Exploratory Data Analysis of pipeline incidents in Canada

Ismat Ara Khan

Background:

The Canada Energy Regulator (CER) has a mandate to protect people and the environment during construction, operation, and abandonment of oil and gas pipelines and associated facilities. Despite its best efforts in prevention and mitigation, sometimes incidents that lead to adverse effects to people and the environment can happen. During the period from 2008 to 2020 there have been 723 incidents that involved release of substance.

Descriptive Analysis:

This is the basic and most commonly used analysis technique in Statistics. In this study descriptive methods will be employed mainly to identify the distribution of the number of incidents during the period from 2008 to 2020. Figures which illustrate the relationships between number of incidents and below factors are presented here.

Merging data

```
library(readxl)
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.3.3
```

```
##
## 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
```

```
pipeline_original=read_xlsx("C:/Users/Ismat/OneDrive/Desktop/fall24/Stat consulting/ass 1/pipeline-incic
case_study <- read_xlsx("C:/Users/Ismat/OneDrive/Desktop/fall24/Stat consulting/ass 1/ssc2021_case_study

case_study <- case_study |> rename(Incident_Number = `Event Number`)

pipeline_original <- pipeline_original |> rename(Incident_Number = `Incident Number`)

merged_data <- left_join(case_study, pipeline_original, by = "Incident_Number")
```

Cleaning data

Let's delete some columns which have too many NA values:

Clean the characteric ones:

```
# List of columns to keep
keep_columns <- c("Product Category", "Product Type", "Volume Released", "Latitude.x", "Longitude.x")

# Calculate missing values per column (works for both numeric and character columns)
missing_values <- sapply(merged_data, function(x) sum(is.na(x)))

# Identify columns to keep: either in 'keep_columns' or have <= 1000 missing values
columns_to_keep <- names(missing_values)[missing_values <= 1000 | names(missing_values) %in% keep_colum

# Subset the data with the columns to keep
cleaned_data <- merged_data[, columns_to_keep]

# View the cleaned data structure
str(cleaned_data)
```

```
## tibble [723 x 108] (S3: tbl_df/tbl/data.frame)
##  $ Incident_Number                                                    : chr [1:723] "INC2
##  $ Occurred/Discovered Date and Time                                  : POSIXct[1:723], f
##  $ Product Category                                                   : chr [1:723] "Liqu
##  $ Product Type                                                       : chr [1:723] "Crud
##  $ Volume Released                                                    : num [1:723] 8 100
##  $ Latitude.x                                                         : num [1:723] 49.7
##  $ Longitude.x                                                        : num [1:723] -101.
##  $ Incident Types                                                     : chr [1:723] "Relea
##  $ Reported Date                                                      : POSIXct[1:723], f
##  $ Nearest Populated Centre                                           : chr [1:723] "Crom
##  $ Province                                                           : chr [1:723] "Mani
##  $ Company                                                            : chr [1:723] "Enbr
##  $ Status                                                             : chr [1:723] "Clos
##  $ Latitude.y                                                         : num [1:723] 49.8
##  $ Longitude.y                                                        : num [1:723] -101.
##  $ Approximate Volume Released (m3)                                   : chr [1:723] "8" "
##  $ Substance                                                          : chr [1:723] "Crud
##  $ Release Type                                                       : chr [1:723] "Liqu
##  $ Significant                                                        : chr [1:723] "No"
##  $ Year                                                               : num [1:723] 2008
##  $ Occurrence Date and Time                                           : chr [1:723] "2008
##  $ Discovered Date and Time                                           : chr [1:723] "2008
##  $ Detailed what happened                                             : chr [1:723] "Dama
##  $ What happened category                                             : chr [1:723] "Corr
##  $ Detailed why it happened                                           : chr [1:723] "Job
##  $ Why it happened category                                           : chr [1:723] "Main
##  $ Duration of interruption of pipeline operations                    : chr [1:723] NA NA
##  $ Pipeline or Facility Type                                          : chr [1:723] NA NA
##  $ Activity being performed at time of incident                       : chr [1:723] "Oper
##  $ How the incident was discovered                                    : chr [1:723] "Empl
##  $ Closed Date                                                        : POSIXct[1:723], f
##  $ Pipeline or facility equipment involved                            : chr [1:723] "Yes"
##  $ Rupture                                                            : chr [1:723] "No"
```

```
##  $ Incident type                                                                : chr [1:723] "Relea
##  $ Released substance type                                                      : chr [1:723] "Crude
##  $ Released volume (m3)                                                         : num [1:723] 8 100
##  $ Pipe body release                                                            : chr [1:723] "No" "
##  $ Residual effects on the environment                                          : chr [1:723] "No" "
##  $ Source of Explosion                                                          : chr [1:723] NA NA
##  $ Source of Fire                                                               : chr [1:723] NA NA
##  $ Activity at time of fatality                                                 : chr [1:723] NA NA
##  $ Number of fatalities                                                         : num [1:723] NA NA
##  $ Type of Injury                                                               : chr [1:723] NA NA
##  $ Number of individuals injured                                                : num [1:723] NA NA
##  $ Workdays lost                                                                : num [1:723] NA NA
##  $ Work restricted by injury                                                    : chr [1:723] NA NA
##  $ Conditions that resulted in the operation beyond  limits                     : chr [1:723] NA NA
##  $ Conditions that resulted in adverse effects on the environment               : chr [1:723] NA NA
##  $ Number of people evacuated                                                   : num [1:723] 0 0 0
##  $ Regulation                                                                   : chr [1:723] "OPR"
##  $ Pipeline Name                                                                : chr [1:723] NA NA
##  $ Pipeline outside diameter (NPS)                                              : chr [1:723] NA NA
##  $ Pipeline length (km)                                                         : num [1:723] NA NA
##  $ Substance carried                                                            : chr [1:723] NA NA
##  $ Facility Name                                                                : chr [1:723] "Crom
##  $ Facility Type                                                                : chr [1:723] NA NA
##  $ Facility latitude                                                            : num [1:723] NA NA
##  $ Facility longitude                                                           : num [1:723] NA NA
##  $ Country                                                                      : chr [1:723] "Canad
##  $ Affects Company Property                                                     : chr [1:723] "Yes"
##  $ Off Company Property                                                         : chr [1:723] "No" "
##  $ Affects Pipeline right-of-way                                                : chr [1:723] "No" "
##  $ Affects off Pipeline right-of-way                                            : chr [1:723] "No" "
##  $ Land Use                                                                     : chr [1:723] "Devel
##  $ Population Density                                                           : chr [1:723] "Unkno
##  $ Kilometre post                                                               : chr [1:723] NA "M
##  $ Emergency Level                                                              : chr [1:723] "Level
##  $ Investigation Type                                                           : chr [1:723] "Stand
##  $ Was NEB Staff Deployed                                                       : chr [1:723] "No" "
##  $ Related NEB event number                                                     : chr [1:723] NA NA
##  $ Equipment or component involved                                              : chr [1:723] "Stati
##  $ Design standard                                                              : chr [1:723] NA NA
##  $ Nominal pipe size                                                            : chr [1:723] NA "NI
##  $ Material                                                                     : chr [1:723] NA NA
##  $ Material grade                                                               : chr [1:723] NA NA
##  $ Schedule                                                                     : num [1:723] NA NA
##  $ Design wall thickness (mm)                                                   : chr [1:723] NA NA
##  $ Custom design wall thickness (mm)                                            : num [1:723] NA NA
##  $ Actual wall thickness (mm)                                                   : num [1:723] NA NA
##  $ Licensed maximum operating pressure (kPa)                                    : num [1:723] NA 0 0
##  $ Restricted operating pressure (kPa)                                          : num [1:723] NA NA
##  $ Actual operating pressure at time of failure (kPa)                           : num [1:723] NA NA
##  $ Designed depth of cover (m)                                                  : num [1:723] NA NA
##  $ Actual depth of cover (m)                                                    : num [1:723] NA NA
##  $ Year of manufacture                                                          : num [1:723] NA NA
##  $ Year of installation                                                         : num [1:723] NA NA
##  $ Year when put into service                                                   : num [1:723] NA NA
```

```
##  $ Most recent cathodic protection reading at incident site (mV vs. Cu/CuSO4)    : num [1:723] NA NA
##  $ Weld type                                                                     : chr [1:723] NA NA
##  $ Seam type                                                                     : chr [1:723] NA NA
##  $ Seam joining method                                                           : chr [1:723] NA NA
##  $ Seam clock position                                                           : chr [1:723] NA NA
##  $ Coating location                                                              : chr [1:723] NA NA
##  $ Coating type                                                                  : chr [1:723] NA NA
##  $ Coating condition                                                             : chr [1:723] NA NA
##  $ Application method                                                            : chr [1:723] NA NA
##  $ Year when the coating was applied                                             : num [1:723] NA NA
##  $ Insulation installed                                                          : chr [1:723] "No" "
##  $ Repair type                                                                   : chr [1:723] "Type
##   [list output truncated]
```

```
sum(is.na(cleaned_data))
```

```
## [1] 34483
```

```
head(cleaned_data)
```

```
## # A tibble: 6 x 108
##   Incident_Number Occurred/Discovered Date a~1 `Product Category` `Product Type`
##   <chr>           <dttm>                       <chr>              <chr>
## 1 INC2008-001     2008-01-02 11:30:00          Liquid             Crude Oil - S~
## 2 INC2008-004     2008-01-23 06:05:00          Liquid             Crude Oil - S~
## 3 INC2008-008     2008-01-26 10:45:00          Gas                Natural Gas -~
## 4 INC2008-009     2008-01-27 12:01:00          Miscellaneous      Potassium Hyd~
## 5 INC2008-128     2008-02-04 12:00:00          Gas                Natural Gas -~
## 6 INC2008-016     2008-02-22 12:00:00          Gas                Natural Gas -~
## # i abbreviated name: 1: `Occurred/Discovered Date and Time`
## # i 104 more variables: `Volume Released` <dbl>, Latitude.x <dbl>,
## #   Longitude.x <dbl>, `Incident Types` <chr>, `Reported Date` <dttm>,
## #   `Nearest Populated Centre` <chr>, Province <chr>, Company <chr>,
## #   Status <chr>, Latitude.y <dbl>, Longitude.y <dbl>,
## #   `Approximate Volume Released (m3)` <chr>, Substance <chr>,
## #   `Release Type` <chr>, Significant <chr>, Year <dbl>, ...
```

