AIT – 580 PROJECT

Asmita Singh

Deliverable 1 – Dataset Selection

Briefly describe the dataset: *size* (required storage), metadata (data items' meanings and types), structure. Who (company, agency, organization) collected the data? Who they are, what do they do? What is their role/purpose?

The dataset selected is based on United States Tornadoes from 1950-2018. It has been collected from NOAA's national Weather Service Storm Prediction Center.

The Storm Prediction Center (SPC) is a part of the National Weather Service (NWS) and is one of nine National Centers for Environmental Prediction. Their mission is to provide timely and accurate forecasts and watches for severe thunderstorms and tornadoes over the contiguous United States. The SPC also issues forecasts for hazardous winter and fire weather.[1]

Link for dataset - https://www.spc.noaa.gov/wcm/#data
Name of the dataset – U.S. TORNADOES* (1950-2018)
Year of the dataset – 2018
Size – 6977 KB
Number of Records – 64826
Number of Columns – 29

Name of the Columns and its datatype -

Column Names	Datatypes	Column Names	Datatypes
om	Nominal	slat	Interval
yr	Interval	slon	Interval
mo	Interval	elat	Interval
dy	Interval	elon	Interval
date	Interval	len	Ratio
time	Interval	wid	ratio
tz	Nominal	ns	Ordinal
st	Nominal	sn	Ordinal
stf	Nominal	sg	Ordinal
stn	Ratio	f1	Nominal
mag	Ordinal	f2	Nominal
inj	Ratio	f3	Nominal
fat	Ratio	f4	Nominal
loss	Ratio	fc	Ordinal
closs	Ratio		

[1] NOAA's national Weather Service Storm Prediction Center [Website]. (n.d.). Retrieved October 20, 2019, from https://www.spc.noaa.gov/faq/#1.1

Metadata Details

Event Details File (1950-2018_all_tornadoes.csv) [2] -

1. **om –** A count of the number of tornadoes during the year. However, before 2007, these numbers were assigned as the information arrived in the NWS database. Since 2007, the numbers have been assigned in a sequential manner after event date-times are converted to CST. This field should not be used to count the number of tornadoes. Example 1,2,3
2. **yr** – Year, four digits are recorded in this field. Example 2018, 2007
3. **mo** – The number of the month for the event in this record. The values range from 1-12 (1=January, 12=December)
4. **dy** – The number of days in the month for the event in this record. The value ranges from 1-31.
5. **date** – the date when the tornado was recorded by SPC. The format is mm/dd/yyyy.
6. **time** – the time when the storm was recorded by SPC. The format is hh:mm:ss
7. **tz** – Time Zone for the County. Eastern Standard Time (EST), Central Standard Time (CST), Mountain Standard Time (MST), Greenwich Mean Time (GMT). All times, except for, ?=unknown and 9=GMT, were converted to 3=CST.
8. **st** – The state name where the event occurred. It is a two-letter postal abbreviation, PR=Puerto Rico, VI=Virgin Islands.
9. **stf** - The FIPS number of the county entered by the continuing tornado segment as it crossed from one county to another. The following FIPS number is provided within this field. Example, 1, 72, 56.
10. **stn** – State Number – Number of tornados, in this state, in this year: May not be sequential in some years, discontinued in 2008. This number can be calculated in the spreadsheet by sorting and after accounting for border crossing tornadoes and 4+ county segments.
11. **mag** - the measured extent of the magnitude type of tornado. values -9,0,1,2,3,4,5 (-9=unknown)
12. **inj -** Number of injuries related to the event of a tornado for the year
13. **fat** - Number of fatalities related to the event of a tornado for the year.
14. **loss** - Estimated property loss information in millions of dollars occurred by the tornado event. Prior to 1996, this is a categorization of tornado damage by dollar amount **(**0 = Unknown ,1 < $50 ,2 = $50-500 ,3 = $500-5,000 ,4 = $5,000-50,000 ,5 = $50,000-$500,000 ,6 = $500,000-$5,000,000 ,7 = $5,000,000-$50,000,000 ,8 = $50,000,000-500,000,000 ,9 = $5000,000,000).
15. **closs** - Estimated crop loss in millions of dollars (started in 2007)
16. **slat** - Starting latitude in decimal degrees where the event occurred, includes '-' if its south of the equator.
17. **slon** - Starting longitude in decimal degrees where the event occurred, includes '-' if its West of the Prime Meridian.
18. **elat** - Ending latitude in decimal degrees where the event occurred, includes '-' if its south of the equator.
19. **elon** - Ending longitude in decimal degrees where the event occurred, includes '-' if its West of the Prime Meridian.
20. **len** - Length of the tornado or tornado segment while on the ground (in miles). Ex: 0.66, 1.05, 0.48

[2] NOAA's national Weather Service Storm Prediction Center [Website]. (n.d.). Retrieved October 20, 2019, from https://www.spc.noaa.gov/wcm/data/SPC_severe_database_description.pdf

21. **wid** - Width of the tornado or tornado segment while on the ground (in yards). Ex: 150, 350
22. **ns** - Number of states affected by this tornado; 1,2 or 3
23. **sn** - State number 1 or 0
24. **sg** - Tornado segment number: 1,2 or -9
25. **f1** - 1st Country FIPS code
26. **f2** - 2nd Country FIPS code
27. **f3** - 3rd Country FIPS code
28. **f4** - 4th Country FIPS code
29. **fc** - fc=0 for unaltered scale rating, fc=1 if previous rating was -9 (unknown). Valid for records altered between 1950-1982.

Describe any privacy, quality, ethical, or other issues with this dataset

Privacy – The dataset has been retrieved from open data sources https://www.spc.noaa.gov/wcm/#data. It has tornado information and the destruction caused by the tornado, which is not a piece of private information. Hence no privacy issue is present.

Quality – There are no special characters or null columns present in the dataset. The tornado length is given in miles and width in yards. Data unit conversion would be required for either of them for analysis.

Ethical – The primary stakeholder (who identified the data) is identified here to support the analysis (SPC). Hence there is no ethical issue present.

What *potential value* can be obtained by studying this data?

List some *specific questions*, and *plan to answer them* in your analysis

The loss caused by the tornado and the states it affected the most could be analyzed using this dataset. Also, the history of tornado can help in predicting the future occurrence of it. The questions which I plan to answer are -

- What is the correlation between injuries, fatalities, loss, crop loss, length of the tornado, the width of the tornado?
- Inspecting the loss caused before and after 1996. Before 1996 the losses were categorized as following - 0 = Unknown ,1 < $50 ,2 = $50-500 ,3 = $500-5,000 ,4 = $5,000-50,000 ,5 = $50,000-$500,000 ,6 = $500,000-$5,000,000 ,7 = $5,000,000-$50,000,000 ,8 = $50,000,000-500,000,000 ,9 = $5000,000,000
- Finding the number of storms based on severity (magnitude) and fc over years.
- Scatterplots for the length for length and width of the storm
- How many numbers of tornadoes came over the years?
- Which is the State with maximum loss?
- How many numbers of fatalities are there by the state?
- How many numbers of Tornado injuries caused by the magnitude of the storm?
- How the latitude and longitude of the tornado changing?
- Check if Tornado's loss is related to injuries, fatalities, and area.
- Which factors are contributing to the loss?

Resources: What software and hardware resources will you need to study this data?

The software I have used are R, Python, and PostgreSQL to perform the analysis.

Background & prior studies

Identify and briefly discuss one or more other similar studies that were done in the domain of your project

Earth's climate system is unpredictable and nonlinear. The evolution of earth's climate change over the years has led to scientists predict climate changes, cyclone prediction, weather forecasting, etc. However, it imposes some limits as the nature of the earth's climate system is highly unpredictable. The climate changes are attributed to volcanic eruptions and El Nino–Southern Oscillation. The long-term prediction of weather gets hampered as scientists are unable to understand the complete phenomena of these factors.

