GROUP PROJECT:- GROUP 4

IT8416: Data Mining

201900615 Oskar Alkhateeb, 201900658 Zainab Salman, 2020002789 Mohamed Adel

Table of Contents

Task 1: Problem Statement and Definition

Introduction

Motivation and Project Definition

Covid-19 has had a devastating impact on the healthcare system and the lives of people globally. It has been fatal to over one million people in the United States of America alone. This project will be a correlation study about the impact of Covid-19 in the United States' healthcare system and how this superpower handled this pandemic.

Throughout the pandemic, the United States was known to be one of the countries who handled the situation poorly. With a severe delay in mask mandates, insufficient quarantining procedures, lack of contact tracing as well as the population refusing to isolate, it was a recipe for disaster. The nonchalant response to the pandemic from both the public and the nation's government was appalling and had contributed to the fatalities and amount of people who were infected. The Centres for Disease Control and Prevention did not have a centralised point or group who were able to correctly handle the pandemic and were slow and problematic when it came to testing the public.

These situations alone caused the health sector huge issues as people who likely had Covid-19 and were experiencing severe symptoms were going to emergency rooms. The influx in patients caused many fatalities as hospitals were not equipped to deal with many severe cases at once. This is proven by the fact that the annual death count in the years of the pandemic had drastically deviated from the standard deviation. (CDC, 2019)

Objectives

We use data mining to analyse and assess the trends of this past data. It will be formed into a case study to show future leaders and the population of this powerful nation what the effects of poorly handling a pandemic would constitute to. It could be used as a warning to future generations. The impact of this data would likely be bittersweet as it would teach us about what could happen if this situation was repeated with a different, inevitable pandemic but also allows us to visualise the lives that were lost and showing us the true nature of the pandemic.

Techniques Used

For this project, we will be using a hybrid of techniques. We will first investigate the past data statistics to obtain a summary of the pandemic and have a breakdown of what it shows. This will then be used to find trends in the data to predict the future changes and what could happen if it were to repeat itself in the future.

Expected Result

Our main prediction is that unless powerful countries, such as the United States, do not take the pandemic seriously, there will be dire consequences globally should another pandemic occurs. Especially if it is more fatal and contagious than Covid-19, there will be a bigger death toll. We will also see that as the variants come out, there will be a decrease in the death toll but an increase in number of cases. We would also be able to see the changes when vaccinations come into play. Having this data stood for this way will make it easier to look through and visualise outcomes with data to back it up.

Commented [MH1]: add link

Commented [ZS2R1]: Send the link

Commented [OA3R1]: @Mohamed Adel Jaafar Abdulredha Hasan my free what link?

Commented [MH4R1]: Data Download | The COVID Tracking Project

Task 2: Selection of the Data Set and Data Collection

Data Set Breakdown

Source

We have selected this dataset as it satisfies all our needs and is supplied by a lot of useful attributes and data, also covers every single state in the United States of America which will result in providing us with an accurate representation of the real time covid 19 cases, the presence of data loss and redundant data will also help us to recover those data and fix this dataset. (The Covid Tracking Project, 2021)

We decided to go with the United States as a source of data as the states were a perfect an example for a country that didn't take serious actions against the virus and the number of cases and death was outrages and for some period it topped the world charts with its covid-19 numbers.

Hopefully, we learn something from what the states did against covid-19 and help more countries and people by raising awareness that we could lessen the damages of such global health threats by analysing and visualizing this dataset.

The dataset holds 41 attributes of both types qualitative and quantitative such as date, state, death, hospitalized, negative Cases, Positive and much more. We have 20780 instances including duplicated instances and missing data which is ideal to implement data cleaning techniques on it.

Commented [MH5]: qualitative / quantitive

Commented [ZS6R5]: Done

Visualisation

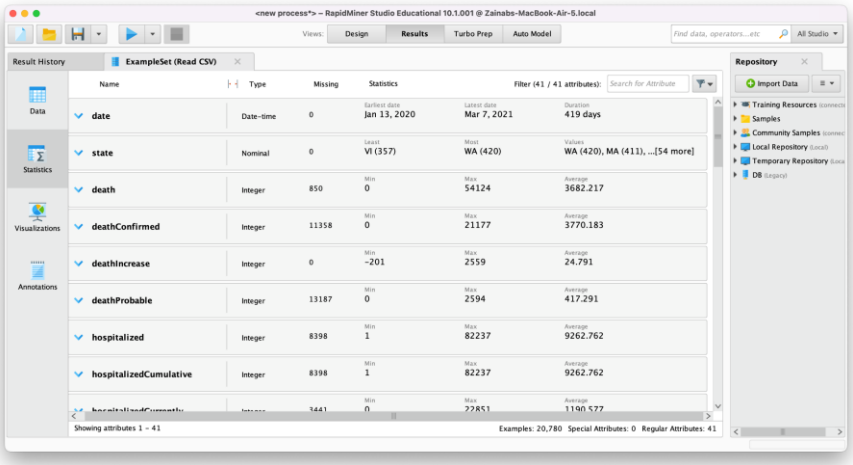


Figure 1: image shows statistics about the dataset including number of instances and attributes alongside redundant attributes and missing values.

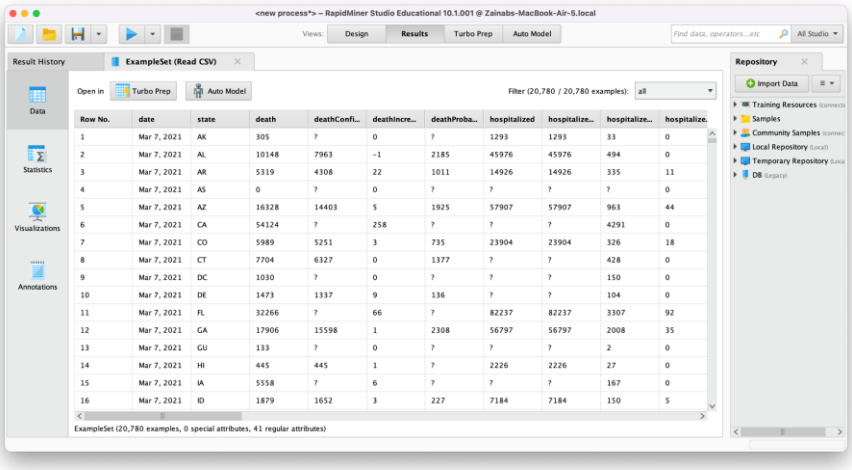


Figure 2: Shows some of the data from the dataset with the presence of missing and redundant data.

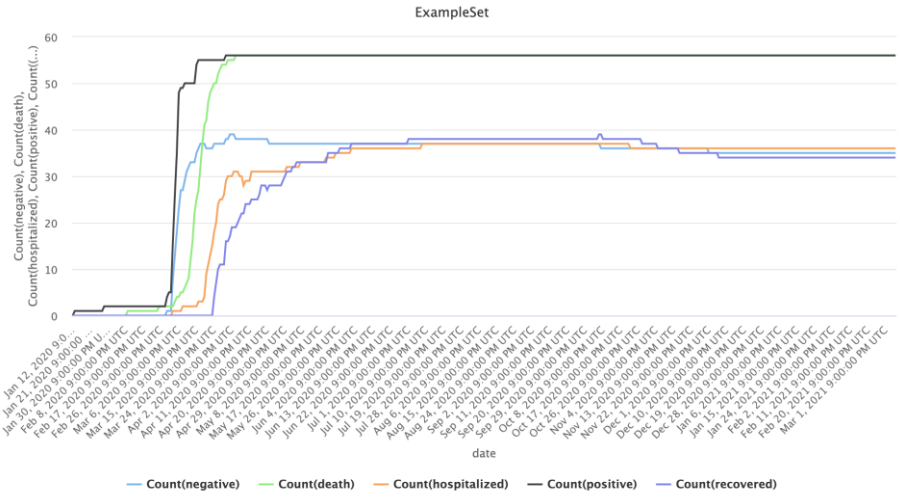


Figure 3

Figure 3 represents a graph illustrating the relationship between the date of data entry (x-axis) and the number of recorded deaths (y-axis). Prior to March 6th, the graph shows a slope that is relatively constant, implying that the number of deaths was stable and/or consistent during this time period. However, on after March 6th, there is a noticeable and sudden change in the slope. The rapid increase can be attributed to the outbreak or COVID-19 that ultimately caused a higher number of deaths. After the increase, it seems as though the slope stabilised at a new, higher level suggesting that a large number of deaths within the population. This indicates that the pandemic had caused a sustained and ongoing issue which contributed to the higher death toll.

