**CAPSTONE REPORT ON COVID CASES**

**Presented by:**

**THANKYOU FELIX MUBULA**

**DATA SCIENCE FELLOW**

**FE/23/55996836**

**Fig.1. Heatmap: Correlation of Features**

**Confirmed and Confirmed Last Week:**

These features have a perfect correlation of 1.00, indicating that the number of confirmed cases this week is directly related to the number of confirmed cases last week.

**1 Week Change:**

This feature has a high correlation of 0.95 with Confirmed cases, suggesting that changes in the number of cases over a week are strongly related to the total number of confirmed cases.

**Deaths and Active Cases:**

Both Deaths and Active cases have a correlation of 0.93 with Confirmed cases. This indicates a strong relationship between the number of confirmed cases and the number of deaths and active cases.

**Recovered:**

The feature Recovered has a correlation of 0.91 with Confirmed cases, showing a strong relationship between the number of confirmed cases and the number of recoveries.

**New Deaths and New Recovered:**

These features have lower correlations of 0.87 and 0.86, respectively, with Confirmed cases. This indicates a moderate relationship between the number of confirmed cases and the number of new deaths and recoveries.

This heatmap provides valuable insights into how different features related to confirmed COVID-19 cases are interrelated. The high correlations between Confirmed cases and other features like Confirmed Last Week, 1 Week Change, Deaths, Active cases, and Recovered highlight the interconnectedness of these metrics. Understanding these relationships can be useful for analyzing the dynamics of the pandemic and for making informed decisions related to public health.

**Fig.2. Bar chart: Feature Importance for Predicting 'New cases**

**Top Predictors:**

The most important feature is "New deaths," suggesting a strong correlation between new deaths and new cases. This makes sense as increased deaths could reflect severe outbreaks.

**1 week change:** is the second most important feature, indicating that recent trends over the past week are crucial for prediction.

**New recovered:** follows, pointing to the importance of recovery rates in understanding new case trends.

**Lesser Importance Features:**

Lower on the list are features like "Active," "Confirmed," and "Deaths." While they contribute to the model, their impact is comparatively less significant.

**Confirmed last week:** is the least important feature, suggesting past confirmed cases are not as predictive as recent changes and new data.

**Implications:**

The prominence of "New deaths" highlights the critical need for effective mortality management to control new cases.

The importance of recent trends ("1 week change") emphasizes the necessity for timely data and rapid response strategies.

The role of "New recovered" suggests that boosting recovery rates could be a significant factor in managing outbreaks.

The chart clearly shows that dynamic and recent data (like new deaths and weekly changes) are more predictive of new cases than static or cumulative data. This can help health officials prioritize their focus on current trends and real-time data to manage the spread of the disease effectively.

**Fig.3. Box plots:Checking for Outliers in Data**

These box plots are used to visualize the distribution and spread of the data, highlighting the median, quartiles, and potential outliers. Here is a detailed analysis of each box plot:

Key Insights

**Confirmed:**

Y-axis: Confirmed

The plot shows a large number of outliers, with most data points clustered near the lower end of the scale. The median is very low compared to the maximum values, suggesting that while most regions report low confirmed cases, some have exceptionally high counts.

**Deaths:**

Y-axis: Deaths

Similar to the confirmed cases, there are many outliers. Most data points are near the lower end, indicating that most regions have relatively low death counts, but a few have very high counts.

**Recovered:**

Y-axis: Recovered

This plot also shows many outliers, with most data points near the lower end. The median is low, but some regions report very high recovery numbers, indicating variability in recovery rates.

**Active:**

Y-axis: Active

The distribution is similar to confirmed cases, with many outliers and most data points near the lower end. The median is low compared to the maximum values.

**New cases:**

Y-axis: New cases

This plot shows a large number of outliers, with most new cases being relatively low but a few regions experiencing very high numbers of new cases.

**New deaths:**

Y-axis: New deaths

The distribution is similar to new cases, with many outliers and most data points near the lower end. The median is low compared to the maximum values.

**New recovered:**

Y-axis: New recovered

This plot shows many outliers, with most new recoveries being relatively low but a few regions experiencing very high numbers of new recoveries.

**Deaths / 100 Cases:**

Y-axis: Deaths / 100 Cases

The plot shows a more balanced distribution with fewer outliers. The median is around 5 deaths per 100 cases, with some regions experiencing higher rates.