I cleaned the same columns which are Approximate Volume released (m3),Latitude.y, longitude.y

```
# Install and load dplyr package if not already installed


library(dplyr)

# Remove the specified columns from merged_data
cleaned_data <- cleaned_data %>%
  select(-`Approximate Volume Released (m3)`, -Latitude.y, -Longitude.y)

# View the updated dataset
head(cleaned_data)
```

```
## # A tibble: 6 x 105
##   Incident_Number Occurred/Discovered Date a~1 `Product Category` `Product Type`
##   <chr>           <dttm>                       <chr>              <chr>
## 1 INC2008-001     2008-01-02 11:30:00          Liquid             Crude Oil - S~
## 2 INC2008-004     2008-01-23 06:05:00          Liquid             Crude Oil - S~
## 3 INC2008-008     2008-01-26 10:45:00          Gas                Natural Gas -~
```

```
## 4 INC2008-009     2008-01-27 12:01:00          Miscellaneous      Potassium Hyd~
## 5 INC2008-128     2008-02-04 12:00:00          Gas                Natural Gas -~
## 6 INC2008-016     2008-02-22 12:00:00          Gas                Natural Gas -~
## # i abbreviated name: 1: `Occurred/Discovered Date and Time`
## # i 101 more variables: `Volume Released` <dbl>, Latitude.x <dbl>,
## #   Longitude.x <dbl>, `Incident Types` <chr>, `Reported Date` <dttm>,
## #   `Nearest Populated Centre` <chr>, Province <chr>, Company <chr>,
## #   Status <chr>, Substance <chr>, `Release Type` <chr>, Significant <chr>,
## #   Year <dbl>, `Occurrence Date and Time` <chr>,
## #   `Discovered Date and Time` <chr>, `Detailed what happened` <chr>, ...
```

