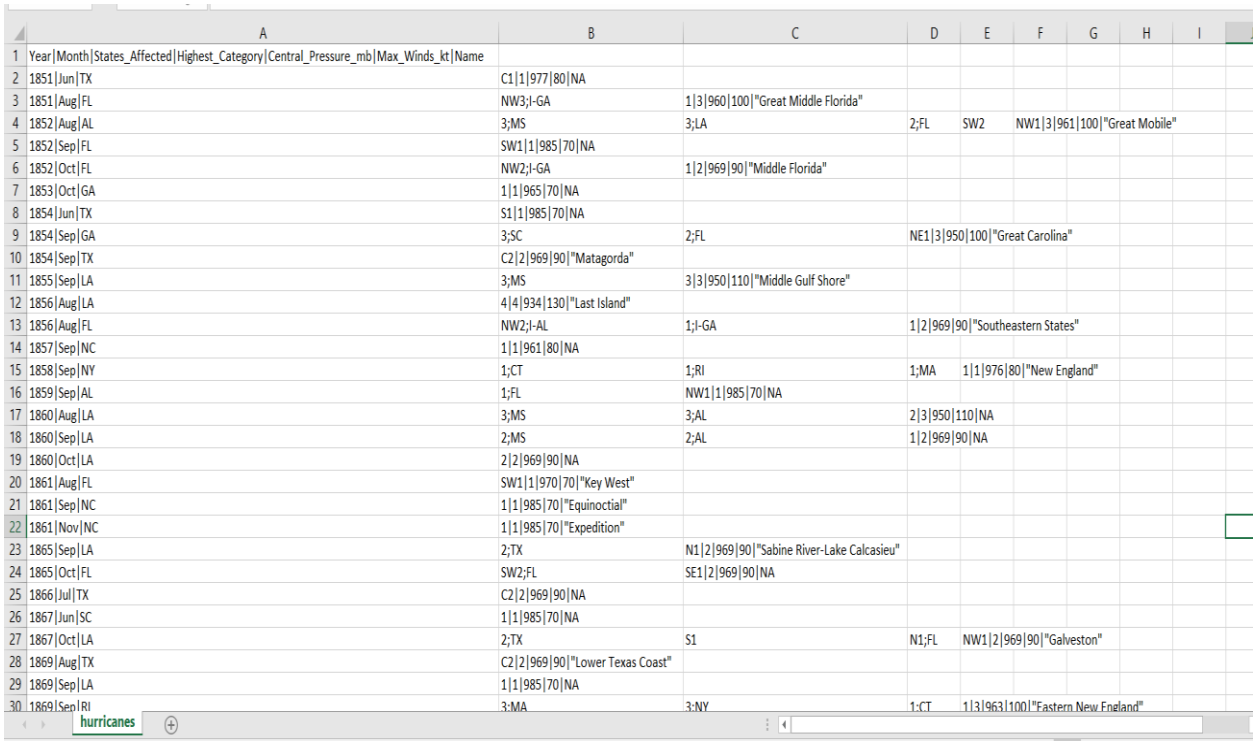
Assignment 6 - Data Exploration with Excel

- **Review IDMA Chapter 5 and author slide presentation**

Reviewed IDMA Chapter 5 and author slide presentation

- **Using the `hurricanes.csv` dataset, load the data into an Excel spreadsheet.**
  - **Display a screenshot of the spreadsheet (don't submit the entire spreadsheet)**

Step 1: If opened the .csv file directly in Excel:

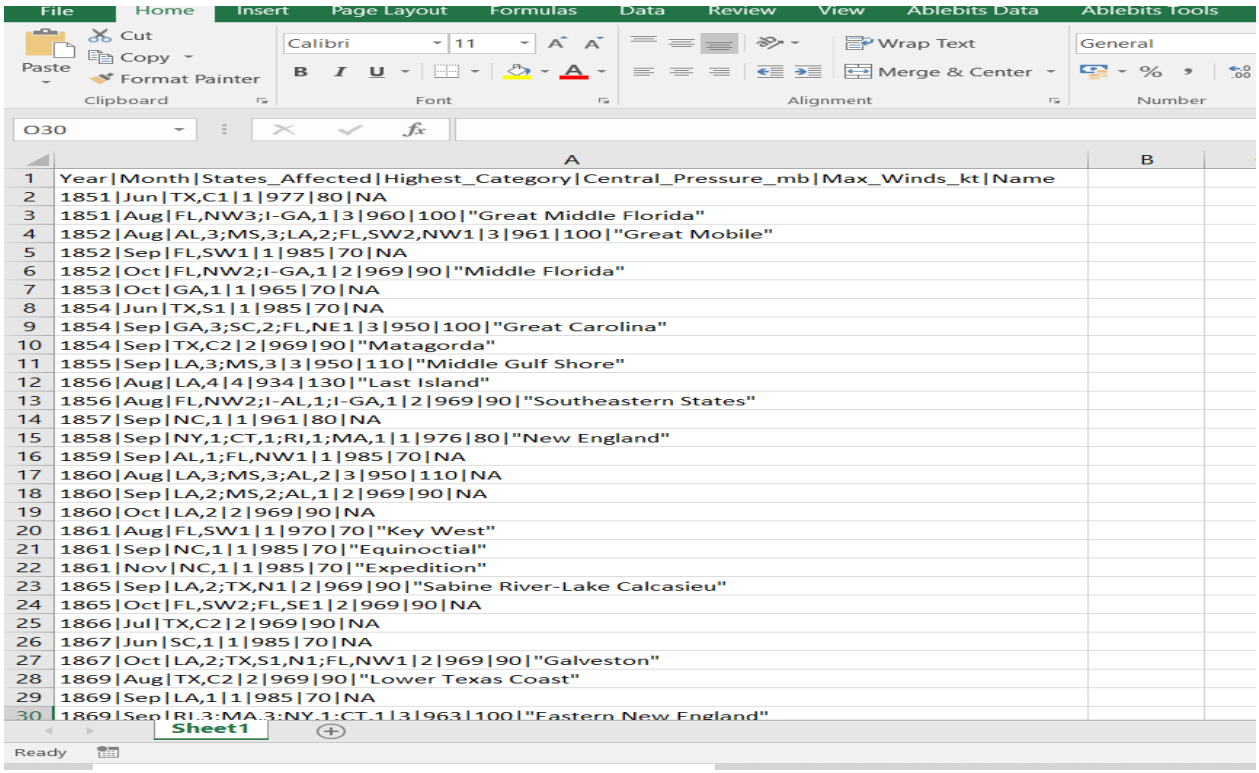


	A	B	C	D	E	F	G	H	I
1	Year|Month|States_Affected|Highest_Category|Central_Pressure_mb|Max_Winds_kt|Name								
2	1851|Jun|TX	C1|1|977|80|NA							
3	1851|Aug|FL	NW3;I-GA	1|3|960|100|"Great Middle Florida"						
4	1852|Aug|AL	3;MS	3;LA	2;FL	SW2	NW1|3|961|100|"Great Mobile"			
5	1852|Sep|FL	SW1|1|985|70|NA							
6	1852|Oct|FL	NW2;I-GA	1|2|969|90|"Middle Florida"						
7	1853|Oct|GA	1|1|965|70|NA							
8	1854|Jun|TX	S1|1|985|70|NA							
9	1854|Sep|GA	3;SC	2;FL	NE1|3|950|100|"Great Carolina"					
10	1854|Sep|TX	C2|2|969|90|"Matagorda"							
11	1855|Sep|LA	3;MS	3|3|950|110|"Middle Gulf Shore"						
12	1856|Aug|LA	4|4|934|130|"Last Island"							
13	1856|Aug|FL	NW2;I-AL	1;I-GA	1|2|969|90|"Southeastern States"					
14	1857|Sep|NC	1|1|961|80|NA							
15	1858|Sep|NY	1;CT	1;RI	1;MA	1|1|976|80|"New England"				
16	1859|Sep|AL	1;FL	NW1|1|985|70|NA						
17	1860|Aug|LA	3;MS	3;AL	2|3|950|110|NA					
18	1860|Sep|LA	2;MS	2;AL	1|2|969|90|NA					
19	1860|Oct|LA	2|2|969|90|NA							
20	1861|Aug|FL	SW1|1|970|70|"Key West"							
21	1861|Sep|NC	1|1|985|70|"Equinoctial"							
22	1861|Nov|NC	1|1|985|70|"Expedition"							
23	1865|Sep|LA	2;TX	N1|2|969|90|"Sabine River-Lake Calcasieu"						
24	1865|Oct|FL	SW2;FL	SE1|2|969|90|NA						
25	1866|Jul|TX	C2|2|969|90|NA							
26	1867|Jun|SC	1|1|985|70|NA							
27	1867|Oct|LA	2;TX	S1	N1;FL	NW1|2|969|90|"Galveston"				
28	1869|Aug|TX	C2|2|969|90|"Lower Texas Coast"							
29	1869|Sep|LA	1|1|985|70|NA							
30	1869|Sep|RI	3;MA	3;NY	1;CT	1|3|963|100|"Eastern New England"				

hurricanes

Mismatched attribute names and its values can be seen above.

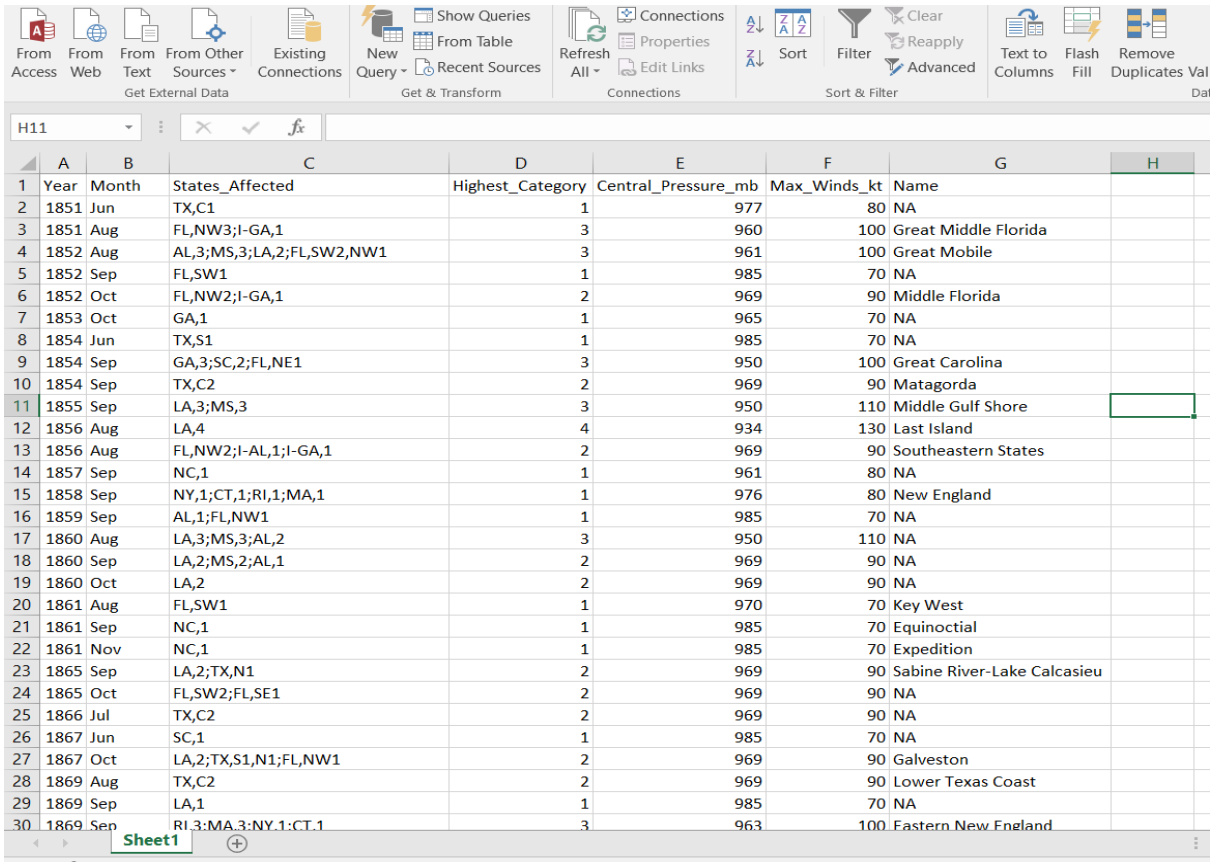
Step 2: Opened the hurricanes.csv file in notepad and pasted this data into a new Microsoft Excel Worksheet (.xlsx) and the data is loaded as following:



	A
1	Year|Month|States_Affected|Highest_Category|Central_Pressure_mb|Max_Winds_kt|Name
2	1851|Jun|TX,C1|1|977|80|NA
3	1851|Aug|FL,NW3;I-GA,1|3|960|100|"Great Middle Florida"
4	1852|Aug|AL,3;MS,3;LA,2;FL,SW2,NW1|3|961|100|"Great Mobile"
5	1852|Sep|FL,SW1|1|985|70|NA
6	1852|Oct|FL,NW2;I-GA,1|2|969|90|"Middle Florida"
7	1853|Oct|GA,1|1|965|70|NA
8	1854|Jun|TX,S1|1|985|70|NA
9	1854|Sep|GA,3;SC,2;FL,NE1|3|950|100|"Great Carolina"
10	1854|Sep|TX,C2|2|969|90|"Matagorda"
11	1855|Sep|LA,3;MS,3|3|950|110|"Middle Gulf Shore"
12	1856|Aug|LA,4|4|934|130|"Last Island"
13	1856|Aug|FL,NW2;I-AL,1;I-GA,1|2|969|90|"Southeastern States"
14	1857|Sep|NC,1|1|961|80|NA
15	1858|Sep|NY,1;CT,1;RI,1;MA,1|1|976|80|"New England"
16	1859|Sep|AL,1;FL,NW1|1|985|70|NA
17	1860|Aug|LA,3;MS,3;AL,2|3|950|110|NA
18	1860|Sep|LA,2;MS,2;AL,1|2|969|90|NA
19	1860|Oct|LA,2|2|969|90|NA
20	1861|Aug|FL,SW1|1|970|70|"Key West"
21	1861|Sep|NC,1|1|985|70|"Equinoctial"
22	1861|Nov|NC,1|1|985|70|"Expedition"
23	1865|Sep|LA,2;TX,N1|2|969|90|"Sabine River-Lake Calcasieu"
24	1865|Oct|FL,SW2;FL,SE1|2|969|90|NA
25	1866|Jul|TX,C2|2|969|90|NA
26	1867|Jun|SC,1|1|985|70|NA
27	1867|Oct|LA,2;TX,S1,N1;FL,NW1|2|969|90|"Galveston"
28	1869|Aug|TX,C2|2|969|90|"Lower Texas Coast"
29	1869|Sep|LA,1|1|985|70|NA
30	1869|Sep|RI,3;MA,3;NY,1;CT,1|3|963|100|"Eastern New England"

