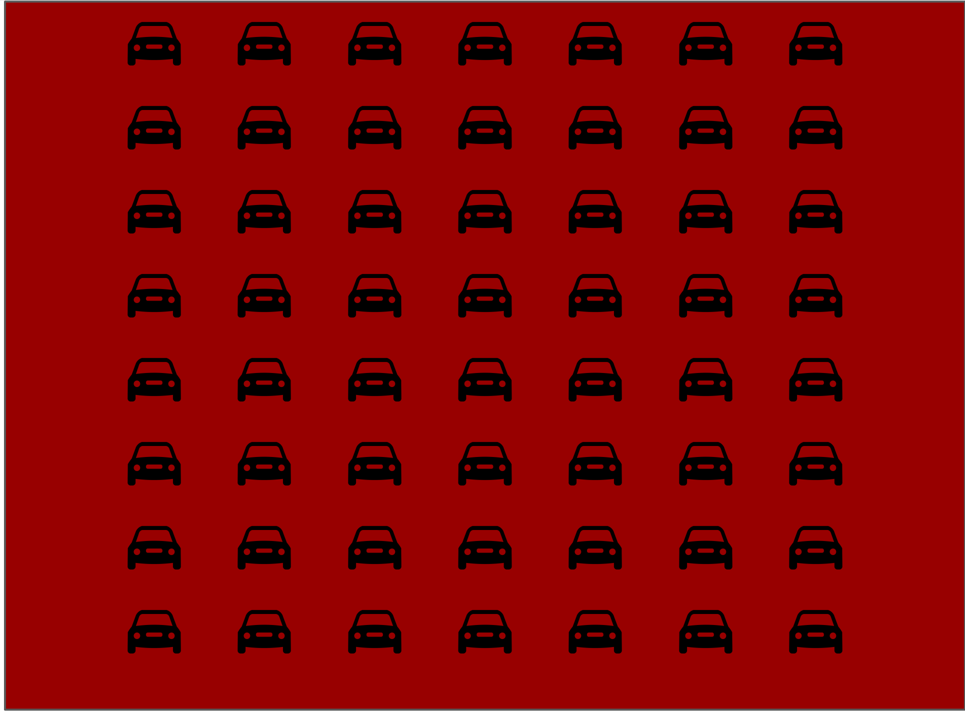
Eeshan Agarwal

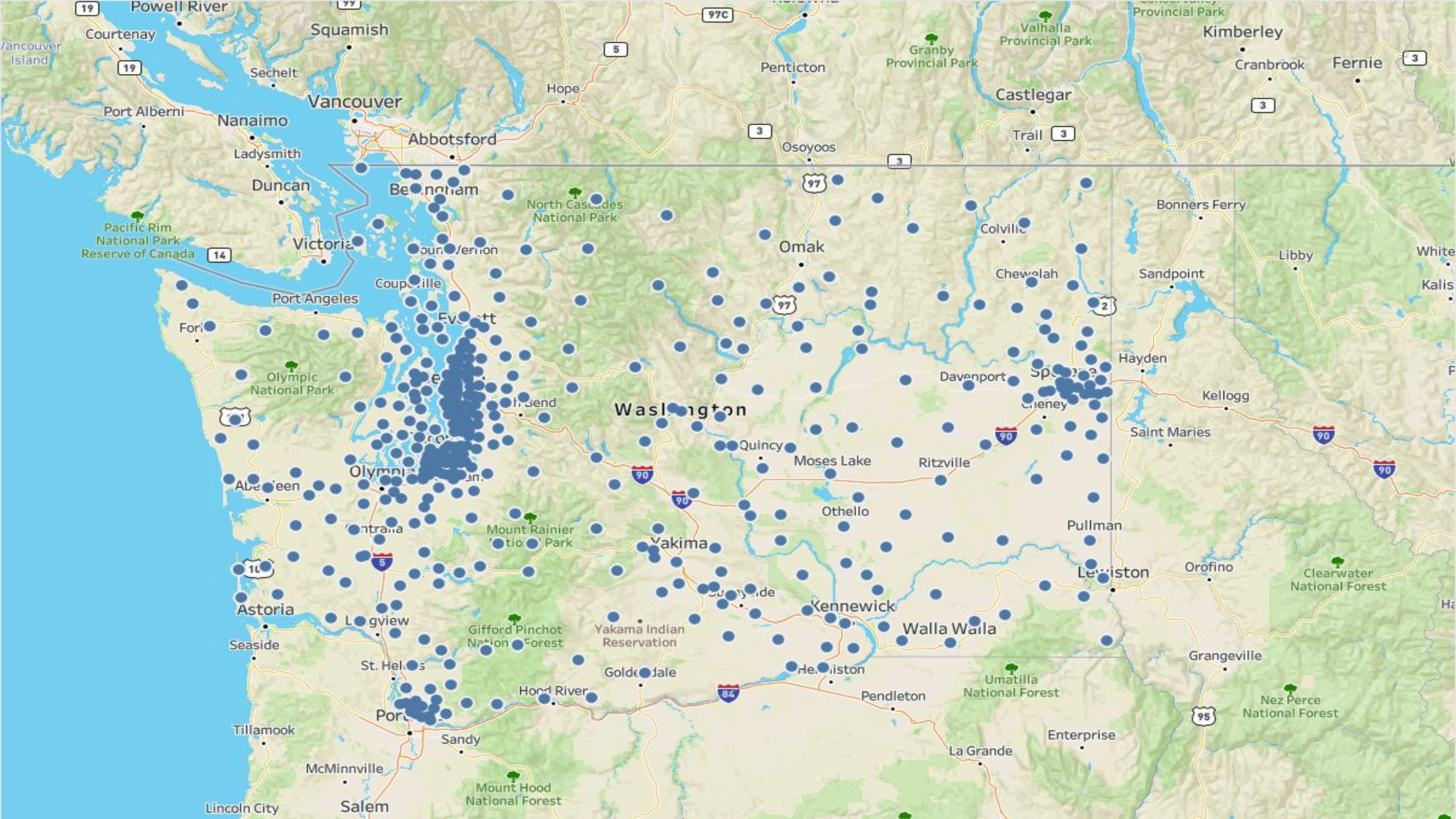
On average, 534 fatal crashes are reported in Washington State yearly

