PDA_Project_US_Car_Accidents_Analysis

Group13

2023-11-28

US Car Accident Analysis

Inlcuded Libraries

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.3     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.3     ✓ tibble    3.2.1
## ✓ lubridate 1.9.2     ✓ tidyr     1.3.0
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
## Loading required package: ISLR
##
##
## Attaching package: 'MASS'
##
##
## The following object is masked from 'package:dplyr':
##
##     select
##
##
##
## Attaching package: 'MLmetrics'
##
##
## The following object is masked from 'package:base':
##
##     Recall
```

```
## Warning: package 'corrplot' was built under R version 4.3.2
```

```
## corrplot 0.92 loaded
```

PART I: Data Preperation and cleaning

```
# read the file into a variable
USAcaraccidents = read.csv("C:/Users/anush/OneDrive/Documents/Data Analytics/US_Accidents_March23_sampled_500k.csv")
```

```
head(USAcaraccidents)
```

```
##          ID  Source Severity                   Start_Time
## 1 A-2047758 Source2        2           2019-06-12 10:10:56
## 2 A-4694324 Source1        2 2022-12-03 23:37:14.000000000
## 3 A-5006183 Source1        2 2022-08-20 13:13:00.000000000
## 4 A-4237356 Source1        2           2022-02-21 17:43:04
## 5 A-6690583 Source1        2           2020-12-04 01:46:00
## 6 A-1101469 Source2        2           2021-03-29 07:03:58
##                        End_Time Start_Lat  Start_Lng  End_Lat    End_Lng
## 1           2019-06-12 10:55:58  30.64121  -91.15348       NA         NA
## 2 2022-12-04 01:56:53.000000000  38.99056  -77.39907 38.99004  -77.39828
## 3 2022-08-20 15:22:45.000000000  34.66119 -120.49282 34.66119 -120.49244
## 4           2022-02-21 19:43:23  43.68059  -92.99332 43.68057  -92.97222
## 5           2020-12-04 04:13:09  35.39548 -118.98518 35.39548 -118.98600
## 6           2021-03-29 08:51:01  42.53208  -70.94427       NA         NA
##   Distance.mi.
## 1        0.000
## 2        0.056
## 3        0.022
## 4        1.054
## 5        0.046
## 6        0.000
##                                                  Description
## 1           Accident on LA-19 Baker-Zachary Hwy at Lower Zachary Rd.
## 2 Incident on FOREST RIDGE DR near PEPPERIDGE PL Drive with caution.
## 3       Accident on W Central Ave from Floradale Ave to Western Ave.
## 4             Incident on I-90 EB near REST AREA Drive with caution.
## 5               RP ADV THEY LOCATED SUSP VEH OF 20002 - 726 CRAWFORD
## 6                                Accident on Forest St at Lowell St.
##             Street        City           County State    Zipcode Country
## 1       Highway 19     Zachary East Baton Rouge    LA 70791-4610      US
## 2  Forest Ridge Dr    Sterling          Loudoun    VA 20164-2813      US
## 3    Floradale Ave      Lompoc    Santa Barbara    CA      93436      US
## 4       14th St NW      Austin            Mower    MN      55912      US
## 5       River Blvd Bakersfield             Kern    CA 93305-2649      US
## 6        Lowell St     Peabody            Essex    MA 01960-4275      US
##     Timezone Airport_Code   Weather_Timestamp Temperature.F. Wind_Chill.F.
## 1 US/Central         KBTR 2019-06-12 09:53:00             77             77
## 2 US/Eastern         KIAD 2022-12-03 23:52:00             45             43
## 3 US/Pacific         KLPC 2022-08-20 12:56:00             68             68
## 4 US/Central         KAUM 2022-02-21 17:35:00             27             15
## 5 US/Pacific         KBFL 2020-12-04 01:54:00             42             42
## 6 US/Eastern         KBVY 2021-03-29 06:53:00             42             35
##   Humidity... Pressure.in. Visibility.mi. Wind_Direction Wind_Speed.mph.
## 1          62        29.92             10             NW               5
## 2          48        29.91             10              W               5
## 3          73        29.79             10              W              13
## 4          86        28.49             10            ENE              15
## 5          34        29.77             10           CALM               0
## 6          58        29.37             10              W              13
##   Precipitation.in. Weather_Condition Amenity  Bump Crossing Give_Way Junction
## 1                 0              Fair   False False    False    False    False
## 2                 0              Fair   False False    False    False    False
```

```
## 3                0             Fair   False False    False    False    False
## 4                0       Wintry Mix   False False    False    False    False
## 5                0             Fair   False False    False    False    False
## 6                0             Fair   False False    False    False    False
##   No_Exit Railway Roundabout Station  Stop Traffic_Calming Traffic_Signal
## 1   False   False      False   False False           False           True
## 2   False   False      False   False False           False          False
## 3   False   False      False   False False           False           True
## 4   False   False      False   False False           False          False
## 5   False   False      False   False False           False          False
## 6   False   False      False   False False           False           True
##   Turning_Loop Sunrise_Sunset Civil_Twilight Nautical_Twilight
## 1        False            Day            Day              Day
## 2        False          Night          Night            Night
## 3        False            Day            Day              Day
## 4        False            Day            Day              Day
## 5        False          Night          Night            Night
## 6        False            Day            Day              Day
##   Astronomical_Twilight
## 1                   Day
## 2                 Night
## 3                   Day
## 4                   Day
## 5                 Night
## 6                   Day
```

```
#Checking for Null Values
sum(is.na(USAcaraccidents))
```

```
## [1] 791189
```

```
#Removing null value rows
USAcaraccidents <- USAcaraccidents[complete.cases(USAcaraccidents), ]
```

```
#Confirming cleanup
sum(is.na(USAcaraccidents))
```

```
## [1] 0
```

```
#Total size
nrow(USAcaraccidents)
```

```
## [1] 232130
```

```
summary(USAcaraccidents)
```

```
##       ID               Source             Severity      Start_Time
##  Length:232130      Length:232130      Min.   :1.000   Length:232130
##  Class :character   Class :character   1st Qu.:2.000   Class :character
##  Mode  :character   Mode  :character   Median :2.000   Mode  :character
##                                        Mean   :2.076
##                                        3rd Qu.:2.000
##                                        Max.   :4.000
##    End_Time           Start_Lat       Start_Lng          End_Lat
##  Length:232130      Min.   :24.57   Min.   :-124.50   Min.   :24.57
##  Class :character   1st Qu.:33.22   1st Qu.:-117.55   1st Qu.:33.22
##  Mode  :character   Median :36.06   Median : -87.36   Median :36.06
##                     Mean   :36.13   Mean   : -95.25   Mean   :36.13
##                     3rd Qu.:40.13   3rd Qu.: -80.22   3rd Qu.:40.13
##                     Max.   :48.99   Max.   : -67.48   Max.   :49.00
##     End_Lng         Distance.mi.      Description           Street
##  Min.   :-124.50   Min.   :  0.0000   Length:232130      Length:232130
##  1st Qu.:-117.55   1st Qu.:  0.0670   Class :character   Class :character
##  Median : -87.37   Median :  0.2660   Mode  :character   Mode  :character
##  Mean   : -95.25   Mean   :  0.8579
##  3rd Qu.: -80.21   3rd Qu.:  0.9270
##  Max.   : -67.48   Max.   :149.6900
##      City               County             State              Zipcode
##  Length:232130      Length:232130      Length:232130      Length:232130
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##    Country            Timezone         Airport_Code       Weather_Timestamp
##  Length:232130      Length:232130      Length:232130      Length:232130
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Temperature.F.   Wind_Chill.F.     Humidity...      Pressure.in.
##  Min.   :-29.00   Min.   :-52.00   Min.   :  1.00   Min.   :19.36
##  1st Qu.: 48.00   1st Qu.: 46.00   1st Qu.: 47.00   1st Qu.:29.18
##  Median : 63.00   Median : 63.00   Median : 66.00   Median :29.72
##  Mean   : 61.03   Mean   : 59.69   Mean   : 63.78   Mean   :29.35
##  3rd Qu.: 76.00   3rd Qu.: 76.00   3rd Qu.: 83.00   3rd Qu.:29.96
##  Max.   :140.00   Max.   :140.00   Max.   :100.00   Max.   :30.95
##  Visibility.mi.    Wind_Direction     Wind_Speed.mph.  Precipitation.in.
##  Min.   :  0.000   Length:232130      Min.   :  0.00   Min.   :0.000000
##  1st Qu.: 10.000   Class :character   1st Qu.:  3.00   1st Qu.:0.000000
##  Median : 10.000   Mode  :character   Median :  7.00   Median :0.000000
##  Mean   :  9.054                      Mean   :  7.44   Mean   :0.005643
##  3rd Qu.: 10.000                      3rd Qu.: 10.00   3rd Qu.:0.000000
##  Max.   :100.000                      Max.   :142.00   Max.   :9.960000
##  Weather_Condition    Amenity              Bump             Crossing
##  Length:232130      Length:232130      Length:232130      Length:232130
##  Class :character   Class :character   Class :character   Class :character
```

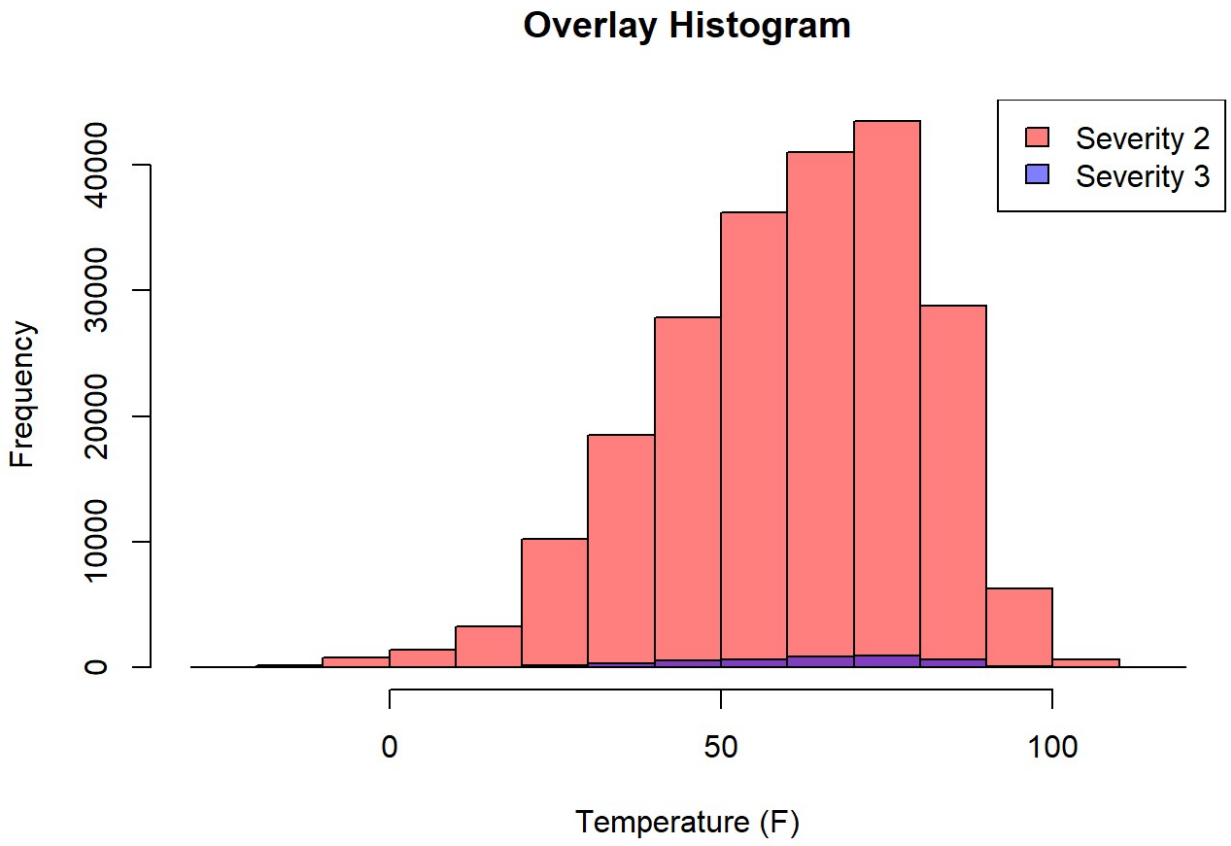
```
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##    Give_Way           Junction           No_Exit            Railway
##  Length:232130      Length:232130      Length:232130      Length:232130
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##   Roundabout          Station              Stop           Traffic_Calming
##  Length:232130      Length:232130      Length:232130      Length:232130
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Traffic_Signal     Turning_Loop       Sunrise_Sunset     Civil_Twilight
##  Length:232130      Length:232130      Length:232130      Length:232130
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Nautical_Twilight  Astronomical_Twilight
##  Length:232130      Length:232130
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

