Contents

1. Abstract

Train is the most common mode of travel for long distance journey in India. This year India has seen major accidents in railways due to several factors including natural causes and human error. We have tried to study the causes of train accidents for year 2002-2017, and other factors which trigger train accidents in different parts of the country. This report includes major accidents till 2017 which impacted the customer experience or say made it unsafe to travel in train.

Keywords: Train accidents; inferential statistics; data visualisation; ggplot2

2. Introduction

Data was collected from various news websites, online news archives, Wikipedia articles [1], etc. Data was available on the internet from year 1890-2017 but analysis is carried out for 2002- 2017 because news reports on many old incidents needed citations and some of them were reported in less detail on news websites. To cover major features affecting the accident rate in India and to closely compare them, 2002-2017 year range provided enough observations (114) to make some statistical conclusion. First, let's perform some preliminary steps of data wrangling to convert our data to tidy form.

3. Collecting data

Data was collected from online news archives [2] and converted to tidy form in RStudio. RStudio is a free and open-source integrated development environment for R, a programming language for statistical computing and graphics. Here is a preview of top few rows of data set:

```
##   X                          environment         train_name injured killed
## 1 1 open rail track|over 2000 attackers  Sabarmati Express      43     58
## 2 2                     open rail track Shramjeevi Express      80     12
##                                                              triggering_factor
## 1 attack by a mob at Godhra station in Gujarat and four coaches were set on fire
## 2        crash occurred when sabotage derailed the Shramjeevi express at Jaunpur
##               time railway_division  cause  env
## 1 27-02-2002 08:30                w attack open
## 2 13-05-2002 03:00                n attack open
```

In order to generate useful results this dataset needs to be in structured format. `Environment` column contains entries which are in sentence form. First we will check for data types and then use some of the descriptive columns to create factor variables.

4. Data pre-processing

Checking for data types:

```
## 'data.frame':    114 obs. of  10 variables:
##  $ X               : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ environment     : Factor w/ 44 levels "","an old weld gave away due to the weight of the power ca
##  $ train_name      : Factor w/ 98 levels " Varanasi-Allahabad Passenger",..: 81 86 56 41 27 65 48 80
##  $ injured         : int  43 80 29 150 15 62 50 50 100 0 ...
##  $ killed          : int  58 12 49 140 36 14 37 13 16 0 ...
##  $ triggering_factor: Factor w/ 88 levels "","  crash occurred when sabotage derailed the Shramjeevi e
##  $ time            : Factor w/ 114 levels "01-01-2011 06:00",..: 100 45 17 33 53 58 51 102 15 95 ..
```

```
##  $ railway_division : Factor w/ 16 levels "c","e","ec","eco",..: 15 6 8 3 6 5 6 6 7 14 ...
##  $ cause            : Factor w/ 7 levels "attack","fire",..: 1 1 6 1 2 6 3 1 5 4 ...
##  $ env              : Factor w/ 10 levels "bad track","busy",..: 5 5 2 10 5 4 5 5 5 4 ...
```

Data types for all variables are correct except `time` column. `Time` column needs to be a datetime object. Converting time column to datetime object:

```
tr$time <- as.POSIXlt(tr$time, format = "%d-%m-%Y %H:%M")
```

Verifying:

```
## [1] "POSIXlt" "POSIXt"
```

5. Feature engineering

Data type correction being done, next we will create some new columns from the existing ones to create meaningful visualisations.

5.1. Hour, daytime and season

`Time` column will be helpful in creating these columns:

5.1.1. Hour

Creating hour column:

```
tr$hour <- hour(tr$time)
```

For comparing accident trends over a single day, only the hour column is not sufficient. It will produce numerical values with 24 levels to consider for each day.

It is better to divide each day in interval of 6 hours to better compare accidents over a single day. These time intervals are chosen to see the effect of visibility on accident.

5.1.2. Daytime

Creating daytime column:

```
tr$daytime <- cut(tr$hour, c(-1,6,12,18,23))
levels(tr$daytime)[1] <- "[0,6]"
levels(tr$daytime)[4] <- "(18,24)"
levels(tr$daytime)
```

```
## [1] "[0,6]"   "(6,12]"  "(12,18]" "(18,24)"
```

This feature will be helpful in studying the daily trends in accident.

5.1.3. Season

Creating season column:

```
for(i in c(1:length(tr$time))){
    if(month(tr$time[i]) %in% c(4:6)){
        tr$season[i] <- "summer"
    }
    else if(month(tr$time[i]) %in% c(7:9)){
        tr$season[i] <- "monsoon"
    }
    else if(month(tr$time[i]) %in% c(10:11)){
        tr$season[i] <- "autumn"
    }
    else{
        tr$season[i] <- "winter"
    }
}

# converting to factor variable
tr$season <- factor(tr$season)
levels(tr$season)
```

```
## [1] "autumn"  "monsoon" "summer"  "winter"
```

The month value used as boundary condition for any two consecutive season just for convenience to create `season` column effectively with code. These intervals are taken from Wikipedia article [3] on **Climate of India**.

5.2. Type of train

`train_name` column is just a label and will not help in studying accidents. It will be difficult to study number of accidents for each and every train. Instead, we will create a new column which will show type of train involved in accident, like express, passenger, freight, etc.

Creating `train_type` column:

```
for(i in c(1:length(tr$train_name))){
    for(j in c("Express", "Passenger", "Freight", "Mail", "others")){
        if(grepl(j, as.character(tr$train_name[i]))){
            tr$train_type[i]<-as.character(strapply(as.character(tr$train_name[i]), j))
        }
    }
}

# converting to factor variable
tr$train_type <- factor(tr$train_type)
levels(tr$train_type)
```

```
## [1] "Express"   "Freight"   "Mail"      "others"    "Passenger"
```

5.3. Country region

Now we have `railway_division` column. 17 values for railway division column are too much and will make our graph look messy. Categorising them into North, South, East, West and Central will make our study simple. For example, n, nc, ne, nef, nw (abbreviations are provided at the end of report) will go in North area, and similarly for other regions.

Creating **region** column:

```
for(i in c(1:length(tr$railway_division))){
    if(tr$railway_division[i] %in% c('n', 'nc', 'ne', 'nef', 'nw')){
        tr$region[i] <- 'North'
    }
    else if(tr$railway_division[i] %in% c('s', 'sc', 'se', 'sw')){
        tr$region[i] <- 'South'
    }
    else if(tr$railway_division[i] %in% c('e', 'ec', 'eco')){
        tr$region[i] <- 'East'
    }
    else if(tr$railway_division[i] %in% c('c')){
        tr$region[i] <- 'Central'
    }
    else{
        tr$region[i] <- 'West'
    }
}

# converting to factor variable
tr$region <- factor(tr$region)
levels(tr$region)
```

```
## [1] "Central" "East"    "North"   "South"   "West"
```

6. Feature selection

Next part is including relevant features for visualisation purpose. We have 15 variables in our data.

```
colnames(tr)
```

```
##  [1] "X"                 "environment"       "train_name"
##  [4] "injured"           "killed"            "triggering_factor"
##  [7] "time"              "railway_division"  "cause"
## [10] "env"               "hour"              "daytime"
## [13] "season"            "train_type"        "region"
```

```
## [1] 15
```

We are removing columns which are not important for analysis. We are removing these columns due to the following reasons:

- they are not in a structured format.
- some of them were in complex form and new derived columns are used instead.

The columns which are to be removed are:

6.1. Environment

`Environment` column is descriptive in nature, `env` column is more specific in describing the environment of accident spot, like open area, level-crossing, etc.

6.2. Train_name

Too many distinct values are there for this variable. This was used to create `train_type` column previously.

6.3. Triggering_factor

This one is also descriptive in nature and `cause` column shows the specific cause, like human error, technical fault, etc.

Removing the above mentioned columns:

```
tr$environment <- NULL
tr$triggering_factor <- NULL
tr$train_name <- NULL
```

We have done these tasks till now :

- Correcting data types
- Creating useful columns from the given ones
- Removing columns that are not important for analysis

Next part is generating bivariate and multivariate plots to derive a conclusion about year by year trends in train accidents.

7. Exploratory data analysis

After correcting data types and performing feature extraction on the data set, next part is creating visualisations to get insights about what are the major causes of accidents, where they are more frequent and many more features which show different frequency of accidents.

We will use this data to answer the following questions:

Q1. On which time of day more accidents are happening? Does visibility play a major role in accidents?

Q2. In which season accidents are more frequent?

Q3. Which region of the country has experienced more accidents?

Q4. What is the environment of accident spot (like open area, flood affected area or bad track?

