COVID-19

Data Analysis

ABDEL SINI GEORGE (N01530574)

FEMY MARIA JOSE(N01531559)

GOUTHAM PRASAD(N01531148)

SHILPA ANNIE JOHN(N01531556)

SHRUTHY CHANDRAN(N01530149)

Case Introduction

During the past **COVID-19** outbreak global agencies faced serious issues regarding the improper resource allocations. This led to **increased fatality rates.** This plight could have been avoided with optimal usage of resources.

In future, they can **minimize the fatality rates** by quickly identifying the hotspots based on **real time analysis** of COVID statistics and allocate resources accordingly.

Our project is to build a model to **provide insights** to minimize the consequences of the future outbreak.

QUESTIONS

"How to estimate the total number of deaths in certain countries for which the total number of death is unknown ?"

"How to identify the COVID hotspots in different continents to facilitate the government decisions regarding:

→ Vaccination

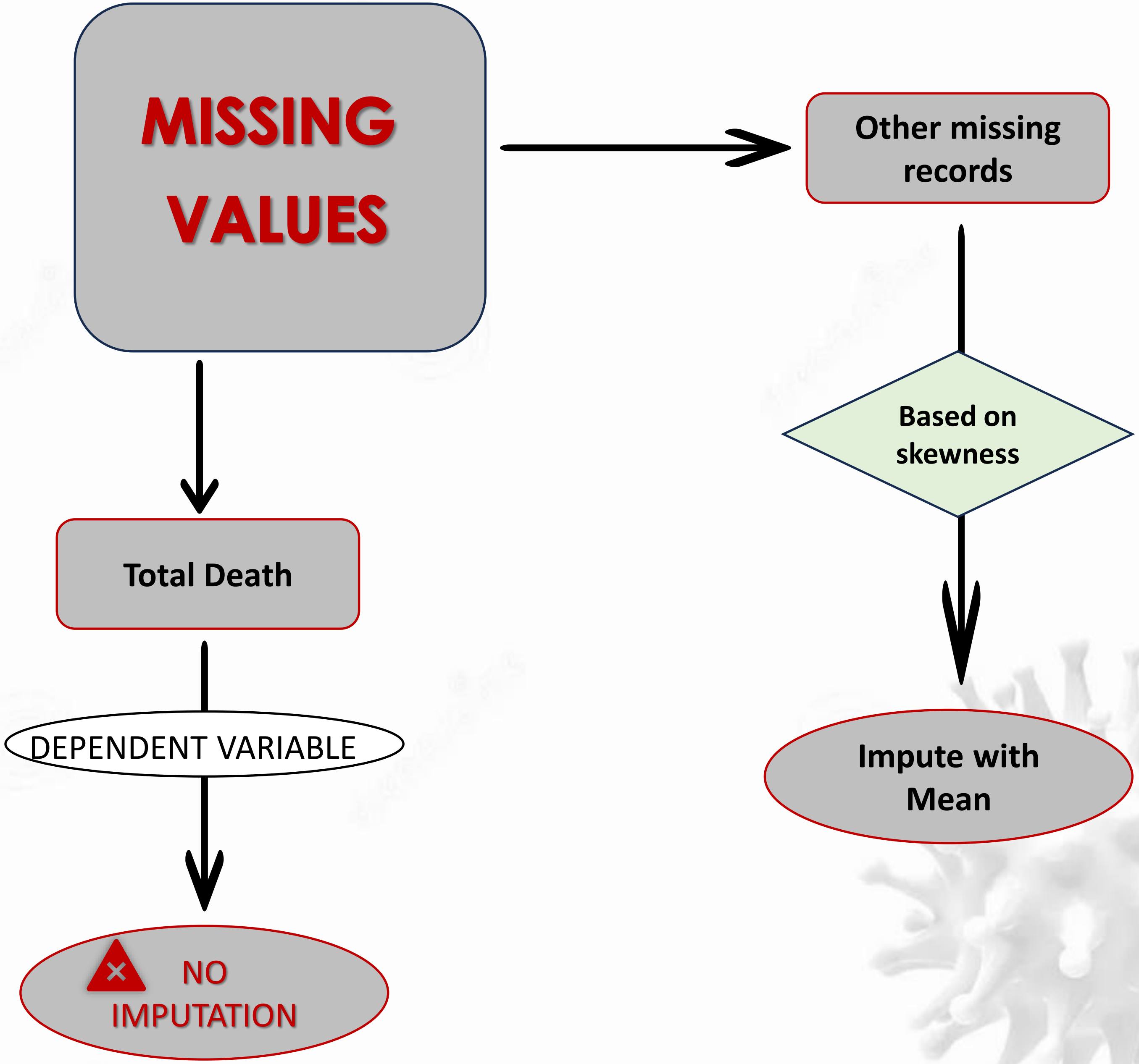
→ Other emergency services"

Data Collection

- The dataset we used is **The complete Our World in Data COVID-19 dataset.**
- Link: [Coronavirus (COVID-19) Deaths - Our World in Data]
- It is a time series data of **COVID global statistics** and **total number of deaths**.

