Road Accidents in Kenya (2012–2023) Analysis

Eann Baraka

2025-08-12

Abstract

This project explores road traffic accident data in Kenya (2012–2023) to uncover trends in time, location and accident characteristics. Using Excel, R, and Tableau, the dataset was cleaned, transformed, and visualized to provide insights into the most accident-prone roads, counties and time periods. The results highlight key risks and provide recommendations to inform policymakers, road users and transport authorities.

Introduction

Road traffic accidents are a major concern in Kenya. This report analyzes crash data from 2012–2023, identifies high-risk roads, peak accident times, and days and provides recommendations for improved road safety.

1. Ask (Problem Statement)

Guiding Question: Which roads, times, and days are most prone to accidents, and how can stakeholders act to reduce them?

2. Prepare (Data Collection)

Citation of source:

Milusheva, S. (2024). Road Traffic Crashes 2012-2023 [Data set]. World Bank, Development Data Group. https://doi.org/10.48529/ZJMW-PJ61

Fields: Crash ID, Crash date_time, Crash Date, latitude, longitude, n_crash_reports, contains_fatality_words, contains_pedestrian_words, contains_matatu_words, contains_motorcycle_words.

Tools Used: Excel, R and Tableau.

3. Process (Data Cleaning & Transformation in Excel and R)

(a) Initial Cleaning in Excel

Removed duplicate records.

Corrected obvious typos or missing values in categorical columns.

Saved cleaned file as crashes_Acode.csv

```r
library(tmaptools)
```

```
## Warning: package 'tmaptools' was built under R version 4.5.1
```

```r
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 4.5.1
```

```r
library(readr)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

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

```r
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.5.1
```

```
##
## 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
```

```r
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.5.1
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats 1.0.0     v stringr 1.5.1
## v ggplot2 3.5.2     v tibble  3.3.0
## v purrr   1.0.4     v tidyr   1.3.1
```

```
## -- 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 become errors
```

```r
#Read file
df <- read.csv("crashes_Acode.csv")

head(df)
```

```
##   crash_id      crash_datetime crash_date  latitude longitude n_crash_reports
## 1        1 2018-06-06 20:39:54 2018-06-06 -1.263030  36.76437               1
## 2        2 2018-08-17 06:15:54 2018-08-17 -0.829710  37.03782               1
## 3        3 2018-05-25 17:51:54 2018-05-25 -1.125301  37.00330               1
## 4        4 2018-05-25 18:11:54 2018-05-25 -1.740958  37.12903               1
## 5        5 2018-05-25 21:59:54 2018-05-25 -1.259392  36.84232               1
## 6        6 2018-05-26 07:11:54 2018-05-26 -1.215499  36.83515               1
##   contains_fatality_words contains_pedestrian_words contains_matatu_words
## 1                       0                         0                     0
## 2                       1                         0                     0
## 3                       0                         0                     0
## 4                       0                         0                     0
## 5                       1                         0                     0
## 6                       0                         0                     0
##   contains_motorcycle_words
## 1                         0
## 2                         0
## 3                         0
## 4                         0
## 5                         0
## 6                         0
```

(b) Converting Latitude & Longitude to Locations in R

Since the dataset only contained crash coordinates (latitude and longitude), I used the tmaptools library in R to perform reverse geocoding and obtain descriptive location names such as roads and counties.

Due to API request limits, the conversion was done in batches of 3,000 rows, resulting in 10 separate data frames with the extracted location information.

Note: The reverse geocoding process is time-consuming, so the code should not be re-run; instead, the extracted results can be viewed directly using the 'view()' function

```r
# geo for 1-3000
#location <- rev_geocode_OSM(df$longitude[1:3000], df$latitude[1:3000])
view(location)
# ... etc

colnames(location)
```

```
## NULL
```

(c) Merging All Location Data-frames

After getting all the 10 data-frames, I had to put them all together to form 1 data-frame

