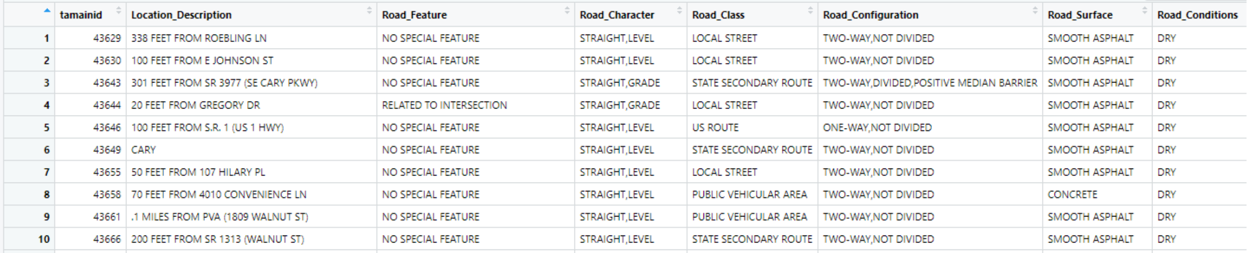
Data Analysis on Crash Data Collected from the Town of Cary in North Carolina

Halima Sadiq

Facts and information about the dataset:

- Source: https://data.townofcary.org/explore/dataset/cpd-crash-incidents
- 26,476 records of five years from 01/01/2016 to 12/02/2021 (as of 12/05/2021)
- Data is constantly updated to include new entries and remove errors
- 47 variables ranging from crash location, time, vehicles involved, fatalities, injuries, road conditions, weather conditions and more

First few rows of dataset showing some columns and row values:



	tamainid	Location_Description	Road_Feature	Road_Character	Road_Class	Road_Configuration	Road_Surface	Road_Conditions
1	43629	338 FEET FROM ROEBLING LN	NO SPECIAL FEATURE	STRAIGHT,LEVEL	LOCAL STREET	TWO-WAY,NOT DIVIDED	SMOOTH ASPHALT	DRY
2	43630	100 FEET FROM E JOHNSON ST	NO SPECIAL FEATURE	STRAIGHT,LEVEL	LOCAL STREET	TWO-WAY,NOT DIVIDED	SMOOTH ASPHALT	DRY
3	43643	301 FEET FROM SR 3977 (SE CARY PKWY)	NO SPECIAL FEATURE	STRAIGHT,GRADE	STATE SECONDARY ROUTE	TWO-WAY,DIVIDED,POSITIVE MEDIAN BARRIER	SMOOTH ASPHALT	DRY
4	43644	20 FEET FROM GREGORY DR	RELATED TO INTERSECTION	STRAIGHT,GRADE	LOCAL STREET	TWO-WAY,NOT DIVIDED	SMOOTH ASPHALT	DRY
5	43646	100 FEET FROM S.R. 1 (US 1 HWY)	NO SPECIAL FEATURE	STRAIGHT,LEVEL	US ROUTE	ONE-WAY,NOT DIVIDED	SMOOTH ASPHALT	DRY
6	43649	CARY	NO SPECIAL FEATURE	STRAIGHT,LEVEL	STATE SECONDARY ROUTE	TWO-WAY,NOT DIVIDED	SMOOTH ASPHALT	DRY
7	43655	50 FEET FROM 107 HILARY PL	NO SPECIAL FEATURE	STRAIGHT,LEVEL	LOCAL STREET	TWO-WAY,NOT DIVIDED	SMOOTH ASPHALT	DRY
8	43658	70 FEET FROM 4010 CONVENIENCE LN	NO SPECIAL FEATURE	STRAIGHT,LEVEL	PUBLIC VEHICULAR AREA	TWO-WAY,NOT DIVIDED	CONCRETE	DRY
9	43661	.1 MILES FROM PVA (1809 WALNUT ST)	NO SPECIAL FEATURE	STRAIGHT,LEVEL	PUBLIC VEHICULAR AREA	TWO-WAY,NOT DIVIDED	SMOOTH ASPHALT	DRY
10	43666	200 FEET FROM SR 1313 (WALNUT ST)	NO SPECIAL FEATURE	STRAIGHT,LEVEL	STATE SECONDARY ROUTE	TWO-WAY,NOT DIVIDED	SMOOTH ASPHALT	DRY