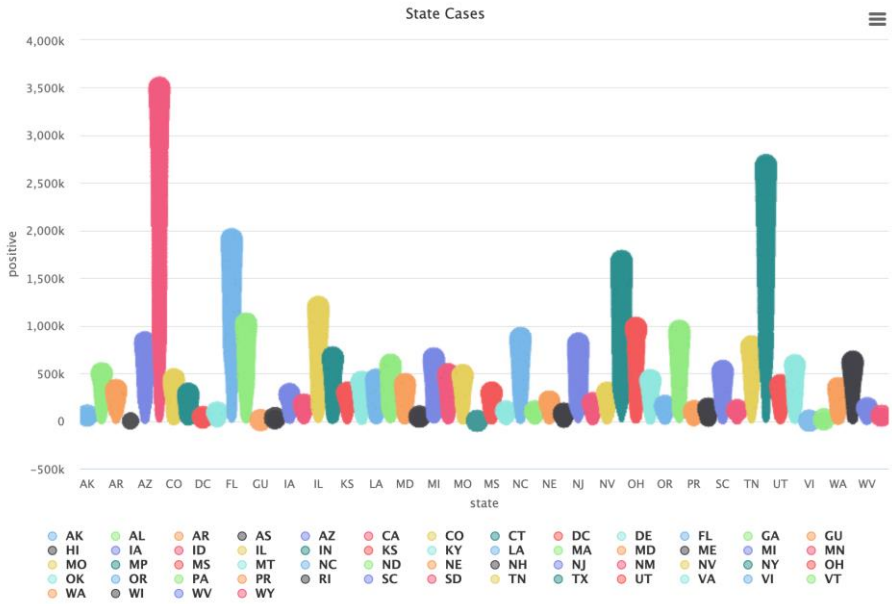


Figure 4: Here, we have a scatter type plot that shows the number of positive COVID-19 cases per state. As per the graph, it is noticeable that the state of California (CA) overtakes the rest of the state as it is largest recorded case state.

Distribution of Data

Skewness

The data skewness helps determine if a distribution lean more towards one end rather than another. In this context, it can be seen that a higher concentration of cases was found in California, Texas, Florida, and New York. This indication of a positively skewed distribution is likely due to the higher density of the population which would mean that the virus would be spread more quickly. Considering this, we are able to say that the negatively skewed values are when there were not as many cases recorded.

Central Tendency

The central tendency represents the mean, median and mode of the cases. It is more or less an average amount at where the cases sat at throughout the whole of the United States. The mean would represent the average as a whole, while the median represents the middle value and mode represents the most common value. Each of these averages are important as it gives us a rough idea of what the cases were generally like.

Outliers

Outliers are extreme values that are significantly deviated from the rest of the data. In the context of COVID-19, the outliers may represent extreme highs or lows in the number of cases recorded per day. Outliers could also indicate data quality issues such as discrepancies or unique events that

caused those certain spikes. In order to ensure that our data has integrity and that these outliers are not skewing out results positively or negatively, we must identify and analyse their value.

Range

Having a range indicated the magnitude of the cases in the United States. We would take the maximum and minimum away from each other to gain a simple measure of the spread of cases across the country. Should there be a larger range, it would indicate a wider dispersion of cases.

Spread

Having a spread of data indicates the variability of our data points. Measures such as standard deviation or the interquartile range is how we would be able to visualise this data. In the context of COVID-19, it would help us assess how widely the cases were distributed across the different time periods and states.

Task 3: Preparation and Pre-Processing of Selected Data

Commented [MH7]: add kdd thing

Commented [ZS8R7]: Yeh ik I had it in mind

Commented [ZS9R7]: Done

Visualisation Before Data Preparation

Data processing and preparation refers to the set of activities that are performed on raw data to transform it into a format that is suitable for analysis. The goal is to ensure that data is accurate, complete, consistent, and formatted in a way that can be easily used for analysis. This is important because it can have a significant impact on the accuracy and reliability of the results of data analysis. By ensuring that data is accurate, complete, consistent, and formatted correctly. As mentioned in the KDD architecture for data mining that the next step after selection of the data is data cleaning and pre-processing which is very crucial to be done right to ensure the quality if the analysis.

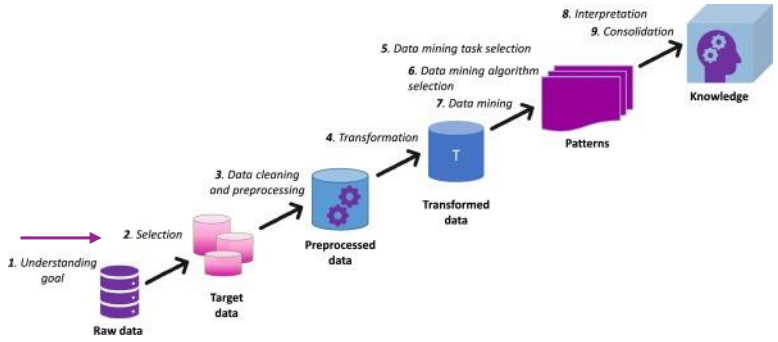


Figure 5: Steps of KDD architecture in data mining.

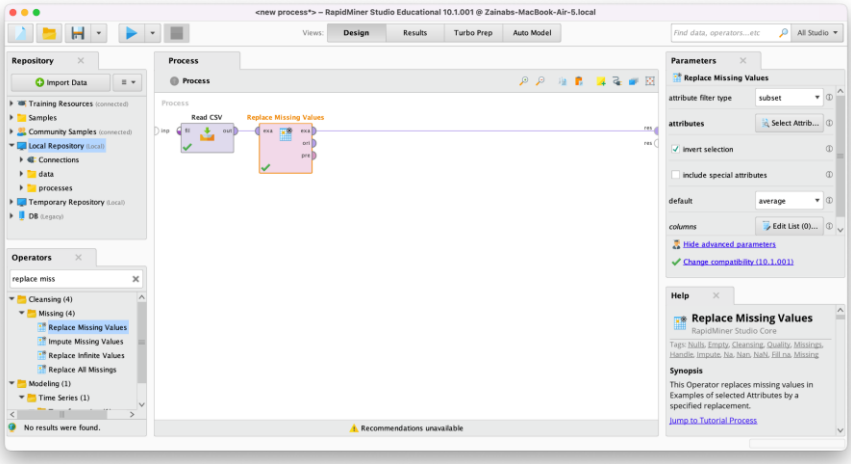


Figure 6: The image below shows the implementation of "Replace Missing values" Operator on the dataset.

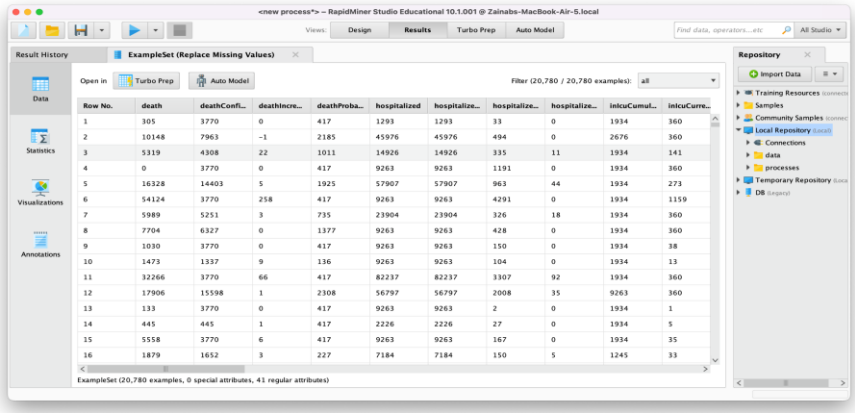


Figure 7: As you can see, we have replaced all missing values with the average of the column to reduce the impact of missing values on the overall distribution of the data.

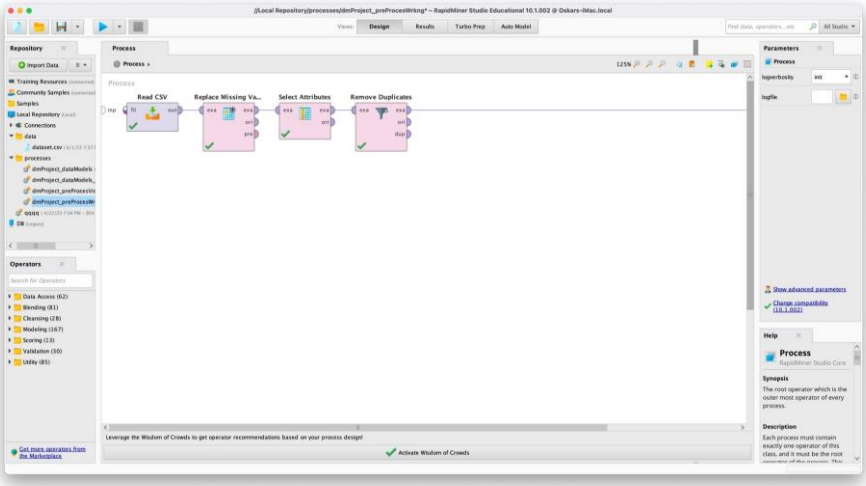


figure 6: We have added the "select attributes" operator to identify the most important attribute to reduce the dimensionality and remove any redundant attributes that are irrelevant to our analysis.in addition to that we have removed some of the attributes manually as well as using the Remove Correlated Attributes operator which will be shows in later figures, as well as the "Remove duplicates" operator to reduce duplicate records that may impact our analysis in a negative manner. Adding both the operators increases the quality of the analysis.

