**Assignment 5**

**a) Identify the analytical data type (NOIR) of each data item; explain your reasons. Note that States_Affected is a compound data item.**

Ans:

**Year:** Ordinal and interval – The year data can be ordered and quantitative in nature. The difference between each value is the same and hence it is interval. Also, the interval scale does not give any sense of ratio between the two measurements.

**Month:** Ordinal and interval - The month data can be ordered. The difference between each month is the same and hence it is interval. Also, the interval scale does not give any sense of ratio between the two measurements.

**States_Affected - States:** - Nominal - These values are categorical in nature.

**States_Affected – Category of Storm**: Ordinal and Interval. The category is quantitative. They can be ordered and hence ordinal. The difference between each value is the same and hence it is interval. Also, the interval scale does not give any sense of ratio between the two measurements.

**Highest Category –** Ordinal and Interval - The category is quantitative. They can be ordered and hence ordinal. The difference between each value is the same and hence it is interval. Also, the interval scale does not give any sense of ratio between the two measurements.

**Central Pressure MB –** Ordinal and ratio: The central pressure is quantitative. They can be ordered and hence ordinal. A pressure of zero is significant and hence is a ratio. Ratio scales have true zero value.

**Max_winds_kt –** Ordinal and ratio - The values in Max_winds_kt is quantitative. They can be ordered and hence ordinal. A rating of zero is significant and hence is a ratio. Ratio scales have true zero value and have meaning when ratio between two measurements is calculated.

**Name –** Nominal - These values are categorical in nature and cannot be ordered.s

**b) Load the dataset into Python; display a few records.**

The below screenshots are the head and the tail of the data obtained by loading the dataset and displaying the first and the last rows.

Head:

| | Year | Month | States_Affected | Highest_Category | \ |
|---|---|---|---|---|---|
| 0 | 1851 | Jun | TX,C1 | 1 | |
| 1 | 1851 | Aug | FL,NW3;I-GA,1 | 3 | |
| 2 | 1852 | Aug | AL,3;MS,3;LA,2;FL,SW2,NW1 | 3 | |
| 3 | 1852 | Sep | FL,SW1 | 1 | |
| 4 | 1852 | Oct | FL,NW2;I-GA,1 | 2 | |

| | Central_Pressure_mb | Max_Winds_kt | Name |
|---|---|---|---|
| 0 | 977.0 | 80.0 | NaN |
| 1 | 960.0 | 100.0 | Great Middle Florida |
| 2 | 961.0 | 100.0 | Great Mobile |
| 3 | 985.0 | 70.0 | NaN |
| 4 | 969.0 | 90.0 | Middle Florida |

**Assignment 5**

Tail:

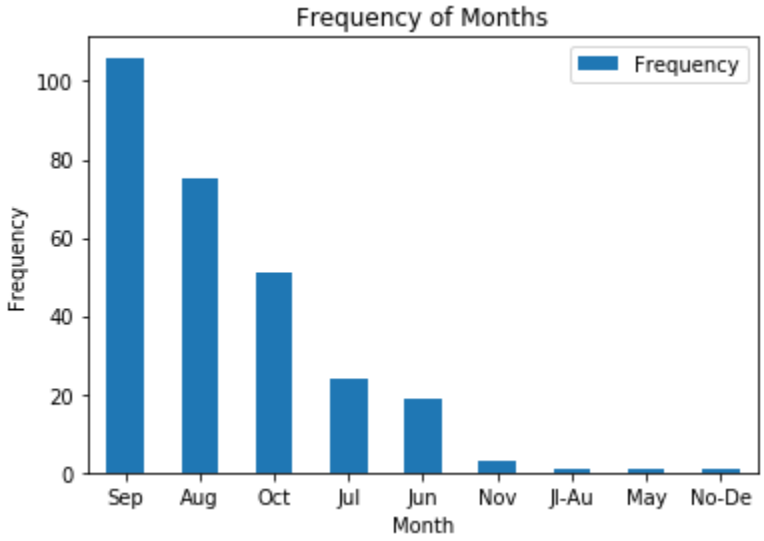
| | Year | Month | States_Affected | Highest_Category | Central_Pressure_mb | Max_Winds_kt | Name |
|---|---|---|---|---|---|---|---|
| 276 | 2005 | Aug | FL,SE1,SW1;LA,3;MS,3;AL,1 | 3 | 920.0 | 110.0 | Katrina |
| 277 | 2005 | Sep | NC,1 | 1 | 982.0 | 65.0 | Ophelia |
| 278 | 2005 | Sep | FL,SW1;LA,3;TX,N2 | 3 | 937.0 | 100.0 | Rita |
| 279 | 2005 | Oct | FL,SW3;FL,SE2 | 3 | 950.0 | 105.0 | Wilma |
| 280 | 2007 | Sep | TX,N1;LA,1 | 1 | 985.0 | 80.0 | Humberto |

