**Final Analysis of the Task – 2**

**(Group - 3)**

**Initial Analysis**

1. **Dropping the unwanted columns:**
   * We have dropped many columns from the dataset as there are many null/missing/irrelevant/duplicated information present. Here is the list of columns dropped:

['continentcode', 'countrycode', 'source\_link', 'source\_name', 'storm\_name','photos\_link','key','version','geolocation','time','changeset\_id','cat\_id', 'cat\_src','adminname1','adminname2','tstamp','location\_accuracy','location\_description', 'near'].

1. **Doing some visualizations:**
   * We have plotted Top 20 countries according to landslide count which gives us the idea about the places where landslides occur the most.
   * Then we have plot the boxplot of fatalities v/s landslide size which gives the insight about outliers in the data.
   * Then we did outlier removal for numeric columns but the data is very much skewed so the outlier removal has also not helped that much.
   * After removal we plotted more plots of Top 20 countries v/s fatalities and Top 20 countries v/s injuries.

**Datetime analysis**

1. **Date column conversion to datetime:**
   * Converted the date column from categorical to datetime to use different analysis related to date column such as day of the month, days, yearly trend, monthly trend, weekly trend, etc.
   * Plotted many bar plots associated with the above analysis and also shown the heatmap correlation matrix of the numerical columns.

**Feature Selection & Engineering**

1. **Selecting features:**
   * Selected the most relevant and interdependent features which will be helpful for further analysis. List of selected columns are:

['date', 'country', 'hazard\_type', 'landslide\_type', 'trigger', 'fatalities', 'injuries', 'latitude', 'longitude', 'distance', 'population']

* + Then created many features based on the selected features that will also help for further analysis.

1. **Feature Engineering:**
   * Performed Label Encoding on the categorical features selected and done imputation on the numerical features selected.

**Univariate Analysis**

1. **Fatalities and Injuries:**
   * The distribution of fatalities and injuries is highly skewed, with most events resulting in few or no casualties. This indicates that while landslides can be deadly, the majority of events do not result in high casualties.
2. **Latitude and Longitude:**
   * The distribution of latitude and longitude shows the geographical spread of landslides. Most events are concentrated in specific regions, which could be due to geographical or climatic factors.
3. **Distance and Population:**
   * The distance feature, representing the radius of the affected area, shows a right-skewed distribution, indicating that most landslides affect a relatively small area.
   * Population distribution is also skewed, with many events occurring in less populated areas.
4. **Total Casualties and Population Density:**
   * Total casualties, a combination of fatalities and injuries, also shows a skewed distribution, reinforcing the observation that most landslides have a limited human impact.
   * Population density, calculated as population per affected area, shows variability, suggesting that landslides occur in both densely and sparsely populated areas.

**Bivariate Analysis**

1. **Fatalities vs. Injuries:**
   * The scatter plot shows a positive correlation between fatalities and injuries, as expected. Events with higher fatalities tend to have higher injuries as well.
2. **Distance vs. Population Density:**
   * The scatter plot shows no clear relationship between distance and population density, suggesting that the size of the affected area does not directly correlate with the population density.
3. **Distance vs. fatalities:**
   * The scatter plot shows relationship between the distance of the impact of landslide with fatalities happened. By the graph we understood that the fatalities are very less at the greater distance as compared to nearer distance.
4. **Distance vs. injuries:**
   * The scatter plot shows relationship between the distance of the impact of landslide with injuries happened. By the graph we understood that the injuries are very less at the greater distance as compared to nearer distance.
5. **Distance vs. casualties:**
   * The scatter plot shows relationship between the distance of the impact of landslide with casualties happened. By the graph we understood that the casualties are very less at the greater distance as compared to nearer distance.
6. **Total casualties vs. Population Density:**
   * The scatter plot shows no relationship between the Total casualties due to landslide with population density. Suggesting that the size of the Total casualties does not directly correlate with the population density.

**Correlation Analysis**

* The correlation heatmap reveals:
  + A strong positive correlation between fatalities and injuries, indicating that events with higher fatalities also tend to have higher injuries.
  + Weak correlations between other numeric features, suggesting that they may not have strong linear relationships with each other.

**Insights and Implications**

* **Geographical Hotspots:** The concentration of landslides in specific regions suggests the need for targeted mitigation efforts in these areas.
* **Casualty Patterns:** The skewed distribution of casualties indicates that while landslides can be deadly, most events have limited human impact. This could inform resource allocation for emergency response.
* **Population Density:** The variability in population density suggests that landslides can affect both urban and rural areas, highlighting the need for diverse mitigation strategies.
* **Feature Relationships:** The weak correlations between most features suggest that a complex interplay of factors influences landslide occurrences and impacts, which may require advanced modeling techniques to fully understand.