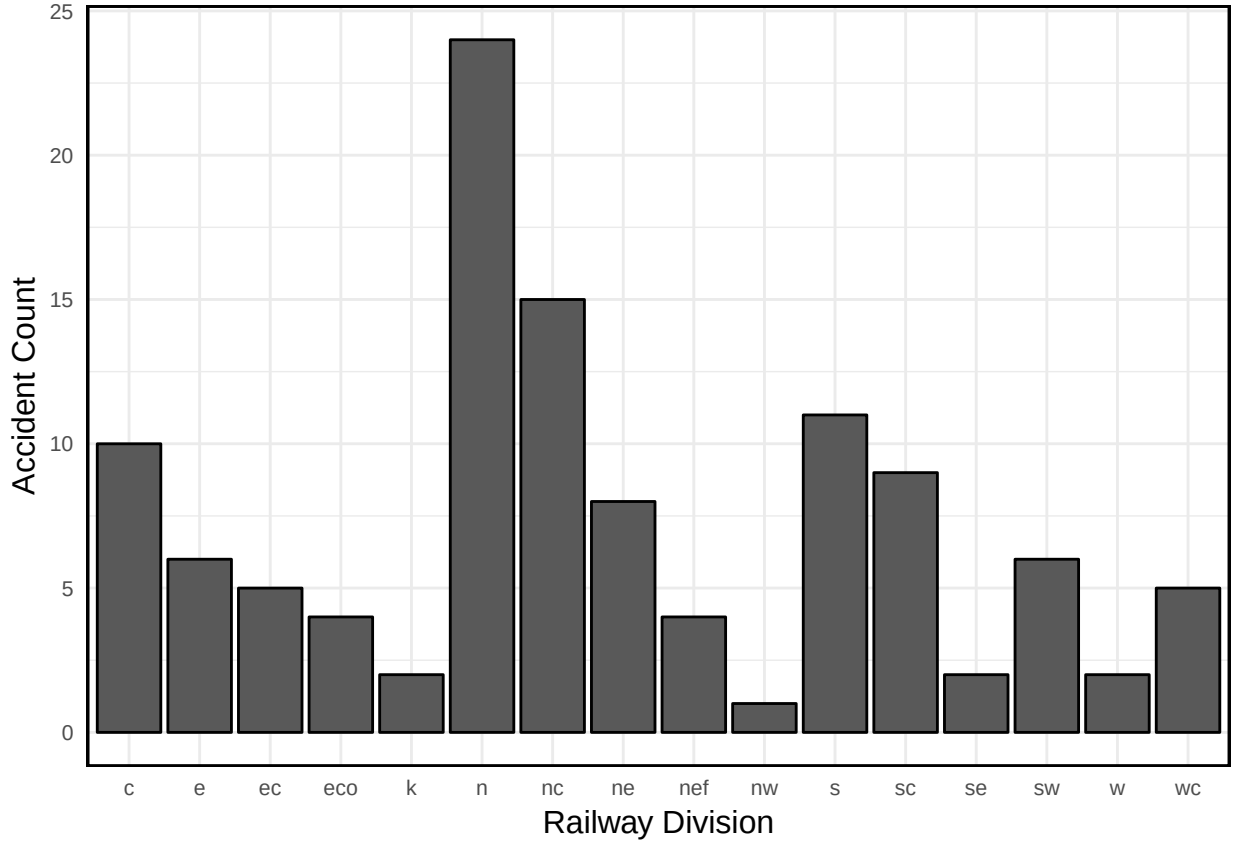
Q5. Which type of trains suffer more accidents(express, passenger, freight, etc.)?

Q6. What are the major causes for accidents? Is there any significant difference across different causes? Most of the factors result in derailment of train as we see in news reports, but here we will try to investigate the cause for these derailments. Whether it is human error, natural cause or technical glitch.

We have 6 features in our data which will help in answering these questions. First we will see the distribution of accident count based on each feature and then will include more features to better understand the trends.

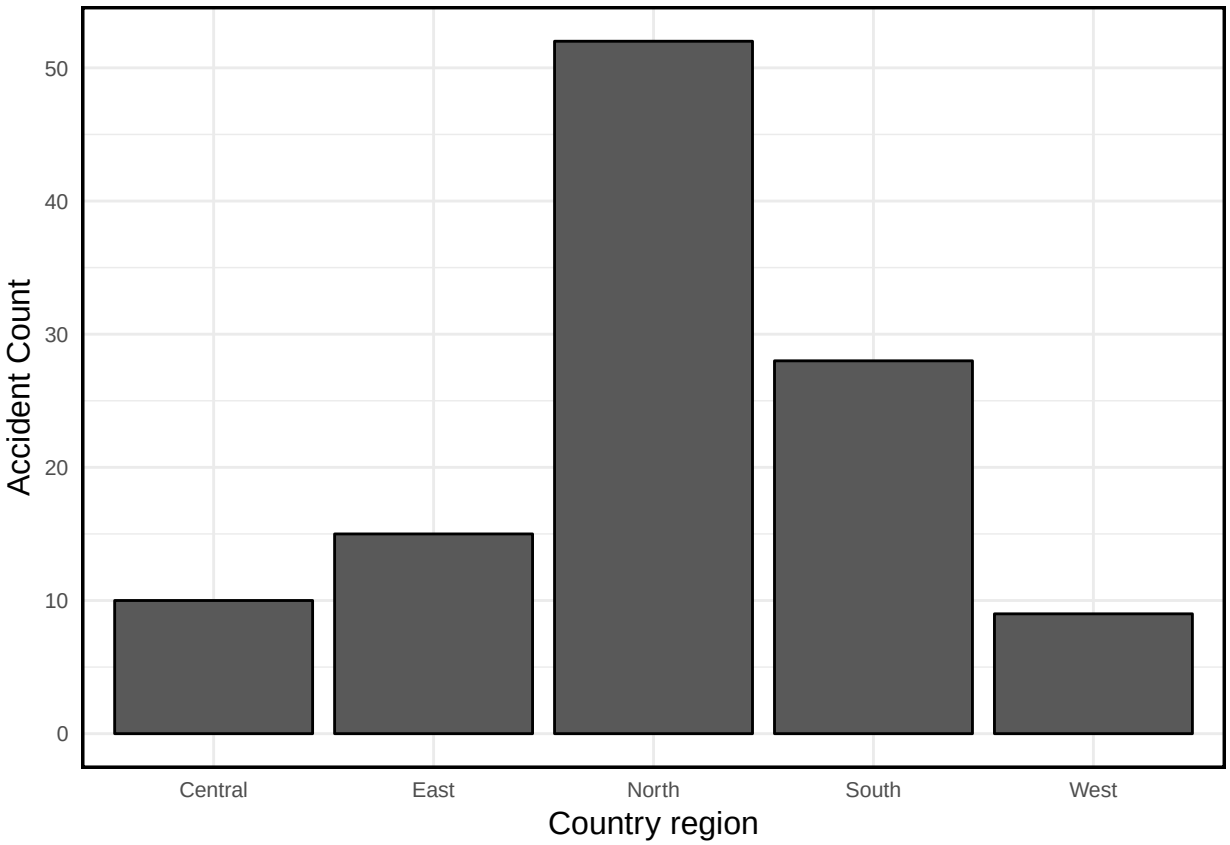
We have 6 features on which we can categorise our analysis:

7.1. Region



The distribution based on railway division shows peaks at regions which are more general (say larger in area), like if we consider **n, nc, ne, nef** and **nw** then they collectively represent accidents that occur in North India. Some of the accidents are hard to categorize based on such granular distinction.

To tackle this problem, we will see accidents in different parts of India, i.e., North, East, West, South and Central to classify them more clearly with less overlapping regions.

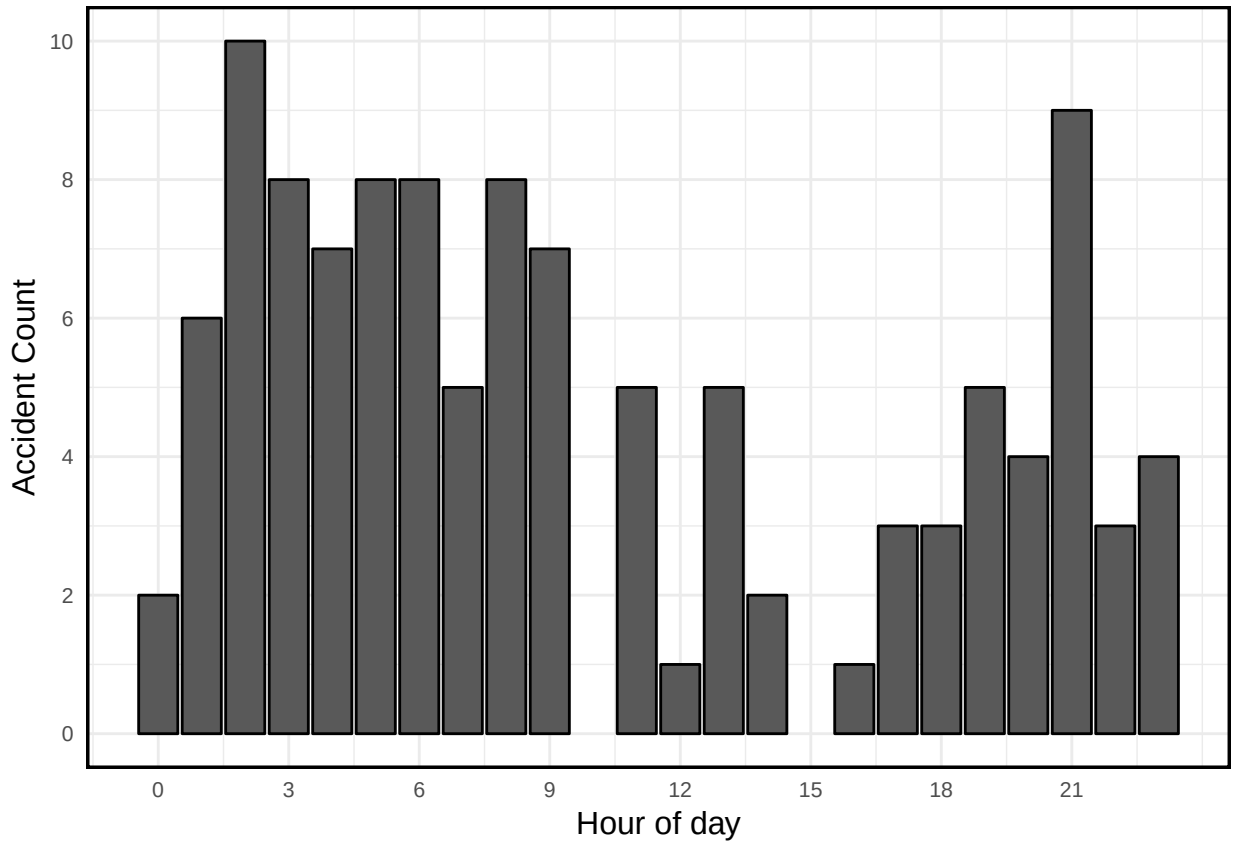


Accident count in South is almost half of that in North region. In North region, the state which contributed largest to the accident count is U.P.. Accidents in U.P. are so frequent that there is an entire Wikipedia article [4] on it.

