UK Road Safety

Arijit Guchhait

2026-02-14

ETL

Loading Required library

```
library(tidyverse) # For dplyr pipe and ggplot
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.6
v forcats   1.0.1     v stringr   1.6.0
v ggplot2   4.0.1     v tibble    3.3.1
v lubridate 1.9.4     v tidyr     1.3.2
v purrr     1.2.1
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(corrplot) # For co-relation matrix
```

```
corrplot 0.95 loaded
```

Data Extrcation And Loading from Source

```r
if (!file.exists("data/raw/collision.csv")) {
  df <- read_csv("[https://data.dft.gov.uk/road-accidents-safety-data/](https://data.dft.gov
  write.csv(df, "data/raw/collision.csv", row.names = FALSE)
} else {
  df <- read_csv("data/raw/collision.csv")
}
```

```
New names:
Rows: 48472 Columns: 45
-- Column specification
-------------------------------------------------------- Delimiter: "," chr
(7): collision_index, collision_ref_no, date, local_authority_ons_dist... dbl
(37): ...1, collision_year, location_easting_osgr, location_northing_os... time
(1): time
i Use `spec()` to retrieve the full column specification for this data. i
Specify the column types or set `show_col_types = FALSE` to quiet this message.
* `` -> `...1`
```

```r
cat("File extrcated at:",as.character(Sys.Date())," to data/raw/collision.csv")
```

```
File extrcated at: 2026-02-14  to data/raw/collision.csv
```

Data Transformation

```r
df_clean <- df %>%
  na.omit() %>%
  mutate(
    date=as.Date(date,format="%d/%m/%Y"),
    month=month(date),
    hour =as.numeric(substr(time,1,2)) # Using Base R function to get the hour information
    # from the time columns
    )%>%
    # Selecting only relevant columns for analysis
  select(
    collision_index,collision_severity,number_of_vehicles,
    number_of_casualties,day_of_week,month,hour,
    weather_conditions,light_conditions,location_easting_osgr,
    location_northing_osgr,road_type,speed_limit,road_surface_conditions,
    junction_detail,police_force
```

```
  )%>%
  # Remove rows with -1 (missing values)
filter(
  speed_limit != -1,
  road_surface_conditions != -1,
  junction_detail != -1,
  weather_conditions != -1,
  light_conditions != -1
) %>%
mutate(
 weather_condition_legend =factor(
    weather_conditions,
    levels = c(1,2,3,4,5,6,7,8,9),
    labels=c(
      "fine","raining","snowing",
       "fine+winds","raining+winds",
      "snowing+winds","fog/mist","other",
      "unknown"
      )
    )
  ) %>%
mutate(
  light_conditions_legend=factor(
    light_conditions,
    levels=c(1,4,5,6,7),
    labels=c(
      "daylight",
      "darkness_lights_lit",
      "darkness_lights_unlit",
      "darkness_no_lighting",
      "darkness_lighting_unknown"
      )
    )
  ) %>%
mutate(
  hour_legend=factor(
    case_when(
      hour >=6 & hour <12 ~ "morning",
      hour >= 12 & hour < 18 ~ "afternoon",
      hour >= 18 & hour <= 22 ~ "evening",
      hour < 6 | hour >22~ "night"
    ),
```

```
      levels =c("morning","afternoon","evening","night")
    )
  )%>%
  mutate(
    day_of_week_legend=factor(
      day_of_week,
      levels=1:7,
      labels=c(
        "Sunday","Monday","Tuesday",
        "Wednesday","Thursday","Friday","Saturday"
        )
      ),
    month_legend=factor(
      month,levels=1:12,
      labels=c("Jan", "Feb", "Mar", "Apr", "May", "Jun",
               "Jul", "Aug", "Sep", "Oct", "Nov", "Dec"
               )
      )
    ) %>%
  mutate(
    collision_severity_legend=factor(
      collision_severity,
      levels =1:3,
      labels =c("Fatal","Serious","Slight")
    )
  ) %>%
  select(
    # 1. Collision related Columns:
    collision_index,collision_severity,number_of_casualties,number_of_vehicles,collision_seve
    # 2. Date Time Columns:
    hour,hour_legend,day_of_week,day_of_week_legend,month,month_legend,
    # 3. Environmental Columns:
    weather_conditions,weather_condition_legend,light_conditions,light_conditions_legend,
    #4. Traffic Columns:
    road_type,speed_limit,road_surface_conditions,police_force,
    junction_detail,
    #5. Geographical Columns:
    location_easting_osgr,location_northing_osgr
    )
write.csv(df_clean,"data/processed/clean.csv")
cat("Cleaned dataset with", nrow(df_clean), "rows and", ncol(df_clean), "columns.")
```

```
Cleaned dataset with 42135 rows and 22 columns.
```

