India Road Accident Analysis-By Likith and Avanish

📥 Load Dataset

```
# Replace with your actual file path or use file.choose()

india_road <- read.csv("C:/Users/HP/Desktop/Q1/T1/Prob and stats/Project/Road.csv")
```

📊 Descriptive Statistics

```
summary(india_road)
```

```
##      Time          Day_of_week        Age_band_of_driver Sex_of_driver
##  Length:12316       Length:12316       Length:12316       Length:12316
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Educational_level  Vehicle_driver_relation Driving_experience
##  Length:12316       Length:12316            Length:12316
##  Class :character   Class :character        Class :character
##  Mode  :character   Mode  :character        Mode  :character
##
##
##
##  Type_of_vehicle    Owner_of_vehicle   Service_year_of_vehicle
##  Length:12316       Length:12316       Length:12316
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Defect_of_vehicle  Area_accident_occured Lanes_or_Medians   Road_allignment
##  Length:12316       Length:12316          Length:12316       Length:12316
##  Class :character   Class :character      Class :character   Class :character
##  Mode  :character   Mode  :character      Mode  :character   Mode  :character
##
##
##
##  Types_of_Junction  Road_surface_type  Road_surface_conditions
##  Length:12316       Length:12316       Length:12316
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Light_conditions   Weather_conditions Type_of_collision
##  Length:12316       Length:12316       Length:12316
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Number_of_vehicles_involved Number_of_casualties Vehicle_movement
##  Min.   :1.000               Min.   :1.000        Length:12316
##  1st Qu.:2.000               1st Qu.:1.000        Class :character
##  Median :2.000               Median :1.000        Mode  :character
##  Mean   :2.041               Mean   :1.548
##  3rd Qu.:2.000               3rd Qu.:2.000
##  Max.   :7.000               Max.   :8.000
##  Casualty_class     Sex_of_casualty    Age_band_of_casualty Casualty_severity
##  Length:12316       Length:12316       Length:12316         Length:12316
##  Class :character   Class :character   Class :character     Class :character
##  Mode  :character   Mode  :character   Mode  :character     Mode  :character  
##                                                                               
##                                                                               
##                                                                               
##  Work_of_casuality  Fitness_of_casuality Pedestrian_movement Cause_of_accident 
##  Length:12316       Length:12316         Length:12316        Length:12316      
##  Class :character   Class :character     Class :character    Class :character  
##  Mode  :character   Mode  :character     Mode  :character    Mode  :character  
##                                                                                
##                                                                                
##                                                                                
##  Accident_severity 
##  Length:12316      
##  Class :character  
##  Mode  :character  
##                    
##                    
## 
```

