Data Analysis of Fatal Crashes

Data From Washington Crash Survey

Team IC23031

Questions

Demographics

What are the demographics of ZIP codes with high risk drivers

Proportionality

What percentage of fatal crashes occur locally to the driver's home

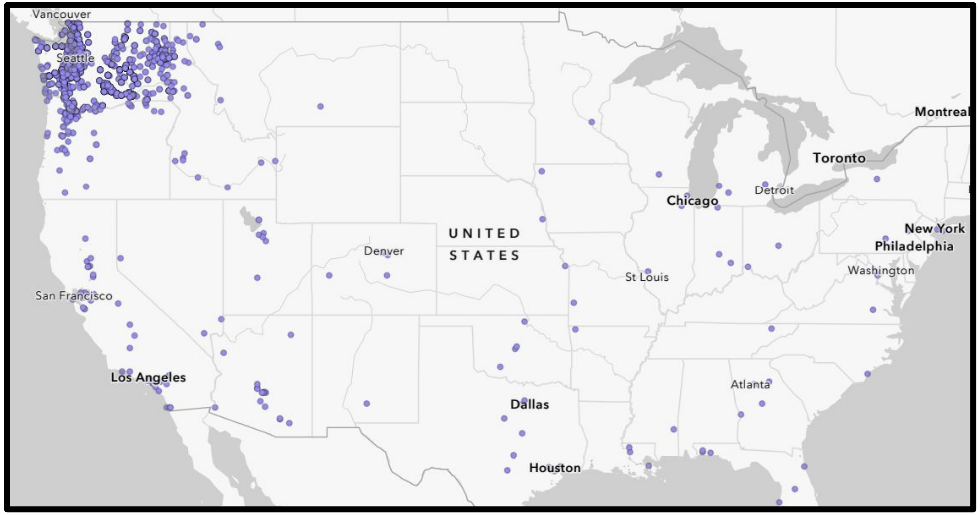
Risk-Zones

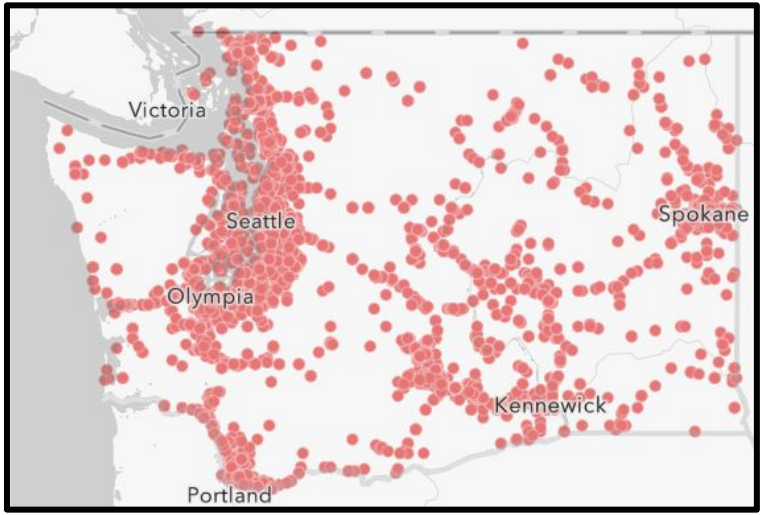
Which ZIP codes produce the highest risk drivers

Behavior

What are the behavioral differences between people who crashed locally vs non locally

Data Exploration





- Driver Location given in residency ZIP code
  - Scattered across US
- Accident Address given in geolocation coordinates
  - Clustered around Seattle