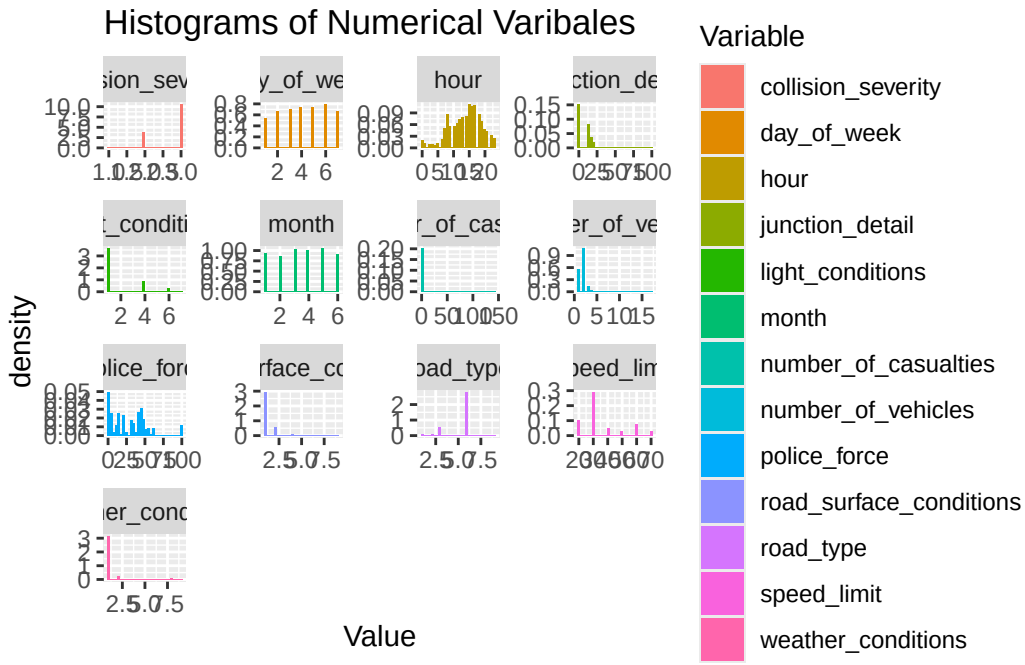
EDA

```r
# Define numeric columns for analysis
numeric_col <- df_clean %>%
  select(where(is.numeric),-location_northing_osgr,
         -location_easting_osgr) %>%
  names()
# Statistical Summary
numeric_col_summary <- data.frame(
  Variable = numeric_col,
  first_quartile = sapply(df_clean[numeric_col], function(x) quantile(x, 0.25)),
  third_quartile = sapply(df_clean[numeric_col], function(x) quantile(x, 0.75)),
  IQR = sapply(df_clean[numeric_col], IQR),
  SD = sapply(df_clean[numeric_col], sd),
  Mean = sapply(df_clean[numeric_col], mean),
  row.names = NULL
)
numeric_col_summary
```