DATA CLEANING STEPS

- Missing Values
- Irrelevant and Incorrect Data



MISSING VALUES

Other missing records

Based on skewness

Impute with Mean

Total Death

DEPENDENT VARIABLE

NO IMPUTATION

Irrelevant & Incorrect Data

The irrelevant and incorrect records were identified and removed.

Data Analytics Models

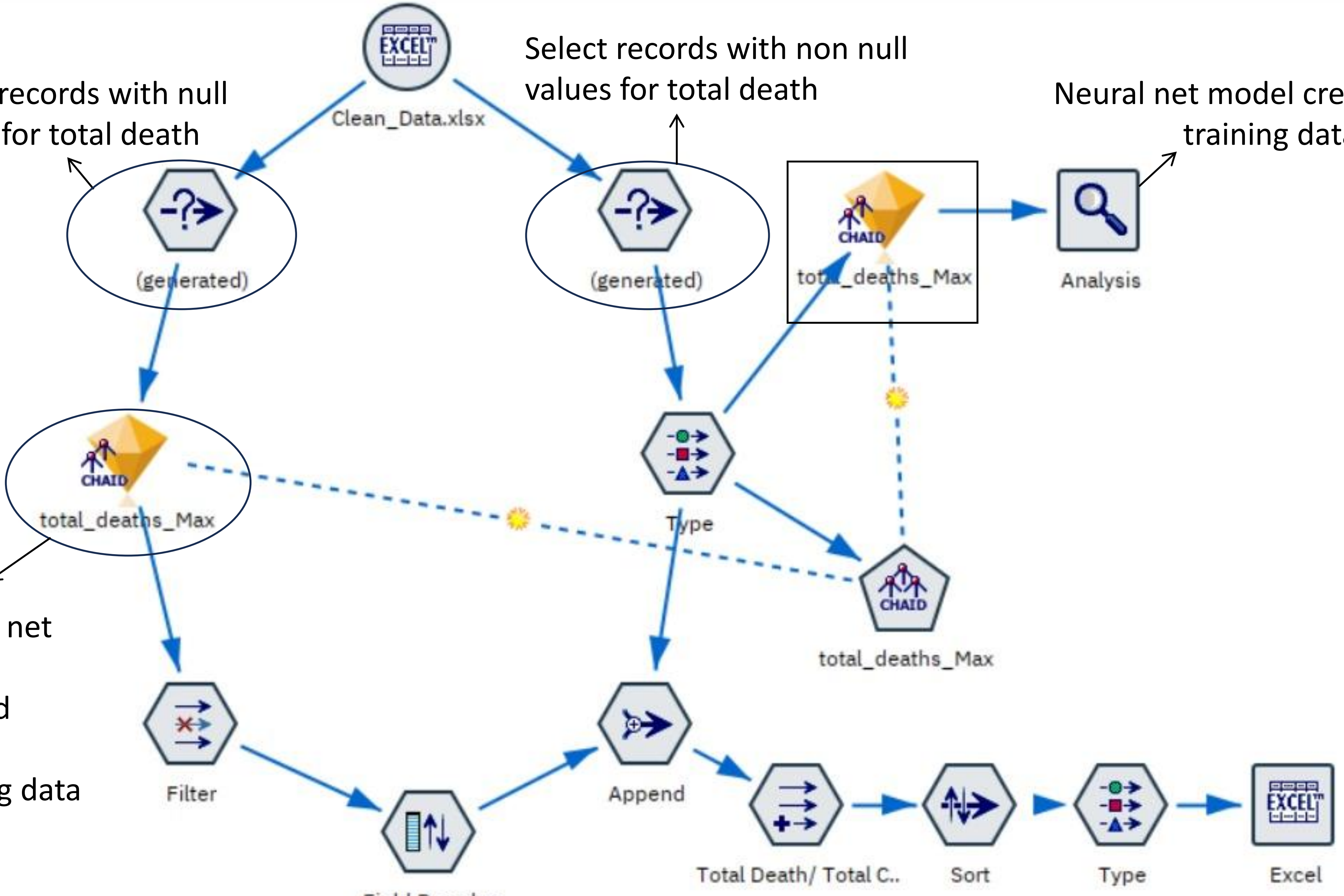
1. CHAID node

   Neural Net

2. K-Means Clustering

   Two-Steps Clustering

CHAID node



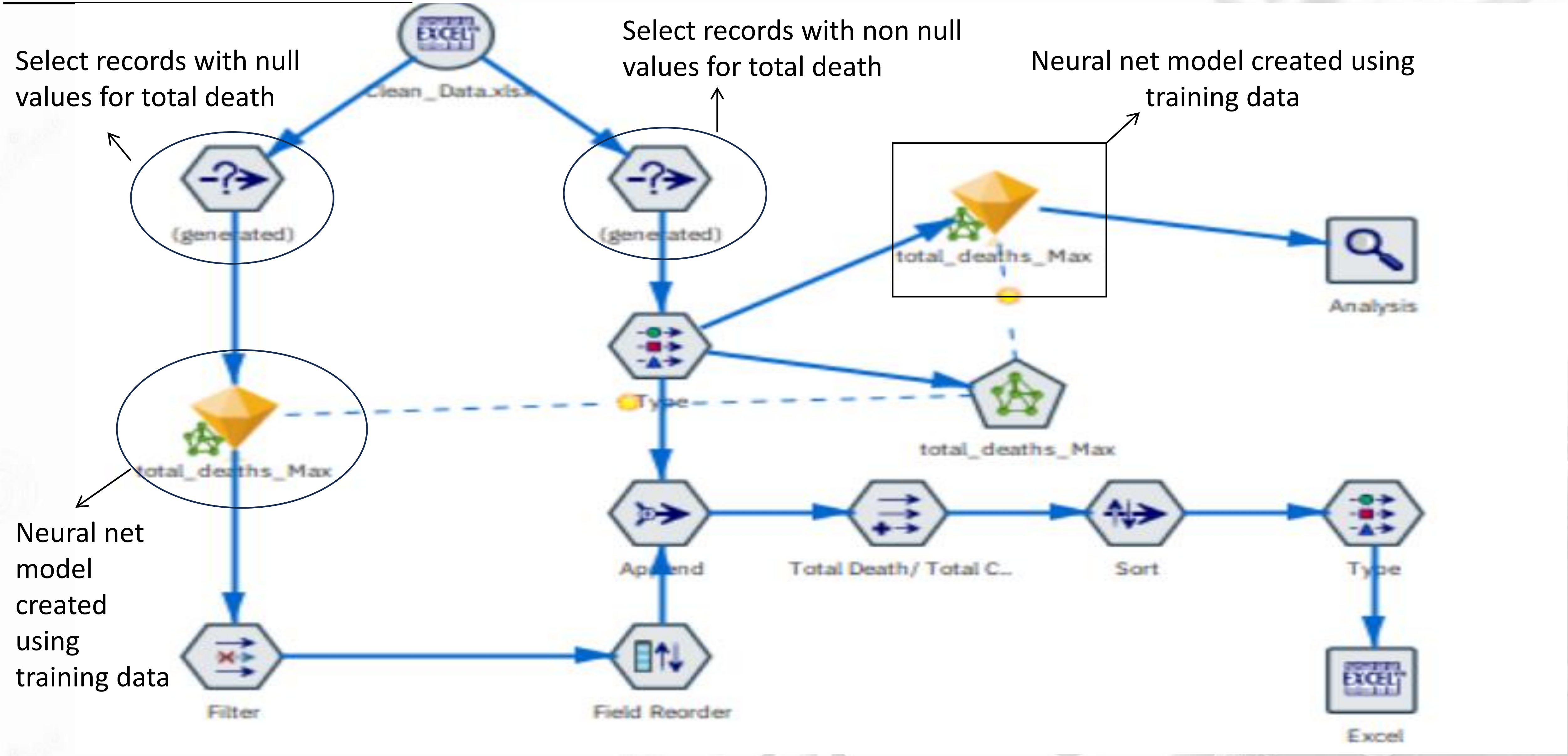
Select records with null values for total death

