

**TASK**

**CAPSTONE PROJECT II**

[](http://www.hyperiondev.com/portal/)

**Introduction**

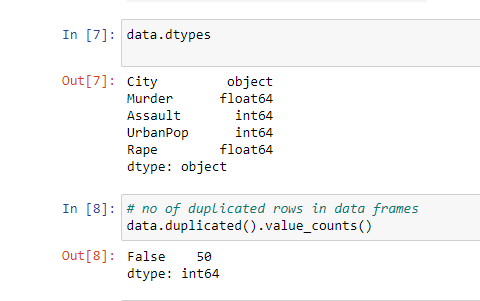
The data set I will be exploring today is the UsArrests data set. “This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas.” In this task, we explore the number and different types per city vs the population of the city using unsupervised learning methods such as Principal Component Analysis (PCA) and various clustering techniques.

Table

Description automatically generated

**DATA CLEANING**

The next step is to clean our data and to check the attributes of our data so that we know what we are dealing with and how to clean it in order for it to be useful for our analysis.



According to the findings above our datatypes are float 64, int 64 and int for the city column which in this case is our target column. We also notice that there are no duplicates in our data set.

MISSING DATA

Missing data is the most common problems when it comes to data analysis as it affects the quality of the result and the data itself. There is it’s important to check the data for any missing values before we proceed to any stages of our analysis.



From the above information, it shows that our data does not have any missing values in any of the columns.

CORRELATION ANALYSIS

Correlation is a statistical technique that can show whether and how strongly pairs of variables are related/interdependent. When we look at two variables over time if one variable change how does this affect change in another variable.

According to the diagram below you will notice that there is a strong positive correlation between murder and assault and an average relationship between rape and murder. I n other words we are saying the crime variables are correlated with each other.

From this information we can conclude that most murder cases were because of assault. This means rape and assault would be the cause of murder maybe due to resistance or fear to be reported if the victim has seen the face of the person.

This correlation suggests that the victims would end up being murdered but the initial intent would be assault and rape.

Chart, bar chart

Description automatically generated

PCA: UNSTANDARDISED DATA

Principal Component Analysis helps us to speed up the machine learning algorithm. If your learning algorithm is too slow because the input dimension is too high, then using PCA to speed it up can be a reasonable choice.

Table

Description automatically generated

pca.explained\_variance\_ratio

[0.62006039, 0.24744129, 0.0891408]

pca.explained\_variance\_ratio\_

[0.62006039, 0.24744129, 0.0891408]

According to the information given above, the first components are using up the most percentage of our data and these are the crime variables

Thus Principal Component Analysis is used to remove the redundant features from the datasets without losing much information. These features are low dimensional in nature. The first component has the highest variance followed by second, third and so on.PCA works best on data set having 3 or higher dimensions. Because, with higher dimensions, it becomes increasingly difficult to make interpretations from the resultant cloud of data

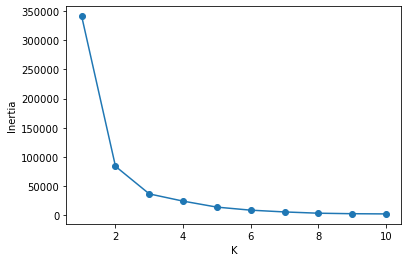
CLUSTER ANALYSIS

For this analysis we will use K-Means and Hierarchical clustering.

**K-means**

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. Typically, unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labelled, outcomes. A cluster refers to a collection of data aggregated because of similarities.

For illustration purposes, we will also show you the so-called Elbow method for choosing a value for K. The Elbow method is commonly used strategy in which the algorithm is run multiple times with increasing values of k. This is then plotted against the sum of squared distances of samples to their closest cluster center (called inertia). The value of k to choose is identified from the resulting plot where there is an 'elbow' in the graph.



From the diagram above, the elbow runs between 2 and 4. For this analysis we will set our k to 3.

Chart, scatter chart

Description automatically generated

Based on the clusters above, murder is determined by assault. Thus, we see that the best silhouette score is actually achieved using k=6.

Chart, scatter chart

Description automatically generated

**AGGLOMERATIVE HIERARCHICAL CLUSTERING**

Hierarchical clustering is an algorithm that builds, as the name suggests, a hierarchy of clusters, which form a tree-like structure called a “dendrogram”

Initially, each individual data point is assigned to a mini-cluster of its own (at the “leaves” of the dendrogram, A-F) and a measure of dissimilarity (such as Euclidean distance between their features) is defined. Using this dissimilarity metric, the algorithm consists of iteratively merging the most similar two clusters into one larger cluster, starting from the leaves and working upwards until all observations belong to a single cluster. The measure of the dissimilarity determines how far up the dendrogram the merge takes place - greater dissimilarity places the merge higher up the dendrogram

For the linkage criteria we use the complete and average criterias.

**Complete linkage is**  the dissimilarity metric is calculated between all samples in the one cluster with all samples in the other cluster, and the largest dissimilarity is chosen. **Average linkage** is the dissimilarity metric is calculated between all samples in the one cluster with all samples in the other cluster, and the average dissimilarity is chosen.

Chart, scatter chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

The data points in the bottom right belong to the victims that were not entirely murdered due to assaults. These are victims whose murder was intentional. The victims at top right have high chances of having been assaulted before being murdered.

Chart, scatter chart

Description automatically generated

[Exploring Correlation in Python: Pandas, SciPy (re-thought.com)](https://re-thought.com/exploring-correlation-in-python/)

[Principal Component Analysis explained | Kaggle](https://www.kaggle.com/nirajvermafcb/principal-component-analysis-explained)

[Understanding K-means Clustering in Machine Learning | by Dr. Michael J. Garbade | Towards Data Science](https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1)

**THIS REPORT WAS WRITTEN BY : GEMARIAH MADINDI**

