{Fill in title}

Machine Learning, Spring 2021

{Fill in names}

Abstract

{Fill in your abstract}

Introduction

{Fill in your introduction}

Setup

```
set.seed(315)
# Load necessary packages
```

Loading and Preprocessing the Data

1. Begin by loading the dataset you've chosen. Depending on which dataset you've chosen, it may be helpful to consider a subset of the data for your analyses. Give an explanation as to why you did or did not subset your data.

```
# Load the data.
```

2. Next, provide some basic information about the data. What are the samples and what are the variables? How many samples are there? How many variables are there? How many of those variables are numerical vs categorical? How many of the numerical variables are discrete? Discuss any other information about the data that you think is relevant.

```
# Investigate the characteristics of the data.
```

3. In some scenarios, it may be helpful to transform or preprocess your data in some way before embarking on your analysis. For example, one may wish to center and standardize each numerical variable before performing certain regression analyses. On the other hand, one may wish to group some factors of a categorical variable together. Perform any preprocessing that you feel is appropriate and discuss the motivation for this choice. If you decide not to, explain why. Note that it may be helpful to omit this step and return at a later point once you have explored the data a bit.

```
# Preprocess the data.
```

Exploratory Data Analysis

4. Provide a statistical summary of the data as well as a summary of the data in words. Note that it is not necessary to list every summary statistic for every variable. Rather, try to give the reader a sense of what kinds of values the variables take on. And don't forget any categorical variables!

```
# Summarize the data.
```

5. Pick three variables in your dataset and visualize the distribution of each in separate plots. Remember to properly title and label each plot. Compare them and comment on what you see. Do any of the distributions appear similar or different? In what ways?

```
# Visualize distributions of the data.
```

6. Pick two numerical variables in your dataset. What is their correlation? Visualize their joint behavior in a single plot. Does this suggest that there is a strong relationship between the two?

```
# Compute correlation.
# Plot joint behavior.
```

7. Pick two numerical variables in your dataset. Perform a hypothesis test to check whether their means are equal. Discuss what test you performed and why. What is the result of the test? Note that such a test won't make sense if you have centered your data so you should apply this to your data without any centering.

```
# Perform hypothesis test.
```

8. In a single plot, visualize the relationship between the means or medians and the standard deviations of each variable. Do you notice any outliers?

```
# Plot means or medians vs sds.
```

Principal Components Analysis

9. Next, we would like to reduce the dimension of the data. Explain what the "dimension" of a dataset is. Discuss why one would want to reduce it. Provide an intuitive explanation of how PCA can achieve this.

10. Run PCA on the dataset. Decide whether to set $scale = TRUE$ or $scale = FALSE$ and explain your choice. Provide a numerical summary of the first 5 PCs. Note that you should leave any categorical variables out of this analysis.

```
# Run PCA.
```

11. Plot a screeplot of the PCs. How many principal components are required to explain at least 80% of the variation in the data? Based on this and the plot, does it seem like PCA has done a good job in reducing the dimensionality of the data?

```
# Plot screeplot.
```

12. Visualize the distributions of the first 3 PCs in separate plots and comment. Are there any apparent clusters? Do any of the first 3 PCs appear most helpful for separating the data?

```
# Plot first 3 PCs.
```

13. Plot the first 3 PCs against one another (see the Biplots section of Computing Assignment 3). Are there any apparent clusters?

```
# Plot first 3 PCs against one another.
```

Clustering

14. Next we would like to perform a cluster analysis on the data. Explain intuitively what that means. What exactly is a cluster? What are the objects being clustered? Why might this be helpful for this dataset?

15. Apply one of the clustering algorithms you learned about in class to your data. Depending on the algorithm you use, you may need to specify a cutpoint to obtain a single cluster label for each sample.

Note also that depending on the dimension of your dataset, it may be preferable to apply the clustering algorithm to the first few PCs of your data rather than the data itself. Elaborate upon your choice.

```
# Apply the clustering algorithm.
```

16. What did the clustering algorithm find? Are the clusters relatively homogeneous or heterogeneous? Summarize the results.

```
# Summarize results of clustering.
```

17. What is the within-cluster sum of squares for the clusters you found? How does this compare to the total sum of squares?

```
# Compute sums of squares.
```

18. Plot the first 3 PCs against one another again but include the cluster label for each point. Do any patterns emerge? Comment on what you see.

```
# Plot PCs with clusters.
```

Classification

19. Next we would like to perform classification on the data. Explain intuitively what that means. What does it mean to classify the data? What are the objects being classified? Why might this be of interest for this dataset?
20. Identify a categorical variable in the data to correspond to the class of each sample in the dataset. In particular, is there a variable that might be interesting to predict from the other variables?
21. Construct a table that describes how many samples in each class fall into each cluster. How well was your previous cluster analysis able to identify the classes you chose? Are any of the clusters comprised mostly by one class?

```
# Construct table.
```

22. Randomly break your dataset into a training set (roughly 80% of the data) and a test set (roughly 20% of the data). What is the point of doing this before doing classification?

```
# Break into train and test sets.
```

23. Using the class labels based on the categorical variable you chose earlier, apply a classification method to the training set. Be sure to specify any free parameters that were chosen. What is the accuracy of the fitted model on the training set?

```
# Run classification algorithm.
# Compute training set accuracy.
```

24. What is the accuracy of the fitted model on the test set? How does this compare to the accuracy on the training set? Does this make sense? Explain why or why not.

```
# Compute accuracy on test set.
```

25. Construct a confusion matrix based on the results of classification. Were there any classes on which your algorithm performed better than others? What about worse than others? Does this make sense based on the nature of the data?

```
# Find classes where the algorithm performed best / worst.
```

Conclusion

{Fill in your conclusion}