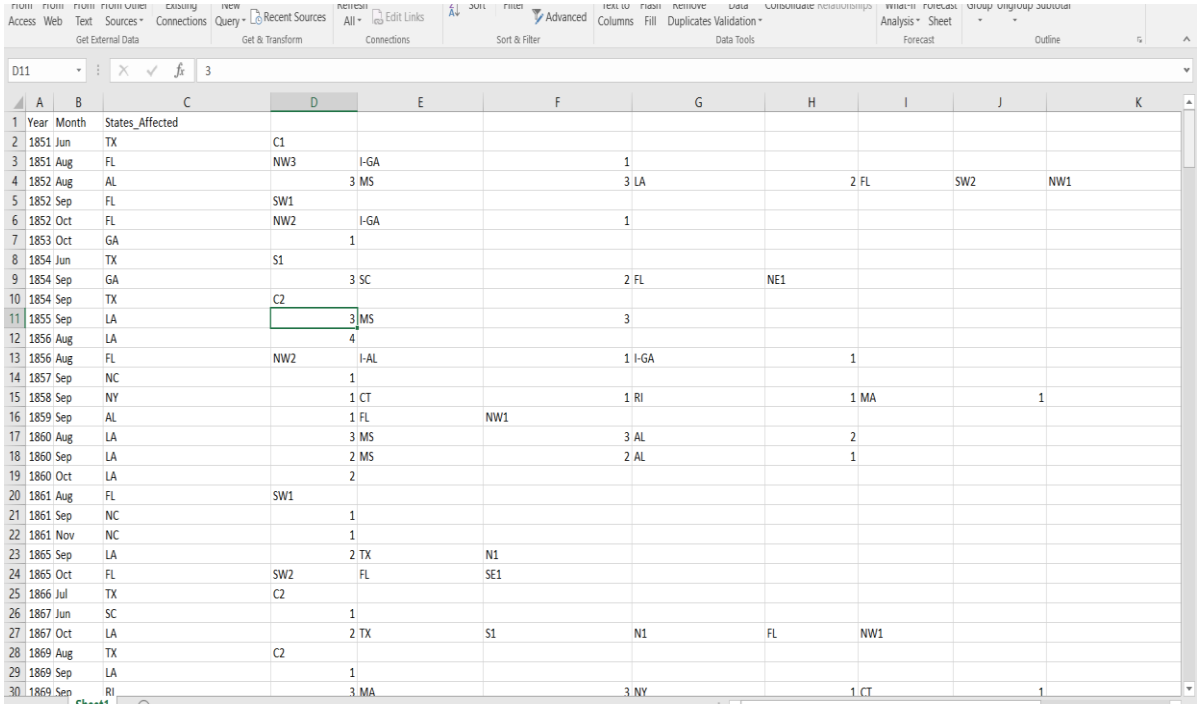
Entire Data is fallen into a single column.

Step 3: So in order to separate each column, selected the entire range and used Text to Columns function under Data tab in Excel with delimiter '|'. Hence the following:



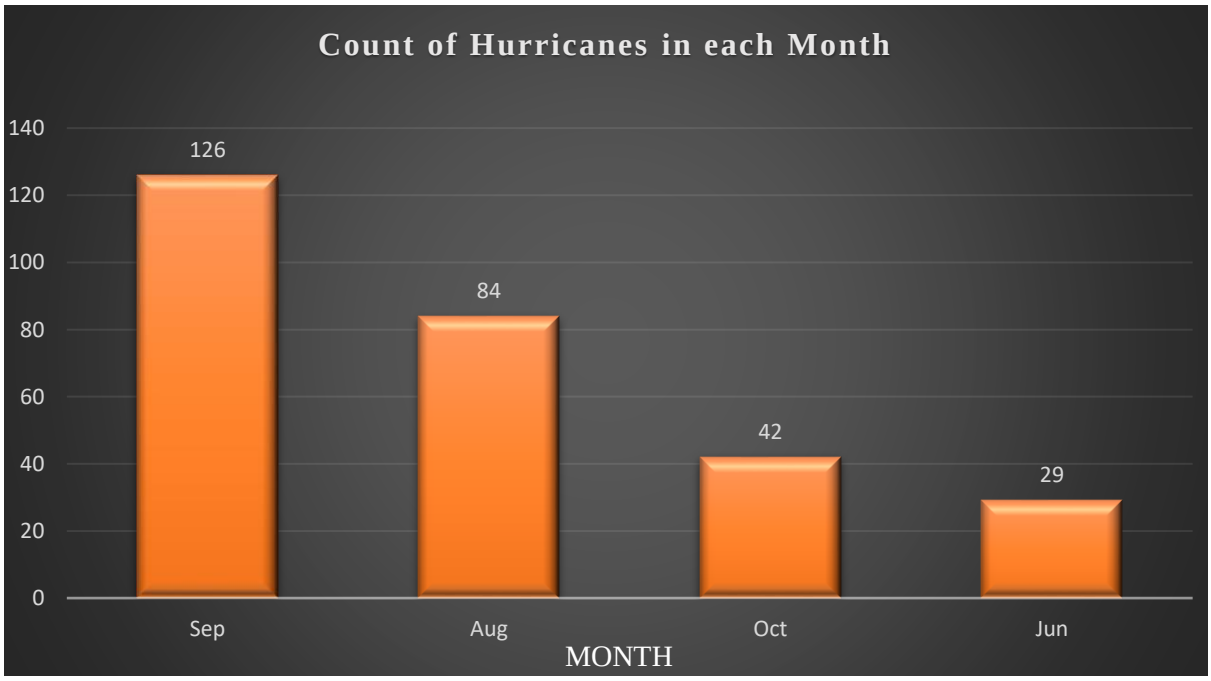
	A	B	C	D	E	F	G	H
1	Year	Month	States_Affected	Highest_Category	Central_Pressure_mb	Max_Winds_kt	Name	
2	1851	Jun	TX,C1	1	977	80	NA	
3	1851	Aug	FL,NW3;I-GA,1	3	960	100	Great Middle Florida	
4	1852	Aug	AL,3;MS,3;LA,2;FL,SW2,NW1	3	961	100	Great Mobile	
5	1852	Sep	FL,SW1	1	985	70	NA	
6	1852	Oct	FL,NW2;I-GA,1	2	969	90	Middle Florida	
7	1853	Oct	GA,1	1	965	70	NA	
8	1854	Jun	TX,S1	1	985	70	NA	
9	1854	Sep	GA,3;SC,2;FL,NE1	3	950	100	Great Carolina	
10	1854	Sep	TX,C2	2	969	90	Matagorda	
11	1855	Sep	LA,3;MS,3	3	950	110	Middle Gulf Shore	
12	1856	Aug	LA,4	4	934	130	Last Island	
13	1856	Aug	FL,NW2;I-AL,1;I-GA,1	2	969	90	Southeastern States	
14	1857	Sep	NC,1	1	961	80	NA	
15	1858	Sep	NY,1;CT,1;RI,1;MA,1	1	976	80	New England	
16	1859	Sep	AL,1;FL,NW1	1	985	70	NA	
17	1860	Aug	LA,3;MS,3;AL,2	3	950	110	NA	
18	1860	Sep	LA,2;MS,2;AL,1	2	969	90	NA	
19	1860	Oct	LA,2	2	969	90	NA	
20	1861	Aug	FL,SW1	1	970	70	Key West	
21	1861	Sep	NC,1	1	985	70	Equinoctial	
22	1861	Nov	NC,1	1	985	70	Expedition	
23	1865	Sep	LA,2;TX,N1	2	969	90	Sabine River-Lake Calcasieu	
24	1865	Oct	FL,SW2;FL,SE1	2	969	90	NA	
25	1866	Jul	TX,C2	2	969	90	NA	
26	1867	Jun	SC,1	1	985	70	NA	
27	1867	Oct	LA,2;TX,S1,N1;FL,NW1	2	969	90	Galveston	
28	1869	Aug	TX,C2	2	969	90	Lower Texas Coast	
29	1869	Sep	LA,1	1	985	70	NA	
30	1869	Sep	RI,3;MA,3;NY,1;CT,1	3	963	100	Eastern New England	