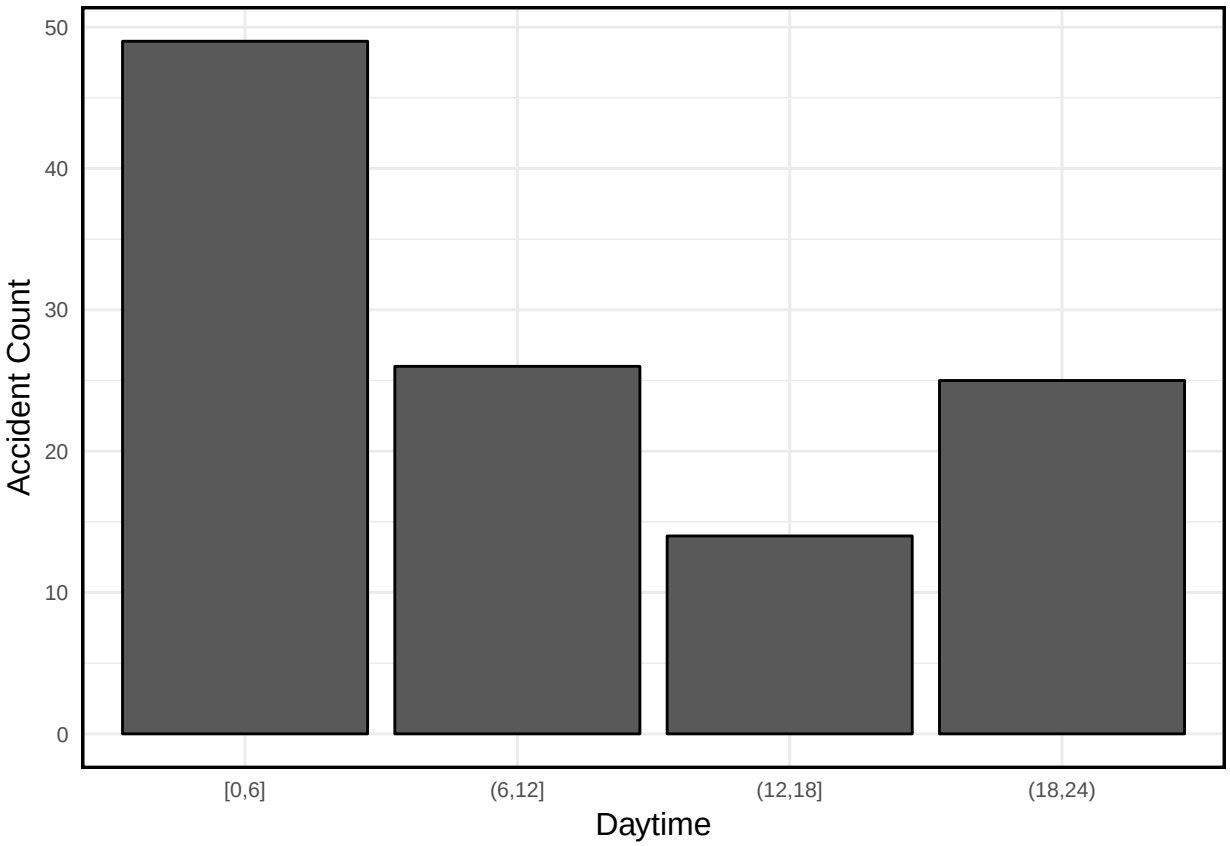
7.2. Daytime

Another column which has many levels is **hour** column.

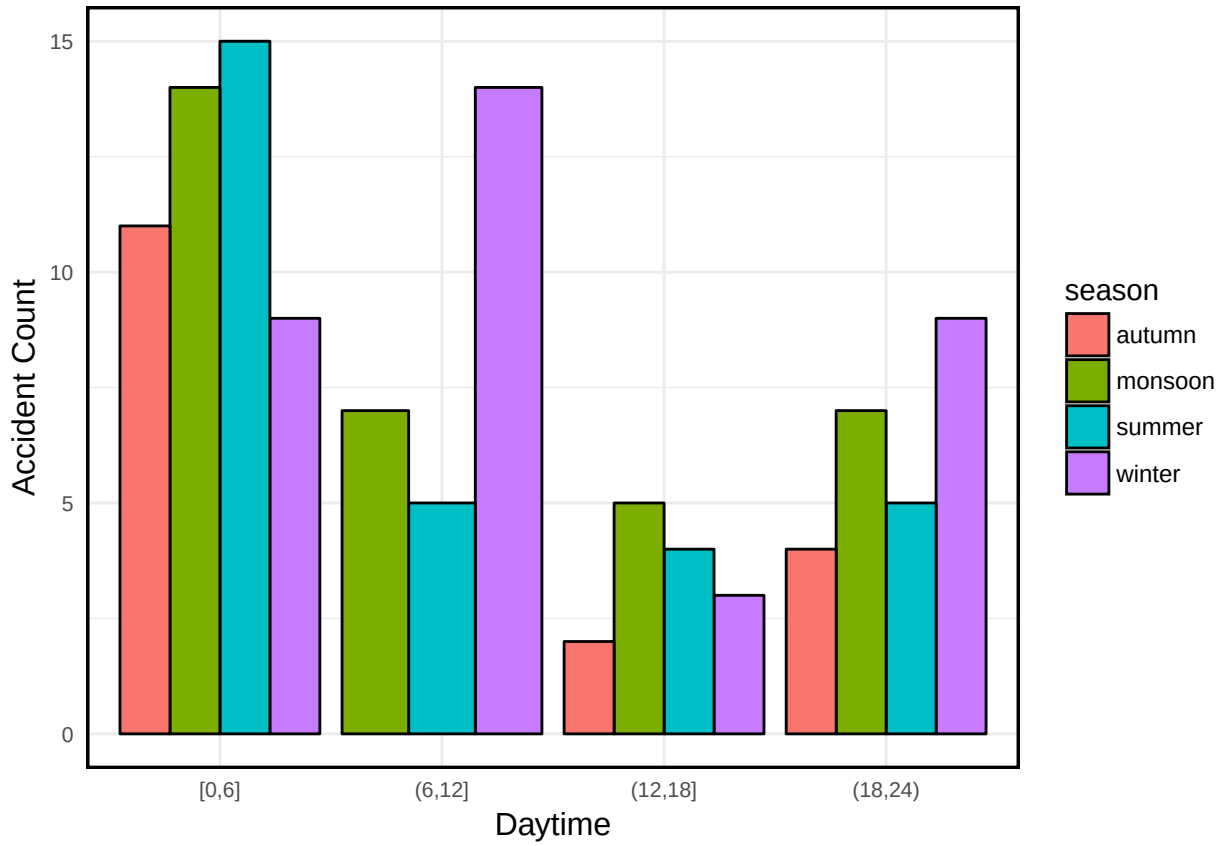
Let us see the trends based on hour and then we will use **daytime** column for considering visibility effect on accidents.



This graph shows that most of the accidents take place when there is low visibility, i.e. night time or wee hours. We will not go deep into exact hour at which accident took place because the sources which provided this information gave an estimate of accident timings. That is why 10:00 hrs and 15:00 hrs have no count. Some accidents took place around evening so we put them in 17-18 hrs range, and similar is the case for 10:00 hrs. We are not interested in exact time of accident, for this purpose we have created **daytime** column previously to compare trends which shows effect of visibility based on hour of accident.

