MCIS6273 Data Mining (Prof. Maull) / Fall 2022 / HW2

Points Possible	Due Date	Time Commitment (estimated)
20	Monday, Oct 24 @ Midnight	*up to* 24 hours

- **GRADING:** Grading will be aligned with the completeness of the objectives.
- **INDEPENDENT WORK:** Copying, cheating, plagiarism and academic dishonesty *are not tolerated* by University or course policy. Please see the syllabus for the full departmental and University statement on the academic code of honor.

OBJECTIVES

- Learn more about the analysis of the FARS database.
- Recreate data analysis in a published paper.
- Perform cluster analysis on NHTSA FARS data.
- Resubmit your PCA analysis from HW1.

WHAT TO TURN IN

You are being encouraged to turn the assignment in using the provided Jupyter Notebook. To do so, make a directory in your Lab environment called `homework/hw2`. Put all of your files in that directory. Then zip that directory, rename it with your name as the first part of the filename (e.g. `maull_hw2_files.zip`), then download it to your local machine, then upload the `.zip` to Blackboard.

If you do not know how to do this, please ask, or visit one of the many tutorials out there on the basics of using zip in Linux.

If you choose not to use the provided notebook, you will still need to turn in a `.ipynb` Jupyter Notebook and corresponding files according to the instructions in this homework.

ASSIGNMENT TASKS

(30%) Learn more about the analysis of the FARS database.

In the last homework we learned about the FARS database, and actually USED it!

While it might seem like a nice dataset to have (and it is), researchers have been using it for years to study transportation issues in this country, with the hopes of using the outcomes of these analyses for safety policy decisions.

You will get a little more cozy with such research by reading the paper:

> U. Roy, "Comparative Analysis of Fatal Pedestrian Crashes between Kansas and USA," JTTs, vol. 09, no. 03, pp. 381–396, 2019, doi: 10.4236/jtts.2019.93024.

As you will read, the phenomenon of pedestrian fatalities (involving automobiles) have been studied (and are relatively low, but not non-existent) has in recent years been on the rise, with no single reason for the change.

It is important as a data scientist, that you be able to reference literature and assess its relevance to analysese you may be doing as a professional, and with datasets such as those in FARS, there are plenty of researchers trying to understand correlation and causes between the multitude of changes in the data.

In this paper (Roy, 2019), you see some familiar concepts and some unfamiliar, but you may come understand a component of the data that we did not explore in the prior homework.

You can download the paper free from one of the following sources:

- SemanticScholar link

Read the paper (it will take about an hour) and answer the questions below.

§ Task: Read the paper and write a 4-7 sentence (about a paragraph) *summary*. State in your own words what you learned, what expanding your knowledge of the topic and what you found *interesting* about the information you received. Please include the major points of the paper, and any weaknesses the authors point out with their research.

§ Task: Answer the following questions:

(a) What time of day is most common for pedestrian fatalities in Kansas (over all years)?
(b) How does this compare with the most common time of day for the US overall?
(c) Looking at figure 11, would you say poor atmospheric conditions have a significant impact on pedestrial fatalities?
(d) On page 392, the author states *"For Kansas, speed limits between 30 mph and 40 mph account for 52% of total crashes (26% crashes for 30 mph and 26% for 35 mph or 40mph), ..."*. Why is this statement as written incorrect?
(e) The authors go on to explain the abnormally high number of fatalities at higher speeds with *"... Kansas has lot of rural roads, where the speed limit is high and in rural roads, laws are not strictly enforced, all of which might lead to a larger number of fatal pedestrian crashes."*. Which of the suggested countermeasures would you think might successfully addres this issue? If you do not find anything sufficient, what might you recommend instead?

(20%) Recreate data analysis in a published paper.

As a data scientist, you must be prepared to defend the analyses and models that you create. This is often done with workflow tools and processes which capture the full range of activities performed during an analysis. Commercial tools are very good at this, but often simply organizing your notebooks will go a long way to keeping track of your analyses.

In academic publishing, it is becoming more common for authors to submit data used in their papers. If the data was not already public, is it customary to ensure the released publication data does not violate the privacy or personally identifying information of others. It is also customary to make certain data was obtained in a way that the participants were aware that the data was going to be shared or otherwise given an option to decline to have such data publicly released.