Step 4: When the States_Affected column is further divided into different columns using delimiters ’,’ and ‘; ‘ the following is displayed with many empty cells



Year	Month	States_Affected								
1851	Jun	TX	C1							
1851	Aug	FL	NW3	I-GA		1				
1852	Aug	AL	3	MS		3	LA	2	FL	SW2	NW1
1852	Sep	FL	SW1							
1852	Oct	FL	NW2	I-GA		1				
1853	Oct	GA	1							
1854	Jun	TX	S1							
1854	Sep	GA	3	SC		2	FL	NE1		
1854	Sep	TX	C2							
1855	Sep	LA	3	MS		3				
1856	Aug	LA	4							
1856	Aug	FL	NW2	I-AL		1	I-GA	1		
1857	Sep	NC	1							
1858	Sep	NY	1	CT		1	RI	1	MA	1
1859	Sep	AL	1	FL	NW1					
1860	Aug	LA	3	MS		3	AL	2		
1860	Sep	LA	2	MS		2	AL	1		
1860	Oct	LA	2							
1861	Aug	FL	SW1							
1861	Sep	NC	1							
1861	Nov	NC	1							
1865	Sep	LA	2	TX	N1					
1865	Oct	FL	SW2	FL	SE1					
1866	Jul	TX	C2							
1867	Jun	SC	1							
1867	Oct	LA	2	TX	S1		N1	FL	NW1	
1869	Aug	TX	C2							
1869	Sep	LA	1							
1869	Sep	RI	3	MA		3	NY	1	CT	1

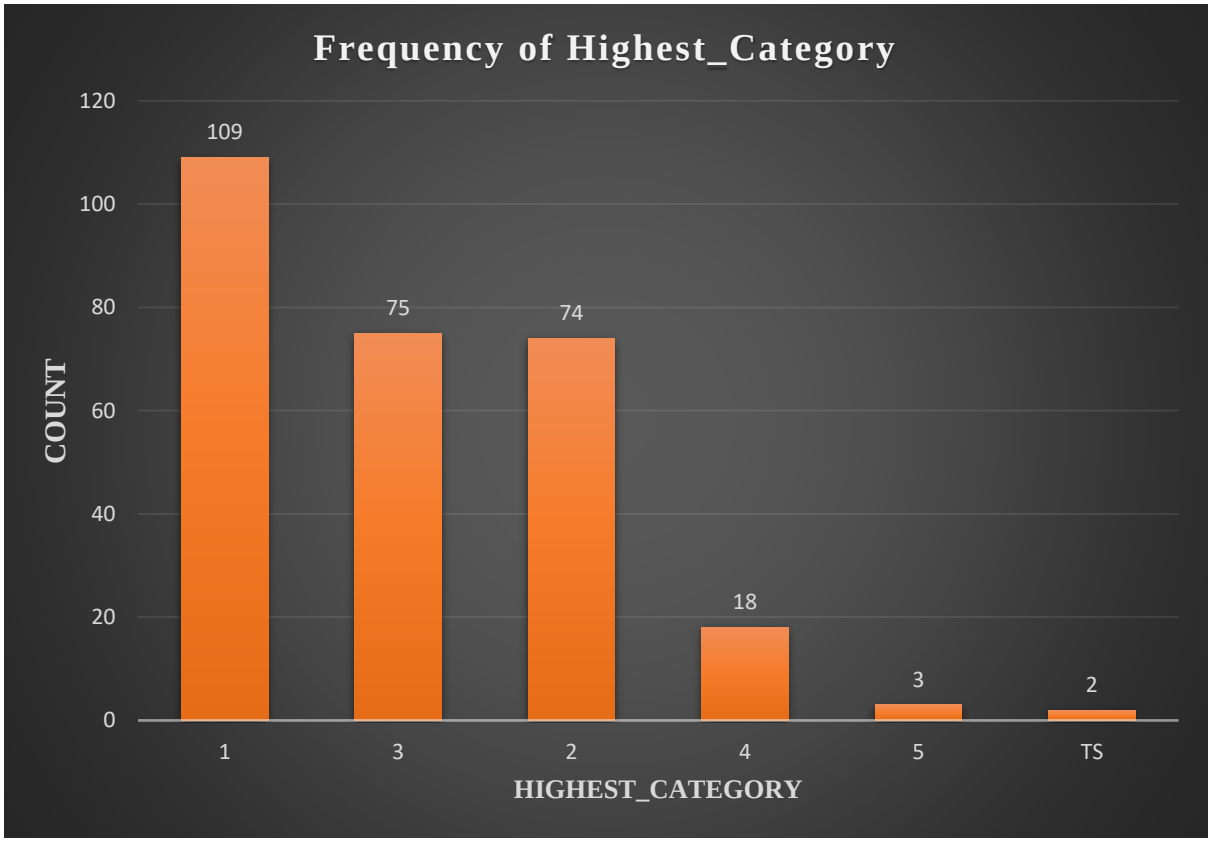
- **Prepare visualizations showing:**
  - **the number of hurricanes for each month**