Data of more than 4000 cities, counties and zip codes was analyzed in this analysis for insights, patterns and trends.

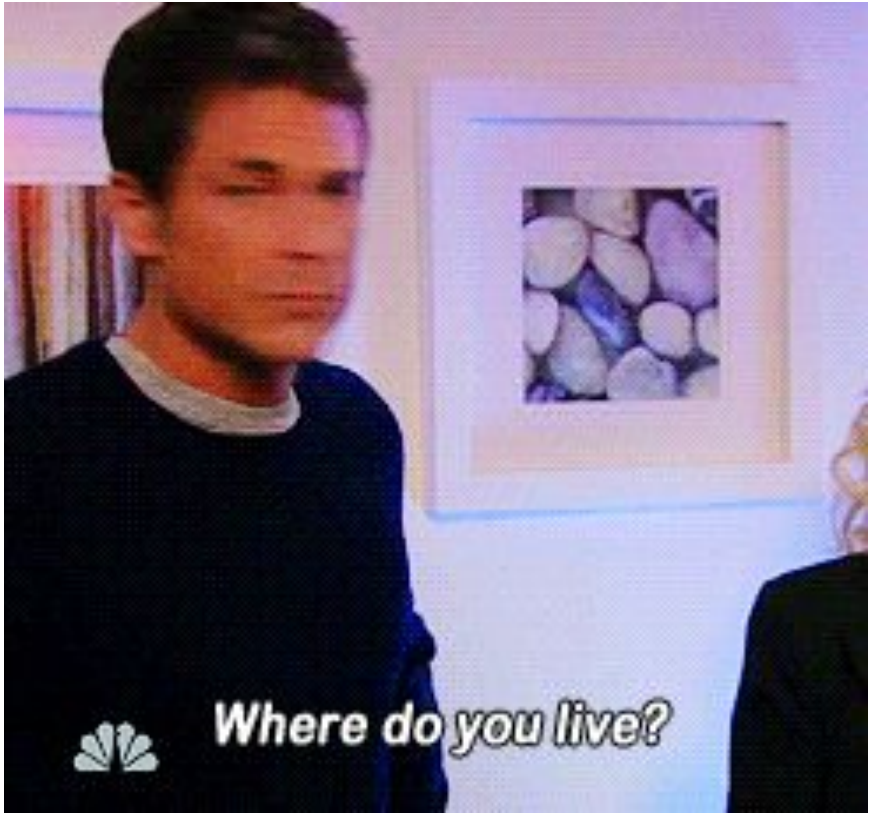
Meta data included 250+ variables for each data point (i.e., fatal crash incident)

The operative term “***fatal***” in this analysis means a crash which resulted in either death or serious injury




Powell River
Squamish
Courtenay
Vancouver Island
Sechelt
Vancouver
Nanaimo
Abbotsford
Hope
Port Alberni
Ladysmith
Duncan
Pacific Rim National Park Reserve of Canada
Victoria
Port Angeles
Bellingham
North Cascades National Park
Mount Vernon
Coupeville
Everett
Forks
Olympic National Park
Bend
Olympia
Aberdeen
Centralia
Mount Rainier National Park
Astoria
Longview
Seaside
St. Helens
Portland
Tillamook
Sandy
McMinnville
Lincoln City
Salem
Mount Hood National Forest
Gifford Pinchot National Forest
Hood River
Goldendale
Yakama Indian Reservation
Yakima
Sunnyside
Washington
Quincy
Moses Lake
Othello
Kennewick
Hermiston
Pendleton
Walla Walla
Umatilla National Forest
La Grande
Enterprise
Penticton
Osoyoos
Omak
Granby Provincial Park
Valhalla Provincial Park
Castlegar
Trail
Kimberley
Cranbrook
Fernie
Colville
Chewelah
Davenport
Spokane
Cheney
Ritzville
Pullman
Lewiston
Bonners Ferry
Sandpoint
Hayden
Kellogg
Saint Maries
Libby
Orofino
Clearwater National Forest
Grangeville
Nez Perce National Forest
19
5
97C
3
97
14
101
2
90
84
95

Among drivers involved in fatal crashes, what proportion are involved in crashes in communities where they live?