El Nino, during the past 40 years, has affected the South American coast. The 1982-83 El Nino was by far the strongest and was not predicted. The scientists did not anticipate this El Nino and climates uneven behavior led to the loss of human life and economy imbalance.[3]

The volcanic eruptions can be predicted if volcanologists have a thorough understanding of volcanic eruptions history if they can install proper equipment well in advance of eruption, and they continuously monitor and interpret the data coming from the instruments. The timely prediction of 1980 Mt. St. Helens volcanic eruption saved 20,000 lives.[4]

The National Weather Service (an agency within NOAA) collects and interprets rainfall data throughout the United States, and issues flood watches and warnings as appropriate. Based on rainfall prediction, river flow direction, and storm prediction, the flood situation can be predicted, and necessary evacuation steps can be taken. For example, Japan's heavy rain forecast was, as a result of the typhoon, led to a flood-like situation. The early predictions helped in saving more lives.[5]

Deliverable 2 – Data Analysis

EXPLORATORY DATA ANALYSIS USING PYTHON

The loss values from 1950-1996 were given as numbers between 0-9. The loss from 1996 to 2015 was given as loss*100,000 million dollars. The loss from 2016 to 2018 was given in million dollars. To make the unit same, the loss values have converted according to the metadata given.

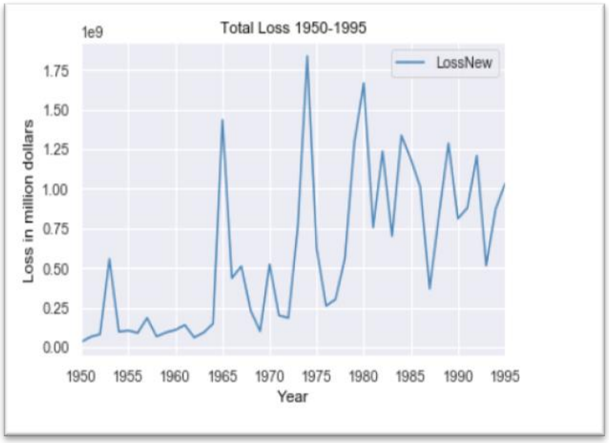
[3] El Nino and Climate Prediction. [Website]. (n.d.). Retrieved October 19, 2019 from https://atmos.washington.edu/gcg/RTN/rtnt.html

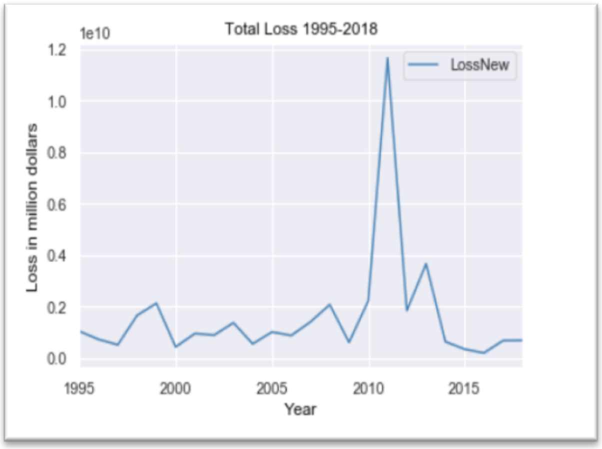
[4] Tyson,P. (2019). *Can We Predict Eruptions?* [Website]. Retrieved October 19, 2019 from https://www.pbs.org/wgbh/nova/vesuvius/predict.html

[5] Can floods be predicted? [Website]. (2017). Retrieved October 19, 2019 from https://www.americangeosciences.org/critical-issues/faq/can-floods-be-predicted

Inspecting the loss caused before and after 1996.

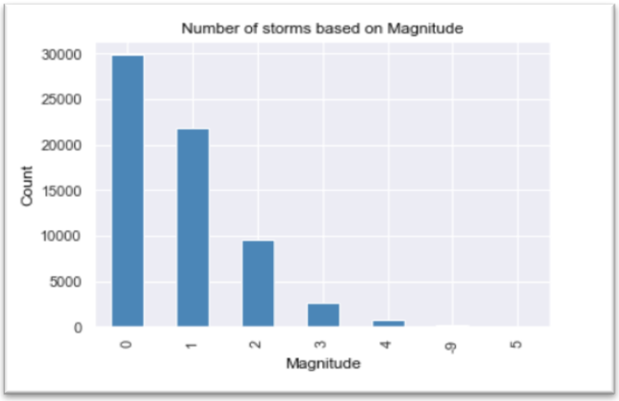
The highest loss was reported in the year 1975 before the year 1996 whereas the highest loss was reported in the year 2011 between 1996-2018.

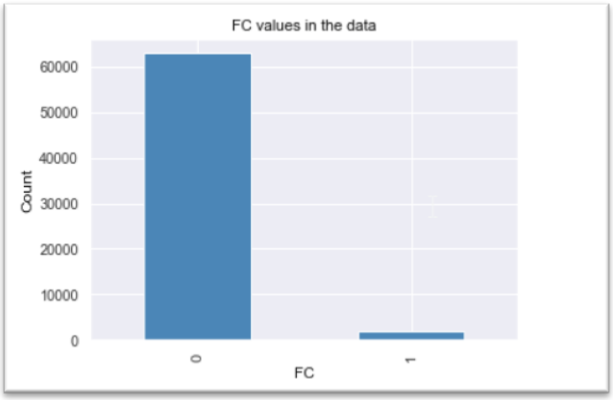




Finding the number of storms based on severity (magnitude) and FC over the years.

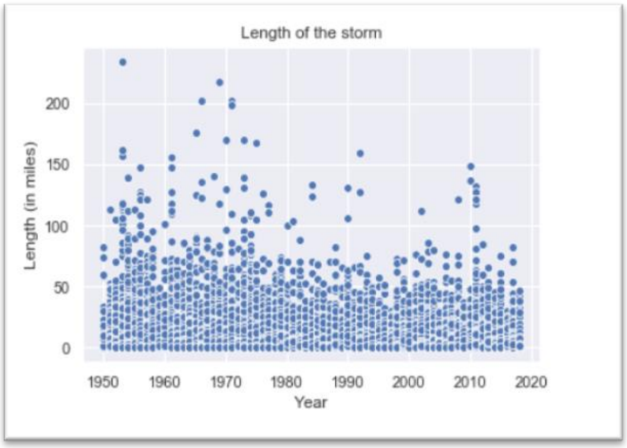
The highest count for the storm was for magnitude 0 (29884). With respect to FC values, the scale rating FC=0 had the highest count (62961).

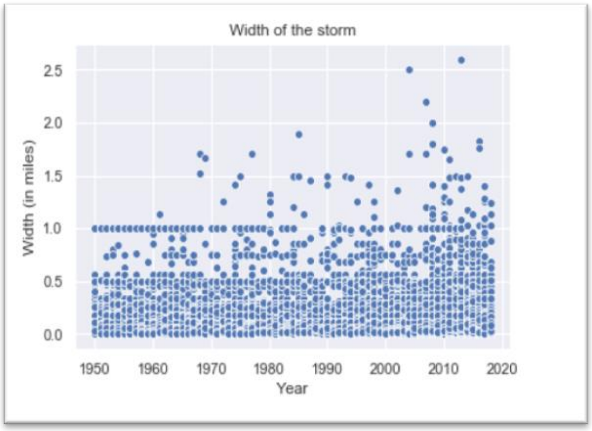




Scatterplots for the length for length and width of the storm

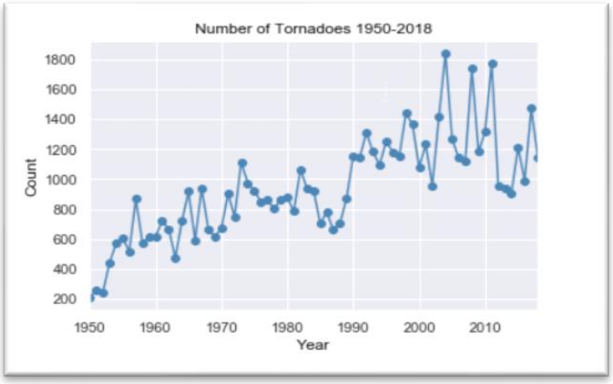
From the scatterplot for length of the storm, it is observed that the length of the storm mostly lay in the range of 0-150 miles. There some outliers in the graph. The width of the storm was given in yards. Hence it was converted from yards to miles before plotting. It is observed that the width of the storm mostly lay in the range of 0-1 mile with few outliers like 2.5 miles.