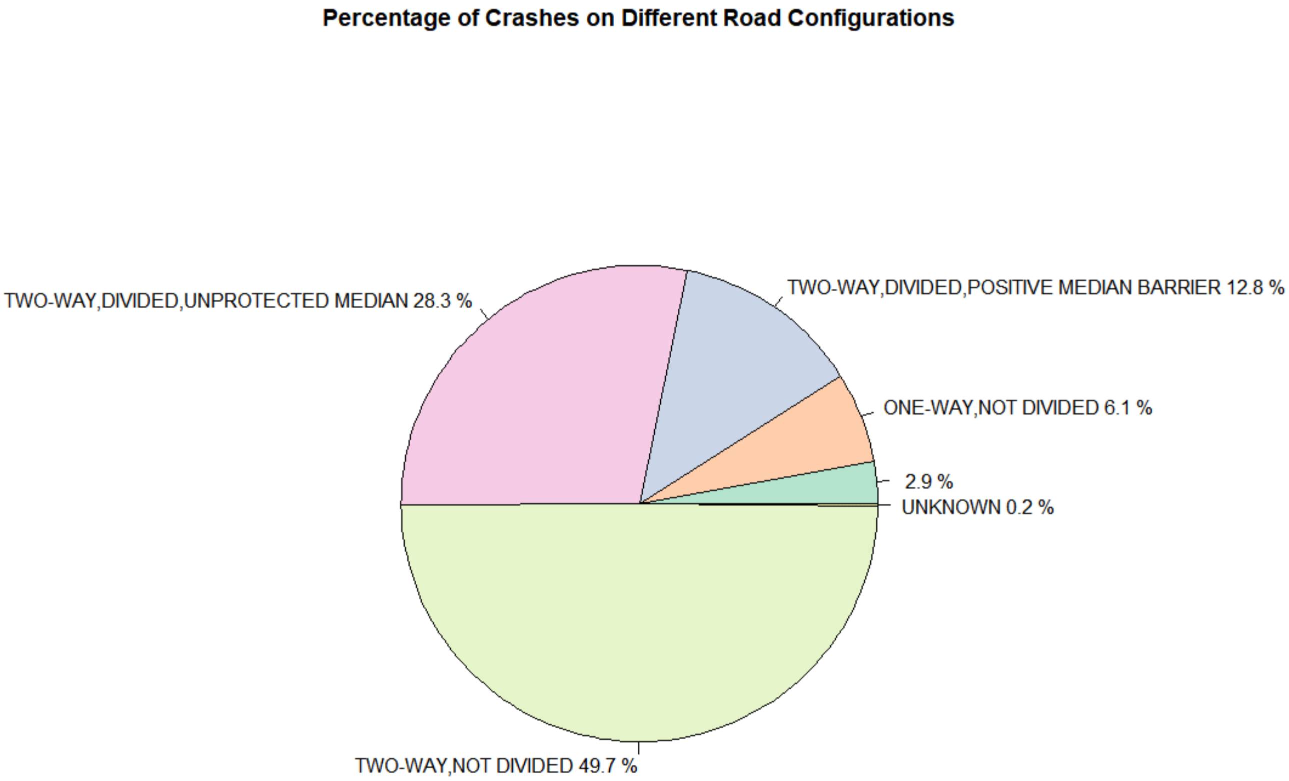
Techniques or topics used for data analysis:

- Data Visualization
- Hypothesis Testing
- Predictions

Learning more about the data through visualizations:

1. **Percentage of crashes at different road configurations:**

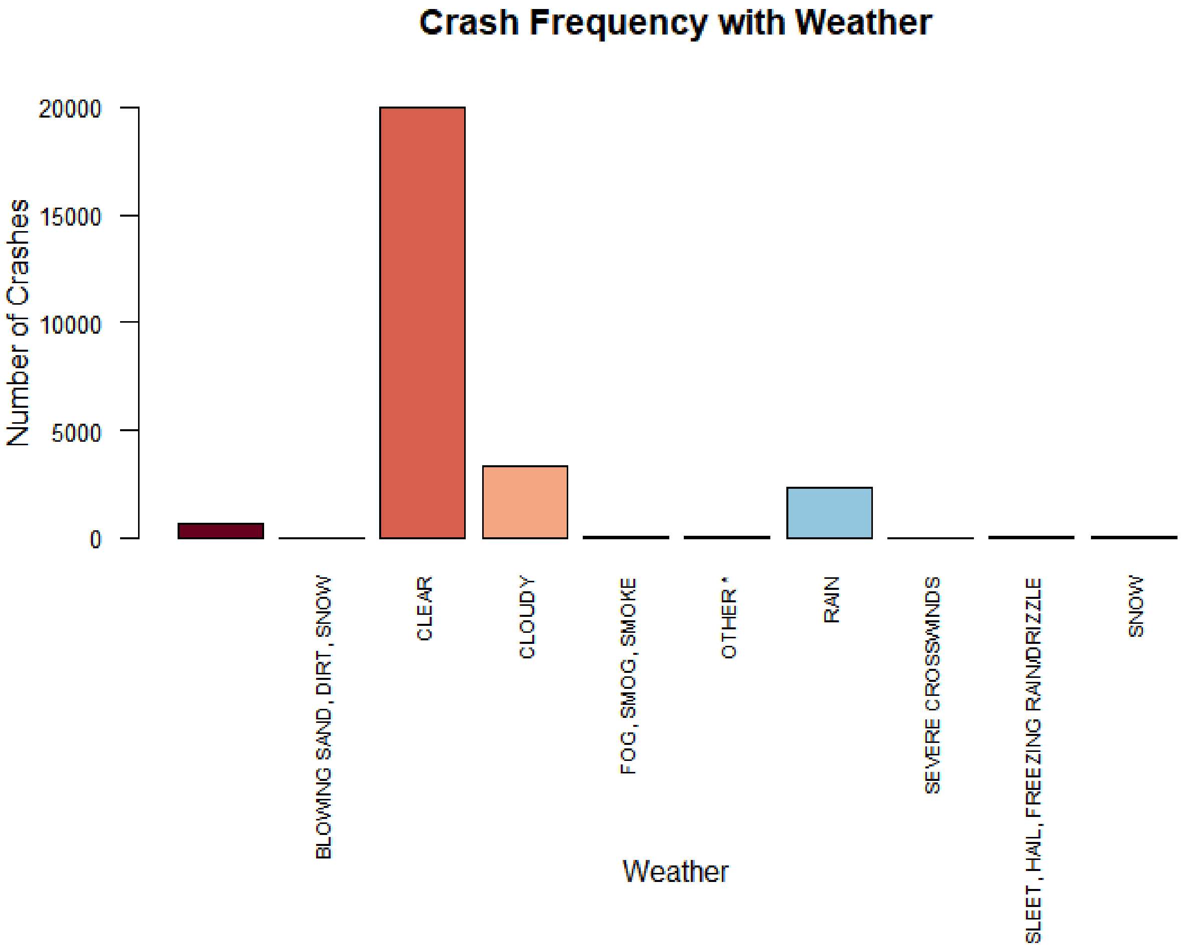
```
> x = data.frame(table(crashData[['Road_Configuration']]))
> piepercent = round(100*x$Freq/sum(x$Freq),1)
> pie(x$Freq, labels = paste(x$Var1, sep = ' ', piepercent, '%'), main = "Percentage of
Crashes on Different Road Configurations", col = brewer.pal(6,"Pastel2"))
```



The highest percentage of crashes occur at two-way, not divided roads followed by two-way divided roads with an unprotected median. For drivers, this visualization may encourage them to drive more carefully on two-way not divided roads.

