Data Wrangling Mini Project

For my Capstone Project #1, I downloaded a data with approximately 200.000 observations in 60 columns in which has texts in some of them. I am following the explained steps below:

1. The website had a limitation of 5000 observation for each download. In the query page, I chose date criteria for all data to be downloaded. The first file was 2019. Then I chose January-June in the first chunk, and July-December in the second chunk each year. As a result, except 2019, 1999, 1995 and 1988, all years have two files with approximately 3500 observations. 1999 and 1995 have three files. In total 64 csv files obtained with exact same formatting.

2. All files were combined by using following commands with a for loop and can be executed like;
   np.vstack([File1,File2])
   or;
   np.concatenate([File1,File2], axis=0)
   or;
   file1.append(file2).append(file3).append(file4)……append(file64)

3. After proper indexing (datetime, report accession number) and sorting, I got rid of the duplicates with the following command;
   combined_df.drop_duplicates( )

4. In order to classify the data frame according to “Events” column, I created multiple data frames. The classification of the data is as follows:
   a. Airspace conflict issues,
   b. Abnormal equipment/activity due to emergency situation,
   c. Flight cabin event,
   d. Ground issues.

   In events column 55 different explanation word taxonomy used. Each taxonomy falls under one of the classifications above.
5. For missing values;
   a. in “Events” column; “Synopsis” is including very brief information about the safety issue. I will develop some criteria to classify the row under one of the data frames declared in the previous bullet.
   b. in other columns; “Synopsis” and “Narrative” columns includes lots of information related with concerning safety issue. Most of the cells can be filled with these two columns.

c. if also “Synopsis”, “Narrative” columns are missing, I will drop the rows with following command:

df.loc [: , df.isnull( ).any( )

6. For each classified data frame, outliers supposed to be treated differently. If there is an error each outlier needs to be checked individually.