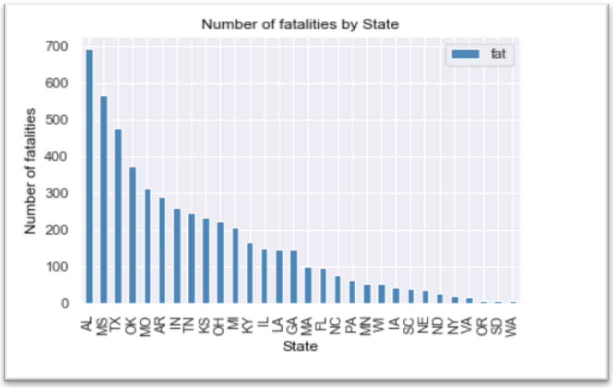
Number of tornadoes came over the years from 1950-2018

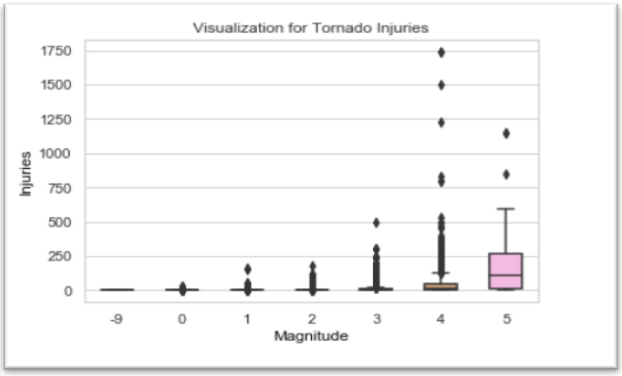
The highest number of storms came in the year 2004 with a count of 1842.



Number of fatalities per state and injuries based on magnitude from 1950-2018

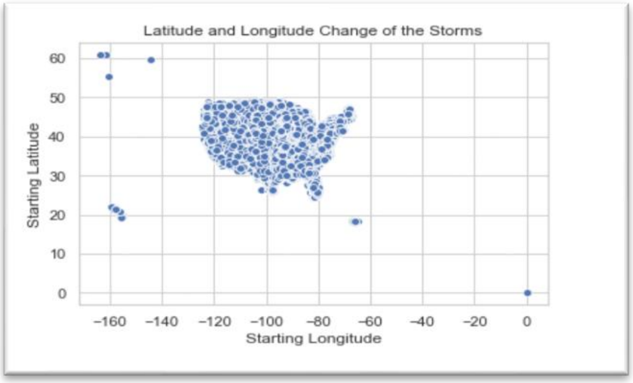
The bar graph shows highest number of fatalities was from state AL (Alabama) with a count of 692. The boxplot for storm injuries based on magnitude shows that there is a more significant variability for storm mag=5 with few outliers. The storm with mag=4 has more substantial outliers, which imply that the storm has caused more number of injuries than the usual range for mag=4 storm.





Change in latitude and longitude of the tornado

The latitude and longitude scatterplot show the area covered by storms in the US. It depicts that storms have covered almost all the states of the US.

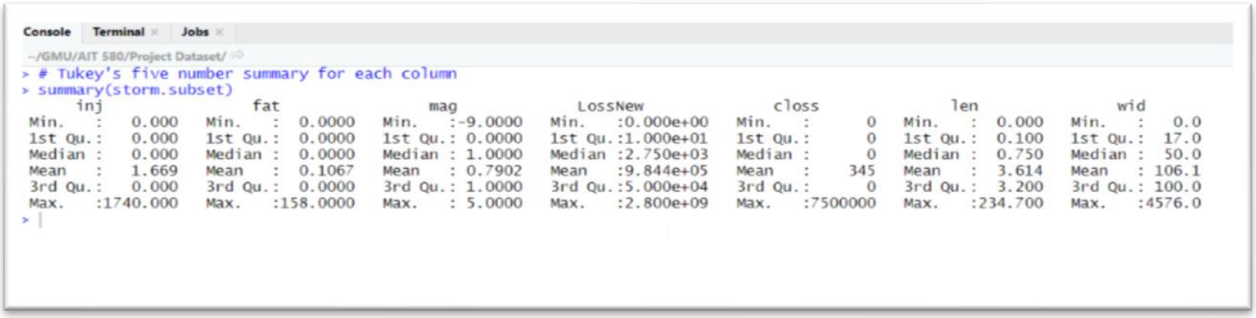


CORRELATION, REGRESSION ANALYSIS, HYPOTHESIS TESTING USING R

There are 29 columns in the data, so I have created a subset of the data on which I performed Correlation analysis. The columns included are columns included injuries, fatalities, magnitude, lossnew, crop loss, length and width of the storm.

Summary Statistics

Tukey's five-number summary for each column of the subset created



```
> # Tukey's five number summary for each column
> summary(storm.subset)
      inj                fat                mag              LossNew               closs               len              wid        
 Min.   :   0.000   Min.   :  0.0000   Min.   :-9.0000   Min.   :0.000e+00   Min.   :      0   Min.   :  0.000   Min.   :   0.0  
 1st Qu.:   0.000   1st Qu.:  0.0000   1st Qu.: 0.0000   1st Qu.:1.000e+01   1st Qu.:      0   1st Qu.:  0.100   1st Qu.:  17.0  
 Median :   0.000   Median :  0.0000   Median : 1.0000   Median :2.750e+03   Median :      0   Median :  0.750   Median :  50.0  
 Mean   :   1.669   Mean   :  0.1067   Mean   : 0.7902   Mean   :9.844e+05   Mean   :    345   Mean   :  3.614   Mean   : 106.1  
 3rd Qu.:   0.000   3rd Qu.:  0.0000   3rd Qu.: 1.0000   3rd Qu.:5.000e+04   3rd Qu.:      0   3rd Qu.:  3.200   3rd Qu.: 100.0  
 Max.   :1740.000   Max.   :158.0000   Max.   : 5.0000   Max.   :2.800e+09   Max.   :7500000   Max.   :234.700   Max.   :4576.0  
> |
```

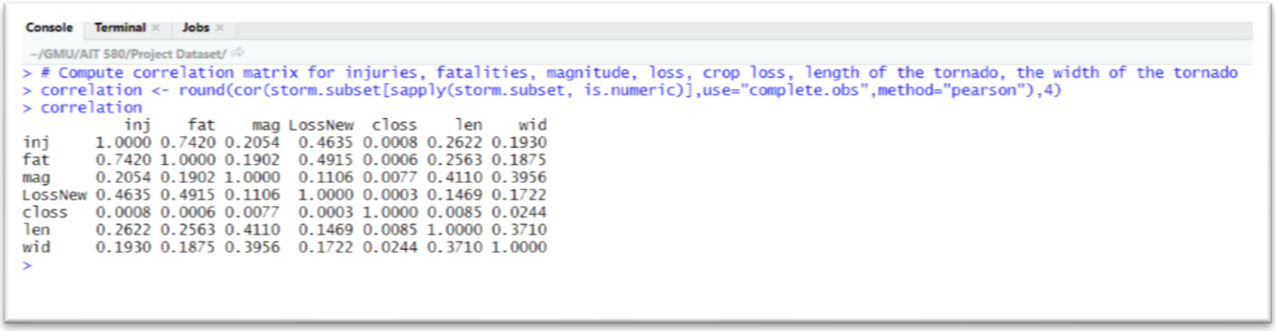
Analysis of summary statistics

- The injuries vary from 0 to 1740.
- The fatalities range from 0 to 158.
- The crop loss ranges from 0 to 7500000.
- The length of the storm ranges from 0 to 234.7 miles.
- The width of the storm ranges from 0 to 4576 yards.

