



AMPURE
CAPITAL
Car Insurance Data
Analysis

Presented by David Burns
12/6/2022

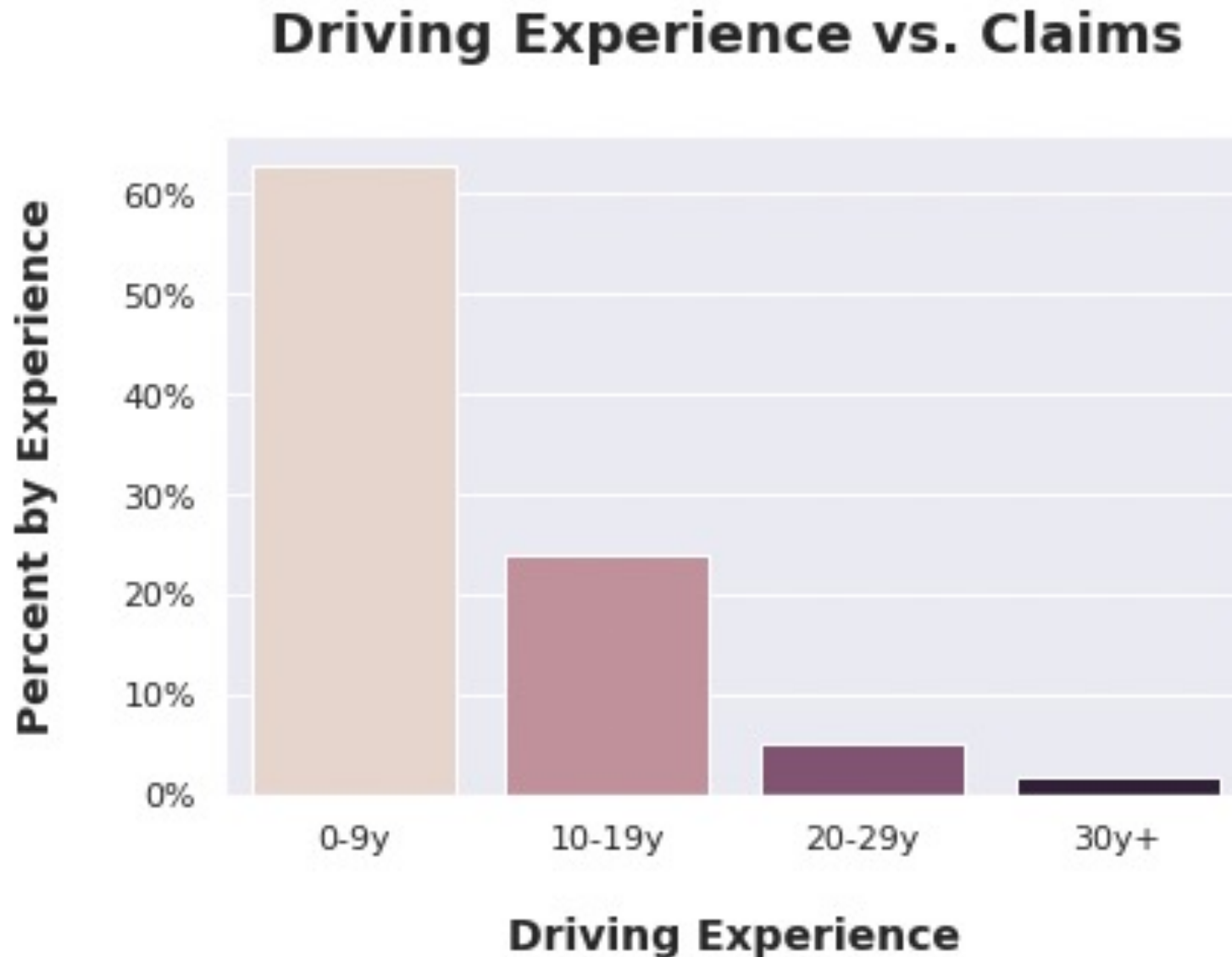
Data Set Analyzed

Data Set

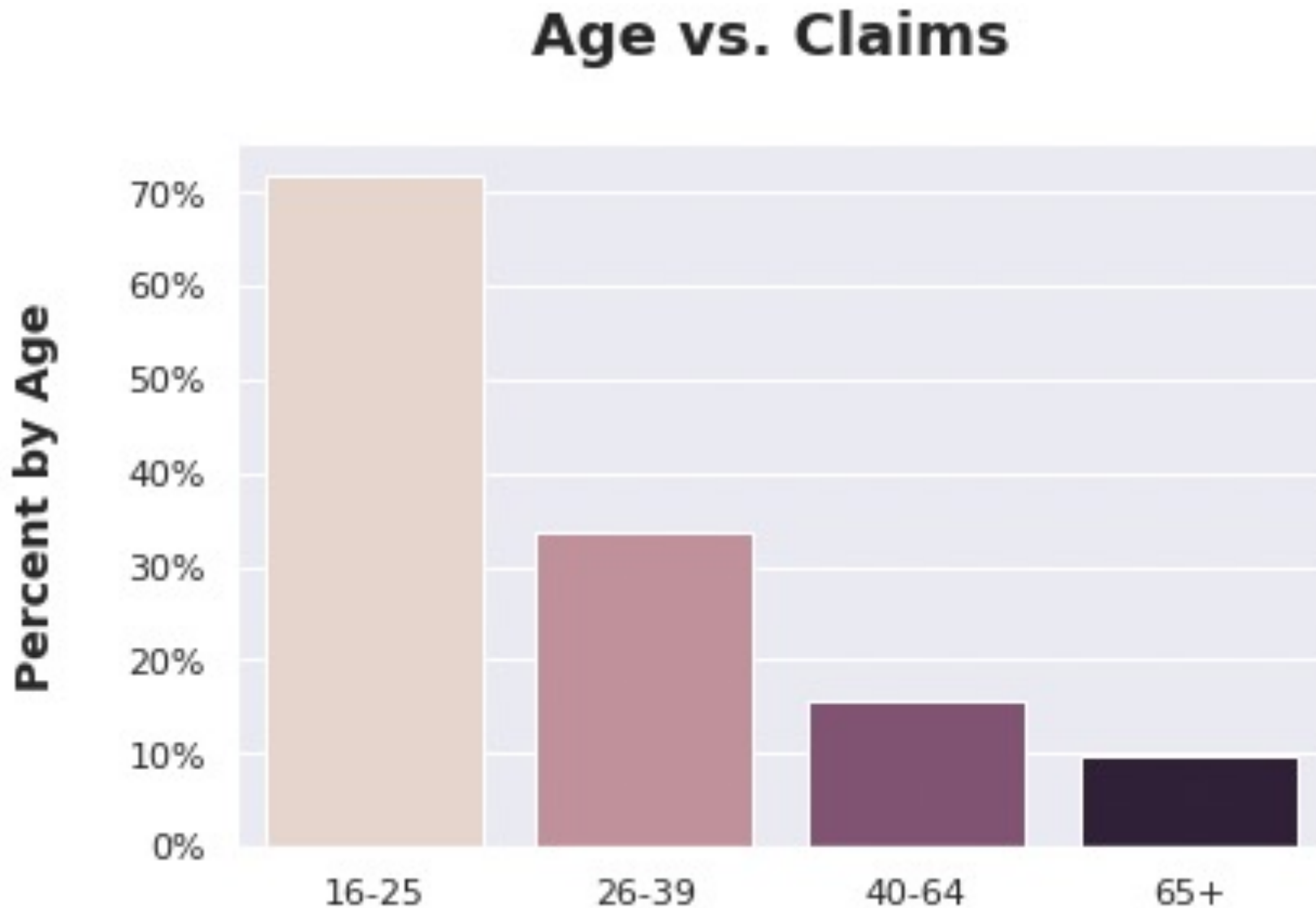
- 10,000 samples
- 18 columns of variables that could impact claims
 - Age of driver
 - Driving experience
 - Type of car
 - Tickets, accidents, DUIs, Etc.
- 1 column of 'outcome,' or claim data (0 or 1)

The original source of the data is Kaggle

Lack of driving Experience Biggest Risk



Age is Biggest Driver of Claims



Analysis - Predictive Model

- Binary classification problem – will or won't there be a claim
- Focused on reducing false negatives
 - False negatives are more expensive than missed revenue
- Model accuracy was 85% with a 15% false negative rate

Analysis - Clustering

Two primary clusters

- Young people with lots of claims
- Older people with many fewer claims

Including race duplicated these clusters

Interesting “unclustered” were older, unmarried, mostly men with lots of speeding tickets and DUIs who drove sports cars!

Recommendations

Speeding tickets, DUIs, claims, etc. were lifetime total

- Older people had more speeding tickets, etc.
- Fewer claims

Recommend dropping claims, tickets, and DUIs off the record over time

Differentiate younger people

- Include school grades, drivers' education, etc. which indicate