PART II: Exploratory Data Analysis

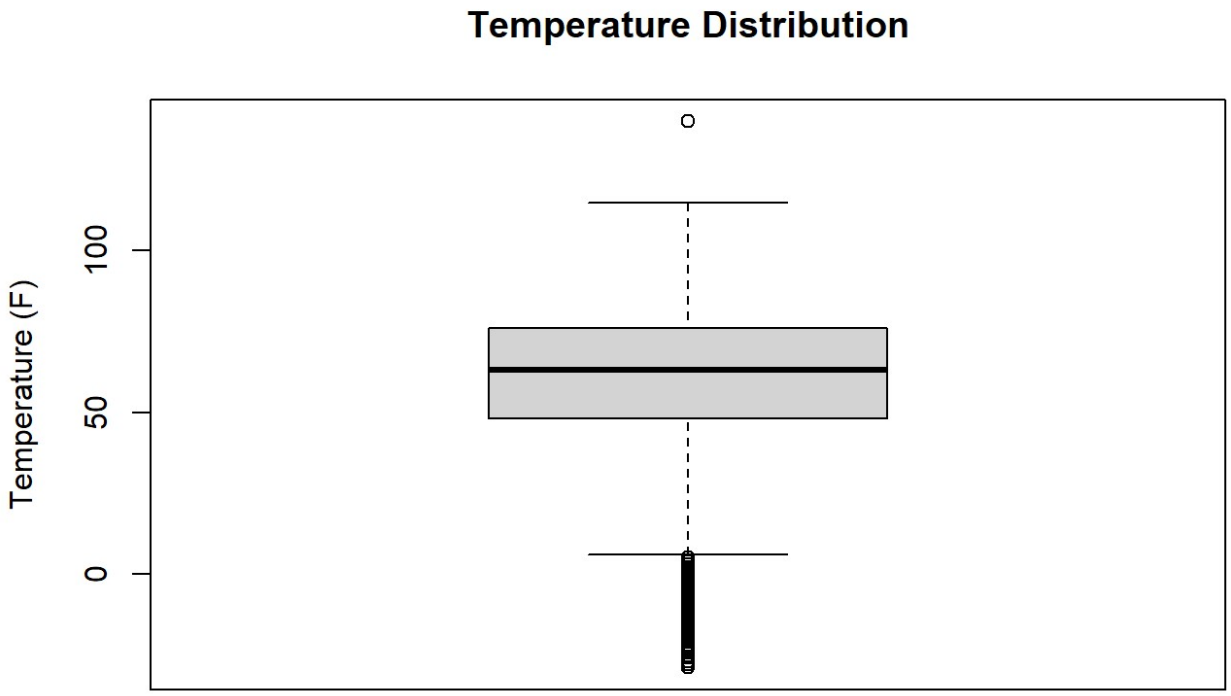
1.Temperature Distribution Analysis:

- The overlay histogram and boxplot reveal that accidents with severity 2 tend to occur more frequently in extreme temperature conditions compared to severity 3 accidents.

```
# Overlay Histogram with Severity
hist(USAcaraccidents$Temperature[USAcaraccidents$Severity == 2], col = rgb(1, 0, 0, 0.5), main =
"Overlay Histogram", xlab = "Temperature (F)", ylab = "Frequency")
hist(USAcaraccidents$Temperature[USAcaraccidents$Severity == 3], col = rgb(0, 0, 1, 0.5), add =
TRUE)
legend("topright", legend = c("Severity 2", "Severity 3"), fill = c(rgb(1, 0, 0, 0.5), rgb(0, 0,
1, 0.5)))
```



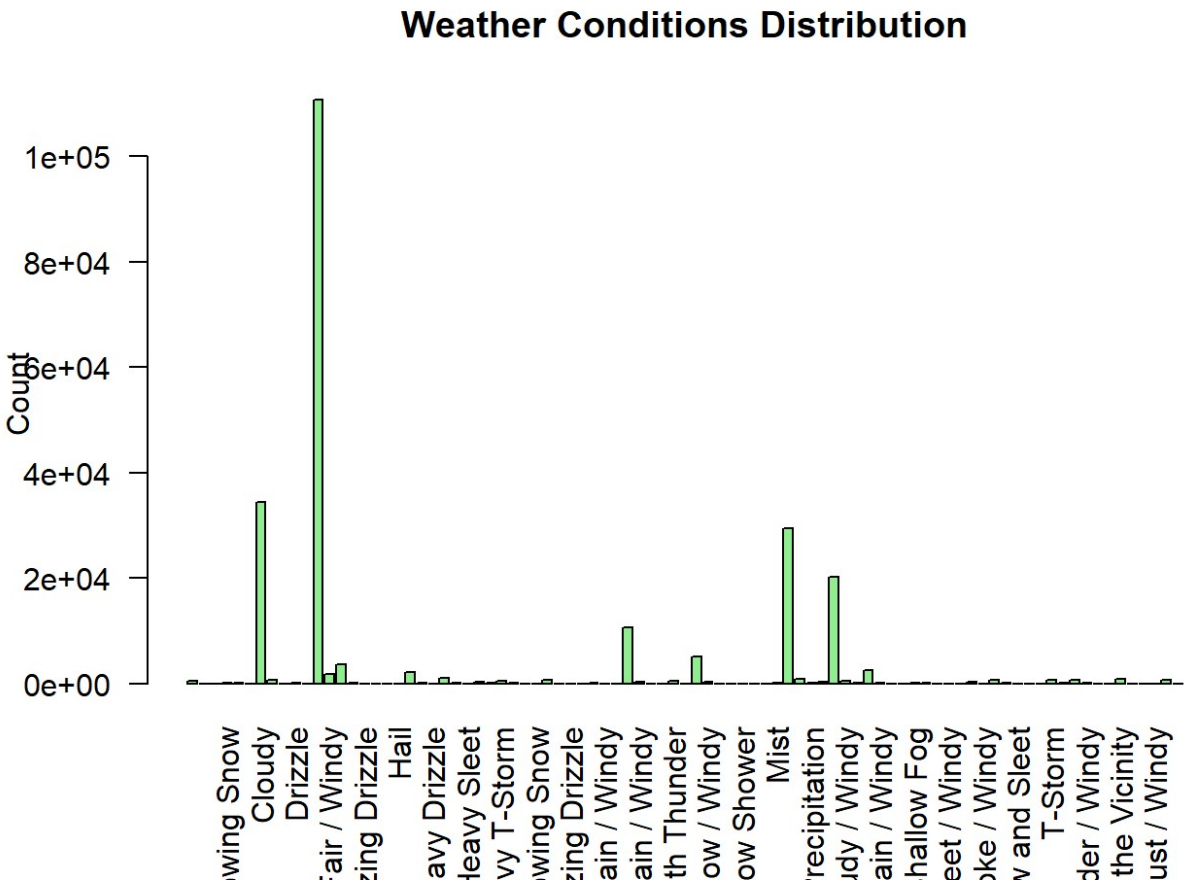

```
# Boxplot for Temperature
boxplot(USAcaraccidents$Temperature, main = "Temperature Distribution", ylab = "Temperature
(F)")
```



2.Weather Conditions Distribution Bar Plot:

- Fair/Windy weather conditions dominate the dataset, followed by cloudy conditions. Adverse weather conditions are less common, indicating that most accidents occur in relatively clear weather.

```
# Bar plot for Weather Conditions
weather_counts <- table(USAcaraccidents$Weather_Condition)
barplot(weather_counts, main = "Weather Conditions Distribution", ylab = "Count", col = "lightgr
een", las = 2)
```