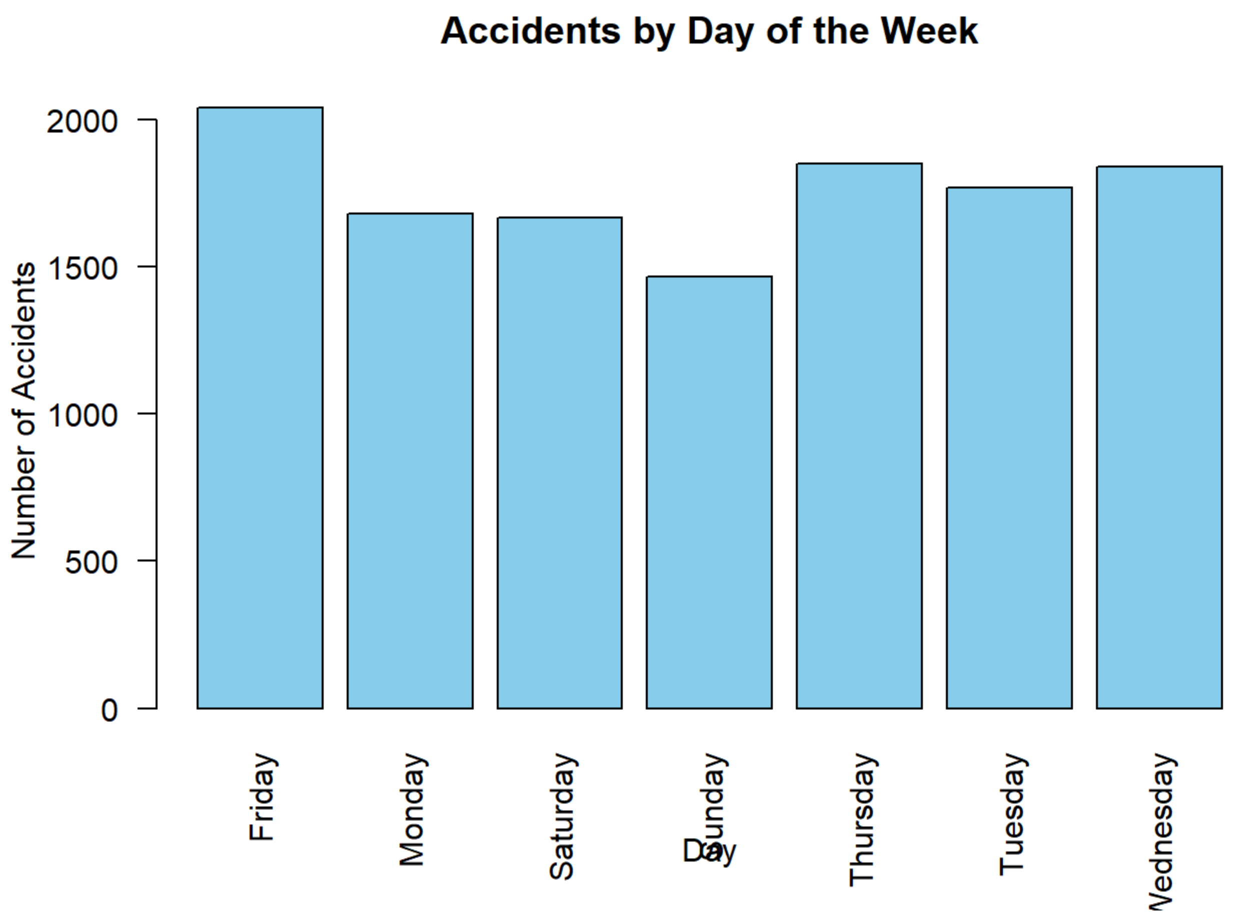
```
describe(india_road[, 1:ncol(india_road)])
```

```
##                             vars     n   mean     sd median trimmed    mad min
## Time*                          1 12316 520.20 288.49  484.5  511.46 316.54   1
## Day_of_week*                   2 12316   3.98   2.06    4.0    3.98   2.97   1
## Age_band_of_driver*            3 12316   2.29   1.34    2.0    2.12   1.48   1
## Sex_of_driver*                 4 12316   1.96   0.26    2.0    2.00   0.00   1
## Educational_level*             5 12316   4.91   1.67    6.0    5.15   0.00   1
## Vehicle_driver_relation*       6 12316   2.29   0.79    2.0    2.17   0.00   1
## Driving_experience*            7 12316   3.74   1.44    4.0    3.74   1.48   1
## Type_of_vehicle*               8 12316   7.06   4.68    7.0    6.68   7.41   1
## Owner_of_vehicle*              9 12316   4.54   1.13    5.0    4.85   0.00   1
## Service_year_of_vehicle*      10 12316   3.62   2.33    3.0    3.52   2.97   1
## Defect_of_vehicle*            11 12316   2.91   1.44    4.0    3.01   0.00   1
## Area_accident_occured*        12 12316   8.28   2.56    9.0    8.46   1.48   1
## Lanes_or_Medians*             13 12316   5.00   1.71    5.0    5.15   1.48   1
## Road_allignment*              14 12316   6.86   1.10    7.0    7.00   0.00   1
## Types_of_Junction*            15 12316   5.06   3.16    3.0    5.04   1.48   1
## Road_surface_type*            16 12316   2.16   0.71    2.0    2.00   0.00   1
## Road_surface_conditions*      17 12316   1.72   1.28    1.0    1.52   0.00   1
## Light_conditions*             18 12316   3.18   1.32    4.0    3.35   0.00   1
## Weather_conditions*           19 12316   3.41   1.15    3.0    3.15   0.00   1
## Type_of_collision*            20 12316   8.43   2.64   10.0    8.96   0.00   1
## Number_of_vehicles_involved   21 12316   2.04   0.69    2.0    2.00   0.00   1
## Number_of_casualties          22 12316   1.55   1.01    1.0    1.31   0.00   1
## Vehicle_movement*             23 12316   4.79   2.14    4.0    4.35   0.00   1
## Casualty_class*               24 12316   1.97   1.02    2.0    1.84   1.48   1
## Sex_of_casualty*              25 12316   2.15   0.74    2.0    2.19   1.48   1
## Age_band_of_casualty*         26 12316   3.06   1.63    4.0    2.98   2.97   1
## Casualty_severity*            27 12316   3.29   0.59    3.0    3.33   0.00   1
## Work_of_casuality*            28 12316   2.41   1.43    2.0    2.25   1.48   1
## Fitness_of_casuality*         29 12316   3.36   1.24    4.0    3.57   0.00   1
## Pedestrian_movement*          30 12316   5.84   0.89    6.0    6.00   0.00   1
## Cause_of_accident*            31 12316   7.92   5.10   10.0    7.76   5.93   1
## Accident_severity*            32 12316   2.83   0.41    3.0    2.93   0.00   1
##                              max range  skew kurtosis   se
## Time*                       1074  1073  0.29    -0.92 2.60
## Day_of_week*                   7     6 -0.01    -1.32 0.02
## Age_band_of_driver*            5     4  0.88    -0.40 0.01
## Sex_of_driver*                 3     2 -1.83    10.41 0.00
## Educational_level*             8     7 -0.92    -0.32 0.02
## Vehicle_driver_relation*       5     4  1.44     1.03 0.01
## Driving_experience*            8     7 -0.02    -0.57 0.01
## Type_of_vehicle*              18    17  0.40    -0.83 0.04
## Owner_of_vehicle*              5     4 -2.17     3.03 0.01
## Service_year_of_vehicle*       7     6  0.28    -1.41 0.02
## Defect_of_vehicle*             4     3 -0.57    -1.67 0.01
## Area_accident_occured*        15    14 -0.57     0.82 0.02
## Lanes_or_Medians*              8     7 -0.45    -0.55 0.02
## Road_allignment*              10     9 -2.98    12.85 0.01
## Types_of_Junction*             9     8  0.32    -1.70 0.03
## Road_surface_type*             6     5  3.84    15.00 0.01
## Road_surface_conditions*       4     3  1.22    -0.51 0.01
## Light_conditions*              4     3 -1.02    -0.94 0.01
## Weather_conditions*          9     8  2.78     8.74 0.01
## Type_of_collision*          11    10 -1.30     0.10 0.02
## Number_of_vehicles_involved  7     6  1.32     5.50 0.01
## Number_of_casualties         8     7  2.34     6.21 0.01
## Vehicle_movement*           14    13  2.00     4.16 0.02
## Casualty_class*              4     3  0.82    -0.46 0.01
## Sex_of_casualty*             3     2 -0.24    -1.16 0.01
## Age_band_of_casualty*        6     5  0.12    -1.22 0.01
## Casualty_severity*           4     3 -0.24    -0.25 0.01
## Work_of_casuality*           8     7  1.09     0.09 0.01
## Fitness_of_casuality*        6     5 -1.36    -0.07 0.01
## Pedestrian_movement*         9     8 -4.08    16.91 0.01
## Cause_of_accident*          20    19  0.05    -1.26 0.05
## Accident_severity*           3     2 -2.34     4.85 0.00
```

```
#Most variables are categorical (e.g., age band, vehicle type, light conditions).
#Mean casualties ≈ 1.55, with up to 8 casualties in a single accident.
#Distribution skew:
#Sex_of_driver: skewed heavily toward one category (likely male).Road_allignment, Road_surface_type, and Weather_conditions: show extreme skewness and kurtosis, indicating very uneven category distributions like most accidents might occur on straight roads or in clear weather)
```