Select records with non null values for total death

Neural net model created using training data

Neural net model created using training data

Neural Net



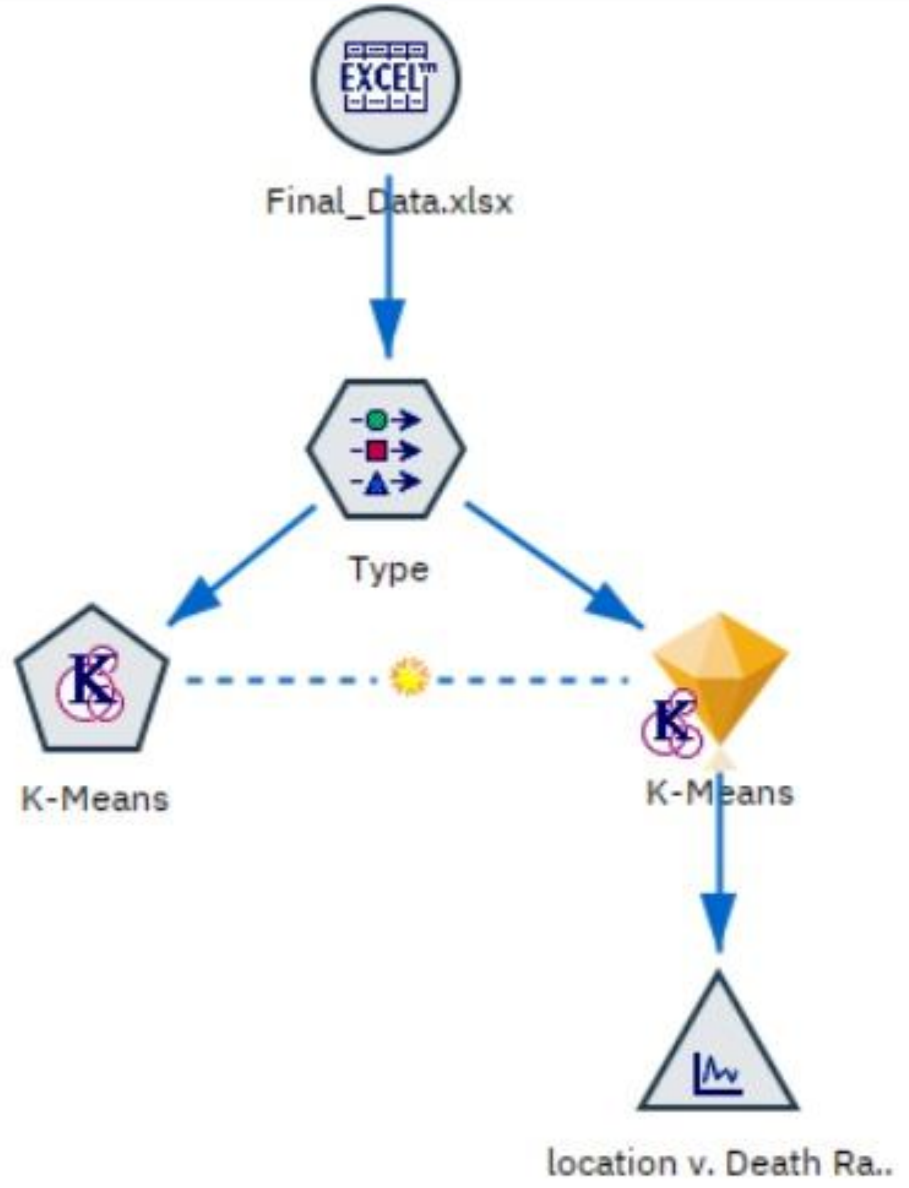
Select records with null values for total death

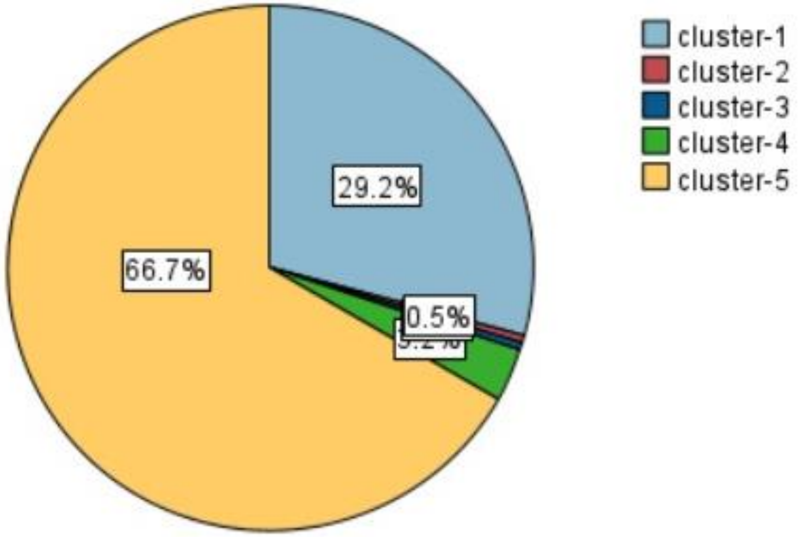
Select records with non null values for total death

