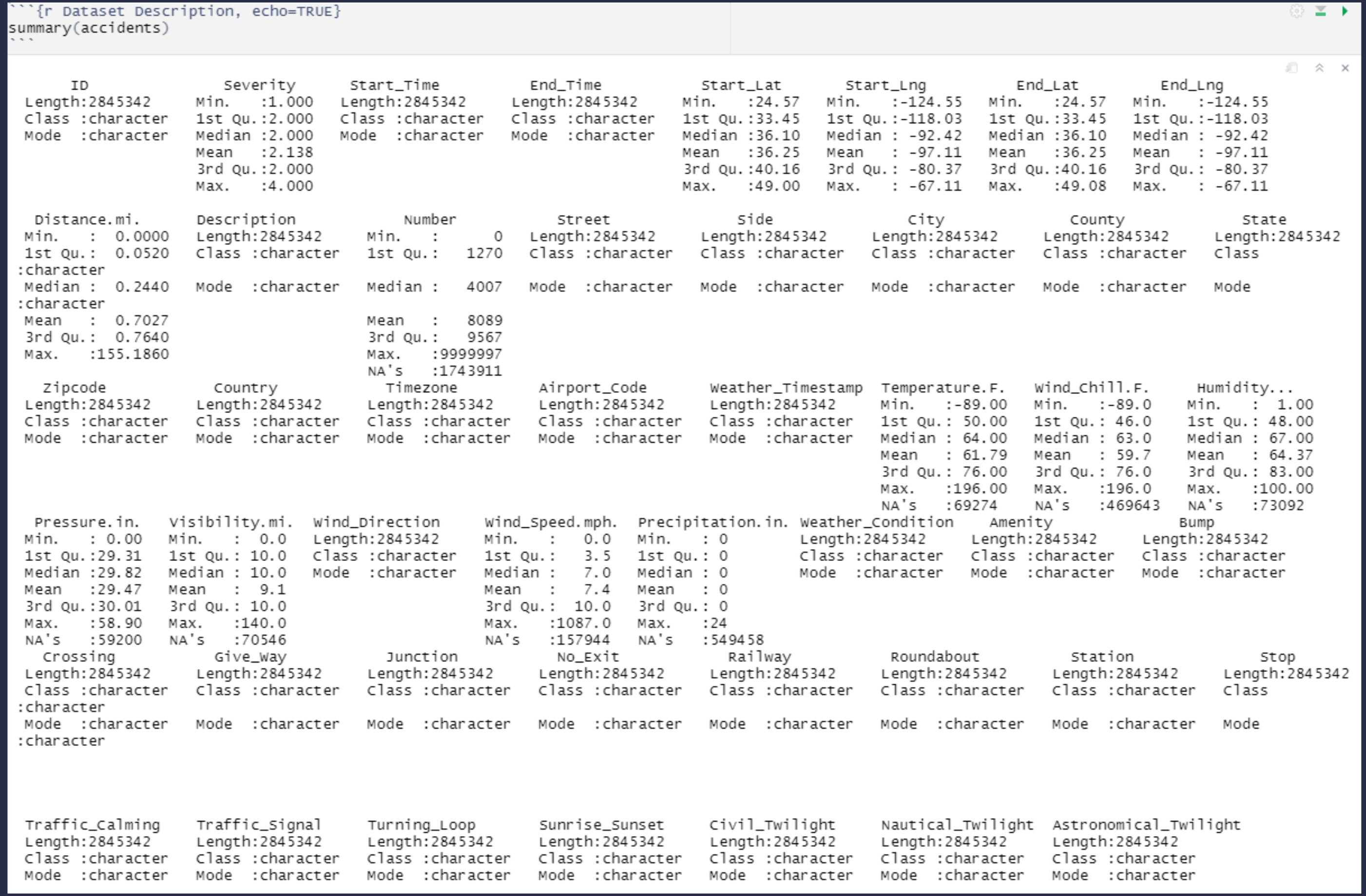
US Accidents

A COUNTRYWIDE TRAFFIC ACCIDENT DATASET

CHIH-YUN,LIU (DATA ANALYSIS)

Data Set - Summary



~~~
```{r Dataset Description, echo=TRUE}
summary(accidents)
```
~~~

```
      ID               Severity      Start_Time           End_Time           Start_Lat       Start_Lng          End_Lat         End_Lng       
 Length:2845342     Min.   :1.000   Length:2845342     Length:2845342     Min.   :24.57   Min.   :-124.55   Min.   :24.57   Min.   :-124.55  
 Class :character   1st Qu.:2.000   Class :character   Class :character   1st Qu.:33.45   1st Qu.:-118.03   1st Qu.:33.45   1st Qu.:-118.03  
 Mode  :character   Median :2.000   Mode  :character   Mode  :character   Median :36.10   Median : -92.42   Median :36.10   Median : -92.42  
                    Mean   :2.138                                         Mean   :36.25   Mean   : -97.11   Mean   :36.25   Mean   : -97.11  
                    3rd Qu.:2.000                                         3rd Qu.:40.16   3rd Qu.: -80.37   3rd Qu.:40.16   3rd Qu.: -80.37  
                    Max.   :4.000                                         Max.   :49.00   Max.   : -67.11   Max.   :49.08   Max.   : -67.11  

  Distance.mi.       Description            Number             Street              Side               City              County              State          
 Min.   :  0.0000   Length:2845342     Min.   :      0   Length:2845342     Length:2845342     Length:2845342     Length:2845342     Length:2845342    
 1st Qu.:  0.0520   Class :character   1st Qu.:   1270   Class :character   Class :character   Class :character   Class :character   Class 
:character  
 Median :  0.2440   Mode  :character   Median :   4007   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  
:character  
 Mean   :  0.7027                      Mean   :   8089                                                                                                  
 3rd Qu.:  0.7640                      3rd Qu.:   9567                                                                                                  
 Max.   :155.1860                      Max.   :9999997                                                                                                  
                                       NA's   :1743911                                                                                                  
   Zipcode            Country            Timezone          Airport_Code       Weather_Timestamp  Temperature.F.    Wind_Chill.F.      Humidity...    
 Length:2845342     Length:2845342     Length:2845342     Length:2845342     Length:2845342     Min.   :-89.00   Min.   :-89.0    Min.   :  1.00  
 Class :character   Class :character   Class :character   Class :character   Class :character   1st Qu.: 50.00   1st Qu.: 46.0    1st Qu.: 48.00  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median : 64.00   Median : 63.0    Median : 67.00  
                                                                                                Mean   : 61.79   Mean   : 59.7    Mean   : 64.37  
                                                                                                3rd Qu.: 76.00   3rd Qu.: 76.0    3rd Qu.: 83.00  
                                                                                                Max.   :196.00   Max.   :196.0    Max.   :100.00  
                                                                                                NA's   :69274    NA's   :469643   NA's   :73092   
  Pressure.in.   Visibility.mi.  Wind_Direction      Wind_Speed.mph.  Precipitation.in. Weather_Condition    Amenity                Bump          
 Min.   : 0.00   Min.   :  0.0   Length:2845342     Min.   :   0.0   Min.   : 0        Length:2845342     Length:2845342     Length:2845342    
 1st Qu.:29.31   1st Qu.: 10.0   Class :character   1st Qu.:   3.5   1st Qu.: 0        Class :character   Class :character   Class :character  
 Median :29.82   Median : 10.0   Mode  :character   Median :   7.0   Median : 0        Mode  :character   Mode  :character   Mode  :character  
 Mean   :29.47   Mean   :  9.1                      Mean   :   7.4   Mean   : 0                                                                   
 3rd Qu.:30.01   3rd Qu.: 10.0                      3rd Qu.:  10.0   3rd Qu.: 0                                                                   
 Max.   :58.90   Max.   :140.0                      Max.   :1087.0   Max.   :24                                                                   
 NA's   :59200   NA's   :70546                      NA's   :157944   NA's   :549458                                                               
   Crossing           Give_Way           Junction            No_Exit            Railway           Roundabout           Station              Stop          
 Length:2845342     Length:2845342     Length:2845342     Length:2845342     Length:2845342     Length:2845342     Length:2845342     Length:2845342    
 Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   Class 
:character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  
:character  




 Traffic_Calming    Traffic_Signal     Turning_Loop       Sunrise_Sunset     Civil_Twilight     Nautical_Twilight  Astronomical_Twilight
 Length:2845342     Length:2845342     Length:2845342     Length:2845342     Length:2845342     Length:2845342     Length:2845342    
 Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character  
```