**Recovered / 100 Cases:**

Y-axis: Recovered / 100 Cases

This plot shows a wider distribution with fewer outliers. The median is around 60 recovered per 100 cases, indicating a higher recovery rate in most regions.

**Deaths / 100 Recovered:**

Y-axis: Deaths / 100 Recovered

The plot shows a large number of outliers, with most data points near the lower end. The median is low, indicating that most regions have a low death rate compared to recoveries.

**1 week change:**

Y-axis: 1 week change

This plot shows many outliers, with most data points near the lower end. The median is low, but some regions have experienced significant changes in one week.

**1 week % increase:**

Y-axis: 1 week % increase

The distribution is similar to the 1 week change, with many outliers and most data points near the lower end. The median is low, but some regions have experienced significant percentage increases in one week.

These box plots are interesting and relevant as they provide a visual summary of the distribution and variability of COVID-19 metrics across different regions. They highlight the presence of outliers and the overall trends in the data, which can be useful for identifying regions with extreme values and understanding the spread of the pandemic.

**Fig.4. Distribution of Daily Growth Rate (%) in Confirmed Cases.**

**Main Clusters:**

The majority of the daily growth rates are clustered around the lower end of the scale, with a very high frequency at the 0% growth rate.

This indicates that most of the confirmed cases have relatively low daily growth rates.

**Outliers:**

There are very few instances of higher growth rates, as evidenced by the sparse bars.

The density plot overlaying the histogram also tapers off quickly, reinforcing that higher growth rates are rare.

**Prevalence of Low Growth Rates:**

The high frequency at 0% suggests effective control measures in place, leading to low daily growth rates for the majority of cases.

**Rarity of High Growth Rates:**

The sparse bars for higher growth rates indicate that outliers with significantly higher growth rates are not common. This could be due to localized outbreaks or anomalies.

**Healthcare and Public Health Implications:**

Understanding the distribution can help in resource allocation. With most cases exhibiting low growth rates, resources can be concentrated on areas showing higher growth rates to prevent further spread.

This histogram provides a visual representation of how the daily growth rates in confirmed cases are distributed, highlighting areas where interventions might be most needed.

**Fig.5. Histogram: Distribution of Cases per 100,000 Population**

**Right-Skewed Distribution:**

The histogram shows a right-skewed distribution, with most data concentrated at the lower end of the x-axis. This indicates that the majority of the population has relatively few cases per 100,000.

**High Frequency at Lower Values:**

The highest frequency occurs at the lowest range of cases per 100,000. This suggests that a significant portion of the population experiences fewer cases.

**Decreasing Frequency:**

As the number of cases per 100,000 increases, the frequency decreases. This pattern highlights that higher case numbers are less common in the population.

Conclusion

The histogram effectively illustrates the skewed nature of the distribution, emphasizing that most individuals experience lower case numbers per 100,000. This visual representation is crucial for understanding the spread and density of cases within the population, aiding in targeted interventions and resource allocation.

**Fig.6. Histogram: Distribution of Mortality Ratio (%) across Countries**

**Skewness:**

The distribution is right-skewed. This means most countries have a lower mortality ratio, with fewer countries experiencing higher mortality ratios.

**Peak Frequency:**

The highest frequency is observed in the lowest mortality ratio range (0-1%), with over 40 countries falling into this category.

**Decline in Frequency:**

As the mortality ratio increases, the number of countries decreases sharply. This suggests that higher mortality ratios are less common.

**Long Tail:**

The long tail extending towards higher mortality ratios indicates a few countries have significantly higher mortality ratios. These are outliers and not the norm.

**Outliers:**

A few countries have mortality ratios above 20%, making them rare occurrences within the dataset.

This histogram gives a clear visual representation of the mortality ratios across different countries, highlighting the prevalence of lower mortality ratios and the rarity of higher ones. It's a powerful way to understand how mortality is distributed globally.

**Fig.7. Scatter plot: Daily Growth Rate vs Mortality Ratio by WHO Region**

**Clustered Data Points:**

Most data points are clustered near the origin, indicating that the majority of regions have low daily growth rates and low mortality ratios.

**Outliers:**

There are a few outliers with significantly higher daily growth rates and mortality ratios.

**Notably:**

One data point from the Eastern Mediterranean region has a mortality ratio above 25%.

A few data points from the Americas have daily growth rates exceeding 100,000%.

**Regional Differences:**

Europe has several data points with mortality ratios between 5% and 15%.

