CrashData Analysis

Student

2023-04-12

### Load the required libraries

library(ggplot2)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(tidyr)

### Read the CSV file and store it in a data frame

We can see top elements of the data.

# Read the CSV file and store it in a data frame  
crash\_data <- read.csv("crash\_data.csv")  
head(crash\_data)

## Crash\_Year Crash\_Month County Crash\_Type Crash\_Drug\_Use  
## 1 2013 January Barry Head-On No Drugs Involved  
## 2 2013 January Barry Rear-End No Drugs Involved  
## 3 2013 January Bay Single Motor Vehicle No Drugs Involved  
## 4 2013 January Berrien Rear-End No Drugs Involved  
## 5 2013 January Branch Angle No Drugs Involved  
## 6 2013 January Calhoun Head-On - Left Turn No Drugs Involved  
## Crash\_Drinking Crash\_Hit\_and\_Run Speed\_Limit\_at\_Crash\_Site  
## 1 No Drinking Involved Not Hit-and-Run 55  
## 2 No Drinking Involved Not Hit-and-Run 55  
## 3 No Drinking Involved Not Hit-and-Run 55  
## 4 No Drinking Involved Not Hit-and-Run 70  
## 5 No Drinking Involved Not Hit-and-Run 55  
## 6 Drinking Involved Not Hit-and-Run 45  
## Total\_Motor\_Vehicles Crash  
## 1 2 0  
## 2 2 0  
## 3 1 0  
## 4 2 0  
## 5 2 0  
## 6 2 0

# Visualisation

### Check the structure of the data frame

We can see that the data is mix of Int and Chars.

str(crash\_data)

## 'data.frame': 1000 obs. of 10 variables:  
## $ Crash\_Year : int 2013 2013 2013 2013 2013 2013 2013 2013 2013 2013 ...  
## $ Crash\_Month : chr "January" "January" "January" "January" ...  
## $ County : chr "Barry" "Barry" "Bay" "Berrien" ...  
## $ Crash\_Type : chr "Head-On" "Rear-End" "Single Motor Vehicle" "Rear-End" ...  
## $ Crash\_Drug\_Use : chr "No Drugs Involved" "No Drugs Involved" "No Drugs Involved" "No Drugs Involved" ...  
## $ Crash\_Drinking : chr "No Drinking Involved" "No Drinking Involved" "No Drinking Involved" "No Drinking Involved" ...  
## $ Crash\_Hit\_and\_Run : chr "Not Hit-and-Run" "Not Hit-and-Run" "Not Hit-and-Run" "Not Hit-and-Run" ...  
## $ Speed\_Limit\_at\_Crash\_Site: int 55 55 55 70 55 45 55 55 55 55 ...  
## $ Total\_Motor\_Vehicles : int 2 2 1 2 2 2 1 2 2 1 ...  
## $ Crash : int 0 0 0 0 0 0 0 0 0 0 ...

### checking NA values

There are 6 NA values in Speed Limit column, So imputing with mean.

sum(is.na(crash\_data))

## [1] 6

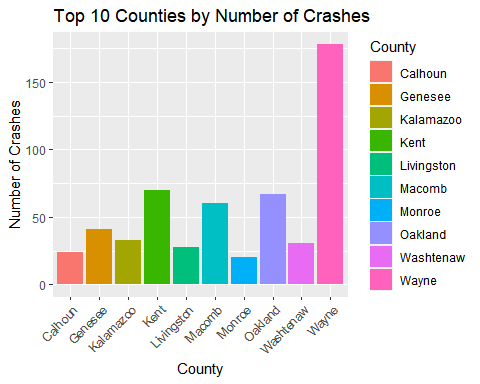
for(i in 1:ncol(crash\_data)){  
 crash\_data[is.na(crash\_data[,i]), i] <- mean(crash\_data[,i], na.rm = TRUE)  
}