Descriptive Statistics

```
# Check the structure of the merged data
str(cleaned_data)
```

```
## tibble [723 x 105] (S3: tbl_df/tbl/data.frame)
##  $ Incident_Number                                                : chr [1:723] "INC2
##  $ Occurred/Discovered Date and Time                              : POSIXct[1:723], f
##  $ Product Category                                               : chr [1:723] "Liqu
##  $ Product Type                                                   : chr [1:723] "Crud
##  $ Volume Released                                                : num [1:723] 8 100
##  $ Latitude.x                                                     : num [1:723] 49.7 
##  $ Longitude.x                                                    : num [1:723] -101.
##  $ Incident Types                                                 : chr [1:723] "Rele
##  $ Reported Date                                                  : POSIXct[1:723], f
##  $ Nearest Populated Centre                                       : chr [1:723] "Crom
##  $ Province                                                       : chr [1:723] "Mani
##  $ Company                                                        : chr [1:723] "Enbr
##  $ Status                                                         : chr [1:723] "Clos
##  $ Substance                                                      : chr [1:723] "Crud
##  $ Release Type                                                   : chr [1:723] "Liqu
##  $ Significant                                                    : chr [1:723] "No" 
##  $ Year                                                           : num [1:723] 2008 
##  $ Occurrence Date and Time                                       : chr [1:723] "2008
##  $ Discovered Date and Time                                       : chr [1:723] "2008
##  $ Detailed what happened                                         : chr [1:723] "Dama
##  $ What happened category                                         : chr [1:723] "Corr
##  $ Detailed why it happened                                       : chr [1:723] "Job 
##  $ Why it happened category                                       : chr [1:723] "Main
##  $ Duration of interruption of pipeline operations                : chr [1:723] NA NA
##  $ Pipeline or Facility Type                                      : chr [1:723] NA NA
##  $ Activity being performed at time of incident                   : chr [1:723] "Oper
##  $ How the incident was discovered                                : chr [1:723] "Empl
##  $ Closed Date                                                    : POSIXct[1:723], f
##  $ Pipeline or facility equipment involved                        : chr [1:723] "Yes"
##  $ Rupture                                                        : chr [1:723] "No" 
##  $ Incident type                                                  : chr [1:723] "Rele
##  $ Released substance type                                        : chr [1:723] "Crud
##  $ Released volume (m3)                                           : num [1:723] 8 100
##  $ Pipe body release                                              : chr [1:723] "No" 
##  $ Residual effects on the environment                            : chr [1:723] "No" 
##  $ Source of Explosion                                            : chr [1:723] NA NA
```

```
##  $ Source of Fire                                                                : chr [1:723] NA NA
##  $ Activity at time of fatality                                                  : chr [1:723] NA NA
##  $ Number of fatalities                                                          : num [1:723] NA NA
##  $ Type of Injury                                                                : chr [1:723] NA NA
##  $ Number of individuals injured                                                 : num [1:723] NA NA
##  $ Workdays lost                                                                 : num [1:723] NA NA
##  $ Work restricted by injury                                                     : chr [1:723] NA NA
##  $ Conditions that resulted in the operation beyond  limits                      : chr [1:723] NA NA
##  $ Conditions that resulted in adverse effects on the environment                : chr [1:723] NA NA
##  $ Number of people evacuated                                                    : num [1:723] 0 0 0
##  $ Regulation                                                                    : chr [1:723] "OPR"
##  $ Pipeline Name                                                                 : chr [1:723] NA NA
##  $ Pipeline outside diameter (NPS)                                               : chr [1:723] NA NA
##  $ Pipeline length (km)                                                          : num [1:723] NA NA
##  $ Substance carried                                                             : chr [1:723] NA NA
##  $ Facility Name                                                                 : chr [1:723] "Crom
##  $ Facility Type                                                                 : chr [1:723] NA NA
##  $ Facility latitude                                                             : num [1:723] NA NA
##  $ Facility longitude                                                            : num [1:723] NA NA
##  $ Country                                                                       : chr [1:723] "Canad
##  $ Affects Company Property                                                      : chr [1:723] "Yes"
##  $ Off Company Property                                                          : chr [1:723] "No"
##  $ Affects Pipeline right-of-way                                                 : chr [1:723] "No"
##  $ Affects off Pipeline right-of-way                                             : chr [1:723] "No"
##  $ Land Use                                                                      : chr [1:723] "Devel
##  $ Population Density                                                            : chr [1:723] "Unkno
##  $ Kilometre post                                                                : chr [1:723] NA "M
##  $ Emergency Level                                                               : chr [1:723] "Level
##  $ Investigation Type                                                            : chr [1:723] "Stand
##  $ Was NEB Staff Deployed                                                        : chr [1:723] "No"
##  $ Related NEB event number                                                      : chr [1:723] NA NA
##  $ Equipment or component involved                                               : chr [1:723] "Stati
##  $ Design standard                                                               : chr [1:723] NA NA
##  $ Nominal pipe size                                                             : chr [1:723] NA "NI
##  $ Material                                                                      : chr [1:723] NA NA
##  $ Material grade                                                                : chr [1:723] NA NA
##  $ Schedule                                                                      : num [1:723] NA NA
##  $ Design wall thickness (mm)                                                    : chr [1:723] NA NA
##  $ Custom design wall thickness (mm)                                             : num [1:723] NA NA
##  $ Actual wall thickness (mm)                                                    : num [1:723] NA NA
##  $ Licensed maximum operating pressure (kPa)                                     : num [1:723] NA 0 0
##  $ Restricted operating pressure (kPa)                                           : num [1:723] NA NA
##  $ Actual operating pressure at time of failure (kPa)                            : num [1:723] NA NA
##  $ Designed depth of cover (m)                                                   : num [1:723] NA NA
##  $ Actual depth of cover (m)                                                     : num [1:723] NA NA
##  $ Year of manufacture                                                           : num [1:723] NA NA
##  $ Year of installation                                                          : num [1:723] NA NA
##  $ Year when put into service                                                    : num [1:723] NA NA
##  $ Most recent cathodic protection reading at incident site (mV vs. Cu/CuSO4)    : num [1:723] NA NA
##  $ Weld type                                                                     : chr [1:723] NA NA
##  $ Seam type                                                                     : chr [1:723] NA NA
##  $ Seam joining method                                                           : chr [1:723] NA NA
##  $ Seam clock position                                                           : chr [1:723] NA NA
##  $ Coating location                                                              : chr [1:723] NA NA
```

```
##  $ Coating type                                                                  : chr [1:723] NA NA
##  $ Coating condition                                                             : chr [1:723] NA NA
##  $ Application method                                                            : chr [1:723] NA NA
##  $ Year when the coating was applied                                             : num [1:723] NA NA
##  $ Insulation installed                                                          : chr [1:723] "No" "
##  $ Repair type                                                                   : chr [1:723] "Type
##  $ Repair date                                                                   : POSIXct[1:723], fo
##  $ Equipment or component has never been inspected                               : chr [1:723] "No" "
##  $ Most recent inspection date for the failed equipment or component             : POSIXct[1:723], fo
##   [list output truncated]
```