```
#To merge all the data-frames from location to location 10 into one data-frame
#all_dfs <- list(location, location2, location3, location4, location5,
#                location6, location7, location8, location9, location10)
#combined_df <- bind_rows(all_dfs)

#combined_df
```

Following, had to make one data-frame on the location and Original data-frame. For location had to extract the name, city, city district, road, state, suburb since they are the most important columns

(d) First to add ID row number to both dfs (The original data-frame and the new data-frame with exact locations) so as to join them easily

```
combined_df_new <- combined_df %>% mutate(id = row_number())
view(combined_df_new)

df_new <- df %>% mutate(id = row_number())
view(df_new)

colnames(combined_df_new)
colnames(df_new)
```

— Select desired columns from each dataframe to make into one dataframe

```
combined_df_new2 <- combined_df_new %>% select(id, name, city, city_district,
                                               road, state, suburb)
df_new2 <- df_new %>% select(id,crash_datetime, crash_date, n_crash_reports, contains_fatality_words,
                             contains_matatu_words, contains_pedestrian_words,
                             contains_motorcycle_words)
```

— Merge both dataframes using ID column

```
final_df <- inner_join(combined_df_new2, df_new2, by = "id")
view(final_df)

To export file
write_csv(final_df, "final_df.csv")
```

(e) Datetime Formatting in Excel

Imported final_df.csv into Excel

Reformatted crash_date_time to store time only for easier time-of-day analysis.

Exported as final_accidents_main.csv

Assumptions

- The dataset doesn't provide all the 47 counties in Kenya only those counties around Nairobi where the major roads are.

- Location information obtained via reverse geocoding may not be 100% accurate for all points due to API limitations.
- The presence of "contains_matatu_words," "contains_motorcycle_words," and "contains_pedestrian_words" was used as a proxy for vehicle involvement. This assumes that crash descriptions are reliable and consistently recorded.

4. Analyze

The analysis phase aimed to uncover trends and patterns from the cleaned accident dataset using **R**.

```
accidents <- read_excel("final_accidents_main.xlsx")
```

4.1 Temporal Analysis

(a) First analysis was to investigate the top timings prone to the accidents and categorise the times during the day and night

```
# Parse crash_time and extract hour
accidents <- accidents %>%
  mutate(
    crash_time = parse_time(crash_time),  # convert to time object
    hour = hour(crash_time)               # extract hour
  )

# Categorize hours into time periods
accidents <- accidents %>%
  mutate(time_period = case_when(
    hour >= 0 & hour < 1  ~ "Midnight (12 AM - 1 AM)",
    hour >= 1 & hour < 6  ~ "Early Morning (1 AM - 5:59 AM)",
    hour >= 6 & hour < 12 ~ "Morning (6 AM - 11:59 AM)",
    hour >= 12 & hour < 17 ~ "Afternoon (12 PM - 4:59 PM)",
    hour >= 17 & hour < 19 ~ "Evening (5 PM - 6:59 PM)",
    TRUE ~ "Night (7 PM - 11:59 PM)"
  ))

# Count accidents per time period
time_period_accidents <- accidents %>%
  count(time_period) %>%
  arrange(desc(n))

time_period_accidents
```

```
## # A tibble: 6 x 2
##   time_period                        n
##   <chr>                          <int>
## 1 Morning (6 AM - 11:59 AM)      12081
## 2 Afternoon (12 PM - 4:59 PM)     6938
## 3 Night (7 PM - 11:59 PM)         6182
## 4 Evening (5 PM - 6:59 PM)        4391
## 5 Early Morning (1 AM - 5:59 AM)  1201
## 6 Midnight (12 AM - 1 AM)          271
```

(b) Daily and monthly patterns to highlight the seasonal trends and peak accident days

```r
daily_accidents <- accidents %>%
  count(day_of_week) %>%
  arrange(desc(n))

daily_accidents
```

```
## # A tibble: 7 x 2
##   day_of_week     n
##   <chr>       <int>
## 1 Sat          4560
## 2 Mon          4541
## 3 Thu          4475
## 4 Sun          4428
## 5 Tue          4400
## 6 Fri          4350
## 7 Wed          4310
```