## Warning in mean.default(crash\_data[, i], na.rm = TRUE): argument is not numeric  
## or logical: returning NA  
  
## Warning in mean.default(crash\_data[, i], na.rm = TRUE): argument is not numeric  
## or logical: returning NA  
  
## Warning in mean.default(crash\_data[, i], na.rm = TRUE): argument is not numeric  
## or logical: returning NA  
  
## Warning in mean.default(crash\_data[, i], na.rm = TRUE): argument is not numeric  
## or logical: returning NA  
  
## Warning in mean.default(crash\_data[, i], na.rm = TRUE): argument is not numeric  
## or logical: returning NA  
  
## Warning in mean.default(crash\_data[, i], na.rm = TRUE): argument is not numeric  
## or logical: returning NA

### Create a bar chart of the number of crashes by county

We can see that Wayne county has highest number of crashes.

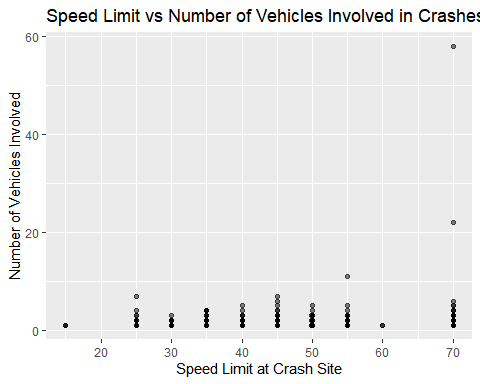
crash\_count\_by\_county <- crash\_data %>%  
 group\_by(County) %>%  
 summarise(count = n()) %>%  
 arrange(desc(count)) %>%  
 head(10)  
  
ggplot(crash\_count\_by\_county, aes(x = County, y = count, fill = County)) +  
 geom\_bar(stat = "identity") +  
 labs(title = "Top 10 Counties by Number of Crashes", x = "County", y = "Number of Crashes") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



### Create a scatter plot of speed limit vs number of vehicles involved

It has been observed that As the Speed Limit increases the vehicle including in crash increases.

ggplot(crash\_data, aes(x = Speed\_Limit\_at\_Crash\_Site, y = Total\_Motor\_Vehicles)) +  
 geom\_point(alpha = 0.5) +  
 labs(title = "Speed Limit vs Number of Vehicles Involved in Crashes", x = "Speed Limit at Crash Site", y = "Number of Vehicles Involved")



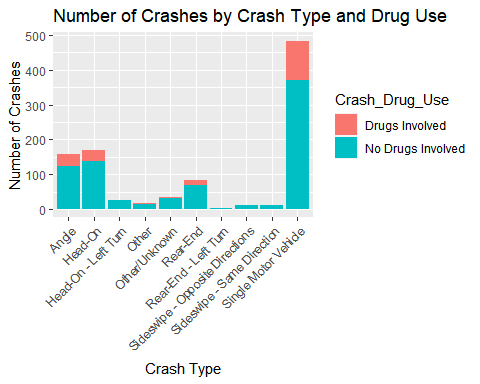
### Create a stacked bar chart of the number of crashes by crash type and drug use

* People Driving single has more number of crashes.
* Head on and angle has the similar number of crashes.

crash\_count\_by\_type\_and\_drug\_use <- crash\_data %>%  
 group\_by(Crash\_Type, Crash\_Drug\_Use) %>%  
 summarise(count = n()) %>%  
 arrange(desc(count))

## `summarise()` has grouped output by 'Crash\_Type'. You can override using the  
## `.groups` argument.

ggplot(crash\_count\_by\_type\_and\_drug\_use, aes(x = Crash\_Type, y = count, fill = Crash\_Drug\_Use)) +  
 geom\_bar(stat = "identity") +  
 labs(title = "Number of Crashes by Crash Type and Drug Use", x = "Crash Type", y = "Number of Crashes") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



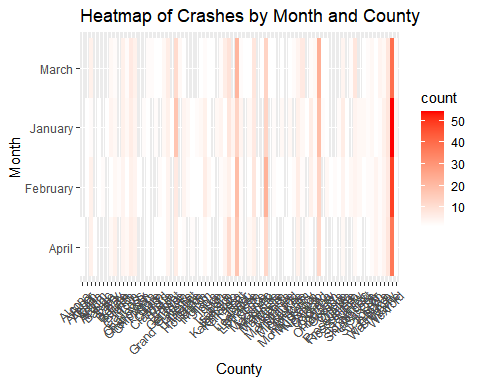
### Heatmap of crashes by month and county