**c) Create summary tables and appropriate plots for Month, and for Highest_Category. (Note: TS < Cat 1)**

**Summary table for Month:**

```
summary tables for Month
count     281
unique      9
top       Sep
freq      106
Name: Month, dtype: object
```

**Visualization for Month:**

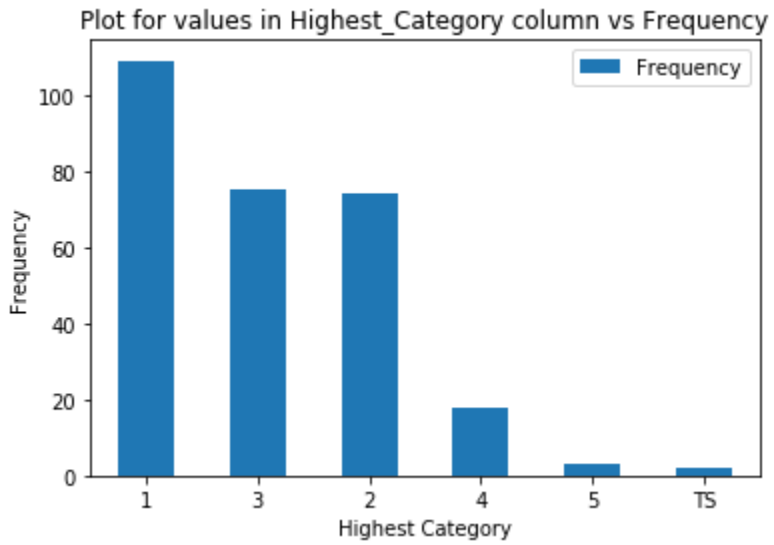


**Assignment 5**

**Summary table for Highest_Accuracy:**

```
summary tables for Highest Accuracy
count     281
unique      6
top         1
freq      109
Name: Highest_Category, dtype: object
```

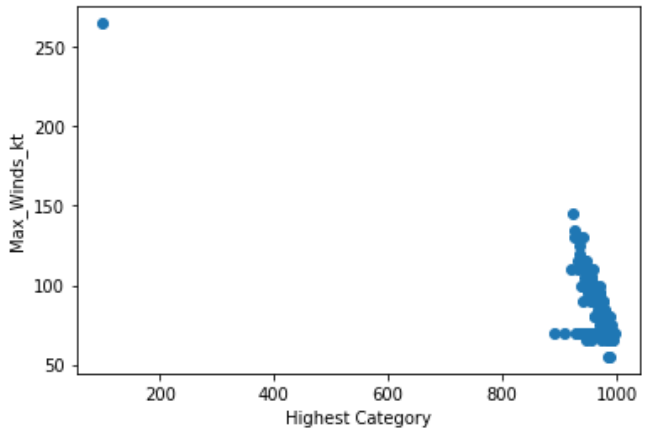
**Visualization for Highest_Accuracy:**




**d) Is there a relationship between Central_Pressure_mb and Max_Winds_kt? Explain your analysis and answer.**

**Scatterplot between Central_Pressure_mb and Max_Winds_kt**

**Assignment 5**




**Pearson's correlation coefficient between Central_Pressure_mb and Max_Winds_kt**

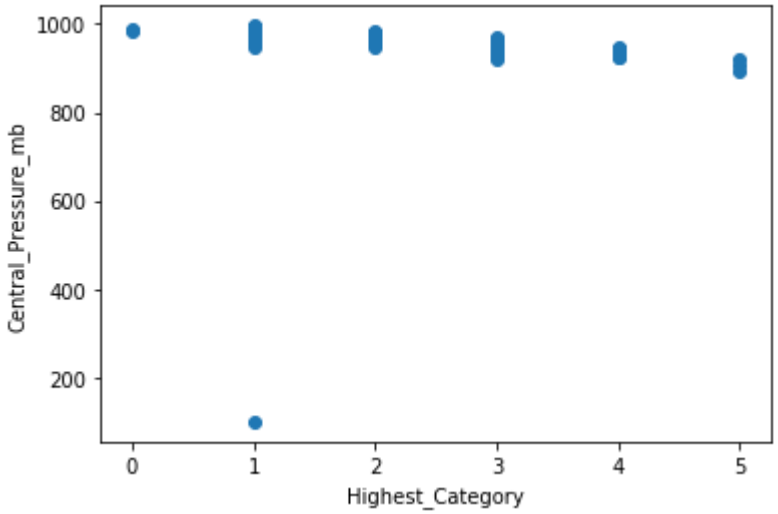
```
Pearson Correlation Coefficient Value between Central_Pressure_mb and Max_Winds_kt
-0.6869652573925864
```