In the case of this paper, both datasets (the KARS and FARS) are public.

For this part of the assignment you will recreate some of the graphs in tha paper from part 1 above. For all graph outputs, **you do not need to recreate colors – you can use the standard colors in Pandas and matplotlib.**

NOTE: You will need to refer back to, and otherwise reuse, your work from your first homework to complete this part.

§ Task: Recreate the graph (with actual FARS data) in Figure 15. You do not need to show the Kansas data. **BONUS (up to to 3 points extra):** You will earn up to 3 bonus points added to this assignment if you access the KARS data set and perform the same analysis (in other words, you cannot just put the numbers in figure 15 in the graph and earn the bonus).

§ Task: Recreate Figure 4 and Figure 9.

§ Task: Take the data from Figure 4 and Figure 9 and combine them. What you will end up with is a grouping by age band, then by time of day. Your final graph will be able to answer questions about which age group is more or less likely to experience a pedestrian fatality during which time of day.

(50%) Perform cluster analysis on NHTSA FARS data.

Now that we have data, let's analyze it.

You will, by now, have watched the lectures on clustering and unsupervised learning algorithms. You will know that clustering provides a significant tool for understanding underlying patterns in data and can be powerful in explaining the various aspects of your data.

In this part, we will go one step further and begin to explore the patterns in the FARS dataset.

§ Task: *(Reshape the US dataset from HW1)*

You will go back to the dataset in HW1 and begin the process of reshaping the data so that you will only include pedestrian fatalities, that is to say, you will only include fatalites with pedestrians.

You will also need to make sure of the following:

- you are using data from the 1975-2020 that you found in HW1

- you are including all states data **except Kansas**
- restrict to *numeric features*, eliminate year as a feature
- scale the data using scikit-learn Standard Scaler

You will end up with a new dataset which will only have pedestrian fatalities, numeric features and data scaled for 1975-2020.

§ Task: *(Perform ad hoc K-Means clustering)*

You will now take the dataset from the first part and begin the process of clustering.

To be successful, please study the following:

- K-Means in scikit-learn
- K-Means example notebook

You will choose a three K for clusters 5, 10 and 12. You will need to report the centroids for each cluster and in words how you would describe that cluster. I will give more guidance on this in a mini-session.

§ Task: *(Perform elbow analysis to find optimal cluster size)*

In the first part, we chose the cluster size K. Another way to do this is to analyze the change in within cluster sum of squares and determine when such value fails to change significantly. In other words, when the addition of another cluster fails to significantly change the within cluster sum of squares, then you can be confident more clusters won't make a difference (increasing K will no longer be relevant).

This is often referred to as "Elbow Analysis" or the "Elbow Method" because you will visually find the elbow in a plot of the sum of squares and choose K based on that.

Study the following code, implement it, and find the optimal K based on it.

Your answer must include:

- the elbow graph
- the optimal K
- the reanalysis of the previous answer based on the optimal K (re-run your clusters and report their centroid characteristics)

Here is the code to help you:

```
max_clusters = 15
css = [] # within cluster sum of squares

for k in range(1,max_clusters):
  kmeans = KMeans(n_clusters=k, 'k-means++', max_iter=200, n_init=10, random_state=0)
  kmeans.fit(d) # where d is the dataset you have standardized in the first part of this
  css.append(kmeans.inertia_)

# now make a line plot of all the values in css
...
```

§ Task: *(Find out where Kansas fits in)*

Using the optimal K found in your elbow analysis and understanding the characteristics of each representative in each cluster, which cluster would Kansas likely belong to?

By extension, which states are most like Kansas (belong to the same cluster) regarding pedestrian fatalities?

§ Task: *(Perform Agglomorative Clustering)*

One method also worth implementing is the hierarchical technique agglomorative clustering. This clustering method has the benefit of interpretability and as mentioned in the lectures, can be a powerful technique to show how a cluster is partitioned and can give specific insights to the features of clusters themselves. Most often the output of the cluster is in the form of a dendrogram, but often these can be complex to show on a single screen or page without scrolling.

To be successful you will need to study the following:

- main sklearn api to AgglomerativeClustering
- plotting the dendrogram

Set the parameter `n_clusters` to the same value found in your work above.