```
# Summary statistics for all columns
summary(cleaned_data)
```

```
##  Incident_Number    Occurred/Discovered Date and Time Product Category
##  Length:723         Min.   :2008-01-02 11:30:00.00    Length:723
##  Class :character   1st Qu.:2011-02-14 15:42:30.00    Class :character
##  Mode  :character   Median :2013-07-08 02:00:00.00    Mode  :character
##                     Mean   :2013-09-21 22:10:35.19
##                     3rd Qu.:2016-03-26 03:52:30.00
##                     Max.   :2020-06-24 10:30:00.00
##
##  Product Type       Volume Released      Latitude.x     Longitude.x
##  Length:723         Min.   :       0   Min.   :42.28   Min.   :-126.89
##  Class :character   1st Qu.:       1   1st Qu.:49.19   1st Qu.:-120.47
##  Mode  :character   Median :      10   Median :51.44   Median :-111.69
##                     Mean   :   91334   Mean   :51.52   Mean   :-102.51
##                     3rd Qu.:     349   3rd Qu.:55.56   3rd Qu.: -84.10
##                     Max.   :16500000   Max.   :65.29   Max.   : -61.61
##                     NA's   :251
##  Incident Types     Reported Date                     Nearest Populated Centre
##  Length:723         Min.   :2008-01-02 00:00:00.00   Length:723
##  Class :character   1st Qu.:2011-02-16 00:00:00.00   Class :character
##  Mode  :character   Median :2013-07-08 00:00:00.00   Mode  :character
##                     Mean   :2013-09-24 15:46:03.49
##                     3rd Qu.:2016-03-27 12:00:00.00
##                     Max.   :2020-06-25 00:00:00.00
##
##    Province           Company             Status           Substance
##  Length:723         Length:723         Length:723         Length:723
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  Release Type       Significant              Year      Occurrence Date and Time
##  Length:723         Length:723         Min.   :2008   Length:723
##  Class :character   Class :character   1st Qu.:2011   Class :character
##  Mode  :character   Mode  :character   Median :2013   Mode  :character
##                                        Mean   :2013
##                                        3rd Qu.:2016
##                                        Max.   :2020
##
##  Discovered Date and Time Detailed what happened What happened category
```

```
##  Length:723               Length:723             Length:723
##  Class :character         Class :character       Class :character
##  Mode  :character         Mode  :character       Mode  :character
##
##
##
##
##  Detailed why it happened Why it happened category
##  Length:723               Length:723
##  Class :character         Class :character
##  Mode  :character         Mode  :character
##
##
##
##
##  Duration of interruption of pipeline operations Pipeline or Facility Type
##  Length:723                                      Length:723
##  Class :character                                Class :character
##  Mode  :character                                Mode  :character
##
##
##
##
##  Activity being performed at time of incident How the incident was discovered
##  Length:723                                   Length:723
##  Class :character                             Class :character
##  Mode  :character                             Mode  :character
##
##
##
##
##   Closed Date                     Pipeline or facility equipment involved
##  Min.   :2009-05-29 00:00:00.00   Length:723
##  1st Qu.:2013-01-02 00:00:00.00   Class :character
##  Median :2015-03-03 00:00:00.00   Mode  :character
##  Mean   :2015-06-28 10:01:29.63
##  3rd Qu.:2017-08-15 00:00:00.00
##  Max.   :2024-07-05 00:00:00.00
##
##    Rupture          Incident type      Released substance type
##  Length:723         Length:723         Length:723
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  Released volume (m3) Pipe body release  Residual effects on the environment
##  Min.   :       0     Length:723         Length:723
##  1st Qu.:       0     Class :character   Class :character
##  Median :       1     Mode  :character   Mode  :character
##  Mean   :   78855
##  3rd Qu.:      64
##  Max.   :16500000
```

```
##  NA's   :31
##  Source of Explosion Source of Fire     Activity at time of fatality
##  Length:723          Length:723         Length:723
##  Class :character    Class :character   Class :character
##  Mode  :character    Mode  :character   Mode  :character
##
##
##
##
##  Number of fatalities Type of Injury     Number of individuals injured
##  Min.   : NA          Length:723         Min.   :2.0
##  1st Qu.: NA          Class :character   1st Qu.:2.5
##  Median : NA          Mode  :character   Median :3.0
##  Mean   :NaN                             Mean   :3.0
##  3rd Qu.: NA                             3rd Qu.:3.5
##  Max.   : NA                             Max.   :4.0
##  NA's   :723                             NA's   :721
##  Workdays lost Work restricted by injury
##  Min.   : NA   Length:723
##  1st Qu.: NA   Class :character
##  Median : NA   Mode  :character
##  Mean   :NaN
##  3rd Qu.: NA
##  Max.   : NA
##  NA's   :723
##  Conditions that resulted in the operation beyond  limits
##  Length:723
##  Class :character
##  Mode  :character
##
##
##
##
##  Conditions that resulted in adverse effects on the environment
##  Length:723
##  Class :character
##  Mode  :character
##
##
##
##
##  Number of people evacuated  Regulation        Pipeline Name
##  Min.   :  0.0000            Length:723         Length:723
##  1st Qu.:  0.0000            Class :character   Class :character
##  Median :  0.0000            Mode  :character   Mode  :character
##  Mean   :  0.5864
##  3rd Qu.:  0.0000
##  Max.   :350.0000
##
##  Pipeline outside diameter (NPS) Pipeline length (km) Substance carried
##  Length:723                      Min.   :   0.206     Length:723
##  Class :character                1st Qu.:  40.424     Class :character
##  Mode  :character                Median : 384.474     Mode  :character
##                                  Mean   : 796.715
```

```
##                                3rd Qu.:1274.364
##                                Max.   :3393.426
##                                NA's   :437
##  Facility Name      Facility Type      Facility latitude Facility longitude
##  Length:723         Length:723         Min.   :43.00     Min.   :-122.70
##  Class :character   Class :character   1st Qu.:49.96     1st Qu.:-121.94
##  Mode  :character   Mode  :character   Median :54.39     Median :-120.52
##                                        Mean   :53.11     Mean   :-110.86
##                                        3rd Qu.:55.81     3rd Qu.:-110.04
##                                        Max.   :58.65     Max.   : -61.61
##                                        NA's   :601       NA's   :601
##    Country          Affects Company Property Off Company Property
##  Length:723         Length:723               Length:723
##  Class :character   Class :character         Class :character
##  Mode  :character   Mode  :character         Mode  :character
##
##
##
##
##  Affects Pipeline right-of-way Affects off Pipeline right-of-way
##  Length:723                    Length:723
##  Class :character              Class :character
##  Mode  :character              Mode  :character
##
##
##
##
##    Land Use         Population Density Kilometre post     Emergency Level
##  Length:723         Length:723         Length:723         Length:723
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  Investigation Type Was NEB Staff Deployed Related NEB event number
##  Length:723         Length:723             Length:723
##  Class :character   Class :character       Class :character
##  Mode  :character   Mode  :character       Mode  :character
##
##
##
##
##  Equipment or component involved Design standard    Nominal pipe size
##  Length:723                      Length:723         Length:723
##  Class :character                Class :character   Class :character
##  Mode  :character                Mode  :character   Mode  :character
##
##
##
##
##    Material         Material grade        Schedule
##  Length:723         Length:723         Min.   : 20.00
##  Class :character   Class :character   1st Qu.: 80.00
```

```