2. Impact of Weather on Crash Frequency (if any):

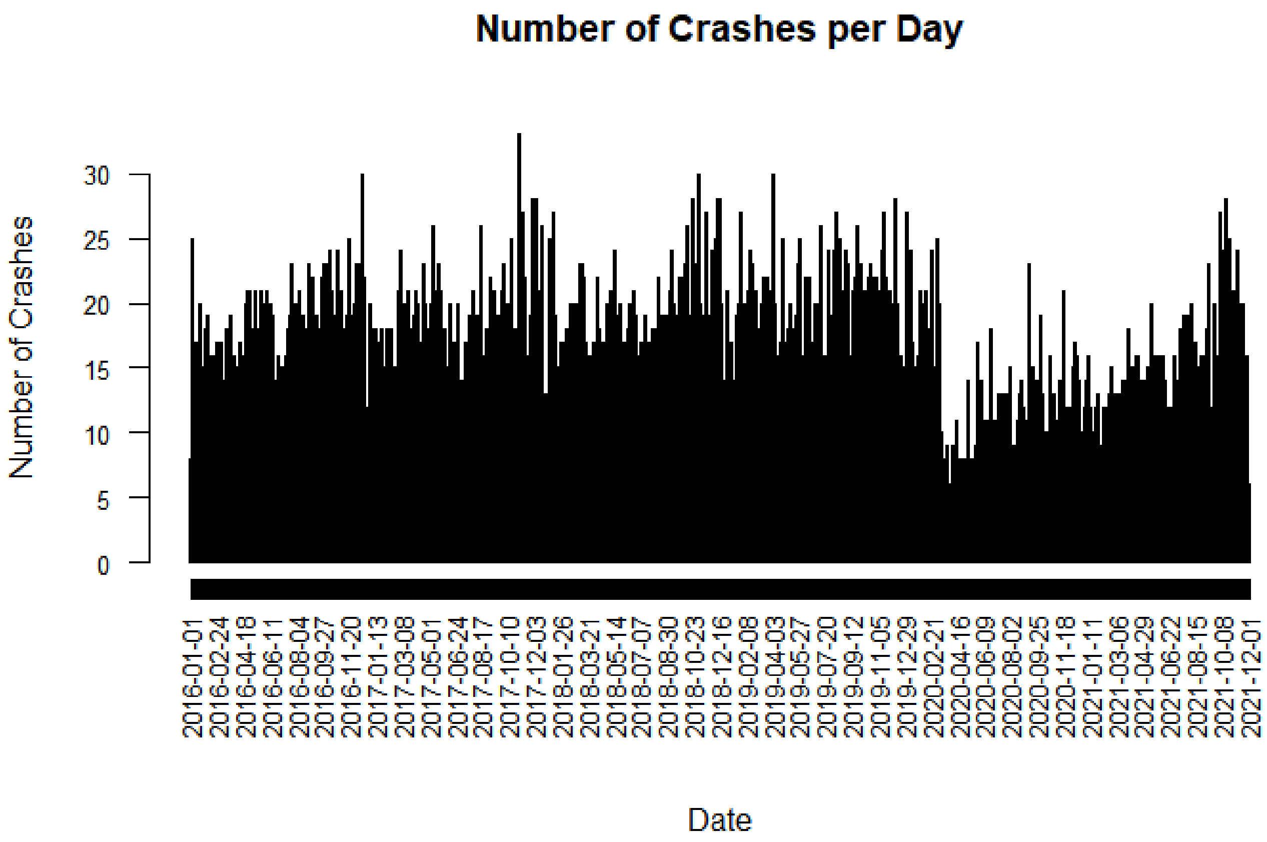
```
> w = table(crashData$Weather)
> extraMargin = par(mar = c(8,4,4,2) + 0.1)
> par(extraMargin)
> barplot(w,main = "Crash Frequency with Weather",xlab = " ",ylab="Number of
Crashes",las=2,cex.names = 0.7,cex.axis = 0.8, col = brewer.pal(10,"RdBu"))
> mtext("Weather", side=1, line=9)
> par(mar=c(12,5,5,4))
```



One common thought may be that crashes are more likely on difficult weather conditions. However, this visualization shows that this is not the case and crashes occur most frequently on clear days. Therefore, it is important to always drive safe and vigilantly.

3. **Number of crashes per day over the years:**

```
> c = table(crashData$TA_Date)
> View(c)
> plot(c, las=2,cex.axis = 0.8, xlab = " ",ylab ="Number of Crashes",main="Number of
Crashes per Day")
> mtext("Date", side=1, line=6)
```



This visualization shows an interesting trend when the number of crashes occurring per day dropped around March 2021. This can be linked to the start of COVID-19 pandemic and the lockdown. As less vehicles were on the road, the number of crashes occurring also decreased. However, with time as everything returned to normal, the frequency of crashes also increased.

Hypothesis Testing:

Hypothesis: Crashes are deadlier at night than at day as driver may be speeding or be more reckless and crashes are more severe.

Null Hypothesis: Crash fatalities remain the same throughout the day.