**Explanation:**

The Pearson's correlation coefficient value between the features is -0.6869 which is very close to -0.7. It can be stated that the strength of correlation between the two variables is High in the negative direction when the correlation coefficient is rounded. Else, it can be considered as medium strength in the negative direction. Also, the scatterplot shows that the degree of correlation is high in the inverse direction.

**e) Is there a relationship between Highest_Category and Central_Pressure_mb? Explain your analysis and answer.**

**Scatterplot between Highest_Category and Central_Pressure_mb:**




**Pearson's correlation coefficient between Highest_Category and Central_Pressure_mb:**

```
Pearson Correlation Coefficient Value between Highest_Category and Central_Pressure_mb
-0.2306239845236448
```

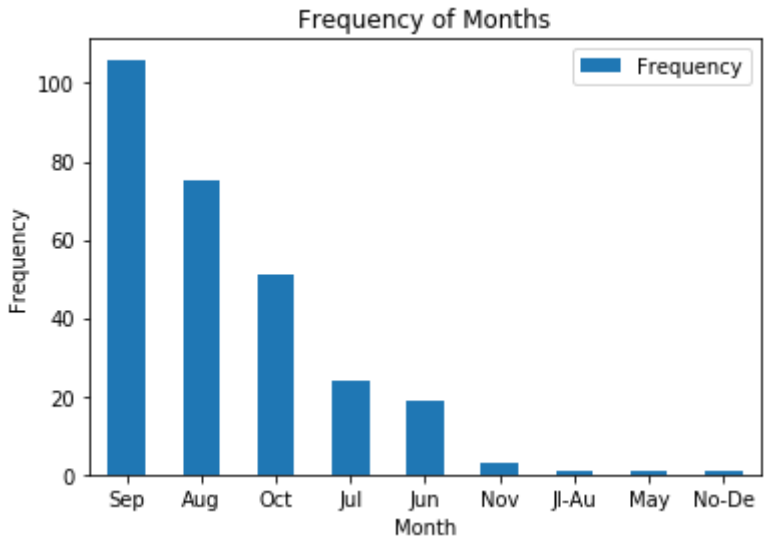
The Pearson's correlation coefficient value between the features is -0.2731. It can be stated that the strength of correlation between the two variables is very low in the negative/inverse direction. However, the scatterplot does not give away any information regarding the direction of correlation. This can be because of that the values in the '**Highest_Category'** are categorical in nature.

**f) Display a table and visualization of Months. Explain the results.**

**Table:**

| | Month | Frequency |
|---|---|---|
| 0 | Sep | 106 |
| 1 | Aug | 75 |
| 2 | Oct | 51 |
| 3 | Jul | 24 |
| 4 | Jun | 19 |
| 5 | Nov | 3 |
| 6 | Jl-Au | 1 |
| 7 | May | 1 |
| 8 | No-De | 1 |

Visualization of months:



- It is observed that the majority number of hurricanes have occurred in the month of September.
- The next highest hurricanes have been recorded in August, October and July.
- It can be stated that 82.5% of hurricanes occurred in the months of August, September and October months from the period of 1851 to 2007.
- Only 17.5% of hurricanes have occurred in the months of November, Jl-Aug, May and Nov-Dec.
- Also, the majority of the hurricanes have occurred in the months of June, July, August, September and October.
- The period from November – December experienced very a smaller number of hurricanes.
- According to the data, January, February, March, April and May did not experience any hurricanes at all.

**g) Parse and summarize the data in States_Affected; explain your method and results. Challenging!**

**Method:**

- The below are the steps preformed to extract the relevant numbers:
- By using df.apply() method, the cells of each row belonging to the column 'States_Affected' are iterated.
- For each cell, the value in it is split by ';'
- The resultant list contains the base state and a number determining the category of hurricane
- The state and the categories have been separated and stored in a dictionary.
- The dictionary was updated continuously while iterating through the rows.
- The dictionary has been converted to a dataframe so that it can be represented in a table.
- The table and its summary are displayed below.

