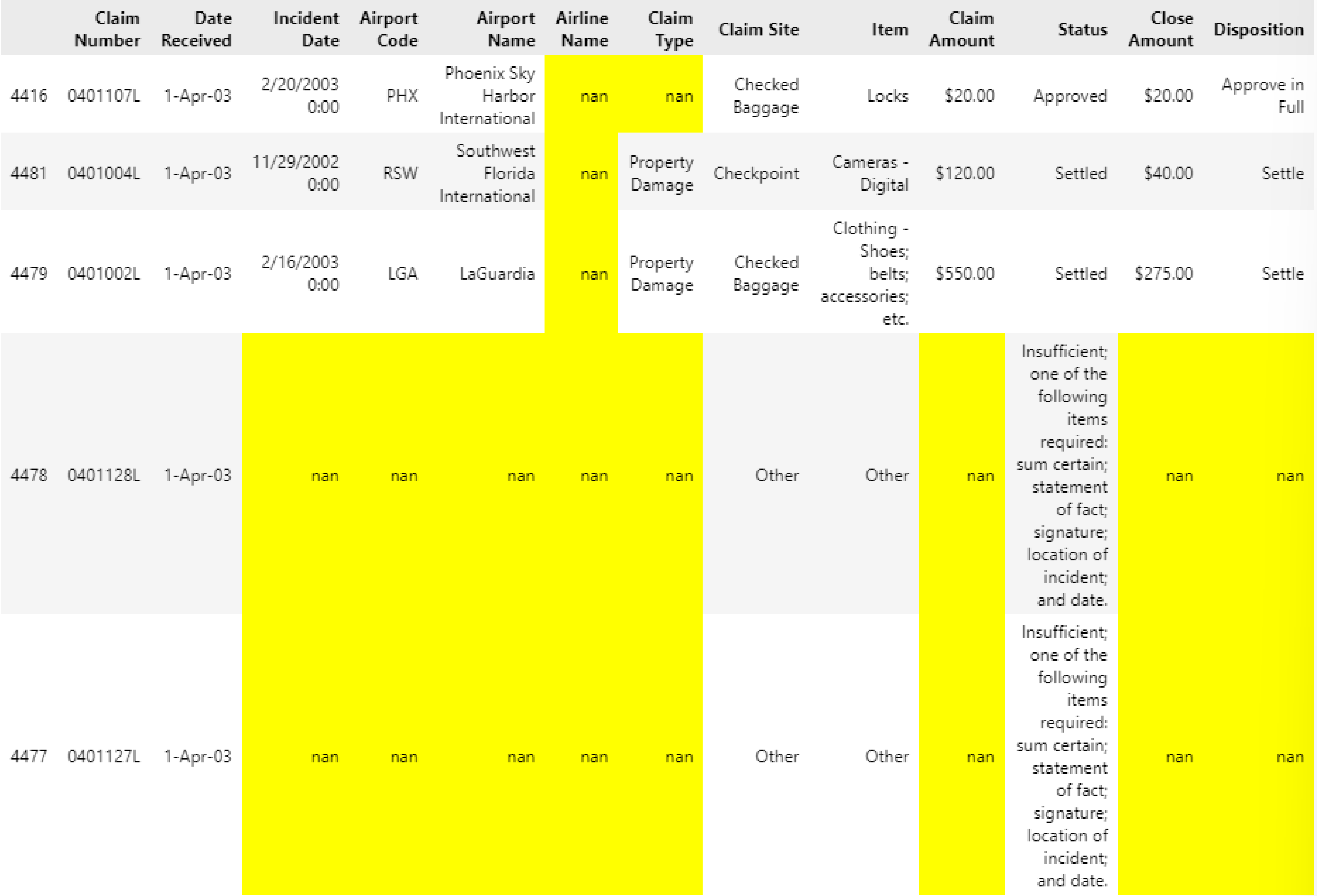
# MISSING VALUES HANDLING: CASE STUDY ON TSA CLAIMS DATABASE

by: Noor Kharismawan Akbar

Purwadhika
Digital Technology School



| | Claim Number | Date Received | Incident Date | Airport Code | Airport Name | Airline Name | Claim Type | Claim Site | Item | Claim Amount | Status | Close Amount | Disposition |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4416 | 0401107L | 1-Apr-03 | 2/20/2003 0:00 | PHX | Phoenix Sky Harbor International | nan | nan | Checked Baggage | Locks | $20.00 | Approved | $20.00 | Approve in Full |
| 4481 | 0401004L | 1-Apr-03 | 11/29/2002 0:00 | RSW | Southwest Florida International | nan | Property Damage | Checkpoint | Cameras - Digital | $120.00 | Settled | $40.00 | Settle |
| 4479 | 0401002L | 1-Apr-03 | 2/16/2003 0:00 | LGA | LaGuardia | nan | Property Damage | Checked Baggage | Clothing - Shoes; belts; accessories; etc. | $550.00 | Settled | $275.00 | Settle |
| 4478 | 0401128L | 1-Apr-03 | nan | nan | nan | nan | nan | Other | Other | nan | Insufficient; one of the following items required: sum certain; statement of fact; signature; location of incident; and date. | nan | nan |
| 4477 | 0401127L | 1-Apr-03 | nan | nan | nan | nan | nan | Other | Other | nan | Insufficient; one of the following items required: sum certain; statement of fact; signature; location of incident; and date. | nan | nan |

DF.NAN

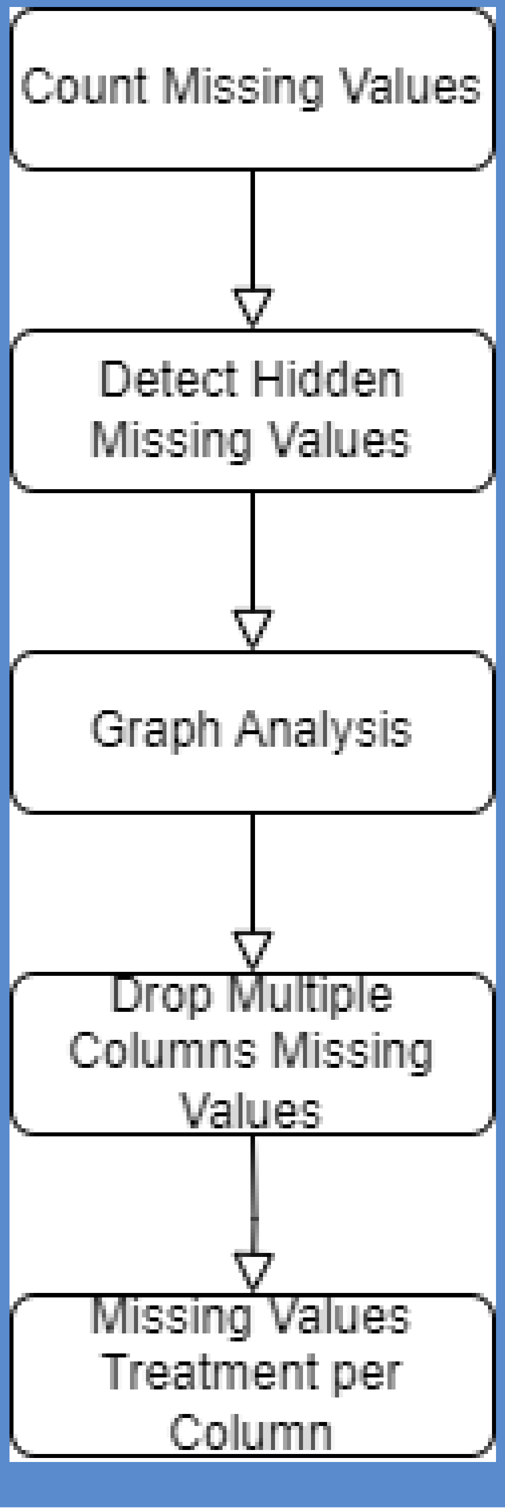
# BACKGROUND

**Missing data** occur when no data values are stored for the variable under observation in statistics. This problem is quite common in many real-life datasets and can have a significant effect on the conclusions that can be drawn from it.