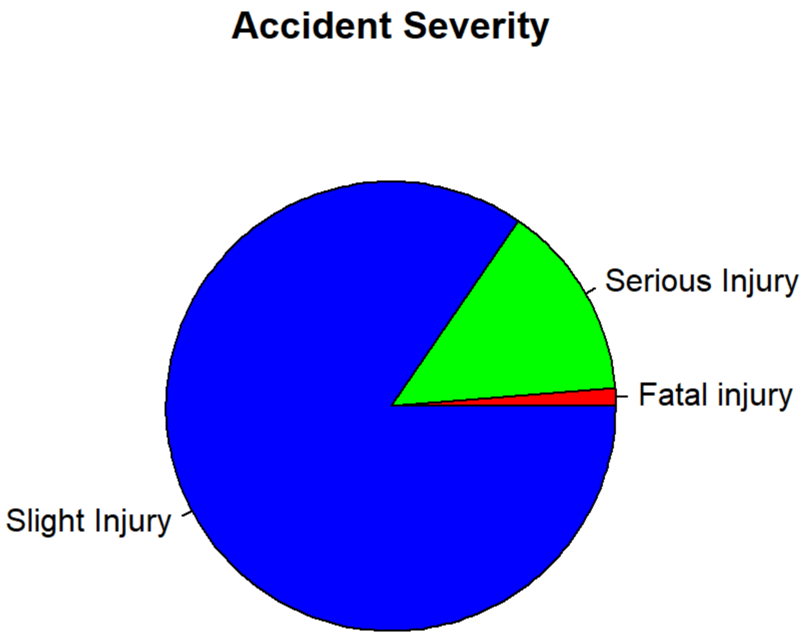
📈 Bar Plot: Accidents by Day

```
accident_counts <- table(india_road$Day_of_week)
barplot(accident_counts,
        main = "Accidents by Day of the Week",
        xlab = "Day",
        ylab = "Number of Accidents",
        col = "skyblue", las = 2)
```