```r
monthly_accidents <- accidents %>%
  count(month) %>%
  arrange(desc(n))

monthly_accidents
```

```
## # A tibble: 12 x 2
##    month     n
##    <chr> <int>
##  1 Mar    2887
##  2 May    2744
##  3 Nov    2696
##  4 Jul    2672
##  5 Dec    2622
##  6 Oct    2571
##  7 Feb    2566
##  8 Jun    2556
##  9 Aug    2528
## 10 Jan    2455
## 11 Sep    2416
## 12 Apr    2351
```

4.2 Geographic Analysis

Top counties, cities and roads with the highest accident counts were identified.

```r
top_counties <- accidents %>%
  count(county) %>%
  arrange(desc(n)) %>%
  head(10)

top_counties
```

```
## # A tibble: 10 x 2
##    county        n
##    <chr>     <int>
##  1 Nairobi   23432
##  2 Kiambu     4700
##  3 Machakos   1805
##  4 Kajiado     645
##  5 Murang`a    309
##  6 Nakuru       55
##  7 Kirinyaga    47
##  8 Makueni      36
##  9 Nyandarua    16
## 10 Nyeri         7
```

```r
top_roads <- accidents %>%
  drop_na(road) %>%
  count(road) %>%
  arrange(desc(n)) %>%
  head(10)

top_roads
```

```
## # A tibble: 10 x 2
##    road                   n
##    <chr>              <int>
##  1 Thika Road          4057
##  2 NA                  3358
##  3 Mombasa Road        1563
##  4 Nairobi Expressway  1415
##  5 Waiyaki Way         1347
##  6 Ngong Road           989
##  7 Langata Road         958
##  8 Jogoo Road           797
##  9 Airport North Road   537
## 10 Outer Ring Road      503
```

4.3 Accident Characteristics

Fatality analysis: Computed the fatality rate per county.

Vehicle involvement: Identified accident counts involving motorcycles, matatus, and pedestrians.

County–vehicle prone mapping: Mapped counties to the types of vehicles most involved.

```r
fatality_analysis <- accidents %>%
  group_by(county) %>%
  summarise(total_accidents = n(),
            fatal_accidents = sum(contains_fatality_words),
            fatality_rate = fatal_accidents / total_accidents) %>%
  arrange(desc(fatality_rate))

fatality_analysis
```

```
## # A tibble: 13 x 4
##    county    total_accidents fatal_accidents fatality_rate
##    <chr>               <int>           <dbl>         <dbl>
##  1 Makueni                36              10        0.278
##  2 Kirinyaga              47              10        0.213
##  3 Kitui                   5               1        0.2
##  4 Murang`a              309              30        0.0971
##  5 Machakos             1805             172        0.0953
##  6 Kiambu               4700             381        0.0811
##  7 Kajiado               645              49        0.0760
##  8 Nairobi             23432            1629        0.0695
##  9 Nyandarua              16               1        0.0625
## 10 Nakuru                 55               1        0.0182
## 11 Embu                    1               0        0
## 12 Narok                   6               0        0
## 13 Nyeri                   7               0        0
```

Key outputs from this analysis:

- Most accidents occurred on Mondays and Saturdays.
- The most accidents occured during the morning hours from 6am - 12pm
- Nairobi County recorded the highest number of crashes.
- Thika Road and Mombasa Road were the most accident-prone roads.
- Matatus had the most involvement in the accidents.

5. Share & Act

Exported the file to create visuals and provide recommendations using tableau

```r
#TO visualize in Tableau we have to Report file
write_csv(accidents, "final_roadaccidents.csv")
```

Interactive Tableau Story

You can view the interactive dashboard here:
Click to open Tableau Story

Recommendations

1. **Enforcement & Awareness**
   - Deploy more traffic enforcement and road safety campaigns during high-risk times (morning rush hours, weekends).

2. **Targeted Interventions**
   - Nairobi and Kiambu should receive prioritized interventions, given their high accident counts.