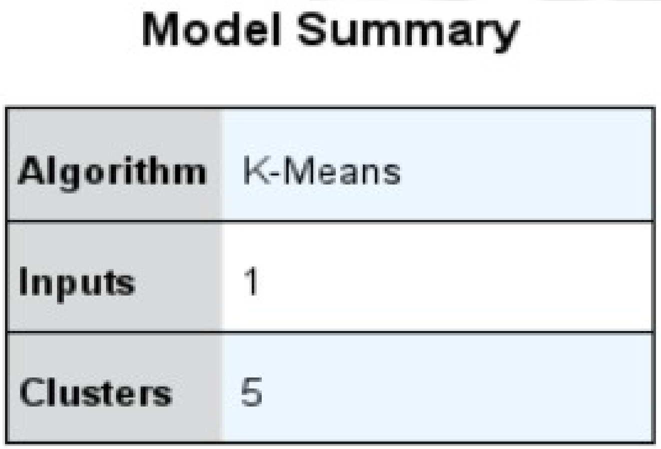
Neural net model created using training data

Neural net model created using training data

K-Means Clustering

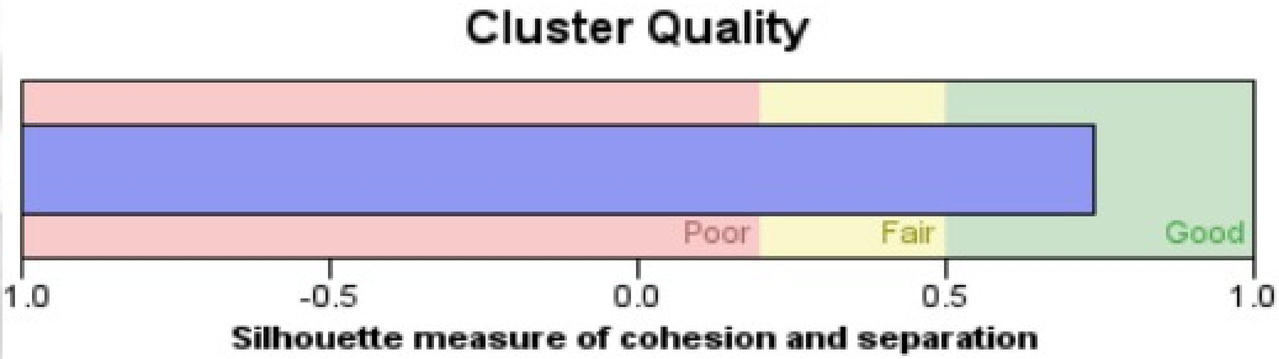




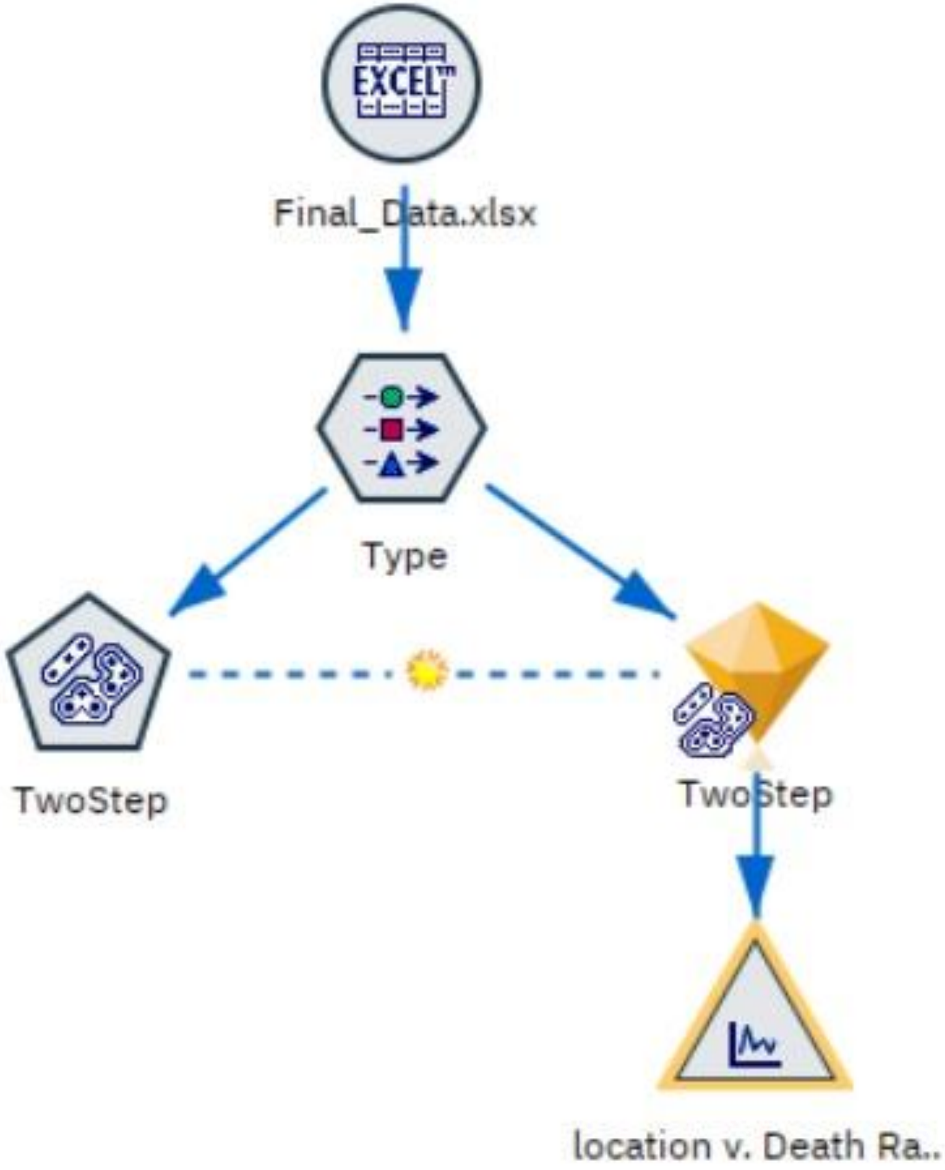


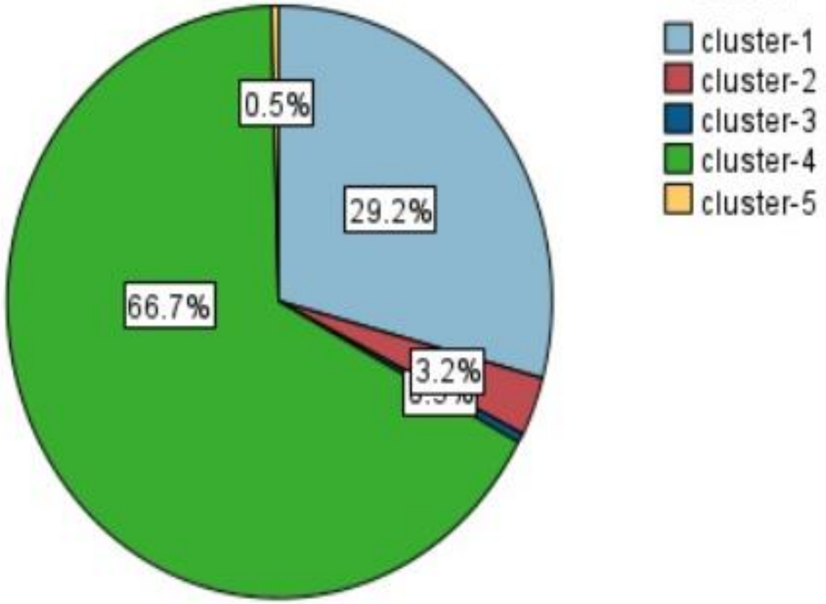