Month	Count of Month
Jun	29
Aug	84
Sep	126
Oct	42

September has the highest hurricanes and June had the lowest count.

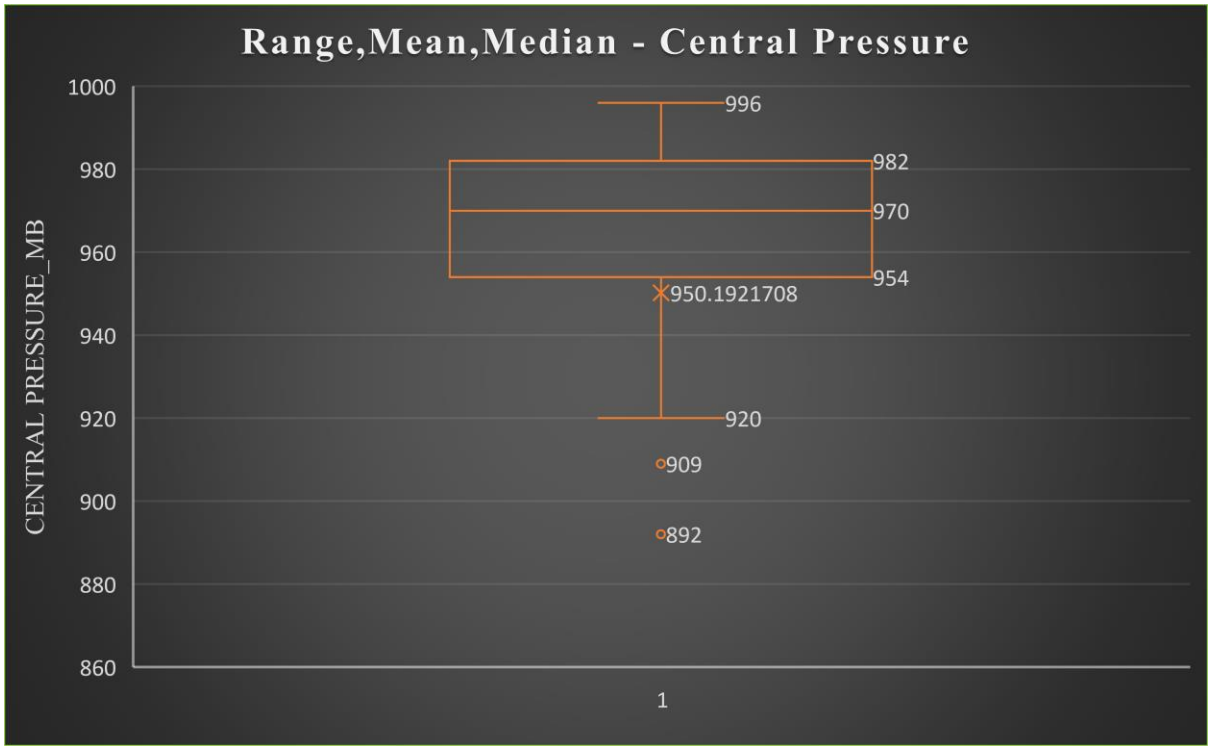
- **the frequency of highest categories**



Highest_Category	Count of Highest_Category
1	109
2	74
3	75
4	18
5	3
TS	2

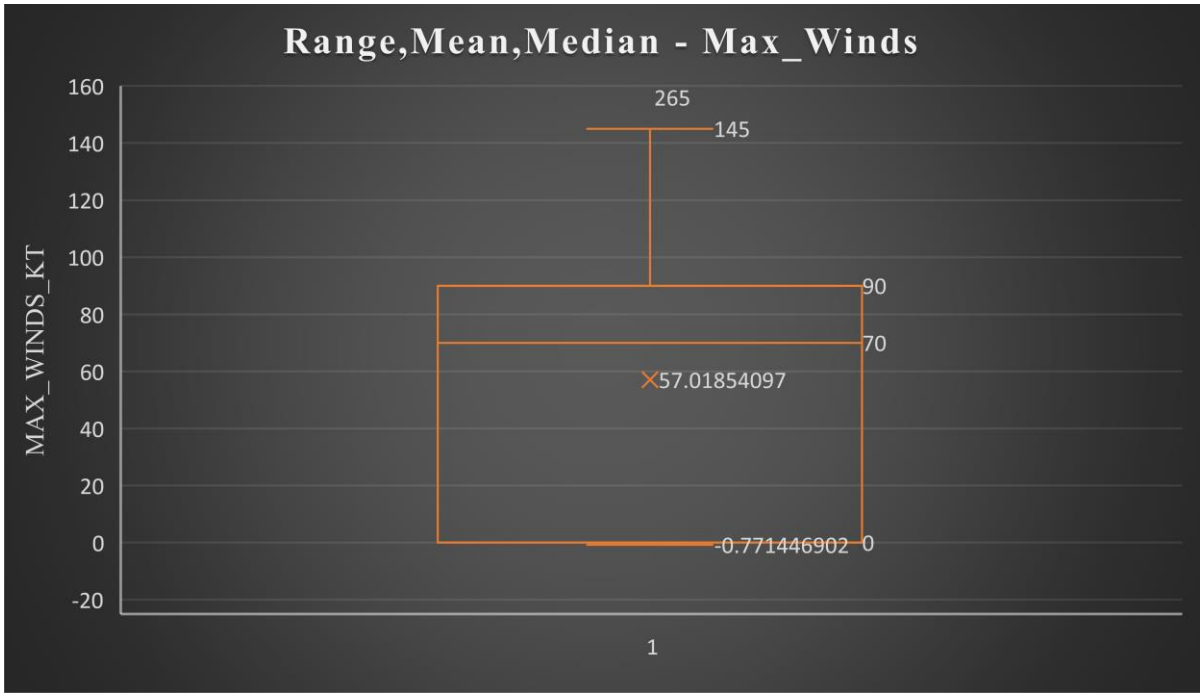
Most of the hurricanes fall under Category 1 with least in Category ‘TS’ – Tropical Storm