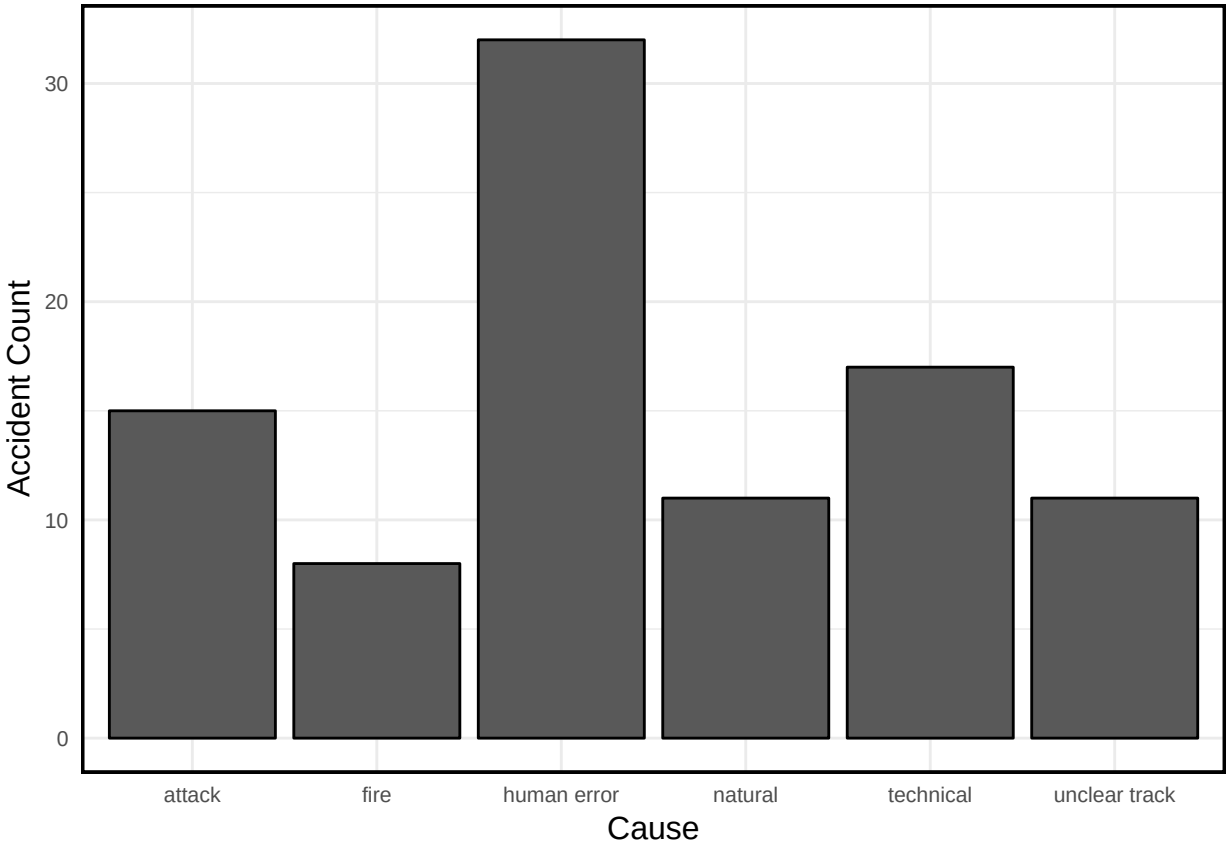


12:00 to 18:00 hours is the time when less accidents happen. Visibility factor can be inferred from this relation. Accidents are more likely to occur when there is low visibility. Most of the accidents occur during night time, as seen from the data. Day time has also considerable amount of acccidents but on the basis of visibility criteria it is almost half compared to night time. Increasing security during night time can help in preventing these mishaps.



Daytime has second larget number of accidents because of accident in day in winter season.

7.3. Cause



Let us see what each value on x-axis stands for:

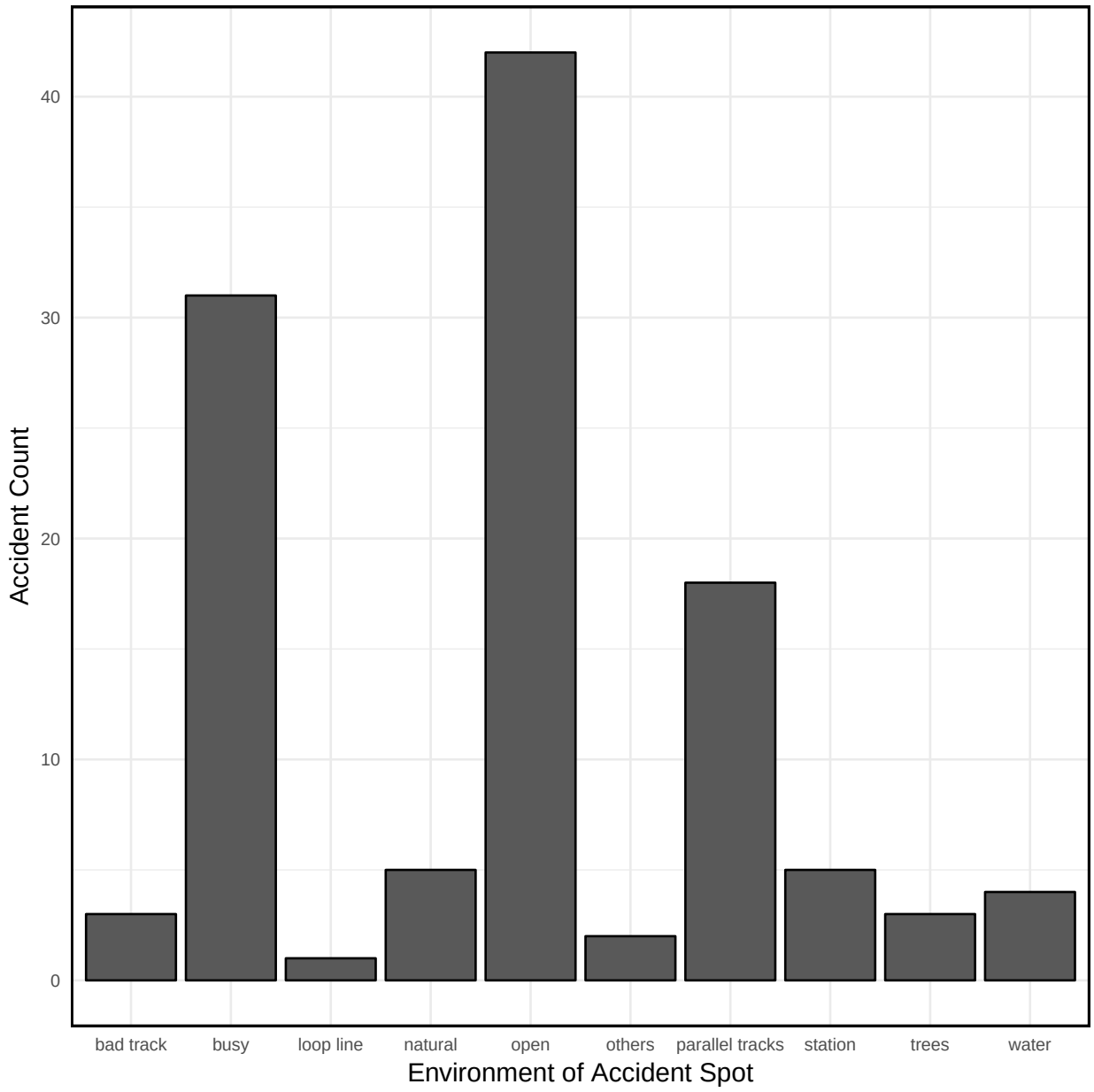
- **attack**: this includes all external forces like:
  - bombing by terrorists;
  - attack by local people;
  - sabotage;
  - one case is of hijack [5].
- **fire**: fire caught inside train led to few accidents. This is the least frequent cause of accident because it was caused when passengers unintentionally carried some inflammable item with them. Electrical short circuits also caused fire.
- **human error**: this includes human negligence at large scale. This includes:
  - improper signalling
  - speeding by driver
  - unmanned level crossing
  - negligence about train timings on level-crossing
  - one incident includes people watching sample fireworks while standing on railway track [6]
- **natural**: these are beyond one's control. Natural causes are:
  - heavy rain
  - flash flood
  - dense fog
- **technical**: this includes malfunctioning of railway system:

  - no prior alert for driver to stop;
  - poor maintenance of track;
  - some tracks are pending for reconstruction [7];
  - brake fail;
  - no alert about poor tracks ahead.

- **unclear track**: technically this should also come under **technical** cause but track was occupied in many cases so we are considering it as a separate factor:
  - boulders on track
  - some vehicle was already on track(includes trains and roadways vehicles)
  - some roadways vehicle stuck on track

Mostly human error is reponsible for train accidents because they are in some way directly related to the short-comings in system. Mistakes like improper signalling, poor maintenance of track are very sensitive areas when it comes to railway security. In some cases train driver overshot the red signal. These mistakes should not be tolerated at such large scale. Speeding can be easily avoided if there is proper timing for crossings and loop-line entry.

7.4. Environment



There is no perfect conclusion from this graph but if we see the environment which has most number of accidents then open area, busy track, and parallel track have high frequency of accident in the last 16 years.

Let us see what these features mean:

- Open area:

This includes tracks which were away from residential areas and surrounded by non-agricultural land.

