Problem:

In this dataset, I will be attempting to cluster together similar instances of covid statistics from throughout the world. In doing this, I hope to discover how many distinct clusters can be meaningfully distinguished, as well as try to find out what characteristics are shared within those clusters. My hypothesis is that there are distinct clusters separating rich and poor nations, but that there will be some distinct outliers for rich nations that did not implement the safety precautions that they have the resources for, such as the United States.

Clustering Explanation:

Kmeans:

Kmeans clustering is a clustering method where firstly the person has to pick a number, k of centroids, which are basically just center points around which other points cluster. From there, three random spots for the centroids are selected, and the distance from each point to each centroid is calculated. The centroid of the shortest distance is assigned as being the centroid representing that point. After every point is assigned, the mean value of each centroid is calculated and assigned as the new location of that centroid. This process is repeated until no points are reassigned.

Agglomerative Clustering:

With agglomerative clustering, all points start off as their own cluster, then as similarities between points are found, clusters are consolidated down until the danger of overfitting exceeds the utility of fewer clusters. This tradeoff is measured using a dendrogram.

Dataset Introduction:

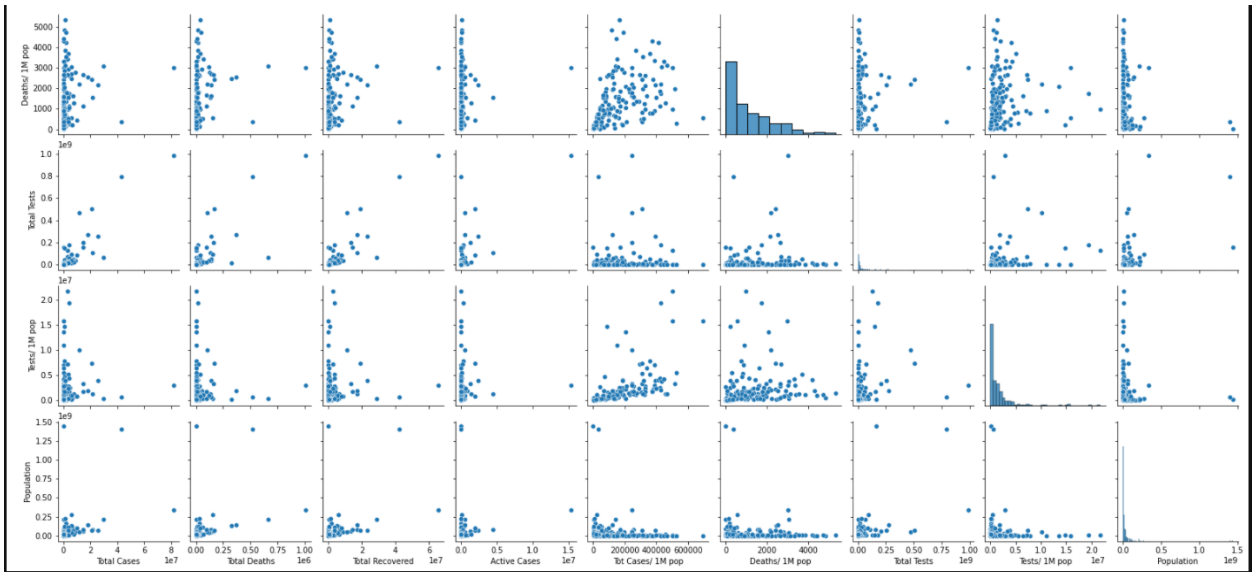
Country - Name of world countries
Total Cases - Total number of Covid-19 cases
Total Deaths - Total number of Deaths
Total Recovered - Total number of recovered cases
Active Cases - Total number of Active cases
Total Cases/1 mil population- Total Cases per 1 million of the population
Death/1 mil population - Total Deaths per 1 million of the population
Total Tests - Total number of Covid tests done
Tests/1 mil population - Covid tests done per 1 million of the population
Population - Population of the country

Link : https://www.worldometers.info/coronavirus/#countries

Data Understanding/Visualization:

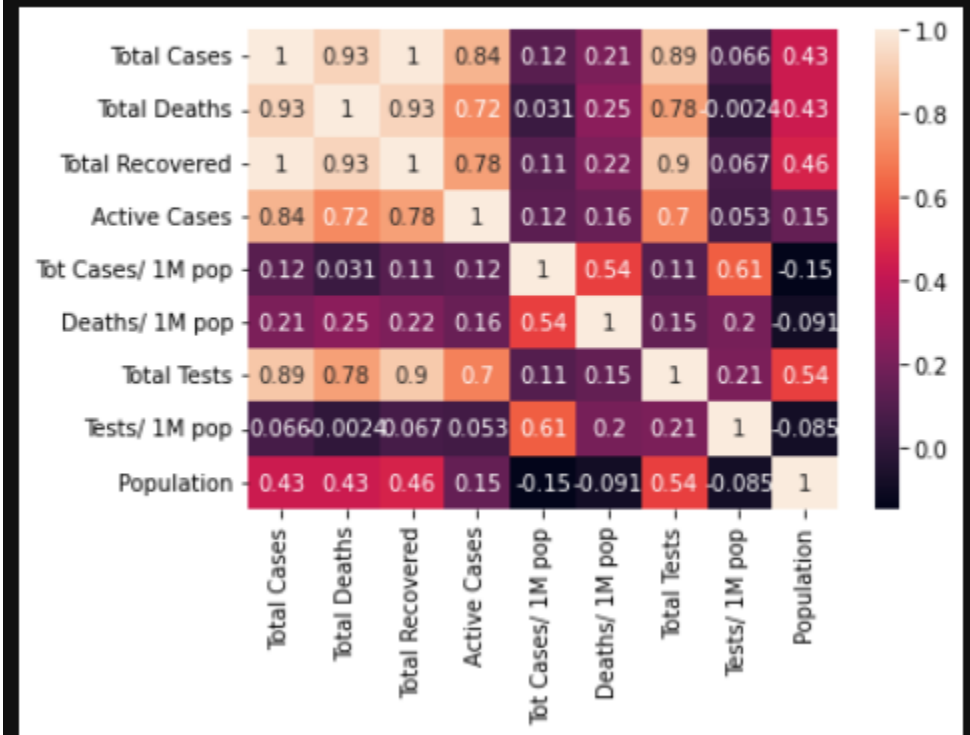
For visualizations, I used a pair plot and a heatmap to get a really broad overview of the relationships between the variables. What we can see between some of our correlations is that there are strong linear relationships between some of the variables, as would be expected. Things like covid per population and death per population are positively correlated for instance. Covid is also correlated with deaths. A lot of no-brainers here.

There is a relationship between tests and cases as well, which is interesting. This makes sense because places with more cases need more tests, but tests are also useful for informing people they need to isolate, which should lead to a lessening of the correlation. A naive perspective would be that covid testing causes more covid, but that is extremely unlikely because tests per population have almost no relationship with total cases.




```
sns.heatmap(df.corr(), annot=True)
```

```
<AxesSubplot:>
```



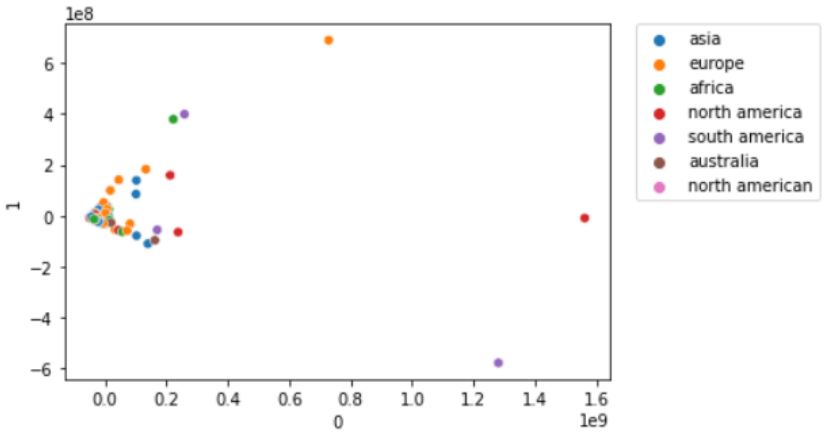
Pre-Processing:

For the first part of the preprocessing, I dropped nulls to avoid errors in the later steps. I also used principal component analysis to reduce the number of dimensions to two, to make visualization possible. The below visualization uses PCA, with the colors being representative of

the different continents, which, since the dataset was originally countries, required sorting all of the countries into continents. Very exciting. The results are underwhelming. Most of the points are clustered together into a big ball with a few outliers, one of which is in North america.

```
In [21]: sns.scatterplot(x = pca_df[0], y = pca_df[1], hue=df['continent'])
         plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left', borderaxespad=0)
```