##  Mode  :character   Mode  :character   Median : 80.00
##                                        Mean   : 87.59
##                                        3rd Qu.:100.00
##                                        Max.   :160.00
##                                        NA's   :694
##  Design wall thickness (mm) Custom design wall thickness (mm)
##  Length:723                 Min.   :  0.000
##  Class :character           1st Qu.:  3.075
##  Mode  :character           Median :  4.000
##                             Mean   : 11.231
##                             3rd Qu.:  6.175
##                             Max.   :114.000
##                             NA's   :707
##  Actual wall thickness (mm) Licensed maximum operating pressure (kPa)
##  Min.   :  0.900            Min.   :    0
##  1st Qu.:  3.125            1st Qu.:    0
##  Median :  4.000            Median : 1440
##  Mean   :  7.554            Mean   : 3800
##  3rd Qu.:  7.750            3rd Qu.: 8262
##  Max.   :114.000            Max.   :14400
##  NA's   :673                NA's   :551
##  Restricted operating pressure (kPa)
##  Min.   :   0
##  1st Qu.:4400
##  Median :5897
##  Mean   :5235
##  3rd Qu.:7026
##  Max.   :8935
##  NA's   :712
##  Actual operating pressure at time of failure (kPa) Designed depth of cover (m)
##  Min.   :   5                                       Min.   :0.000
##  1st Qu.:3709                                       1st Qu.:0.600
##  Median :5170                                       Median :0.860
##  Mean   :4779                                       Mean   :0.815
##  3rd Qu.:6630                                       3rd Qu.:1.050
##  Max.   :8960                                       Max.   :1.500
##  NA's   :651                                        NA's   :711
##  Actual depth of cover (m) Year of manufacture Year of installation
##  Min.   :0.000             Min.   :1952        Min.   : 197
##  1st Qu.:0.450             1st Qu.:1970        1st Qu.:1972
##  Median :1.150             Median :1994        Median :1980
##  Mean   :1.054             Mean   :1988        Mean   :1968
##  3rd Qu.:1.355             3rd Qu.:2009        3rd Qu.:2008
##  Max.   :4.500             Max.   :2019        Max.   :2019
##  NA's   :687               NA's   :686         NA's   :627
##  Year when put into service
##  Min.   :1953
##  1st Qu.:1972
##  Median :1981
##  Mean   :1986
##  3rd Qu.:2008
##  Max.   :2019
##  NA's   :636
##  Most recent cathodic protection reading at incident site (mV vs. Cu/CuSO4)
```

```
##  Min.   :     -2124
##  1st Qu.:     -1351
##  Median :     -1118
##  Mean   :  62105804
##  3rd Qu.:      -688
##  Max.   :1863200035
##  NA's   :693
##   Weld type          Seam type         Seam joining method Seam clock position
##  Length:723         Length:723         Length:723          Length:723
##  Class :character   Class :character   Class :character    Class :character
##  Mode  :character   Mode  :character   Mode  :character    Mode  :character
##
##
##
##
##  Coating location   Coating type       Coating condition  Application method
##  Length:723         Length:723         Length:723         Length:723
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  Year when the coating was applied Insulation installed Repair type
##  Min.   :1952                      Length:723           Length:723
##  1st Qu.:1968                      Class :character     Class :character
##  Median :1974                      Mode  :character     Mode  :character
##  Mean   :1976
##  3rd Qu.:1981
##  Max.   :2017
##  NA's   :686
##   Repair date                  Equipment or component has never been inspected
##  Min.   :2008-01-11 00:00:00   Length:723
##  1st Qu.:2009-09-21 12:00:00   Class :character
##  Median :2012-03-19 00:00:00   Mode  :character
##  Mean   :2013-03-23 23:30:00
##  3rd Qu.:2016-09-07 12:00:00
##  Max.   :2020-07-24 00:00:00
##  NA's   :579
##  Most recent inspection date for the failed equipment or component
##  Min.   :2006-01-02 00:00:00
##  1st Qu.:2015-03-27 06:00:00
##  Median :2016-09-12 00:00:00
##  Mean   :2016-07-11 13:40:48
##  3rd Qu.:2017-12-09 18:00:00
##  Max.   :2020-06-24 00:00:00
##  NA's   :623
##  Type of most recent inspection
##  Length:723
##  Class :character
##  Mode  :character
##
##
##
```

```
##
##  Most recent inspection part of the routine inspection program
##  Length:723
##  Class :character
##  Mode  :character
##
##
##
##
##  No maintenance done on this equipment or component
##  Length:723
##  Class :character
##  Mode  :character
##
##
##
##
##  Date of the most recent maintenance work for the failed equipment or component
##  Min.   :2006-01-02 00:00:00.00
##  1st Qu.:2009-09-07 00:00:00.00
##  Median :2012-03-02 00:00:00.00
##  Mean   :2012-10-16 16:24:10.35
##  3rd Qu.:2015-07-11 12:00:00.00
##  Max.   :2020-06-24 00:00:00.00
##  NA's   :584
##  Most recent maintenance Type
##  Length:723
##  Class :character
##  Mode  :character
##
##
##
##
##  Most recent maintenance work part of the routine maintenance program
##  Length:723
##  Class :character
##  Mode  :character
##
##
##
##
```

```
# To get more specific descriptive statistics for numerical columns (e.g., mean, sd, etc.)
# Use the sapply function for numeric columns only
numeric_columns <- cleaned_data[, sapply(cleaned_data, is.numeric)]
descriptive_stats <- sapply(numeric_columns, function(x) c(mean = mean(x, na.rm = TRUE),
                                                           sd = sd(x, na.rm = TRUE),
                                                           min = min(x, na.rm = TRUE),
                                                           max = max(x, na.rm = TRUE),
                                                           median = median(x, na.rm = TRUE)))
```

```
## Warning in min(x, na.rm = TRUE): no non-missing arguments to min; returning Inf
```

```
## Warning in max(x, na.rm = TRUE): no non-missing arguments to max; returning
## -Inf
```

```
## Warning in min(x, na.rm = TRUE): no non-missing arguments to min; returning Inf
## Warning in max(x, na.rm = TRUE): no non-missing arguments to max; returning
## -Inf
```

```
# Display the descriptive statistics
descriptive_stats
```

```
##        Volume Released Latitude.x Longitude.x        Year Released volume (m3)
## mean      9.133354e+04  51.516134  -102.50543 2013.229599             78854.53
## sd        8.517898e+05   4.374145    20.39376    3.129231            827047.52
## min       1.000000e-04  42.281736  -126.88705 2008.000000                 0.00
## max       1.650000e+07  65.286034   -61.61268 2020.000000          16500000.00
## median    1.000000e+01  51.444184  -111.69054 2013.000000                 1.00
##        Number of fatalities Number of individuals injured Workdays lost
## mean                    NaN                      3.000000           NaN
## sd                       NA                      1.414214            NA
## min                     Inf                      2.000000           Inf
## max                    -Inf                      4.000000          -Inf
## median                   NA                      3.000000            NA
##        Number of people evacuated Pipeline length (km) Facility latitude
## mean                     0.5864454          796.7148074         53.108079
## sd                      13.1774248          933.5780802          3.628778
## min                      0.0000000            0.2057038         43.004754
## max                    350.0000000         3393.4264203         58.654722
## median                   0.0000000          384.4735081         54.385738
##        Facility longitude  Schedule Custom design wall thickness (mm)
## mean           -110.85559  87.58621                           11.23125
## sd               17.25530  42.89981                           27.54085
## min            -122.70420  20.00000                            0.00000
## max             -61.61268 160.00000                          114.00000
## median         -120.52312  80.00000                            4.00000
##        Actual wall thickness (mm) Licensed maximum operating pressure (kPa)
## mean                       7.5540                                  3800.133
## sd                        15.6613                                  4098.083
## min                        0.9000                                     0.000
## max                      114.0000                                 14400.000
## median                     4.0000                                  1439.500
##        Restricted operating pressure (kPa)
## mean                               5235.455
## sd                                 2903.868
## min                                   0.000
## max                                8935.000
## median                             5897.000
##        Actual operating pressure at time of failure (kPa)
## mean                                             4778.574
## sd                                               2355.537
## min                                                 5.000
## max                                              8960.000
## median                                           5170.500
##        Designed depth of cover (m) Actual depth of cover (m)
## mean                     0.8150000                 1.0538889
## sd                       0.4089121                 0.8781353
## min                      0.0000000                 0.0000000
## max                      1.5000000                 4.5000000
```

```
## median                   0.8600000                 1.1500000
##        Year of manufacture Year of installation Year when put into service
## mean            1987.59459             1967.698                 1986.33333
## sd                21.66213              183.708                   19.72878
## min             1952.00000              197.000                 1953.00000
## max             2019.00000             2019.000                 2019.00000
## median          1994.00000             1980.500                 1981.00000
##        Most recent cathodic protection reading at incident site (mV vs. Cu/CuSO4)
## mean                                                                  62105803.7
## sd                                                                   340172392.7
## min                                                                      -2124.0
## max                                                                  1863200035.1
## median                                                                   -1117.5
##        Year when the coating was applied
## mean                            1975.6757
## sd                                15.0574
## min                             1952.0000
## max                             2017.0000
## median                          1974.0000
```