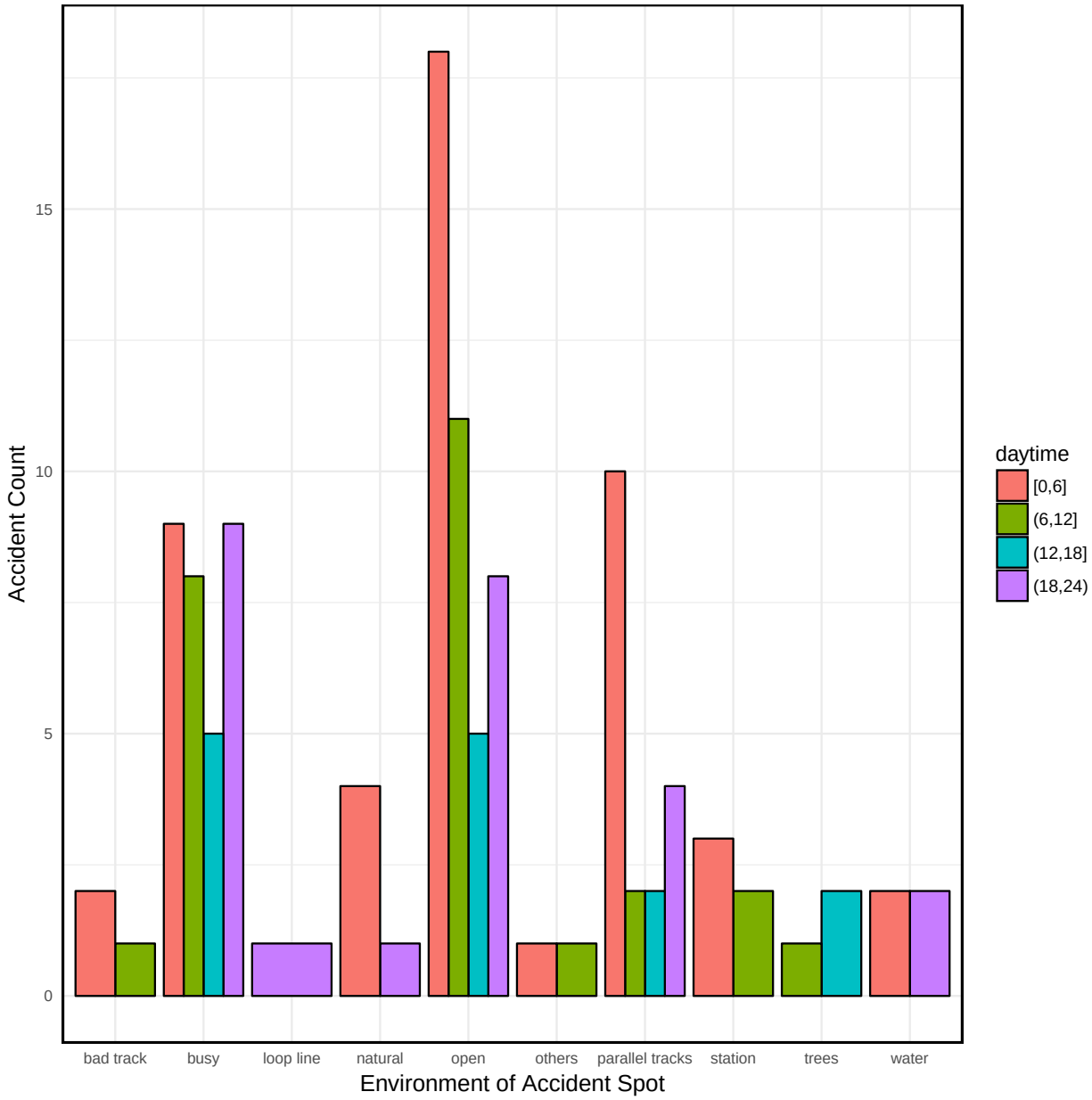
- Busy track:

It includes all the cases when the track was occupied, whether it is roadways vehicle, obstacle, another train, and humans[6].

- Parallel tracks:

More than one track on the route.

Possible reasons why there are more accident can be seen if we include the daytime factor:



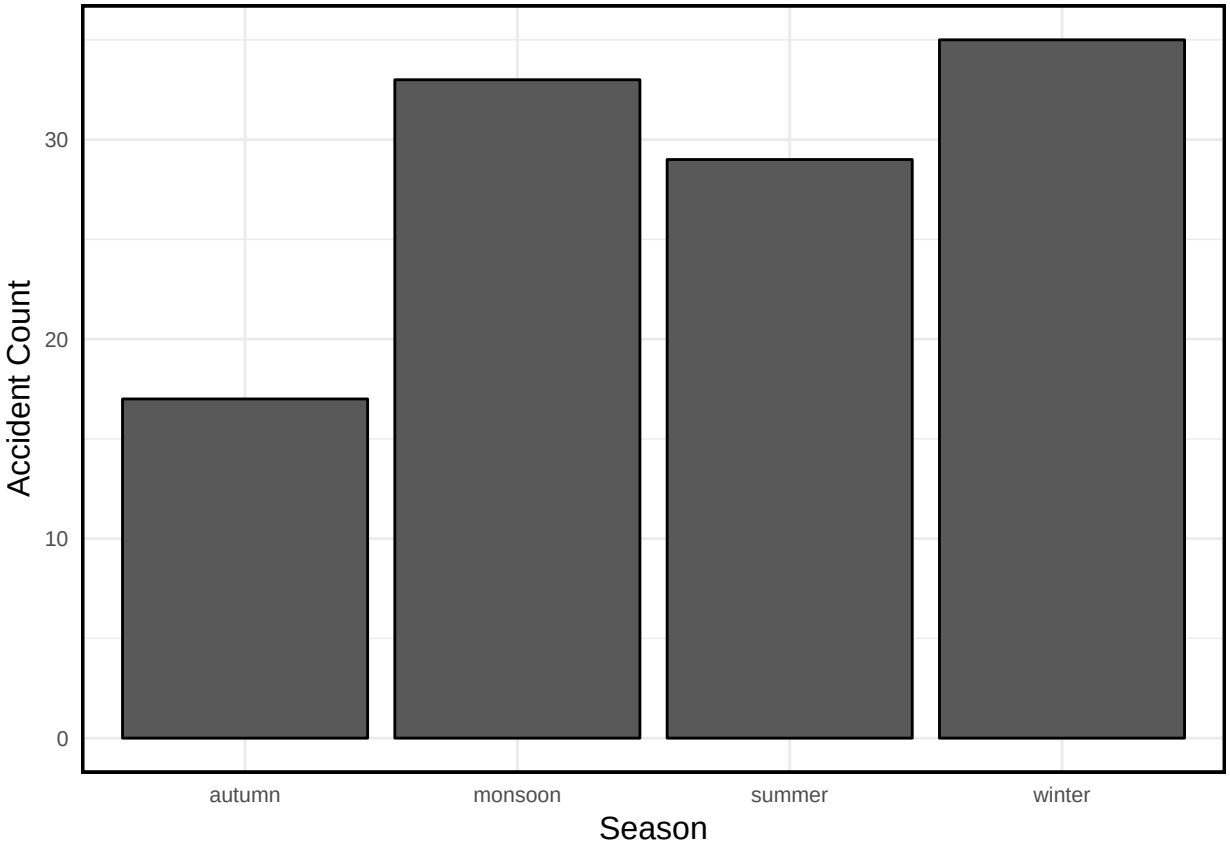
We can clearly see that visibility is a key factor here. Such a large peak for night time in every environment shows how frequently accidents occur in low light. More on this in **season** section.

Let us perform a Chi-squared test of dependence for **environment and daytime**:

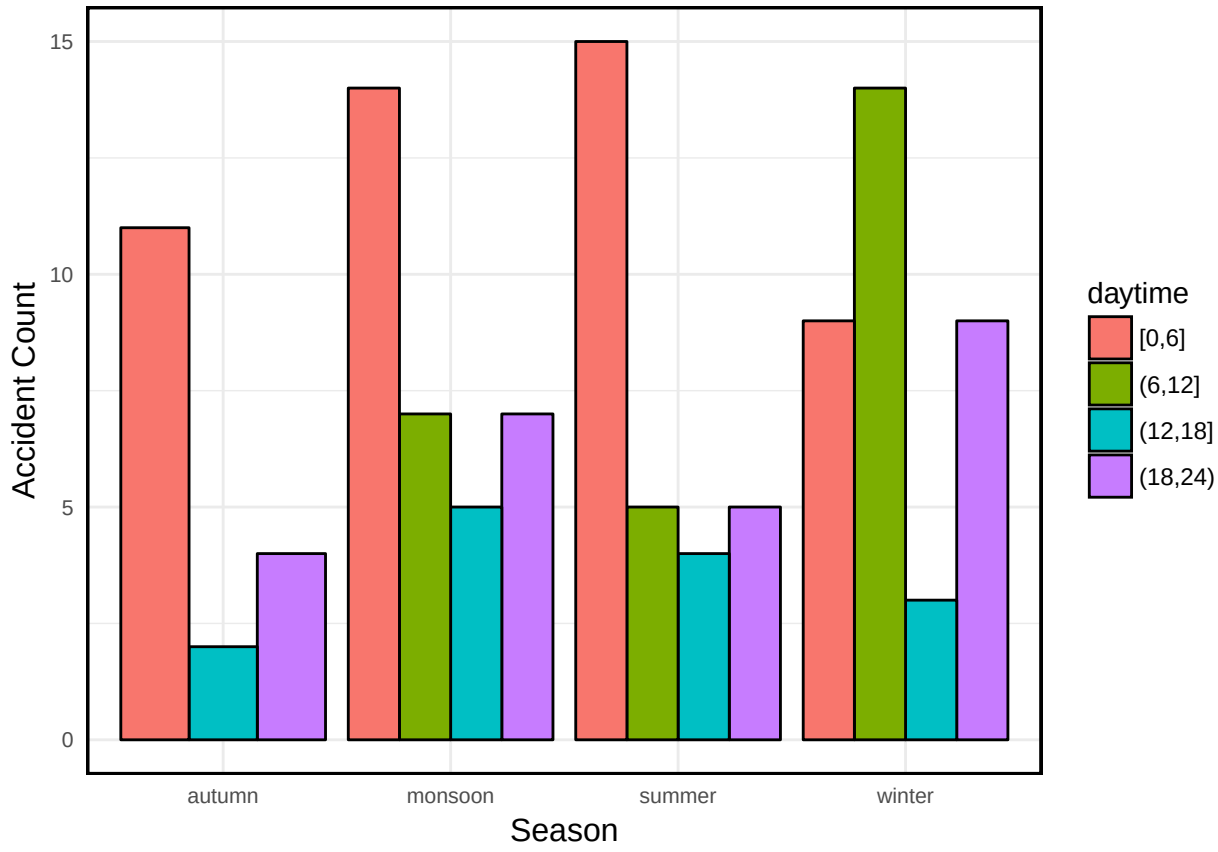
```
## [1] 0.3377004
```

As the p-value is greater than the significance level 0.01, we do not reject the null hypothesis that the environment of accident is independent of daytime. We are not taking significance level of 0.05 because this is a practical situation and 0.05 is a high threshold for significance level.

