STAT437 – Unsupervised Learning - Final Project – 200 Points

Due: December 15th by 11:59pm CST on Canvas.

1. Working in **groups**
2. Main **purpose** of the analysis
3. Thinking about what makes a **cluster(ing) "meaningful"**
4. **Additional** analyses
5. **Intended audience**
6. Project **AI tools policies**
7. **Qualitative** evaluation criteria
8. Project **format**
    a. Report
    b. Presentation
    c. Presentation peer evaluation questions
    d. Report peer evaluation questions
    e. Individual contribution questions
9. **Dataset** options
10. **Report** Specifications + Rubric
11. **Presentation** Rubric
12. **Pre-Report Team** questions
13. ***Report* peer evaluation** questions
14. ***Presentation* peer evaluation** questions
15. ***Individual contribution*** report questions

1. Working in Groups

You should work in groups of 3-4 people

You must do at least 25% of the work in order to get full credit.

To receive full credit, you should follow the steps and answer the questions given in this document for your project. However, if you think that there are additional questions or analyses that would add additional insights to your overall research goal, you're more than welcome to pursue these in addition to what is stipulated in this document.

2. Main Purpose of the Analysis

Identifying, Exploring, and Describing the "Inherent" Clusters

The purpose of this analysis is to learn as much as you can about the dataset, the clustering structure, and the clusters that exist in your dataset. Ideally, you will be able to identify and describe each of the "inherent" clusters that exist in your high dimensional dataset.

Actionable Insights

In your research motivation you should identify at least one type of person/application etc. in which the insights extracted from your clusters would be useful and actionable.

3. Thinking about what makes a cluster(ing) "meaningful"?

Different Cluster Definitions

For instance, one *could* apply k-means to an unclusterable dataset and return, say, k=3 clusters. But would these k=3 clusters be considered "meaningful"? Depending on research motivation for the cluster analysis, perhaps not. If the person that you describe in your research practically considered a "meaningful" cluster as a set of observations that were relatively well-separated from other observations, then this set of k-means clusters would be considered not meaningful and perhaps misleading about the nature of the dataset in this scenario.

Variable Scaling

Furthermore, you would want to consider if the type of person/application described in your research motivation would want the clustering structure of the dataset to be dominated by higher scale variables, or would they want the contribution of each variable to contribute to the clustering structure equally. If so, then you should scale your variables first.

Irrelevant Clustering Structure Contributions

By performing basic descriptive analytics on your dataset (even after variable scaling), do you think that the **clustering structure** detected by some of your clustering related algorithms **may *still* be dominated by some variables** more than others?

- For instance, do you suspect that the clustering structure (or clustering algorithm results) might be dominated by the categorical variables? If so, what are some work-arounds you can try for this?
- Do you have some numerical variables that only have a few distinct values, but the gaps between these values don't necessarily indicate any meaningful clusters?
    - Ex: a clustering structure that is just defined by the discrete value gaps of 4 shoe sizes (7, 8, 9, 10) doesn't necessarily indicate a meaningful clustering structure.

Clustering Results that Only Exist because of some Data Preprocessing Property/Decision

For instance, does your dataset have a **suspicious amount of 0 values** for a given variable, whereas most of the rest of the variable values are much higher than 0? Do you suspect that whoever made/curated the dataset might have imputed missing values with 0's? If so, you might want to consider dropping these rows (or columns) because a clustering algorithm that simply finds clusters of observations that had missing values is not very interesting.

Masking Variables

A **masking variable** in a dataset that you intend to cluster is variable that does not contribute anything to the clustering structure and can weaken the ability of an algorithm to detect the clustering structure. Do you think that your dataset has any masking variables? If so, you might consider deleting them or giving them smaller scaling weights.

4. Additional Analyses

If you think that there are additional questions or analyses that would add additional insights to your overall research goal, you're more than welcome to pursue these in addition to what is stipulated in this document.

5. Intended Audience/Reader of your Project

The intended audience of your report/presentations should be someone who taken this STAT437 class. **Theoretically, you should be able to send/present your report to one of your classmates (who is not on your team), and they should be able to understand everything that you did and the claims that you are making.**

6. Project AI Tools Policy

Common STAT437 AI Slop Types + Project Penalties

Please review the Unit 0 AI Slop tables to see what kind of "AI slop" we will be on the lookout for in your projects.

- For each instance of AI slop of this nature that we find, we will take off all points in the corresponding rubric cell.
- We will also take of -3 *additional* points for each distinct type of AI slop that is found.

Project Group Tiers

You will work in groups of 3-4 to complete the final project. To mirror professional penalties, at the end of the semester, students will end up in one of two tiers.

- **Tier 1 Students:** *At most 2 assignments* in which there was an "AI slop warning".
- **Tier 0 Students:** *More than 2 assignments* in which there was an "AI slop warning".

Forming Project Groups Based on Tiers

- A group of ALL **tier 1** students can work together.
- A group of ALL **tier 0** students can work together.
- In order for a mixed group with **tier 1** AND **tier 0** students to work together, **EACH tier 1 student in this proposed group needs to *individually* email me saying that this is ok.**

  If a group is already formed, and then a student gets more than 2 "AI slop warnings", I will email this group getting verification that this intended group is still fine.