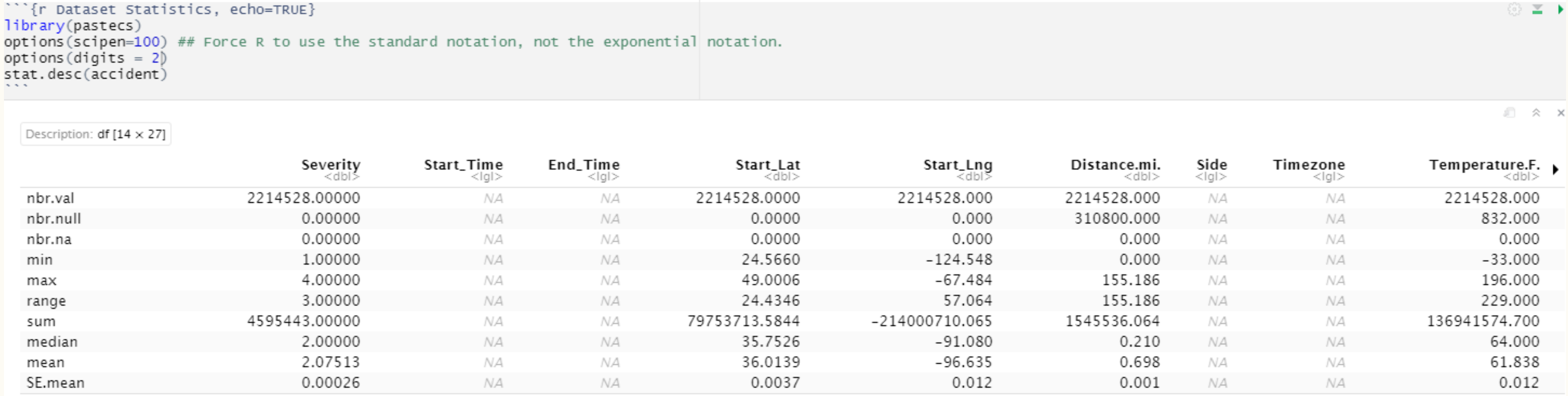
- This is a countrywide car accident dataset, which covers 49 states of the USA.
- Data collected from February 2016 to Dec 2021 (about 2.8 million accident records in this dataset), using multiple APIs that provide streaming traffic incident (or event) data.
- 2845342 instances & 47 columns

(Reference: https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents)

Descriptive Statistics

(Data Pre-processing: Remove unneeded variables)

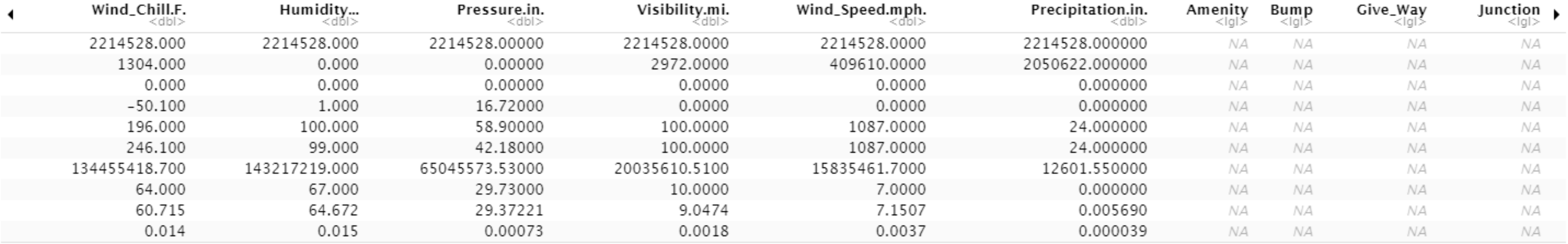
- Sample dimensions now : 2214528 x 27



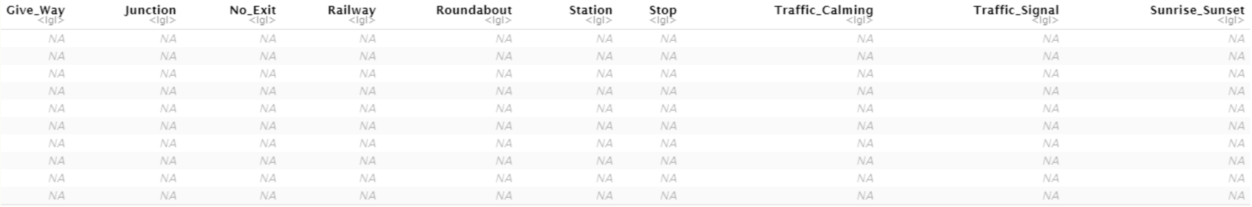
~~~
```{r Dataset Statistics, echo=TRUE}
library(pastecs)
options(scipen=100) ## Force R to use the standard notation, not the exponential notation.
options(digits = 2)
stat.desc(accident)
```
~~~

Description: df [14 × 27]

	Severity <dbl>	Start_Time <lgl>	End_Time <lgl>	Start_Lat <dbl>	Start_Lng <dbl>	Distance.mi. <dbl>	Side <lgl>	Timezone <lgl>	Temperature.F. <dbl>
nbr.val	2214528.00000	NA	NA	2214528.0000	2214528.000	2214528.000	NA	NA	2214528.000
nbr.null	0.00000	NA	NA	0.0000	0.000	310800.000	NA	NA	832.000
nbr.na	0.00000	NA	NA	0.0000	0.000	0.000	NA	NA	0.000
min	1.00000	NA	NA	24.5660	-124.548	0.000	NA	NA	-33.000
max	4.00000	NA	NA	49.0006	-67.484	155.186	NA	NA	196.000
range	3.00000	NA	NA	24.4346	57.064	155.186	NA	NA	229.000
sum	4595443.00000	NA	NA	79753713.5844	-214000710.065	1545536.064	NA	NA	136941574.700
median	2.00000	NA	NA	35.7526	-91.080	0.210	NA	NA	64.000
mean	2.07513	NA	NA	36.0139	-96.635	0.698	NA	NA	61.838
SE.mean	0.00026	NA	NA	0.0037	0.012	0.001	NA	NA	0.012