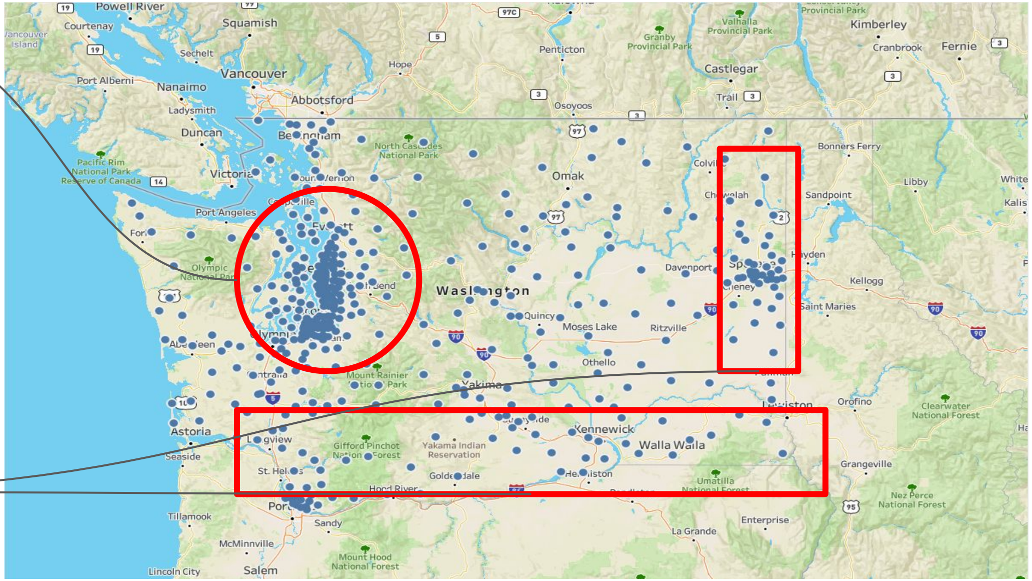
Among drivers involved in fatal crashes, what proportion are involved in crashes in communities where they live?

23%

Only 23% of the people who crash in a community belong to the very same community they crashed in. By "community", we assume zip code. MAPBOX API was used to convert x, y coordinates of crash location to match it to its specific zip code, which was then compared with the driver's zip code to calculate the proportion. This proves the hypothesis that an overwhelming majority of people involved in fatal crashes in a community are not resident of that area.

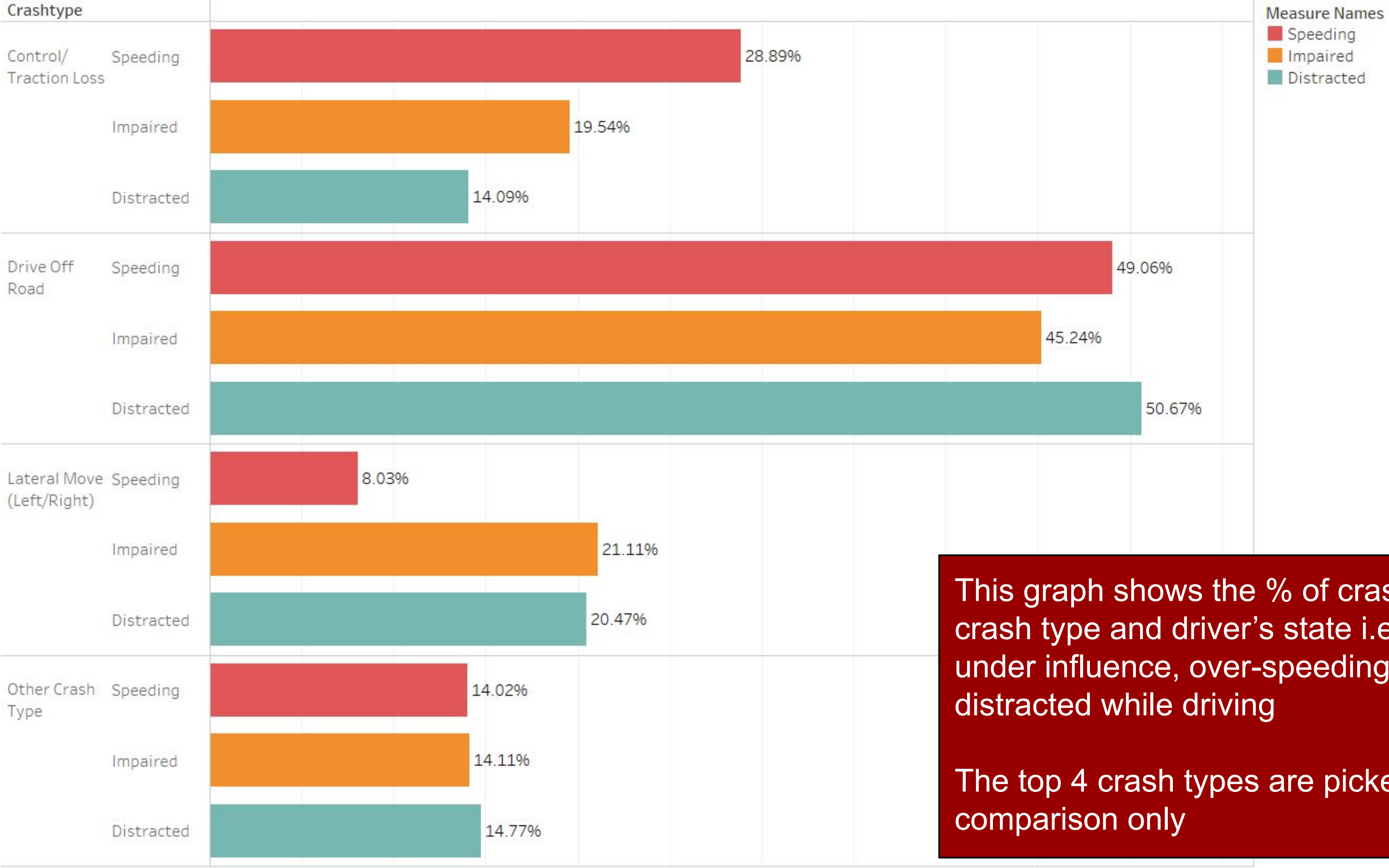
The % stays the same for border areas while it changes significantly for Seattle