In this article, we will use the **TSA Claims Database** as a case study.
Dataset Link:
https://www.kaggle.com/datasets/terminal-security-agency/tsa-claims-database

# MISSING VALUES HANDLING FLOW

Following are the steps for handling missing values in this case study.

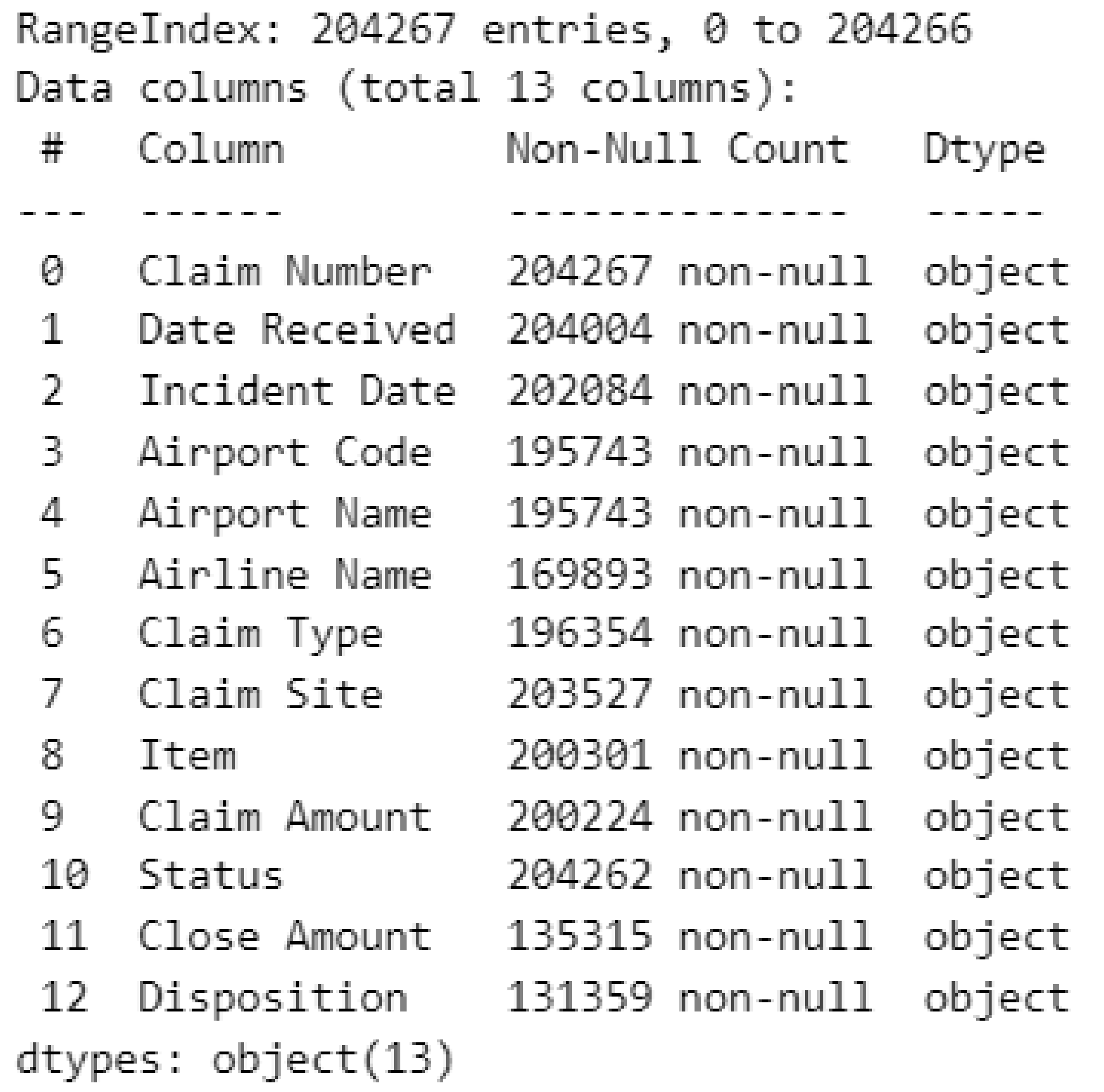


Count Missing Values
Detect Hidden Missing Values
Graph Analysis
Drop Multiple Columns Missing Values
Missing Values Treatment per Column