**Assignment 5**

Parsed and Summarized the data in States_Affected

| | 1 | 2 | 3 | 4 | 5 | 0 | States |
|---|---|---|---|---|---|---|---|
| 0 | 8 | 5.0 | 3.0 | 2.0 | 0.0 | 0.0 | TX C |
| 1 | 26 | 18.0 | 14.0 | 0.0 | 0.0 | 0.0 | FL NW |
| 2 | 7 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | I-GA |
| 3 | 11 | 5.0 | 5.0 | 0.0 | 0.0 | 0.0 | AL |
| 4 | 2 | 6.0 | 8.0 | 0.0 | 1.0 | 0.0 | MS |
| 5 | 20 | 14.0 | 16.0 | 3.0 | 1.0 | 0.0 | LA |
| 6 | 17 | 10.0 | 8.0 | 4.0 | 1.0 | 0.0 | FL SW |
| 7 | 6 | 5.0 | 2.0 | 1.0 | 0.0 | 0.0 | GA |
| 8 | 7 | 7.0 | 7.0 | 1.0 | 0.0 | 0.0 | TX S |
| 9 | 17 | 7.0 | 4.0 | 2.0 | 0.0 | 0.0 | SC |
| 10 | 13 | 8.0 | 1.0 | 0.0 | 0.0 | 0.0 | FL NE |
| 11 | 3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | I-AL |
| 12 | 21 | 14.0 | 11.0 | 1.0 | 0.0 | 1.0 | NC |
| 13 | 6 | 1.0 | 5.0 | 0.0 | 0.0 | 0.0 | NY |
| 14 | 5 | 3.0 | 3.0 | 0.0 | 0.0 | 0.0 | CT |
| 15 | 3 | 2.0 | 4.0 | 0.0 | 0.0 | 0.0 | RI |
| 16 | 5 | 2.0 | 3.0 | 0.0 | 0.0 | 1.0 | MA |
| 17 | 14 | 7.0 | 3.0 | 4.0 | 0.0 | 0.0 | TX N |
| 18 | 13 | 14.0 | 11.0 | 3.0 | 1.0 | 0.0 | FL SE |
| 19 | 5 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | ME |
| 20 | 5 | 2.0 | 1.0 | 0.0 | 0.0 | 0.0 | VA |
| 21 | 1 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | MD |
| 22 | 2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | DE |
| 23 | 2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | NJ |
| 24 | 1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | I-PA |
| 25 | 2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | I-NC |
| 26 | 2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | I-VA |
| 27 | 1 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | NH |
| 28 | 1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | FL I-AL |

**Results:**

- Florida North West has experienced the largest number of hurricanes in total.
- MS, LA, FL SW, FL SE have experienced highest category of hurricane severity which is 5.
- NC and MA are the only states that have experience TS level hurricane category.
- LA experienced the second highest number of hurricanes in total across all categories.
- The mean for the 1st level hurricane category is greater than all the other categories.
- The same is true for the standard deviation too.

**Assignment 5**

- The maximum number of hurricanes for 1, 2, 3, 4 and 5 categories are 26, 18, 16, 4, 1 and 1 respectively.
- All the states in the dataset experienced a hurricane of at least 1 category.