AI Slop Penalties in the Group Project

In order to ensure fairness in group contributions, if a group member significantly lowers the project grade due to **excessive AI slop** (or not completing the prompts in their assigned section):

- this **teammate will be graded exclusively on their contributions** and receive that as their final project grade
- the remaining teammates will be graded on the collective grade (minus this teammate's contributions).

Reviewing your Teammates Work

- **Professional Expectation for Group Work:** The expectation is that you will read over your teammates work before submitting the final report.

- **Opportunity to not Get Penalized for Significantly Poor-Quality Teammate Work:**

    - In the individual contributions report, there will be a question asking if you have reviewed your teammates work and have noticed anything significantly wrong.
    - IF your teammate significantly lowers the overall project grade due to AI slop, missing significant parts of the rubric, or very low-quality work, BUT **you *do not* indicate that anything was wrong**:
        - you may be graded collectively without this teammates contributions, BUT you
        - may lose some professionalism points (see rubric).

7. Qualitative Evaluation Criteria

In addition to being graded for correctness and completion, this project will also be graded on a qualitative basis. Qualitatively, we will be looking for the following things.

Clarity about Analyses, Algorithms, and Data Choices

- Someone who has taken this class should be able to read through your report and/or watch your presentation and easily be able to do the following.
  - Replicate what you did in your analyses.
  - Know why you made the choices that you did in your analyses.

Clarity about Motivation (ie. the "so what?") of your Analyses

- *Beginning of the Report and Presentation:*
  - Someone who is **about to** read your report and watch your presentation should be able to clearly answer the questions.
    - *"Why should I (or someone else) care about the report that I am about to read/listen to?"*
    - *"What research questions do they intend to answer?"*
    - *"How do these research questions relate to their motivation?"*
  - Therefore, in the introduction of your report and presentation you should make this clear.
- *Middle of the Report and Presentation:*
  - While **in the middle of** your report and presentation, your audience should be able to clearly answer the question.
    - *"How do each of these analyses/algorithms/data choices that they're making/using tie back into the overarching motivation of this whole analysis?"*
  - Therefore, for each new analysis/model/algorithm/data choice that you make, you should explain this and make it clear to your audience.
- *End of the Report and Presentation:*
  - Someone who has **just finished** reading your report and watching your presentation should be able to clearly answer the questions:
    - *"Why should I (or someone else) care about the analysis that I just read/listened to?"*
    - *"Did their analyses and conclusions answer the research questions that they stated at the beginning of the report/presentation? If so, how? What were the answers to these research questions?"*
    - *"How would the results/answers to these research questions be useful to someone?"*
  - Therefore, in the conclusion of your report and presentation you should make this clear.

8. Project Format [5 components]

Pre-Report Team Questions [5 pts]

***Deadline*: November 18 11:59pm CST on Canvas.**

- **Steps**:
  - Before beginning significant work on the project your **team should all meet** to discuss expectations and plans for the group work. See the pre-report team questions list below for things that you should discuss.
  - Afterward, each teammates should fill out the **Pre-Report Team Questions**
- **Graded**:
  - For completeness

Project Report [152 pt]

Deadline: December 15th 11:59pm CST on Canvas.

Should contain: Everything stipulated in the **Project Report Specifications** discussed below.

Format:

- Jupyter notebook.
- This should look like a clean data analysis report that you would theoretically submit to an employer (not a homework assignment). Thus, at the very least, your report should have:
  - a title
  - headings for each of your sections
  - You should write paragraphs and in complete sentences.
- You can use and modify the attached project **project_template_YOURNAMESHERE.ipynb** file as a template for this report if you'd like. You can add and delete as many cells as you'd like in this file.

Graded:

- See "Project Report Specifications" section below for point breakdown.

Project Presentation [28 pt]

Presentation Date: December 16 7-10pm CST online.

Format:

- **Your presentation should be no more than 9 minutes.**
- You must present some part of the presentation in order to get full presentation credit.
- Presentation should be presented in **slides** (not the Jupyter notebook).

Graded:

- See attached **presentation rubric** for what you should present and how you will be graded.

Report Peer Evaluation [8 pts]

***Deadline*: December 18 11:59pm CST on Canvas.**

- **Steps**:
  - You will be randomly assigned to **read** another group's report *(as an individual).*
  - After reading their report you will fill out a survey form on **Canvas**, which will ask you the following questions (see last pages of this document).
- **Graded**:
  - For completeness

Presentation Peer Evaluation [5 pts]

***Deadline*: December 18 11:59pm CST on Canvas.**

- **Steps**:
  - You will be randomly assigned to **watch** another group's presentation *(as an individual).* It will not be the same group that you read the report for.
  - After watching their presentation, you will fill out a survey form on **Canvas**, which will ask you the following questions (see last pages of this document).
- **Graded**:
  - For completeness

Individual Research Impact Questions [2 pts]

***Deadline*: December 18 11:59pm CST on Canvas.**

- **Steps**:
  - **As an individual, you must do at least 25% of the work in your team to get full credit.**
  - You will be asked a few questions about the work that you *individually* contributed to your group and if you met your pre-arranged group deliverables.
  - You should have an understanding as to how your individual contributions influenced and were influenced by the insights and decisions made by your teammates. (See questions in last pages of this document)
- **Graded**:
  - For completeness

9. Dataset Options

You can choose your own dataset or you can choose from one of the three supplied datasets below. The csvs for each of these datasets are located in the same folder that this document is in. There is more information about each of these datasets below.