FLOWCHART

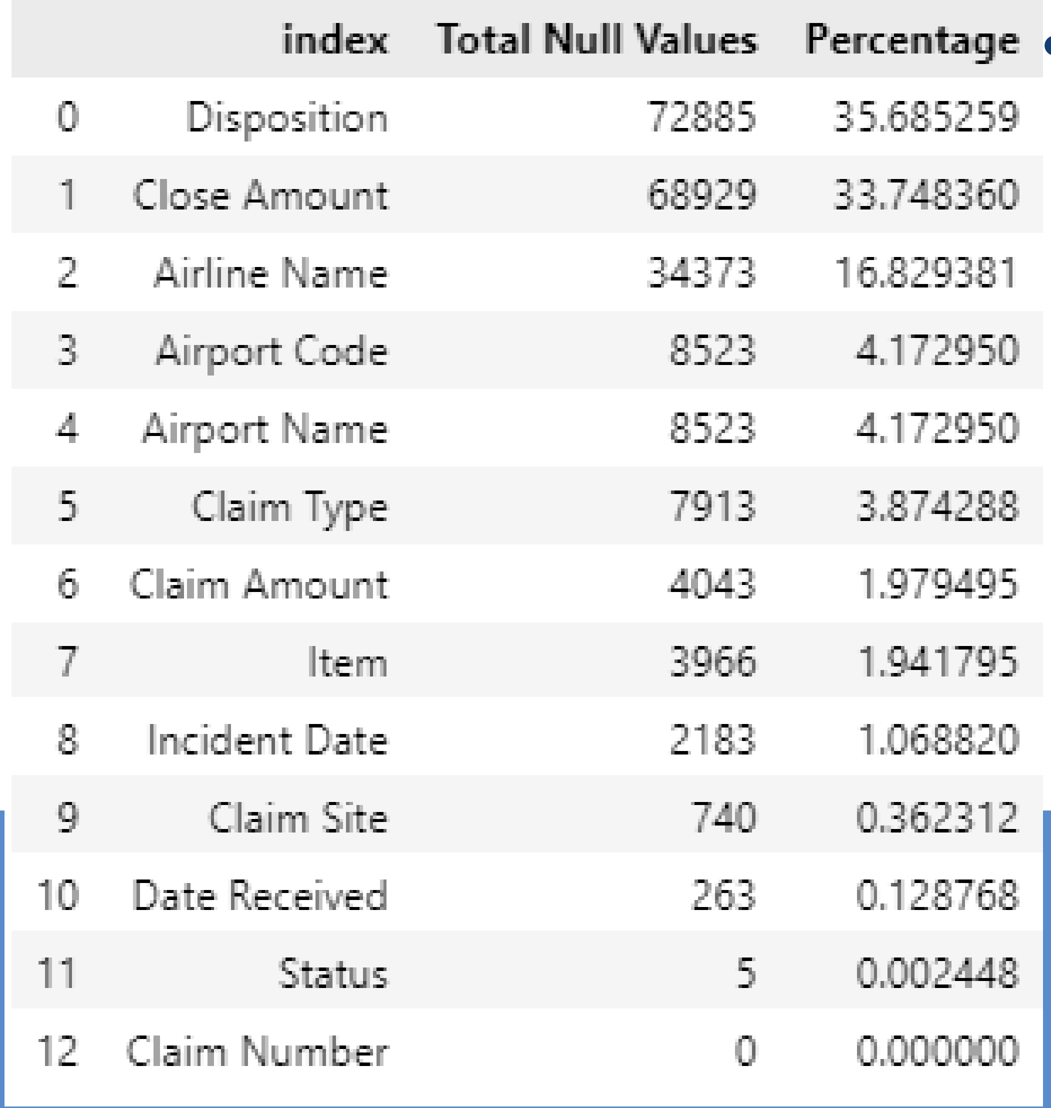
# MISSING VALUES CHECK

- The dataset consists of **204,267 rows & 13 columns**. It can also be seen that there are several missing values. All existing data types are in the form of object.
- It can be seen that the missing values have the **largest** value at **~30%** for Close Amount & Disposition values. If a variable has **less than 50%** missing values, **imputation** might be a viable option.



```
RangeIndex: 204267 entries, 0 to 204266
Data columns (total 13 columns):
 #   Column          Non-Null Count   Dtype 
---  ------          --------------   ----- 
 0   Claim Number    204267 non-null  object
 1   Date Received   204004 non-null  object
 2   Incident Date   202084 non-null  object
 3   Airport Code    195743 non-null  object
 4   Airport Name    195743 non-null  object
 5   Airline Name    169893 non-null  object
 6   Claim Type      196354 non-null  object
 7   Claim Site      203527 non-null  object
 8   Item            200301 non-null  object
 9   Claim Amount    200224 non-null  object
 10  Status          204262 non-null  object
 11  Close Amount    135315 non-null  object
 12  Disposition     131359 non-null  object
dtypes: object(13)
```

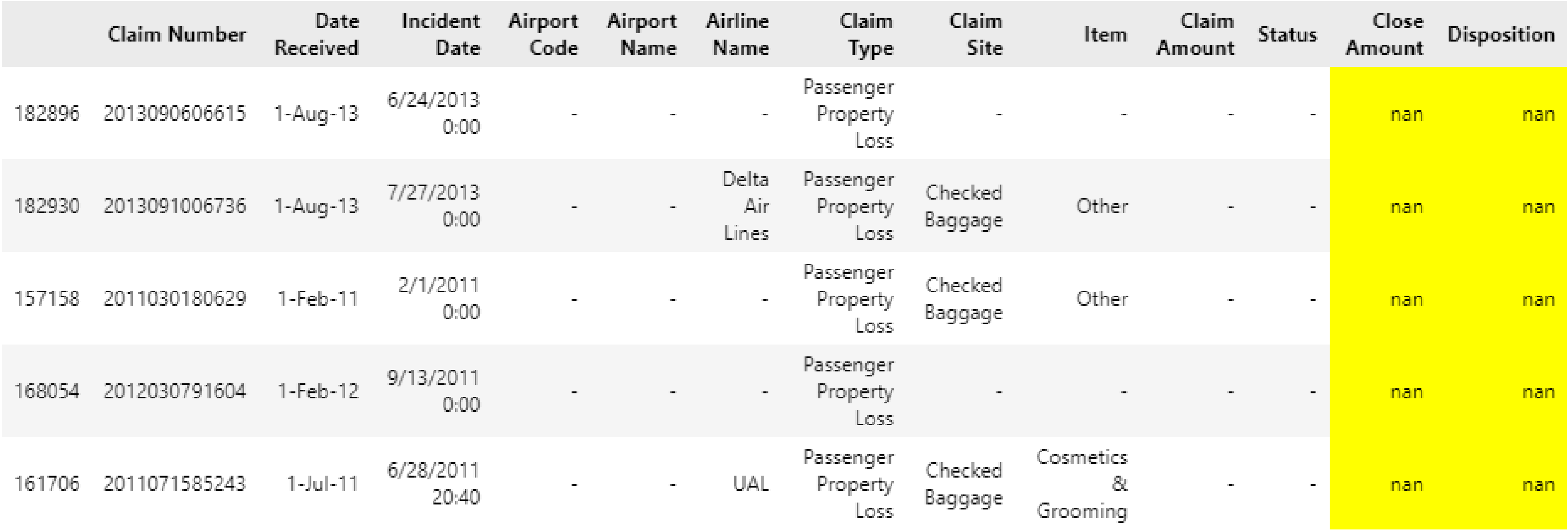
DF.INFO



| | index | Total Null Values | Percentage |
|---|---|---|---|
| 0 | Disposition | 72885 | 35.685259 |
| 1 | Close Amount | 68929 | 33.748360 |
| 2 | Airline Name | 34373 | 16.829381 |
| 3 | Airport Code | 8523 | 4.172950 |
| 4 | Airport Name | 8523 | 4.172950 |
| 5 | Claim Type | 7913 | 3.874288 |
| 6 | Claim Amount | 4043 | 1.979495 |
| 7 | Item | 3966 | 1.941795 |
| 8 | Incident Date | 2183 | 1.068820 |
| 9 | Claim Site | 740 | 0.362312 |
| 10 | Date Received | 263 | 0.128768 |
| 11 | Status | 5 | 0.002448 |
| 12 | Claim Number | 0 | 0.000000 |