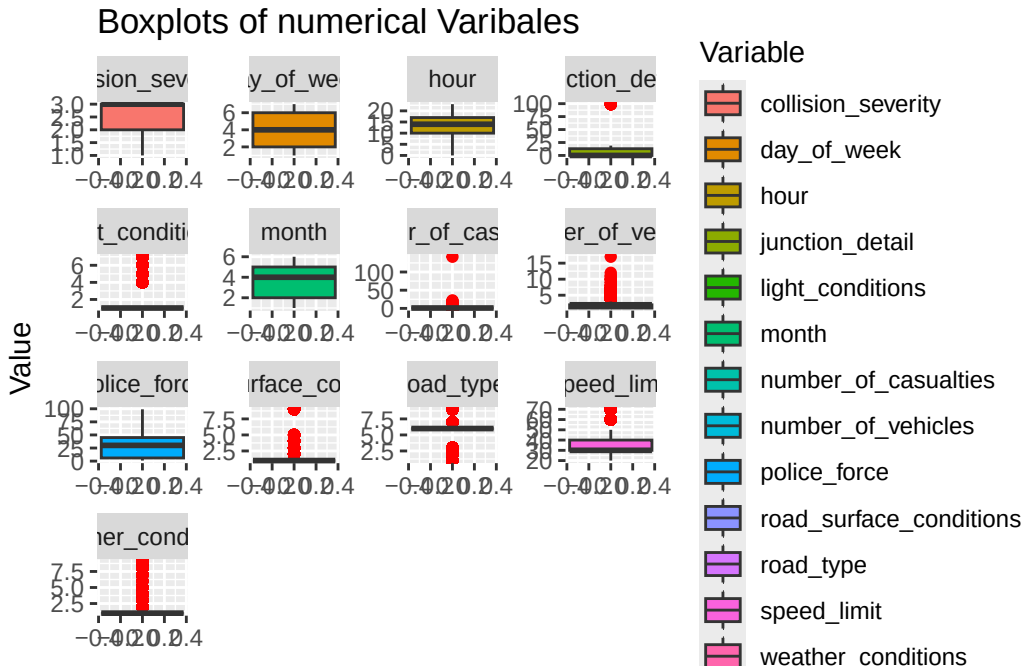
```
                 Variable first_quartile third_quartile IQR         SD
1       collision_severity              2              3   1  0.4870695
2     number_of_casualties              1              1   0  0.9742074
3       number_of_vehicles              1              2   1  0.7115610
4                     hour             10             17   7  5.0634852
5              day_of_week              2              6   4  1.9313053
6                    month              2              5   3  1.6787457
7       weather_conditions              1              1   0  1.6447450
8         light_conditions              1              1   0  1.6387375
9                road_type              6              6   0  1.4968886
10             speed_limit             30             40  10 14.6430679
11 road_surface_conditions              1              1   0  0.8803086
12            police_force              6             45  39 23.7966220
13         junction_detail              0             13  13  7.7074312
        Mean
1   2.709766
2   1.273312
3   1.811273
4  13.711404
```

```
5   4.135683
6   3.530129
7   1.482141
8   1.898066
9   5.340857
10 36.195562
11  1.290898
12 28.994209
13  7.170784
```

```r
# Faceted Histograms showing distribution
df_clean %>%
  select(where(is.numeric),-location_northing_osgr,
         -location_easting_osgr) %>%
  pivot_longer(
    cols = everything(),
    names_to ="Variable",
    values_to = "Value"
  ) %>%
  ggplot(aes(x=Value,fill=Variable))+
  geom_histogram(aes(y=after_stat(density)),bins=30)+
  facet_wrap(~Variable,scales="free")+
  labs(title="Histograms of Numerical Varibales")
```

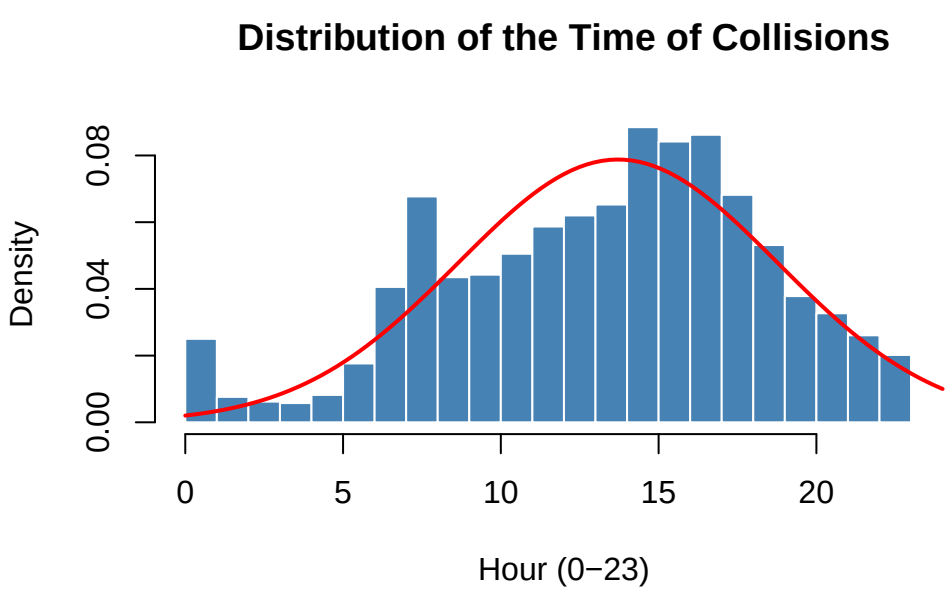


```r
# Faceted Boxplots showing IQR and distribution
df_clean %>%
  select(where(is.numeric),-location_northing_osgr,
         -location_easting_osgr) %>%
  pivot_longer(
    cols=everything(),
    names_to = "Variable",
    values_to = "Value"
  ) %>%
  ggplot(aes(y=Value,fill=Variable))+
  geom_boxplot(outlier.color = 'red')+
  facet_wrap(~Variable,scale='free')+
  labs(title="Boxplots of numerical Varibales")
```