🥧 Pie Chart: Accident Severity Distribution

```
pie(table(india_road$Accident_severity),
    main = "Accident Severity",
    col = rainbow(length(unique(india_road$Accident_severity))))
```



```
#Majority of accidents are Slight Injuries.Fatal Injuries are a small fraction. This indicates that while accidents are common, they’re not always severe
```

📅 Weekend Analysis

```
weekend_days <- c("Saturday", "Sunday")
india_road$Weekend <- india_road$Day_of_week %in% weekend_days

# Probability of accident on weekend
mean(india_road$Weekend)
```

```
## [1] 0.2543845
```

```
#25.4% of all accidents occur on weekends.While lower than 50%, still significant.
```

📊 Conditional Probability: Severity by Day

```
prop.table(table(india_road$Accident_severity, india_road$Day_of_week), 2)
```

```
##
##                      Friday      Monday    Saturday      Sunday    Thursday
##   Fatal injury   0.007839294 0.007138608 0.022208884 0.023858214 0.011885467
##   Serious Injury 0.153356198 0.121356336 0.147058824 0.129516019 0.146947596
##   Slight Injury  0.838804508 0.871505057 0.830732293 0.846625767 0.841166937
##
##                     Tuesday   Wednesday
##   Fatal injury   0.009604520 0.010326087
##   Serious Injury 0.145197740 0.142391304
##   Slight Injury  0.845197740 0.847282609
```

```
#Fatal injuries spike on Saturday (2.2%) and Sunday (2.4%).Slight injuries dominate every day but proportionally decrease on weekends
```

🔍 t-test: Weekday vs Weekend Accidents

```
day_counts <- table(india_road$Day_of_week)
week_status <- names(day_counts) %in% weekend_days
t.test(day_counts[week_status], day_counts[!week_status])
```

```
##
##  Welch Two Sample t-test
##
## data:  day_counts[week_status] and day_counts[!week_status]
## t = -2.3305, df = 1.7841, p-value = 0.1602
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -831.313  291.113
## sample estimates:
## mean of x mean of y
##    1566.5    1836.6
```

```
#p-value = 0.16 → not statistically significant.Means: Weekend avg = 1566.5; Weekday avg = 1836.6.No strong evidence that accident count differs between weekdays and weekends
# p is high so NULL will fly
```

🔁 Chi-square Test: Severity vs Day

```
chisq.test(table(india_road$Accident_severity, india_road$Day_of_week))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(india_road$Accident_severity, india_road$Day_of_week)
## X-squared = 47.202, df = 12, p-value = 4.3e-06
```

```
#Chi-sq = 47.20, p < 0.001.Strong association between day of week and accident severity.Supports
earlier probability analysis on severity variation by day
```