NULL VAL CHECK

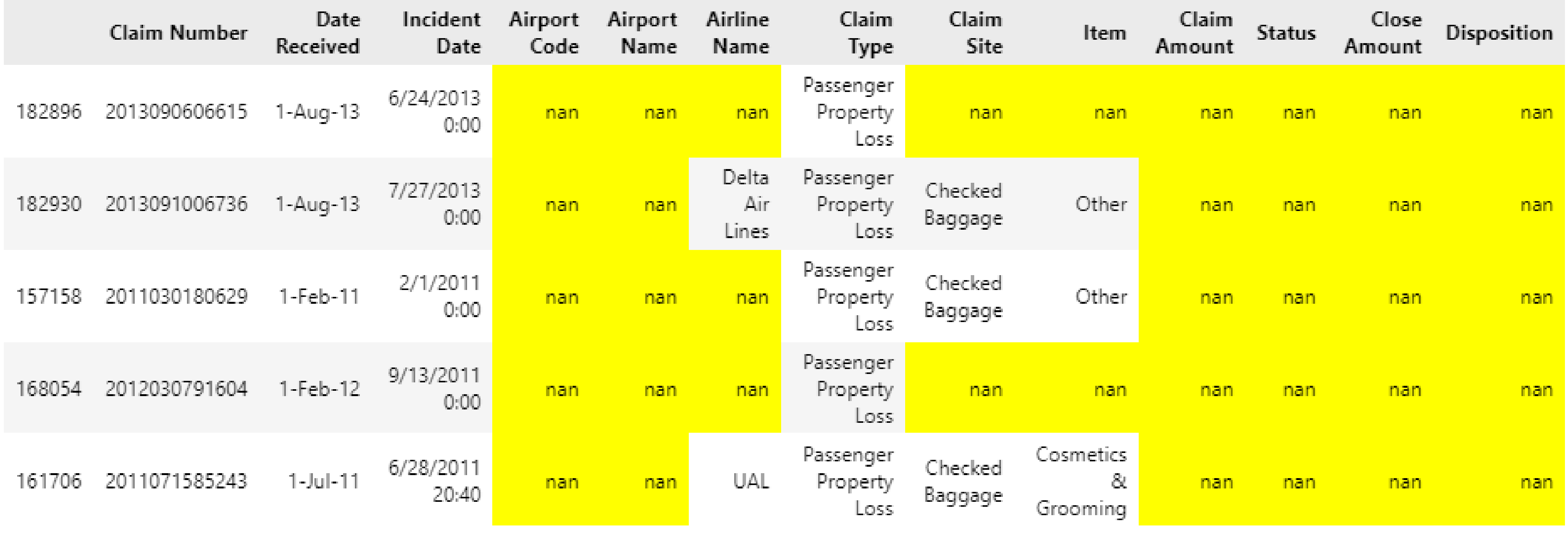
# HIDDEN MISSING VALUES



| | Claim Number | Date Received | Incident Date | Airport Code | Airport Name | Airline Name | Claim Type | Claim Site | Item | Claim Amount | Status | Close Amount | Disposition |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 182896 | 2013090606615 | 1-Aug-13 | 6/24/2013 0:00 | - | - | - | Passenger Property Loss | - | - | - | - | nan | nan |
| 182930 | 2013091006736 | 1-Aug-13 | 7/27/2013 0:00 | - | - | Delta Air Lines | Passenger Property Loss | Checked Baggage | Other | - | - | nan | nan |
| 157158 | 2011030180629 | 1-Feb-11 | 2/1/2011 0:00 | - | - | - | Passenger Property Loss | Checked Baggage | Other | - | - | nan | nan |
| 168054 | 2012030791604 | 1-Feb-12 | 9/13/2011 0:00 | - | - | - | Passenger Property Loss | - | - | - | - | nan | nan |
| 161706 | 2011071585243 | 1-Jul-11 | 6/28/2011 20:40 | - | - | UAL | Passenger Property Loss | Checked Baggage | Cosmetics & Grooming | - | - | nan | nan |

**HIDDEN MISS. VAL**

↓

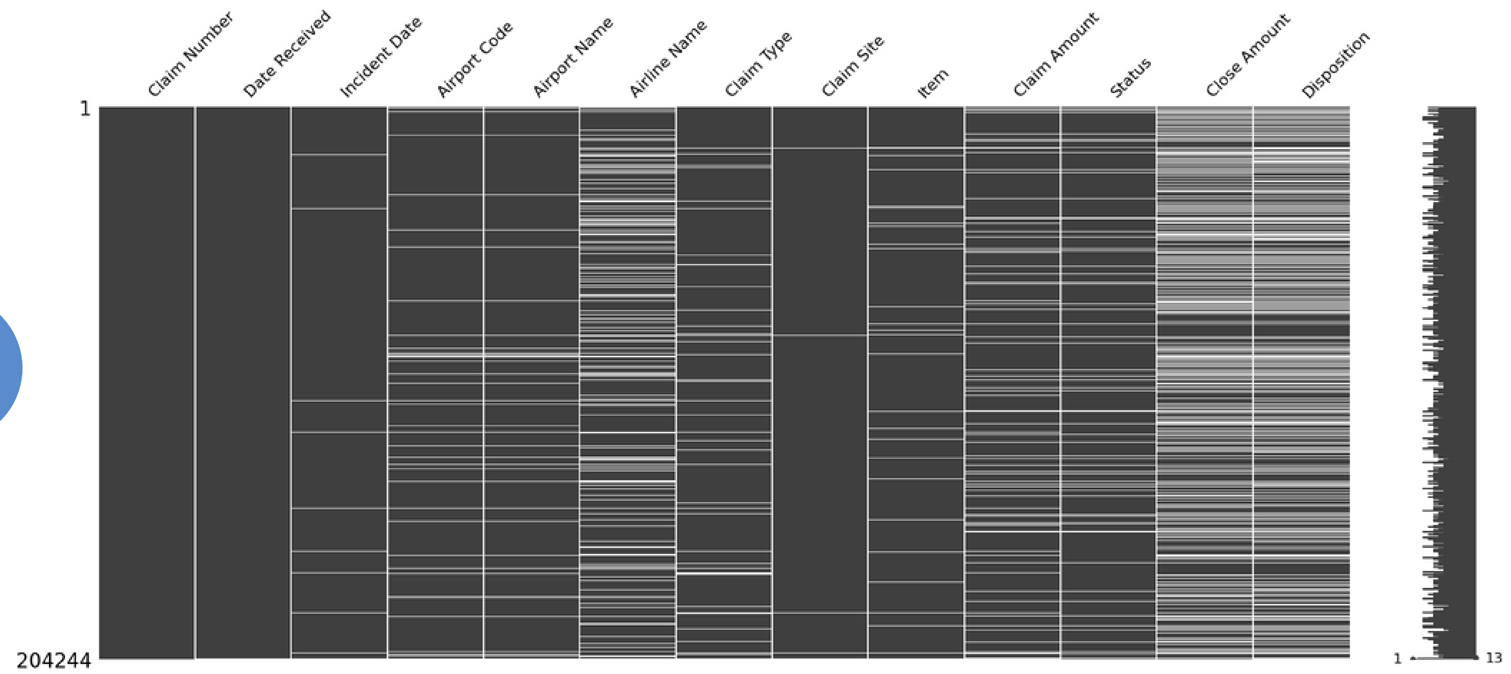


| | Claim Number | Date Received | Incident Date | Airport Code | Airport Name | Airline Name | Claim Type | Claim Site | Item | Claim Amount | Status | Close Amount | Disposition |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 182896 | 2013090606615 | 1-Aug-13 | 6/24/2013 0:00 | nan | nan | nan | Passenger Property Loss | nan | nan | nan | nan | nan | nan |
| 182930 | 2013091006736 | 1-Aug-13 | 7/27/2013 0:00 | nan | nan | Delta Air Lines | Passenger Property Loss | Checked Baggage | Other | nan | nan | nan | nan |
| 157158 | 2011030180629 | 1-Feb-11 | 2/1/2011 0:00 | nan | nan | nan | Passenger Property Loss | Checked Baggage | Other | nan | nan | nan | nan |
| 168054 | 2012030791604 | 1-Feb-12 | 9/13/2011 0:00 | nan | nan | nan | Passenger Property Loss | nan | nan | nan | nan | nan | nan |
| 161706 | 2011071585243 | 1-Jul-11 | 6/28/2011 20:40 | nan | nan | UAL | Passenger Property Loss | Checked Baggage | Cosmetics & Grooming | nan | nan | nan | nan |