7.5. Season



These peaks do not directly convey the number of accidents in each season because every season has different time-span. Autumn lasts for shortest period and so count is also less for it. We can include more features in this plot to generalise our idea.



Now let us perform Chi-squared test of dependence for **season and daytime**:

```
## [1] 0.09443166
```

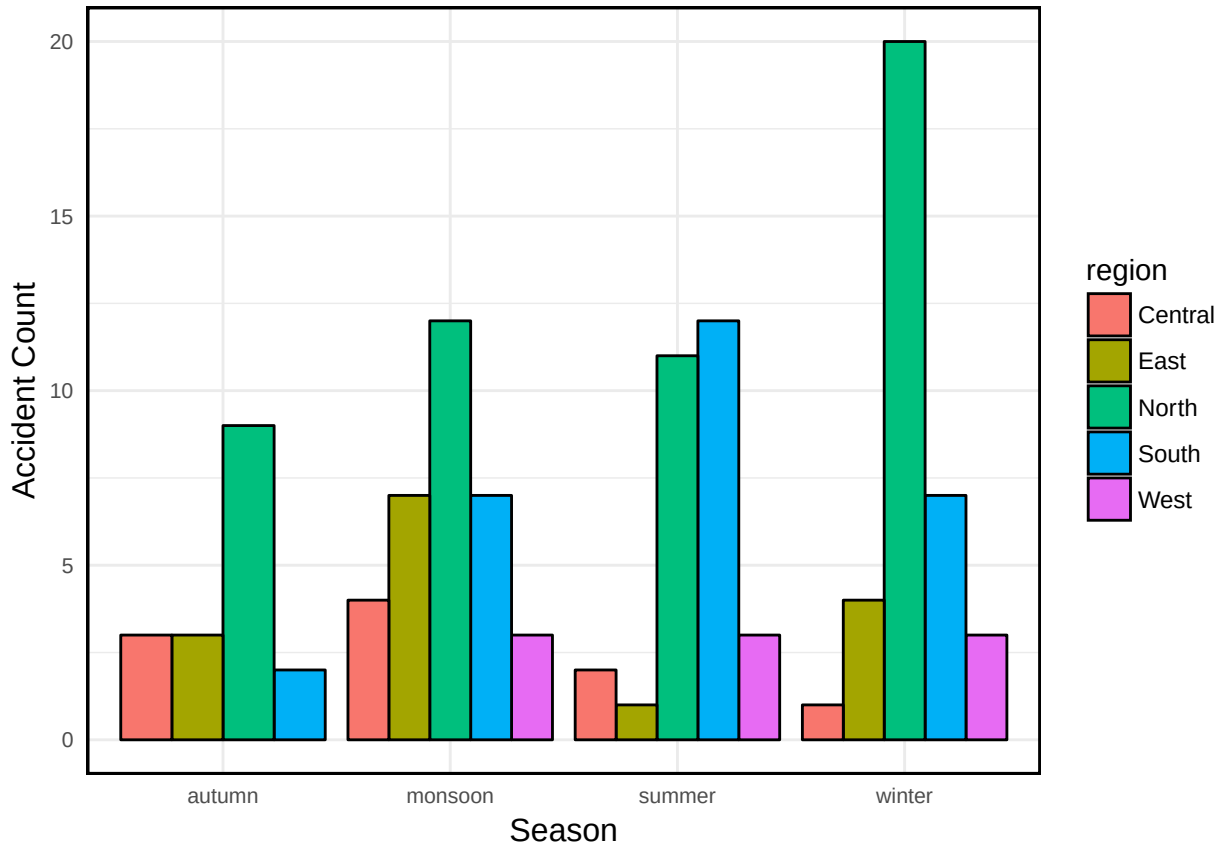
```
100*(test_env_dt$p.value - test_season_dt$p.value)/test_env_dt$p.value
```

```
## [1] 72.03685
```

As the p-value here is also greater than the 0.01 significance level, we do not reject the null hypothesis that season and daytime are independent. But, if we compare it with environment-daytime pair, p-value for season-daytime is 72.04% less than that of environment-daytime pair. This does mean that season-daytime pair are more close to being interdependent pair(or say correlated pair) than environment-daytime pair.

Reflect back on the environment criteria and daytime criteria. Visibility played a major role there for causing accidents.

Here, each season is affected during night time. In winter, however, accidents are still frequent during morning time due to dense fog. In winter, accidents are more in morning than in night time because precautions are taken during night time to avoid accidents due to fog. Though in daytime, some cases are there when dense fog causes visibility problems resulting in accidents.



Let us see perform Chi-squared test of dependence for season and region:

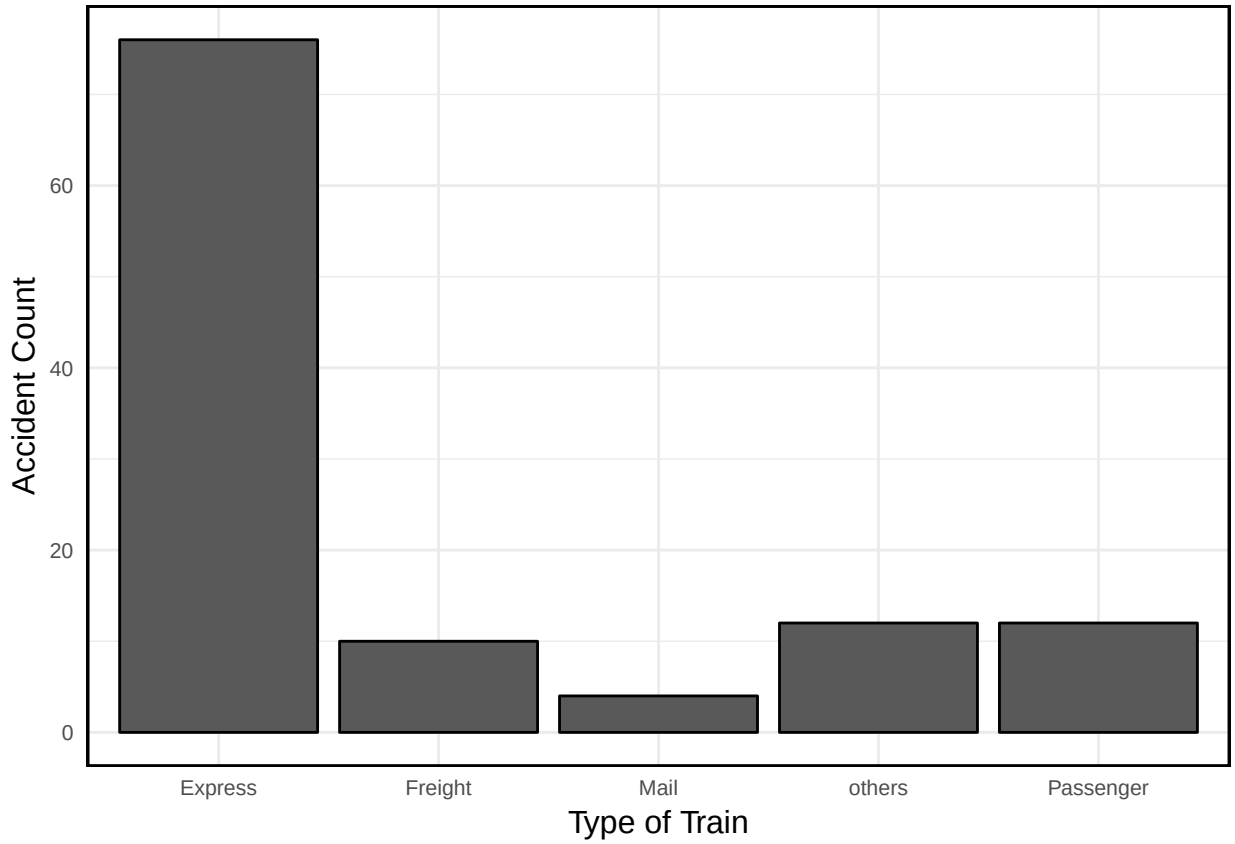
```
## [1] 0.1786312
```

As the p-value for this pair is 0.17 which is higher than the significance level 0.01, we cannot reject the null hypothesis that these two variable are independent of each other.

In South, accidents are more common in summer because of hot climate there. North, on the other hand, has harsh climatic conditions here which is a more severe cause than tropical Southern India. Dense fog in winter affects the timing of train.

Overall, winter sees most number of accidents, reason is decrease in visibility due to dense fog.