- **the range, mean, and median for central pressure and for max winds**



Range = max-min = 996 – 920 = 76

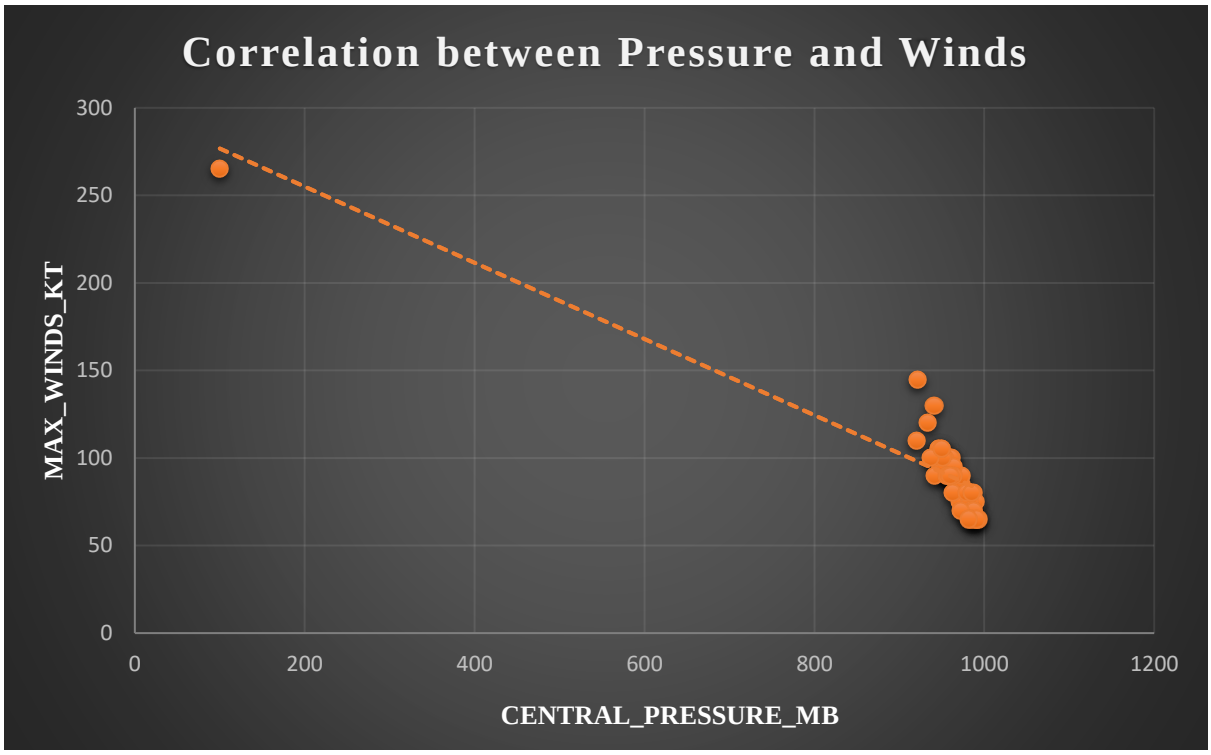
Mean – 950.2 , Median - 970



Range = max-min = 265 – 0 = 265

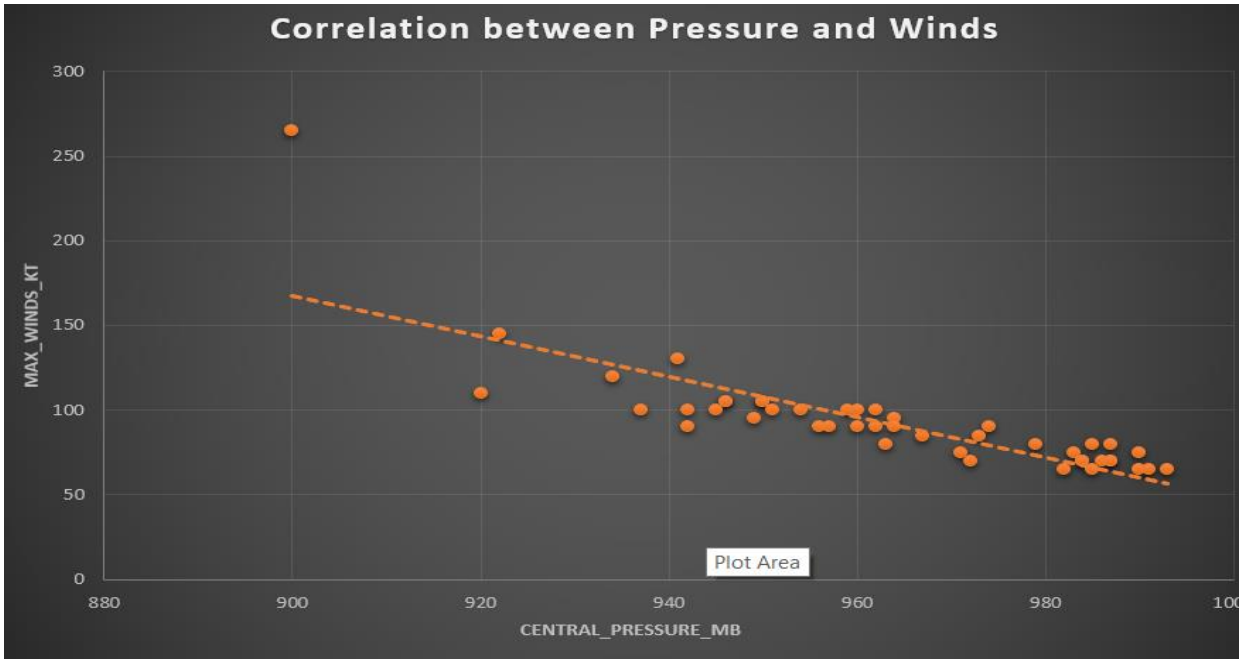
Mean – 57.2 , Median – 70

- **the relationship between central pressure and max winds**



=CORREL(E2:E282,F2:F282)

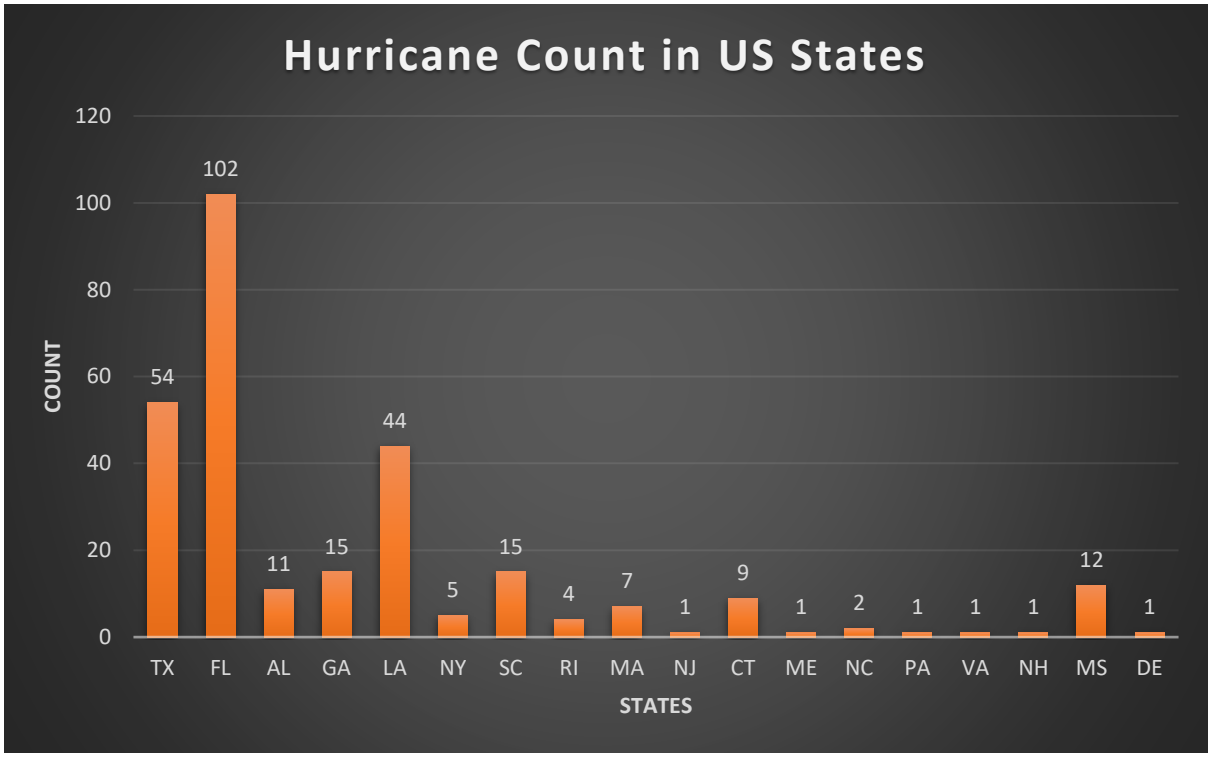
Correlation Coefficient is -0.771446902. Here the correlation coefficient is negative, which means there is a negative linear relationship between max_winds and the central pressure. If we remove that one outlier at central pressure 100(which may be because incorrect data), then the plot looks as follows, where the negative linear relationship between two continuous variables can be seen clearly.



This almost moderate negative correlation signifies that as the pressure increases, the max_winds decreases (and vice versa).

- **the frequency of occurrence for each state (think!)**

Florida has the highest hurricanes and next comes Texas followed by LA according to the below chart.



state	hurricane count
TX	54
FL	102
AL	11
GA	15
LA	44
NY	5
SC	15
RI	4
MA	7
NJ	1
CT	9

ME	1
NC	2
PA	1
VA	1
NH	1
MS	12
DE	1