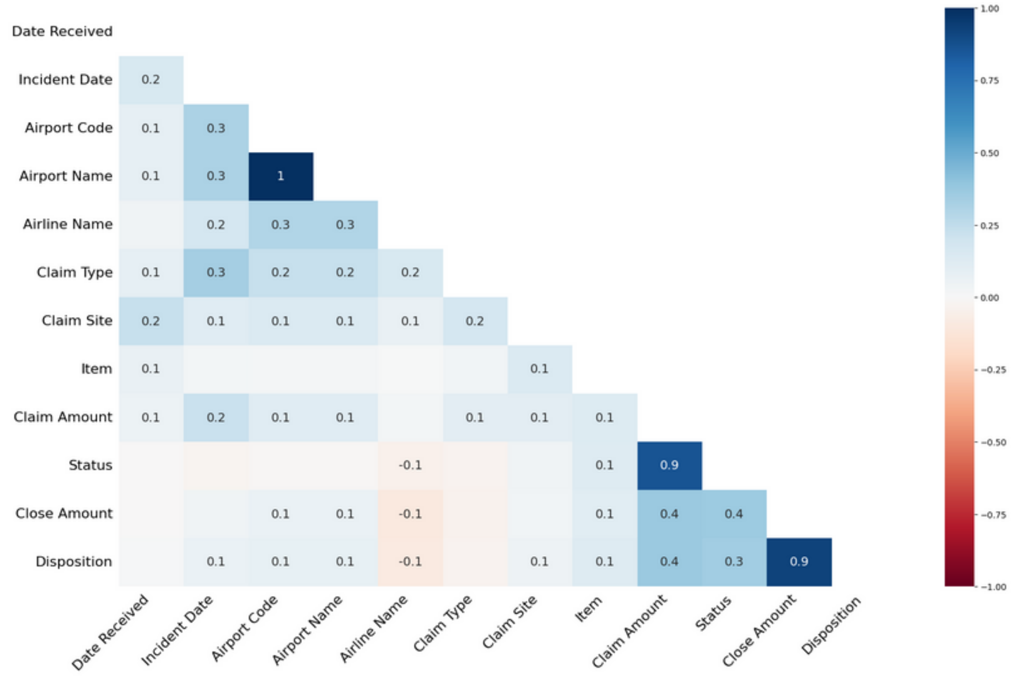
**NP.NAN MISS. VAL**

- There are some missing values, we will handle it. But, the initial step is identifying **hidden missing values (that is '-')**.
- That type of missing values must be **changed to numpy missing values** using np.nan.