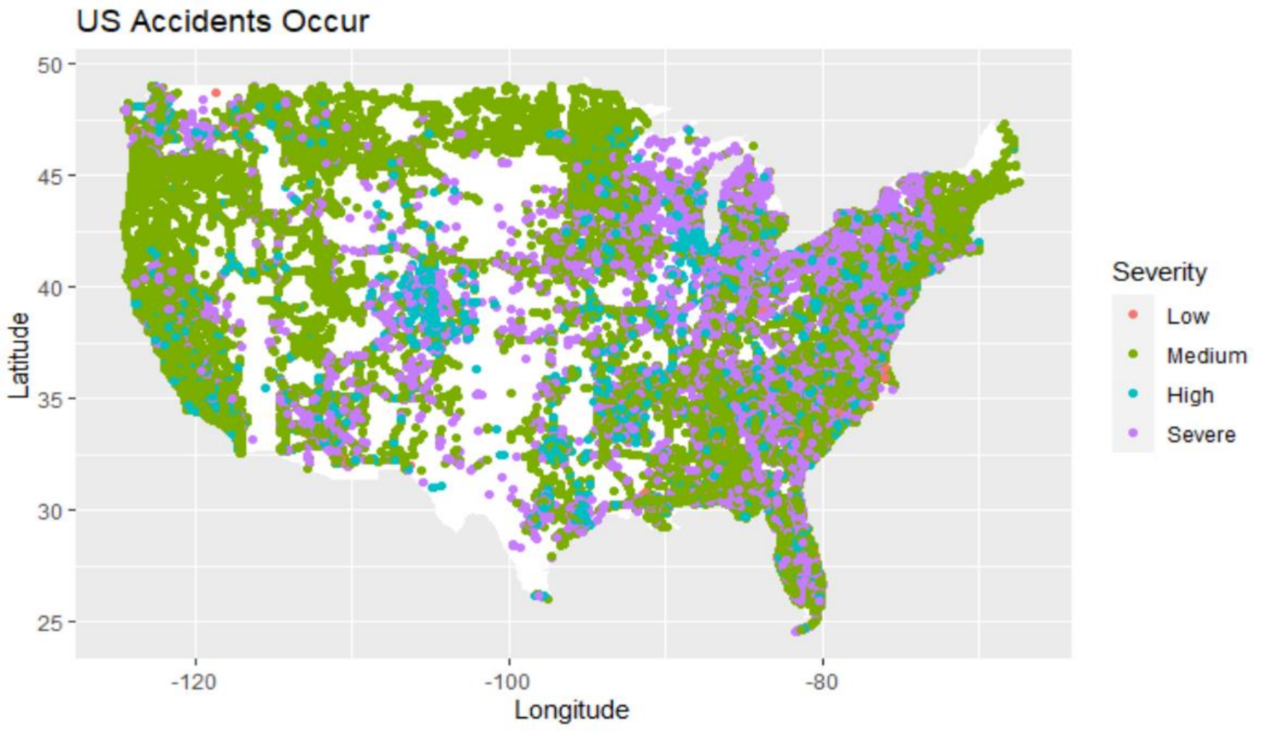
Wind_Chill.F. <dbl>	Humidity... <dbl>	Pressure.in. <dbl>	Visibility.mi. <dbl>	Wind_Speed.mph. <dbl>	Precipitation.in. <dbl>	Amenity <lgl>	Bump <lgl>	Give_Way <lgl>	Junction <lgl>
2214528.000	2214528.000	2214528.00000	2214528.0000	2214528.0000	2214528.000000	NA	NA	NA	NA
1304.000	0.000	0.00000	2972.0000	409610.0000	2050622.000000	NA	NA	NA	NA
0.000	0.000	0.00000	0.0000	0.0000	0.000000	NA	NA	NA	NA
-50.100	1.000	16.72000	0.0000	0.0000	0.000000	NA	NA	NA	NA
196.000	100.000	58.90000	100.0000	1087.0000	24.000000	NA	NA	NA	NA
246.100	99.000	42.18000	100.0000	1087.0000	24.000000	NA	NA	NA	NA
134455418.700	143217219.000	65045573.53000	20035610.5100	15835461.7000	12601.550000	NA	NA	NA	NA
64.000	67.000	29.73000	10.0000	7.0000	0.000000	NA	NA	NA	NA
60.715	64.672	29.37221	9.0474	7.1507	0.005690	NA	NA	NA	NA
0.014	0.015	0.00073	0.0018	0.0037	0.000039	NA	NA	NA	NA

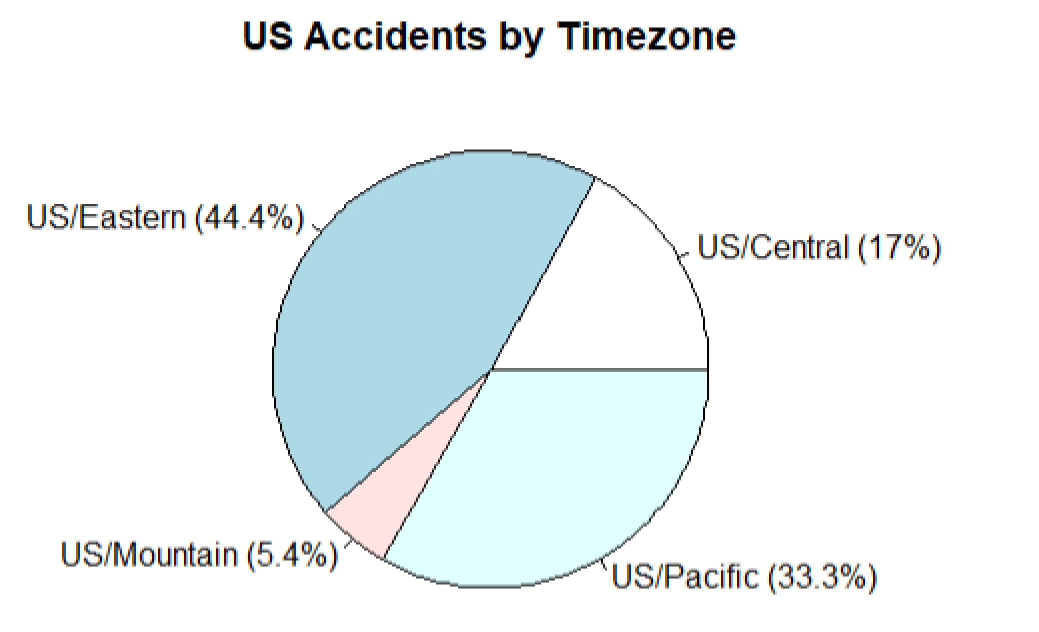


Give_Way <lgl>	Junction <lgl>	No_Exit <lgl>	Railway <lgl>	Roundabout <lgl>	Station <lgl>	Stop <lgl>	Traffic_Calming <lgl>	Traffic_Signal <lgl>	Sunrise_Sunset <lgl>
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Car Accidents Occur in US

Which areas of the United States have more of the reported accidents in this data set?