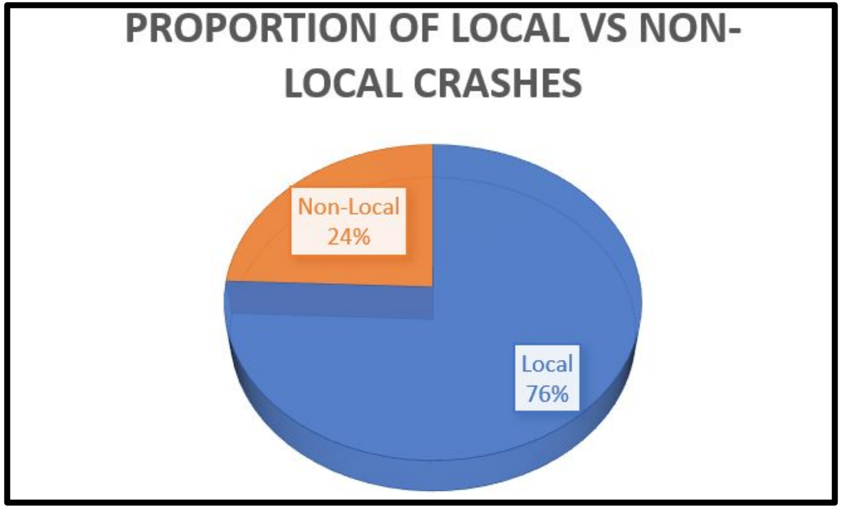
Data Cleaning and Preparation

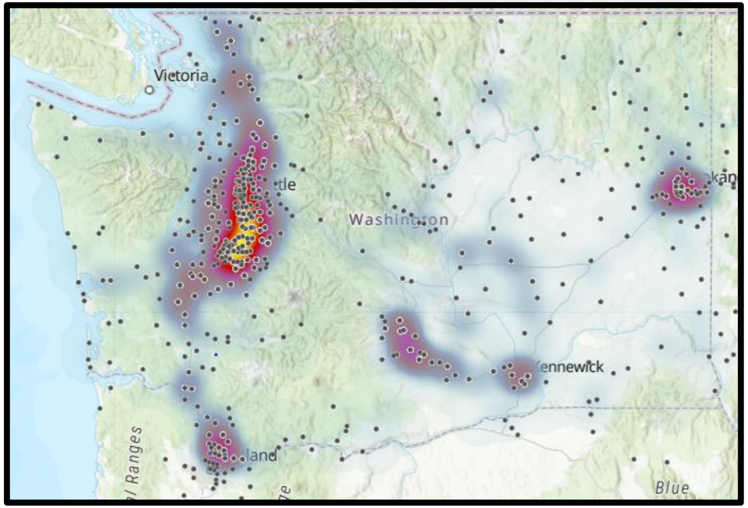
- Used Fatal Crash Survey Data and removed extraneous data values
    - Weather,light,route, etc.
- Converted driver ZIP code to geolocation using pgeocode
- Compared driver geolocation to accident geolocation and got distance using Haversine formula
- Identified whether accident was local to driver based on if location was with 45 kilometers

```
import pandas as pd
from arcgis.geocoding import reverse_geocode
from arcgis.geometry import Geometry
from arcgis.geocoding import geocode
from arcgis.gis import GIS
import pandas as pd
import pgeocode
```

- Python libraries used for geocoding

What Percentage of Fatal Crashes Occur Locally to the Driver's Home?