In Seattle, the proportion for residents and non-residents is 18% and 82% respectively, a change of 5 percentage points



However in border areas the ratio is similar to the state average i.e., 23% resident, 77% non-residents

Are there differences in the types of crashes and behavior factors in those crashes among “residents” versus those deemed to be not “from” the area?



Distracted, Impaired and Speeding for each Crashtype. Color shows details about Distracted, Impaired and Speeding. The data is filtered on Is Resident and Crashtype Set. The Is Resident filter keeps False and True. The Crashtype Set filter keeps 4 members.

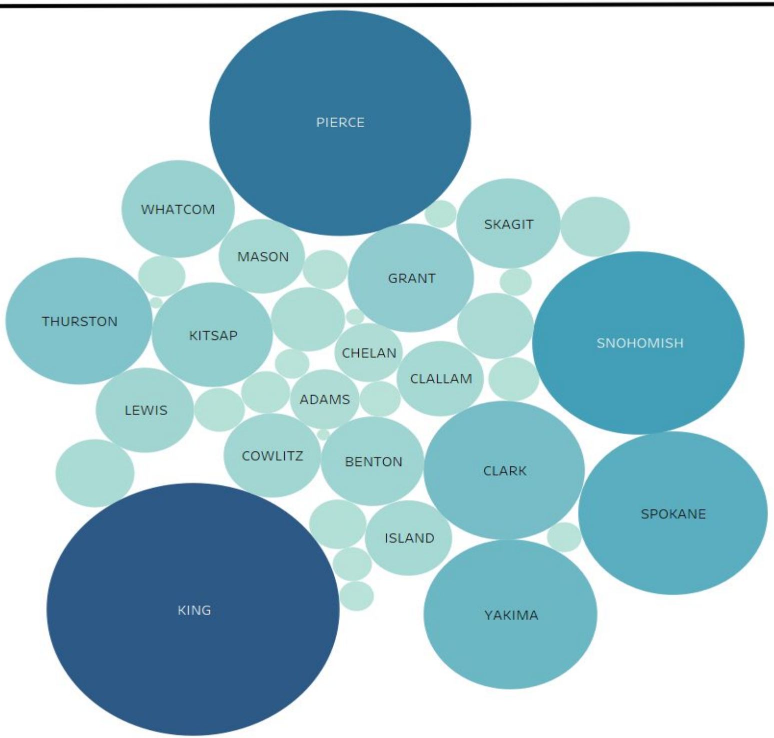
This graph shows the % of crashes by crash type and driver's state i.e., under influence, over-speeding and/or distracted while driving

The top 4 crash types are picked for comparison only

Are there specific resident ZIP Codes that tend to produce higher-risk drivers that are involved in fatal crashes at a higher rate?