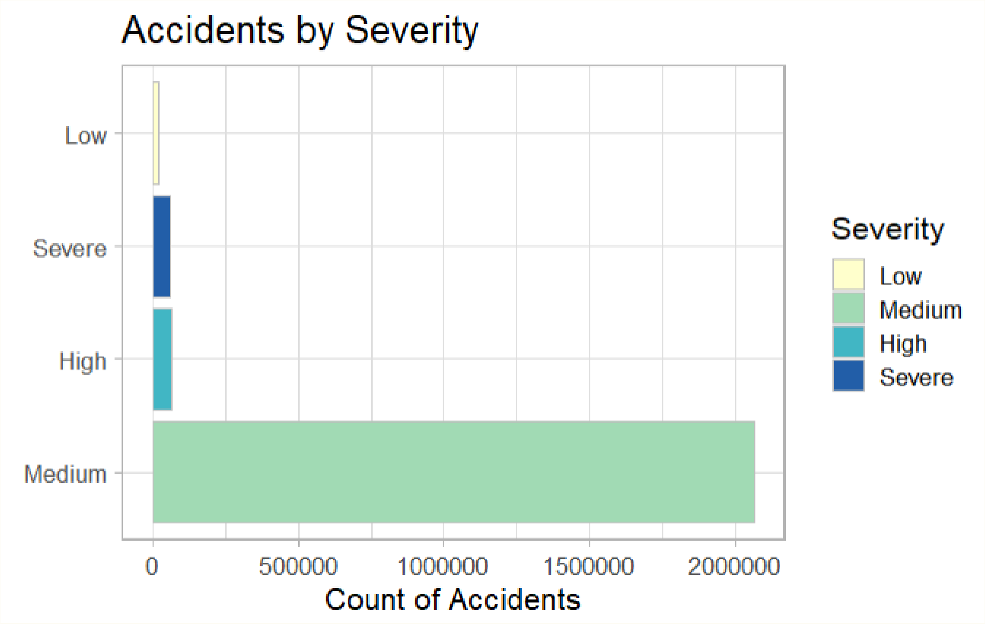




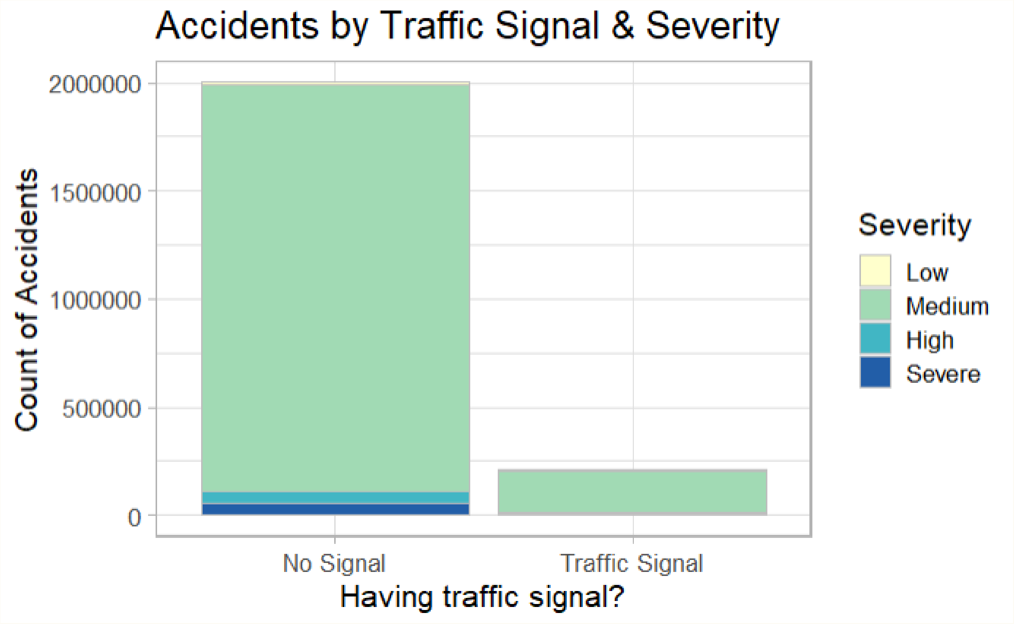
Most of the accidents occurred in the eastern region.

Visualize the data

The highest frequency of severity is Medium

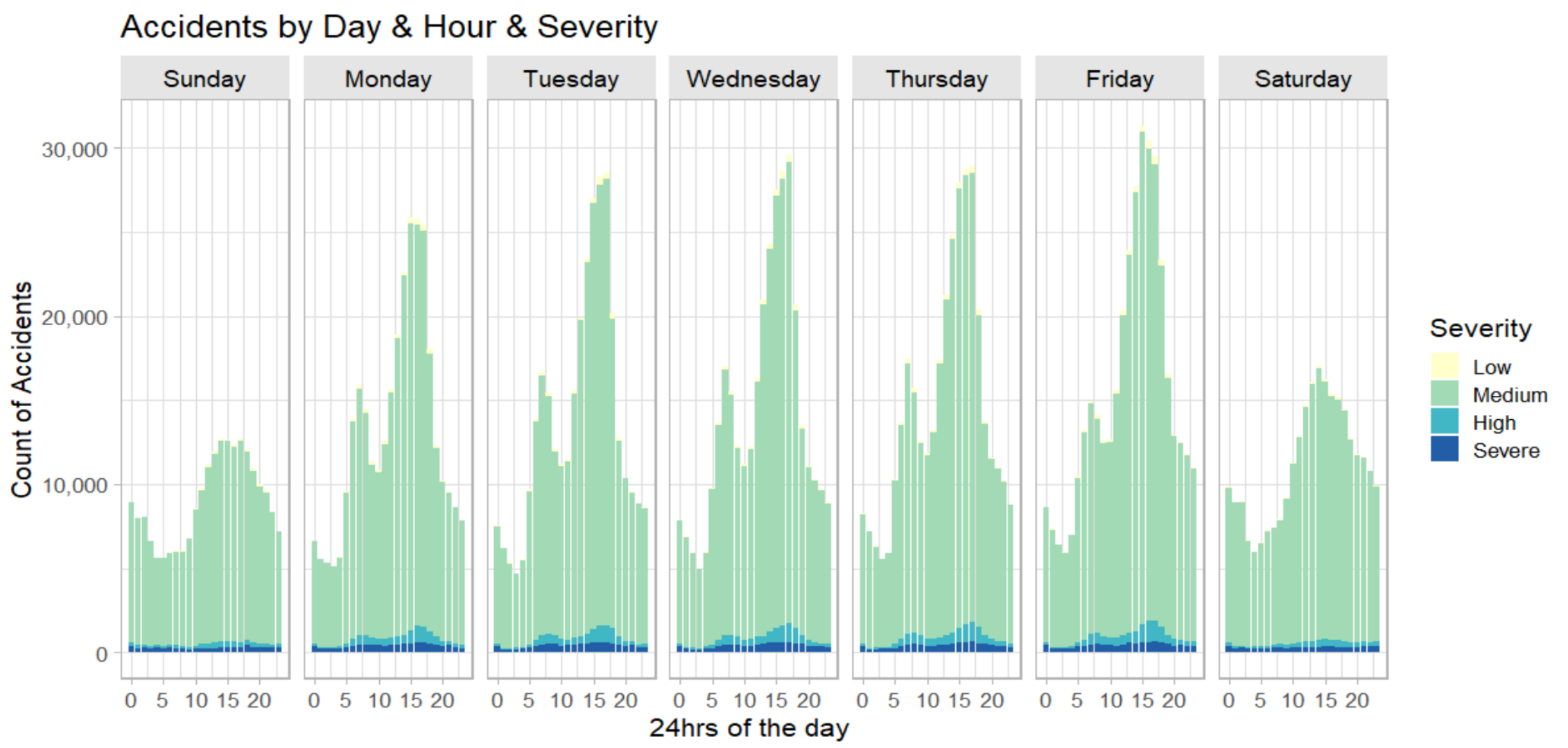


Especially when there is no traffic signal !