Exploratory Data Analysis

For maps, I will use this:

```
library(geodata)
```

```
## Warning: package 'geodata' was built under R version 4.3.3
```

```
## Loading required package: terra
```

```
## Warning: package 'terra' was built under R version 4.3.3
```

```
## terra 1.8.5
```

```
library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```
## Warning: package 'tibble' was built under R version 4.3.3
```

```
## Warning: package 'tidyr' was built under R version 4.3.3
```

```
## Warning: package 'readr' was built under R version 4.3.3
```

```
## Warning: package 'purrr' was built under R version 4.3.3
```

```
## Warning: package 'stringr' was built under R version 4.3.3
```

```
## Warning: package 'lubridate' was built under R version 4.3.3
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v readr     2.1.5
## v ggplot2   3.5.1     v stringr   1.5.1
## v lubridate 1.9.4     v tibble    3.2.1
## v purrr     1.0.4     v tidyr     1.3.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x tidyr::extract() masks terra::extract()
## x dplyr::filter()  masks stats::filter()
## x dplyr::lag()     masks stats::lag()
```

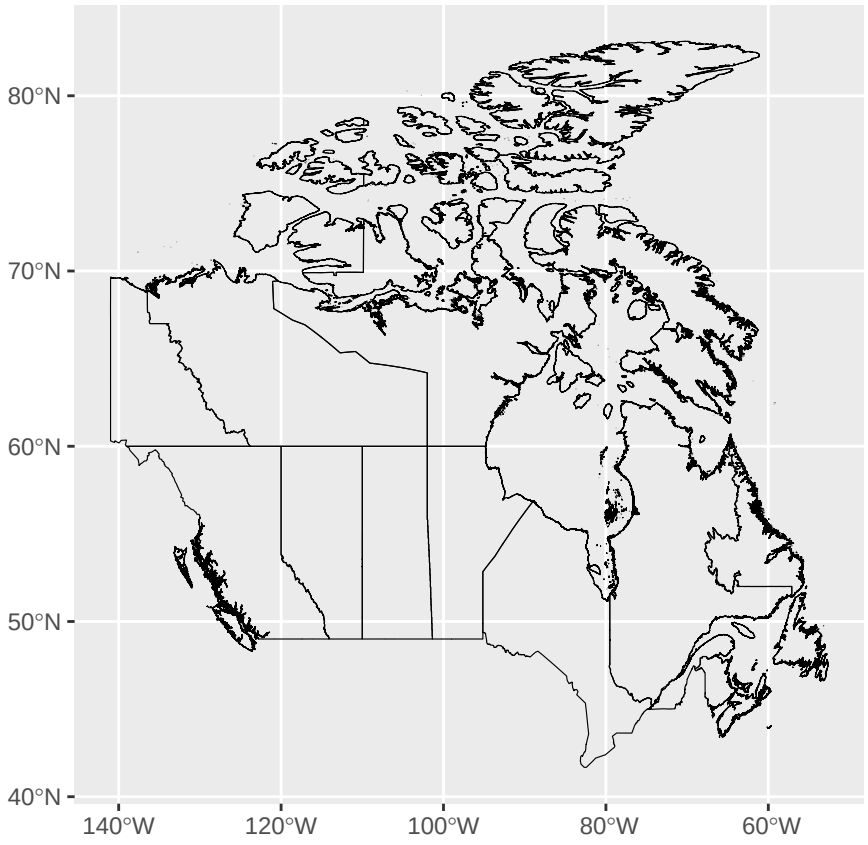
```
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(sf)
```

```
## Warning: package 'sf' was built under R version 4.3.3
```

```
## Linking to GEOS 3.11.2, GDAL 3.8.2, PROJ 9.3.1; sf_use_s2() is TRUE
```

```
can1 <- geodata::gadm(country = "CAN", level = 1, path = tempdir(), version = "latest")
Can1_sf <- st_as_sf(can1)
Can1_sf <- st_transform(Can1_sf, 4269)

ggplot() +
  geom_sf(data = Can1_sf, fill = NA, color = "black")
```



general look of incident across Canada

```
# Load necessary libraries
library(ggplot2)
library(readxl)
library(sf)
library(rnaturalearth)
```

```
## Warning: package 'rnaturalearth' was built under R version 4.3.3
```

```
library(rnaturalearthdata)
```

```
## Warning: package 'rnaturalearthdata' was built under R version 4.3.3
```

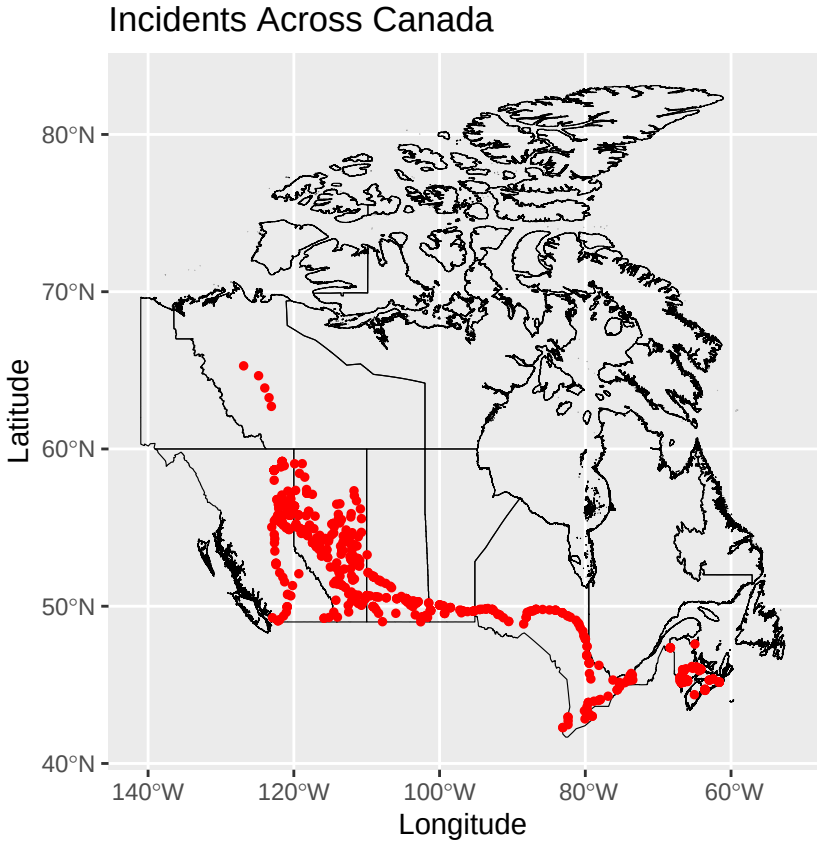
```
##
## Attaching package: 'rnaturalearthdata'
```

```
## The following object is masked from 'package:rnaturalearth':
##
##     countries110
```

```
library(dplyr)
library(ggrepel)
```

```
## Warning: package 'ggrepel' was built under R version 4.3.3
```