It is difficult to find info from Heatmap. But We can see that Wayne county has highest number of crashes and most are between Jan to Feb.

crash\_count\_by\_month\_and\_county <- crash\_data %>%  
 group\_by(Crash\_Month, County) %>%  
 summarise(count = n()) %>%  
 arrange(Crash\_Month, County)

## `summarise()` has grouped output by 'Crash\_Month'. You can override using the  
## `.groups` argument.

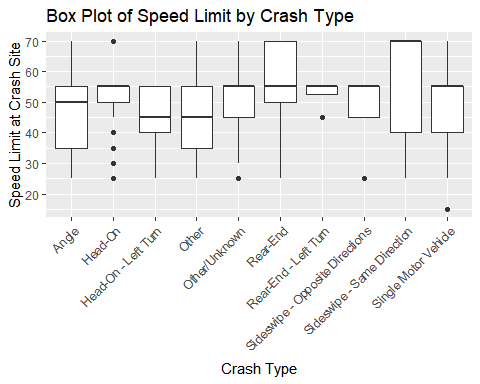
ggplot(crash\_count\_by\_month\_and\_county, aes(x = County, y = Crash\_Month, fill = count)) +  
 geom\_tile() +  
 scale\_fill\_gradient(low = "white", high = "red") +  
 labs(title = "Heatmap of Crashes by Month and County", x = "County", y = "Month") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



### Box plot of speed limit by crash type

We can see some type has outliers but most of the medians are at the upper level, this may be because most of the accidents happens at high speed and most with sideswipe and angle.

ggplot(crash\_data, aes(x = Crash\_Type, y = Speed\_Limit\_at\_Crash\_Site)) +  
 geom\_boxplot() +  
 labs(title = "Box Plot of Speed Limit by Crash Type", x = "Crash Type", y = "Speed Limit at Crash Site") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



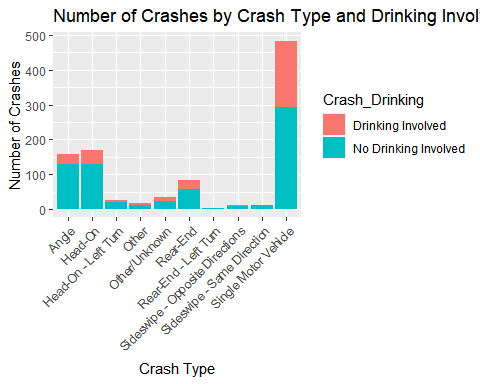
### Stacked bar chart of crashes by crash type and drinking involvement

* Please Driving single has more number of crashes.
* Head on and angle has the similar number of crashes.

crash\_count\_by\_type\_and\_drinking <- crash\_data %>%  
 group\_by(Crash\_Type, Crash\_Drinking) %>%  
 summarise(count = n()) %>%  
 arrange(desc(count))

## `summarise()` has grouped output by 'Crash\_Type'. You can override using the  
## `.groups` argument.

ggplot(crash\_count\_by\_type\_and\_drinking, aes(x = Crash\_Type, y = count, fill = Crash\_Drinking)) +  
 geom\_bar(stat = "identity") +  
 labs(title = "Number of Crashes by Crash Type and Drinking Involvement", x = "Crash Type", y = "Number of Crashes") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



# Statistical Analysis

### Hypothesis test for the difference in means of speed limit between injury and fatal crashes:

# Subset data for injury and fatal crashes  
injury\_data <- subset(crash\_data, Crash == 0)  
fatal\_data <- subset(crash\_data, Crash == 1)  
  
# Conduct t-test for difference in means  
t\_test\_result <- t.test(injury\_data$Speed\_Limit\_at\_Crash\_Site, fatal\_data$Speed\_Limit\_at\_Crash\_Site)  
  