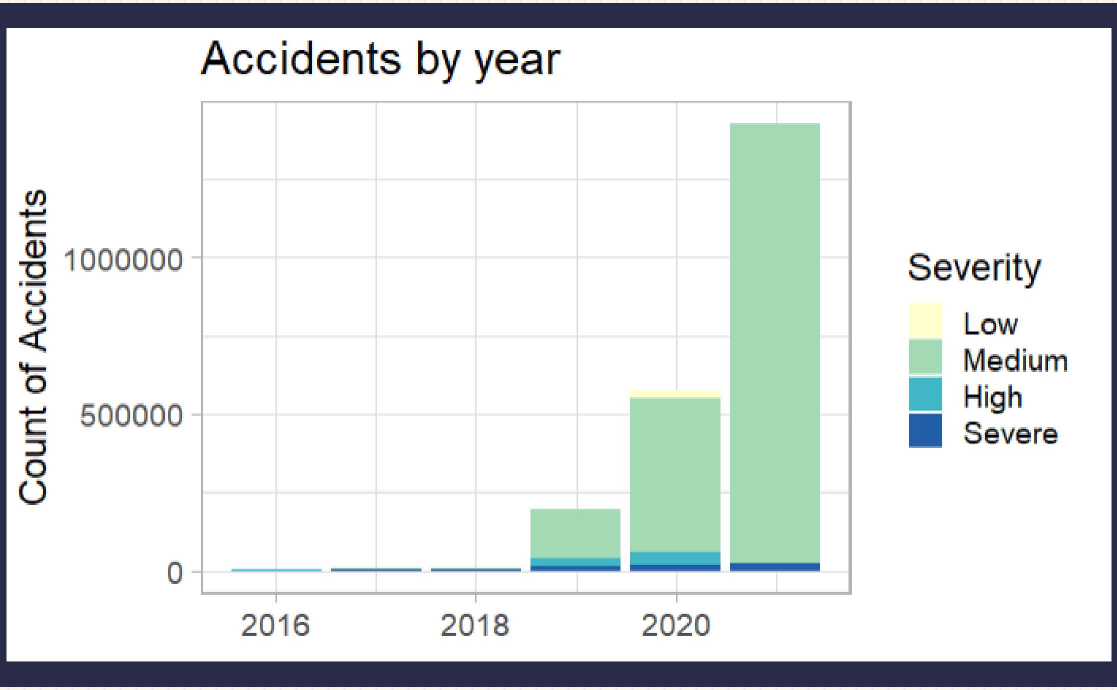
Display the time of accidents

Most accidents happened in the weekday around 16:00~17:00, it might because of the commuting time.

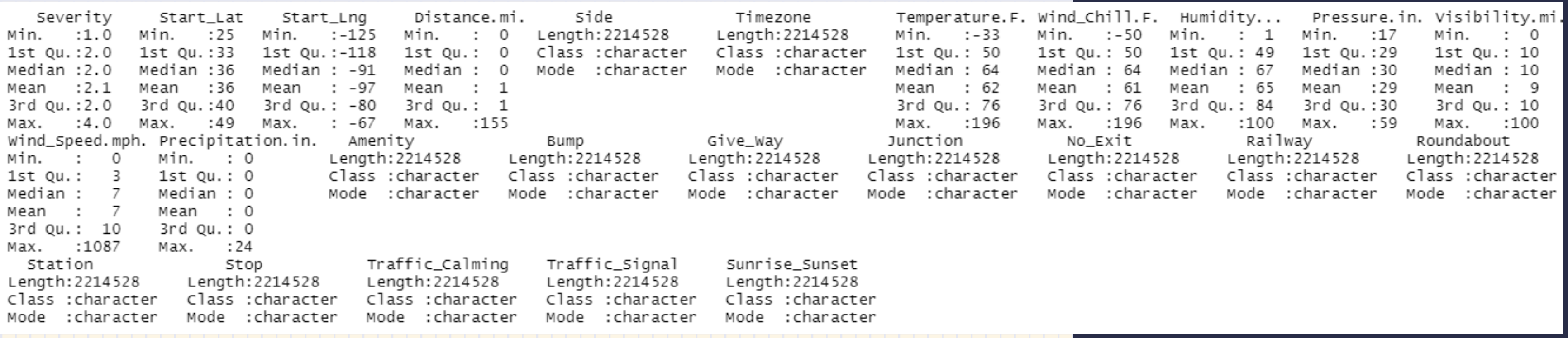


Display the US accidents by Year

- 2020 & 2021 have most accidents might because of the COVID.