Compute correlation matrix for injuries, fatalities, magnitude, loss, crop loss, length of the tornado, the width of the tornado

From the correlation matrix, we can interpret that there is a strong correlation between
Inj and fat – 0.74
LossNew and inj – 0.46
Wid and mag – 0.39
Len and mag – 0.41

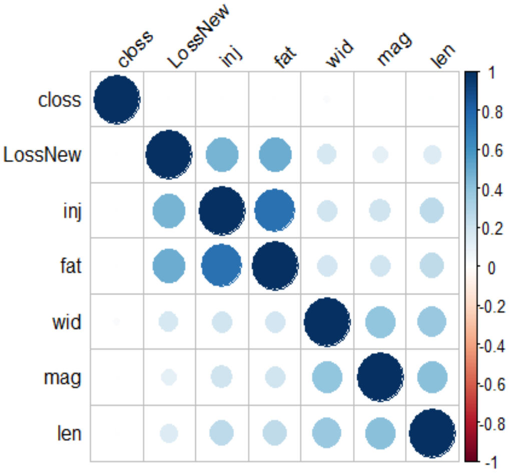
Len and wid – 0.37



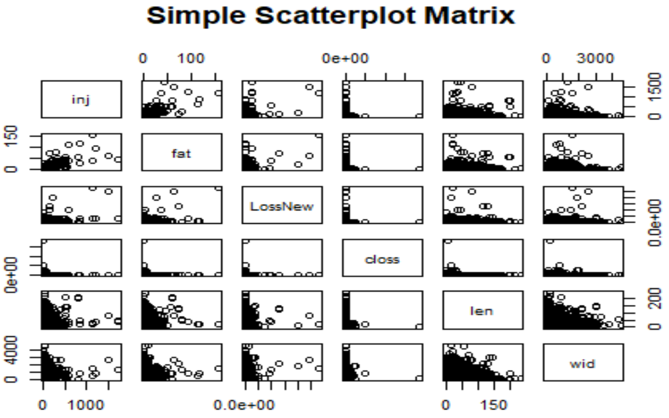
```
> # Compute correlation matrix for injuries, fatalities, magnitude, loss, crop loss, length of the tornado, the width of the tornado
> correlation <- round(cor(storm.subset[sapply(storm.subset, is.numeric)],use="complete.obs",method="pearson"),4)
> correlation
           inj    fat    mag LossNew  closs    len    wid
inj     1.0000 0.7420 0.2054  0.4635 0.0008 0.2622 0.1930
fat     0.7420 1.0000 0.1902  0.4915 0.0006 0.2563 0.1875
mag     0.2054 0.1902 1.0000  0.1106 0.0077 0.4110 0.3956
LossNew 0.4635 0.4915 0.1106  1.0000 0.0003 0.1469 0.1722
closs   0.0008 0.0006 0.0077  0.0003 1.0000 0.0085 0.0244
len     0.2622 0.2563 0.4110  0.1469 0.0085 1.0000 0.3710
wid     0.1930 0.1875 0.3956  0.1722 0.0244 0.3710 1.0000
>
```

Draw correlogram for injuries, fatalities, magnitude, loss, crop loss, length of the tornado, the width of the tornado

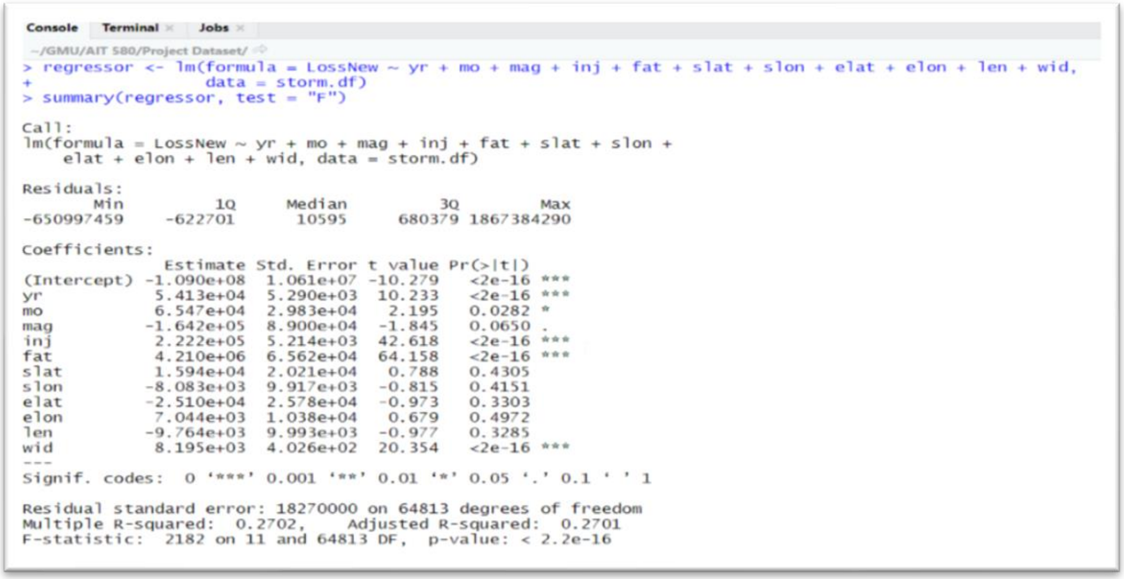
The correlation is shown below using correlogram and pairwise distribution plot.



Pairwise distribution plot for each interesting pair of columns



Linear Regression for the losses:

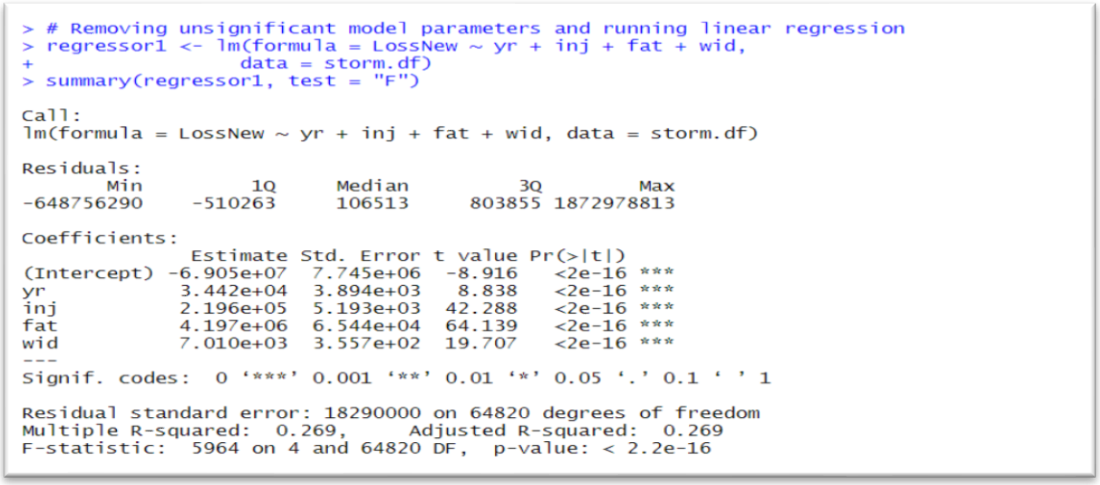


```
> regressor <- lm(formula = LossNew ~ yr + mo + mag + inj + fat + slat + slon + elat + elon + len + wid,
+                 data = storm.df)
> summary(regressor, test = "F")

Call:
lm(formula = LossNew ~ yr + mo + mag + inj + fat + slat + slon + 
    elat + elon + len + wid, data = storm.df)

Residuals:
       Min         1Q     Median         3Q        Max 
-650997459    -622701      10595     680379 1867384290 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.090e+08  1.061e+07 -10.279   <2e-16 ***
yr           5.413e+04  5.290e+03  10.233   <2e-16 ***
mo           6.547e+04  2.983e+04   2.195   0.0282 *  
mag         -1.642e+05  8.900e+04  -1.845   0.0650 .  
inj          2.222e+05  5.214e+03  42.618   <2e-16 ***
fat          4.210e+06  6.562e+04  64.158   <2e-16 ***
slat         1.594e+04  2.021e+04   0.788   0.4305    
slon        -8.083e+03  9.917e+03  -0.815   0.4151    
elat        -2.510e+04  2.578e+04  -0.973   0.3303    
elon         7.044e+03  1.038e+04   0.679   0.4972    
len         -9.764e+03  9.993e+03  -0.977   0.3285    
wid          8.195e+03  4.026e+02  20.354   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18270000 on 64813 degrees of freedom
Multiple R-squared:  0.2702,	Adjusted R-squared:  0.2701 
F-statistic:  2182 on 11 and 64813 DF,  p-value: < 2.2e-16
```

Summary of Linear Regression 1:

- A high value of F statistic, with a very low p-value (<2.2e-16), implies that the null hypothesis can be rejected. This means there is a potential relationship between the predictors and the outcome.
- A significant value of Residual standard error 18270000 means there is a high deviation of the model from the regression line.
- The value of adjusted R-squared (0.2701) shows that more than 27% of the variance in the data is explained by the model.
- In this case, the value of adjusted R-squared is low which shows the model will not explain the variability in the outcome. Hence removing the insignificant parameters in next step.
-