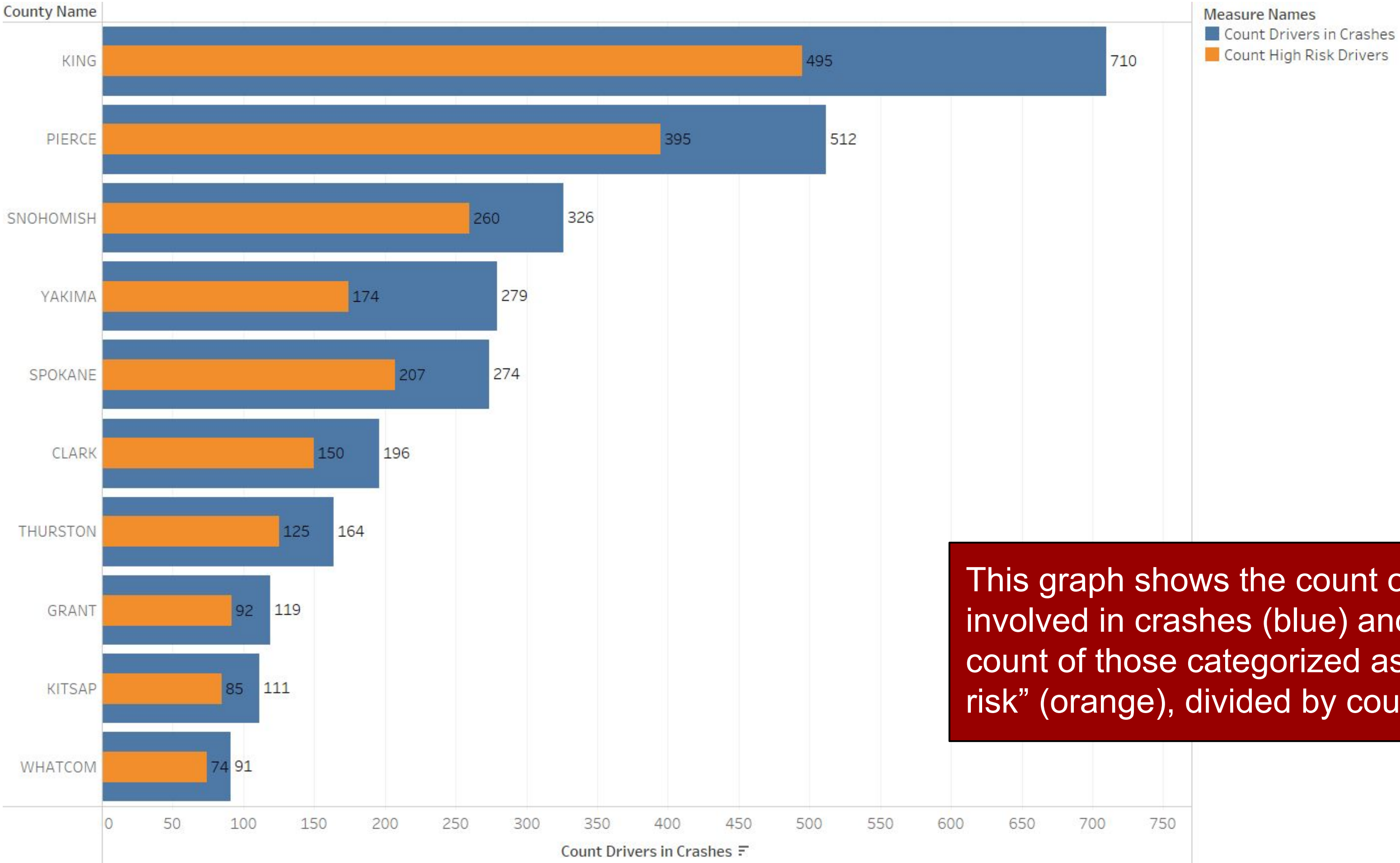
Certain zip codes, and hence the counties, stood out as the ones producing higher-risk drivers

We consider a driver as higher-risk if they are involved in an activity while driving which puts their as well as others' lives at risk such as drunk driving, over speeding etc.



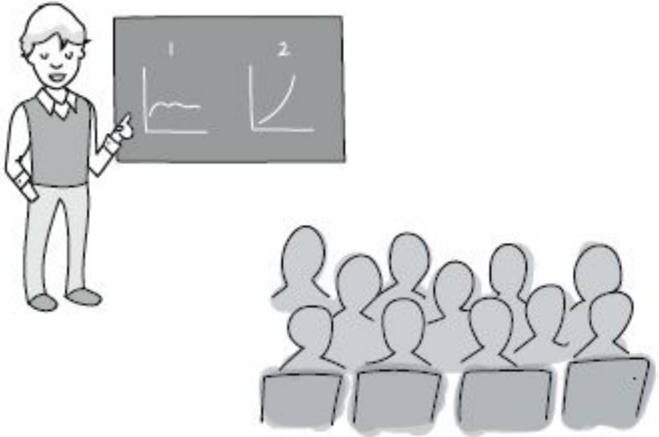
County Name. Color shows sum of Count High Risk Drivers. Size shows sum of Count High Risk Drivers. The marks are labeled by County Name.

Team IC23001
Washington Fatal Crash Files
UMD Information Challenge 2023



Count Drivers in Crashes and Count High Risk Drivers for each County Name. Color shows details about Count Drivers in Crashes and Count High Risk Drivers. The data is filtered on County Name Set, which keeps 10 members.

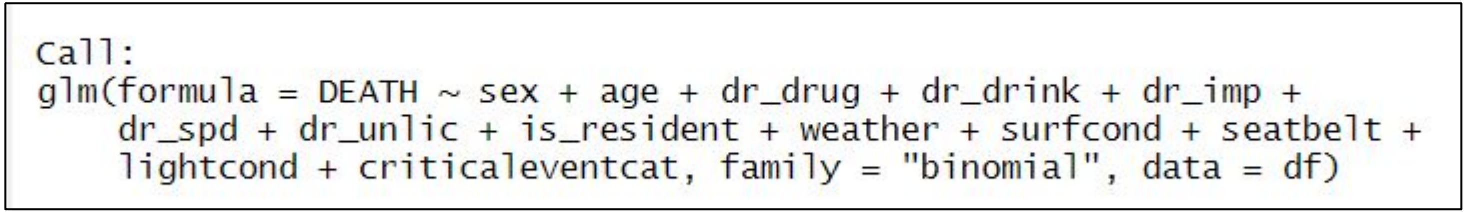
This graph shows the count of drivers involved in crashes (blue) and the count of those categorized as “high risk” (orange), divided by counties