Data Cleaning



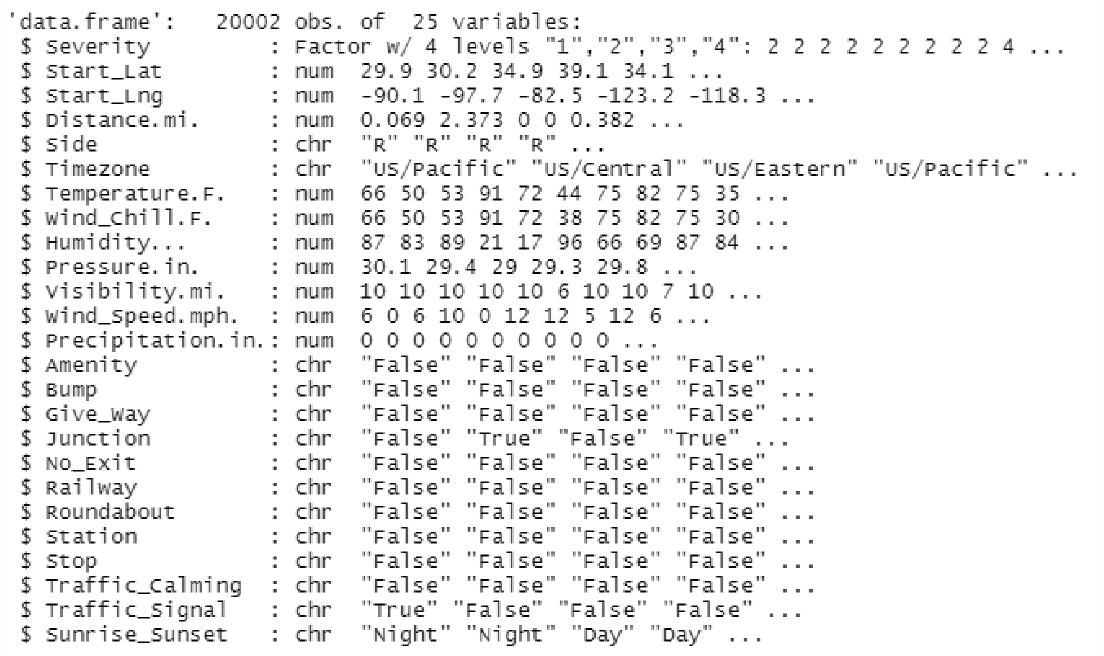
```
   Severity     Start_Lat    Start_Lng     Distance.mi.     Side             Timezone         Temperature.F. Wind_Chill.F.  Humidity...   Pressure.in. Visibility.mi.
Min.   :1.0   Min.   :25   Min.   :-125   Min.   :  0   Length:2214528     Length:2214528     Min.   :-33    Min.   :-50   Min.   :  1   Min.   :17    Min.   :  0
1st Qu.:2.0   1st Qu.:33   1st Qu.:-118   1st Qu.:  0   Class :character   Class :character   1st Qu.: 50    1st Qu.: 50   1st Qu.: 49   1st Qu.:29    1st Qu.: 10
Median :2.0   Median :36   Median : -91   Median :  0   Mode  :character   Mode  :character   Median : 64    Median : 64   Median : 67   Median :30    Median : 10
Mean   :2.1   Mean   :36   Mean   : -97   Mean   :  1                                         Mean   : 62    Mean   : 61   Mean   : 65   Mean   :29    Mean   :  9
3rd Qu.:2.0   3rd Qu.:40   3rd Qu.: -80   3rd Qu.:  1                                         3rd Qu.: 76    3rd Qu.: 76   3rd Qu.: 84   3rd Qu.:30    3rd Qu.: 10
Max.   :4.0   Max.   :49   Max.   : -67   Max.   :155                                         Max.   :196    Max.   :196   Max.   :100   Max.   :59    Max.   :100
Wind_Speed.mph. Precipitation.in.   Amenity              Bump             Give_Way           Junction           No_Exit            Railway            Roundabout
Min.   :   0    Min.   : 0        Length:2214528     Length:2214528     Length:2214528     Length:2214528     Length:2214528     Length:2214528     Length:2214528
1st Qu.:   3    1st Qu.: 0        Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character
Median :   7    Median : 0        Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character
Mean   :   7    Mean   : 0
3rd Qu.:  10    3rd Qu.: 0
Max.   :1087    Max.   :24
  Station              Stop           Traffic_Calming    Traffic_Signal     Sunrise_Sunset
Length:2214528     Length:2214528     Length:2214528     Length:2214528     Length:2214528
Class :character   Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

- Remove unneeded variables
- Define NA values

Data Partition

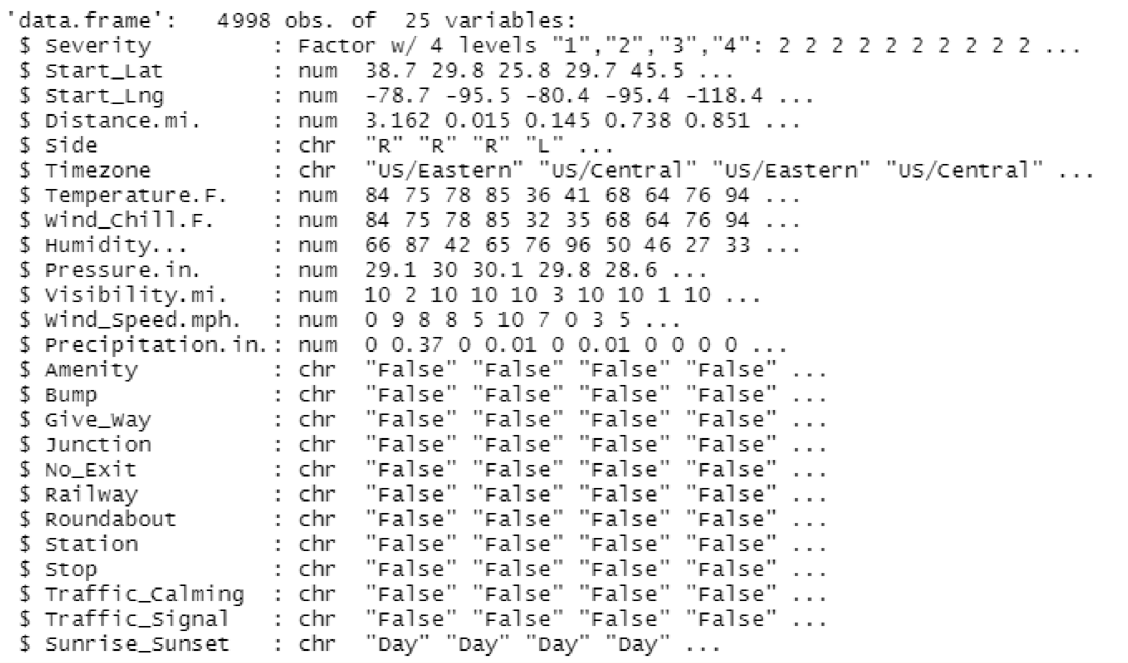
The data is shortened to 25000 observations with random sampling.

TRAINING DATA (80%)