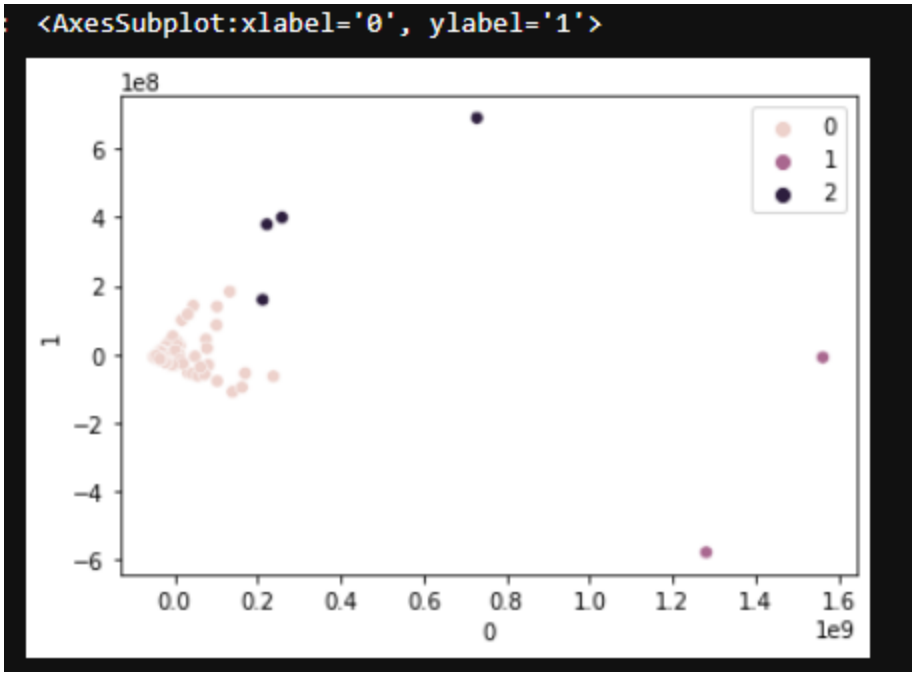
```
Out[21]: <matplotlib.legend.Legend at 0x7f7d3c383f70>
```



Modeling:

K Mean Clustering:

The results for the K Means clustering are a little odd. It interpreted some of the points at around .2 on the X axis as black, whereas it makes more intuitive sense to classify them as the lightest color.

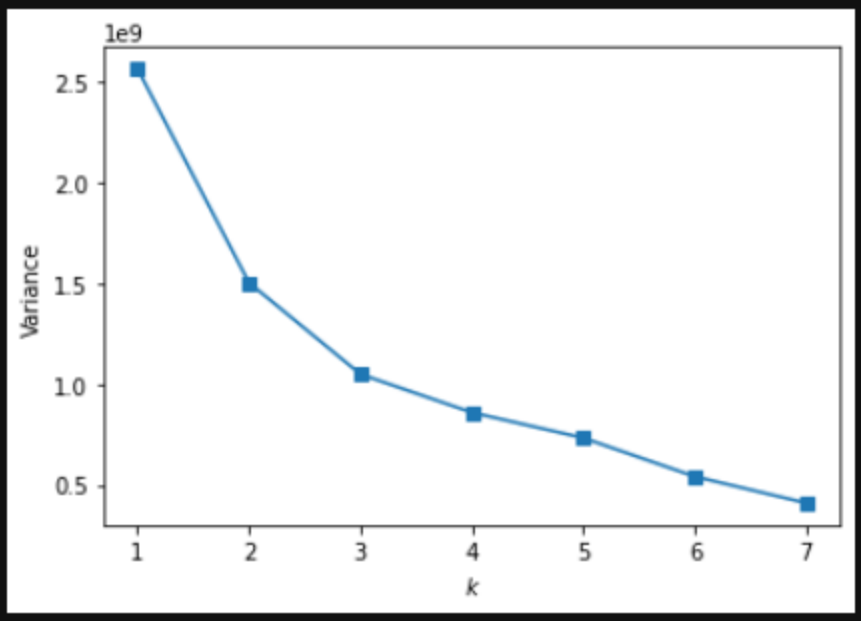


This is a graph of the variance in relation to k. Either two or three is the correct choice, as the rate that the variance goes down goes down to a significantly lower level. I chose three.

```
[22]: inertia = []
      for k in range(1,8):
          kmeans = KMeans(n_clusters=k, random_state=1).fit(X)
          inertia.append(np.sqrt(kmeans.inertia_))

[23]: plt.plot(range(1, 8), inertia, marker='s');
      plt.xlabel('$k$')
      plt.ylabel('Variance')

[23]: Text(0, 0.5, 'Variance')
```