The Americas show data points with both high daily growth rates and mortality ratios.

**Low Mortality Ratios:**

Many regions, particularly in the Western Pacific and South-East Asia, have low mortality ratios, often below 5%.

**High Growth Rates:**

Some regions, particularly in the Americas, exhibit very high daily growth rates, indicating rapid spread in those areas.

The scatter plot highlights areas with higher risks and potential needs for targeted interventions. Low mortality ratios in the Western Pacific and South-East Asia may suggest effective healthcare systems or early interventions.

High daily growth rates in the Americas could indicate a need for more stringent public health measures.

This scatter plot provides a visual representation of the relationship between the daily growth rate of a disease and its mortality ratio across different WHO regions, helping to identify and address areas of concern.

**Fig.8. Bar chart : Top 10 Countries by Confirmed COVID-19 Cases**

**United States:**

The United States has the highest number of confirmed cases, reaching nearly 4 million. This indicates a significant impact of the pandemic in the country, likely due to factors such as population size and the spread of the virus in various states.

**Brazil and India:**

Brazil and India follow, both with confirmed cases exceeding 3 million. These countries have faced major challenges in controlling the spread due to dense populations and healthcare infrastructure limitations.

**Russia and South Africa:**

Russia and South Africa have substantial numbers of confirmed cases, highlighting the pandemic's reach into different continents and the global nature of the crisis.

**Mexico and Peru:**

These countries, with cases in the range of 1 to 2 million, show significant infection rates in the Latin American region. The high numbers reflect regional outbreaks and the struggle to manage the virus spread effectively.

**Chile and United Kingdom:**

Both countries report significant numbers, with the United Kingdom slightly lower than Chile. This indicates challenges faced by European nations and emphasizes the necessity for robust public health measures.

**Iran:**

Iran, at the bottom of the top 10 list, still reports a high number of cases, highlighting ongoing pandemic management challenges in the Middle East.

This chart effectively visualizes the global distribution of COVID-19 cases, with the United States leading by a significant margin. The high numbers in countries across different continents demonstrate the widespread impact of the pandemic. Understanding these numbers can help in comparative analysis of public health responses and preparedness across different regions.

**Fig.9. Bar chart: Least Countries by Confirmed COVID-19 Cases**

**Western Sahara:**

Western Sahara has the fewest confirmed COVID-19 cases among the listed countries. This might be due to its lower population and limited international connectivity.

**Vatican City:**

With its small population and highly controlled environment, Vatican City also reports low COVID-19 cases. Its unique position allows for stringent health measures.

**Greenland:**

Greenland's isolation and lower population density contribute to its low confirmed case count. Effective early interventions and limited exposure might have played a role.

**Saint Kitts and Nevis:**

As a small island nation, Saint Kitts and Nevis has managed to keep its case numbers low, likely through rigorous travel restrictions and public health policies.

**Dominica:**

Dominica’s low numbers could be attributed to its effective pandemic response and public health measures.

**Laos:**

Despite being in a region with high COVID-19 activity, Laos has maintained lower case numbers. This might be due to early and effective government interventions.

**Grenada:**

Grenada’s low case count can be linked to its geographical isolation and efficient health measures.

**Saint Lucia:**

Similar to other island nations, Saint Lucia has benefited from travel restrictions and public health campaigns to keep cases low.

**Timor-Leste:**

Timor-Leste’s lower case numbers suggest successful management and preventive measures in place.

**Fiji:**

Among the listed countries, Fiji has the highest number of confirmed cases. Despite being an island nation, it may have faced challenges due to international travel and local outbreaks.

The bar chart effectively highlights how geographical isolation, population size, and stringent health measures contribute to the low number of confirmed COVID-19 cases in these countries. It underscores the importance of early intervention, strict travel restrictions, and effective public health strategies in managing the spread of the virus.

**Fig.10. Scatter plot: Death Rate vs. Recovery Rate by Country**

**Clustering of Data Points:**

Most of the data points are clustered towards the left side of the plot, indicating that the majority of countries have a low death rate per 100 cases.

**High Recovery Rates:**

A significant number of countries have high recovery rates, with many data points near the top of the plot (close to 100 recovered per 100 cases).

**Outliers:**

There are a few outliers with higher death rates (above 10 deaths per 100 cases) and varying recovery rates.

**Regional Distribution:**

The color coding shows that countries from different WHO regions are spread across the plot, with no single region dominating any particular area. However, there are some regional trends visible, such as a few European countries having higher death rates.