Commented [MH10]: here explain that we removed some of the attr from the list manually in addition to using the other operators (the remove corelated thing)

Commented [ZS11R10]: Done

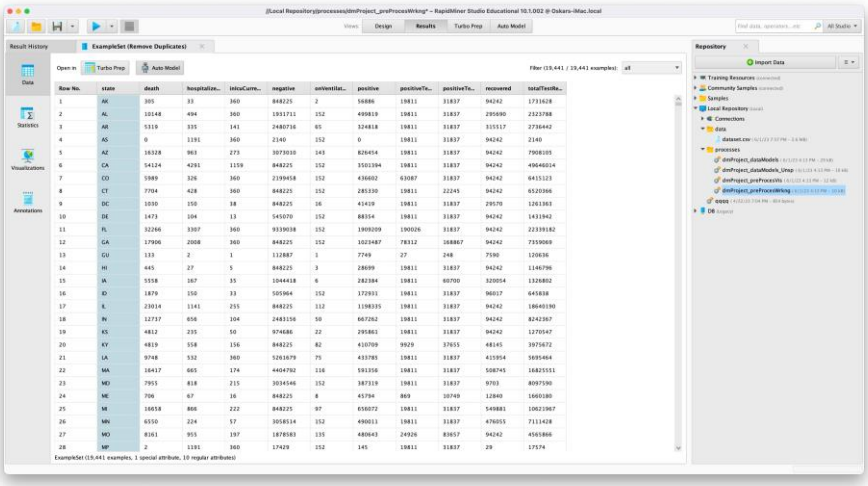


Figure 7: As shown above we can notice that the instance number went down from 20,780 to 19,441 records due to the removal of data. Also note that there was no use PCA as its not needed and that the manual removal of attributes did the job as well as using the Remove Correlated Attributes operator which will be shows in later figures.

Commented [MH12]: mention no pca was used as we dont need it and justify that our manual attribute select is good engh and re mention the manual reduction + remove correlated attr thing

talk about normalization and explain that we use it cuz if not it will take forever to run any operatoion.

Commented [ZS13R12]: Done

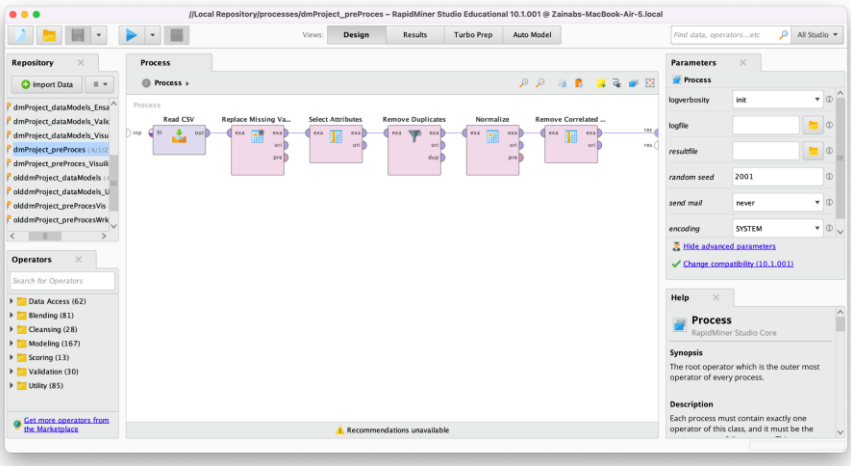


Figure 8: We can see the use of the normalization operator using Z-transformation method to reduce the range of data as if kept without normalization any operation will take lots of time to run, and the **Remove Correlated Attributes operator to decrease the risk of overfitting.**

Commented [MH14]: add some photos of the plots after then bring some plots before the process and compare how they look - they should look the same almost

use a box plot b3d and mention how we can draw a basic outline that may lead to the petronel clusters.

Commented [ZS15R14]: Done

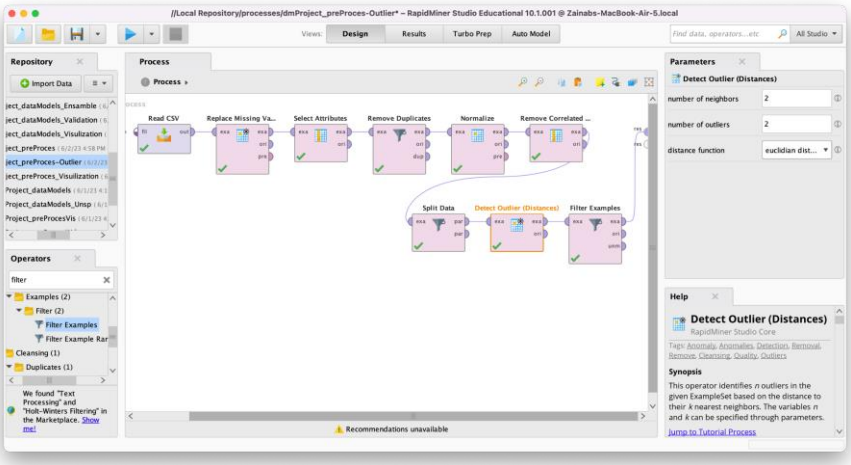


Figure 9: This figure shows that we used the detect outlier operator using Euclidian distance as the distance function as outliers can be error in data entries so to keep the quality of the analysis, we remove them using the filter operator shown in the next steps.

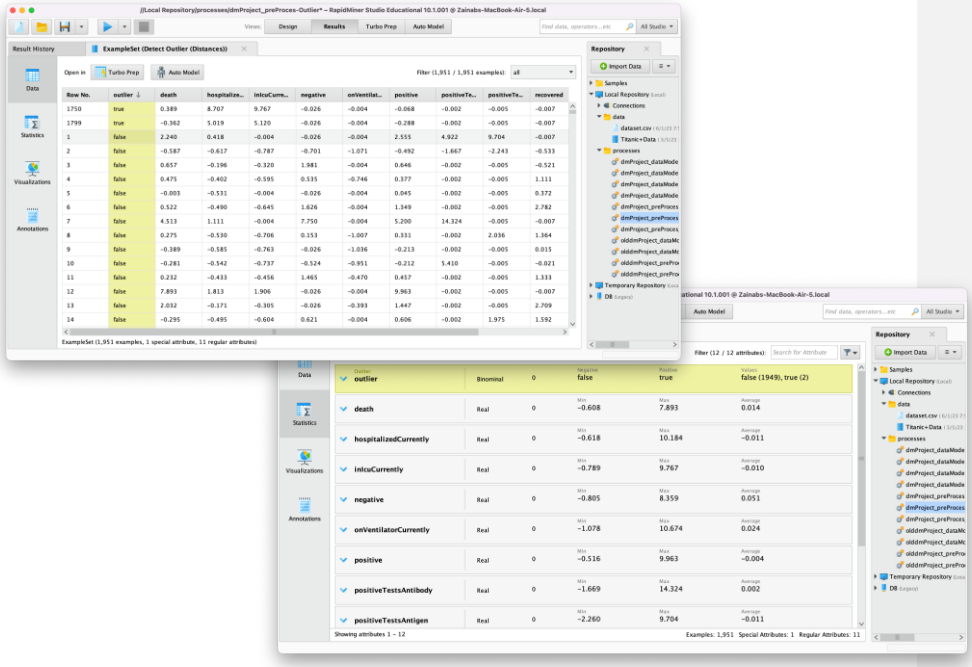


Figure 10: Here a statical representation of the outliers detected before removal as the records with the true value for outlier is an example of outliers detected.

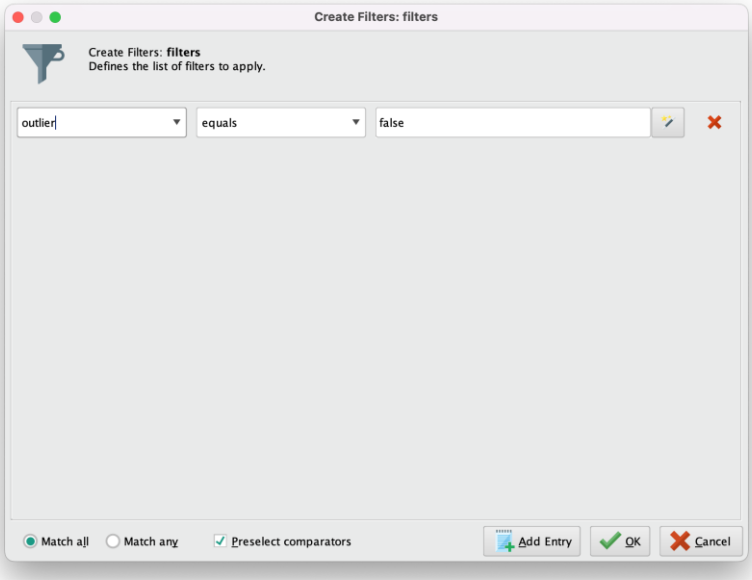


Figure 11: The figure shows the use of the filter operator as we specified the condition is if the outlier is equal to false it will keep the record, therefor any value of outlier returning true will be removed.

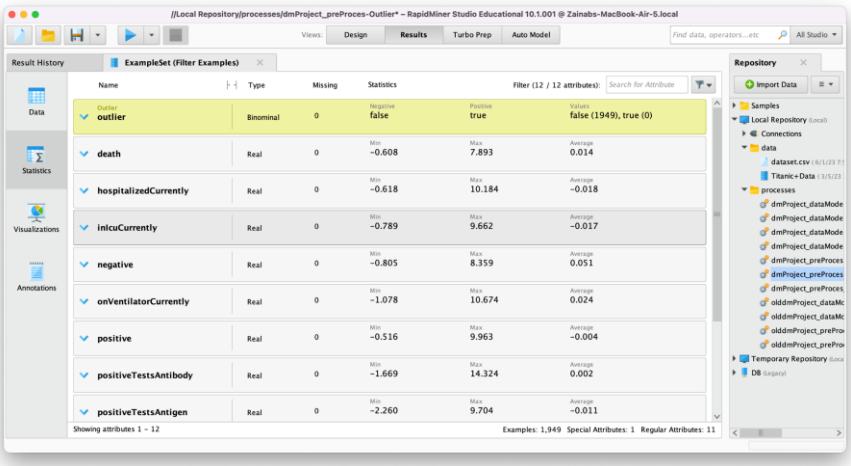


Figure 12: Now we see we have 0 true values for outliers.