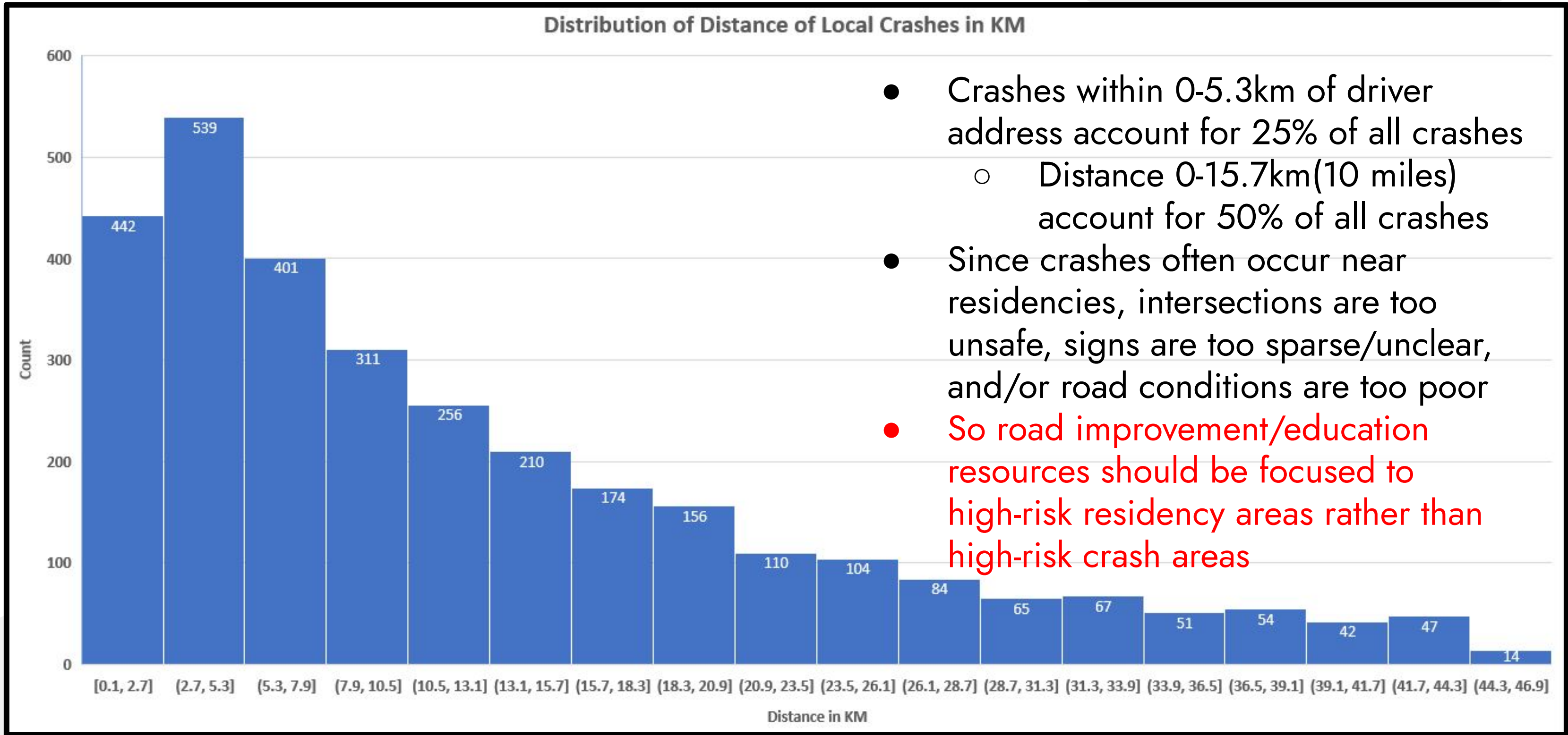




- So 76% of accidents occur within 45km of driver's home
- But how does this proportion change as distance changes?

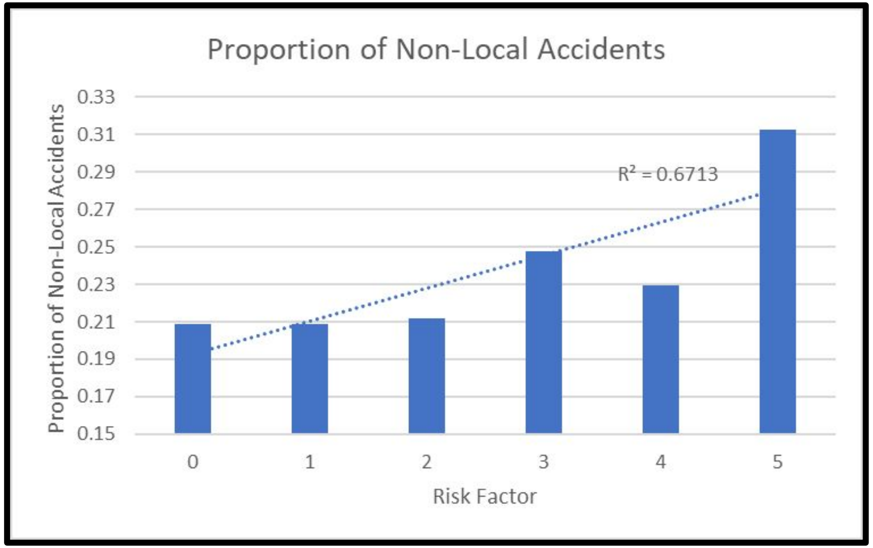
- Driver address noted as black dot, accident location noted as heatmap
- Clusters of driver address largely align with hotspots of accident location