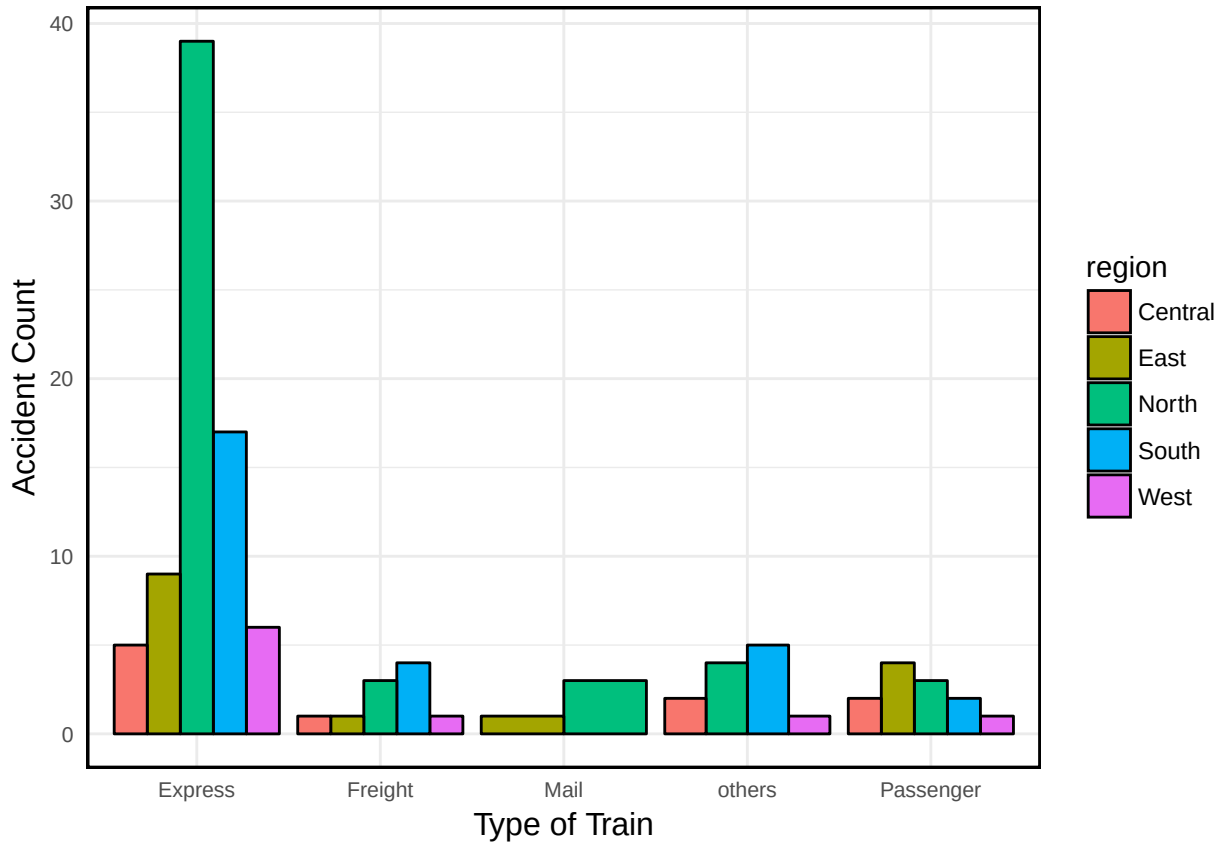
7.6. Train type



This observation is as expected. Major causes are violation of signals and speeding.

Compared to other trains, it is difficult for express and freight trains to lower their speed when they see obstacles in their way. Obstacles can be like:

- boulders on track;
- some vehicle on track; and
- unexpected engagement of roadways vehicle on level-crossing.



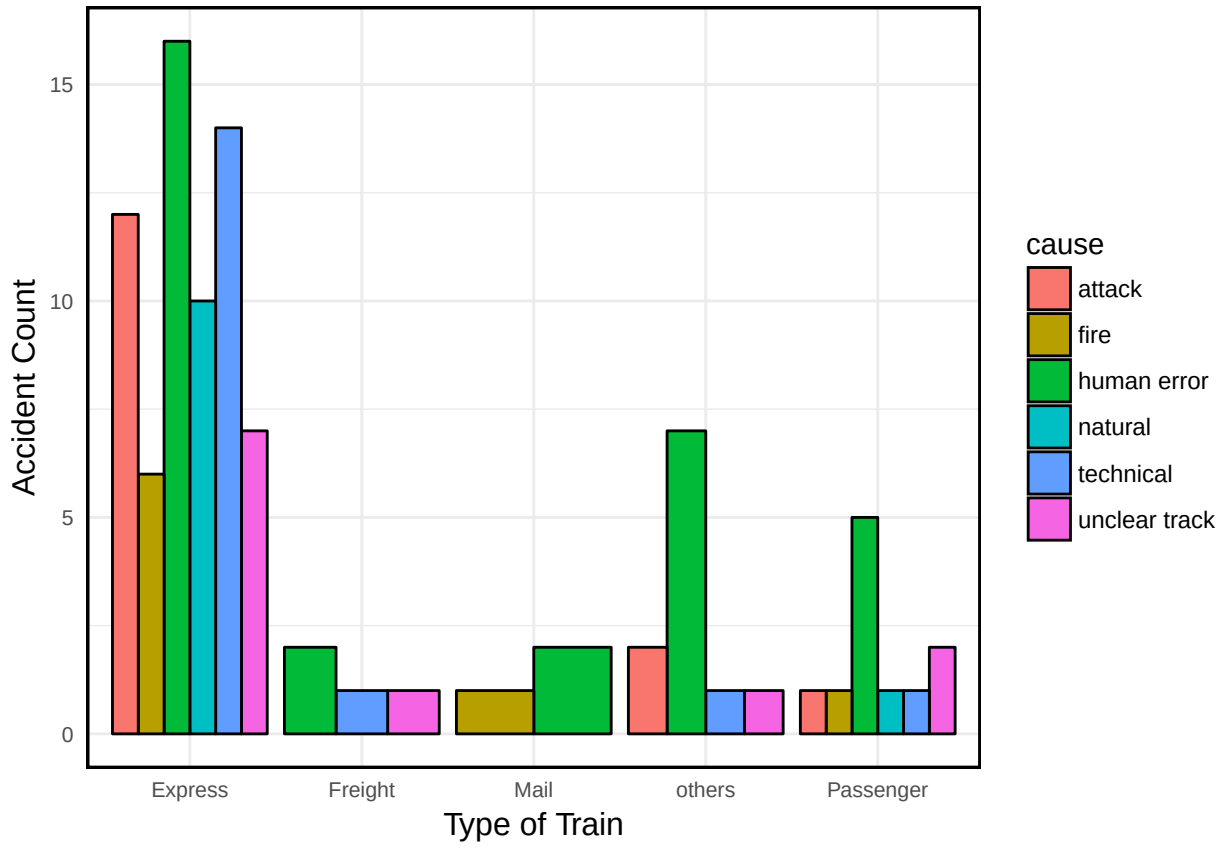
Let us perform Chi-squared test of dependence for train type and region:

```
## [1] 0.4568128
```

No strong relation between train type and region of accident. A p-value of 0.438 means that we cannot reject the null hypothesis that season and train types are independent of each other. We can conclude from this statistic that there is no specific part of India where one type of train is running more frequently.

2 accidents occurred where passenger were sitting on the top of train. This is the result of **dense population** and in order to accommodate facility for increasing population, new railway tracks are laid wherever necessary [8].

On the other side, it is important to maintain the track also. As of 2014 statistics, 5300 km of track was due for renewal [8]. This target keeps increasing if laying of new tracks and maintenance of old tracks are not balanced.



Let us perform Chi-squared test of dependence for train type and cause of accident:

```
## [1] 0.136919
```

```
##
##   Express   Freight      Mail    others Passenger
##        76        10         4        12        12
```

As the p-value is still higher than 0.01 significance level, but if we relax the significance level criteria to 0.1 then we can conclude to some extent that cause and train type are related with each other. But 0.1 level of significance is very high. Chance of accident is high if we are testing for express trains.

High green peaks show that human error is most responsible for almost every type of train accident. Driver's mistake comes under this case but there is no right way to avoid accidents cause by driver's mistake. Some ways can be deploying experienced drivers, changing drivers on shift-basis for long journey trains, checking physical fitness of driver, etc.

Next are technical and external attacks. Attacks on railway system has decreased in recent years but this attack factor counts for sabotage cases also. Technical fault was for speeding and improper signalling, that's why technical error is largest for express trains. Drivers get signals late, some times they are unable to receive signal at right time. Accidents caused by unclear tracks can be avoided by providing information about bad track at right time. People should follow rules when they are near level crossing.

8. Answers

Q1. On which time of day more accidents are happening? Does visibility play a major role in accidents?

A1.

```
##
##   [0,6]  (6,12] (12,18] (18,24)
##      49      26      14      25
```

Accidents are more during night time. Out of 114 observations, accidents in night time (49 + 25 = 74) are greater than that in daytime (= 40). They can be prevented by avoid speeding at night time.

To increase efficiency and to enhance safety in train operations, Advanced Signaling System with Panel/Route Relay/Electronic Interlocking (PI/RRI/EI) along with Multi Aspect Colour Light Signals have been progressively provided at 5,317 stations i.e. about 85% of Broad Gauge stations of Indian Railways, replacing outdated Multi Cabin Mechanical Signaling system involving a large number of human interfaces. Route Relay Interlocking/Electronic Interlocking at 8 major stations namely Barddhaman, Firozpur, Jakhal, Tambaram, Nagda, Gomoh, Agra Cant & Chheoki have been provided during the year 2014-15 [12].

Q2. In which season accidents are more frequent?

A2.

```
##
##  autumn monsoon  summer  winter
##      17      33      29      35
```

Accidents are more frequent in winter season, followed by monsoon. The reason is climatic condition. Heavy rain in monsoon wets railway track, flash floods disrupt train route, dense fog affects movement of train.

Indian Railways is developing an effective accident prevention system to help save lives and increase the efficiency of its service. The fleet of locomotives will be equipped with a 'third eye' to avert train collisions, derailment and accidents on unmanned railway crossings. The technology involves a radar-based device that alerts drivers to any physical obstructions on railway tracks ahead, preventing accidents and breakdowns. This will prove to be more useful at night and in foggy conditions when drivers have to constantly look outside the locomotive to assess weather conditions [13].

The system called 'Terrain Imaging for Diesel Drivers - Infra-red Enhanced Optical and radar assisted' (Tri-netra) has been developed for India by foreign agencies [13].

Q3. Which region of the country has experienced more accidents?

A3.

```
##
## Central    East   North   South    West
##      10      15      52      28       9
```

Northern region has experienced more accidents. This is due to high population density in northern states. As of 2011 Census of India, U.P. alone contributes to 16.5 % of nation's population [9]. Higher population density means more number of railway tracks for accommodation.