Choosing your Own Dataset

There are several places you can go to find interesting datasets, but here are some places you can start.

https://www.kaggle.com/datasets

https://archive.ics.uci.edu/ml/datasets.php

https://corgis-edu.github.io/corgis/csv/

https://data.world/datasets/clustering

https://github.com/fivethirtyeight/data

For students interested in sports data:

a. NFL: https://www.nflfastr.com/
b. MLB and other baseball: https://billpetti.github.io/baseballr/
c. CFB: https://saiemgilani.github.io/cfbfastR/index.html
d. More sports stuff: https://sportsdataverse.org/

If you decide to choose your own dataset, it must meet the following specifications.

1. Dataset Size Specifications

Your dataset should have:

- at least 5 attributes (not including the pre-assigned class labels if there are any) and
- at least 50 rows

(If there's a dataset that doesn't meet these specifications, but you're really interested in you can talk to me about it).

2. Data Cleaning and Scaling Considerations

Before moving onto to checking whether the dataset is clusterable below, you should think about any type of cleaning and scaling that would need to be done in order to create an insightful analysis. Do this cleaning and scaling before moving on to the clusterability check.

3. Clusterability and Cluster Algorithm Fit Specifications

In this project you will be asked to apply at least two clustering algorithms to this dataset. Thus, before proceeding with further analysis, you should do the following.

- **First, test whether your dataset is clusterable**.

    - You should apply the t-SNE algorithm on your **scaled** and/or **unscaled** dataset (depending on what you intend to use).
    - If the t-SNE algorithm suggests that there is a clustering structure, your dataset has passed this check.
- **<u>Clustering Algorithm Suitability</u>**
    - Next, you want to make sure that you *know of* at least two clustering algorithms that will able to cluster this particular type of dataset. For instance, if this is a numerical, structured dataset, then you know many clustering algorithms that can take this dataset as input. (You are not constrained to the clustering algorithms that we have learned in this class. However, ensuring that there are at least two clustering algorithms that we have learned in this class that will cluster your dataset can be a useful backup just in case the work in this project takes longer than you expected.)

Dataset Options (if you don't want to choose your own)

1. **U.S. College Dataset:** The data in the IPEDS_data.csv in the zip file contains information about 1534 U.S. colleges. You can find more information about the dataset here: https://www.kaggle.com/datasets/sumithbhongale/american-university-data-ipeds-dataset

   a. **Subsets of Rows and Columns:** Based on your research goals, you might choose to only cluster just a subset of the rows (or columns) in this dataset. In fact, I'd highly suggest only clustering a subset of meaningful columns. Clustering all columns in this dataset will not yield the most insightful clusters/results.

   If you choose to use a subset of this dataset, make sure to explain why and how it relates to the goal/motivation of your analysis.

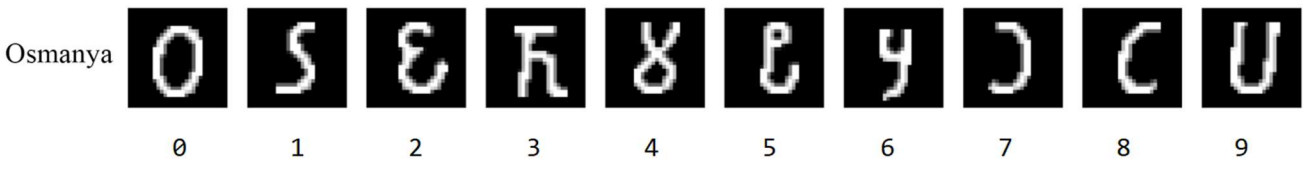
   Note: I can't guarantee that ALL subsets of columns and rows of this dataset will produce a clusterable dataset, but many will. Ask me if you are completely stumped when it comes to finding a clusterable subset of rows and columns. Ideally though, you'll be able to find a clusterable subset of rows and columns that are meaningful and sensible to your particular research goal.

   b. **Missing Values**: This dataset has missing values. You'll need to decide what to do with these missing values. Note that some columns have way more missing values than others.

   c. **Data Cleaning**: I'd highly suggest converting your "number" variables like "Full-time enrollment" to "Percent Full-time enrollment" to prevent these more nuanced variable from being dominated by schools with larger enrollments, for instance.

   df['Percent Full-time enrollment'] = df['Full-time enrollment']/ df['Total enrollment']

2. **Sample of the Afro-MNIST Dataset:** The observations in the osmanya_MNIST_sample.csv in the zip file contain a random sample of 1000 28-by-28 pixel images of digits ("0"-"9") from the Osmanya Afro-MNIST dataset.



   The full dataset and more information about the full dataset can be found here: https://www.kaggle.com/datasets/danjwu/afromnist

3. **U.S. State Substance Abuse Rates:** This dataset is about substance abuse (cigarettes, marijuana, cocaine, alcohol) among different age groups and states. Data was collected from individual states as part of the NSDUH study. The data ranges from 2002 to 2018. Both totals (in thousands of people) and rates (as a percentage of the population) are given.