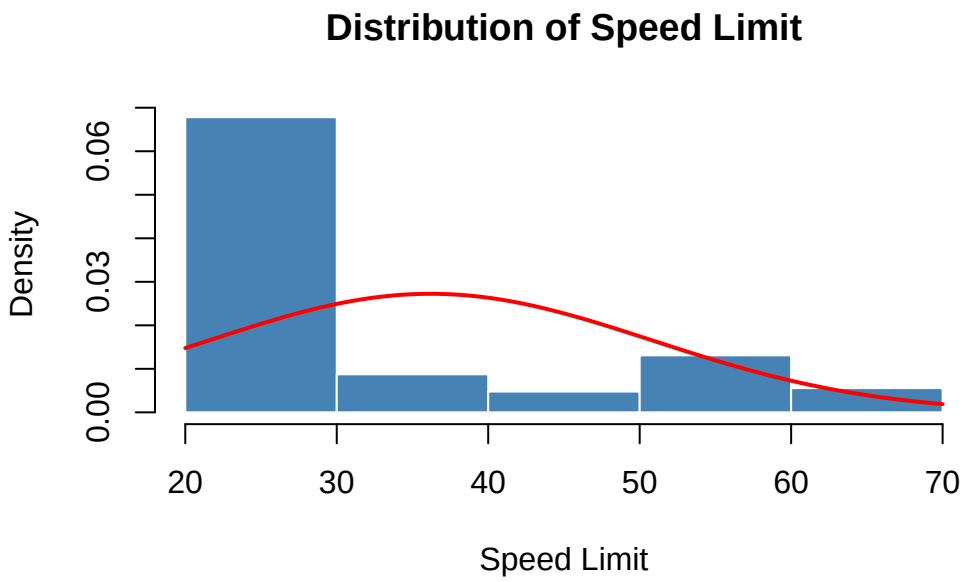



```r
# Distribution of time of collisions

hist(
  df_clean$hour,prob=TRUE,seq(0,24,by=1),
  main= "Distribution of the Time of Collisions",
  xlab= "Hour (0-23)",
  col = "steelblue",
  border = "white"
```

```
)
curve(dnorm(x,mean(df_clean$hour),
            sd(df_clean$hour)),
      add = TRUE,col="red",lwd=2
      )
```

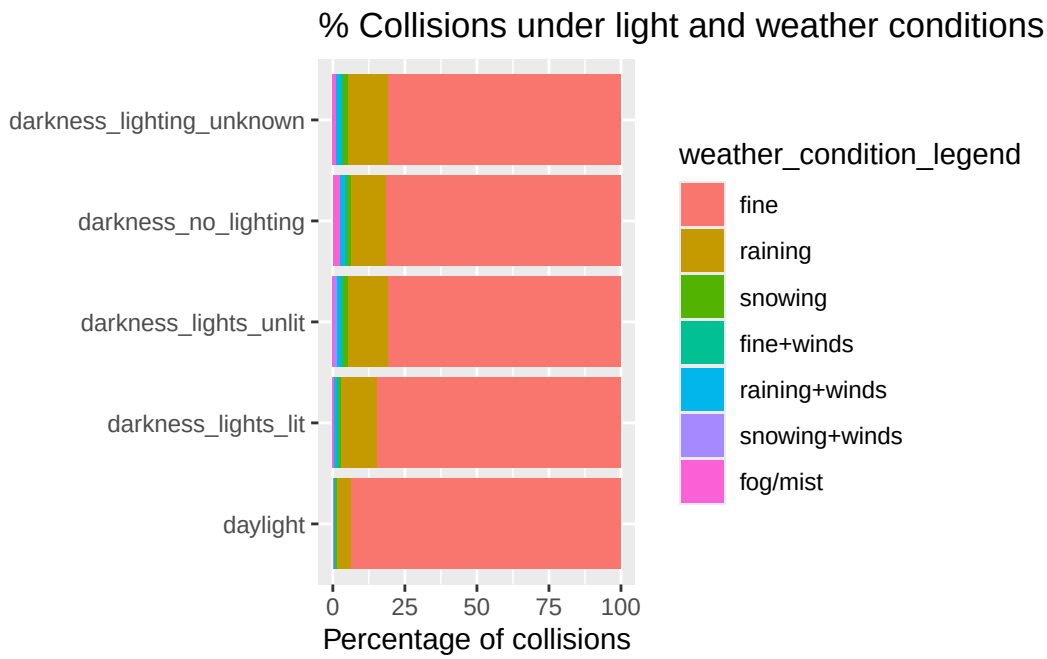



```
# Distribution of Speed Limit

hist(
  df_clean$speed_limit,prob=TRUE,breaks=seq(20,70,by=10),
  col="steelblue",border='white',main = "Distribution of Speed Limit", xlab = "Speed Limit"
)
curve(
  dnorm(x,mean(df_clean$speed_limit),sd(df_clean$speed_limit)),
  add=TRUE,
  col= "red",lwd=2)
```




```
df_clean %>%
  filter(
    !weather_condition_legend %in% c("other","unknown")
    )%>%
  group_by(
    light_conditions_legend,weather_condition_legend
    ) %>%
  summarise(
    Total_collision=n(),.groups='drop',
    ) %>%
  group_by(light_conditions_legend) %>%
  mutate(
    percentage_of_collision=Total_collision*100/sum(Total_collision)
  ) %>%
  ggplot(
    aes(x=light_conditions_legend,
        y=percentage_of_collision,
        fill=weather_condition_legend)
  )+
  geom_col()+
  coord_flip()+
  labs(
    title= '% Collisions under light and weather conditions',
```

```
    x= NULL,
    y= 'Percentage of collisions'
  )
```