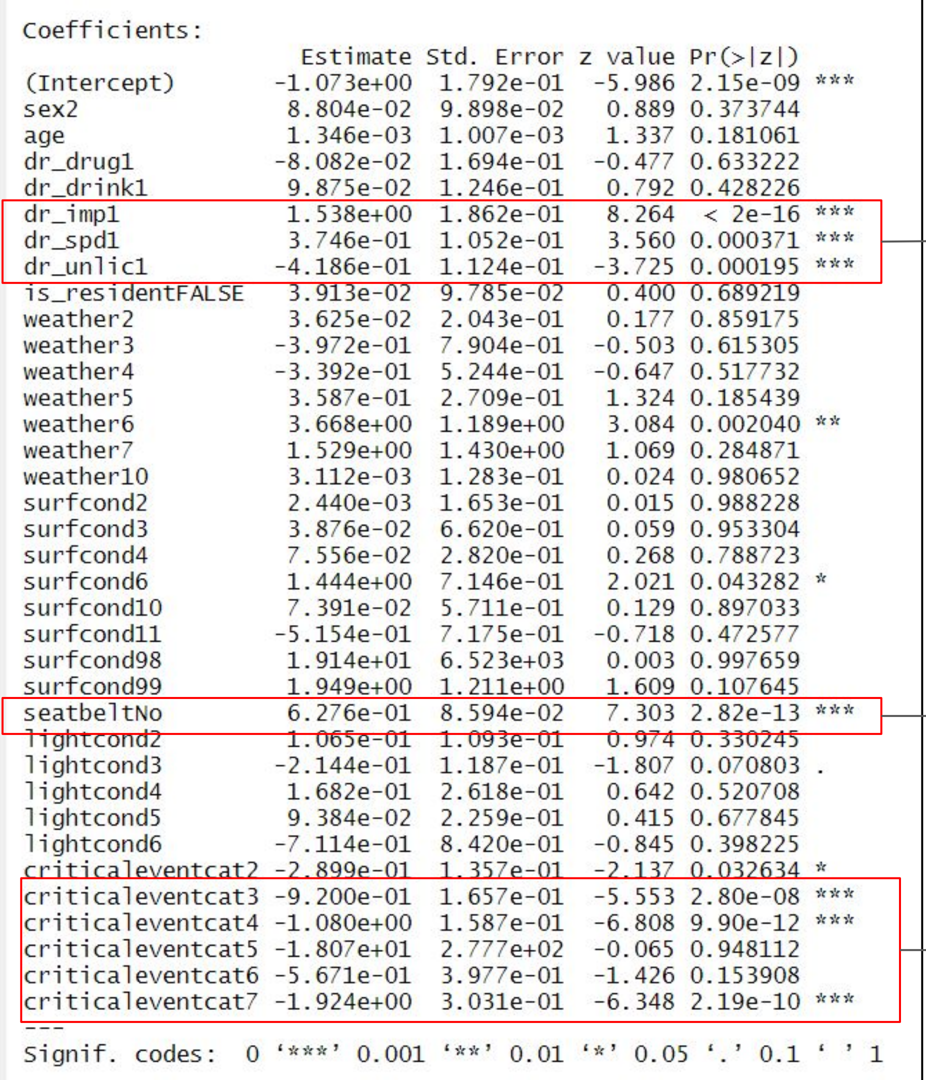
if you torture the data long enough,
it will tell you what you want