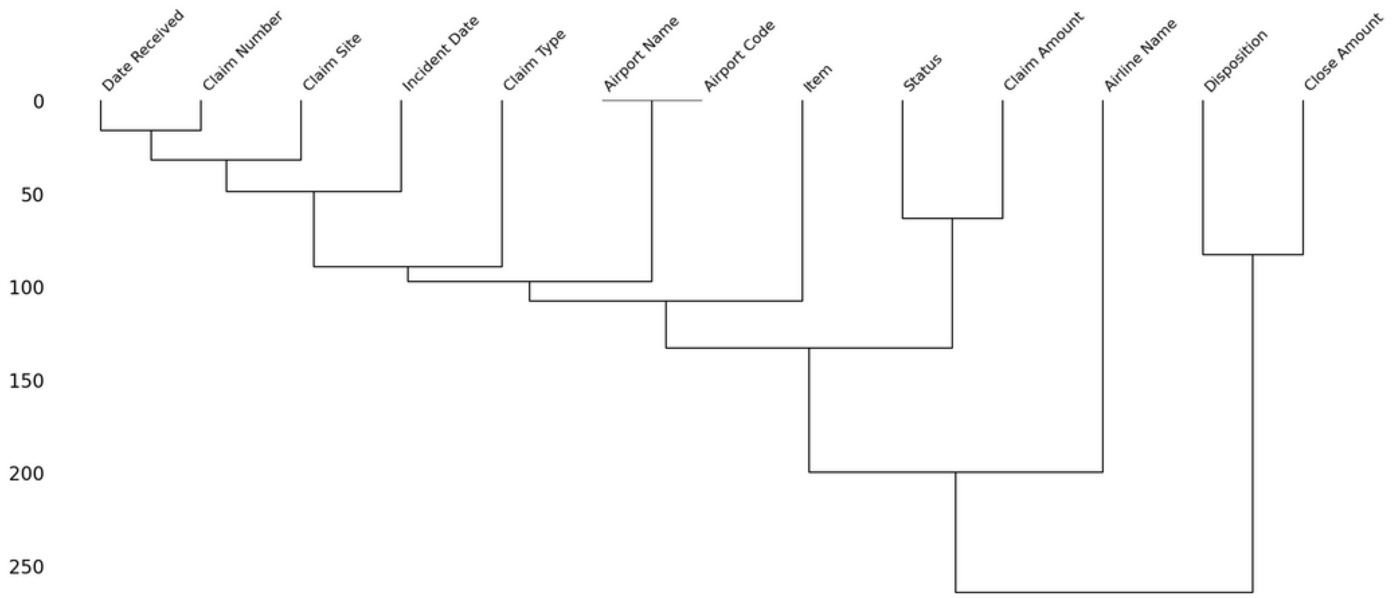
## MATRIX



## HEATMAP



## DENDOGRAM



# GRAPH ANALYSIS

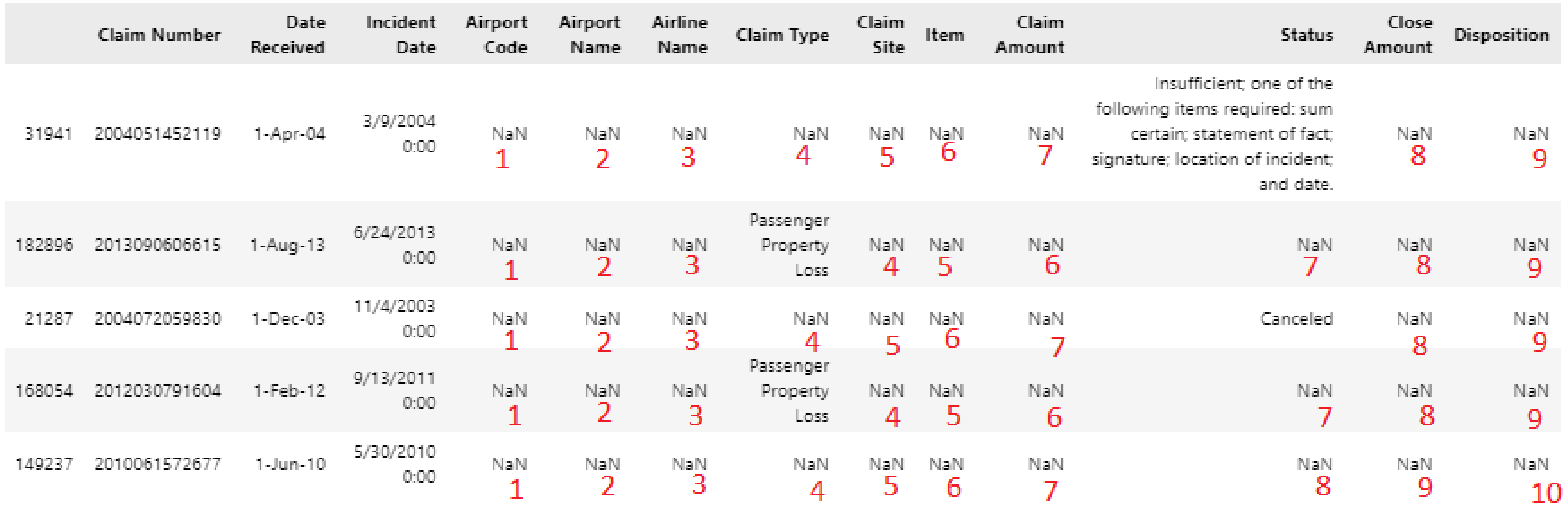
Missing values analysis:

1. Missing values are **scattered**
2. In the **same row**, there are also **several empty columns**
3. The **Disposition**, **Close Amount**, and **Airline Name** columns have many missing values
4. [**Status-Claim Amount**] & [**Close Amount-Disposition**] column has a high missing value correlation (**0.9**).
5. The **Airport Code & Airport Name** column has a missing value correlation with a value of **1**, whereas the two columns have a missing value with the **exact same number & and location**
6. Other columns have a **fairly low correlation** between columns

# DROP MULTIPLE COLUMNS MISSING VALUES

To deal with the large number of empty columns in one row, we will **drop rows that have less than 5 non-null values**.

Because it only contains a **little information** and if inputting is done it will cause a **lot of bias**.

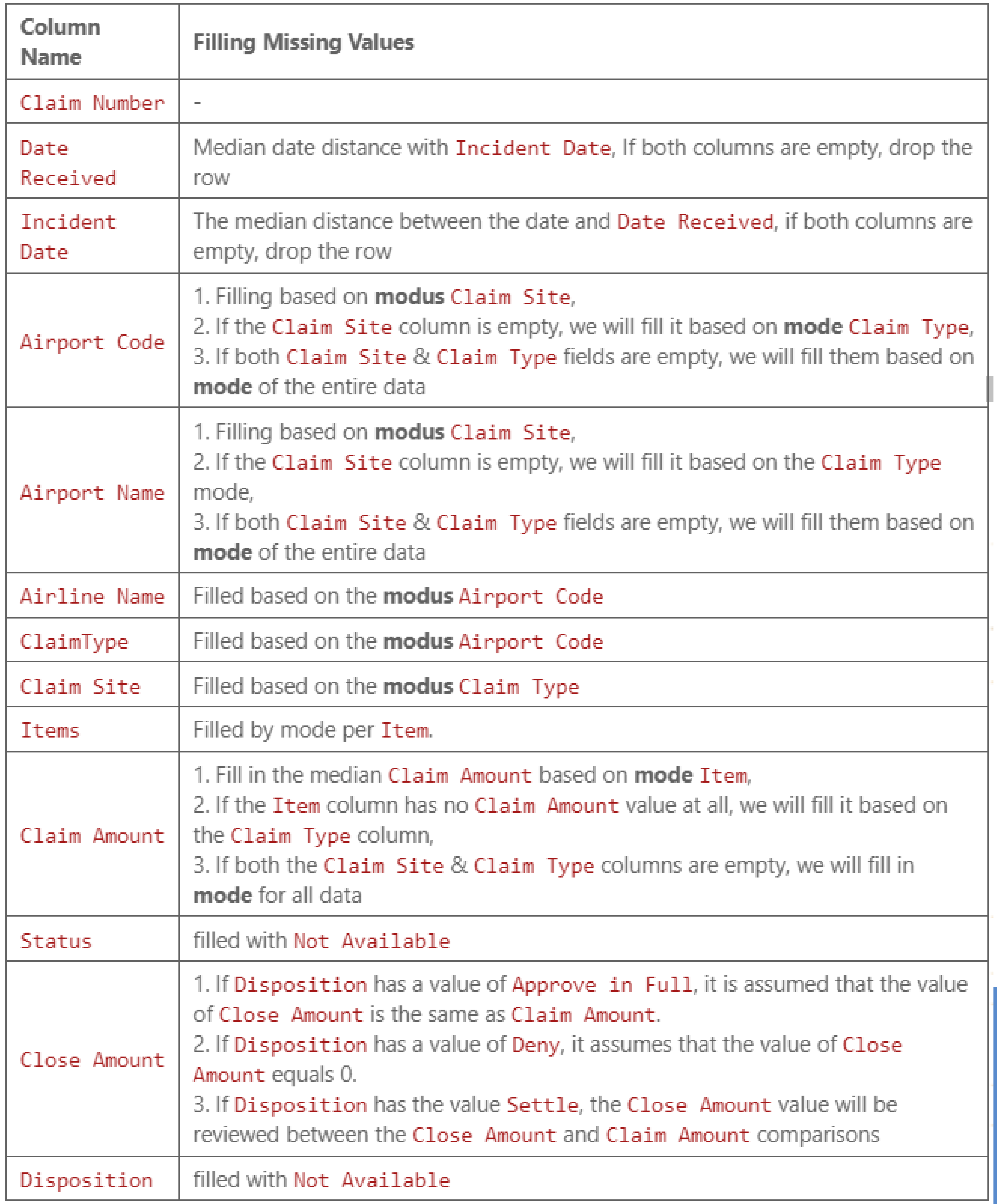


| | Claim Number | Date Received | Incident Date | Airport Code | Airport Name | Airline Name | Claim Type | Claim Site | Item | Claim Amount | Status | Close Amount | Disposition |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 31941 | 2004051452119 | 1-Apr-04 | 3/9/2004 0:00 | NaN<br>1 | NaN<br>2 | NaN<br>3 | NaN<br>4 | NaN<br>5 | NaN<br>6 | NaN<br>7 | Insufficient; one of the following items required: sum certain; statement of fact; signature; location of incident; and date. | NaN<br>8 | NaN<br>9 |
| 182896 | 2013090606615 | 1-Aug-13 | 6/24/2013 0:00 | NaN<br>1 | NaN<br>2 | NaN<br>3 | Passenger Property Loss | NaN<br>4 | NaN<br>5 | NaN<br>6 | NaN<br>7 | NaN<br>8 | NaN<br>9 |
| 21287 | 2004072059830 | 1-Dec-03 | 11/4/2003 0:00 | NaN<br>1 | NaN<br>2 | NaN<br>3 | NaN<br>4 | NaN<br>5 | NaN<br>6 | NaN<br>7 | Canceled | NaN<br>8 | NaN<br>9 |
| 168054 | 2012030791604 | 1-Feb-12 | 9/13/2011 0:00 | NaN<br>1 | NaN<br>2 | NaN<br>3 | Passenger Property Loss | NaN<br>4 | NaN<br>5 | NaN<br>6 | NaN<br>7 | NaN<br>8 | NaN<br>9 |
| 149237 | 2010061572677 | 1-Jun-10 | 5/30/2010 0:00 | NaN<br>1 | NaN<br>2 | NaN<br>3 | NaN<br>4 | NaN<br>5 | NaN<br>6 | NaN<br>7 | NaN<br>8 | NaN<br>9 | NaN<br>10 |