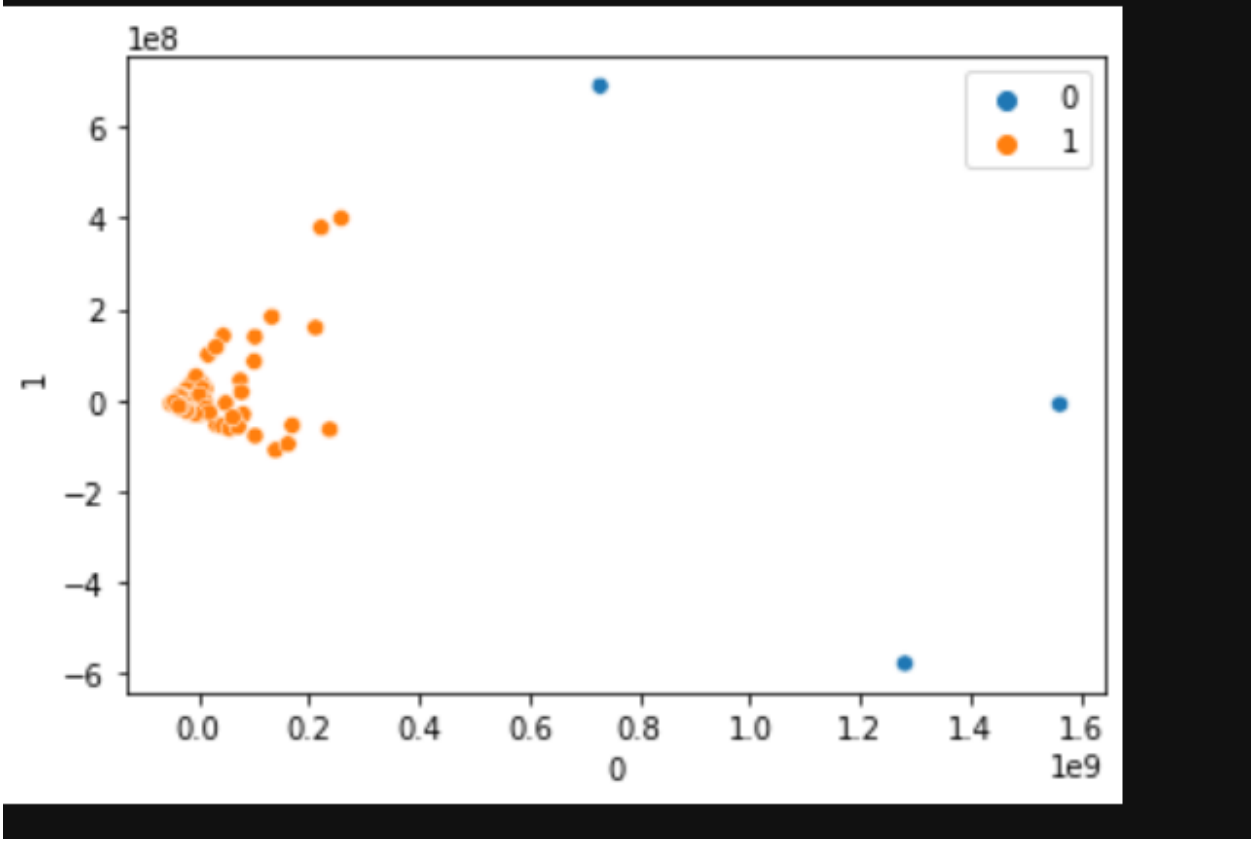


Agglomerative clustering:

The results of the agglomerative clustering are more promising than the Kmeans clustering. The previously mentioned issue was resolved.