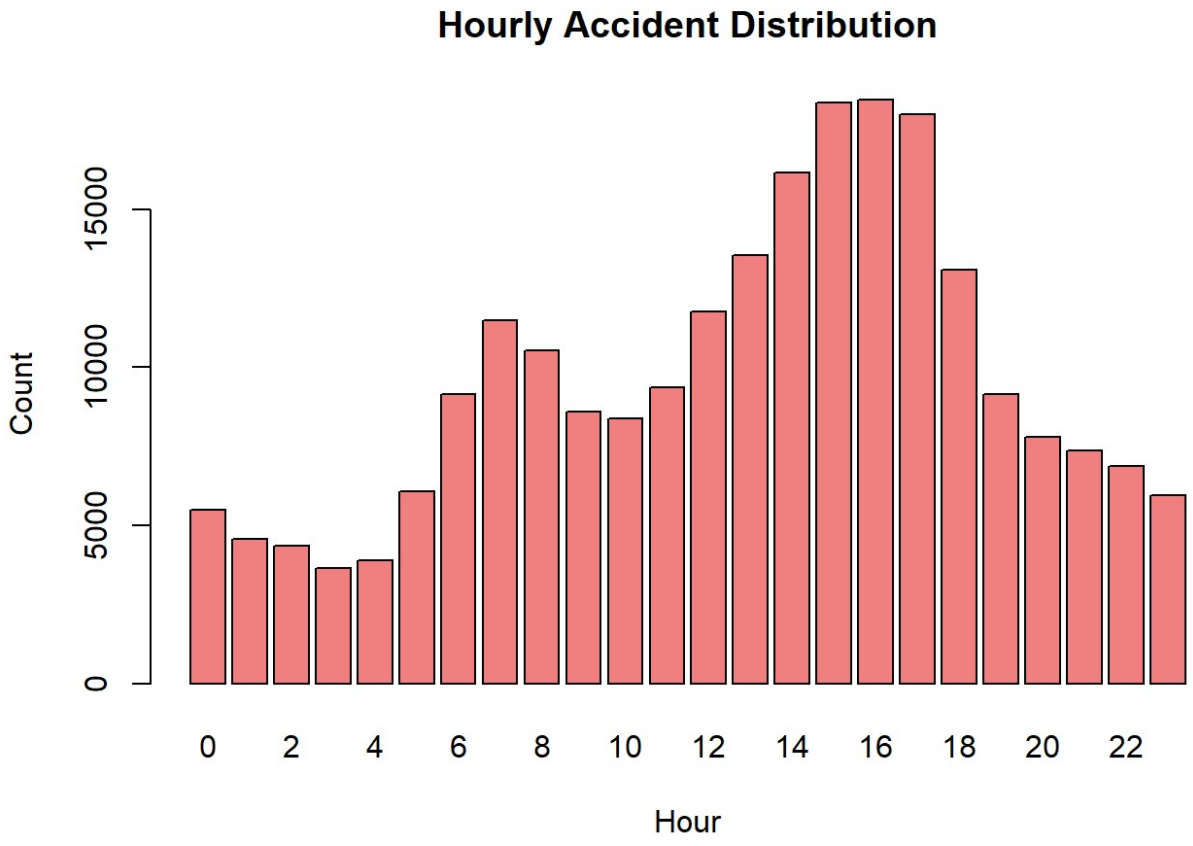


###

3.Hourly Accident Distribution Bar Plot:

- Accidents are relatively evenly distributed across different hours of the day, with a slight increase during the afternoon.

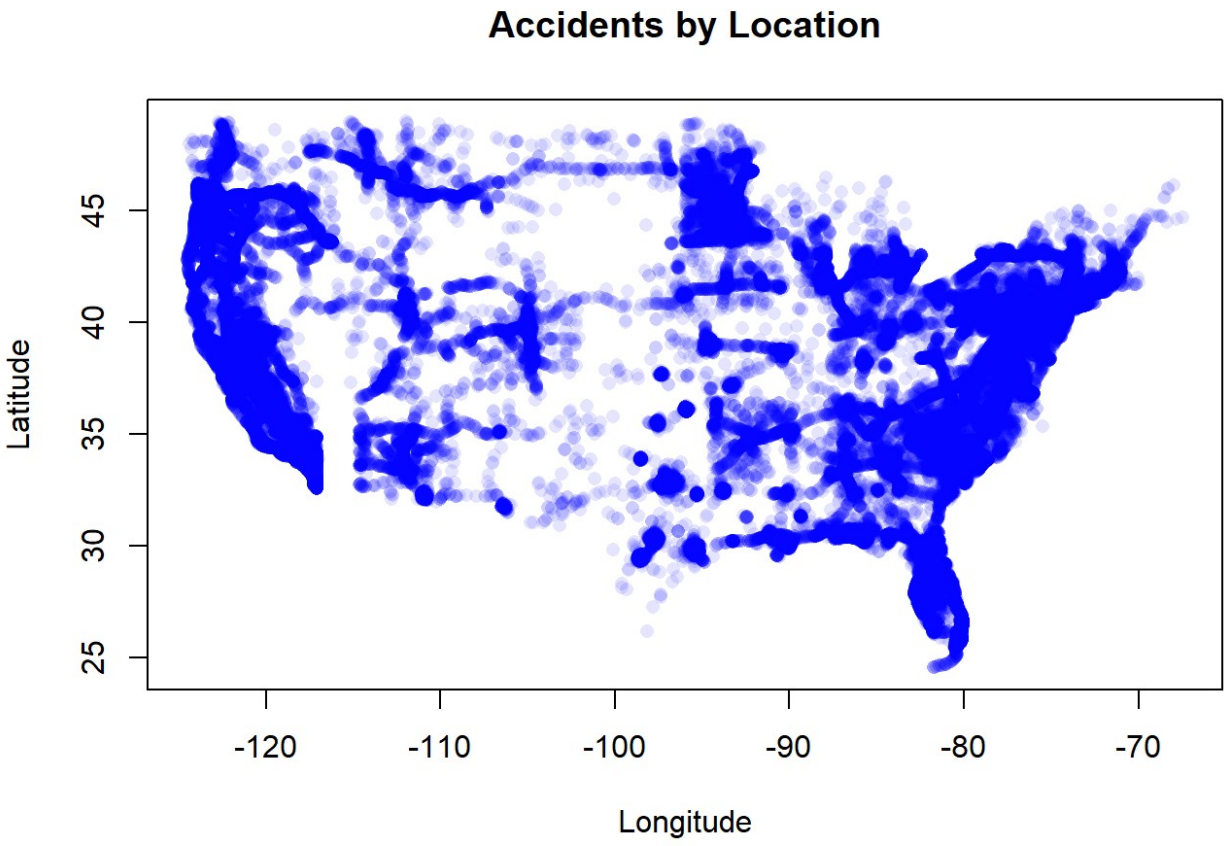
```
# Extracting the hour from Start_Time
USAcaraccidents$Hour <- as.POSIXlt(USAcaraccidents$Start_Time)$hour

# Hourly accident distribution
hourly_counts <- table(USAcaraccidents$Hour)
barplot(hourly_counts, main = "Hourly Accident Distribution", xlab = "Hour", ylab = "Count", col
= "lightcoral")
```



4.Accidents by Location Scatter Plot:

- The scatter plot shows the geographical distribution of accidents, highlighting potential hot-spots or areas with higher accident frequencies.

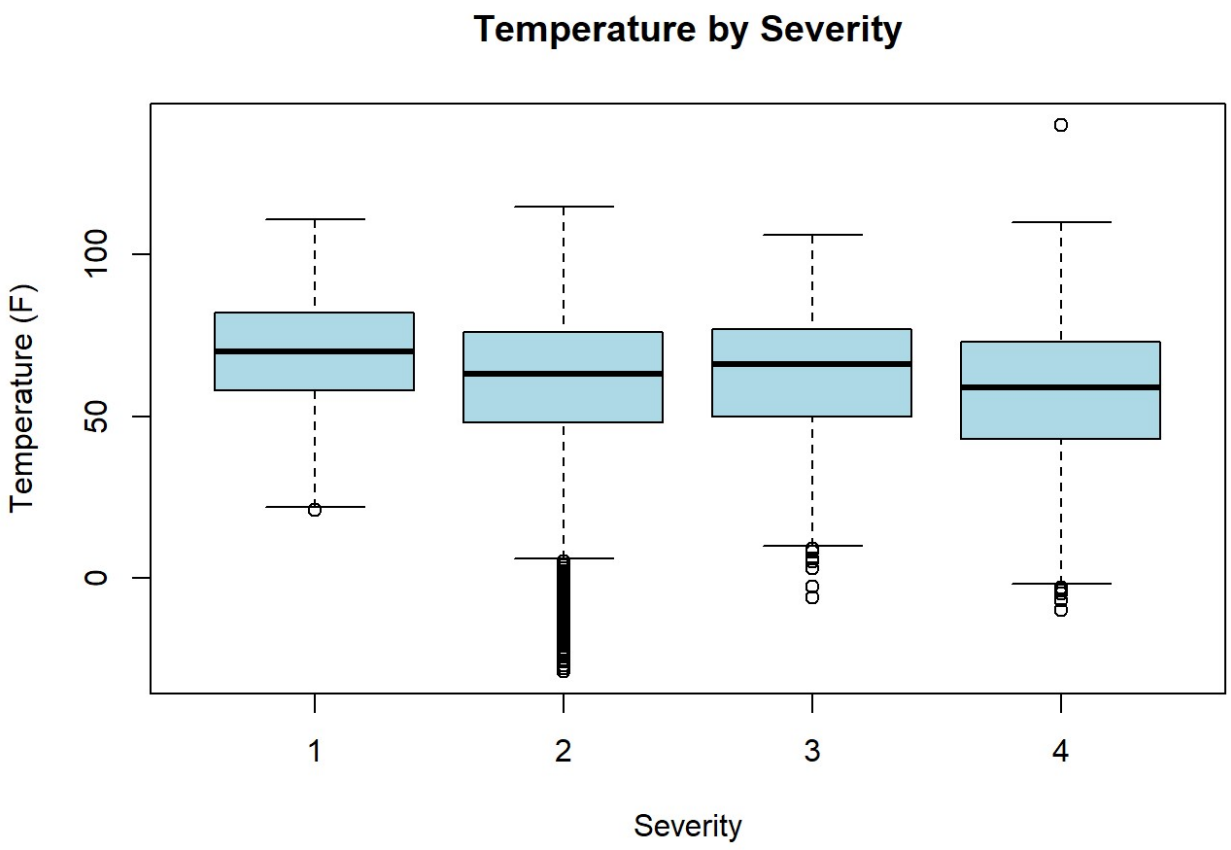
```
# Scatter plot of accidents based on latitude and longitude with transparency and smaller points
plot(USAcaraccidents$Start_Lng, USAcaraccidents$Start_Lat, col = rgb(0, 0, 1, 0.1), pch = 16, ma
in = "Accidents by Location", xlab = "Longitude", ylab = "Latitude")
```



5.Temperature by Severity Boxplot:

The boxplot suggests that there is a variation in temperature for different severity levels, with severity 3 accidents showing a wider range of temperatures.