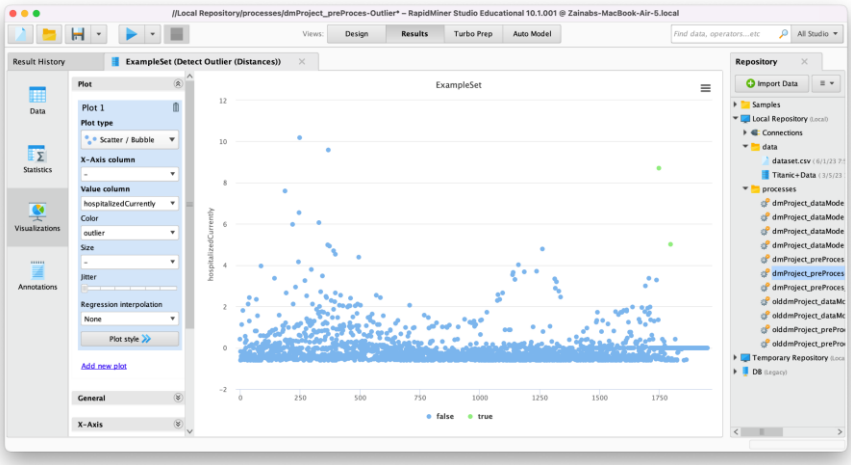


Figure 13: Before the removal of outliers

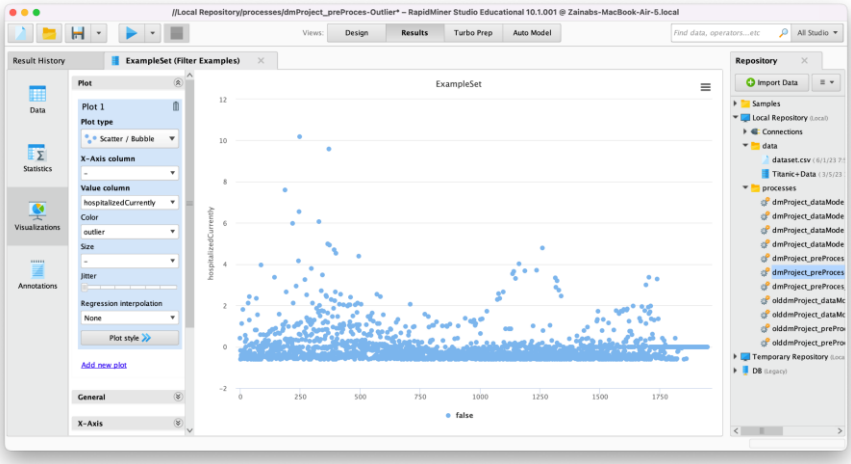


Figure 14: After the removal of the outliers

Visualization After Data Preparation

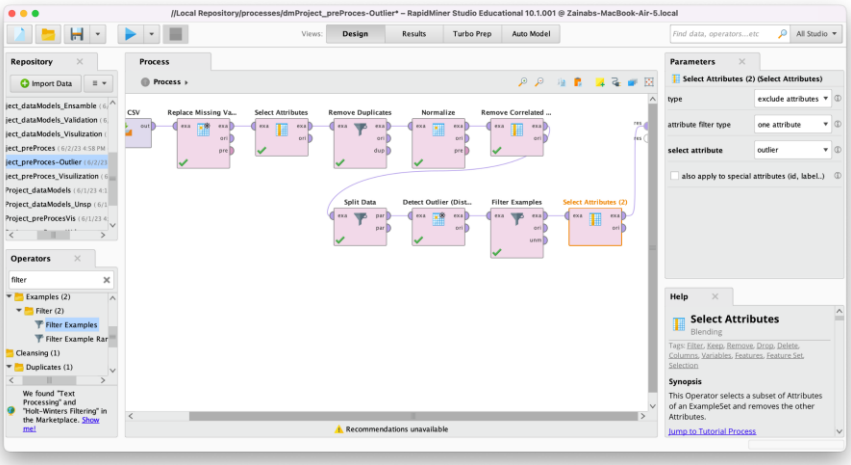


Figure 15: We have added the select attributes again to refine the result and remove the outlier attribute as its not needed.

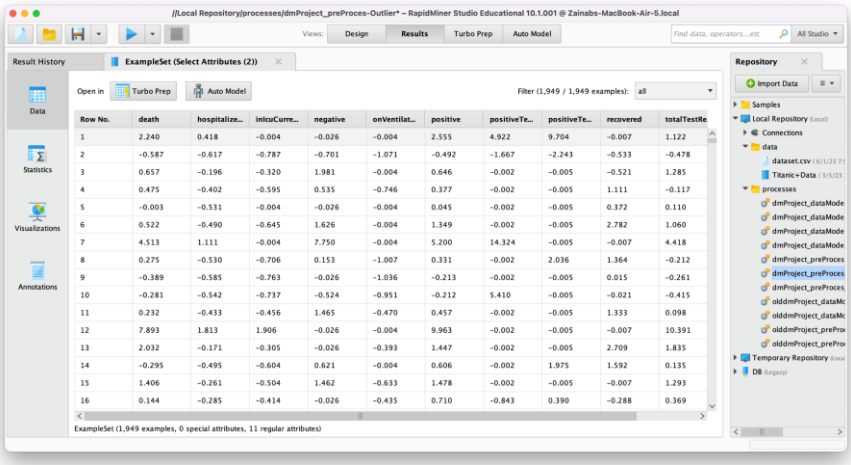


Figure 16: Here we have the final data after the pre-processing.

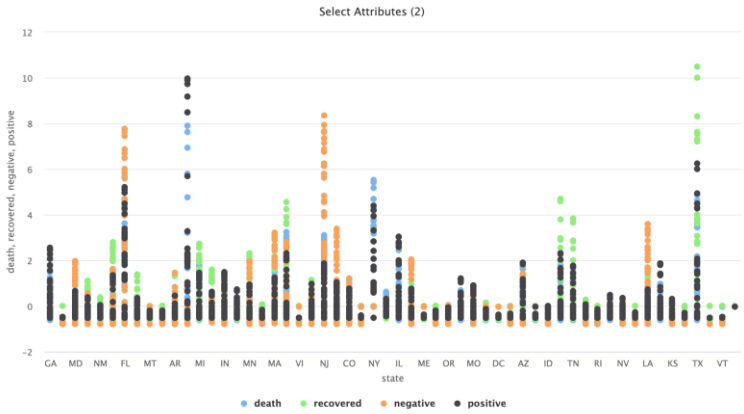


Figure 17: Data before cleaning. Showing Many inconsistencies, outliers, and errors.

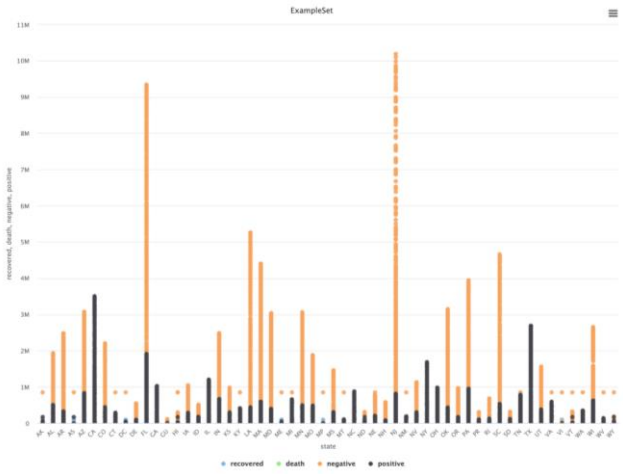


Figure 18

Pre-processing data is an extremely important step when it comes to data analysis. It transforms raw data into a suitable format in order for data analysts to make sense of the data. Usually, raw data contains errors, outliers, and mismatched data. As per our data, we are able to see that it contains a lot of noise, missing values, and inconsistences. If we were to use this data, it would create an analysis that is inaccurate.

After applying data cleaning techniques, normalisations and redefining the attributes, our data is now refined and standardised. This makes it much easier to analyse as we are able to now understand the data. We are also now able to identify the patterns and are now able to compare the datasets with each other. We now also have a dataset that is reliable and trustworthy enough to

make informed decisions. We are able to clearly see the abouts of deaths, recovered cases, positive cases and negative outcomes in each state. This can all be seen in figure 19, located below.

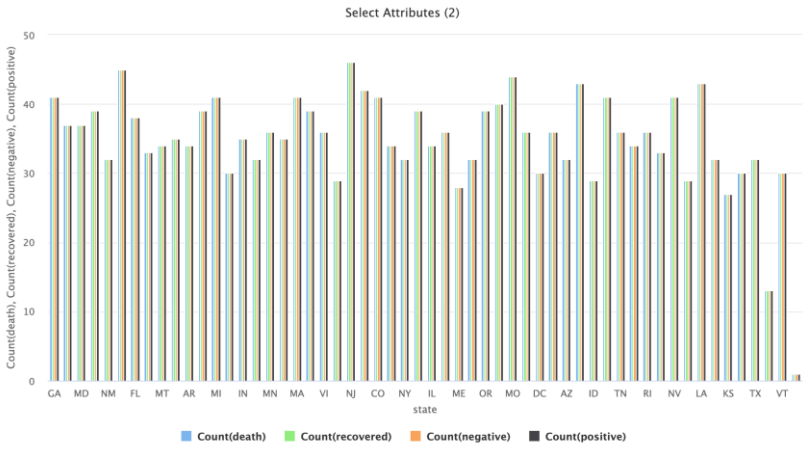


Figure 19: Showing the data after cleaning.