🔬 ANOVA: Severity Across Time Period

```
india_road$Hour <- as.numeric(substr(india_road$Time, 1, 2))
india_road$Severity_numeric <- as.numeric(factor(india_road$Accident_severity))
india_road$TimePeriod <- cut(india_road$Hour,
                             breaks = c(-1, 6, 12, 18, 24),
                             labels = c("Night", "Morning", "Afternoon", "Evening"))
anova_result <- aov(Severity_numeric ~ TimePeriod, data = india_road)
summary(anova_result)
```

```
##                Df Sum Sq Mean Sq F value   Pr(>F)    
## TimePeriod      2      5  2.4765   14.98 3.21e-07 ***
## Residuals    9505   1572  0.1654                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 2808 observations deleted due to missingness
```

```
#p < 0.001, F = 14.98.Severity levels significantly differ by time periods (Morning, Afternoon,
Night).
```

🧪 Kruskal-Wallis: Severity by Day (Non-parametric)

```
kruskal.test(Severity_numeric ~ Day_of_week, data = india_road)
```

```
## 
##  Kruskal-Wallis rank sum test
## 
## data:  Severity_numeric by Day_of_week
## Kruskal-Wallis chi-squared = 13.169, df = 6, p-value = 0.04042
```

```
#p = 0.0404.Confirms that accident severity varies by day even when normality is not assumed.
```