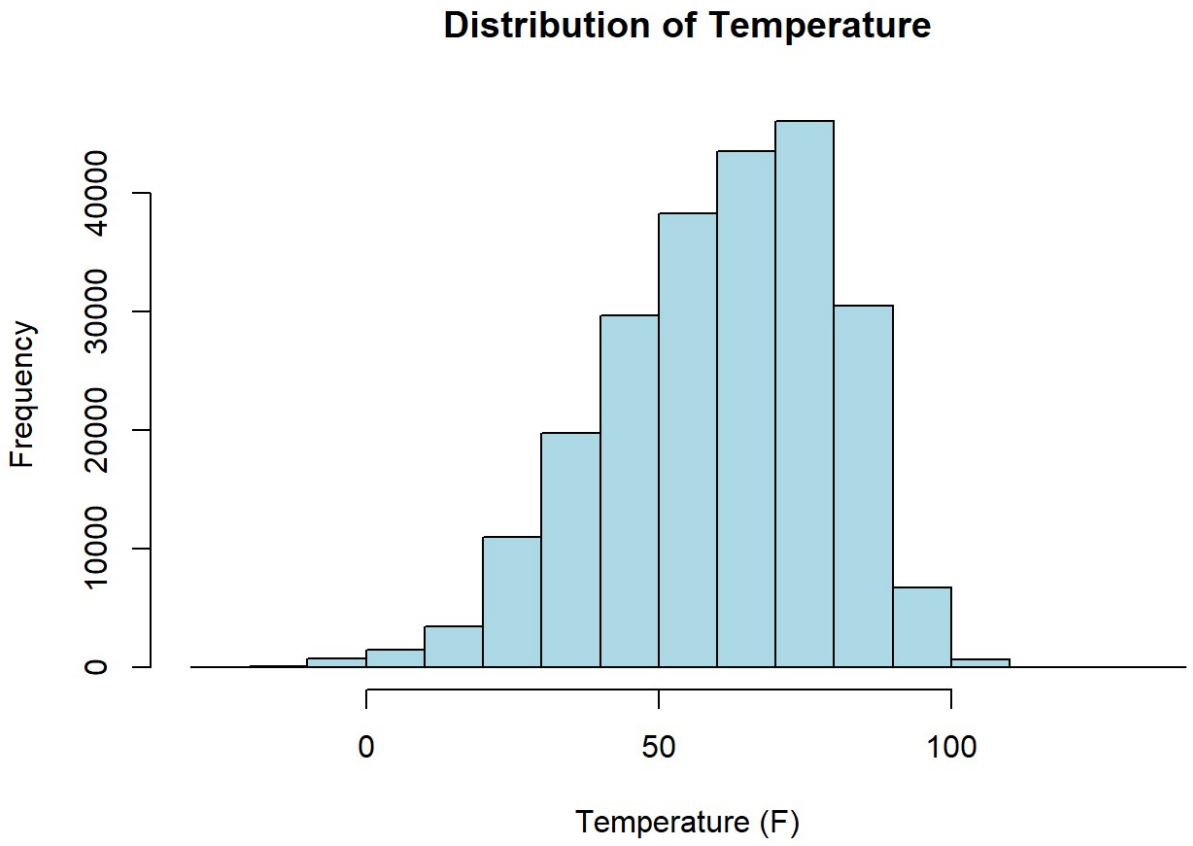
```
boxplot(Temperature.F. ~ Severity, data = USAcaraccidents, main = "Temperature by Severity", xla
b = "Severity", ylab = "Temperature (F)", col = "lightblue")
```



7.Temperature Distribution Histogram:

The histogram provides a clear overview of the temperature distribution in accidents, helping identify common temperature ranges.

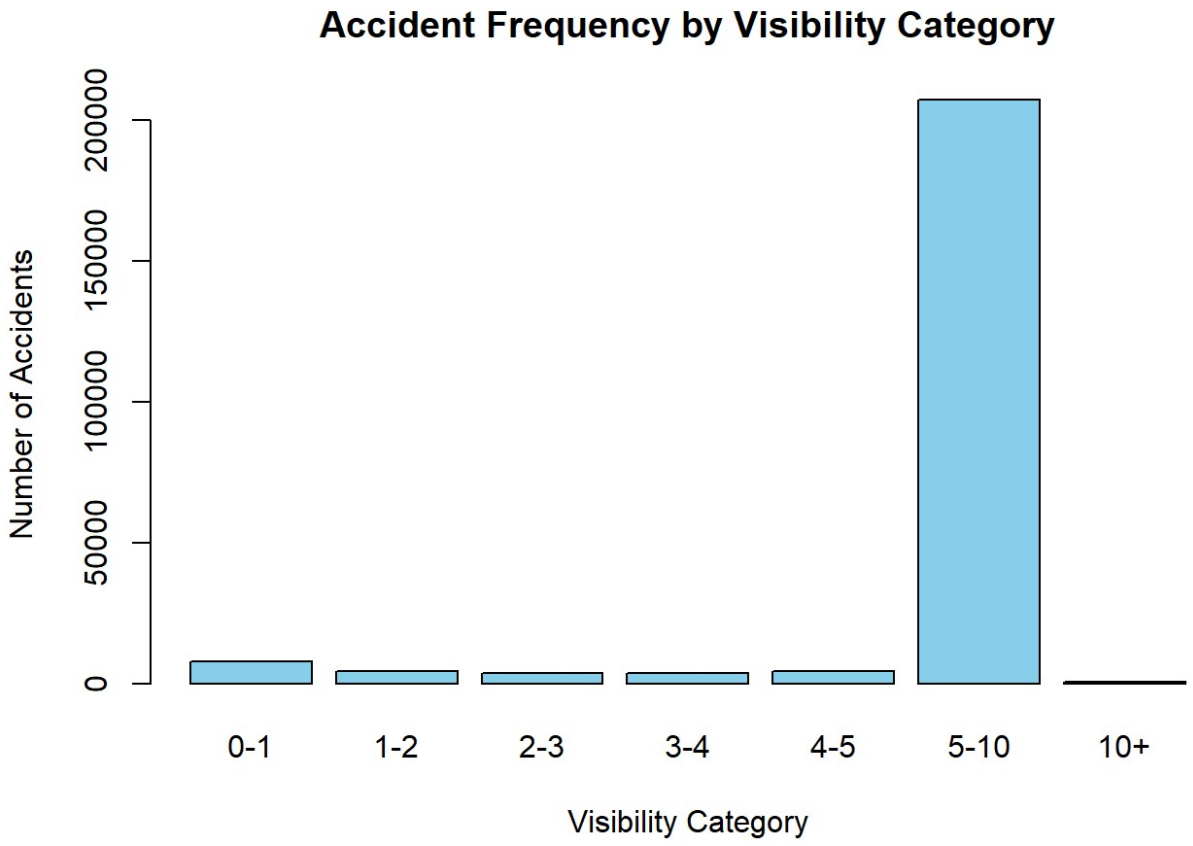
```
# Histogram for Temperature
hist(USAcaraccidents$Temperature.F., main = "Distribution of Temperature", xlab = "Temperature
(F)", col = "lightblue")
```



8.Visibility Category Bar Plot:

Accidents are more frequent in higher visibility conditions, with the majority falling within the 5-10 miles visibility range.

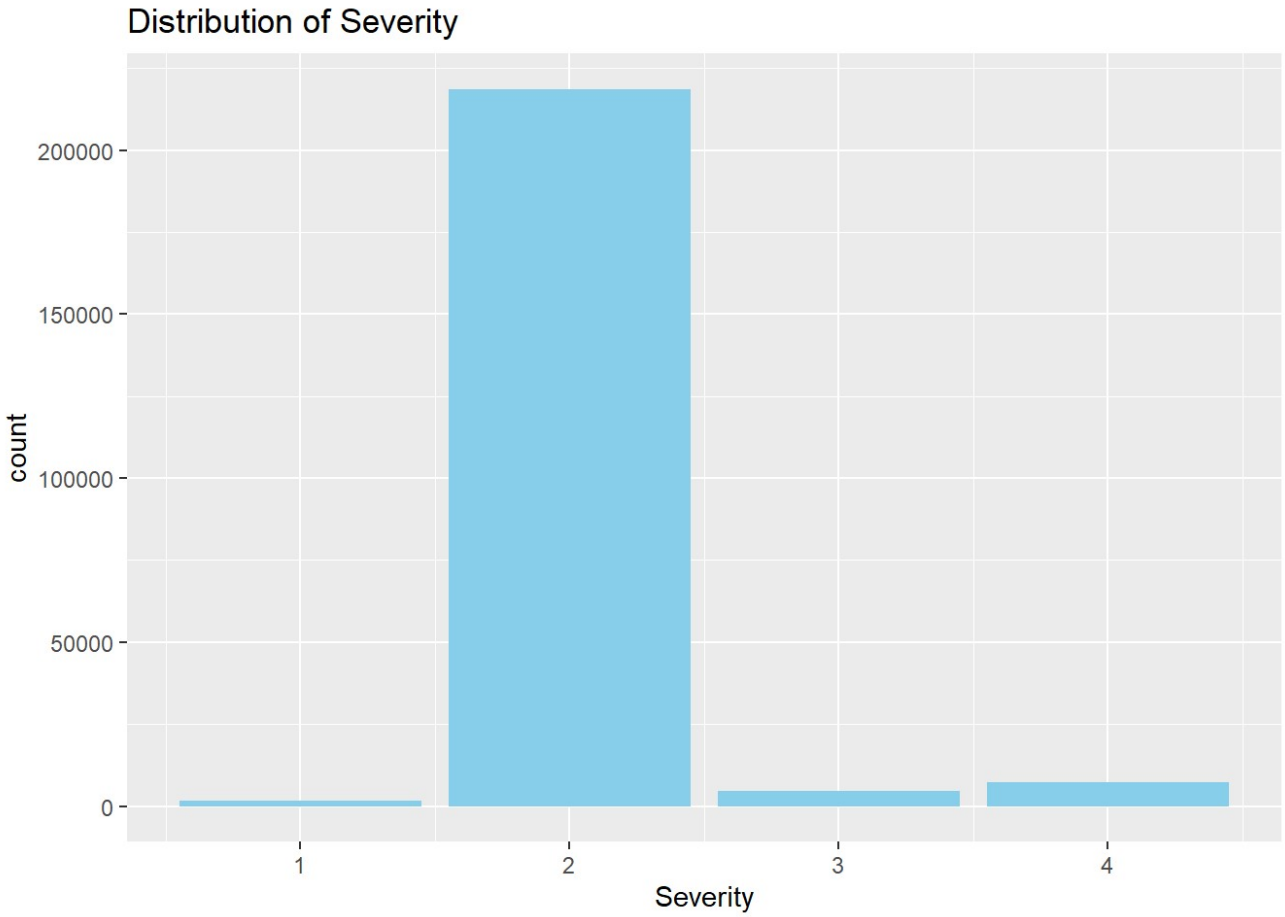
```
# Create Visibility Categories
USAcaraccidents$Visibility_Category <- cut(USAcaraccidents$Visibility.mi., breaks = c(0, 1, 2,
3, 4, 5, 10, Inf), labels = c("0-1", "1-2", "2-3", "3-4", "4-5", "5-10", "10+"))

# Bar Plot for Visibility Categories
barplot(table(USAcaraccidents$Visibility_Category),
        main = "Accident Frequency by Visibility Category",
        xlab = "Visibility Category",
        ylab = "Number of Accidents",
        col = "skyblue")
```



9.Severity Distribution Bar Plot (ggplot2):

Using ggplot2, the bar plot reiterates the distribution of accident severity, providing an alternative visualization.