comic.fosslien.com

Certain behavior factors significantly increase the odds of death in case of an accident



```
Call:
glm(formula = DEATH ~ sex + age + dr_drug + dr_drink + dr_imp + 
    dr_spd + dr_unlic + is_resident + weather + surfcond + seatbelt + 
    lightcond + criticaleventcat, family = "binomial", data = df)
```



```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)    
(Intercept)       -1.073e+00  1.792e-01  -5.986 2.15e-09 ***
sex2               8.804e-02  9.898e-02   0.889 0.373744    
age                1.346e-03  1.007e-03   1.337 0.181061    
dr_drug1          -8.082e-02  1.694e-01  -0.477 0.633222    
dr_drink1          9.875e-02  1.246e-01   0.792 0.428226    
dr_imp1            1.538e+00  1.862e-01   8.264  < 2e-16 ***
dr_spd1            3.746e-01  1.052e-01   3.560 0.000371 ***
dr_unlic1         -4.186e-01  1.124e-01  -3.725 0.000195 ***
is_residentFALSE   3.913e-02  9.785e-02   0.400 0.689219    
weather2           3.625e-02  2.043e-01   0.177 0.859175    
weather3          -3.972e-01  7.904e-01  -0.503 0.615305    
weather4          -3.392e-01  5.244e-01  -0.647 0.517732    
weather5           3.587e-01  2.709e-01   1.324 0.185439    
weather6           3.668e+00  1.189e+00   3.084 0.002040 ** 
weather7           1.529e+00  1.430e+00   1.069 0.284871    
weather10          3.112e-03  1.283e-01   0.024 0.980652    
surfcond2          2.440e-03  1.653e-01   0.015 0.988228    
surfcond3          3.876e-02  6.620e-01   0.059 0.953304    
surfcond4          7.556e-02  2.820e-01   0.268 0.788723    
surfcond6          1.444e+00  7.146e-01   2.021 0.043282 *  
surfcond10         7.391e-02  5.711e-01   0.129 0.897033    
surfcond11        -5.154e-01  7.175e-01  -0.718 0.472577    
surfcond98         1.914e+01  6.523e+03   0.003 0.997659    
surfcond99         1.949e+00  1.211e+00   1.609 0.107645    
seatbeltNo         6.276e-01  8.594e-02   7.303 2.82e-13 ***
lightcond2         1.065e-01  1.093e-01   0.974 0.330245    
lightcond3        -2.144e-01  1.187e-01  -1.807 0.070803 .  
lightcond4         1.682e-01  2.618e-01   0.642 0.520708    
lightcond5         9.384e-02  2.259e-01   0.415 0.677845    
lightcond6        -7.114e-01  8.420e-01  -0.845 0.398225    
criticaleventcat2 -2.899e-01  1.357e-01  -2.137 0.032634 *  
criticaleventcat3 -9.200e-01  1.657e-01  -5.553 2.80e-08 ***
criticaleventcat4 -1.080e+00  1.587e-01  -6.808 9.90e-12 ***
criticaleventcat5 -1.807e+01  2.777e+02  -0.065 0.948112    
criticaleventcat6 -5.671e-01  3.977e-01  -1.426 0.153908    
criticaleventcat7 -1.924e+00  3.031e-01  -6.348 2.19e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The fact that whether driver is impaired, speeding, is unlicensed or not wearing a seatbelt has significant impact on the odds of a death in the situation of a crash

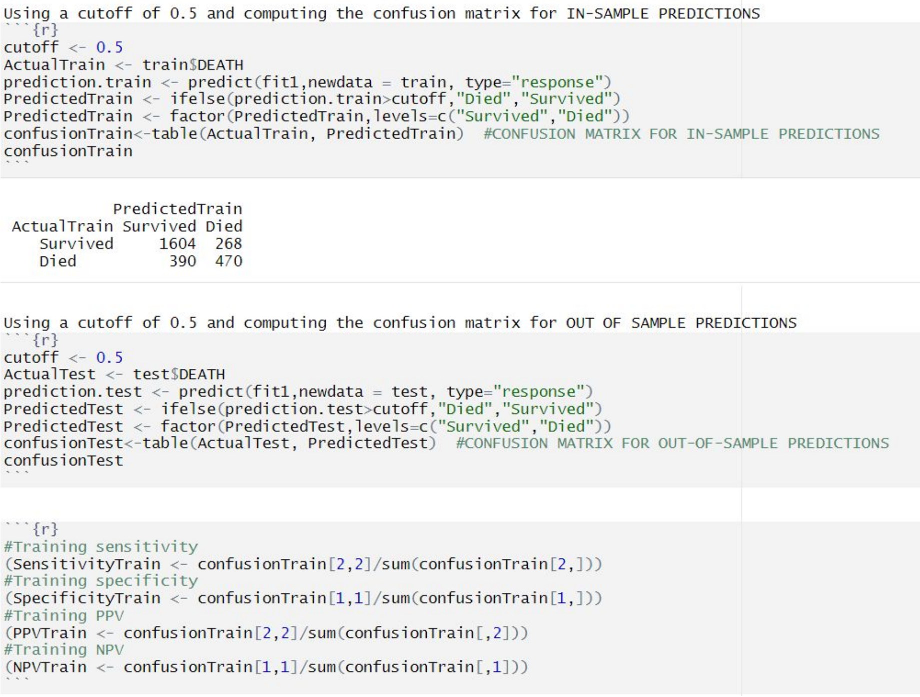
Not wearing a seatbelt increases the odds of death by a multiplicative factor of 1.87, given all else is held constant!

Statistically significant coefficients implying the odds of a death in a crash are relatively less if the critical event leading to the accident falls into either of three

Our prediction model has an an out of sample accuracy of 77%

We partition the data into train and test, and both datasets to predict probabilities. We use the actual and the predicted values to compute a confusion matrix, which we use to find out the accuracy and error rate.

In-sample accuracy is 75% and out-of-sample accuracy is 77%



Using a cutoff of 0.5 and computing the confusion matrix for IN-SAMPLE PREDICTIONS