**Correlation:**

There appears to be a general trend where countries with higher death rates tend to have lower recovery rates, although this is not a strict rule as there are exceptions.

This scatter plot provides a visual representation of the relationship between death rates and recovery rates across different countries and regions.

It can be useful for understanding the impact of health policies and the effectiveness of medical interventions in different parts of the world.

**Fig.11. Scatter plot: Confirmed case Rate and Death rate by WHO Region**

**Clustering near the Origin:**

Most data points are clustered near the origin, indicating that a majority of regions have relatively low numbers of confirmed cases and deaths. This suggests that many regions have been successful in keeping the pandemic under control to some extent.

**High Outliers:**

The Americas and Europe show significant outliers with higher confirmed cases and death rates.

**Americas:** One data point exceeds 4 million confirmed cases and around 140,000 deaths, indicating a severe impact of COVID-19 in this region.

**Europe:** There is also a notable outlier with a high number of deaths but fewer confirmed cases compared to the Americas, suggesting a higher death rate in proportion to confirmed cases.

**Comparative Analysis of Regions:**

**Africa:** Shows lower numbers of confirmed cases and deaths, which might be due to less testing or better management and preventive measures.

**Eastern Mediterranean:** Similar to Africa, with relatively lower case and death numbers.

**Western Pacific:** This region has low confirmed cases and deaths, reflecting effective control measures or lower infection rates.

**South-East Asia:** Also shows lower confirmed cases and deaths compared to the Americas and Europe, indicating better control over the spread of the virus in this region.

**Trend Analysis:**

The scatter plot indicates a positive correlation between the number of confirmed cases and deaths across all regions. This relationship is expected as more cases lead to more deaths, but the degree of correlation varies by region.

The scatter plot provides a visual representation of the COVID-19 impact across different WHO regions, highlighting significant disparities in confirmed cases and death rates. The Americas and Europe are notably more affected, with higher outliers indicating severe impacts. In contrast, regions like Africa, Eastern Mediterranean, Western Pacific, and South-East Asia show lower confirmed cases and deaths, pointing to better management or fewer infections.

This analysis helps in understanding regional differences in the pandemic's impact and can guide public health strategies and resource allocation.

**Fig.12. Bar chart: New Cases, Deaths, and Recoveries by WHO Region**

**Americas:**

This region reports the highest numbers across all three categories: new cases, deaths, and recoveries. The count of new cases is notably high, indicating a significant ongoing spread. The high number of recoveries suggests that while many are getting infected, a large number are also recovering.

**South-East Asia:**

South-East Asia holds the second highest counts, especially in new cases and recoveries. This suggests a substantial impact of COVID-19 in terms of spread, but also a robust recovery rate.

**Africa, Eastern Mediterranean, and Europe:**

These regions show moderate counts of new cases and recoveries, with relatively lower numbers of new deaths. This balanced pattern could indicate effective management in terms of both healthcare and recovery, even though new cases continue to appear.

**Western Pacific:**

The Western Pacific region has the lowest numbers in all categories, suggesting effective control measures or lower infection rates in these areas.

This bar chart provides a clear comparison of how different WHO regions are handling the COVID-19 pandemic. The Americas and South-East Asia are experiencing higher infection rates but also show significant recovery numbers, indicating a robust response. On the other hand, regions like the Western Pacific demonstrate lower numbers across the board, suggesting effective containment strategies.

Understanding these patterns can help in tailoring region-specific strategies to combat the pandemic more effectively.