Removing insignificant model parameters and running a linear regression

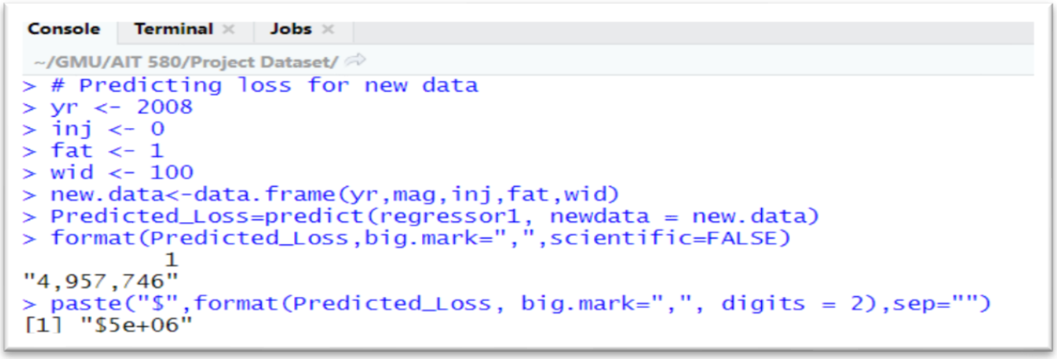


```
> # Removing unsignificant model parameters and running linear regression
> regressor1 <- lm(formula = LossNew ~ yr + inj + fat + wid,
+                  data = storm.df)
> summary(regressor1, test = "F")

Call:
lm(formula = LossNew ~ yr + inj + fat + wid, data = storm.df)

Residuals:
       Min         1Q     Median         3Q        Max 
-648756290    -510263     106513     803855 1872978813 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -6.905e+07  7.745e+06  -8.916   <2e-16 ***
yr           3.442e+04  3.894e+03   8.838   <2e-16 ***
inj          2.196e+05  5.193e+03  42.288   <2e-16 ***
fat          4.197e+06  6.544e+04  64.139   <2e-16 ***
wid          7.010e+03  3.557e+02  19.707   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18290000 on 64820 degrees of freedom
Multiple R-squared:  0.269,	Adjusted R-squared:  0.269 
F-statistic:  5964 on 4 and 64820 DF,  p-value: < 2.2e-16
```

Summary of Linear Regression 2:

- A high value of F statistic, with a very low p-value (<2.2e-16), implies that the null hypothesis can be rejected. This means there is a potential relationship between the predictors and the outcome.
- The tremendous value of Residual standard error 18290000 means there is a high deviation of the model from the regression line.
- The value of adjusted R-squared (0.269) shows that more than 27% of the variance in the data is explained by the model.
- In this case, the value of adjusted R-squared remains the same as before which shows the model will not explain the variability in the outcome.
- The factors that are contributing to loss are year, injuries, fatalities and width of the storm.

Predicting loss for new data using the model developed



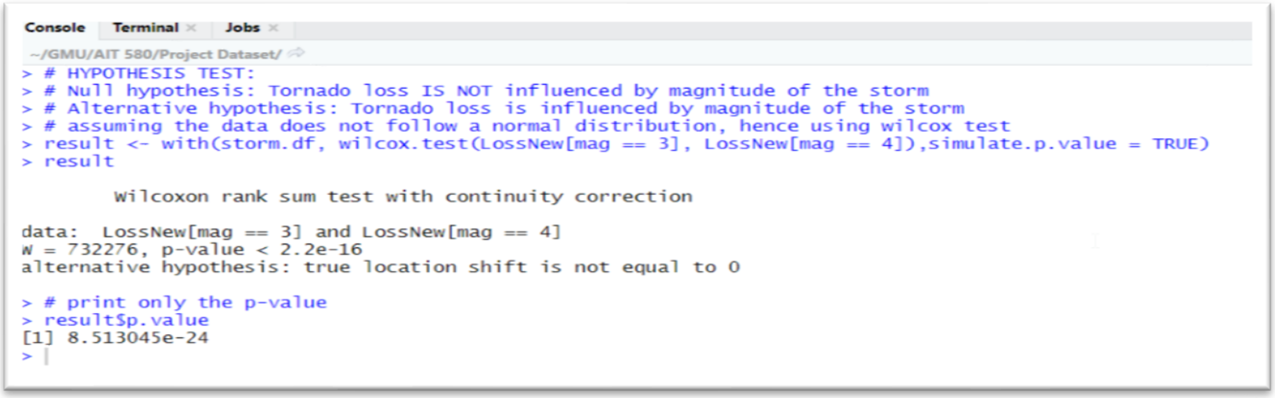
```
> # Predicting loss for new data
> yr <- 2008
> inj <- 0
> fat <- 1
> wid <- 100
> new.data<-data.frame(yr,mag,inj,fat,wid)
> Predicted_Loss=predict(regressor1, newdata = new.data)
> format(Predicted_Loss,big.mark=",",scientific=FALSE)
          1
"4,957,746"
> paste("$",format(Predicted_Loss, big.mark=",", digits = 2),sep="")
[1] "$5e+06"
```

Here we have used the model for prediction. The loss in the year 2008 for storm of width 100 yards with 0 injuries and 1 fatality is $4,957,746.

Hypothesis Test:

We are using Wilcox test, assuming the data is not normally distributed. We are using magnitude 3 and 4 as the Wilcox test accepts 2 levels of the data.

- Null hypothesis: Tornado loss IS NOT influenced by magnitude (3,4) of the storm
- Alternative hypothesis: Tornado loss is influenced by magnitude (3,4) of the storm

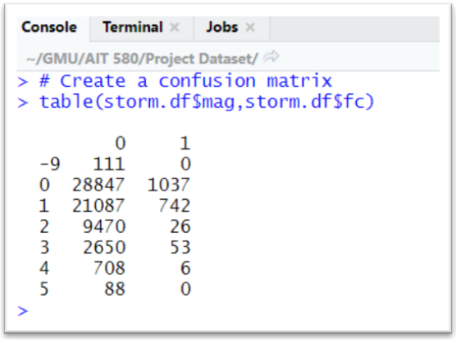


```
> # HYPOTHESIS TEST:
> # Null hypothesis: Tornado loss IS NOT influenced by magnitude of the storm
> # Alternative hypothesis: Tornado loss is influenced by magnitude of the storm
> # assuming the data does not follow a normal distribution, hence using wilcox test
> result <- with(storm.df, wilcox.test(LossNew[mag == 3], LossNew[mag == 4]),simulate.p.value = TRUE)
> result

	Wilcoxon rank sum test with continuity correction

data:  LossNew[mag == 3] and LossNew[mag == 4]
W = 732276, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0

> # print only the p-value
> result$p.value
[1] 8.513045e-24
> |
```

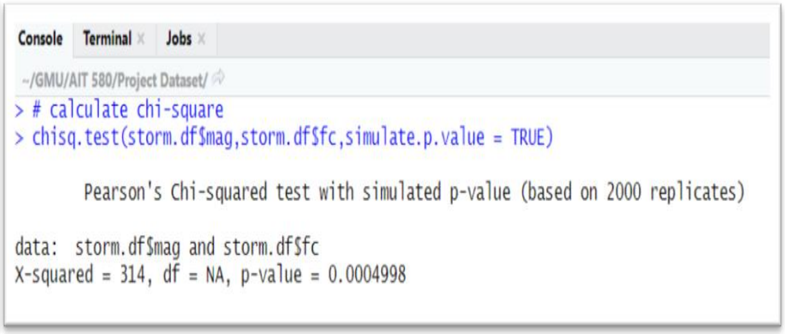
From the results obtained above, we can see the p-value of the test is less than 0, which is less than the significance level alpha = 0.05. We can reject the null hypothesis and conclude that tornado loss is influenced by the magnitude of the storm.

Chi-Square test to check if there is a significance association between FC and MAG.