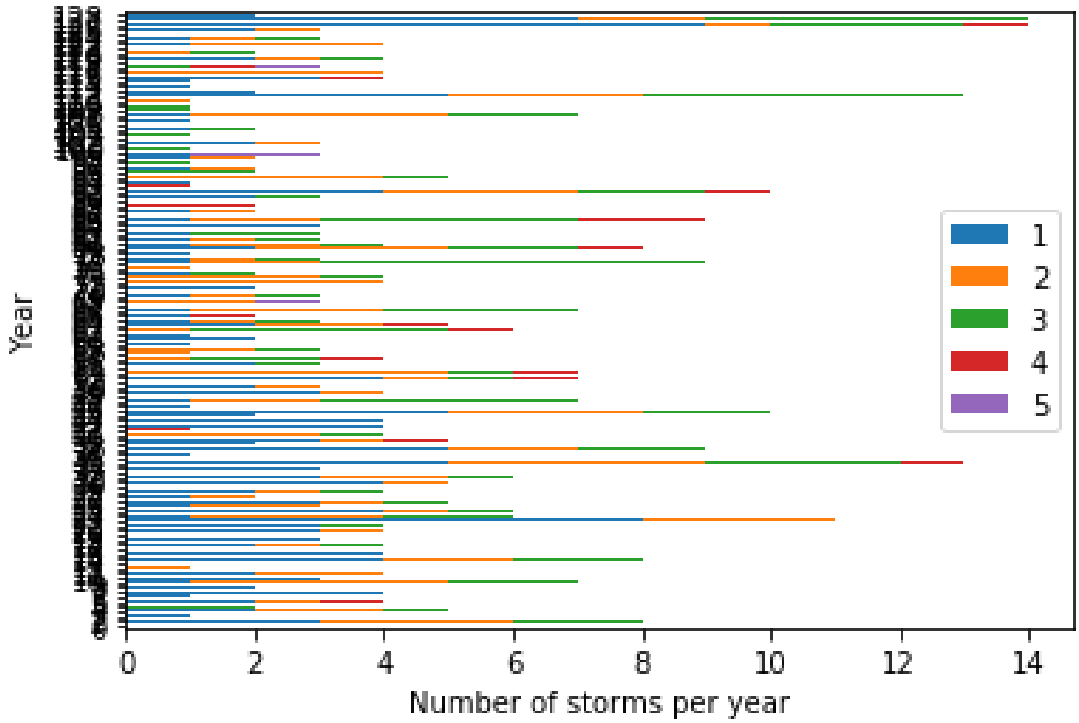
| | 1 | 2 | 3 | 4 | 5 | 0 |
|---|---|---|---|---|---|---|
| count | 29.000000 | 29.000000 | 29.000000 | 29.000000 | 29.000000 | 29.000000 |
| mean | 7.793103 | 4.586207 | 3.758621 | 0.724138 | 0.137931 | 0.068966 |
| std | 6.981503 | 5.172159 | 4.572083 | 1.278854 | 0.350931 | 0.257881 |
| min | 1.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 25% | 2.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 50% | 5.000000 | 2.000000 | 3.000000 | 0.000000 | 0.000000 | 0.000000 |
| 75% | 13.000000 | 7.000000 | 5.000000 | 1.000000 | 0.000000 | 0.000000 |
| max | 26.000000 | 18.000000 | 16.000000 | 4.000000 | 1.000000 | 1.000000 |

**h) Create a table and a visualization showing the number of storms per year for each category. Be creative!**

| Number of storms per year for each category | | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 3 | 2 | 4 | 0 | 5 | Year |
| 0 | 2.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1851 |
| 1 | 3.0 | 2.0 | 3.0 | 0.0 | 0.0 | 0.0 | 1852 |
| 2 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1853 |
| 3 | 2.0 | 1.0 | 2.0 | 0.0 | 0.0 | 0.0 | 1854 |
| 4 | 0.0 | 2.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1855 |
| .. | ... | ... | ... | ... | ... | ... | ... |
| 124 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 2002 |
| 125 | 2.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 2003 |
| 126 | 9.0 | 3.0 | 1.0 | 1.0 | 0.0 | 0.0 | 2004 |
| 127 | 7.0 | 5.0 | 2.0 | 0.0 | 0.0 | 0.0 | 2005 |
| 128 | 2.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 2007 |

Visualization

**Assignment 5**




Due to large number of years, the values on x-zxis have become congested. This can be removed by controlling the xticks.