**Fig.13. Box plot: Mortality Rate Distribution by WHO Region**

**Eastern Mediterranean:** Median: Around 4 deaths per 100 cases.

**IQR (Interquartile Range):** Approximately 2 to 6 deaths per 100 cases.

**Outliers:** A few outliers above 10 deaths per 100 cases.

**Europe:**

**Median:** Also around 4 deaths per 100 cases.

**IQR:** Approximately 2 to 6 deaths per 100 cases.

**Outliers:** Several outliers, with the highest being above 25 deaths per 100 cases.

**Africa:**

**Median:** Around 3 deaths per 100 cases.

**IQR:** Approximately 1 to 5 deaths per 100 cases.

**Outliers:** A few outliers above 10 deaths per 100 cases.

**Americas:**

**Median:** Around 3 deaths per 100 cases.

**IQR:** Approximately 2 to 5 deaths per 100 cases.

**Outliers:** A few outliers above 10 deaths per 100 cases.

**Western Pacific:**

**Median:** Around 2 deaths per 100 cases.

**IQR:** Approximately 1 to 3 deaths per 100 cases.

**Outliers:** A few outliers above 5 deaths per 100 cases.

**South-East Asia:**

**Median:** Around 2 deaths per 100 cases.

**IQR:** Approximately 1 to 3 deaths per 100 cases.

**Outliers:** A few outliers above 5 deaths per 100 cases.

This box plot provides a visual comparison of mortality rates across different WHO regions, showing significant variations. The highest median mortality rates are seen in the Eastern Mediterranean and Europe regions, both at around 4 deaths per 100 cases. The Western Pacific and South-East Asia regions have the lowest median mortality rates, at around 2 deaths per 100 cases. The presence of outliers in all regions indicates that there are some countries with exceptionally high mortality rates, which could be due to various factors such as healthcare quality, demographic differences, or the spread of the virus.

This analysis can help in identifying regions that need more attention and resources to manage COVID-19 effectively.

**Fig.14. Scatter plot: Actual vs. Predicted Values**

**Correlation:**

The plot showcases the relationship between actual and predicted values. Ideally, if the model were perfect, all data points (blue dots) would lie on the red dashed line (y = x). This line represents a perfect prediction scenario where actual values match predicted values exactly.

**Accuracy:**

Many of the data points are clustered around the red dashed line, indicating that the model's predictions are fairly accurate for those points. However, there are several points that deviate significantly from this line, suggesting discrepancies between actual and predicted values.

**Outliers:**

There are notable outliers, especially in the higher range of values, where the predicted values are either much higher or lower than the actual values. These outliers indicate instances where the model's predictions are not accurate, possibly due to data anomalies or model limitations.

**Spread:**

The spread of the residuals (differences between actual and predicted values) increases with the actual values. This suggests that the model's prediction error tends to grow as the actual values increase, indicating potential issues with model performance at higher value ranges.

This scatter plot is essential for evaluating the performance of the predictive model. It helps in identifying how well the model predicts actual values and where it falls short. The presence of outliers and increasing spread at higher values indicates areas where the model may need improvement or further fine-tuning.

Understanding these discrepancies can guide efforts to enhance model accuracy and reliability, particularly for higher value predictions.

**Fig.15. Scatter plot: Residual Plot (Predicted vs. Residuals)**

**Pattern in Residuals:**

The residuals are not randomly scattered around the horizontal line at y=0y = 0. There are clusters of points with higher residuals at higher predicted values, suggesting that the model may not be performing well for higher predicted values.

**Heteroscedasticity:**

The spread of the residuals increases as the predicted values increase, indicating heteroscedasticity. This means that the variance of the residuals is not constant across all levels of the predicted values, which can be a problem for certain types of regression models.

**Outliers:**

There are a few points with very high or very low residuals, indicating potential outliers. These points could be investigated further to understand why the model is performing poorly for these observations.

**Model Performance:**

The presence of large residuals and a non-random pattern suggests that the model may not be capturing all the underlying patterns in the data, and there may be room for improvement in the model.

This residual plot indicates that the model may have issues with heteroscedasticity and outliers, and it may not be performing well for higher predicted values. Investigating the outliers and addressing the heteroscedasticity could help improve the model's performance.

**Fig.16. Histogram: Distribution of Residuals**

**Central Tendency:**

The highest frequency is observed at 0 residuals, with around 25 occurrences, indicating that most of the model's predictions are quite accurate, closely matching the actual values.

**Skewness:**

The distribution is right-skewed, with a long tail extending towards the positive residuals. This suggests that while most predictions are close to the actual values, there are instances where the model overestimates.

**Range:**

The residuals range from approximately -1500 to 1000, showcasing a wide spread. This range tells us about the variability in the model's errors.

**Outliers:**

The presence of outliers, particularly on the positive side, indicates certain instances where the model's predictions deviate significantly from the actual values. This could be due to anomalies in the data or limitations in the model.

**Density Function:**

The green line representing a probability density function overlaid on the histogram suggests a comparison to a normal distribution. The skewness indicates that the residuals do not follow a perfect normal distribution, hinting at potential areas for model improvement.

The histogram of residuals shows that while the predictive model generally performs well, there are areas where it could be refined to handle the outliers and reduce the positive skewness. Analyzing such residuals is crucial for diagnosing and improving model performance.