Analysis tips/ideas:

- **Rates:** Given that U.S. states have different populations, a more insightful analysis would focus on just clustering the rates columns in this dataset.
- **51 Rows**: The objects that you cluster in this analysis should be individual states. Otherwise you will most likely get 50 clusters that correspond to each state, which is not very insightful. So your dataset that you cluster should only have 51 rows (50 states + Washington DC).
- **Years:** You might choose to cluster this dataset in the following way.
    - Cluster the rates for just one year (like 2018)
    - Cluster the rates for all years (or multiple years) all at once.
        - `(See code below, for instance, for how you can create columns in a dataframe that corresponds to the 'Rates.Tobacco.Use Past Month.12-17','Rates.Tobacco.Use Past Month.18-25' for each year).`
    - You might choose to cluster the rates for just one year, and then cluster the rates for another year, then compare the clustering results of the two years.
- **Which Rates:** You can choose to cluster all the rate columns. Alternatively, you might choose to focus on a subset of rates like:
    - just one age group
    - just a subset drug or drug activities

The full dataset and more information about the full dataset can be found here: https://corgis-edu.github.io/corgis/csv/drugs/

```
pivoted_df = df.pivot(index='State', columns='Year',
values=['Rates.Tobacco.Use Past Month.12-17','Rates.Tobacco.Use
Past Month.18-25']).reset_index()

pivoted_df.columns = ['_'.join(map(str, col)) for col in
pivoted_df.columns]

pivoted_df
```

10. Project Report Specifications + Rubric

Your report should include the analyses, code, and explanations detailed in each of the following sections.

General Report Professionalism	Points
Report Professionalism * Should not exhibit "ChatGPT linguistic style" * Write in complete sentences * Write text in the markdown files (not code blocks). * You are not copy-pasting the prompts/questions from this rubric and answering them. Rather, you should incorporate the requirements in this rubric naturally into a paragraph. * Appropriate titles/headers are used.	4
1. Introduction You should write an introduction (1-2 paragraphs) for your report. Your introduction should include/incorporate the following things.	
Research Introduction and Motivation * Clearly state the motivation for why someone might want to learn more about the inherent clusters that exist in this dataset. * Describe at least one type of person/application that may find your cluster analysis useful and how they might use it. Be specific about how they might use it. * Your explanation should include: clear "next steps" for what your motivation person might do with the results of your analysis. * Your explanation should include: the downside of a clustering algorithm NOT finding all of the inherent clusters in this particular dataset. * Your explanation should include: the downside of a clustering algorithm erroneously "splitting" an inherent cluster in this particular dataset. * Your explanation should include what kind of clustering result outputs (ie. hard partition, fuzzy clustering, dendrogram) might be useful to your research motivation and why. * You should use at least ONE CITATION that support your motivation/answers in this section. * Make sure that your citations are referenced and cited appropriately in this document.	6
Summarization of Sections * Your introduction should also give a brief description of what you will do in each section of this report (at most 1 sentence for each section).	2
2. Dataset Discussion You should write a paragraph in your report discussing your dataset(s) that you will be using to answer these research questions. This paragraph should include/incorporate the following things.	
Dataset Display * Read your csv file and display the first 5 rows of your dataframe. * How many rows are in your dataframe (originally before any data cleaning)?	1
Dataset Source * State where YOU got this csv file (dataset) from. * Provide a link/reference to where it came from. * State when you downloaded this csv file.	1.5
Original Dataset Information In the place where you found this dataset, try to answer the following questions. If the source does not give the answer to these questions, say so.	4

* What do the rows (ie. observations) represent in this dataset? * How was this dataset collected? * Is this dataset inclusive of ALL possible types of observations that could have been considered in this dataset? If not, what types of observations might be left out?	
<u>Selected Variables</u> * Briefly describe the variables you intend to use in your analysis. * Why do you think the inclusion of *these* variables are best when it comes to your research motivation? * Should each of these variable be considered equally important in the analysis?	3
3. Basic Dataset Cleaning and Exploration You should show and discuss any dataset cleaning decisions that you made in this section.	
<u>Missing Value Detection and Cleaning</u> * Does your dataset have any missing values? * If so, clean these missing values. * Are there any downsides to cleaning the missing values in this particular way?	3
<u>Outlier Identification - Two Variable Outliers (IF YOUR DATASET IS IMAGES, YOU CAN SKIP)</u> *For every pair of numerical explanatory variables that you're using, create a scatterplot. * Are you able to detect any outliers in these plots?	3
<u>Outlier Identification - 3+ Variable Outliers</u> *Use either a (KNN distance plot or a single linkage dendrogram) to look for outliers including those that can only be seen in 3 or more dimensions. * How many outlier/noise points are being suggested?	4
<u>Outlier Consideration</u> *In the context of your research motivation, what do you think should be done with any identified outliers? Explain. - Should they be dropped? If so, what are some of the pros and cons of dropping these outliers? - Should they be clustered in their own singleton clusters? - Should they be clustered with larger clusters that may happen to be further away? * If you identified outliers in your dataset, does this impact the type of clustering algorithms or clustering evaluation metrics that you might use in your analysis? Explain.	4
<u>Noise Consideration and Identification</u> * Use a technique discussed in this class to determine if your dataset has any noise. *In the context of your research motivation, what do you think should be done with any identified noise? Explain. * If you identified noise in your dataset, does this impact the type of clustering algorithms or clustering evaluation metrics that you might use in your analysis? Explain.	2
<u>Other Data Cleaning</u> * Were there any other data cleaning steps that you deemed suitable for this analysis? What were they? Why did you choose to perform this additional data cleaning? * If there are, do so here. * If you dropped rows, how many did you drop?	
4. Basic Descriptive Analytics Before using any unsupervised learning algorithms, you should learn more about your dataset by performing some basic descriptive analytics.	