**MULTIPLE MISSING VALUES**

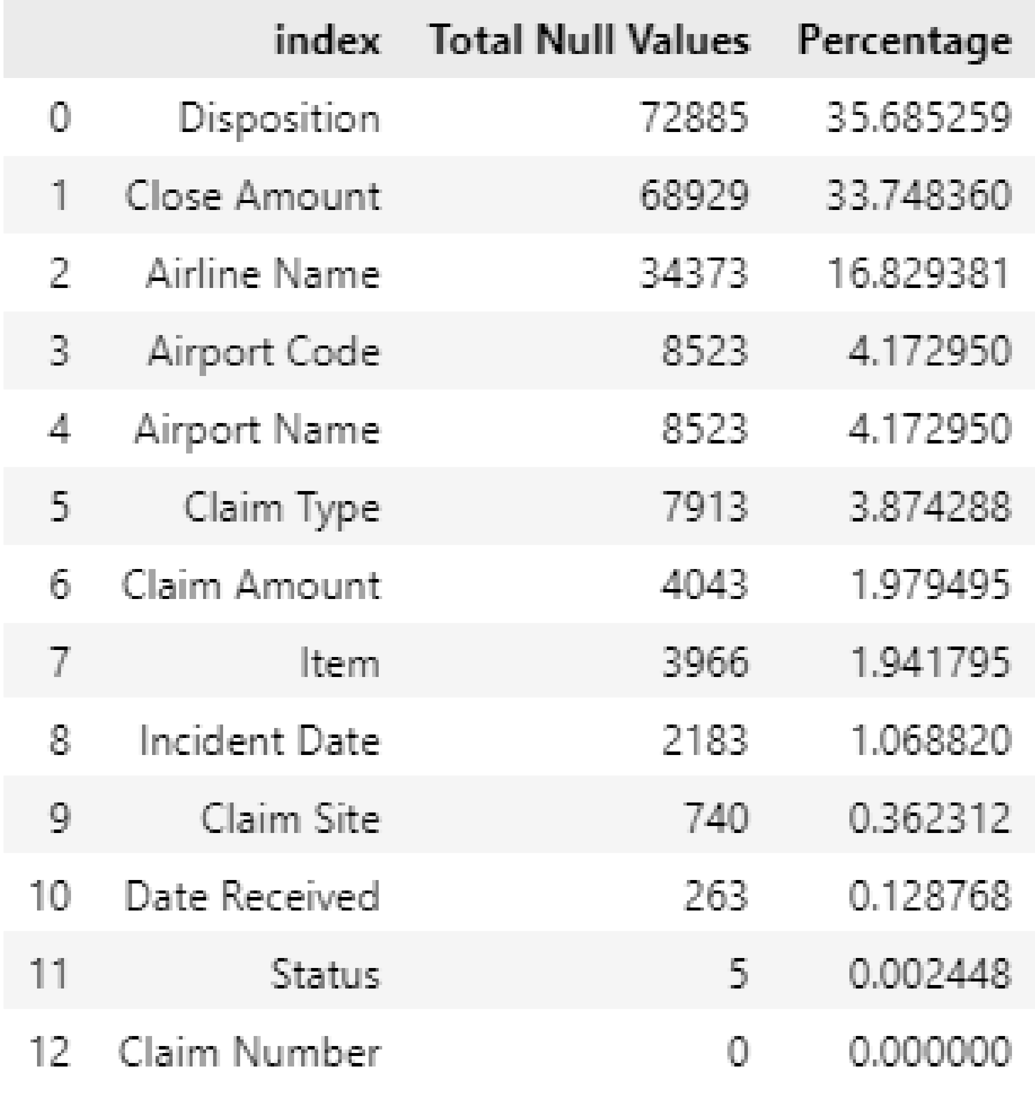
# FILLING BY COLUMNS

The following is a summary of filling in the missing values in this project.