~~~
```{r}
cutoff <- 0.5
ActualTrain <- train$DEATH
prediction.train <- predict(fit1,newdata = train, type="response")
PredictedTrain <- ifelse(prediction.train>cutoff,"Died","Survived")
PredictedTrain <- factor(PredictedTrain,levels=c("Survived","Died"))
confusionTrain<-table(ActualTrain, PredictedTrain)  #CONFUSION MATRIX FOR IN-SAMPLE PREDICTIONS
confusionTrain
```
~~~

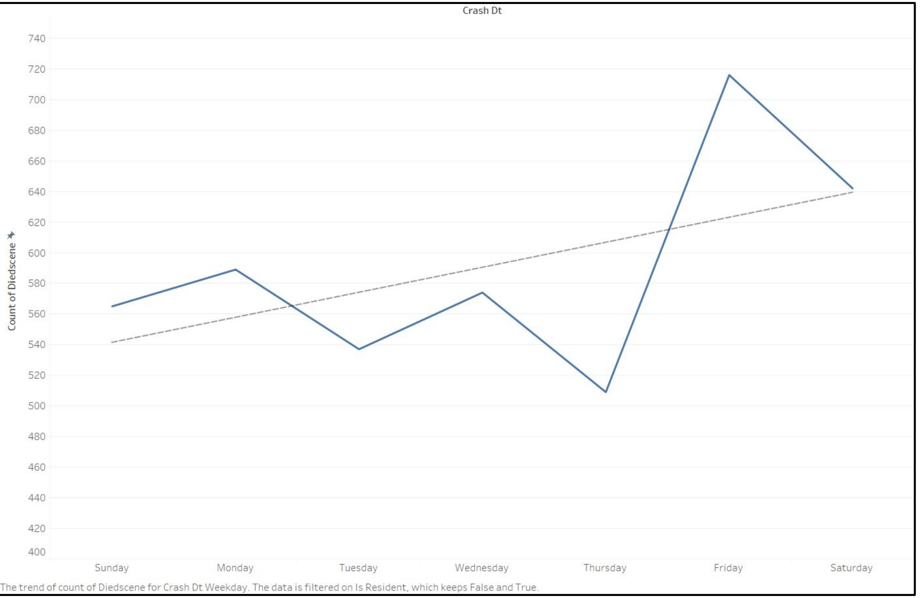
```
             PredictedTrain
 ActualTrain Survived Died
    Survived     1604  268
    Died          390  470
```

Using a cutoff of 0.5 and computing the confusion matrix for OUT OF SAMPLE PREDICTIONS

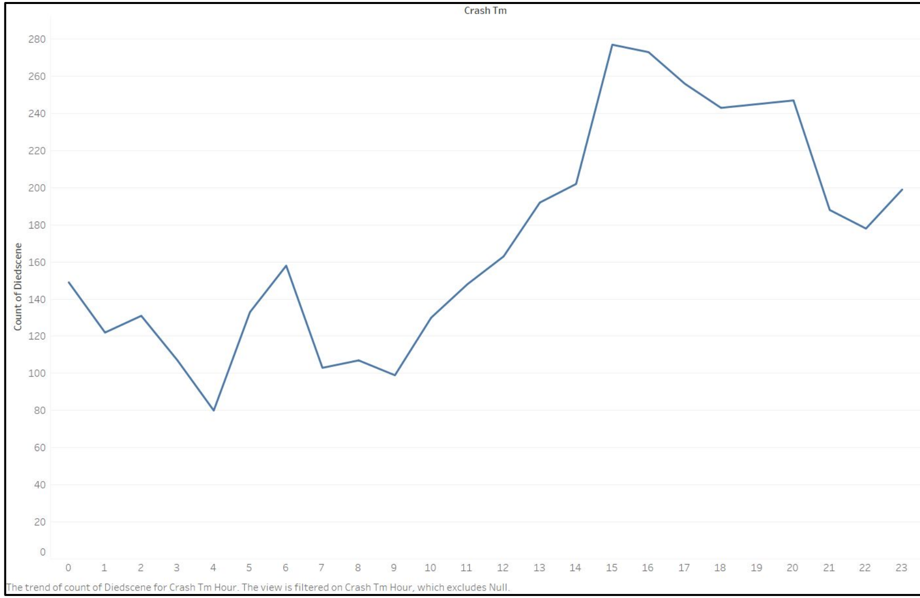
~~~
```{r}
cutoff <- 0.5
ActualTest <- test$DEATH
prediction.test <- predict(fit1,newdata = test, type="response")
PredictedTest <- ifelse(prediction.test>cutoff,"Died","Survived")
PredictedTest <- factor(PredictedTest,levels=c("Survived","Died"))
confusionTest<-table(ActualTest, PredictedTest)  #CONFUSION MATRIX FOR OUT-OF-SAMPLE PREDICTIONS
confusionTest
```
~~~

~~~
```{r}
#Training sensitivity
(SensitivityTrain <- confusionTrain[2,2]/sum(confusionTrain[2,]))
#Training specificity
(SpecificityTrain <- confusionTrain[1,1]/sum(confusionTrain[1,]))
#Training PPV
(PPVTrain <- confusionTrain[2,2]/sum(confusionTrain[,2]))
#Training NPV
(NPVTrain <- confusionTrain[1,1]/sum(confusionTrain[,1]))
```
~~~

Time has an interesting correlation (which, of course, does not imply..causation!)



Frequency of crash by week of the day



Frequency of crash in a 24 hours day

There is a stark difference in the locations where people live and the locations where they get into accidents

We conclude that although there are more non-residents who are involved in accidents in communities where they don’t belong to, the factors contributing to the crash are somewhat similar in both the classes albeit certain differences. Hence, an inclusive rather than targeted approach is the need of the hour.

More data is required for further analysis to ascertain causation for certain response variables.

Thank you for listening to our presentation

Let us know if you have any questions. Thank you!

Acknowledgement

We used the following datasets and APIs in addition to the data already provided:

1. https://www.census.gov/data.html
2. https://www.mapbox.com/
3. https://openai.com/blog/chatgpt
4.