Task 4: Building Data Mining Models

We have used the hybrid way as our data is unlabelled, for the unsupervised we have used both clustering and k means, and x means and random clustering to group related data based on their characteristics and since our data is unlabelled and to build base learners for the prediction models, instead of splitting the data we have implemented cross-validation in order to get the most accurate and precise results, also we have determined k by trial and erroring as we tried with square rooting the number of elements and it produces more 130 clusters which isn't feasible as we have 56 states only.

We varied in our trial-and-error approach from 2 to 7 being the square root of the states, which allowed us to inspect and determine which number neighbouring elements does not lead to overfitted or underfitted clusters.

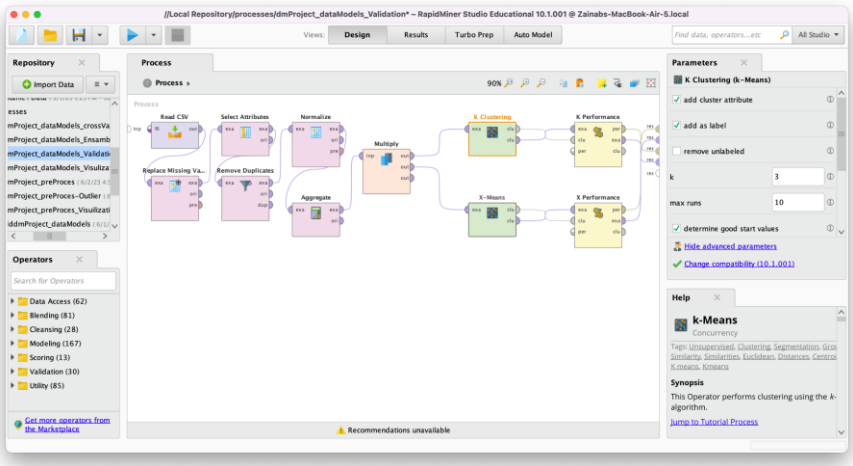


Figure 20: The figure above shows the hybrid method which uses both K Means and X Means. Also used per it measure the distance there is between the clusters, the selected k is 3 after plotting different ranges manually we found 3 to give the optimal results.

Commented [MH16]: we used hybrid, as our data is unlabeled,

for unsupervised talk about the clustering and the 2 techhnics
k means and x means and random just for showing the data and add some nonsense

explain that in k means we used triel and error for the k (by testing and comparing with the normal plots)
mention we did this to ensure no overfitting and shit

and in x it is determent automatically
by a range of minimum to sqrt of n
sqrt(56)

56 being the states plus the occupied lands of us

for the supervissed learning mention that after clustring we use the labels of the clusters to train a knn model and mesure the accuricy of each one

best being k
then x
then random, cuz its random

and then mention ensamble moodle which mixes the stuff and the same situation as above

Commented [MH17R16]: mention that our k is 4

Commented [ZS18R16]: Done

Commented [MAJAH19]: Mention how we compared 3 models and why

Commented [ZS20R19]: Done

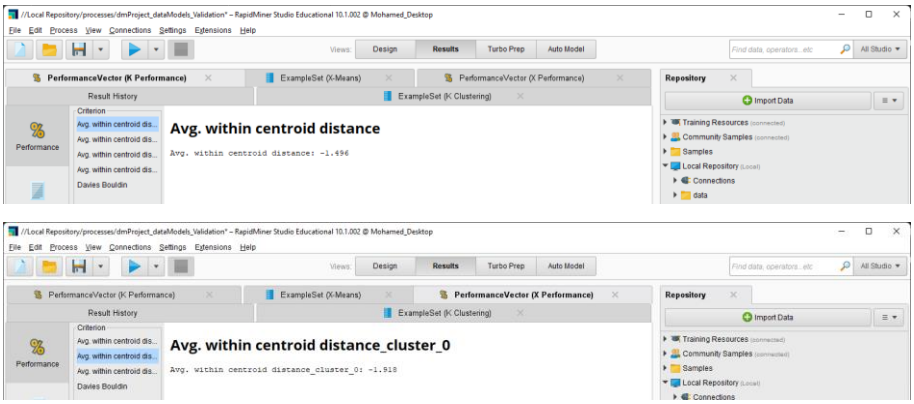


Figure 21

We have k performance centroid distance where the distance withing centroids is: -1.496 and -1.918, as an example of how we selected which one is the best as you can see in the k the distance between the k clusters is less than the x clusters which is why we choose k means over x means as it represents a more realistic number of clusters which overfitting g doesn't exist in this scenario.

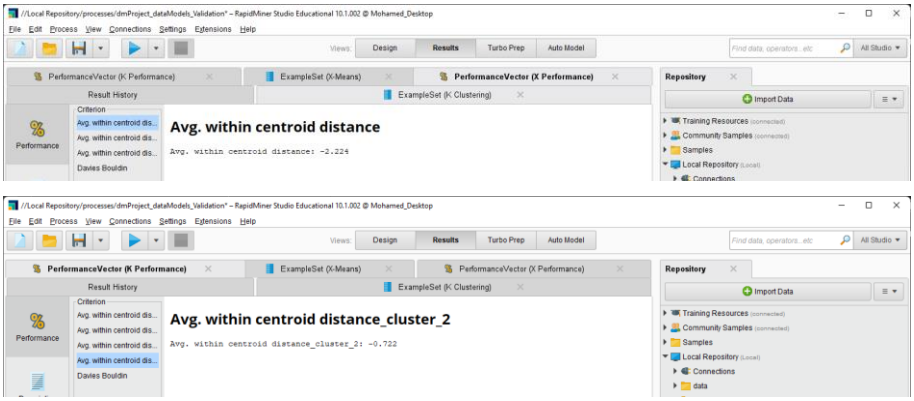


Figure 22: We have X performance centroid distance where the distance withing centroids is: -2.224 and -0.722.

Visualization of Model One

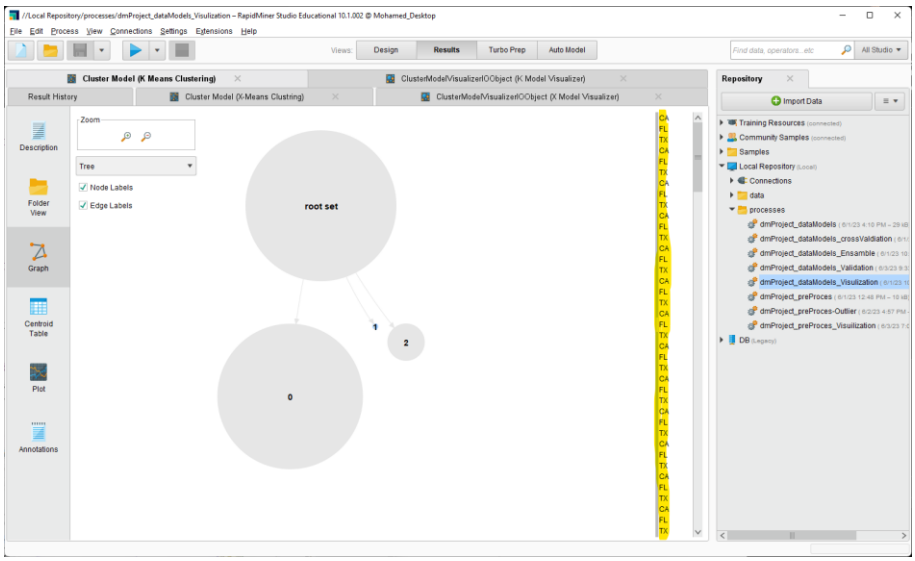


Figure 23: For the k model visualization we notice that we have 3 clusters.

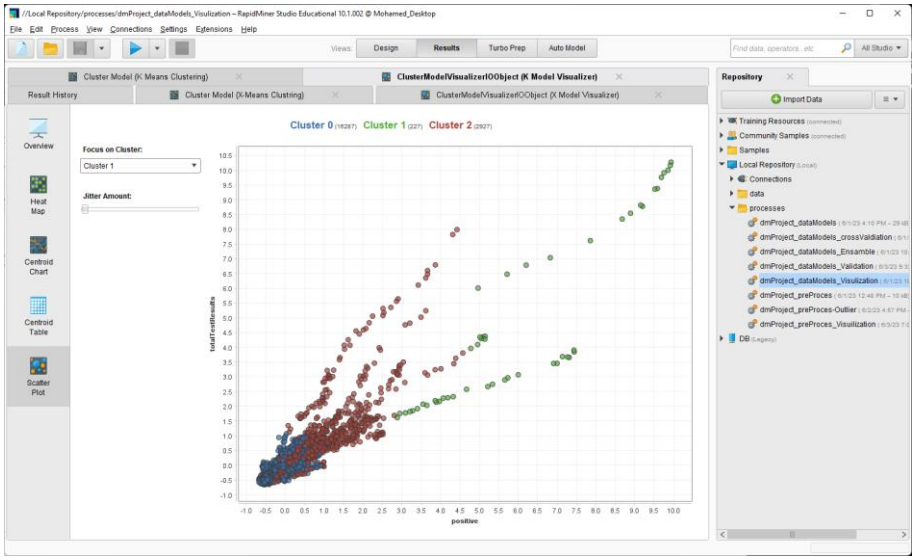


Figure 24: Here, we notice that the distance between the clusters is more realistic than the X means way hence why we went ahead with this methodology.