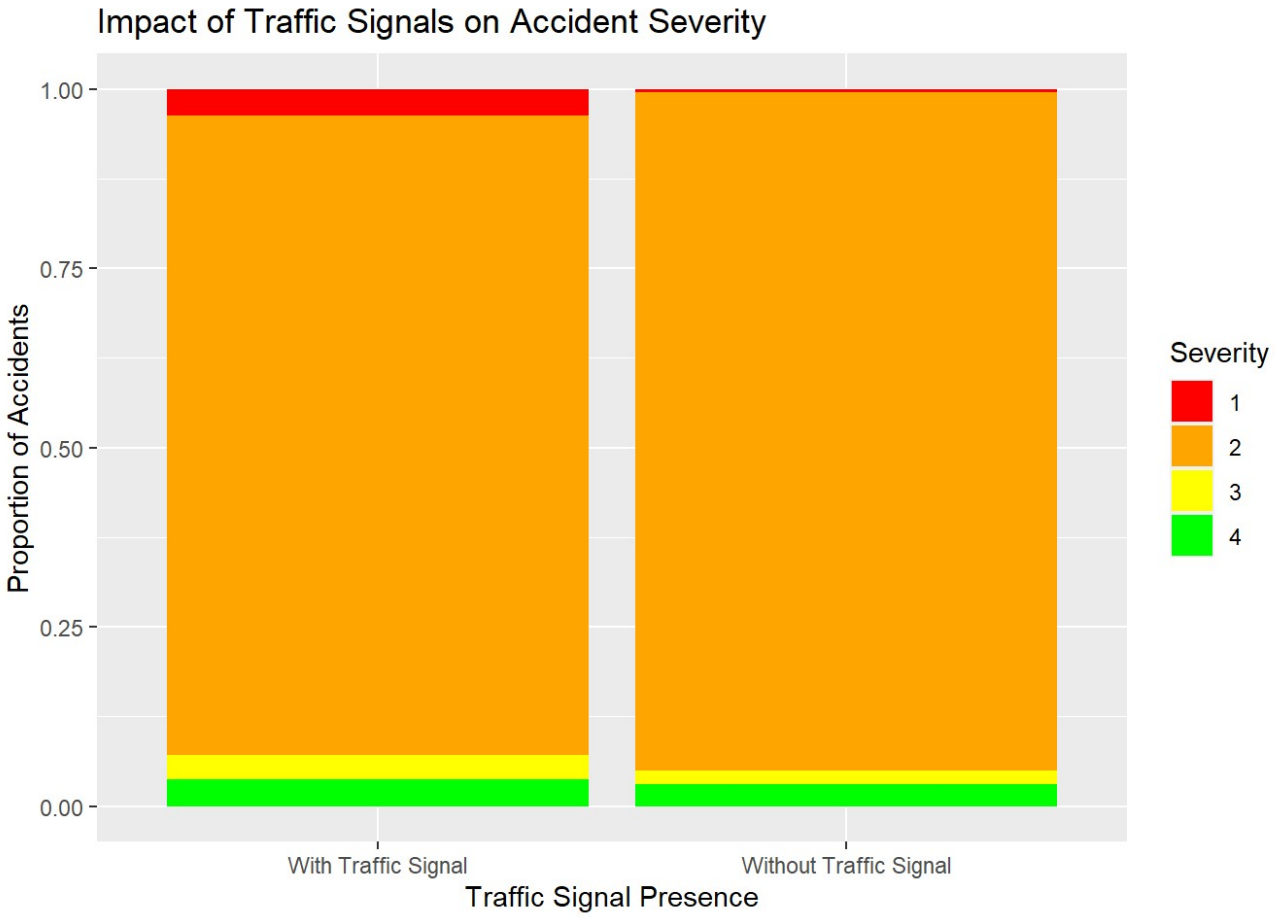
```
ggplot(data = USAcaraccidents, aes(x = Severity)) +
  geom_bar(fill = "skyblue") +
  ggtitle("Distribution of Severity")
```



10.Impact of Traffic Signals on Severity Stacked Bar Plot:

The stacked bar plot indicates that the presence of traffic signals has a marginal impact on accident severity, with severity 2 being slightly higher at locations with traffic signals.

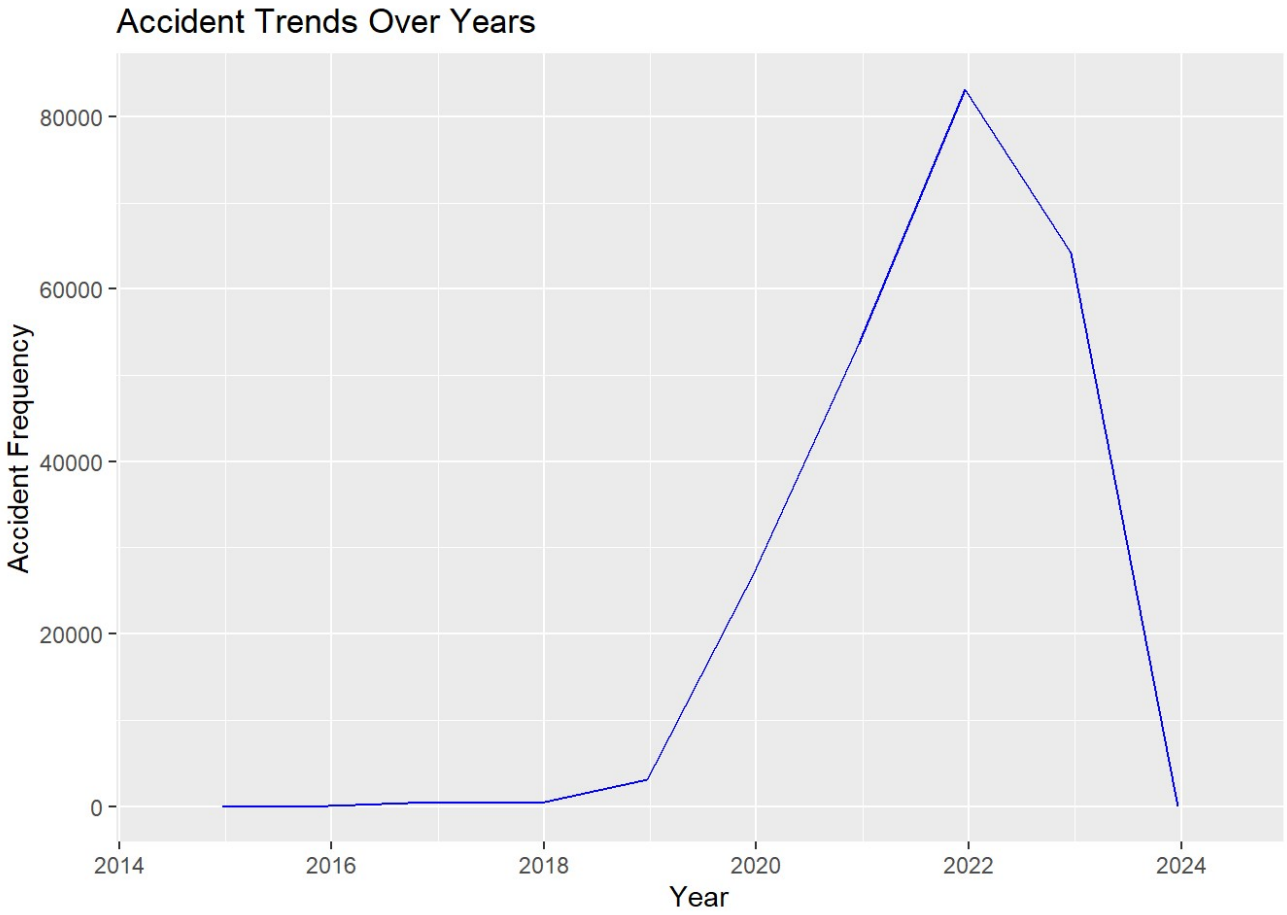
```
# Assuming 'Traffic_Signal' is the column indicating the presence of traffic signals
# Assuming 'Severity' is the column indicating the severity of accidents

# Convert 'Severity' to factor
USAcaraccidents$Severity <- as.factor(USAcaraccidents$Severity)

# Create a new column to categorize locations with and without traffic signals
USAcaraccidents$Signal_Category <- ifelse(USAcaraccidents$Traffic_Signal, "With Traffic Signal",
"Without Traffic Signal")

# Create a bar plot to compare severity distribution
ggplot(data = USAcaraccidents, aes(x = Signal_Category, fill = Severity)) +
  geom_bar(position = "fill", show.legend = TRUE) +
  ggtitle("Impact of Traffic Signals on Accident Severity") +
  xlab("Traffic Signal Presence") +
  ylab("Proportion of Accidents") +
  scale_fill_manual(values = c("1" = "red", "2" = "orange", "3" = "yellow", "4" = "green"))
```



11.Accident Trends Over Years Time Series Plot:

- The time series plot illustrates the overall trend in accident frequency over the years, revealing any notable patterns or changes.