Model Summary

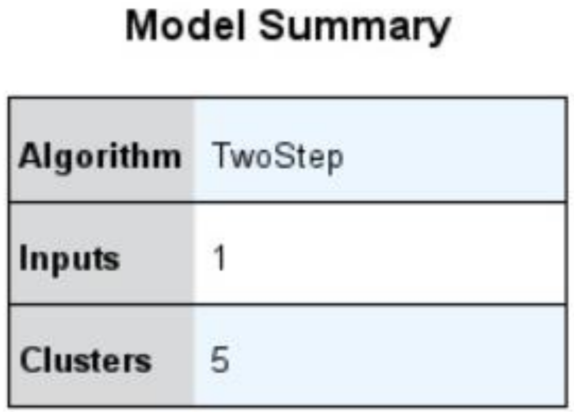
Algorithm	K-Means
Inputs	1
Clusters	5



Two-Steps Clustering

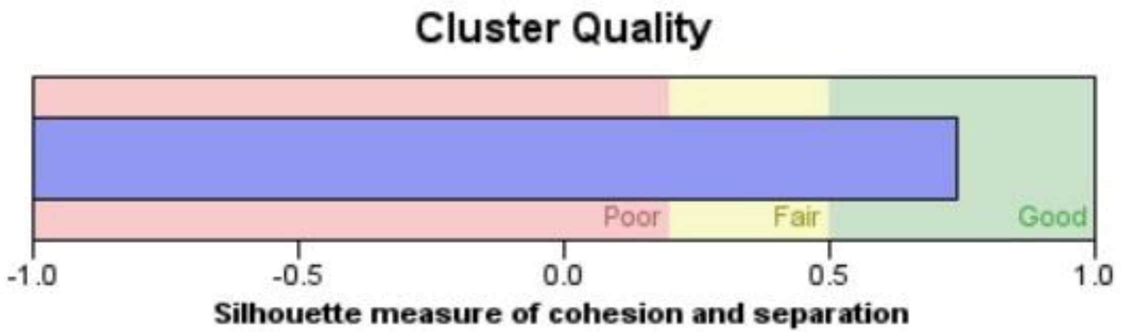






Model Summary

Algorithm	TwoStep
Inputs	1
Clusters	5



Model Comparison

REGRESSION MODELS