```
'data.frame':	20002 obs. of  25 variables:
 $ Severity         : Factor w/ 4 levels "1","2","3","4": 2 2 2 2 2 2 2 2 2 4 ...
 $ Start_Lat        : num  29.9 30.2 34.9 39.1 34.1 ...
 $ Start_Lng        : num  -90.1 -97.7 -82.5 -123.2 -118.3 ...
 $ Distance.mi.     : num  0.069 2.373 0 0 0.382 ...
 $ Side             : chr  "R" "R" "R" "R" ...
 $ Timezone         : chr  "US/Pacific" "US/Central" "US/Eastern" "US/Pacific" ...
 $ Temperature.F.   : num  66 50 53 91 72 44 75 82 75 35 ...
 $ Wind_Chill.F.    : num  66 50 53 91 72 38 75 82 75 30 ...
 $ Humidity...      : num  87 83 89 21 17 96 66 69 87 84 ...
 $ Pressure.in.     : num  30.1 29.4 29 29.3 29.8 ...
 $ Visibility.mi.   : num  10 10 10 10 10 6 10 10 7 10 ...
 $ Wind_Speed.mph.  : num  6 0 6 10 0 12 12 5 12 6 ...
 $ Precipitation.in.: num  0 0 0 0 0 0 0 0 0 0 ...
 $ Amenity          : chr  "False" "False" "False" "False" ...
 $ Bump             : chr  "False" "False" "False" "False" ...
 $ Give_Way         : chr  "False" "False" "False" "False" ...
 $ Junction         : chr  "False" "True" "False" "True" ...
 $ No_Exit          : chr  "False" "False" "False" "False" ...
 $ Railway          : chr  "False" "False" "False" "False" ...
 $ Roundabout       : chr  "False" "False" "False" "False" ...
 $ Station          : chr  "False" "False" "False" "False" ...
 $ Stop             : chr  "False" "False" "False" "False" ...
 $ Traffic_Calming  : chr  "False" "False" "False" "False" ...
 $ Traffic_Signal   : chr  "True" "False" "False" "False" ...
 $ Sunrise_Sunset   : chr  "Night" "Night" "Day" "Day" ...
```

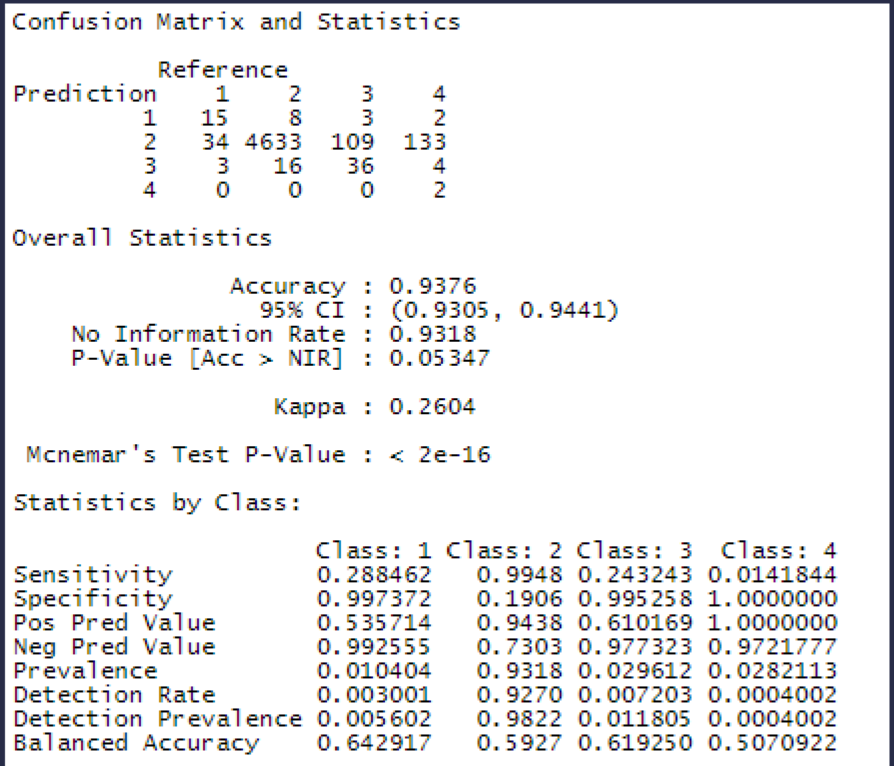
TESTING DATA (20%)



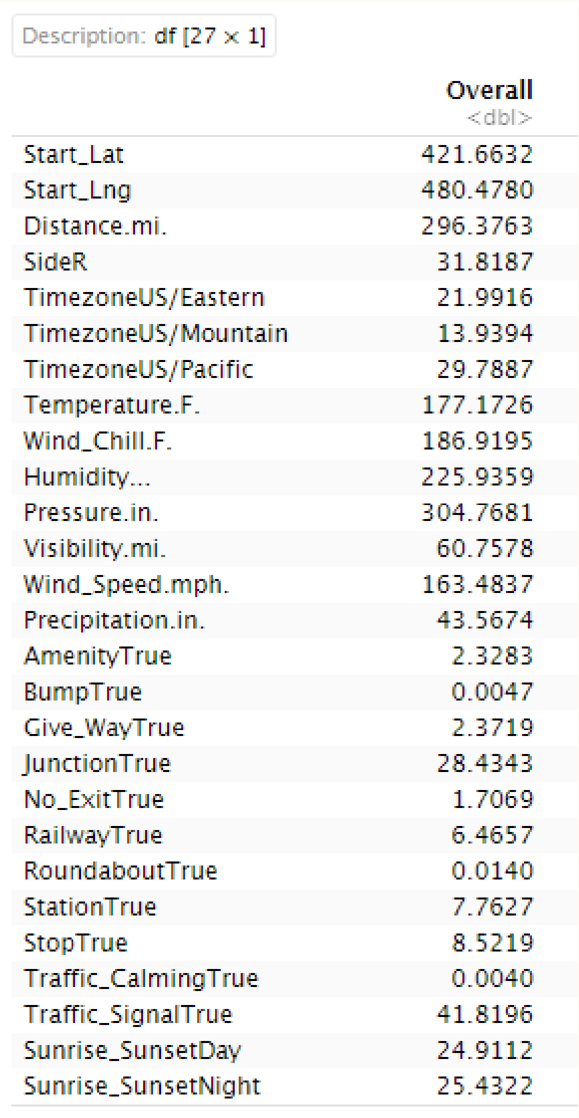
```
'data.frame':	4998 obs. of  25 variables:
 $ Severity         : Factor w/ 4 levels "1","2","3","4": 2 2 2 2 2 2 2 2 2 2 ...
 $ Start_Lat        : num  38.7 29.8 25.8 29.7 45.5 ...
 $ Start_Lng        : num  -78.7 -95.5 -80.4 -95.4 -118.4 ...
 $ Distance.mi.     : num  3.162 0.015 0.145 0.738 0.851 ...
 $ Side             : chr  "R" "R" "R" "L" ...
 $ Timezone         : chr  "US/Eastern" "US/Central" "US/Eastern" "US/Central" ...
 $ Temperature.F.   : num  84 75 78 85 36 41 68 64 76 94 ...
 $ Wind_Chill.F.    : num  84 75 78 85 32 35 68 64 76 94 ...
 $ Humidity...      : num  66 87 42 65 76 96 50 46 27 33 ...
 $ Pressure.in.     : num  29.1 30 30.1 29.8 28.6 ...
 $ Visibility.mi.   : num  10 2 10 10 10 3 10 10 1 10 ...
 $ Wind_Speed.mph.  : num  0 9 8 8 5 10 7 0 3 5 ...
 $ Precipitation.in.: num  0 0.37 0 0.01 0 0.01 0 0 0 0 ...
 $ Amenity          : chr  "False" "False" "False" "False" ...
 $ Bump             : chr  "False" "False" "False" "False" ...
 $ Give_Way         : chr  "False" "False" "False" "False" ...
 $ Junction         : chr  "False" "False" "False" "False" ...
 $ No_Exit          : chr  "False" "False" "False" "False" ...
 $ Railway          : chr  "False" "False" "False" "False" ...
 $ Roundabout       : chr  "False" "False" "False" "False" ...
 $ Station          : chr  "False" "False" "False" "False" ...
 $ Stop             : chr  "False" "False" "False" "False" ...
 $ Traffic_Calming  : chr  "False" "False" "False" "False" ...
 $ Traffic_Signal   : chr  "False" "False" "False" "False" ...
 $ Sunrise_Sunset   : chr  "Day" "Day" "Day" "Day" ...
```

Random Forest model for severity

We see that the accuracy of the model is ~94% and the kappa is ~26%.
Not a good model.



```
Confusion Matrix and Statistics

          Reference
Prediction    1    2    3    4
         1   15    8    3    2
         2   34 4633  109  133
         3    3   16   36    4
         4    0    0    0    2

Overall Statistics
                                          
               Accuracy : 0.9376          
                 95% CI : (0.9305, 0.9441)
    No Information Rate : 0.9318          
    P-Value [Acc > NIR] : 0.05347         
                                          
                  Kappa : 0.2604          
                                          
 Mcnemar's Test P-Value : < 2e-16         

Statistics by Class:

                     Class: 1 Class: 2 Class: 3  Class: 4
Sensitivity          0.288462   0.9948 0.243243 0.0141844
Specificity          0.997372   0.1906 0.995258 1.0000000
Pos Pred Value       0.535714   0.9438 0.610169 1.0000000
Neg Pred Value       0.992555   0.7303 0.977323 0.9721777
Prevalence           0.010404   0.9318 0.029612 0.0282113
Detection Rate       0.003001   0.9270 0.007203 0.0004002
Detection Prevalence 0.005602   0.9822 0.011805 0.0004002
Balanced Accuracy    0.642917   0.5927 0.619250 0.5070922
```



Description: df [27 × 1]

	Overall <dbl>
Start_Lat	421.6632
Start_Lng	480.4780
Distance.mi.	296.3763
SideR	31.8187
TimezoneUS/Eastern	21.9916
TimezoneUS/Mountain	13.9394
TimezoneUS/Pacific	29.7887
Temperature.F.	177.1726
Wind_Chill.F.	186.9195
Humidity...	225.9359
Pressure.in.	304.7681
Visibility.mi.	60.7578
Wind_Speed.mph.	163.4837
Precipitation.in.	43.5674
AmenityTrue	2.3283
BumpTrue	0.0047
Give_WayTrue	2.3719
JunctionTrue	28.4343
No_ExitTrue	1.7069
RailwayTrue	6.4657
RoundaboutTrue	0.0140
StationTrue	7.7627
StopTrue	8.5219
Traffic_CalmingTrue	0.0040
Traffic_SignalTrue	41.8196
Sunrise_SunsetDay	24.9112
Sunrise_SunsetNight	25.4322

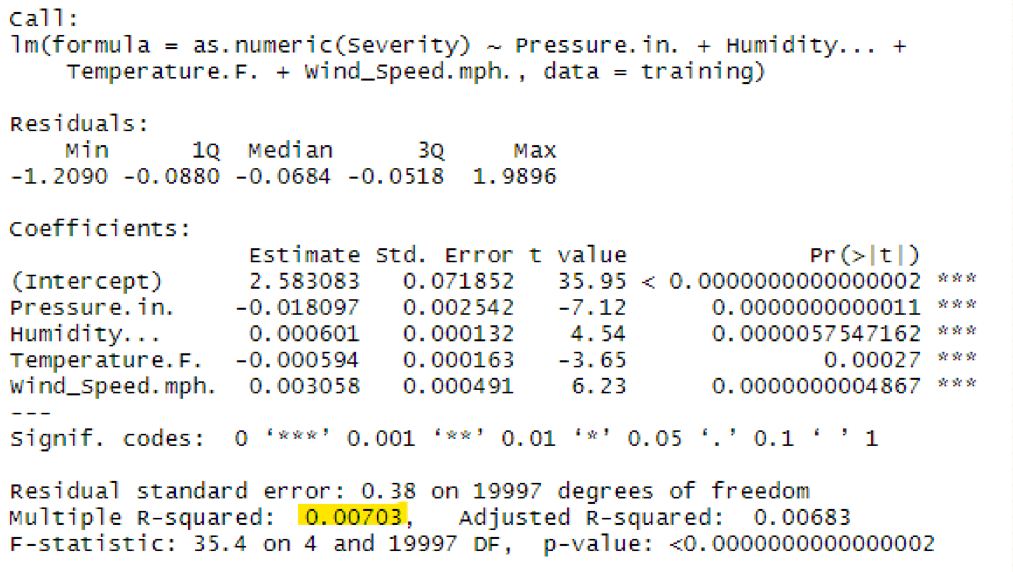
To examine …
which factors for severity accidents most important.

Linear regression model

With more important variables

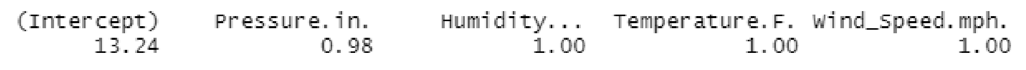
(Severity~ Pressure+Humidity+Temperature+Wind Speed)

Linear Regression



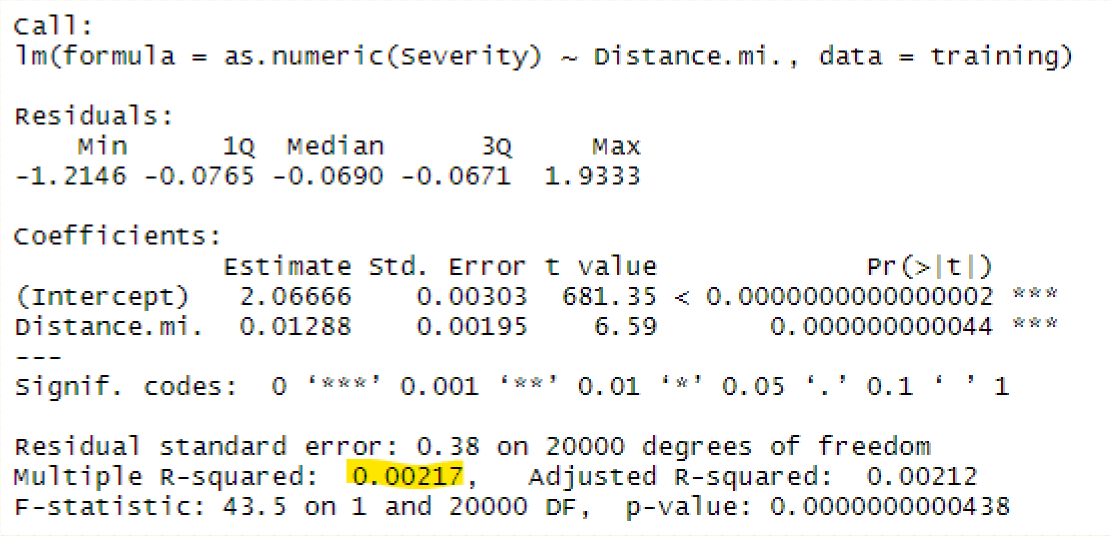
```
Call:
lm(formula = as.numeric(Severity) ~ Pressure.in. + Humidity... +
    Temperature.F. + Wind_Speed.mph., data = training)

Residuals:
    Min      1Q  Median      3Q     Max
-1.2090 -0.0880 -0.0684 -0.0518  1.9896

Coefficients:
                 Estimate Std. Error t value             Pr(>|t|)
(Intercept)      2.583083   0.071852   35.95 < 0.0000000000000002 ***
Pressure.in.    -0.018097   0.002542   -7.12      0.0000000000011 ***
Humidity...      0.000601   0.000132    4.54      0.0000057547162 ***
Temperature.F.  -0.000594   0.000163   -3.65              0.00027 ***
Wind_Speed.mph.  0.003058   0.000491    6.23      0.0000000004867 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.38 on 19997 degrees of freedom
Multiple R-squared:  0.00703,   Adjusted R-squared:  0.00683
F-statistic: 35.4 on 4 and 19997 DF,  p-value: <0.0000000000000002
```

Model Coefficients



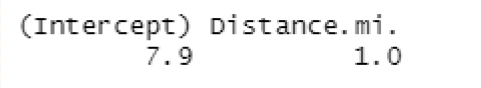
```
    (Intercept)    Pressure.in.     Humidity...  Temperature.F. Wind_Speed.mph.
          13.24            0.98            1.00            1.00            1.00
```

Linear Regression



```
Call:
lm(formula = as.numeric(Severity) ~ Distance.mi., data = training)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.2146 -0.0765 -0.0690 -0.0671  1.9333 

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)    
(Intercept)   2.06666    0.00303  681.35 < 0.0000000000000002 ***
Distance.mi.  0.01288    0.00195    6.59       0.000000000044 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.38 on 20000 degrees of freedom
Multiple R-squared:  0.00217,	Adjusted R-squared:  0.00212 
F-statistic: 43.5 on 1 and 20000 DF,  p-value: 0.0000000000438
```

Model Coefficients



```
(Intercept) Distance.mi. 
        7.9          1.0 
```

Linear regression model

With more important variables

(Severity~ Distance)

Conclusion

Link to RMarkdown.html

- Most car accidents happened in the Eastern United State. (More population there)
- Most car accidents happened when there is no traffic signal.
- Most car accidents happened in the weekday commuting period.
- There are more accidents since 2020, it might due to pandemic and the decrease commuting frequency of public transportation.
- We would need more information to find out what particular factors of car accidents are associated with traffic interference.