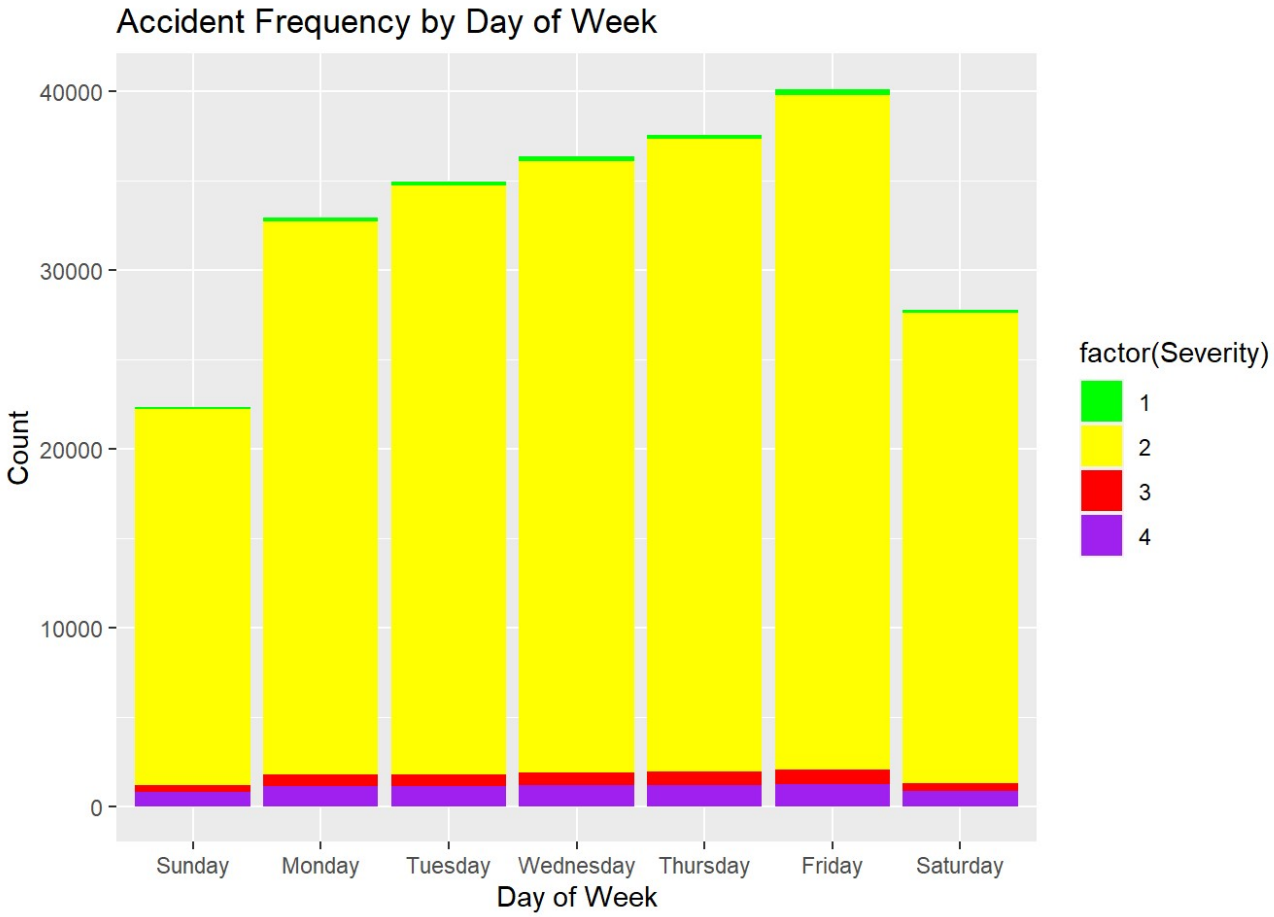
```
# Assuming 'Start_Time' is the column indicating the start time of accidents
# Convert 'Start_Time' to POSIXct format
USAcaraccidents$Start_Time <- as.POSIXct(USAcaraccidents$Start_Time, format = "%Y-%m-%d %H:%M:%
S", tz = "UTC")

# Extract year and month for time series analysis
USAcaraccidents$YearMonth <- format(USAcaraccidents$Start_Time, "%Y-%m")

# Time series plot for accident frequency over years
ggplot(data = USAcaraccidents, aes(x = Start_Time)) +
  geom_freqpoly(binwidth = 60*60*24*365, color = "blue") +
  ggtitle("Accident Trends Over Years") +
  xlab("Year") +
  ylab("Accident Frequency")
```



12.Accident Frequency by Day of Week Stacked Bar Plot:

- Accidents are fairly evenly distributed throughout the week, with a slight increase during weekdays compared to weekends.

```
ggplot(USAcaraccidents, aes(x = factor(wday(Start_Time, label = TRUE, abbr = FALSE)), fill = fac
tor(Severity))) +
  geom_bar(position = "stack") +
  ggtitle("Accident Frequency by Day of Week") +
  xlab("Day of Week") +
  ylab("Count") +
  scale_fill_manual(values = c("1" = "green", "2" = "yellow", "3" = "red", "4" = "purple"))
```



PART III: Hypothesis Testing

```
severity <- USAcaraccidents$Severity
```

```
temperature <- USAcaraccidents$Temperature.F.
```

Null Hypothesis - There is no significant difference in accident severity between tempratures above 50 and temprature 50 or below.
Alternate Hypothesis - There is a significant difference in accident severity between tempratures above 50 and temprature 50 or below.

```
severity_above_threshold <- severity[temperature > 50]
severity_below_threshold <- severity[temperature <= 50]
```

```
severity_above_threshold <- as.numeric(severity_above_threshold)
severity_below_threshold <- as.numeric(severity_below_threshold)
t_test_result <- t.test(severity_above_threshold, severity_below_threshold)
t_test_result
```

```
##
##  Welch Two Sample t-test
##
## data:  severity_above_threshold and severity_below_threshold
## t = -15.109, df = 110141, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.03166971 -0.02439652
## sample estimates:
## mean of x mean of y
##  2.067796  2.095829
```

```
alpha <- 0.05
if (t_test_result$p.value < alpha) {
  print("Reject the null hypothesis. There is a significant difference in accident severity base
d on temperature.")
} else {
  print("Fail to reject the null hypothesis. No significant difference in accident severity base
d on temperature.")
}
```

```
## [1] "Reject the null hypothesis. There is a significant difference in accident severity based
on temperature."
```

```
day_night <- USAcaraccidents$Sunrise_Sunset
```

Null Hypothesis (H0) - There is no significant difference in accident severity between day and night.
Alternate Hypothesis (H1) - There is a significant difference in accident severity between day and night.

```
severity_day <- severity[day_night == "Day"]
severity_night <- severity[day_night == "Night"]
severity_day <- as.numeric(severity_day)
severity_night <- as.numeric(severity_night)
t_test_result <- t.test(severity_day, severity_night)
t_test_result
```

```
##
##  Welch Two Sample t-test
##
## data:  severity_day and severity_night
## t = -8.9268, df = 156459, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.01851469 -0.01184817
## sample estimates:
## mean of x mean of y
##  2.069660  2.084842
```

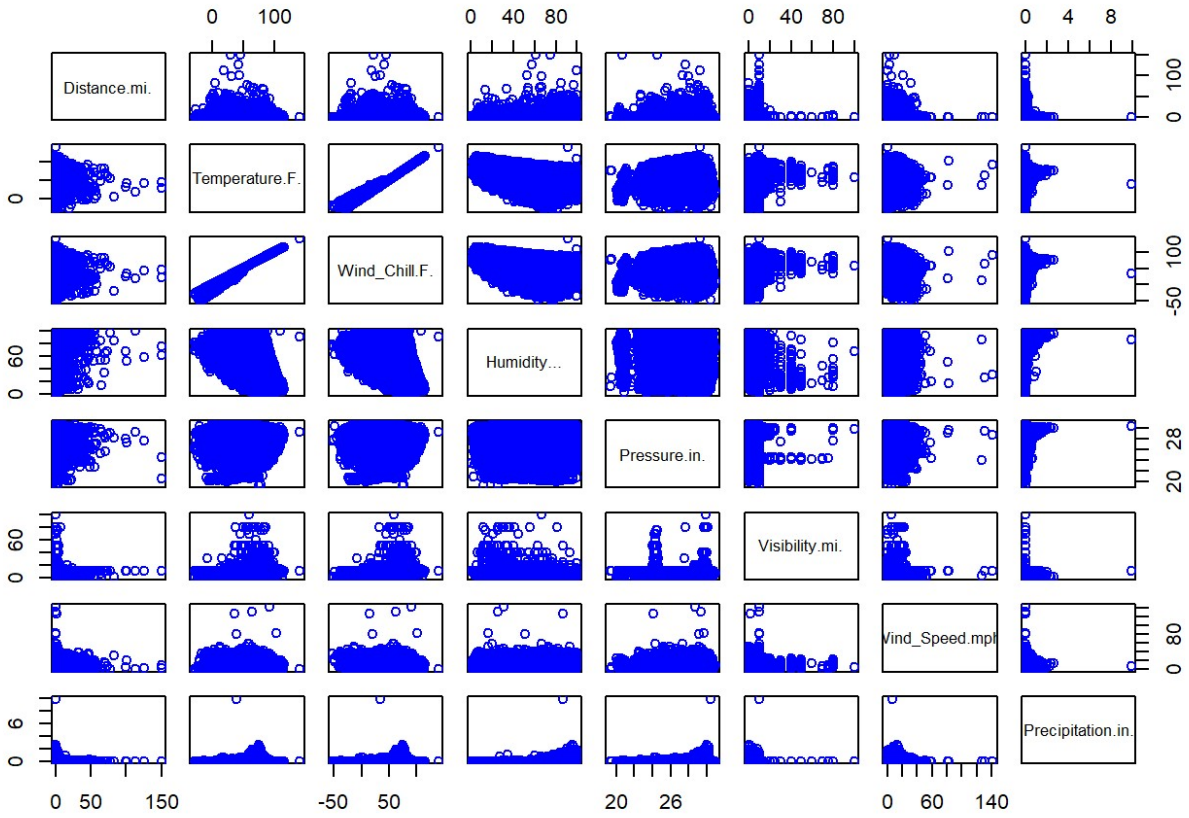
```
alpha <- 0.05
if (t_test_result$p.value < alpha) {
  print("Reject the null hypothesis. There is a significant difference in accident severity betw
een day and night.")
} else {
  print("Fail to reject the null hypothesis. No significant difference in accident severity betw
een day and night.")
}
```

```
## [1] "Reject the null hypothesis. There is a significant difference in accident severity betwe
en day and night."
```

PART IV: Linear Regression Modals

```
#Creating a variable containing only continuous numerical variables
selected_columns <- c('Distance.mi.', 'Temperature.F.', 'Wind_Chill.F.','Humidity...', 'Pressur
e.in.','Visibility.mi.', 'Wind_Speed.mph.', 'Precipitation.in.')
```

```
pairs(USAcaraccidents[selected_columns], col ="blue")
```




```
# Calculate the correlation matrix for numeric variables
correlation_matrix <- cor(USAcaraccidents[selected_columns])
# Print the correlation matrix
print(correlation_matrix)
```

```
##                  Distance.mi. Temperature.F. Wind_Chill.F. Humidity...
## Distance.mi.       1.000000000   -0.063665668  -0.069528338  0.02208069
## Temperature.F.    -0.063665668    1.000000000   0.993449333 -0.36132171
## Wind_Chill.F.     -0.069528338    0.993449333   1.000000000 -0.34481228
## Humidity...        0.022080687   -0.361321707  -0.344812276  1.00000000
## Pressure.in.      -0.091561285    0.197974464   0.208547913  0.11298681
## Visibility.mi.    -0.063329867    0.275930433   0.282055005 -0.39621954
## Wind_Speed.mph.    0.021450572    0.057124308  -0.001174361 -0.17689612
## Precipitation.in.  0.008724501   -0.009672783  -0.009378229  0.14211055
##                  Pressure.in. Visibility.mi. Wind_Speed.mph. Precipitation.in.
## Distance.mi.      -0.091561285  -0.0633298668    0.0214505724       0.008724501
## Temperature.F.     0.197974464   0.2759304333    0.0571243084      -0.009672783
## Wind_Chill.F.      0.208547913   0.2820550048   -0.0011743613      -0.009378229
## Humidity...        0.112986813  -0.3962195440   -0.1768961160       0.142110551
## Pressure.in.       1.000000000   0.0671552680   -0.0585178681       0.003661231
## Visibility.mi.     0.067155268   1.0000000000    0.0009621704      -0.217707402
## Wind_Speed.mph.   -0.058517868   0.0009621704    1.0000000000       0.038456916
## Precipitation.in.  0.003661231  -0.2177074023    0.0384569160       1.000000000
```