What Percentage of Fatal Crashes Occur Locally to the Driver's Home?



- Crashes within 0-5.3km of driver address account for 25% of all crashes
  - Distance 0-15.7km(10 miles) account for 50% of all crashes
- Since crashes often occur near residencies, intersections are too unsafe, signs are too sparse/unclear, and/or road conditions are too poor
- So road improvement/education resources should be focused to high-risk residency areas rather than high-risk crash areas

What are the behavioral differences between people who crashed locally vs non locally



- Risk factor includes drug use, alcohol use, speeding, drowsiness, and distracted driving equally
- Proportion of non-local accidents *generally* are correlated with higher risk factors
- But do all risk factors have the same impact on risk?

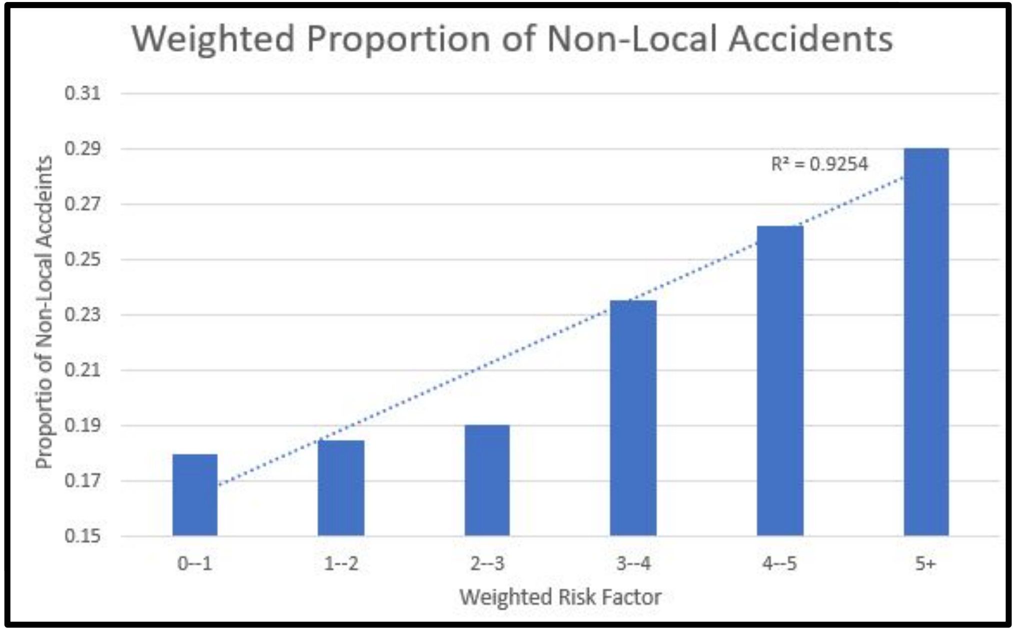
What are the behavioral differences between people who crashed locally vs non locally