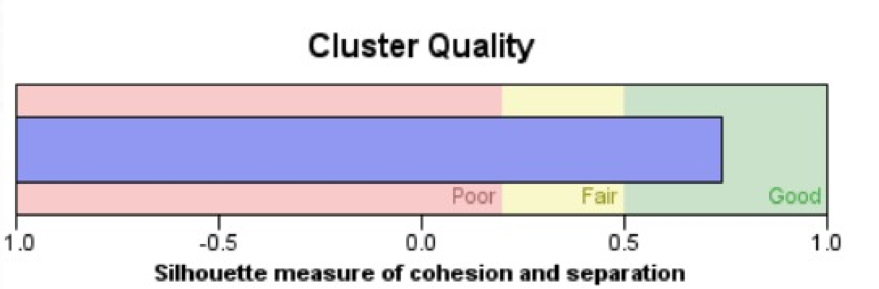
CHAID Model	Neural Net
Mean Absolute Error 14522.343	**Mean Absolute Error** 18240.759

Model conclusion:

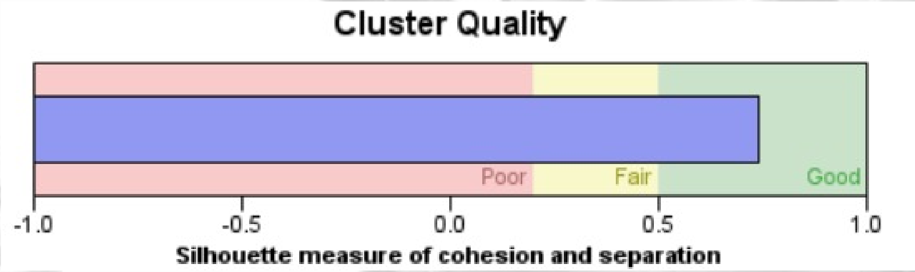
CHAID model has the lower value of MAE
Hence it is the preferred model for the data set.