The contingency table shows the relationship between Magnitude and FC. A chi square test is used to check the association between the two variables- mag and fc.



```
> # Create a confusion matrix
> table(storm.df$mag,storm.df$fc)

          0      1
  -9    111      0
  0   28847   1037
  1   21087    742
  2    9470     26
  3    2650     53
  4     708      6
  5      88      0
>
```

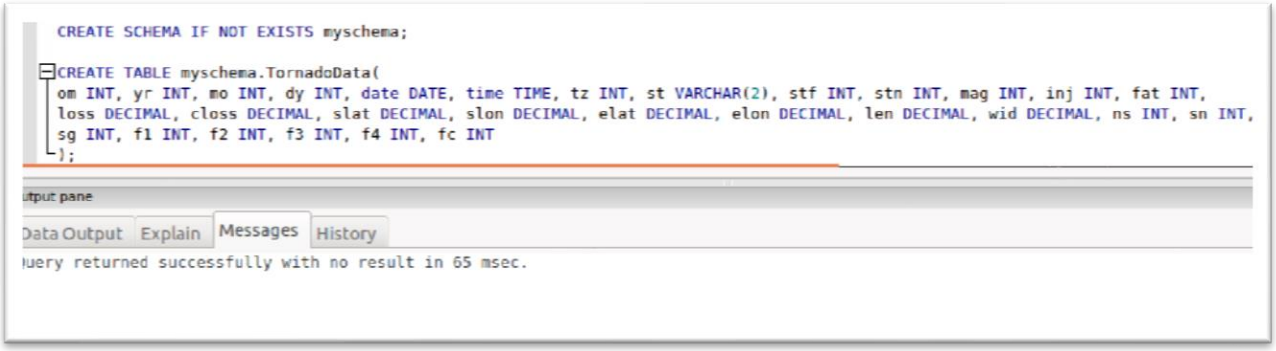


```
> # calculate chi-square
> chisq.test(storm.df$mag,storm.df$fc,simulate.p.value = TRUE)

	Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data:  storm.df$mag and storm.df$fc
X-squared = 314, df = NA, p-value = 0.0004998
```

Since the p-value is 0.0004998, which is less than 0.05, we can conclude that there is significant association between magnitude of the storm and FC.

DATA ANALYSIS USING POSTGRESQL

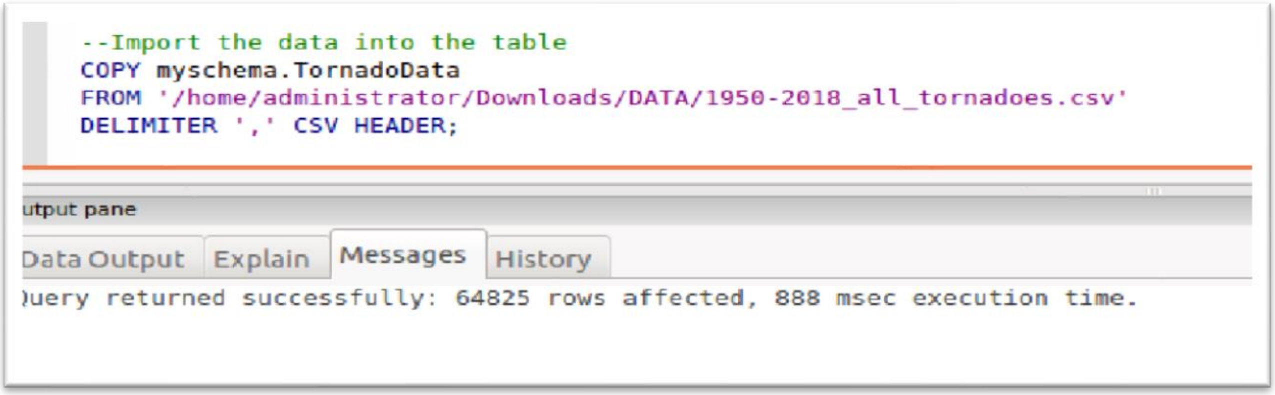
Create schema and table



```
CREATE SCHEMA IF NOT EXISTS myschema;

CREATE TABLE myschema.TornadoData(
om INT, yr INT, mo INT, dy INT, date DATE, time TIME, tz INT, st VARCHAR(2), stf INT, stn INT, mag INT, inj INT, fat INT,
loss DECIMAL, closs DECIMAL, slat DECIMAL, slon DECIMAL, elat DECIMAL, elon DECIMAL, len DECIMAL, wid DECIMAL, ns INT, sn INT,
sg INT, f1 INT, f2 INT, f3 INT, f4 INT, fc INT
);
```

```
uery returned successfully with no result in 65 msec.
```

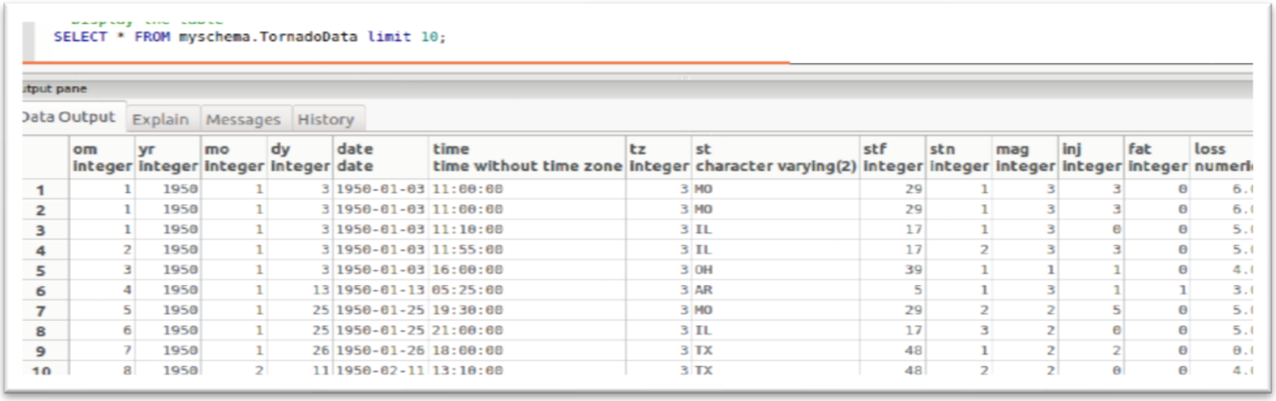
Import the data into the table



```
--Import the data into the table
COPY myschema.TornadoData
FROM '/home/administrator/Downloads/DATA/1950-2018_all_tornadoes.csv'
DELIMITER ',' CSV HEADER;
```

```
Query returned successfully: 64825 rows affected, 888 msec execution time.
```

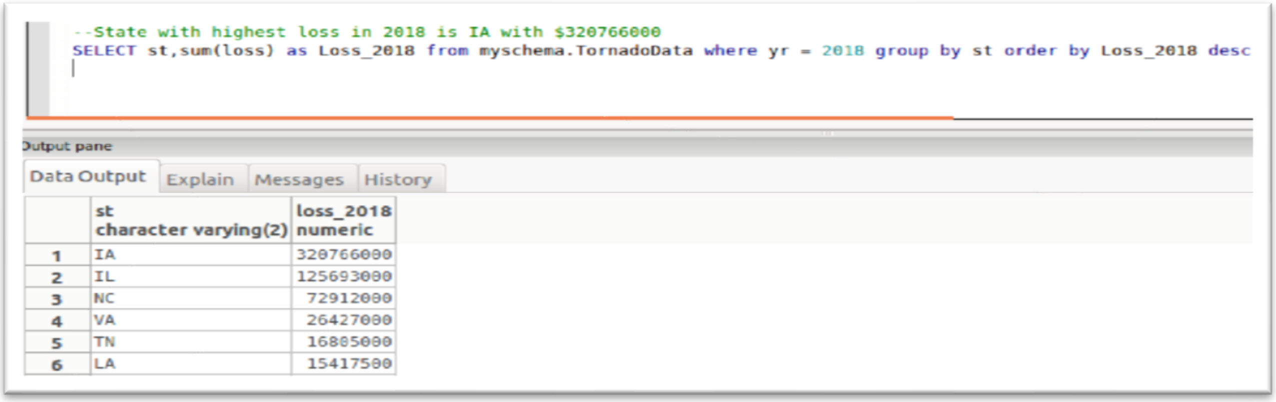
Display the data imported



```
--Display the table
SELECT * FROM myschema.TornadoData limit 10;
```

	om integer	yr integer	mo integer	dy integer	date date	time time without time zone	tz integer	st character varying(2)	stf integer	stn integer	mag integer	inj integer	fat integer	loss numeri
1	1	1950	1	3	1950-01-03	11:00:00	3	MO	29	1	3	3	0	6.
2	1	1950	1	3	1950-01-03	11:00:00	3	MO	29	1	3	3	0	6.
3	1	1950	1	3	1950-01-03	11:10:00	3	IL	17	1	3	0	0	5.
4	2	1950	1	3	1950-01-03	11:55:00	3	IL	17	2	3	3	0	5.
5	3	1950	1	3	1950-01-03	16:00:00	3	OH	39	1	1	1	0	4.
6	4	1950	1	13	1950-01-13	05:25:00	3	AR	5	1	3	1	1	3.
7	5	1950	1	25	1950-01-25	19:30:00	3	MO	29	2	2	5	0	5.
8	6	1950	1	25	1950-01-25	21:00:00	3	IL	17	3	2	0	0	5.
9	7	1950	1	26	1950-01-26	18:00:00	3	TX	48	1	2	2	0	0.
10	8	1950	2	11	1950-02-11	13:10:00	3	TX	48	2	2	0	0	4.