**i) Create a table and a visualization showing the number of storms per state for each category. Be creative**

**Table showing the number of storms per state for each category**

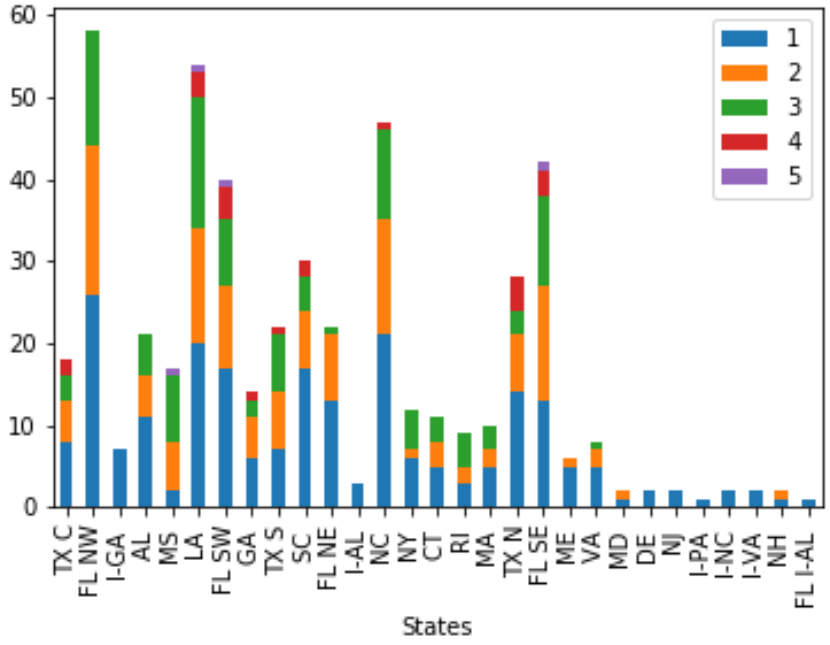
**Assignment 5**

Parsed and Summarized the data in States_Affected

| | 1 | 2 | 3 | 4 | 5 | 0 | States |
|---|---|---|---|---|---|---|---|
| 0 | 8 | 5.0 | 3.0 | 2.0 | 0.0 | 0.0 | TX C |
| 1 | 26 | 18.0 | 14.0 | 0.0 | 0.0 | 0.0 | FL NW |
| 2 | 7 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | I-GA |
| 3 | 11 | 5.0 | 5.0 | 0.0 | 0.0 | 0.0 | AL |
| 4 | 2 | 6.0 | 8.0 | 0.0 | 1.0 | 0.0 | MS |
| 5 | 20 | 14.0 | 16.0 | 3.0 | 1.0 | 0.0 | LA |
| 6 | 17 | 10.0 | 8.0 | 4.0 | 1.0 | 0.0 | FL SW |
| 7 | 6 | 5.0 | 2.0 | 1.0 | 0.0 | 0.0 | GA |
| 8 | 7 | 7.0 | 7.0 | 1.0 | 0.0 | 0.0 | TX S |
| 9 | 17 | 7.0 | 4.0 | 2.0 | 0.0 | 0.0 | SC |
| 10 | 13 | 8.0 | 1.0 | 0.0 | 0.0 | 0.0 | FL NE |
| 11 | 3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | I-AL |
| 12 | 21 | 14.0 | 11.0 | 1.0 | 0.0 | 1.0 | NC |
| 13 | 6 | 1.0 | 5.0 | 0.0 | 0.0 | 0.0 | NY |
| 14 | 5 | 3.0 | 3.0 | 0.0 | 0.0 | 0.0 | CT |
| 15 | 3 | 2.0 | 4.0 | 0.0 | 0.0 | 0.0 | RI |
| 16 | 5 | 2.0 | 3.0 | 0.0 | 0.0 | 1.0 | MA |
| 17 | 14 | 7.0 | 3.0 | 4.0 | 0.0 | 0.0 | TX N |
| 18 | 13 | 14.0 | 11.0 | 3.0 | 1.0 | 0.0 | FL SE |
| 19 | 5 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | ME |
| 20 | 5 | 2.0 | 1.0 | 0.0 | 0.0 | 0.0 | VA |
| 21 | 1 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | MD |
| 22 | 2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | DE |
| 23 | 2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | NJ |
| 24 | 1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | I-PA |
| 25 | 2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | I-NC |
| 26 | 2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | I-VA |
| 27 | 1 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | NH |
| 28 | 1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | FL I-AL |

**Assignment 5**

**Visualization showing the number of storms per state for each category**




**j) Many records have missing values; explain. What you did to address that problem.**

- It was observed that the majority of the rows containing the NaN values were associated with the columns, 'Central_Pressure_mb', 'Max_Winds_kt' and Name.
- The column 'Name' has a lot of null values. Moreover, it is not being used in analysis and hence is completely dropped from the dataframe.
- The null values in 'Central_Pressure_mb' and 'Max_Winds_kt' are replaced by their mode values. We cannot replace it with their respective mean values. This is because, the presence of even a single large value drifts the mean of the column towards itself. Mode is a safer option, as it is related to the number of occurrences of a value.
- Mode is more suitable for data present in 'Central_Pressure_mb' and 'Max_Winds_kt' columns.
- Also, the string TS has been replaced with zero to facilitate mathematical calculations.
- The data type of "Highest_Category" has been change from str to int to facilitate the same.

```
columns containing null vaules
Index(['Central_Pressure_mb', 'Max_Winds_kt', 'Name'], dtype='object')
Number of rows before dropping the column
281
Number of rows after dropping the column
281
```