OPTION 1 **If your dataset is a structured dataset (ie. not image, audio, time-series data etc.), do the following.** * For your numerical attributes, calculate basic summary statistics about each attribute. * For any categorical attributes (including the pre-assigned class labels, if your dataset has any) count up the number of observations of each type. * Determine if there exist are any strong pairwise relationships between the variables in your dataset.	3
OPTION 2 **If your dataset is an image dataset, do the following.** 1. If your dataset has pre-assigned class labels: * Visualize the first few images of each type of class-label. * Discuss how much image variability each of the classes has, and what image elements are different. 2. If your dataset DOES NOT have pre-assigned class labels: * Visualize a random sample of images from this dataset. * Discuss how much image variability the image in your dataset have, and what image elements are different.	
5.Scaling Decisions	
From your analyses conducted here, discuss whether you should scale the dataset or not. Explain why or why not. If you choose to scale, then do so in this section here.	2
6. Clusterability and Clustering Structure Questions	
t-SNE Plots + Clusterability Check * Create some appropriate t-SNE plots (6 perplexity values, 2 random states each). * Select a representative t-SNE plot that displays similar patterns seen across many plots.	2.5
Describe the Underlying Clustering Structure of the Dataset What do your t-SNE plots suggest about the following: * Is your dataset clusterable? (THE ANSWER TO THIS SHOULD BE YES) * Approximately how many underlying clusters does the data have? * What are the shapes of the underlying clusters? * Are the clusters balanced in size? * Are there are any clusters that are not well-separated? * Is there any evidence of nested cluster relationships in this dataset?	6
Clustering Structure and Attribute Association * Is there an association between each of the attributes and the clustering structure suggested by the t-SNE plot? Show the appropriate visualizations to explain. (That is, you should color-code your t-SNE plot by each of your attributes and talk about the amount of association between the attribute labels and the t-SNE plot suggested clusters).	4
Understanding the t-SNE algorithm * Caution your reader about 4 original dataset properties that your t-SNE plots are not able to reveal or represent.	3
7. Clustering Algorithm Selection Motivation	
Clustering Algorithm #1 Explain why you chose to use your first clustering algorithm to cluster this dataset. In your explanation, you should discuss and consider: * your research motivation	5

* the "ideal dataset properties" that this algorithm is designed to work best for (your anaylses above should give you a sense as to whether many of these ideal properties are met or not). Your explanation should give <u>at least 3 distinct reasons.</u> <u>HINT</u>: I OFTEN FIND A LOT OF INCORRECT AI SLOP IN THESE ANSWERS IN THE PAST. BE CAREFUL.	
<u>Clustering Algorithm #2</u> Explain why you chose to use your first clustering algorith to cluster this dataset. In your explanation, you should discuss and consider: * your research motivation * the "ideal dataset properties" that this algorithm is designed to work best for (your anaylses above should give you a sense as to whether many of these ideal properties are met or not). Your explanation should give <u>at least 3 distinct reasons.</u> HINT: I OFTEN FIND A LOT OF INCORRECT AI SLOP IN THESE ANSWERS IN THE PAST. BE CAREFUL.	5

8. Clustering the Dataset and Post-Cluster Analysis for <u>Algorithm 1</u>

<u>Parameter Tuning and Clustering Exploration</u> When it comes to selecting your "best" set of parameters, you should do ALL of the following. *You should try out many combinations of parameters (if your algorithm has more than 1 parameter to tune).*	
<u>Elbow Plot</u> (ONLY IF you can choose the cluster number k) IF your algorithm allows you to preselect the number of clusters, make a corresponding elbow plot. Then interpret this elbow plot. * Is it detecting a clustering structure. * If so, how many clusters is it detecting? <u>t-SNE Plot Corroboration</u> * Cluster your dataset using many parameter values (combinations of parameter values) and color-code your t-SNE plot by the cluster labels. * Which of these clustering(s) has the strongest corroboration with the t-SNE plots? * Do you have reason to believe that these clusterings are in alignment with our research goals and/or are identifying the inherent clusters? Explain. <u>Average Silhouette Score Plot(s)</u> * Cluster your dataset using many parameter values (combinations of parameters values) and calculate the corresponding clustering average silhouette score. Plot these in average silhouette score plot(s). * Highlight a few clusterings with high average silhouette scores. * Do you have reason to believe that these clusterings are in alignment with our research goals and/or are identifying the inherent clusters? Explain. <u>Cluster-Sorted Similarity Matrix</u> * Cluster your dataset using many parameter values (combinations of parameter values) and create cluster-sorted similarity matrices for the clusterings. * Comment on which clustering(s) showcase the most cohesion and separation in the matrices.	16