Visualization of Model Two

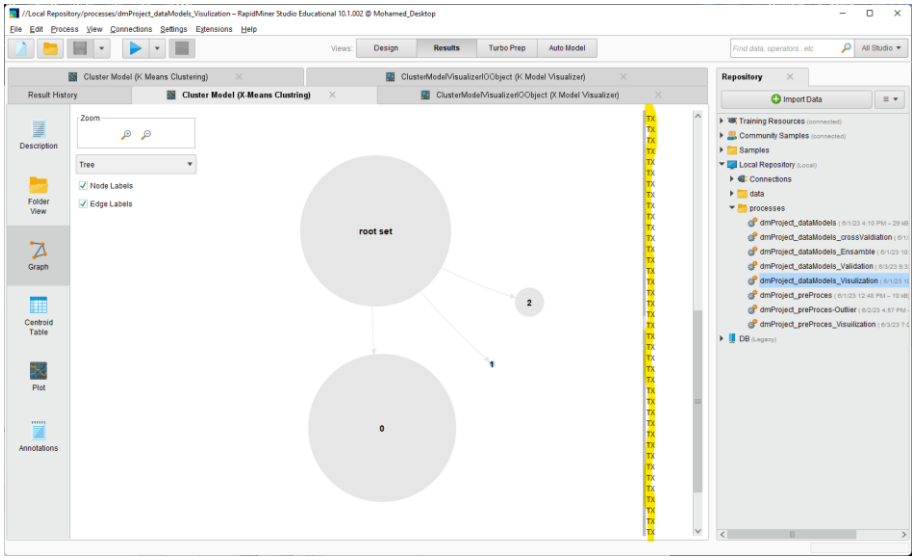


Figure 25: For the x model visualization we notice that we have 3 clusters, however we decided to not go with the x means model as it isn't realistic.

Figure 26: here we notice that the distance between the clusters is dramatically more than the k means way hence why we did not go with this methodology.

Supervised Learning

We added random to test that our methodology is correct, as we used the clusters models as base learners for the supervised learning methodology, we use the labels of the clusters to train a KNN model and measure the accuracy of each one best being k means with a value of 3 for k, then X means and lastly random as it is random.

Commented [MAJAH21]: First talk about the way

Commented [MAJAH22R21]: How we did it and why

Commented [ZS23R21]: Done

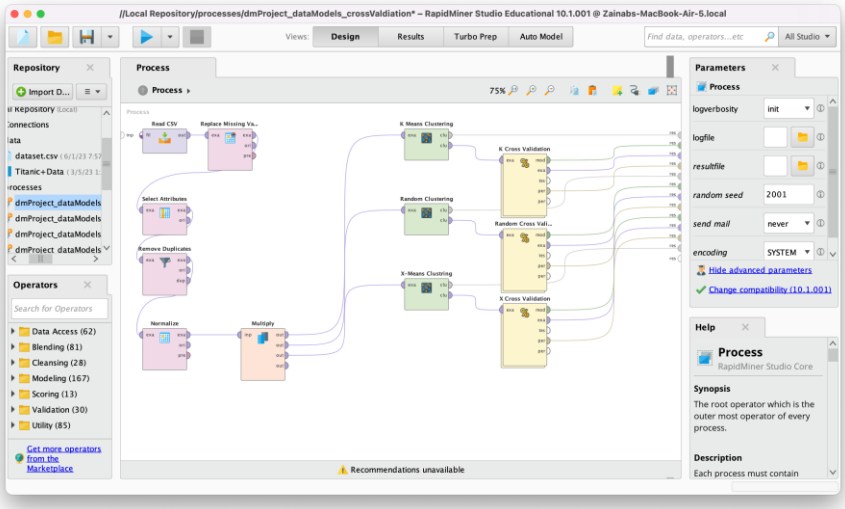


Figure 27: above design shows the hybrid way of supervised learning.

Commented [MAJAH24]: We compared 3 models - adding the random to see if our testing methods are accurite

Commented [ZS25R24]: Done

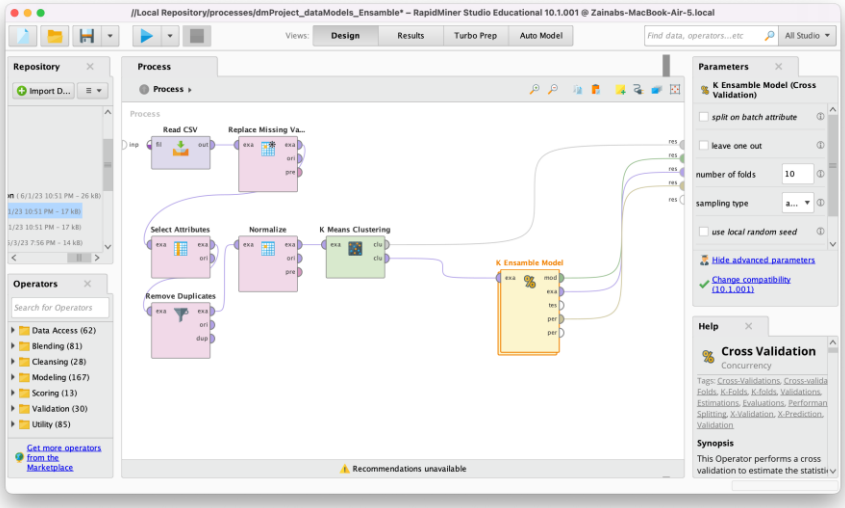


Figure 28: The idea behind assembling is that by combining the estimates generated by different models, the final predictions will be substantially more accurate and precise than those of any single model, above we have the ensemble model by k means.

Visualisation of Ensemble Model

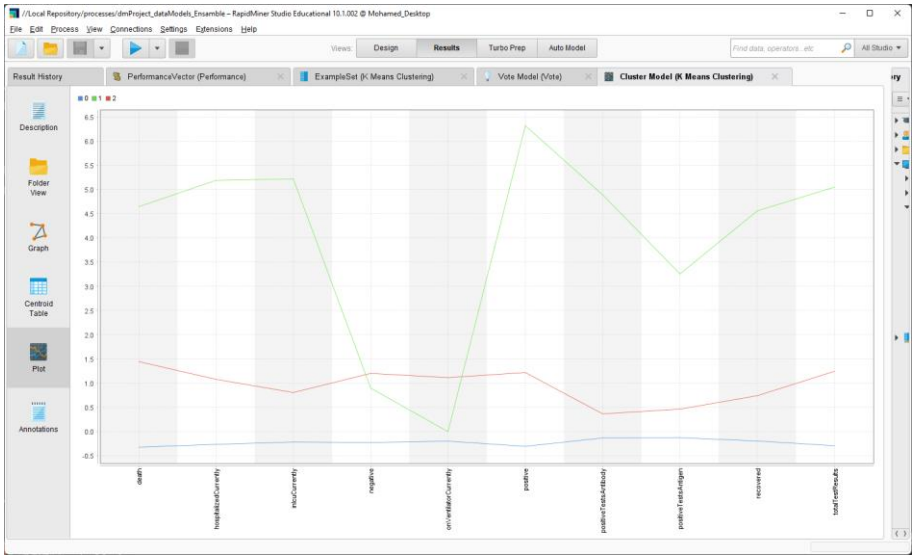


Figure 29: The Visualization shows the result of all ways of data modelling via Ensemble Model, and we can see that k means (green line) has the most efficient and effective output so far.

Commented [MAJAH26]: Add vis of ensamble

Commented [ZS27R26]: Done

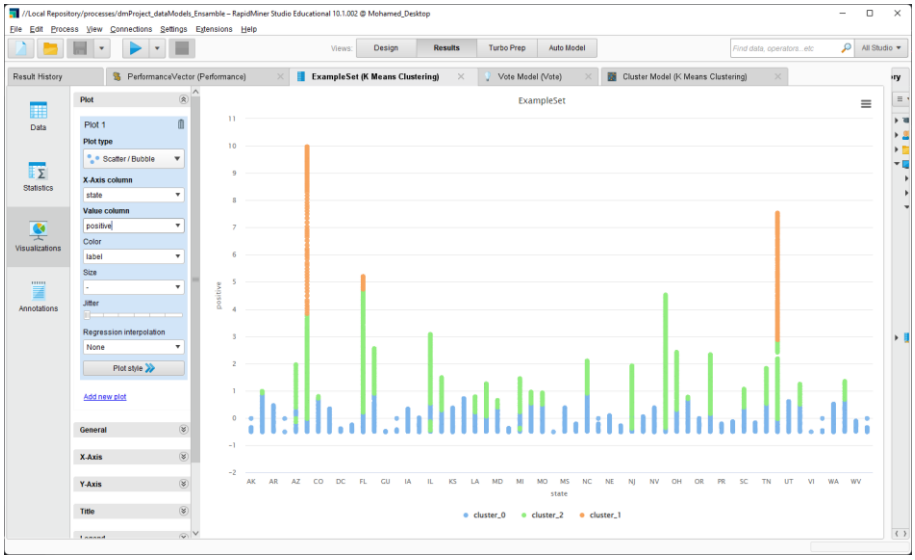


Figure 30: As shown above for the ensemble model we have merged all three ways of modelling, and we can conclude that it matches the expected clusters visualisation.