CLUSTERING MODELS

K-Means Clustering



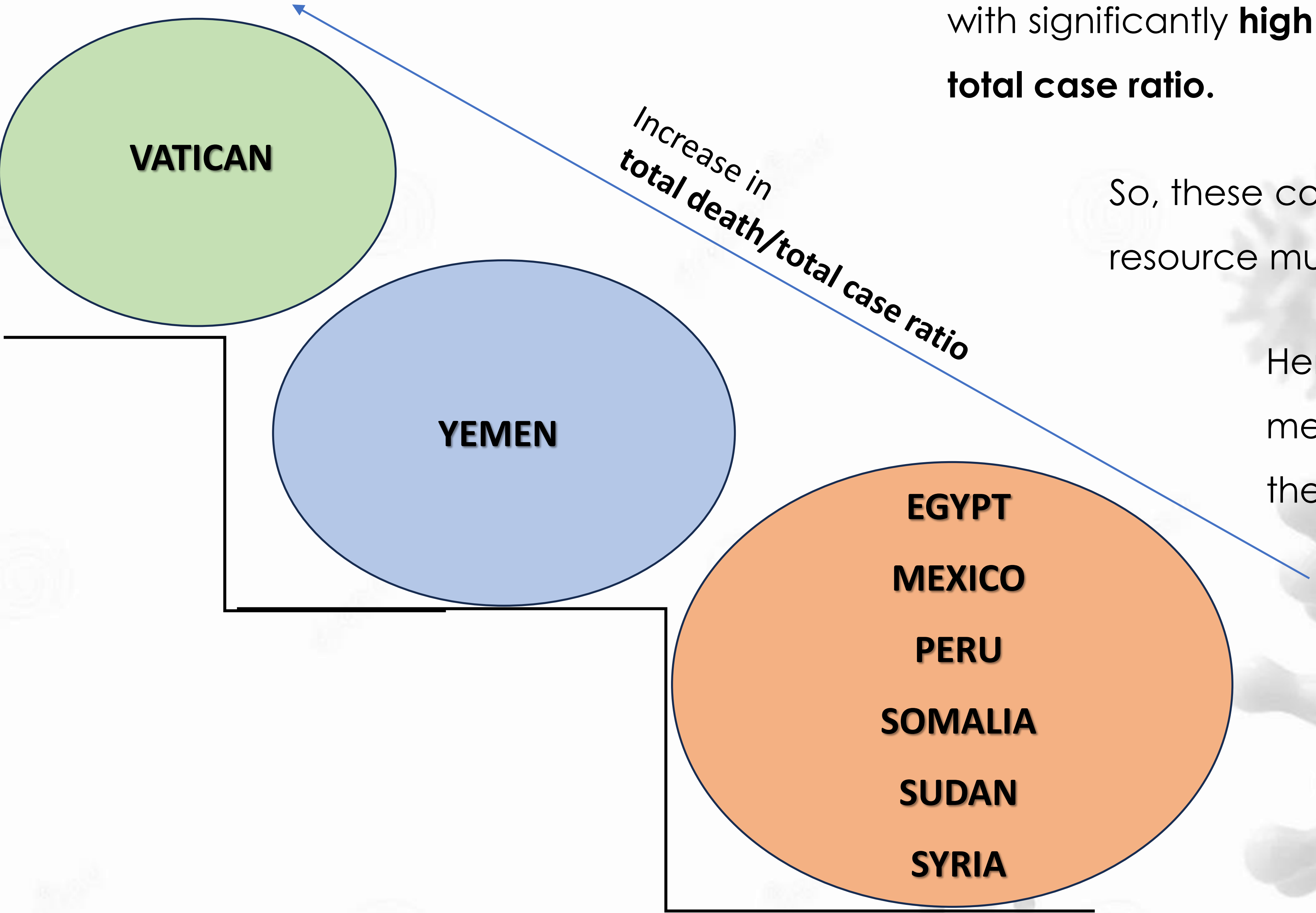
Two-Step Clustering



Model conclusion:

Both models give **similar results** on the dataset.
Hence anyone of the model can be used.

Conclusion



VATICAN

YEMEN

EGYPT

MEXICO

PERU

SOMALIA

SUDAN

SYRIA

Increase in **total death/total case ratio**

Cluster 5, Cluster 3, Cluster 2 are the clusters with significantly **high values of total death/ total case ratio.**

So, these countries should be **prioritized** and resource must be allocated accordingly.

Hence, these countries in the mentioned clusters are identified as the **hotspots**.

References:

- Mathieu, E., Ritchie, H., Rodés-Guirao, L., Appel, C., Giattino, C., Hasell, J., Macdonald, B., Dattani, S., Beltekian, D., Ortiz-Ospina, E., & Roser, M. (2020, March 5). Coronavirus (COVID-19) deaths. Our World in Data. https://ourworldindata.org/covid-deaths

- Shih, D.-H., Shih, P.-L., Wu, T.-W., Li, C.-J., & Shih, M.-H. (2022, July 2). Cluster analysis of US covid-19 infected states for vaccine distribution. Healthcare (Basel, Switzerland).https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9323689/

THANK

YOU