Outlier and Noise Handling * If your dataset had outliers and noise, evaluate whether your clusterings handled the outliers and noise the way that your research motivation wanted them to. Fuzziness (ONLY IF your clustering returns a fuzzy clustering) * If this is fuzzy c-means, you should also evaluate and select your fuzziness parameter p based on whether it has achieved a desireable level of fuzziness. Do your t-SNE plots showcase more fuzzy assignments for probable straddle nodes, and less fuzzy assignments for non-probable straddle nodes? * If this is NMF or GMM, you should also evaluate whether your results achieve a desireable level of fuzziness. Again, you can use t-SNE plot color-coding.	
"Best Clustering(s)" Based on your parameter tuning section above, which clustering(s) do you think provide the strongest evidence that you have: * identified *all* of the inherent clusters * are the most informative * and are most in alignment with your stated research goals and motivations (ex: outlier handling)? Explain your answer.	4
Technique Shortcomings * Elbow plot: If you created an elbow plot, is there anything that has been suggested about this dataset that might suggest that the elbow plot technique may NOT suggest the number of clusters that we want? Explain. * Average Silhouette Score: Is there anything that has been suggested about this dataset that might suggest that the clustering with the highest average silhouette score is NOT the one that identifies all the inherent clusters? Explain. * Cluster-Sorted Similarity Matrices: Is there anything that has been suggested about this dataset that might suggest that the clustering which has found the inherent clusters, may NOT actually display "good" block diagonal form? Explain.	4
"Best" Clustering(s) Algorithm Results Presentation For each of your "best" clusterings that you selected above, cluster the dataset again using these best pararameters and display the results. For instance: • If your clustering is a hard assignment, you can color-code your t-SNE plot with the cluster labels. • In addition, if your clustering algorithm is a hierarchical clustering algorithm, you should ALSO show the dendrogram. • If your clustering is a fuzzy clustering (or has cluster membership scores), you can plot K (# of clusters) t-sne plots, and color code each plot by the cluster membership score for the kth clusters.	2
9. Clustering the Dataset and Post-Cluster Analysis for Algorithm 2	
Parameter Tuning and Clustering Exploration When it comes to selecting your "best" set of parameters, you should do ALL of the following. *You should try out many combinations of parameters (if your algorithm has more than 1 parameter to tune).*	16

Elbow Plot (ONLY IF you can choose the cluster number k) IF your algorithm allows you to preselect the number of clusters, make a corresponding elbow plot. Then interpret this elbow plot. * Is it detecting a clustering structure. * If so, how many clusters is it detecting? t-SNE Plot Corroboration * Cluster your dataset using many parameter values (combinations of parameter values) and color-code your t-SNE plot by the cluster labels. * Which of these clustering(s) has the strongest corroboration with the t-SNE plots? * Do you have reason to believe that these clusterings are in alignment with our research goals and/or are identifying the inherent clusters? Explain. Average Silhouette Score Plot(s) * Cluster your dataset using many parameter values (combinations of parameters values) and calculate the corresponding clustering average silhouette score. Plot these in average silhouette score plot(s). * Highlight a few clusterings with high average silhouette scores. * Do you have reason to believe that these clusterings are in alignment with our research goals and/or are identifying the inherent clusters? Explain. Cluster-Sorted Similarity Matrix * Cluster your dataset using many parameter values (combinations of parameter values) and create cluster-sorted similarity matrices for the clusterings. * Comment on which clustering(s) showcase the most cohesion and separation in the matrices. Outlier and Noise Handling * If your dataset had outliers and noise, evaluate whether your clusterings handled the outliers and noise the way that your research motivation wanted them to. Fuzziness (ONLY IF your clustering returns a fuzzy clustering) * If this is fuzzy c-means, you should also evaluate and select your fuzziness parameter p based on whether it has achieved a desireable level of fuzziness. Do your t-SNE plots showcase more fuzzy assignments for probable straddle nodes, and less fuzzy assignments for non-probable straddle nodes? * If this is NMF or GMM, you should also evaluate whether your results achieve a desireable level of fuzziness. Again, you can use t-SNE plot color-coding.	
"Best Clustering(s)"	4
Based on your parameter tuning section above, which clustering(s) do you think provide the strongest evidence that you have: * identified *all* of the inherent clusters * are the most informative * and are most in alignment with your stated research goals and motivations (ex: outlier handling)? Explain your answer.	
Technique Shortcomings	4