Unsupervised Mode Parameter Setting Adjustment

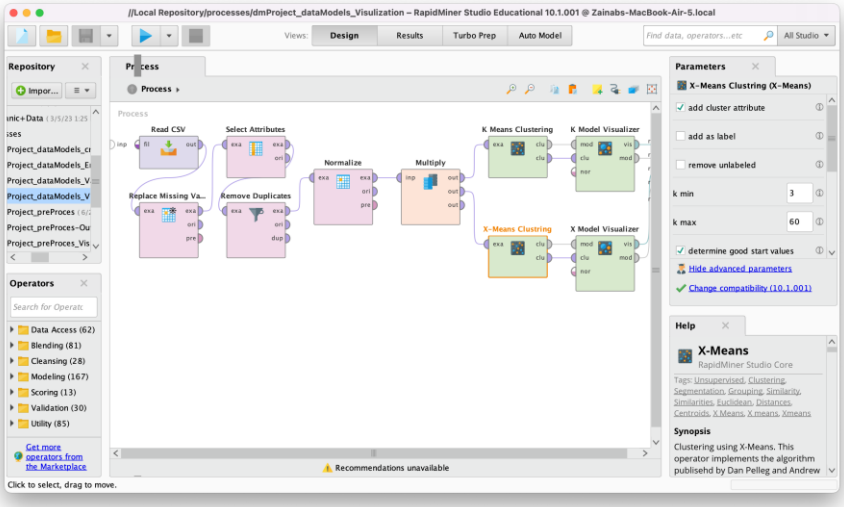


Figure 31: the figure above shows the parameters of the unsupervised model where for the X means we entered the k min as 3 as that the minimum we will accept and 60 as max as we don't want the number of clusters to exceed the number of states in the dataset.

Unsupervised Mode Parameter Setting Adjustment

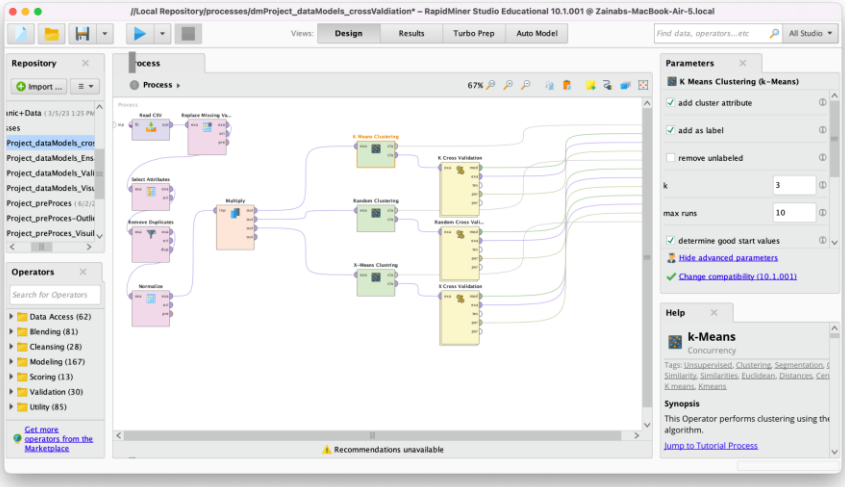


Figure 31: the figure above shows the parameters of the supervised model. The approach could, we changed the max runs to 10 so it doesn't become computationally expensive and take longer to converge if the maximum number of runs is set too high, without necessarily enhancing clustering effectiveness.

Ensemble Mode: Parameter Setting Adjustments

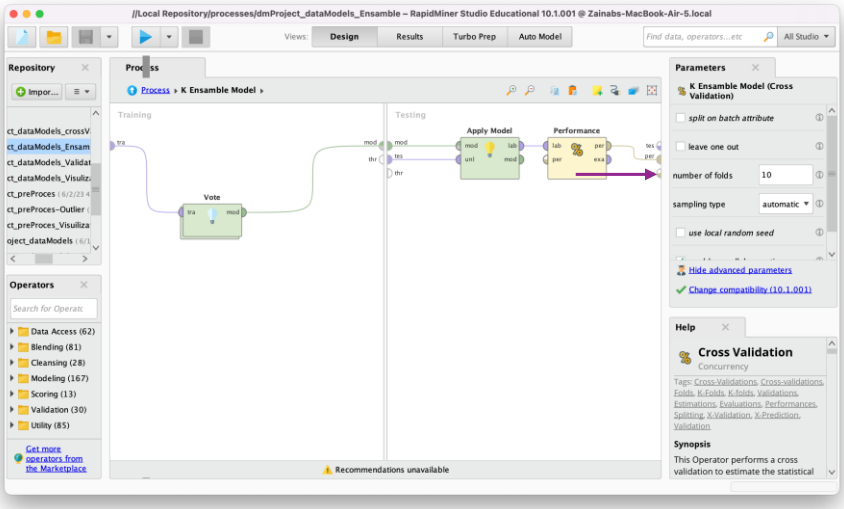
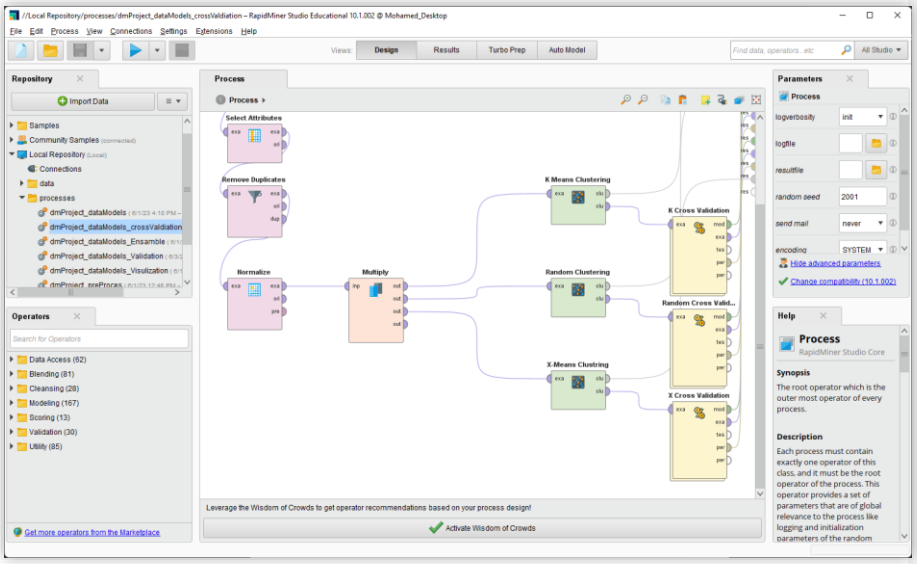


Figure 32: the figure above shows the parameters of the cross validation in ensemble model as we changed the default value of folds to 10 after many trials and error as it doesn't consume a lot of computational resources nor little where we don't get accurate results.

Task 5: Evaluating Data Mining Models

Evaluation Metrics

Since we went with a hybrid approach, our clusters were verified to be of useful means using supervised learning methodologies. The process of verifying and evaluating the results included using the clustered models (un-supervised learning models) as base learners for the “KNN” algorithm, and the ensemble model which includes multiple different algorithms.



K Means Clustering Evaluation

accuracy: 99.85% +/- 0.07% (micro average: 99.85%)

	true cluster_0	true cluster_2	true cluster_1	class precision
pred. cluster_0	16275	15	0	99.91%
pred. cluster_2	12	2911	1	99.56%
pred. cluster_1	0	1	226	99.56%
class recall	99.93%	99.45%	99.56%	

For the K means clustering model it is noticeable that the overall accuracy of the KNN algorithm when compared to the clustering of the base learners is very high (~99.85%) while boosting a very heigh precision and recall rates (~99.45 – 99.91%). This clarifies that the models which were created using the “KNN” supervised learning methods closely match the expected results of the base clusters model. Verifying that the cluster model is valid and robust.

This model clearly represents the base Lerner clusters and robustly resembles the data with minor errors (12 records at max)

Commented [MH28]: 3 things, for the first 2 (clustering) talk about the normal matrix supervised, and the cluster numbers. for the unsupervised

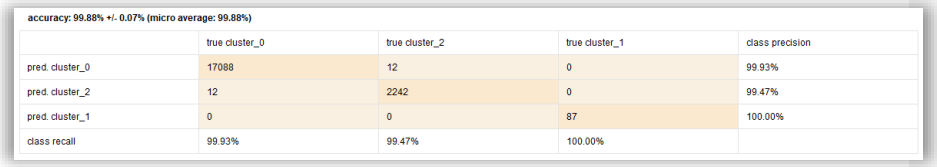
same thing for other one x means

for the ensemble model talk about the supervised learning and add other things im not sure

i think it only needs one per item but explain and add some nonesense

also see what talal group did and copy shwy

X Means Clustering Evaluation



accuracy: 99.88% +/- 0.07% (micro average: 99.88%)

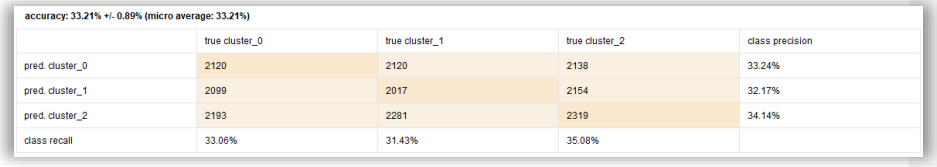
	true cluster_0	true cluster_2	true cluster_1	class precision
pred. cluster_0	17088	12	0	99.93%
pred. cluster_2	12	2242	0	99.47%
pred. cluster_1	0	0	87	100.00%
class recall	99.93%	99.47%	100.00%	

The same situation is prevalent with X clustering, since the algorithm is quite like the K clustering technique, it also boosts a very high percentage of accuracy (~99.88) and a similar but sometimes greater percentage of precision (when comparing to cluster 1).