Table 4. Crash severity estimation results

Variable	Ordered		Generalized Ordered	
	Coefficient	t	Coefficient	t
Characteristics of the driver				
Driver's gender				
Female*	0.491	13.56	0.503	13.71
Driver's age				
16–24	−0.293	−6.51	−0.295	−6.62
60 above*	0.115	2.16	0.096	1.75
Driver's Impairment				
Under the Influence of Alcohol, Drugs*	1.157	11.95	1.003	10.94
Asleep or Fatigue	1.094	7.55	1.039	7.73
Safety Equipment				
Not Using Seatbelts*	1.184	16.21	1.054	14.89
Wrong Use of Equipment	0.942	4.4	0.793	4.07
Driver's Distraction				
Cognitive	−0.421	−4.93	−0.419	−4.93
Passenger Related	−0.780	−2.23	−0.745	−2.17
In-Vehicle Tasks	0.304	2.11	0.290	2.06
Out-Vehicle	−0.483	−2.14	−0.480	−2.15
Cellphone	0.386	2.11	0.397	2.21
Speeding				
Driving Over the Speed Limit*	0.203	3.81	0.173	3.25

- Obtained risk factor data from Cogent Engineering
- Speeding is the most impactful risk factor, drowsiness the least impactful

What are the behavioral differences between people who crashed locally vs non locally



- After applying a weight to risk factors with speeding having the highest weight, positive correlation between risk factor and proportion of Non-Local Crashes increased *dramatically*
- So more speed measures traffic police. cameras and speed-bumps should be implemented to decrease speeding rates in residential areas

Which ZIP codes produce the highest risk drivers

zipcode.py > ...

```python
import pandas as pd
#find highest risk zip codes
df = pd.read_csv('fullcrashdata.csv')
zip_counts = df[df['dzip'] != 99999]['dzip'].value_counts()
sorted_zip_counts = zip_counts.sort_values(ascending=False)
#make new df with zip codes and counts
common = sorted_zip_counts
commonsdf = pd.DataFrame({'dzip': common.index, 'crash_count': common.values})
commonsdf.to_csv('commoncounts.csv' ,index=False)
#print 5 most fatal zips
print(sorted_zip_counts.head(5))
```