1. Extract time of crash and create subsets for day and night crashes:

   ```
   > crashData$Date = str_trunc(crashData$Crash_Date,20,"right",ellipsis=" ")
   > trimws(crashData$Date)
   > crashData$TimeOnly = str_trunc(crashData$Date,10,"left",ellipsis=" ")
   > trimws(crashData$TimeOnly)
   > crashData$TimeOnly = as.POSIXct(crashData$TimeOnly, format ="%H:%M:%S")
   > crashData$TimeOnly = format(crashData$TimeOnly, format ="%H:%M:%S")

   > #Create subset of table with time after 8 pm and before 5 am
   > nightCrash = subset(crashData, crashData$TimeOnly >= "20:00:00")
   > nightCrashBefore5 = subset(crashData, crashData$TimeOnly < "05:00:00")
   > nightCrash = rbind(nightCrash, nightCrashBefore5)
   > #Create subset of table with time before 8 pm and after 5 am
   > dayCrash = subset(crashData, crashData$TimeOnly < "20:00:00")
   > dayCrash = dayCrash[!(dayCrash$TimeOnly %in% nightCrashBefore5$TimeOnly),]
   ```

2. Calculating p value and z-score:

   ```
   > mean.night = mean(nightCrash$Fatality)
   > mean.day = mean(dayCrash$Fatality)
   > sd.night = sd(nightCrash$Fatality)
   > sd.day = sd(dayCrash$Fatality)
   > num.night = length(nightCrash$Fatality)
   > num.day = length(dayCrash$Fatality)
   > zscore = (mean.day - mean.night) / sqrt ( (sd.day^2/num.day) +
   (sd.night^2/num.night))
   > p = pnorm(zscore)
   ```

   Z-score = -1.537371
   P-value = 0.06210129

   We cannot reject the null hypothesis as the p-value is greater than 0.05. Therefore, crash fatalities remain the same throughout the day and nighttime does not increase fatalities.

Predicting Number of Injuries and Fatalities:

Predicting the number of injuries and fatalities by using the number of passengers, number of pedestrians, traffic control, road conditions, light conditions, road configuration and vehicle type.

1. Creating test and train datasets from original dataset:
   ```
   > dt = sort(sample(nrow(crashData),nrow(crashData)*.5))
   > crashDataTrain = crashData[dt,]
   > crashDataTest = crashData[-dt,]
   > crashDataTestWithout = subset(crashDataTest, select = -c(Fatality,Injury))
   ```

2. Cleaning and modifying data for prediction:
   ```
   > crashDataTrain$Weather = as.factor(crashDataTrain$Weather)
   > crashDataTrain$Traffic_Control = as.factor(crashDataTrain$Traffic_Control)
   > crashDataTrain$Road_Configuration = as.factor(crashDataTrain$Road_Configuration)
   > crashDataTrain$Road_Conditions = as.factor(crashDataTrain$Road_Conditions)
   > crashDataTrain$Light_Condition = as.factor(crashDataTrain$Light_Condition)
   > crashDataTrain$Vehicle.Type = as.factor(crashDataTrain$Vehicle.Type)

   #Similary done for crashDataTestWithout dataset, full code in the attached R Code file
   ```

3. Creating the prediction models:
   ```
   > modelFatality = lm(Fatality ~ Weather+Traffic_Control+Road_Configuration+Road_Conditions+Light_Condition+Vehicle.Type,data=crashDataTrain)
   > modelInjury = lm(Injury ~ Weather+Traffic_Control+Road_Configuration+Road_Conditions+Light_Condition+Vehicle.Type,data=crashDataTrain)

   > modelFatality$xlevels[["Weather"]] <- union(modelFatality$xlevels[["Weather"]], levels(crashDataTestWithout[["Weather"]]))
   #Similary done for other predictors, full code in the attached R Code file
   ```

4. Testing the prediction model:

```
> testPredFatality = predict(modelFatality, crashDataTestWithout)
> testPredInjury = predict(modelInjury, crashDataTestWithout)

> regr.error(testPredFatality,crashDataTest$Fatality)
mae            mse            rmse           mape
0.007495737    0.002384604    0.048832408    Inf
> regr.error(testPredInjury,crashDataTest$Injury)
mae          mse          rmse         mape
0.5636078    0.7203997    0.8487636    Inf
```

RMSE for predicting number of fatalities: 0.048832408
RMSE for predicting number of injuries: 0.8487636

Conclusion:

This dataset is very interesting and helpful in understanding the factors and trends that contribute to traffic crashes. Since this is a dynamic and constantly changing dataset, I was also able to detect recent global events such as COVID-19 that impacted the number of crashes. The data visualizations and trends are also very useful for drivers to learn about crash causes and commonalities.

Another interesting find was that crashes do not become deadlier at night, rather the number of fatalities is not affected by time. Therefore, it is important to always drive safely. Lastly, the prediction model can be used to predict if any fatalities or injuries may occur and if yes, how many given multiple predictors such as number of passengers, pedestrians, traffic control, road configurations and more.