* Elbow plot: If you created an elbow plot, is there anything that has been suggested about this dataset that might suggest that the elbow plot technique may NOT suggest the number of clusters that we want? Explain. * Average Silhouette Score: Is there anything that has been suggested about this dataset that might suggest that the clustering with the highest average silhouette score is NOT the one that identifies all the inherent clusters? Explain. * Cluster-Sorted Similarity Matrices: Is there anything that has been suggested about this dataset that might suggest that the clustering which has found the inherent clusters, may NOT actually display perfect block diagonal form? Explain.	
"Best" Clustering(s) Algorithm Results Presentation For each of your "best" clusterings that you selected above, cluster the dataset again using these best pararameters and display the results.	2
For instance: • If your clustering is a hard assignment, you can color-code your t-SNE plot with the cluster labels. • In addition, if your clustering algorithm is a hierarchical clustering algorithm, you should ALSO show the dendrogram. • If your clustering is a fuzzy clustering (or has cluster membership scores), you can plot K (# of clusters) t-sne plots, and color code each plot by the cluster membership score for the kth clusters.	
10. Discussion	
Clustering Comparison: What was different?	4
* Were the "best" results from your two clustering algorithms (1 and 2) different? * To quantify the clustering result differences, calculate and interpret the adjusted rand score between each pair of "best" clusterings. *If one of your clusterings was not a partition-based clustering, convert it into one.* * In addition, more specifically describe what types of observations were different between these two clusterings and how they were different. t-SNE plots color-coded by the cluster labels can help you describe this.	
"Best Clustering(s)"	3
Out of ALL of the clusterings that you have explored (from Algorithm 1 and 2), which clustering(s) do you think provide the strongest evidence that you have: * identified *all* of the inherent clusters * are the most informative * and are most in alignment with your stated research goals and motivations (ex: outlier handling)? Explain your answer.	
Different Clustering Insights	3
* Does clustering algorithm #1 reveal any insights about the data that algorithm #2 does not? If so, what are they? * Does clustering algorithm #2 reveal any insights about the data that algorithm #1 does not? If so, what are they?	
Additional "Best" Clustering(s) Exploration For each of your "best" overall clustering(s) that you selected above you should also use cluster sorted similarity matrices to answer the following questions to describe the clusterings further.	3

Separation: Summarize the distances between the clusters. * Are there clusters that are much further away from the others? Which ones? * Are there clusters that are much closer to other clusters? Which ones? Sparsity: Summarize the sparsity of he clusters. * Are there clusters that are more dense than others? Which ones? * Are there clusters that are more sparse than others? Which ones?	
Additional "Best" Clustering(s) Attribute Descriptions For each of your "best" clusterings that you selected above you should describe what attributes characterize each of the clusters that you found. ***[QUALITIATIVELY, THIS IS THE MOST IMPORTANT PART OF THE PROJECT!]***	
Describing Each of the Clusters Finally, describe what type of attribute values and attribute relationships characterize each of the resulting clusters in your final clustering. You can choose at least one of these options (or pick multiple options to learn more). • Option 1 (don't use this if your dataset is an image dataset): * Create a side-by-side boxplots visualization for each numerical attribute in your dataset (where each cluster label is given a boxplot). * Create a side-by-side barplot visualization for each categorical attribute in your dataset (where each cluster label appears on the x-axis). o Use these plots to **thoroughly describe** which type of attribute values characterize each of the resulting clusters in your final clustering. Ex: **Cluster 1** is characterized by low crime, high pollution, etc. **Cluster 2** is characterized by high crime, medium pollution. You should only leave an attribute out of your descriptions if there is NO association between that attribute and the cluster labels (ie. there is no pair of IQR boxes that don't overlap). • Option 2 (if you used a prototype-based clustering algorithm, and an image dataset): * If your clustering algorithm is a prototype-based clustering algorithm, display (visualize if it's an image dataset) and compare each of the prototypes of the clusters. * Use these prototypes to thoroughly describe which type of attribute values characterize each of the resulting clusters in your final clustering. • Option 3: (If you used NMF) * If you used NMF to cluster the rows of a dataset, display (visualize if it's an image datset) the rows of H. * each of the resulting clusters in your final clustering.	7
11. Conclusion	
Summarize Briefly summarize what you did in this report and what your key findings were.	3
Recommendation * Based on your analyses above, would you recommend that any of the insights that you extracted be used by the person/application discussed in your research motivation? Why or why not? * If so, how might they use these insights? Be specific.	3

Shortcomings/Caveats * Discuss any other shortcomings to your analysis here (all analyses have SOME shortcomings).	3
Future Work * Based on what you observed in your analysis, what are some ideas you might have for future work?	3
Total	**152**

11. Project Presentation Rubric (28 points)

PRESENTATION

Length [1.5 points]

- **The presentation is no longer than 9 minutes.**

Clarity and Motivation [2 points]

- Clearly states their research goals.
- Clearly states and "sells" their motivation for the analysis. The motivation is believable.
- Clearly states the extent to which they achieved their research goals.

Presenting [3 points]

- All team members speak and present **(with cameras on)** some portion of the material.
- Team members speak loud enough for everyone to hear.
- Team members understand the material, they **are not reading directly from a notecard or script.**

SLIDES

Content [14 points] You should present *some* content on each of these topics.

- Motivation and introduction
- *Very* brief display/discussion of the dataset (no more than 30 seconds).
- *Briefly* discusses any data cleaning decisions made (no more than 30 seconds).
- Show noteworthy visualizations from your descriptive analytics section. Why were these visualizations that you showed noteworthy with respect to your research goals?
- Discusses the clustering structure suggested by the t-SNE plot.
- Clustering #1
    - Why did you use this algorithm (brief)?
    - Displays algorithm results.
    - Describes what attributes characterize the clusters in the clustering.
- Clustering #2
    - Why did you use this algorithm (brief)?
    - Displays algorithm results.
    - Describes what attributes characterize the clusters in the clustering.
- Discusses insights and algorithm comparisons made in the discussion section
- Conclusion (including shortcomings)

Correctness [1.5 points]

- Analyses are appropriate for the data, results are interpreted correctly.