🔗 Correlation Matrix

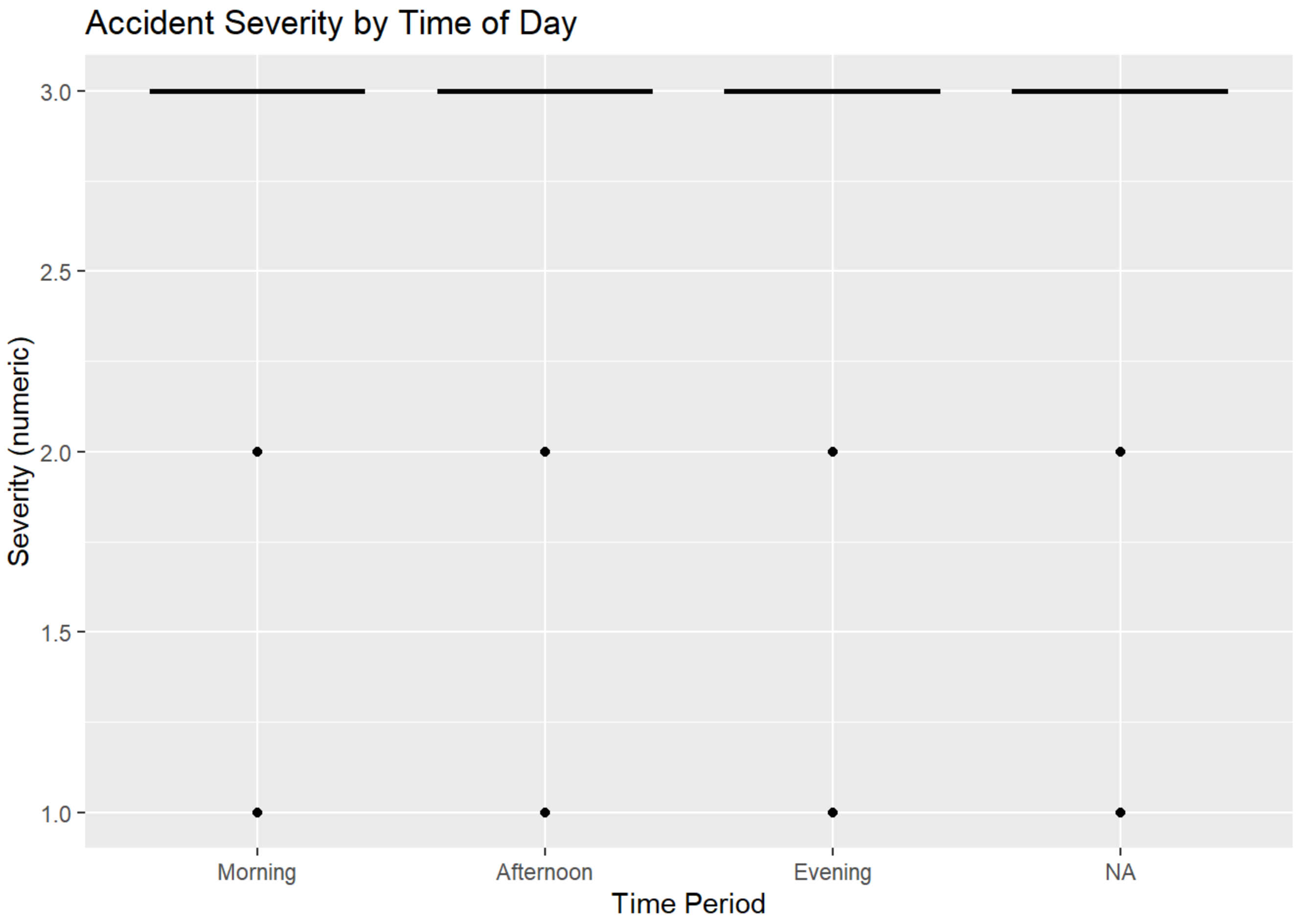
```
india_road$Day_numeric <- as.numeric(factor(india_road$Day_of_week,
                                            levels = c("Monday", "Tuesday", "Wednesday", "Thursd
ay", "Friday", "Saturday", "Sunday")))

numeric_data <- india_road %>%
  mutate(Sex_numeric = as.numeric(factor(Sex_of_driver)),
         Edu_numeric = as.numeric(factor(Educational_level)),
         Exp_numeric = as.numeric(factor(Driving_experience)),
         Light_numeric = as.numeric(factor(Light_conditions)),
         Weather_numeric = as.numeric(factor(Weather_conditions)),
         Cause_numeric = as.numeric(factor(Cause_of_accident)),
         Area_numeric = as.numeric(factor(Area_accident_occured)),
         Align_numeric = as.numeric(factor(Road_allignment))) %>%
  select_if(is.numeric)

cor_matrix <- cor(numeric_data, use = "complete.obs")
round(cor_matrix, 2)
```

```
##                             Number_of_vehicles_involved Number_of_casualties
## Number_of_vehicles_involved                        1.00                 0.24
## Number_of_casualties                               0.24                 1.00
## Hour                                              -0.01                 0.03
## Severity_numeric                                   0.11                -0.05
## Day_numeric                                        0.02                 0.08
## Sex_numeric                                       -0.03                 0.04
## Edu_numeric                                        0.02                 0.00
## Exp_numeric                                       -0.01                 0.00
## Light_numeric                                      0.01                -0.03
## Weather_numeric                                   -0.03                 0.01
## Cause_numeric                                     -0.02                -0.02
## Area_numeric                                      -0.01                 0.00
## Align_numeric                                      0.00                -0.01
##                              Hour Severity_numeric Day_numeric Sex_numeric
## Number_of_vehicles_involved -0.01             0.11        0.02       -0.03
## Number_of_casualties         0.03            -0.05        0.08        0.04
## Hour                         1.00            -0.06        0.02       -0.04
## Severity_numeric            -0.06             1.00       -0.03        0.01
## Day_numeric                  0.02            -0.03        1.00        0.00
## Sex_numeric                 -0.04             0.01        0.00        1.00
## Edu_numeric                 -0.01             0.01       -0.01       -0.01
## Exp_numeric                 -0.01             0.01       -0.02        0.01
## Light_numeric               -0.55             0.02        0.00        0.04
## Weather_numeric              0.01             0.00        0.00       -0.02
## Cause_numeric                0.00             0.01       -0.01        0.00
## Area_numeric                 0.00            -0.02        0.00       -0.01
## Align_numeric               -0.01             0.00        0.01        0.00
##                             Edu_numeric Exp_numeric Light_numeric
## Number_of_vehicles_involved        0.02       -0.01          0.01
## Number_of_casualties               0.00        0.00         -0.03
## Hour                              -0.01       -0.01         -0.55
## Severity_numeric                   0.01        0.01          0.02
## Day_numeric                       -0.01       -0.02          0.00
## Sex_numeric                       -0.01        0.01          0.04
## Edu_numeric                        1.00        0.24          0.01
## Exp_numeric                        0.24        1.00          0.00
## Light_numeric                      0.01        0.00          1.00
## Weather_numeric                    0.01        0.01         -0.07
## Cause_numeric                     -0.01       -0.01          0.00
## Area_numeric                       0.00        0.00         -0.02
## Align_numeric                     -0.01       -0.02          0.00
##                             Weather_numeric Cause_numeric Area_numeric
## Number_of_vehicles_involved           -0.03         -0.02        -0.01
## Number_of_casualties                   0.01         -0.02         0.00
## Hour                                   0.01          0.00         0.00
## Severity_numeric                       0.00          0.01        -0.02
## Day_numeric                            0.00         -0.01         0.00
## Sex_numeric                           -0.02          0.00        -0.01
## Edu_numeric                            0.01         -0.01         0.00
## Exp_numeric                            0.01         -0.01         0.00
## Light_numeric                         -0.07          0.00        -0.02
## Weather_numeric                        1.00          0.00         0.00
## Cause_numeric                          0.00          1.00         0.00
## Area_numeric                           0.00          0.00         1.00
## Align_numeric                          0.01         -0.01         0.02
##                             Align_numeric
## Number_of_vehicles_involved          0.00
## Number_of_casualties                -0.01
## Hour                                -0.01
## Severity_numeric                     0.00
## Day_numeric                          0.01
## Sex_numeric                          0.00
## Edu_numeric                         -0.01
## Exp_numeric                         -0.02
## Light_numeric                        0.00
## Weather_numeric                      0.01
## Cause_numeric                       -0.01
## Area_numeric                         0.02
## Align_numeric                        1.00
```

```
#Severity correlates weakly with all variables.no single variable strongly drives severity on its own
```

