Background Information: Road accidents always rank among the leading cause of death in India. One of the major reasons for this is easy access to vehicles without needing a license. Additionally, the efficiency of the government in implementing strict licensing vary greatly from state to state. Therefore, I wanted to investigate which states are at high risk and profile them into clusters to identify a pattern.

Datasets:
https://www.data.gov.in/files/ogdpv2dms/s3fs-public/RA2021_A7.csv
https://www.data.gov.in/files/ogdpv2dms/s3fs-public/RA2021_A36.csv

Focus Question/Problem Statement: How can the Indian states be clustered based on their outcome of road accidents (fatal, grievous, non-grievous) and their rate of licensed driver accidents vs unlicensed driver accidents.

First, I will examine these datasets by the naked eye to spot any NaN values as they are both small datasets. Then I plan to combine these 2 datasets based on their state columns after standardizing the state column name. When I combined them, I had gotten duplicate state values as some state entries had hanging spaces. I removed those spaces and combined them again to get my cleaned dataset.

Hypothesis: States that are geographically close together will be clustered together. Clusters with a higher proportion of licensed accidents will have lesser proportion of severe outcomes as the drivers are more skilled in evading the worst outcome.

Methods: My plan for creating this model will be to first standardize the data to avoid overfitting as some states have a huge number of accidents while some have very less due to factors such as population and size. Then I will use the K-means algorithm multiple times using different k-values to find the best k-value. I choose the k-means algorithm as it is suited to work with numerical data while also being very simple and easy to understand. It allows me to easily generate many models and find the best one using metrics such as elbow plots of the inertia scores or the silhouette scores.

I chose these metrics because they are easy to interpret visually when plotted on an x-y graph. By examining the graphs, I determined that a K-value of 4 is optimal. Although the inertia score graph suggests a K-value of 3, analyzing both graphs together reveals that 4 represents a middle ground, where both metrics form an elbow or start decreasing at a slower rate, showing that the cluster points are close together while the clusters themselves are separated from each other.

Cluster	Fatal Accidents	Grievous Injury Accidents	Minor Injury Accidents	Non-Injury Accidents	Total Accidents_x	Valid Permanent License - Number	Learner's Licence	Without Licence	Others(Not known)	Total Accidents_y
0	1471.320	974.160	597.72	210.440	3253.640	1912.8	209.6	318.760	812.48	3253.640
1	12776.500	12493.000	24214.00	2796.000	52279.500	38186.5	2245.0	5406.500	6441.50	52279.500
2	7600.125	8180.625	6243.75	1575.875	23600.375	18346.5	642.0	1692.125	2919.75	23600.375
3	19026.000	11609.000	6312.00	782.000	37729.000	19296.0	4318.0	4863.000	9252.00	37729.000

Cluster 0: In this cluster about 75.6% of the accidents were severe (fatal + grievous injuries) while 58.8% of the drivers involved in the accidents had a valid license. High proportion of severe outcomes with low number of licenses.

Cluster 1: In this cluster about of the 48.3% of the accidents were severe (fatal + grievous injuries) while 73% of the drivers involved in the accidents had a valid license. Less proportion of severe outcomes with high number of licenses.

Cluster 2: In this cluster about 66.9%, of the accidents were severe (fatal + grievous injuries) while 77.7% of the drivers involved in the accidents had a valid license. Moderately high proportion of severe outcomes with high number of licenses.

Cluster 3: In this cluster about 81.2%, of the accidents were severe (fatal + grievous injuries) while 51.1% of the drivers involved in the accidents had a valid license. High proportion of severe outcomes with low number of licenses.

	States/UTs	Cluster
0	Andaman and Nicobar Islands	0
33	Tripura	0
29	Sikkim	0
27	Punjab	0
26	Puducherry	0
25	Orissa	0
24	Nagaland	0
23	Mizoram	0
22	Meghalaya	0
21	Manipur	0
18	Lakshadweep	0
35	Uttarakhand	0
14	Jharkhand	0
17	Ladakh	0
12	Himachal Pradesh	0
2	Arunachal Pradesh	0
3	Assam	0
4	Bihar	0
5	Chandigarh	0
13	Jammu and Kashmir	0
7	Dadra and Nagar Haveli	0
8	Delhi	0
36	West Bengal	0
11	Haryana	0
9	Goa	0
19	Madhya Pradesh	1
30	Tamil Nadu	1
20	Maharashtra	2
10	Gujarat	2
16	Kerala	2
28	Rajasthan	2
15	Karnataka	2
31	Telangana	2
1	Andhra Pradesh	2
6	Chhattisgarh	2
34	Uttar Pradesh	3

Clusters 0 and 3 are at the greatest danger and require urgent. Cluster 2 however exhibits a high proportion of fatal accidents despite a high proportion of licensed drivers involved in the accidents. Cluster 1 is at the least risk with a high number of licensed drivers involved in the accidents leading to less severe outcomes.

The hypothesis that geographically proximate locations will be clustered together was proven false. There is an element of randomness to how severe accidents are and how efficiently licenses are distributed.

Links used:
https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.select_dtypes.html
https://vitalflux.com/elbow-method-silhouette-score-which-better/
https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/
https://scikit-learn.org/1.5/auto_examples/text/plot_document_clustering.html#sphx-glr-auto-examples-text-plot-document-clustering-py