State with Highest Loss in 2018

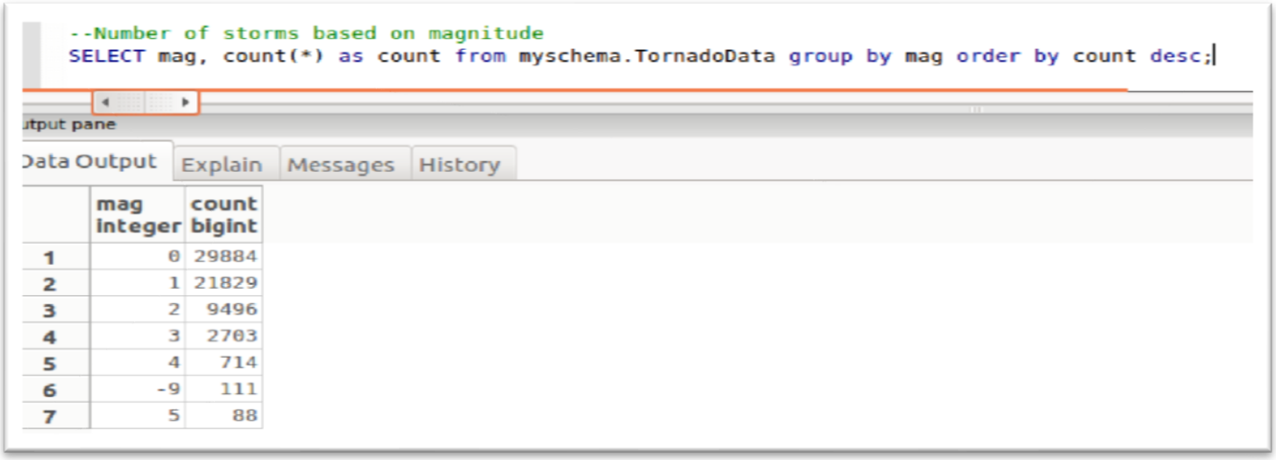


```
--State with highest loss in 2018 is IA with $320766000
SELECT st,sum(loss) as Loss_2018 from myschema.TornadoData where yr = 2018 group by st order by Loss_2018 desc
```

	st character varying(2)	loss_2018 numeric
1	IA	320766000
2	IL	125693000
3	NC	72912000
4	VA	26427000
5	TN	16885000
6	LA	15417500

The state with the highest loss in 2018 was IA (Iowa), with loss of $320766000.

Number of storms based on the magnitude

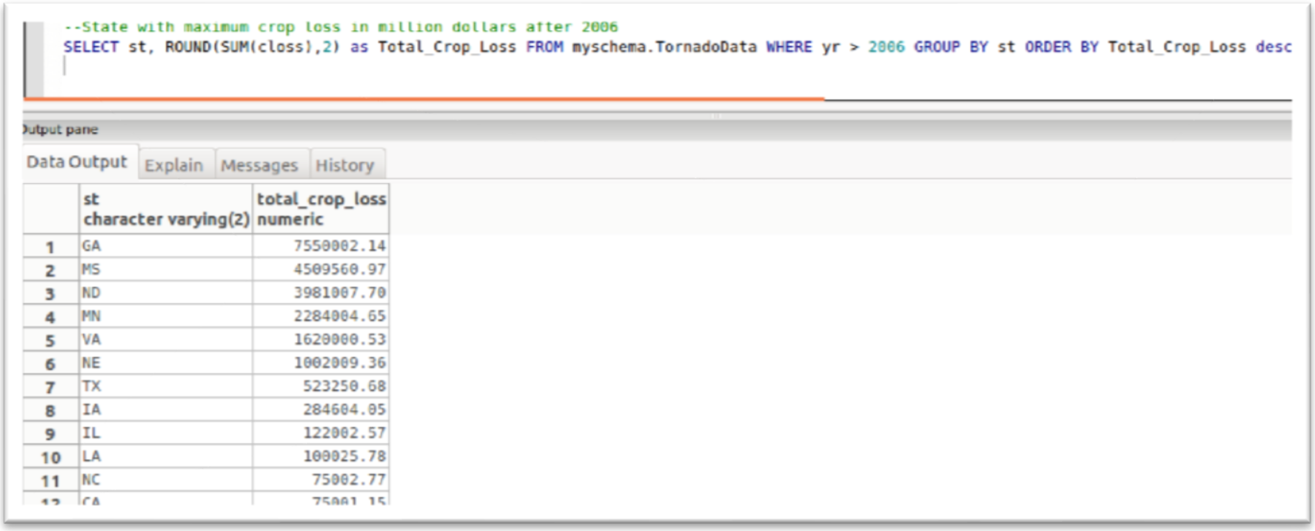


```
--Number of storms based on magnitude
SELECT mag, count(*) as count from myschema.TornadoData group by mag order by count desc;
```

	mag integer	count bigint
1	0	29884
2	1	21829
3	2	9496
4	3	2703
5	4	714
6	-9	111
7	5	88

The storm with magnitude = 0 has the highest number of occurrences, 29884.

State with maximum crop loss

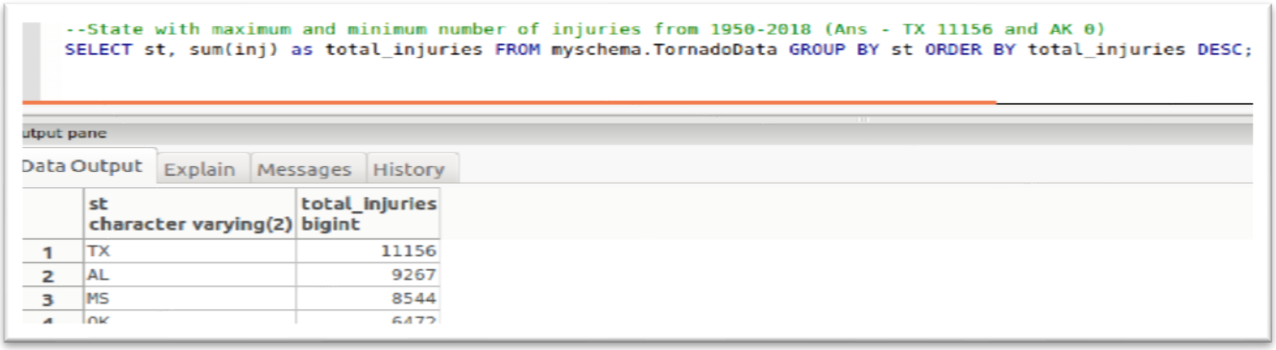


```
--State with maximum crop loss in million dollars after 2006
SELECT st, ROUND(SUM(closs),2) as Total_Crop_Loss FROM myschema.TornadoData WHERE yr > 2006 GROUP BY st ORDER BY Total_Crop_Loss desc
```

	st character varying(2)	total_crop_loss numeric
1	GA	7550002.14
2	MS	4509560.97
3	ND	3981007.70
4	MN	2284004.65
5	VA	1620000.53
6	NE	1002009.36
7	TX	523250.68
8	IA	284604.05
9	IL	122002.57
10	LA	100025.78
11	NC	75002.77
12	CA	75001.15

The state with maximum crop loss is GA (Georgia) with a total loss of $7550002.14