01

99301

47 crashes

Pasco, WA

02

98444

43 crashes

Tacoma, WA

03

98387

35 crashes

Pierce, WA

04

98837

32 crashes

Grant, WA

05

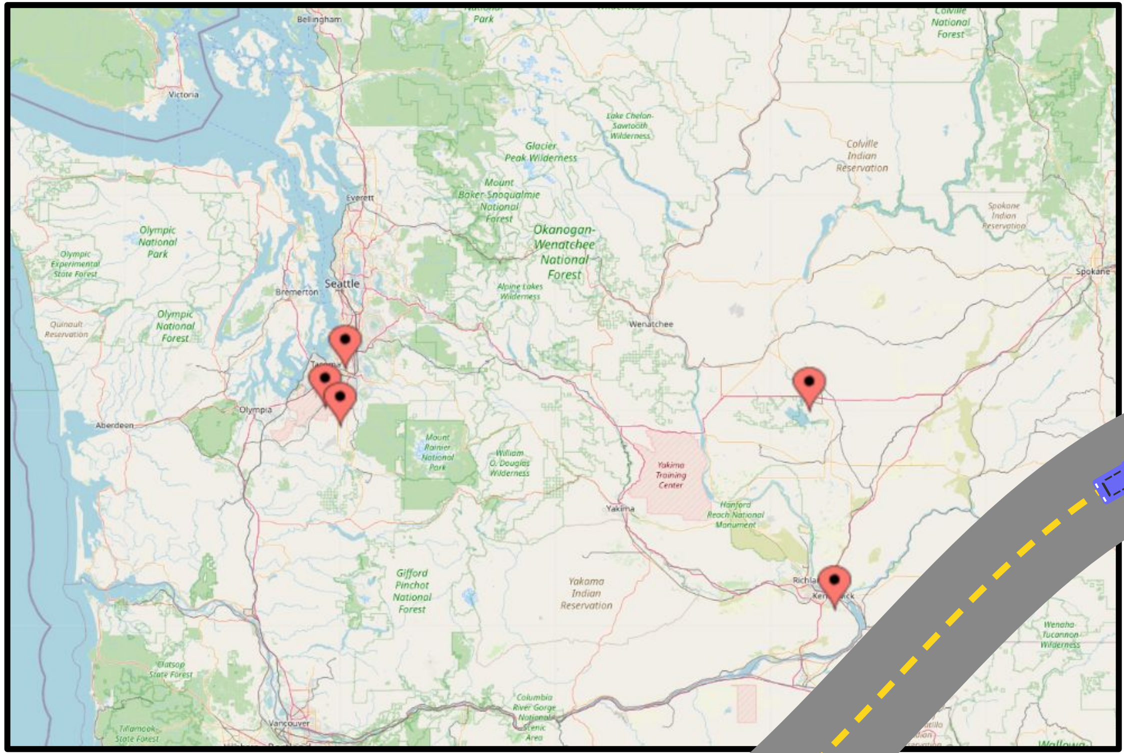
98951

32 crashes

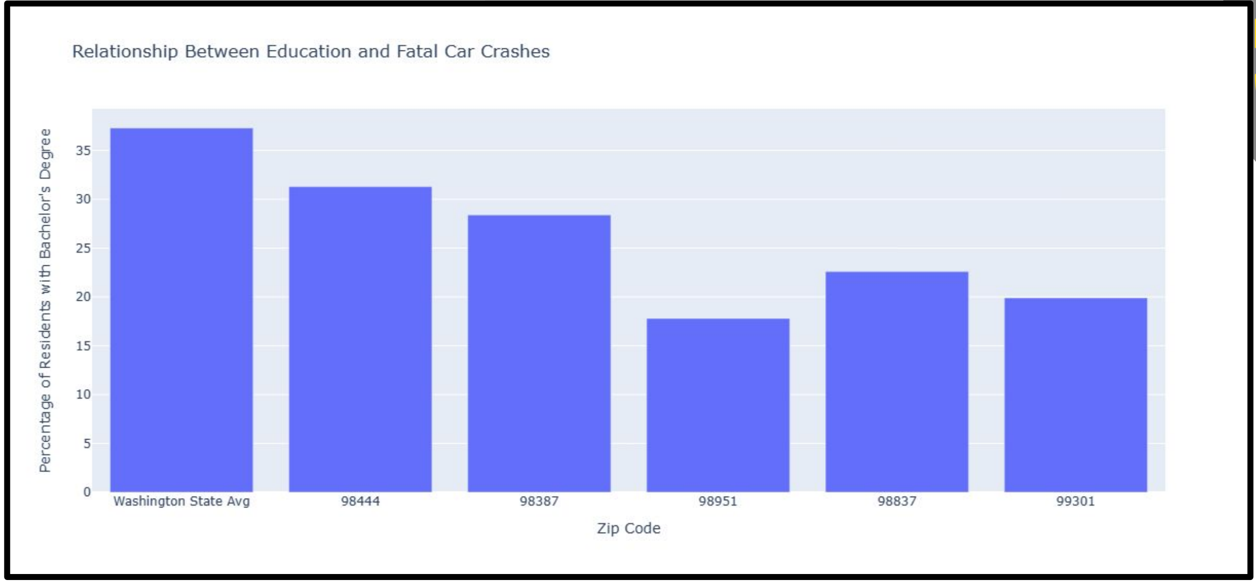
Yakima, WA

Which ZIP codes produce the highest risk drivers

- The demographics of the ZIP codes needed to be investigated to look for commonalities