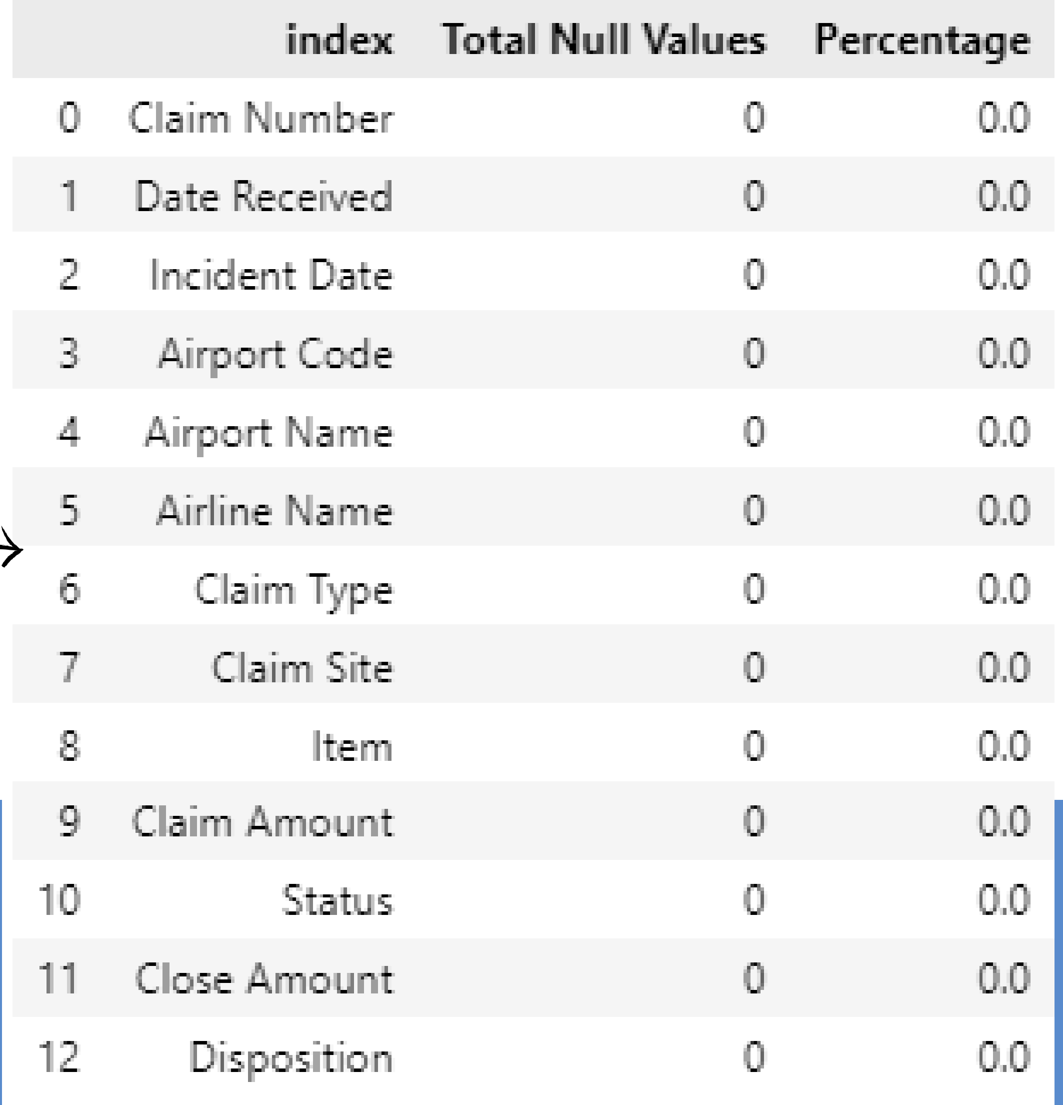
| Column Name | Filling Missing Values |
|---|---|
| `Claim Number` | - |
| `Date Received` | Median date distance with `Incident Date`, If both columns are empty, drop the row |
| `Incident Date` | The median distance between the date and `Date Received`, if both columns are empty, drop the row |
| `Airport Code` | 1. Filling based on **modus** `Claim Site`,<br>2. If the `Claim Site` column is empty, we will fill it based on **mode** `Claim Type`,<br>3. If both `Claim Site` & `Claim Type` fields are empty, we will fill them based on **mode** of the entire data |
| `Airport Name` | 1. Filling based on **modus** `Claim Site`,<br>2. If the `Claim Site` column is empty, we will fill it based on the `Claim Type` mode,<br>3. If both `Claim Site` & `Claim Type` fields are empty, we will fill them based on **mode** of the entire data |
| `Airline Name` | Filled based on the **modus** `Airport Code` |
| `ClaimType` | Filled based on the **modus** `Airport Code` |
| `Claim Site` | Filled based on the **modus** `Claim Type` |
| `Items` | Filled by mode per `Item`. |
| `Claim Amount` | 1. Fill in the median `Claim Amount` based on **mode** `Item`,<br>2. If the `Item` column has no `Claim Amount` value at all, we will fill it based on the `Claim Type` column,<br>3. If both the `Claim Site` & `Claim Type` columns are empty, we will fill in **mode** for all data |
| `Status` | filled with `Not Available` |
| `Close Amount` | 1. If `Disposition` has a value of `Approve in Full`, it is assumed that the value of `Close Amount` is the same as `Claim Amount`.<br>2. If `Disposition` has a value of `Deny`, it assumes that the value of `Close Amount` equals 0.<br>3. If `Disposition` has the value `Settle`, the `Close Amount` value will be reviewed between the `Close Amount` and `Claim Amount` comparisons |
| `Disposition` | filled with `Not Available` |

COLUMN DESC.



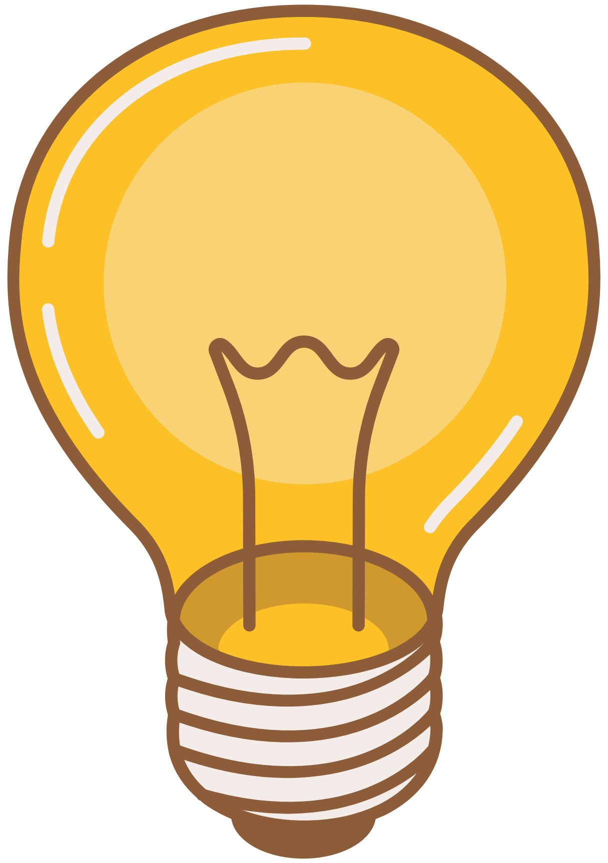
| | index | Total Null Values | Percentage |
|---|---|---|---|
| 0 | Disposition | 72885 | 35.685259 |
| 1 | Close Amount | 68929 | 33.748360 |
| 2 | Airline Name | 34373 | 16.829381 |
| 3 | Airport Code | 8523 | 4.172950 |
| 4 | Airport Name | 8523 | 4.172950 |
| 5 | Claim Type | 7913 | 3.874288 |
| 6 | Claim Amount | 4043 | 1.979495 |
| 7 | Item | 3966 | 1.941795 |
| 8 | Incident Date | 2183 | 1.068820 |
| 9 | Claim Site | 740 | 0.362312 |
| 10 | Date Received | 263 | 0.128768 |
| 11 | Status | 5 | 0.002448 |
| 12 | Claim Number | 0 | 0.000000 |

BEFORE HANDLING



| | index | Total Null Values | Percentage |
|---|---|---|---|
| 0 | Claim Number | 0 | 0.0 |
| 1 | Date Received | 0 | 0.0 |
| 2 | Incident Date | 0 | 0.0 |
| 3 | Airport Code | 0 | 0.0 |
| 4 | Airport Name | 0 | 0.0 |
| 5 | Airline Name | 0 | 0.0 |
| 6 | Claim Type | 0 | 0.0 |
| 7 | Claim Site | 0 | 0.0 |
| 8 | Item | 0 | 0.0 |
| 9 | Claim Amount | 0 | 0.0 |
| 10 | Status | 0 | 0.0 |
| 11 | Close Amount | 0 | 0.0 |
| 12 | Disposition | 0 | 0.0 |

AFTER HANDLING



# WRAP UP

- There are many ways to deal with missing values.
- The first important thing is we need to master the domain knowledge.
- Then, we must look at the many types of missing values.
- After that, we must categorize the missing values and we can do many alternatives such as dropping, inputting, or keeping it blank.
- The Most important thing to consider is the treatment of the missing value must not make the data biased.

# THANK YOU !

**CONTACT ME!**

☎ +6281227223150

✉ akbar.noorkharismawan@gmail.com

http://www.linkedin.com/in/n-k-akbar

https://github.com/baramizzo58

https://public.tableau.com/app/profile/akbar2070

Purwadhika
Digital Technology School