State with maximum injuries are TX with 11156

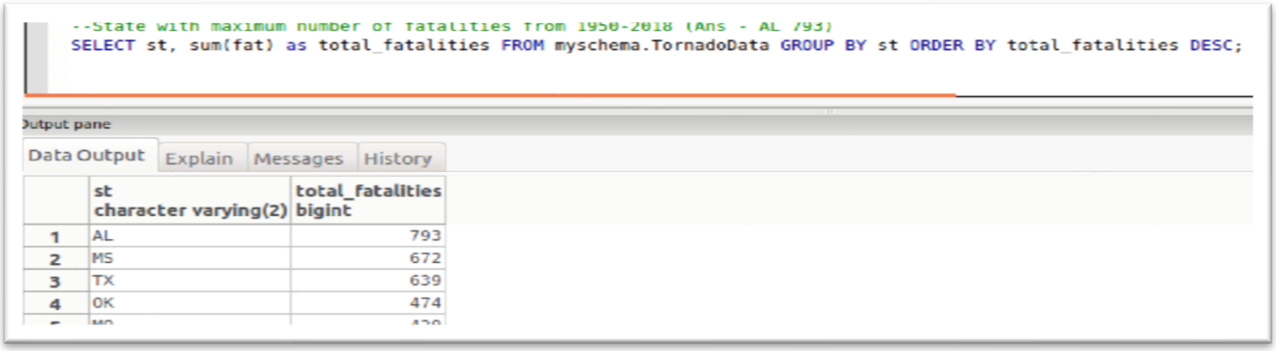


```
--State with maximum and minimum number of injuries from 1950-2018 (Ans - TX 11156 and AK 0)
SELECT st, sum(inj) as total_injuries FROM myschema.TornadoData GROUP BY st ORDER BY total_injuries DESC;
```

	st character varying(2)	total_injuries bigint
1	TX	11156
2	AL	9267
3	MS	8544
4	OK	6472

The state with maximum injuries in Texas, TX = 11156

State with maximum fatalities is AL with 793

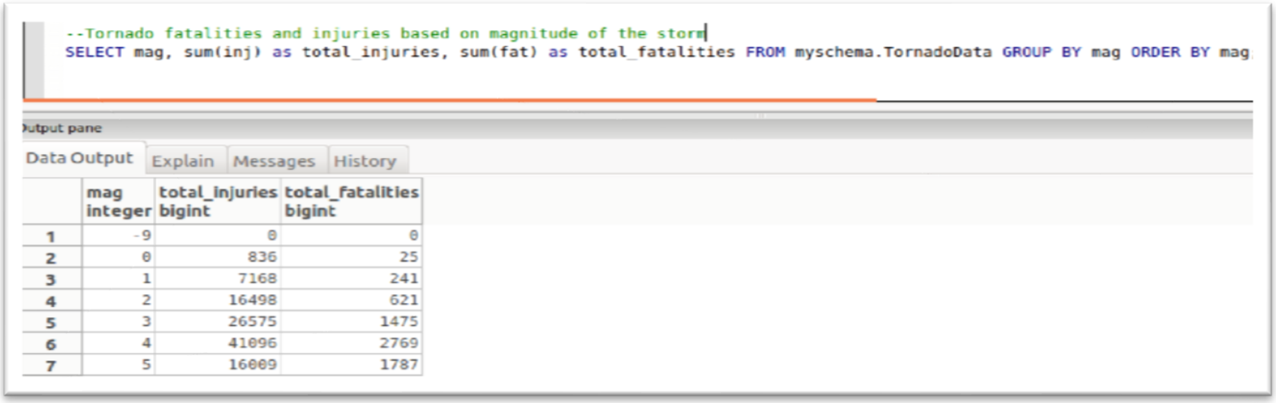


```
--State with maximum number of fatalities from 1950-2018 (Ans - AL 793)
SELECT st, sum(fat) as total_fatalities FROM myschema.TornadoData GROUP BY st ORDER BY total_fatalities DESC;
```

	st character varying(2)	total_fatalities bigint
1	AL	793
2	MS	672
3	TX	639
4	OK	474

The state with maximum fatalities is AL (Alabama) - 793.

Tornado fatalities and injuries based on the magnitude of the storm



```
--Tornado fatalities and injuries based on magnitude of the storm
SELECT mag, sum(inj) as total_injuries, sum(fat) as total_fatalities FROM myschema.TornadoData GROUP BY mag ORDER BY mag
```

	mag integer	total_injuries bigint	total_fatalities bigint
1	-9	0	0
2	0	836	25
3	1	7168	241
4	2	16498	621
5	3	26575	1475
6	4	41096	2769
7	5	16009	1787

The highest number of fatalities, which is 2760 is caused by the storm of magnitude 4. The highest number of injuries which is 41000 caused by the storm of magnitude 4.

Describe the value obtained from the study

The analysis and values obtained from the graphs, SQL queries, Hypothesis test, Regression analysis, and correlation matrix has been discussed with the graphs and analysis plotted. The analysis was done using R, Python, and Postgresql. It helped in knowing the total loss, total crop loss, total injuries per state, fatalities per state, number of storms per year, number of storms with respect to magnitude and factors contributing to the loss.

Include explanations of any technical terms relevant to the project domain.

1. Hypothesis test: It determines the probability that the given hypothesis is correct.
2. Linear Regression: The linear regression is a linear approach to model the response (dependent) and predictor (independent) variables. The case of multiple explanatory variables is called multiple linear regression.

3. Chi-Square Test: The chi-square test of independence is used to analyze the contingency tables formed by two categorical variables. The chi-square test evaluates whether there is significant association between the categories of the two variables.[6]
4. Contingency Table: A contingency table shows the overall summary of the original data in a tabular format.
5. Wilcox Test: Wilcox test is a non-parametric alternative to compare two independent groups of samples.
6. Residual Standard Error: The residual standard deviation is used to describe the difference in standard deviations of observed values versus predicted values, as shown by points in a regression analysis[7].
7. Adjusted R-squared: The adjusted R-squared compares the explanatory power of regression models that contain different numbers of predictors.[8]
8. Multiple R-squared: It is the absolute value of the correlation coefficient.
9. F statistic: F statistic is the value when we run a regression to find out if the means of the two variables are significantly different
10. p-value: A p-value helps in determining the significance of the test. Its value lies between 0 and 1. A small p-value (<0.05) indicates strong evidence against null hypothesis.

Discuss any limitations of your analysis, and recommend future needed analysis

The first variable of the dataset "om" was supposed to capture the count of tornadoes during that particular year. However, before 2007, this column acted like a sequence to the incoming data.

The number of variables is 29, which makes it difficult to analyses each column.

The unit of distance should be kept the same. For example, the length is captured in miles and width in yards.

The loss values are not captured in the same format. From 1950-1995, it is captured between 0-9, 1996-2015 as loss*100,00 and 2016-2018 in million dollars. The unit of the loss should be uniform.

[6] STHDA. [Website]. (n.d.). Retrieved October 24, 2019 from http://www.sthda.com/english/wiki/chi-square-test-of-independence-in-r

[7] Investopedia. [Website]. (n.d.). Retrieved October 24, 2019 from https://www.investopedia.com/terms/r/residual-standard-deviation.asp

[8] The Minitab Blog. [Website]. (n.d.). Retrieved October 24, 2019 from https://blog.minitab.com/blog/adventures-in-statistics-2/multiple-regession-analysis-use-adjusted-r-squared-and-predicted-r-squared-to-include-the-correct-number-of-variables

References –

U.S. TORNADOES* (1950-2018). (2019). [Data File]. Retrieved from https://www.spc.noaa.gov/wcm/#data

NOAA's national Weather Service Storm Prediction Center [Website]. (n.d.). Retrieved October 20, 2019, from https://www.spc.noaa.gov/faq/#1.1

NOAA's national Weather Service Storm Prediction Center [File]. (n.d.). Retrieved October 20, 2019, from https://www.spc.noaa.gov/wcm/data/SPC_severe_database_description.pdf

El Nino and Climate Prediction. [Website]. (n.d.). Retrieved October 19, 2019, from https://atmos.washington.edu/gcg/RTN/rtnt.html

Tyson,P. (2019). Can We Predict Eruptions? [Website]. Retrieved October 19, 2019, from https://www.pbs.org/wgbh/nova/vesuvius/predict.html

Can floods be predicted? [Website]. (2017). Retrieved October 19, 2019, from https://www.americangeosciences.org/critical-issues/faq/can-floods-be-predicted

STHDA. [Website]. (n.d.). Retrieved October 24, 2019 from http://www.sthda.com/english/wiki/chi-square-test-of-independence-in-r

Investopedia. [Website]. (n.d.). Retrieved October 24, 2019 from https://www.investopedia.com/terms/r/residual-standard-deviation.asp

The Minitab Blog. [Website]. (n.d.). Retrieved October 24, 2019 from https://blog.minitab.com/blog/adventures-in-statistics-2/multiple-regession-analysis-use-adjusted-r-squared-and-predicted-r-squared-to-include-the-correct-number-of-variables