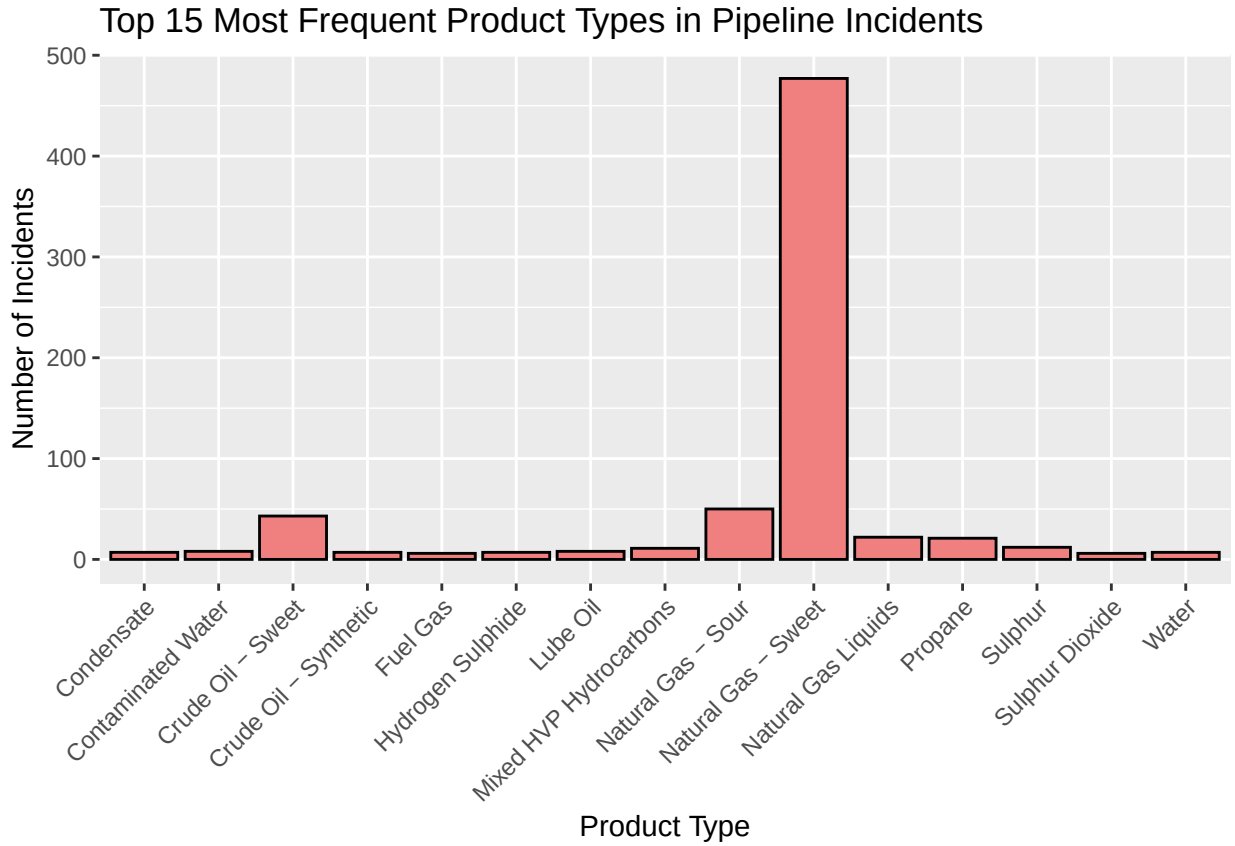
```
cleaned_data$Latitude.x <- as.numeric(cleaned_data$Latitude.x)
cleaned_data$Longitude.x <- as.numeric(cleaned_data$Longitude.x)

# Remove rows with missing lat/long
cleaned_data <- cleaned_data[!is.na(cleaned_data$Latitude.x) & !is.na(cleaned_data$Longitude.x), ]

# Convert the data to an sf object
incidents_sf <- st_as_sf(cleaned_data, coords = c("Longitude.x", "Latitude.x"), crs = 4326)

# Assuming you have Can1_sf as a spatial object for Canada map, we can plot it
ggplot() +
  geom_sf(data = Can1_sf, fill = NA, color = "black") +  # Base map of Canada
  geom_sf(data = incidents_sf, color = "red", size = 1) +  # Incident points on the map
  labs(title = "Incidents Across Canada", x = "Longitude", y = "Latitude")
```



Danger Analysis

#Count of incidents by product type

```r
# Load dplyr for data manipulation
library(dplyr)


# Get the top 15 most frequent Product Types
top_15_product_types <- cleaned_data %>%
  count(`Product Type`) %>%
  top_n(15, n) %>%
  arrange(desc(n)) %>%
  pull(`Product Type`)

# Filter the data for only the top 15 Product Types
filtered_data <- cleaned_data %>%
  filter(`Product Type` %in% top_15_product_types)

# Plot the top 5 most frequent Product Types
ggplot(filtered_data, aes(x = `Product Type`)) +
  geom_bar(fill = "lightcoral", color = "black") +
  labs(title = "Top 15 Most Frequent Product Types in Pipeline Incidents",
       x = "Product Type", y = "Number of Incidents") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate labels for readability
```