We can observe that Temperature, Wind Chill, Pressure and Visibility are inversely related to Distance.
Pressure has the highest correlation with Distance but that does not make sense. In the correlation table we observe that pressure has a positive relation with temperature, which doesn't seem possible. So we can mark our observations here and continue with other

Let's create a simple Linear regression model

```
Numerical_accident_data <- USAcaraccidents[selected_columns]
summary(Numerical_accident_data)
```

```
##   Distance.mi.      Temperature.F.   Wind_Chill.F.     Humidity...
##  Min.   :  0.0000   Min.   :-29.00   Min.   :-52.00   Min.   :  1.00
##  1st Qu.:  0.0670   1st Qu.: 48.00   1st Qu.: 46.00   1st Qu.: 47.00
##  Median :  0.2660   Median : 63.00   Median : 63.00   Median : 66.00
##  Mean   :  0.8579   Mean   : 61.03   Mean   : 59.69   Mean   : 63.78
##  3rd Qu.:  0.9270   3rd Qu.: 76.00   3rd Qu.: 76.00   3rd Qu.: 83.00
##  Max.   :149.6900   Max.   :140.00   Max.   :140.00   Max.   :100.00
##   Pressure.in.   Visibility.mi.    Wind_Speed.mph.  Precipitation.in.
##  Min.   :19.36   Min.   :  0.000   Min.   :  0.00   Min.   :0.000000
##  1st Qu.:29.18   1st Qu.: 10.000   1st Qu.:  3.00   1st Qu.:0.000000
##  Median :29.72   Median : 10.000   Median :  7.00   Median :0.000000
##  Mean   :29.35   Mean   :  9.054   Mean   :  7.44   Mean   :0.005643
##  3rd Qu.:29.96   3rd Qu.: 10.000   3rd Qu.: 10.00   3rd Qu.:0.000000
##  Max.   :30.95   Max.   :100.000   Max.   :142.00   Max.   :9.960000
```

```
# Splitting the dataset into 2
train_index = sample(2,nrow(Numerical_accident_data),replace=TRUE, prob = c(0.8,0.2))
Accident_Training <- Numerical_accident_data[train_index==1,]
Accident_Testing <- Numerical_accident_data[train_index==2,]
dim(Accident_Training)
```

```
## [1] 185892      8
```

```
dim(Accident_Testing)
```

```
## [1] 46238     8
```

```
#Creating a simple linear regression modal
lm_model1 <- lm(Distance.mi. ~ Precipitation.in., data=Accident_Training)
summary(lm_model1)
```

```
##
## Call:
## lm(formula = Distance.mi. ~ Precipitation.in., data = Accident_Training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -3.679  -0.788  -0.591   0.069 148.836
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.853699   0.004567 186.930  < 2e-16 ***
## Precipitation.in. 0.315566   0.094837   3.327 0.000877 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.955 on 185890 degrees of freedom
## Multiple R-squared:  5.956e-05,  Adjusted R-squared:  5.418e-05
## F-statistic: 11.07 on 1 and 185890 DF,  p-value: 0.0008766
```

Let’s answer a couple of question:

- Is there a relationship between the predictor and the response? => Yes
- How strong is the relationship between the predictor and the response? => p -value is close to 0: relationship is strong
- Is the relationship between the predictor and the response positive or negative? => Positive
- In summary, the model suggests that there is a statistically significant positive relationship between Precipitation.in. and Distance.mi., meaning that an increase in precipitation is associated with an increase in distance of accident.

```
# model 2
lm_model2 <- lm(Distance.mi. ~ Visibility.mi., data=Accident_Training)
summary(lm_model2)
```

```
##
## Call:
## lm(formula = Distance.mi. ~ Visibility.mi., data = Accident_Training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -1.261  -0.765  -0.573   0.072 148.877
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.260722   0.016554   76.16   <2e-16 ***
## Visibility.mi. -0.044758   0.001759  -25.45   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.952 on 185890 degrees of freedom
## Multiple R-squared:  0.003472,   Adjusted R-squared:  0.003466
## F-statistic: 647.6 on 1 and 185890 DF,  p-value: < 2.2e-16
```

- Is there a relationship between the predictor and the response? => Yes
- How strong is the relationship between the predictor and the response? => p -value is close to 0: relationship is strong
- Is the relationship between the predictor and the response positive or negative? => Negative
- In summary, this model suggests that visibility has a statistically significant effect on distance. This means that a decrease in visibility tends to lead to larger distance of the accident.

```
# model 3
lm_model3 <- lm(Distance.mi. ~ Wind_Chill.F., data=Accident_Training)
summary(lm_model3)
```

```
##
## Call:
## lm(formula = Distance.mi. ~ Wind_Chill.F., data = Accident_Training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -1.516  -0.757  -0.555   0.067 148.745
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.219766   0.013271   91.91   <2e-16 ***
## Wind_Chill.F. -0.006104   0.000209  -29.20   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.951 on 185890 degrees of freedom
## Multiple R-squared:  0.004565,   Adjusted R-squared:  0.00456
## F-statistic: 852.5 on 1 and 185890 DF,  p-value: < 2.2e-16
```

- In summary, the model suggests a statistically significant relationship between the predictor (Wind_Chill) and the response variable (Distance)

Let's summarize prediction data and calculate MAE (Mean Absolute Error) and MSE (Mean Squared Error) . MAE and MSE are both metrics commonly used to evaluate the performance of a regression model

```
predictions1 <- predict(lm_model1, newdata = Accident_Testing)
predictions2 <- predict(lm_model2, newdata = Accident_Testing)
predictions3 <- predict(lm_model3, newdata = Accident_Testing)
summary(predictions1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.8537  0.8537  0.8537  0.8554  0.8537  1.5669
```

```
summary(predictions2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -2.3199  0.8131  0.8131  0.8555  0.8131  1.2607
```

```
summary(predictions3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3652  0.7559  0.8352  0.8550  0.9390  1.5005
```

```
MAE(y_pred = predictions1, y_true = Accident_Testing$Distance.mi.)
```

```
## [1] 0.9323068
```

```
MAE(y_pred = predictions2, y_true = Accident_Testing$Distance.mi.)
```

```
## [1] 0.9304382
```

```
MAE(y_pred = predictions3, y_true = Accident_Testing$Distance.mi.)
```

```
## [1] 0.929364
```

```
MSE(y_pred = predictions1, y_true = Accident_Testing$Distance.mi.)
```

```
## [1] 3.58659
```

```
MSE(y_pred = predictions2, y_true = Accident_Testing$Distance.mi.)
```

```
## [1] 3.564744
```

```
MSE(y_pred = predictions3, y_true = Accident_Testing$Distance.mi.)
```

```
## [1] 3.565654
```

Multiple Linear Regression

```
multiplelm1 <- lm(Distance.mi.~., data=Accident_Training)
summary(multiplelm1)
```

```
##
## Call:
## lm(formula = Distance.mi. ~ ., data = Accident_Training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -2.809  -0.712  -0.534   0.079 148.250
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        5.0257343  0.1202422  41.797   <2e-16 ***
## Temperature.F.     0.0341590  0.0023856  14.319   <2e-16 ***
## Wind_Chill.F.     -0.0337529  0.0021313 -15.837   <2e-16 ***
## Humidity...        0.0006315  0.0002329   2.711   0.0067 **
## Pressure.in.      -0.1364543  0.0041400 -32.960   <2e-16 ***
## Visibility.mi.    -0.0296511  0.0019817 -14.963   <2e-16 ***
## Wind_Speed.mph.   -0.0009403  0.0009545  -0.985   0.3245
## Precipitation.in. -0.0585772  0.0970306  -0.604   0.5460
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.942 on 185884 degrees of freedom
## Multiple R-squared:  0.01395,    Adjusted R-squared:  0.01392
## F-statistic: 375.8 on 7 and 185884 DF,  p-value: < 2.2e-16
```

```
ypred <-predict(object = multiplelm1, newdata = Accident_Testing)
summary(ypred)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -1.4026  0.7102  0.7724  0.8563  0.9121  2.9183
```

```
MAE(y_pred = ypred, y_true = Accident_Testing$Distance.mi.)
```

```
## [1] 0.9250965
```

```
MSE(y_pred = ypred, y_true = Accident_Testing$Distance.mi.)
```

```
## [1] 3.524409
```