Again, this model / algorithm is also very robust and highly accurate and boosts a clear picture of precision and recall percentages (reaching 100 percent).

This model as well is very clean and clear, it represents an almost perfect match to the base learners with just a tiny margin of expected errors – as low as 0 records and 12 records at max.

Random Clusters

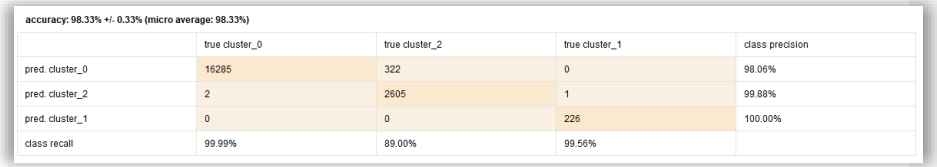


accuracy: 33.21% +/- 0.89% (micro average: 33.21%)

	true cluster_0	true cluster_1	true cluster_2	class precision
pred. cluster_0	2120	2120	2138	33.24%
pred. cluster_1	2099	2017	2154	32.17%
pred. cluster_2	2193	2281	2319	34.14%
class recall	33.06%	31.43%	35.08%	

Finally, the random clustering technique which was used as a sample to verify if the testing algorithm are proper. As seen from the matrix above, the accuracy is 33% at best, which coincides with the fact that 3 clusters are present and are sampled randomly from the whole data set. This model is not used for validating of classifying the data itself, it's just to show how this method of building base models and verifying them with supervised learning gives proper results.

Ensemble Model Evaluation



accuracy: 98.33% +/- 0.33% (micro average: 98.33%)

	true cluster_0	true cluster_2	true cluster_1	class precision
pred. cluster_0	16285	322	0	98.06%
pred. cluster_2	2	2605	1	99.88%
pred. cluster_1	0	0	226	100.00%
class recall	99.99%	89.00%	99.56%	

As stated previously, the ensemble model incorporates multiple different algorithms to produce the most accurate results. Even though the overall percentage is a bit lower than the normal models, that is expected, as multiple algorithms are staged, and the separate clusters are more uniform. As an example, where before we had around 12 records not in the correct cluster for 2 clusters, here most of the clusters have 0 wrongly classified records.

This model could be represented as the most accurate model in terms of the overall classification and is a reliable source of producing studies on.

Task 6: Inferences, Recommendation and Reflection

Commented [MH29]: here we will use some plots and highlight the most notable things plus some بهارات in the mix

sound smart

Commented [OA30R29]: I always do baybeeeeeeeeeee (jk)

Comparison and Contrast

Overall, both main models have produced highly reliable and greatly accurate results, of which we can comprehend the man trend of covid exposure and spread in more than 50 states.

The K means cluster had a small bit of difference in terms of accuracy compared to the X means algorithm, but the algorithm has produced more readable and understandable results which were used for inference.

Using the data that was analysed. It is notable that both models (K means, X means) have classified the states into 3 different categories, the first being the states which were the worst in terms of numbers regarding covid stats (California and New York), this coincides with the real-world expectancy as it is known that these 2 states have a high density compared to others.

However, this cluster represents the lowest set of data. Other clusters have middle or average scoring states, which on average were similar in terms of cases, deaths, recovers, and even antibody test.

Moreover, the last cluster represents the “below average” states, which also is quite tiny in comparison to the “average states” cluster.

On the other hand, the testing algorithm (random cluster) produced very inconsistent and unreliable results, which is as expected acknowledging the fact that this algorithm has no bases to split the data.

Using the information above and the cluster data we could infer those countries with higher densities have a greater chance or higher exposure to any pandemic or health related outbreak. It can also be noticed that bigger states (such as Texas) have lower spread compared to others.

Recommendation

View the analysis that was produced, we can conclude that states or countries with higher populations should implement stricter and wider regulations to combat the spread of any pandemic. Judging by the fact the states were late to implementing curfews and other covid regulations, it is recommended that any country with high densities should implement higher protocols and measures before the spread is uncontrollable.

In addition to that, it is also recommended to keep a steady flow of data to track such outbreaks, as noticeable from the data set, in the beginning stages of the outbreak, data entries about cases were few and far between, this decrease the chance of finding the spread in it early stages before it is able to become a full outbreak.

Reflection

it made us aware that pre-process is important as half of the task won’t even run if the data wasn't pre-processed - example outlier detection and clustering without normalization is hell.

Throughout this project, we had learnt so much about data analysis and the importance of data cleaning. In this reflection I will be discussing sampling data, its relevance, insights that we are able to get from it, and its impact on us understanding the importance of data pre-processing.

Sampling data is such a crucial technique used in data analysis. When one has a large dataset, it is practically impossible to actually make sense of the data. If we were to work with the whole dataset from the larger population, it would be extremely time-consuming, expensive and a massive waste of our resources. When we get a sample, it allows us to select a subset of data points to represent characteristics, in this case it allowed us to see the true impact of COVID-19 on the United States of America. Sampling this data allowed us to gain valuable insights without sacrificing the integrity and accuracy of the data.

When we sample data, it allows us to work with a more manageable dataset. It permits us to be more efficient when cleaning the data, it also allows us to make statistical inferences more accurately and reduces the risk of having biased data. This is one of the reasons why we chose to have our dataset based on one country. It made sense to make assumptions and evaluate data about just one country (which had different states that handled it differently from within) as they had the same basic rules. Some states introduced stricter rules while some were only following basic guidelines. This can be reflected in the number of cases that a state has.

Clean and accurate data supports us greatly when we want to make reliable insights, informed decision making, efficient problem solving and processes and trustworthy predictions. Generally, it is the most crucial step when you want to solve complex problems or make better decisions. If we don't clean our data, it would result in errors due to the outliers, missing values, biases, and inconsistencies being considered when we make those important data driven decisions. It renders our decisions untrustworthy and inaccurate.

This project essentially allowed us to become keener on the importance of data. It made us more aware that pre-processing data is a crucial step (as mentioned and shown in previous sections). Had we normalised our data, most of the operations we ran, such as outlier detection and clustering, wouldn't have been able to run. Even if it had, it would be entirely inaccurate.

Miscellaneous

Log Files

Data Mining Project Preference Document		IT8416		
Mentor: Sini Raj Pulari				
Project Name		**Impact of COVID-19 on Health Sector**		
Group No: 4				
Student Member	**Student ID**	**Student Name**		
1	**Zainab Salman**	**201900658**		
2	**Mohamed Hasan**	**202002789**		
3	**Oskar Alkhateeb**	**201900615**		
Data Mining IT8416 Project				
Date	**Week No**	**Task Name**	**Work Done During Week**	**Issues Experienced (if any)**
12/3/2023 - 19/3/2023	1	Project discussion	Discussed various ideas and topics to be covered and searched for possible data sources and project roles.	-
20/3/2023 - 27/3/2023	2	Project Planning	Evaluated required tasks and split the list of requirements in addition to formatting a timeline to finish the project in a timely manner.	still not sure of some task requirements and needs - will be clarified as course lessons commence
28/03/2023 - 4/4/2023	3	Project Applications	During this week, we worked on task one of the project. As it didn't require any of the models, we came together to formulate an idea on what to write.	-
21/4/2023 - 28/4/2023	4	Project Applications	Throughout this week, we started on the models and other parts of the tasks that did not require the models. This was to ensure that we were able to continue working within our project timeline.	-

29/04/2023 - 6/5/2023	5	Project Applications	During this week, we were coming to the end of our models. We had a few issues that we managed to sort out on our own. We continued to work on the text for each screenshot of the model.	-
7/5/2023 - 15/5/2023	6	Project Applications	During this week, we finished all the models and the explanations for them. We also started a plan for the video.	We were unsure of a certain aspect of the project. We gained clarification and continued moving on.
16/5/2023 - 23/4/2023	7	Project Applications	During this week, we finished the final few tasks. This included screenshotting all of the models and adding an explanation to each one. It also included finishing off the reflection.	-
24/5/2023 - 31/5/2023	8	Project Video	In this week, we recorded the video for the project. We then looked through it to ensure that it met the requirements of the project.	-
1/6/2023 - 5/5/2023	9	Project Finalization	During this week, we completed the project. During the last two days, we did some fine tuning and cosmetic edits to the project document. We went through our project with the rubric to ensure that we did not miss any of the points. We also uploaded the video.	-

GitHub Link

https://github.com/hamodcraft22/DataMinning

YouTube Tutorial

https://youtu.be/EQBUF3B4nnA

Work Cited

CDC. (2019). *Center of Disease Control and Prevention*. Retrieved from COVID-19: https://www.cdc.gov/coronavirus/2019-ncov/index.html

The Covid Tracking Project. (2021). *The Covid Tracking Project.* Retrieved from The Covid Tracking Project: USA: https://covidtracking.com/data/download