**Tools:**

Jyupter Notebook -IDE

MYSQL Workbench 8.0

Kaggle

**Libraries used:**

Pandas

SQL alchemy

MySQL connector

Steps:

At first, I searched for the dataset in Kaggle consisting of major fields and medium sized data. Then the dataset is read into jupyter notebook and performed some basic operations like data available in first 5, last 5, sample, computed the mean, max, std and so on with describe tag, along with the data types of the dataset available with each field. With that null value is checked and sum is obtained.

After some few basic steps cleaning the data is done with the available fields. We can determine the dependent and independent variable as it helps in data reading and cleaning easily. As there was nearabout 88889 data’s and the only few <10 had full values i.e., fully field 31 columns completely so I had to filling the missing values in place of necessary. To fill in the values in number part we used mean values and in string ‘Unknown’. We could ignore those rows but it would affect the total information of the dataset.

So, for this large dataset, I have reduced columns once their purpose is done so that for further process it wouldn’t slow down our process.

Dataset is all about the Aviation Accident ranging up to 2023, so here I compared the factors which could bring accidents like weather condition, location along with the injury severity and calculated the ratio of uninjured along with all the data available.

**Columns added:**

I have added yearly, monthly, daily basis columns which shows the data about the aviation accident and the Ratio uninjured.

As for database I used MySQL with its MySQL workbench as for storing data.

First, I converted our final dataset to csv file and with MySQL.Connector I established connection in between the workbench and jupyter notebook and created the table flight accident and imported the values.

**As for another process**

Here, we create an engine which connects us with the database. After connection is established, it loads the value of our final dataset when the final dataset is called.

**As for creation of table**

We create another table in the same database and we connect it with the engine. Now we, create new table and insert the value from the previous loaded tables as we did before. Then connection is closed.

**Alternative approach:**

With the csv file we simply can import the datasheets as such this process takes time proportional with the dataset size.

**Motive**

So, the main motive for this process was to know the insights about the aviation accident or we can say few problems causing the accident. Similarly, with this we also can know the rate of accident in different countries along with the casualties after the accident. As well as we can also get information about the ‘hotspot’.

Along with this we can use various other python libraries for other computation such as NumPy, seaborn, matplotlib and so on for better presentation of the dataset and make user interactive.