Forward Stepwise regression

```
intercept_only <- lm(Distance.mi. ~ 1, data= Accident_Training)
all <- lm(Distance.mi. ~., data=Accident_Training)
forward <- stepAIC (intercept_only, direction='forward',scope = formula(all))
```

```
## Start:  AIC=249273.3
## Distance.mi. ~ 1
##
##                     Df Sum of Sq    RSS    AIC
## + Pressure.in.       1    6018.4 704584 247694
## + Wind_Chill.F.      1    3244.0 707358 248425
## + Temperature.F.     1    2738.8 707863 248557
## + Visibility.mi.     1    2467.0 708135 248629
## + Humidity...        1     301.0 710301 249197
## + Wind_Speed.mph.    1     298.6 710304 249197
## + Precipitation.in.  1      42.3 710560 249264
## <none>                           710602 249273
##
## Step:  AIC=247694.2
## Distance.mi. ~ Pressure.in.
##
##                     Df Sum of Sq    RSS    AIC
## + Visibility.mi.     1   1996.64 702587 247169
## + Wind_Chill.F.      1   1745.60 702838 247235
## + Temperature.F.     1   1429.57 703154 247319
## + Humidity...        1    695.76 703888 247513
## + Wind_Speed.mph.    1    162.20 704422 247653
## + Precipitation.in.  1     46.93 704537 247684
## <none>                           704584 247694
##
## Step:  AIC=247168.7
## Distance.mi. ~ Pressure.in. + Visibility.mi.
##
##                     Df Sum of Sq    RSS    AIC
## + Wind_Chill.F.      1    945.58 701642 246920
## + Temperature.F.     1    721.93 701865 246980
## + Wind_Speed.mph.    1    165.32 702422 247127
## + Humidity...        1     81.03 702506 247149
## + Precipitation.in.  1      7.80 702579 247169
## <none>                           702587 247169
##
## Step:  AIC=246920.3
## Distance.mi. ~ Pressure.in. + Visibility.mi. + Wind_Chill.F.
##
##                     Df Sum of Sq    RSS    AIC
## + Temperature.F.     1    920.55 700721 246678
## + Wind_Speed.mph.    1    173.96 701468 246876
## <none>                           701642 246920
## + Precipitation.in.  1      1.47 701640 246922
## + Humidity...        1      0.11 701642 246922
##
## Step:  AIC=246678.3
## Distance.mi. ~ Pressure.in. + Visibility.mi. + Wind_Chill.F. +
##     Temperature.F.
##
##                     Df Sum of Sq    RSS    AIC
## + Humidity...        1   29.6259 700691 246672
```

```
## <none>                            700721 246678
## + Wind_Speed.mph.    1    6.7115 700714 246679
## + Precipitation.in.  1    0.7588 700720 246680
##
## Step:  AIC=246672.4
## Distance.mi. ~ Pressure.in. + Visibility.mi. + Wind_Chill.F. +
##     Temperature.F. + Humidity...
##
##                     Df Sum of Sq    RSS    AIC
## <none>                            700691 246672
## + Wind_Speed.mph.    1    3.9595 700688 246673
## + Precipitation.in.  1    1.6745 700690 246674
```

```
summary(forward)
```

```
##
## Call:
## lm(formula = Distance.mi. ~ Pressure.in. + Visibility.mi. + Wind_Chill.F. +
##     Temperature.F. + Humidity..., data = Accident_Training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -2.813  -0.712  -0.534   0.079 148.251
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     5.0241823  0.1202128  41.794  < 2e-16 ***
## Pressure.in.   -0.1364435  0.0041398 -32.959  < 2e-16 ***
## Visibility.mi. -0.0293990  0.0019486 -15.087  < 2e-16 ***
## Wind_Chill.F.  -0.0327255  0.0018546 -17.646  < 2e-16 ***
## Temperature.F.  0.0330010  0.0020787  15.876  < 2e-16 ***
## Humidity...     0.0006466  0.0002306   2.803  0.00506 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.942 on 185886 degrees of freedom
## Multiple R-squared:  0.01395,    Adjusted R-squared:  0.01392
## F-statistic: 525.8 on 5 and 185886 DF,  p-value: < 2.2e-16
```

In summary, the final iteration model includes predictors Pressure.in., Visibility.mi., Wind_Chill.F., Temperature.F., and Humidity…, and it has the lowest AIC among the considered models. This model is chosen as it strikes a balance between model complexity and goodness of fit. It can be noted that the improvement made by adding humidity had a very small improvement.

```
ypred_forward <-predict(object = forward, newdata = Accident_Testing)
MAE(y_pred = ypred_forward, y_true = Accident_Testing$Distance.mi.)
```

```
## [1] 0.9251014
```

```
MSE(y_pred = ypred_forward, y_true = Accident_Testing$Distance.mi.)
```

```
## [1] 3.524496
```

Backward Stepwise Regression

```
backward <- stepAIC (all, direction='backward')
```

```
## Start:  AIC=246675
## Distance.mi. ~ Temperature.F. + Wind_Chill.F. + Humidity... +
##     Pressure.in. + Visibility.mi. + Wind_Speed.mph. + Precipitation.in.
##
##                     Df Sum of Sq    RSS    AIC
## - Precipitation.in.  1       1.4 700688 246673
## - Wind_Speed.mph.    1       3.7 700690 246674
## <none>                           700686 246675
## - Humidity...        1      27.7 700714 246680
## - Temperature.F.     1     772.8 701459 246878
## - Visibility.mi.     1     843.9 701530 246897
## - Wind_Chill.F.      1     945.4 701632 246924
## - Pressure.in.       1    4095.0 704781 247756
##
## Step:  AIC=246673.4
## Distance.mi. ~ Temperature.F. + Wind_Chill.F. + Humidity... +
##     Pressure.in. + Visibility.mi. + Wind_Speed.mph.
##
##                   Df Sum of Sq    RSS    AIC
## - Wind_Speed.mph.  1       4.0 700691 246672
## <none>                         700688 246673
## - Humidity...      1      26.9 700714 246679
## - Temperature.F.   1     775.2 701463 246877
## - Visibility.mi.   1     859.8 701547 246899
## - Wind_Chill.F.    1     949.2 701637 246923
## - Pressure.in.     1    4094.0 704782 247754
##
## Step:  AIC=246672.4
## Distance.mi. ~ Temperature.F. + Wind_Chill.F. + Humidity... +
##     Pressure.in. + Visibility.mi.
##
##                  Df Sum of Sq    RSS    AIC
## <none>                        700691 246672
## - Humidity...     1      29.6 700721 246678
## - Visibility.mi.  1     858.0 701549 246898
## - Temperature.F.  1     950.1 701642 246922
## - Wind_Chill.F.   1    1173.7 701865 246982
## - Pressure.in.    1    4094.7 704786 247754
```

```
summary(backward)
```

```
##
## Call:
## lm(formula = Distance.mi. ~ Temperature.F. + Wind_Chill.F. +
##     Humidity... + Pressure.in. + Visibility.mi., data = Accident_Training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -2.813  -0.712  -0.534   0.079 148.251
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     5.0241823  0.1202128  41.794  < 2e-16 ***
## Temperature.F.  0.0330010  0.0020787  15.876  < 2e-16 ***
## Wind_Chill.F.  -0.0327255  0.0018546 -17.646  < 2e-16 ***
## Humidity...     0.0006466  0.0002306   2.803  0.00506 **
## Pressure.in.   -0.1364435  0.0041398 -32.959  < 2e-16 ***
## Visibility.mi. -0.0293990  0.0019486 -15.087  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.942 on 185886 degrees of freedom
## Multiple R-squared:  0.01395,    Adjusted R-squared:  0.01392
## F-statistic: 525.8 on 5 and 185886 DF,  p-value: < 2.2e-16
```

> The output then shows the stepwise elimination of variables based on the AIC.The first step eliminates the variable “Precipitation.in.”, the second step eliminates “Wind_Speed.mph.”

```
ypred_backward <-predict(object = backward, newdata = Accident_Testing)
MAE(y_pred = ypred_backward, y_true = Accident_Testing$Distance.mi.)
```

```
## [1] 0.9251014
```

```
MSE(y_pred = ypred_backward, y_true = Accident_Testing$Distance.mi.)
```

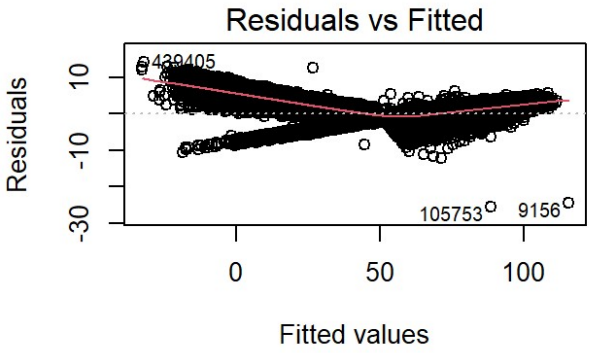
```
## [1] 3.524496
```

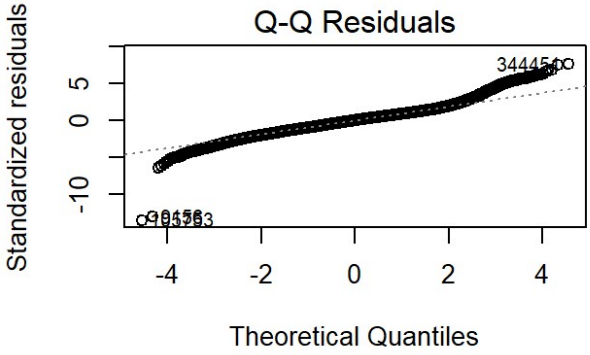
What happens if we don’t include Distance at all? Let’s instead try to use Temperature as the response variable.

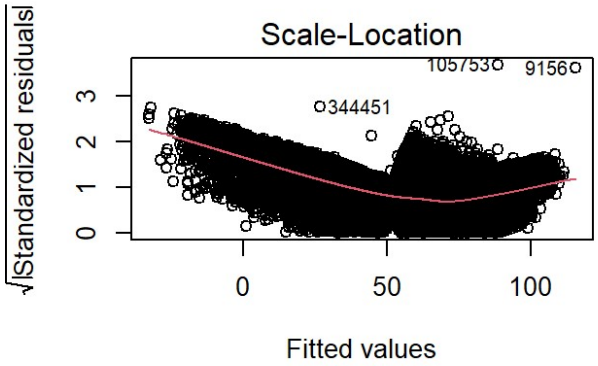
```
multiplelm_withoutdistance <- lm(Temperature.F.~., data=Accident_Training[,2:8])
summary(multiplelm_withoutdistance)
```

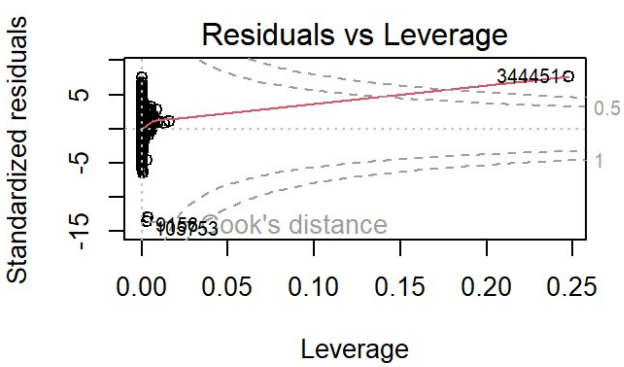
```
##
## Call:
## lm(formula = Temperature.F. ~ ., data = Accident_Training[, 2:8])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.5410  -1.2091   0.0555   1.1867  14.1411
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        9.7262378  0.1147065   84.79   <2e-16 ***
## Wind_Chill.F.      0.8880318  0.0002264 3921.60   <2e-16 ***
## Humidity...       -0.0101559  0.0002252  -45.09   <2e-16 ***
## Pressure.in.      -0.0640550  0.0040223  -15.93   <2e-16 ***
## Visibility.mi.    -0.0703084  0.0019197  -36.62   <2e-16 ***
## Wind_Speed.mph.    0.1963141  0.0008086  242.79   <2e-16 ***
## Precipitation.in. -1.1269681  0.0943004  -11.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.888 on 185885 degrees of freedom
## Multiple R-squared:  0.9905, Adjusted R-squared:  0.9905
## F-statistic: 3.228e+06 on 6 and 185885 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(multiplelm_withoutdistance)
```


Residuals vs Fitted
Residuals
Fitted values
105753
9156


Q-Q Residuals
Standardized residuals
Theoretical Quantiles


Scale-Location
√|Standardized residuals|
Fitted values
105753
9156
344451


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
344451
105753