Q4. What is the environment of accident spot (like open area, flood affected area or bad track?

A4.

```
##
##       bad track            busy       loop line         natural
##               3              31               1               5
##            open          others parallel tracks         station
##              42               2              18               5
```

```
##           trees           water
##               3               4
```

Open area includes tracks away from agricultural field and residential areas. Accidents are frequent in opens area because there is only one track and fast moving train collides before even changing track or slowing down.

It has been decided to progressively eliminate all unmanned level crossings by (i) closing unmanned crossings having NIL/negligible Train Vehicle Units (TVUs), (ii) merger of unmanned level crossing with nearby unmanned/manned gates or Road Under Bridge or Road Over Bridge or Subway by construction of diversion road, (iii) provision of Subways/Road Under Bridges. The Unmanned Level Crossings which cannot be eliminated by the above means will be progressively manned based on the volume of rail road traffic (TVU) and visibility conditions [14].

Q5. Which type of trains suffer more accidents(express, passenger, freight, etc.)?

A5.

```
##
##   Express   Freight      Mail    others Passenger
##        76        10         4        12        12
```

Express trains suffer more accident because of speeding and signal overshooting. Derailment occurs when taking sharp turn at high speed [10]. Many cases include binary collision of express train with other express/passenger train. Passenger train is usually running at low speed when collision takes place. It is the express train which rams into it [11].

Technological aids of Automatic Train Protection System to loco pilots to avoid collisions due to Signal Passing at Danger (SPAD) or over speeding have been put on trial on Indian Railways [12].

Q6. What are the major causes for accidents? Is there any significant difference across different causes? Most of the factors result in derailment of train as we see in news reports. But we will try to investigate the cause for these derailments. Whether it is human error, natural cause or technical glitch.

A6.

```
##
##        attack          fire   human error       natural     technical
##            15             8            32            11            17
## unclear track       unknown
##            11            20
```

Human error is responsible for most of the accidents. It includes drivers' and signalers' fault. Out of 114 observation collected, 28.07% accidents are caused by human error.

Let us perform some post-hoc tests to see whether some specific cause are more important than others. We will see relation between different cause and their affect on casualties.

Filling null values by median for `killed` column and `injured` column:

```
tr$killed[is.na(tr$killed)] <- with(tr, ave(killed, cause,
                                            FUN = function(x) median(x, na.rm = TRUE)))[is.na(tr$killed

tr$injured[is.na(tr$injured)] <- with(tr, ave(injured, cause,
                                              FUN = function(x) median(x, na.rm = TRUE)))[is.na(tr$injure
```

Creating victim column after handling null values in killed column and injured column:

```
tr$victim <- tr$injured + tr$killed
```

Let us perform one-way ANOVA test for victims column:

```r
v <- aov(victim ~ cause, data = tr)
summary(v)
```

```
##             Df  Sum Sq Mean Sq F value Pr(>F)
## cause        5  131849   26370   1.982 0.0891 .
## Residuals   88 1170943   13306
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As the p-value is greater than 0.05, we cannot reject the null hypothesis that victims is independent of cause.

Perform same tests for killed and injured columns also:

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## cause        5  11444    2289   1.996  0.087 .
## Residuals   88 100909    1147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## cause        5  69288   13858   1.852  0.111
## Residuals   88 658596    7484
```

As the p-value is high for both injured and killed column, we cannot reject the null that they are independent of cause.

We are trying to build a linear model to explain the relationship between killed column and `cause`. Now let us see how much significant each cause is:

```
##
## Call:
## lm(formula = killed ~ cause, data = tr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.467 -14.781 -10.144   7.969 165.533
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           43.467      8.743   4.971 3.25e-06 ***
## causefire            -25.717     14.825  -1.735  0.08630 .
## causehuman error     -28.685     10.596  -2.707  0.00815 **
## causenatural         -28.648     13.442  -2.131  0.03586 *
## causetechnical       -29.996     11.996  -2.501  0.01425 *
## causeunclear track   -34.830     13.442  -2.591  0.01120 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.86 on 88 degrees of freedom
## Multiple R-squared:  0.1019, Adjusted R-squared:  0.05083
## F-statistic: 1.996 on 5 and 88 DF,  p-value: 0.087
```

We can see that human error is the most important cause for accidents because it has the least p-value 0.008. It is very much less than 0.01 significance level. It can be considered an important factor. Excpet fire, other causes are also significant in predicting `killed` variable. Fire is unexpected case so it is not so much important.

Train Protection and Warning System (TPWS) is a proven Automatic Train Protection System to avoid

train accidents on account of human error of Signal Passing at Danger (SPAD) or over-speeding. As a pilot project, TPWS has been provided on Chennai-Gummidipundi Suburban Section of Southern Railway, Hazrat Nizamuddin - Agra Section of Northern/North Central Railway and Dum Dum-Kavi Subhash section of Kolkata Metro [14].

One way ANOVA showed that `killed` is independent of `cause`. Now let us finally conduct Tukey Honest Significance test:

```
a1 <- aov(killed ~ cause, data = tr)
summary(a1)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## cause        5  11444    2289   1.996  0.087 .
## Residuals   88 100909    1147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
TukeyHSD(a1)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = killed ~ cause, data = tr)
##
## $cause
##                                  diff       lwr       upr     p adj
## fire-attack                -25.71666667 -68.90905 17.475717 0.5129895
## human error-attack         -28.68541667 -59.55723  2.186400 0.0839898
## natural-attack             -28.64848485 -67.81172 10.514752 0.2811369
## technical-attack           -29.99607843 -64.94540  4.953245 0.1351620
## unclear track-attack       -34.83030303 -73.99354  4.332934 0.1103239
## human error-fire            -2.96875000 -41.96689 36.029389 0.9999239
## natural-fire                -2.93181818 -48.77443 42.910792 0.9999679
## technical-fire              -4.27941176 -46.57883 38.020010 0.9996912
## unclear track-fire          -9.11363636 -54.95625 36.728974 0.9921603
## natural-human error          0.03693182 -34.44541 34.519275 1.0000000
## technical-human error       -1.31066176 -30.92026 28.298933 0.9999948
## unclear track-human error   -6.14488636 -40.62723 28.337457 0.9952887
## technical-natural           -1.34759358 -39.52374 36.828553 0.9999983
## unclear track-natural       -6.18181818 -48.24988 35.886246 0.9981123
## unclear track-technical     -4.83422460 -43.01037 33.341922 0.9990782
```

There is no pair for which the upper and lower range of 95% confidence interval lie on the same side of zero. So there is not any statistical difference across different types of cause.

9. Conclusion

From the data of 16 years, it is clear that some causes can be prevented and some can't. Natural causes are very rare in causing accidents. Every year train accidents take place and exact cause of accident is not known in many cases. If the investigation of accidents are carried out in more depth then improvements can be made effectively. Accidents happening inside train, like fire, are less nowadays. Human negligence is major cause and will remain the cause in future if no special attention is given by train drivers, roadways riders, people crossing tracks, vehicles crossing tracks, etc.

There are some limitations of this data:

- it was not possible to account for each accident because of very brief report of some incidents on news websites
- this report does not address solution for external attacks
- condition (old or newly launched) of trains is not known
- this report addresses accident counts over years 2002-2017. It does not analyse trends in casualties because we wanted to prevent them directly by knowing the cause of accidents.

10. Abbreviations used

- c: Central Railway Zone
- e: Eastern Railway Zone
- ec: East Central Railway Zone
- eco: East Coast Railway Zone
- k: Konkan Railway Zone
- n: Northern Railway Zone
- nc: North Central Railway Zone
- ne: North Eastern Railway Zone
- nef: North East Frontier Railway Zone
- nw: North Western Railway Zone
- s: Southern Railway Zone
- sc: South Central Railway Zone
- se: South Eastern Railway Zone
- sw: South Western Railway Zone
- w: Western Railway Zone
- wc: West Central Railway Zone

11. References

1. [https://en.wikipedia.org/wiki/List_of_Indian_rail_accidents]
2. [https://timesofindia.indiatimes.com/india/586-train-accidents-in-last-5-years-53-due-to-derailments/articleshow/60141578.cms]
3. [https://en.wikipedia.org/wiki/Climate_of_India]
4. [https://en.wikipedia.org/wiki/List_of_Uttar_Pradesh_train_accidents]
5. [http://www.thehindu.com/todays-paper/Hijack-leads-to-train-collision-4-die/article16626772.ece]
6. [https://timesofindia.indiatimes.com/city/kochi/Three-die-in-train-accident/articleshow/12063072.cms]
7. [https://economictimes.indiatimes.com/industry/transportation/railways/railways-target-laying-9-5-km-of-tracks-every-articleshow/57195430.cms]
8. [http://www.indianrailways.gov.in/railwayboard/uploads/directorate/finance_budget/Budget_2015-16/White_Paper-_English.pdf]
9. [https://en.wikipedia.org/wiki/2011_Census_of_India]
10. [http://www.bbc.com/news/uk-34232974]
11. [http://news.bbc.co.uk/2/hi/south_asia/2025207.stm]
12. [http://www.indianrailways.gov.in/railwayboard/uploads/directorate/stat_econ/IRSP_2014-15/IR_Annual_Report%20%26%20Accounts_2014-15/11.pdf]
13. [http://www.dailymail.co.uk/indiahome/indianews/article-3683522/Indian-Railways-use-radar-based-eye-device-warn-dr html]
14. [http://pib.nic.in/newsite/PrintRelease.aspx?relid=155175]