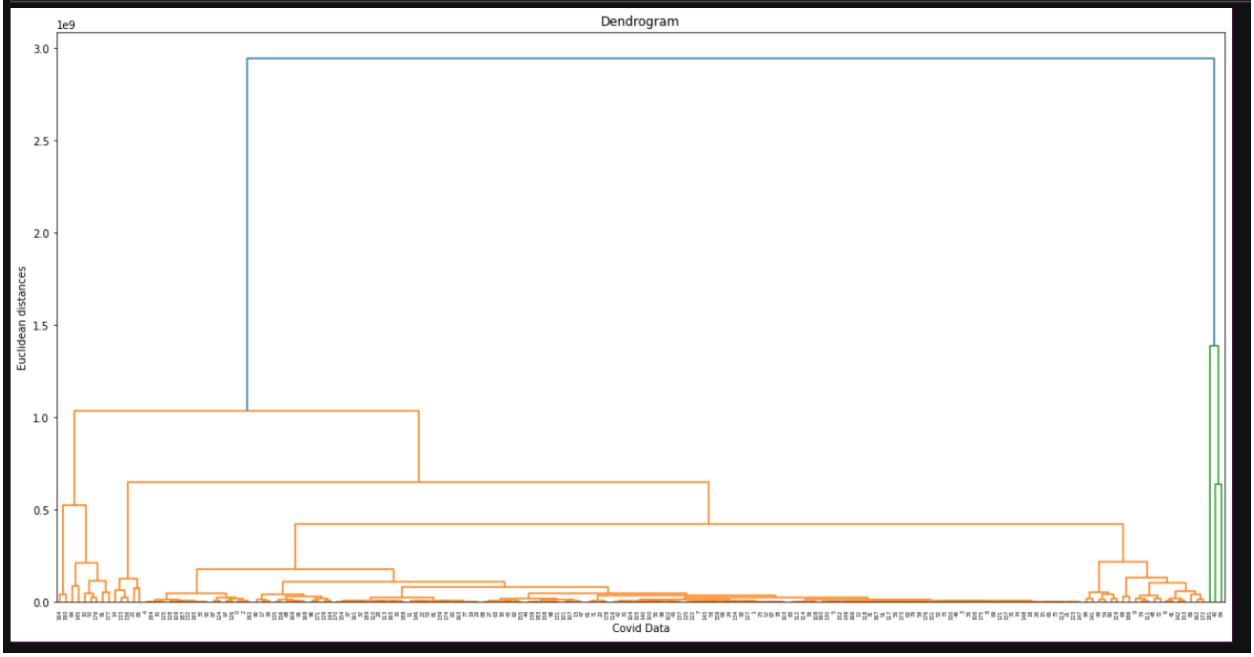
```
sns.scatterplot(x = pca_df[0], y = pca_df[1], hue=y_agglo)
```

```
<AxesSubplot:xlabel='0', ylabel='1'>
```



This is the dendrogram of the appropriate number of clusters, 2.

```
#Create and display a dendrogram
import scipy.cluster.hierarchy as shc
plt.figure(figsize=(20, 10))
plt.title('Dendrogram')
plt.xlabel('Covid Data')
plt.ylabel('Euclidean distances')
dend = shc.dendrogram(shc.linkage(X, method='ward'))
```



Story Telling:

K Means and agglomerative clustering performed quite similarly, with agglomerative clustering performing slightly better. Measurement of performance subjectively based on my own preference, as there isn't any other clear measurement criteria. Using two clusters rather than three seems to be a better decision, and some of the clustering results from KMeans are odd. Several of the dots near the large 0 cluster appear to be mislabeled. I believe agglomerative clustering performed better because the dataset was relatively small, about 200 entries in total. The results indicate that there is one really clear cluster of countries that are all very similar and either one or two more other clusters of countries that are quite varied.

My interpretation of this is that the large cluster is of relatively rich nations, which had access to covid testing and had better production of masks and medicine, and the other cluster(s) are representative of poorer nations. If there are three groups, then there is a gradient in poorer nations, potentially due to some poorer nations being slightly less poor or having cultures more suitable for covid prevention methods. Another interpretation is that the distinction is between rich nations that had the resources to handle the pandemic, but chose not to. A weakness of PCA and clustering is that they are hard to interpret, so it is pretty ambiguous what these clusters are, at the end of the day. At this point, I think that it would be best to try applying other tools to go more in depth than clustering can allow.