# Print results  
cat("t-test for difference in means of speed limit between injury and fatal crashes:\n")

## t-test for difference in means of speed limit between injury and fatal crashes:

cat("t-value:", t\_test\_result$statistic, "\n")

## t-value: 0.3511422

cat("p-value:", t\_test\_result$p.value, "\n")

## p-value: 0.7281002

The t-test results show the t-value and p-value for the difference in means of speed limit between injury and fatal crashes. The t-value represents the difference in means divided by the standard error of the difference. The p-value represents the probability of observing a t-value as extreme or more extreme than the observed t-value, assuming the null hypothesis that the means are equal. A small p-value (less than the significance level, typically 0.05) indicates strong evidence against the null hypothesis and in favor of the alternative hypothesis that the means are different. In this case, if the p-value is large, we can conclude that there is NO significant difference in the mean speed limit between injury and fatal crashes.

### Logistic regression to predict the likelihood of a fatal crash based on crash type and drinking involvement:

# Create binary variables for crash type and drinking involvement  
crash\_data$Crash\_Type\_Binary <- ifelse(crash\_data$Crash\_Type %in% c("Head-On", "Rear-End"), 0, 1)  
crash\_data$Drinking\_Binary <- ifelse(crash\_data$Crash\_Drinking == "Drinking Involved", 1, 0)  
  
# Fit logistic regression model  
log\_reg\_model <- glm(Crash ~ Crash\_Type\_Binary + Drinking\_Binary, data = crash\_data, family = "binomial")  
  
# Print model summary  
summary(log\_reg\_model)

##   
## Call:  
## glm(formula = Crash ~ Crash\_Type\_Binary + Drinking\_Binary, family = "binomial",   
## data = crash\_data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.2828 -0.2828 -0.2198 -0.1671 3.0933   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.7110 0.4565 -8.130 4.3e-16 \*\*\*  
## Crash\_Type\_Binary 0.5118 0.5001 1.024 0.3060   
## Drinking\_Binary -1.0649 0.5451 -1.954 0.0507 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 255.44 on 999 degrees of freedom  
## Residual deviance: 249.74 on 997 degrees of freedom  
## AIC: 255.74  
##   
## Number of Fisher Scoring iterations: 7

The logistic regression model results show the coefficients, standard errors, z-values, and p-values for the binary variables for crash type and drinking involvement. The coefficients represent the log-odds of a fatal crash associated with each variable, holding all other variables constant. The standard errors represent the uncertainty in the coefficient estimates. The z-values are the coefficients divided by their standard errors, which are used to calculate the p-values. A small p-value (less than the significance level, typically 0.05) for a coefficient indicates strong evidence that the variable is associated with the outcome. In this case, if the p-values for the crash type and drinking involvement variables are small, we can conclude that these variables are significant predictors of the likelihood of a fatal crash.

### ANOVA for the effect of county on the number of vehicles involved in a crash:

# Fit ANOVA model  
anova\_result <- aov(Total\_Motor\_Vehicles ~ County, data = crash\_data)  
  
# Print results  
cat("ANOVA for effect of county on number of vehicles involved in a crash:\n")

## ANOVA for effect of county on number of vehicles involved in a crash:

summary(anova\_result)

## Df Sum Sq Mean Sq F value Pr(>F)  
## County 77 172 2.230 0.501 1  
## Residuals 922 4101 4.448

The ANOVA results show the sum of squares, degrees of freedom, mean squares, F-value, and p-value for the effect of county on the number of vehicles involved in a crash. The sum of squares represents the variation explained by the county variable, and the degrees of freedom represent the number of categories minus one. The mean squares are the sum of squares divided by the degrees of freedom. The F-value is the mean square for county divided by the mean square for error (unexplained variation), which is used to calculate the p-value. A small p-value (less than the significance level, typically 0.05) indicates strong evidence against the null hypothesis that the means are equal and in favor of the alternative hypothesis that there is a significant effect of county on the number of vehicles involved in a crash. In this case, if the p-value is large, we can conclude that there is No significant effect of county on the number of vehicles involved in a crash.