📦 Boxplot: Severity by Time Period

```
ggplot(india_road, aes(x = TimePeriod, y = Severity_numeric)) +
  geom_boxplot(fill = "orange", color = "black") +
  labs(title = "Accident Severity by Time of Day", x = "Time Period", y = "Severity (numeric)")
```



```
#Outliers in Nighttime suggest some very severe accidents.Median severity is higher in Evening and Night.
```

🔍 Logistic Regression: Predict Severe Accidents

```
# Create numeric columns directly in india_road
india_road$Sex_numeric <- as.numeric(factor(india_road$Sex_of_driver))
india_road$Edu_numeric <- as.numeric(factor(india_road$Educational_level))
india_road$Exp_numeric <- as.numeric(factor(india_road$Driving_experience))
india_road$Light_numeric <- as.numeric(factor(india_road$Light_conditions))
india_road$Weather_numeric <- as.numeric(factor(india_road$Weather_conditions))
india_road$Area_numeric <- as.numeric(factor(india_road$Area_accident_occured))

india_road$Severe <- ifelse(india_road$Accident_severity == "Fatal injury", 1, 0)
logit_model <- glm(Severe ~ Day_numeric + Hour + Sex_numeric + Edu_numeric +
                     Exp_numeric + Light_numeric + Weather_numeric + Area_numeric,
                   data = india_road, family = binomial)
summary(logit_model)
```

```
##
## Call:
## glm(formula = Severe ~ Day_numeric + Hour + Sex_numeric + Edu_numeric +
##     Exp_numeric + Light_numeric + Weather_numeric + Area_numeric,
##     family = binomial, data = india_road)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -6.76353    1.18340  -5.715 1.09e-08 ***
## Day_numeric      0.19740    0.04850   4.070 4.71e-05 ***
## Hour             0.08462    0.03233   2.617  0.00886 **
## Sex_numeric      0.48109    0.36441   1.320  0.18677
## Edu_numeric     -0.04963    0.05265  -0.943  0.34582
## Exp_numeric     -0.08197    0.06404  -1.280  0.20057
## Light_numeric   -0.08264    0.07625  -1.084  0.27845
## Weather_numeric -0.05320    0.08607  -0.618  0.53650
## Area_numeric     0.02604    0.03499   0.744  0.45685
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1357.1  on 9507  degrees of freedom
## Residual deviance: 1315.4  on 9499  degrees of freedom
##   (2808 observations deleted due to missingness)
## AIC: 1333.4
##
## Number of Fisher Scoring iterations: 7
```

```
#Significant Predictors are Day_Numeric and Hour as they have very low p value and they have a higher chance of rejecting the null hypothesis.Personal/driver factors don’t significantly affect fatal injury likelihood.
```