Layout [6 points]

- **No code shown on the slides!**
- **No irrelevant code output is shown on the slides.**
- Content is well organized.
- Fonts are easy to read.
- Visualizations are not messy and are easy to see and interpret.
- Slides are engaging.
- Slides are not too wordy.
    - Your slides should not contain paragraphs of text.
    - Should use bullet points.
    - Complete sentences are often not needed and can be visually burdensome.

12. Pre-Report Team Questions

Before beginning significant work on the project your **team should all meet** to discuss expectations and plans for the group work. See the pre-report team questions list below for things that you should discuss.

Afterward, each teammates should fill out the **Pre-Report Team Questions**

These questions will be posted on a Canvas quiz for you to submit.

1. **Meeting**: Did your group meet to discuss these questions and expectations for group work? Did you attend the meeting?
2. **Project Specifications:** Did you have the chance to read over the project specifications prior to the meeting?
3. **Workload Distribution**: What is your *current* plan for how to distribute the work for this project? Does this feel equal to you? *[Very brief description] (It's ok if you change this later on).*
4. **AI Expectations:** Did you read over the AI policies for the project (in the project specifications)? Did your group discuss expectations about AI use in the project?
5. **Version Control:** What is your group's plan for combining your work? What tools/platforms might you use?
6. **Group Communication:** How will your group stay in touch about team collaboration?
7. **Contingency Plans:** Has your group discussed contingency plans of what to do if at least one of your teammates is unexpectedly unable to complete their tasks, or non-responsive for whatever reason? *Note: If you have concerns about being unable to meet a deadline or having a very busy work week, you should share these concerns with your group sooner rather than later.*
8. **Checking Cohesiveness:** Which student(s) are in charge of making sure that all of the submitted work is cohesive and consistent? (This may be all students, or perhaps 1-2). *Note*: *This is an important role which requires more work than what you might expect. This should be considered when coming up with the workload distribution.*
9. **Final Look Over + Deadline Buffer:** What deadline has your group set for all of the individual work submissions to be brought together in a final report? Have you given your group a buffer for you to all read over the collective report and point out any potential issues that you might see when it is all put together?

13. Report Peer Evaluation Questions

Your assigned group will see your responses.

These questions will be posted on a Canvas quiz for you to submit.

Motivation

1. What is the **motivation** for the analysis in this report? Or in other words, why should you (or someone else) care about the analysis that you just read?

Research Question/Answer

2. What was their **research questions** or **goals** of the analysis, and what was the **answer/outcome** of their question/goals?

Usefulness

3. For the person/type of person that they talked about in their motivation, what do you think would be their **"next steps"** after reading this analysis?

Correctness

4. Did you catch any **code or interpretation errors** in this analysis? If so, what did you catch?

5. **How confident** are you that you have caught all the code/interpretation errors in this analysis (if any)? Explain your answer. **Describe your process for how you went about looking for errors.**

Robustness

6. Name at least one **step/decision/interpretation** that this person made in their report in which you could envision another data scientist doing something different. Why do you think that this other data scientist might have done something different?

Transparency

7. Were there any analysis decisions made in this report, in which they made a particular decision, but **did not explain** either:
   - that they **made a decision** or
   - **why they made that particular decision?**

   If so, what were they?

8. Are there any **shortcomings** that you can think of (the report should have mentioned some) in which their analyses may not have provided *perfect* answers to their research questions.

Clarity/Readability

9. Roughly how long did it take for you to come up with an answer to each of questions (1)-(8) above?
10. What are some **tips** that you would give to this researcher for how you might have been able to come up with answers (1)-(8) more easily/more quickly?

14. Presentation Peer Evaluation Questions

Your assigned group will see your responses.

These questions will be posted on a Canvas quiz for you to submit.

Motivation

1. What is the **motivation** for the analysis in this report? Or in other words, why should you (or someone else) care about the analysis that you just read?

Research Question/Answer

2. What was their **research questions** or **goals** of the analysis, and what was the **answer/outcome** of their question/goals?

Usefulness

3. For the person/type of person that they talked about in their motivation, what do you think would be their **"next steps"** after watching this presentation?

Correctness

4. Did you catch any **code or interpretation errors** in this analysis? If so, what did you catch?

5. **How confident** are you that you have caught all the code/interpretation errors in this analysis (if any)? Explain your answer. **Describe your process for how you went about looking for errors.**

Robustness

6. Name at least one **step/decision/interpretation** that this person made in their report in which you could envision another data scientist doing something different. Why do you think that this other data scientist might have done something different?

Transparency

7. Were there any analysis decisions made in this report, in which they made a particular decision, but **did not explain** either:
   - that they **made a decision** or
   - **why they made that particular decision?**

   If so, what were they?

8. Are there any **shortcomings** that you can think of (the report should have mentioned some) in which their analyses may not have provided *perfect* answers to their research questions.

Clarity/Readability

9. Roughly how long did it take for you to come up with an answer to each of questions (1)-(8) above?
10. What are some **tips** that you would give to this researcher for how you might have been able to come up with answers (1)-(8) more easily/more quickly?

15. Individual Research Impact Questions

These questions will be posted on a Canvas quiz for you to submit.

1. What were your individual **contributions** to this group project?

2. Did you read over your **teammate's work**? Is your teammate's work **correct**, including all of the **requested specifications**, and is **free of noticeable AI slop**? If not, explain.