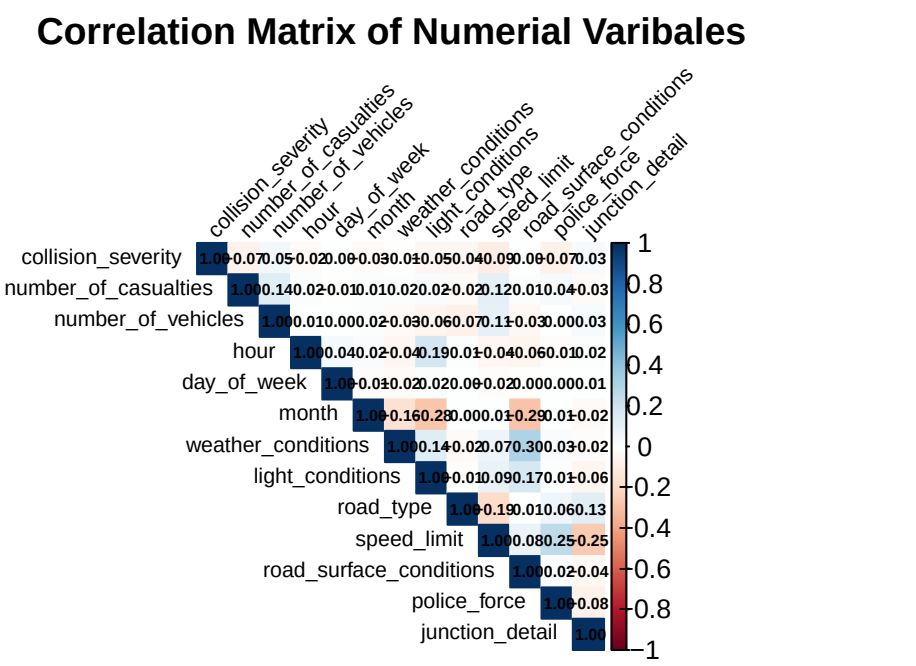
Key observations are:

- Distribution of Time (Hour) shows that most collision happens in afternoon (around 03 pm) and it is left skewed means longer tail in the left region of the distribution curve. So more accident is around 15 hours
- As discussed before, the distribution of speed limit is right skewed means more accidents occur in 30-40 mph and there are outliers (60-70 mph).
- we find how 91% accidents happens in fine weather and daylight also account for only 75%. This shows how relatively "safe" environment can be more dangerous.

Co-relation analysis of variables

```
# Creating co-reelation matrix view
df_filtered <- df_clean |>
  filter(!weather_conditions %in% c(8, 9)) # Better syntax using %in%
```

```r
# Create correlation matrix from FILTERED data
cor_matrix <- cor(df_filtered[numeric_col], use = "pairwise.complete.obs")

# Adjust margins and plot
par(mar = c(2, 2, 4, 2))  # Adjust margins
corrplot(cor_matrix,  # Use cor_matrix, not Cor_M
         method = "color",
         type = "upper",
         tl.col = "black",
         tl.srt = 45,
         tl.cex = 0.7,
         number.cex = 0.5,
         number.digits = 2,
         addCoef.col = "black",
         diag = TRUE,
         title = "Correlation Matrix of Numerial Varibales",
         mar = c(0, 0, 2, 0))  # Title margin
```



Summary:

- Here we can see as expected number of vehicles has a positive co-relation (0.05) with collision_severity.

- Furthermore, oppose to known believe speed_limit shows negative co-relation (-0.09) with `collision_severity` which we will explore further in next model section