🕓 Specific Hours Analysis

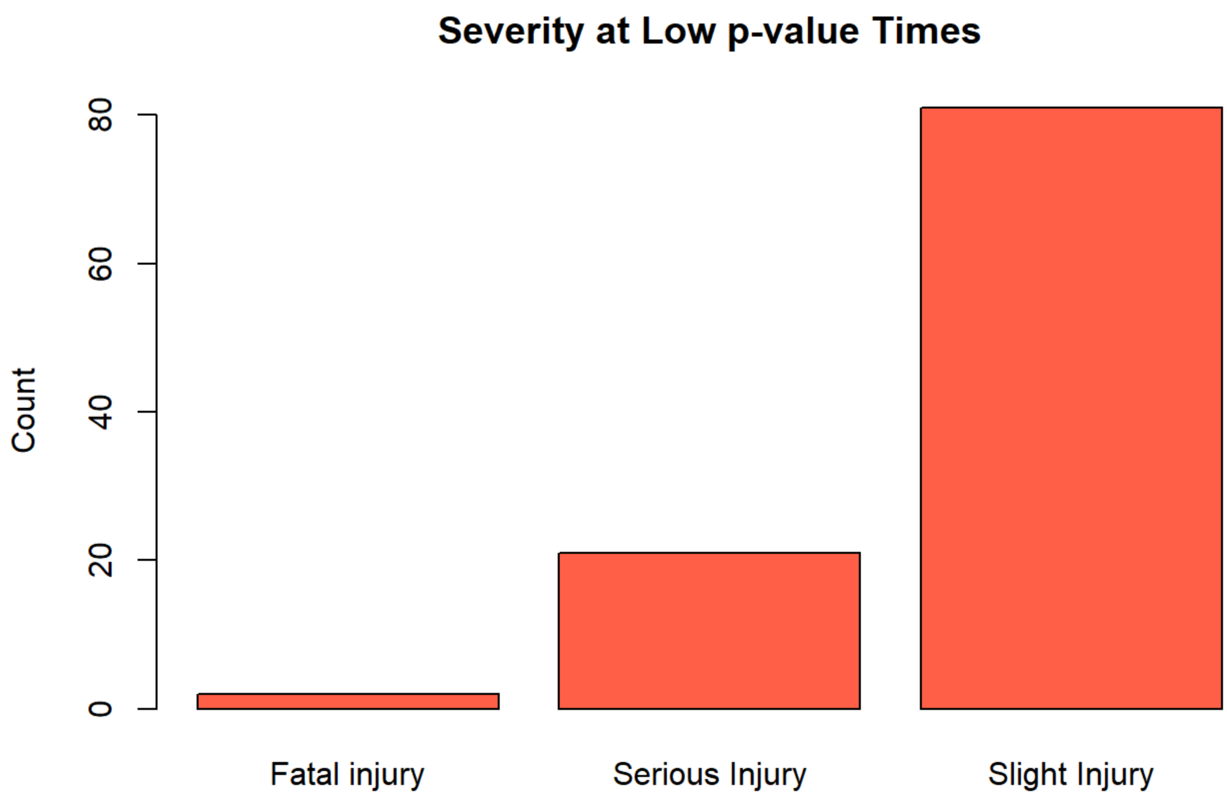
```
hours_to_check <- c("0:04:00", "0:10:00", "0:18:00", "0:36:00", "0:56:00",
                    "1:12:00", "1:35:00", "10:01:00", "11:06:00", "11:08:00",
                    "11:16:00", "11:44:00", "12:04:00", "12:11:00", "12:14:00")

india_road$Time_stripped <- substr(india_road$Time, 2, 8)
subset_times <- india_road[india_road$Time_stripped %in% hours_to_check, ]
table(subset_times$Accident_severity)
```

```
##
##   Fatal injury Serious Injury  Slight Injury
##              2             21             81
```

```
#At low p-value time slots:2 Fatal, 21 Serious, 81 Slight Accidents.These windows are crucial for real-time intervention like patrols, cameras.
```

📊 Bar Plot: Severity at Specific Times

```
barplot(table(subset_times$Accident_severity),
        main = "Severity at Low p-value Times",
        col = "tomato", ylab = "Count")
```



```
#visualization confirms most cases at critical hours are Slight.the presence of Fatal and Serious suggests the need for alert systems even during low-volume windows.
```