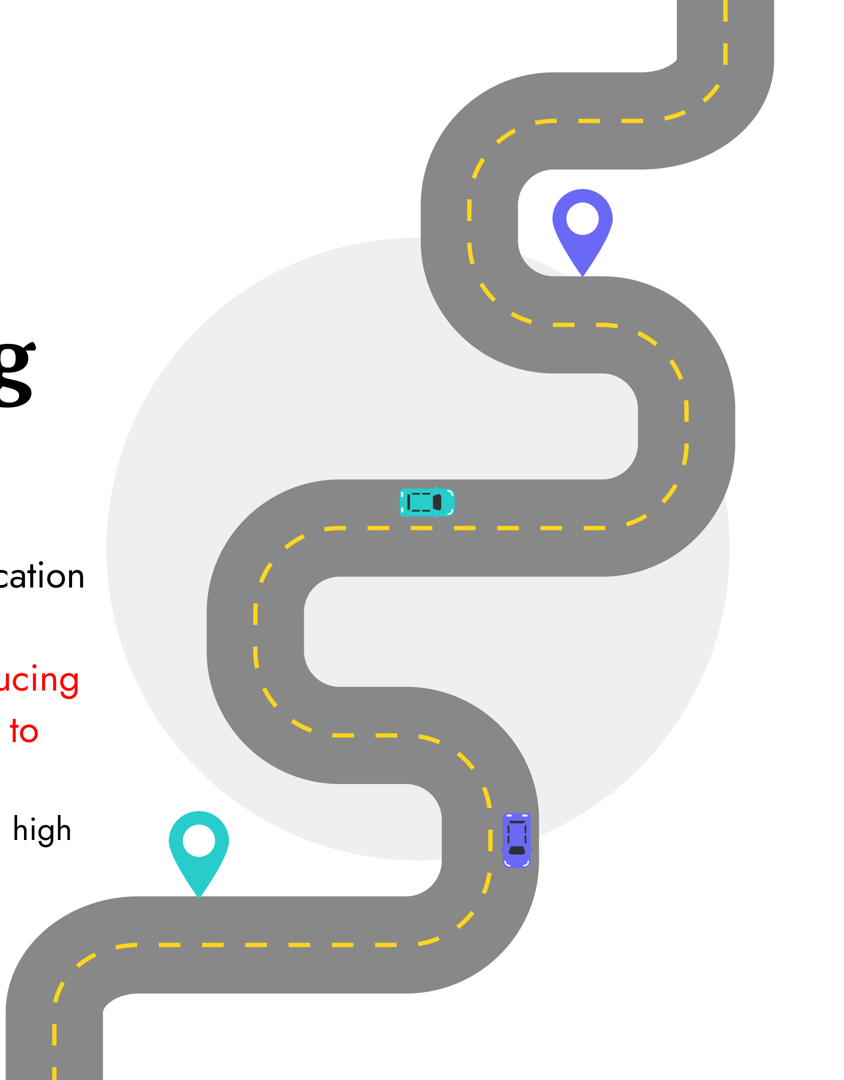
What are the demographics of ZIP codes with high risk drivers



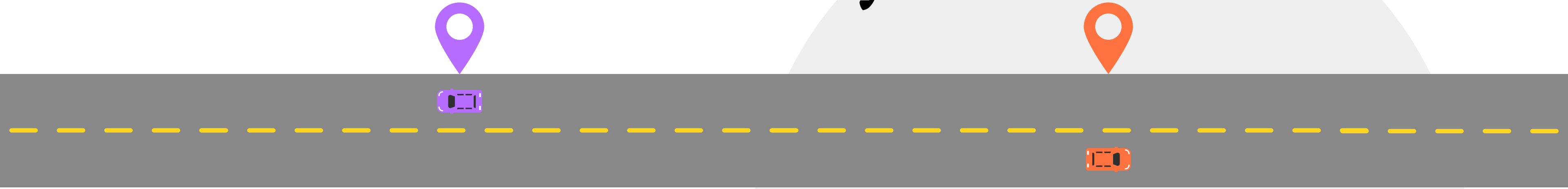
- ZIP codes producing the most fatal drivers tend to have a lower percentage of residents who received a Bachelor's Degree
    - Percentages were compared with Washington state average

High risk ZIP codes: Improved Safe Driving Education for teens

- Increased government subsidized & mandated education on safe driving during driver’s education
- Recommend to high schools within ZIP codes producing high proportions of drivers involved in fatal crashes to revise driving curriculum
  - As relatively low numbers of residents attend college, high school is their final stop in the education process



Thank you



Questions?