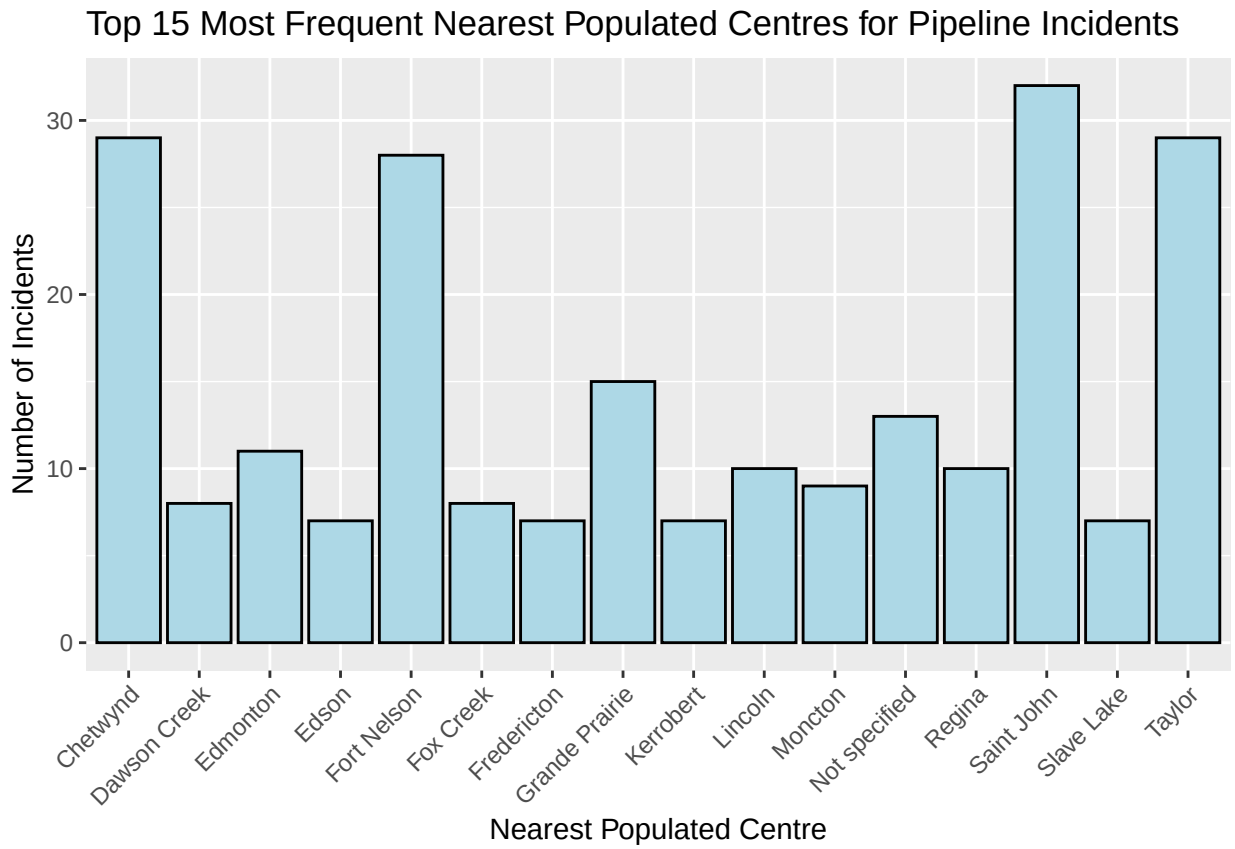
This bar chart displays the number of pipeline incidents for top 15 product type.It is clear that “Natural Gas

- Sweet" has the highest number of incidents which is almost 500. Among others "Natural Gas - Sour" and "Crude Oil - Sweet" have almost the same number of incidents. Though they have the second and the third highest number of incidents but they are only around 50 which is very small number compare to the first one.

We may come to a conclusion that "Natural Gas - Sour" is most dangerous than other products that used so far.

Number of incidents - nearest populated center

```
# Get the top 15 most frequent Nearest Populated Centres
top_15_centres <- cleaned_data %>%
  count(`Nearest Populated Centre`) %>%
  top_n(15, n) %>%
  arrange(desc(n)) %>%
  pull(`Nearest Populated Centre`)

# Filter the data for only the top 15 centres
filtered_data_centres <- cleaned_data %>%
  filter(`Nearest Populated Centre` %in% top_15_centres)

# Plot the top 5 Nearest Populated Centres
ggplot(filtered_data_centres, aes(x = `Nearest Populated Centre`)) +
  geom_bar(fill = "lightblue", color = "black") +
  labs(title = "Top 15 Most Frequent Nearest Populated Centres for Pipeline Incidents",
       x = "Nearest Populated Centre", y = "Number of Incidents") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate labels for readability
```



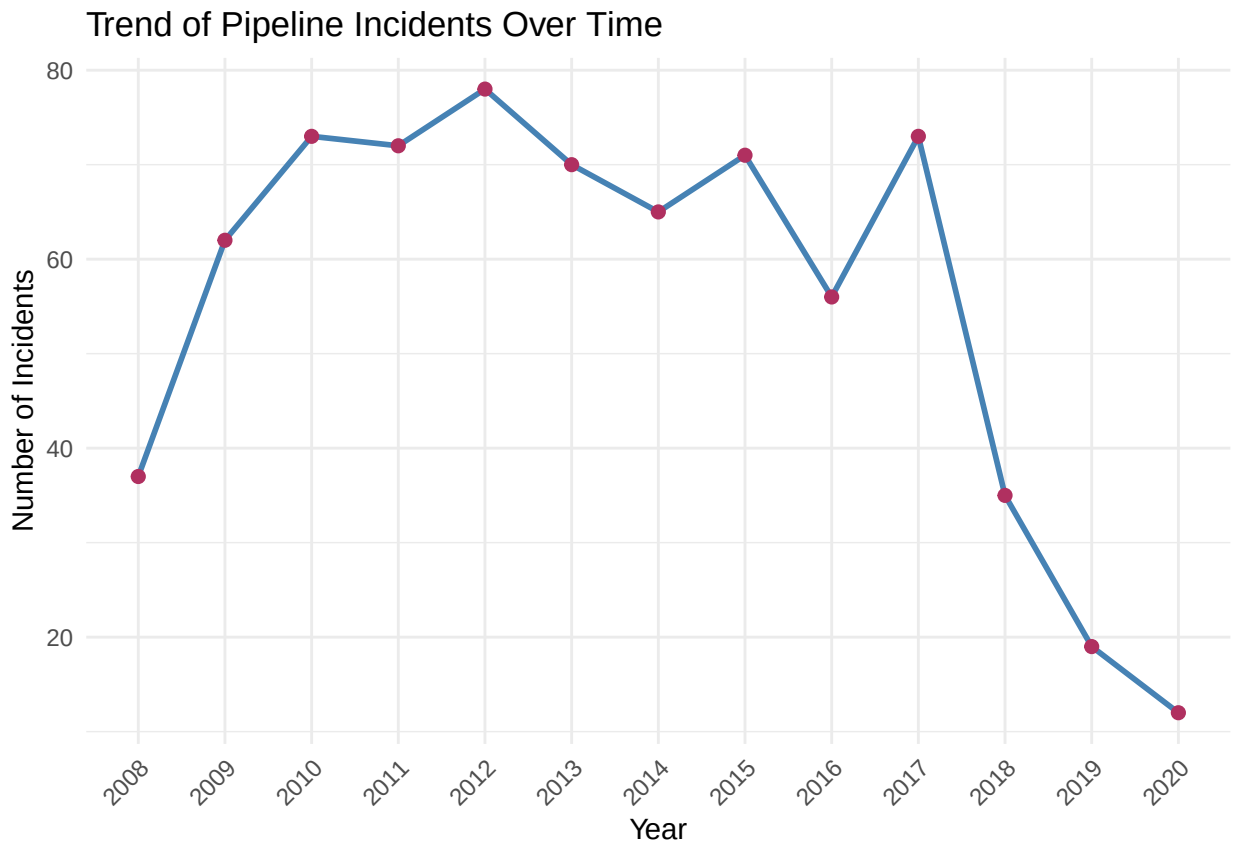
This bar chart displays the number of pipeline incidents based on their proximity to the nearest populated center. The top 15 populated centers with the highest number of incidents are shown. Among them Chetwynd, Edmonton, Saint John and Taylor have experience the highest number of incidents which can interpret as Centers with higher population densities might have more pipeline infrastructure, leading to a higher likelihood of incidents.

Trend Analysis

```
library(readxl)
library(ggplot2)
library(dplyr)


# Convert 'Reported Date' to Date format
cleaned_data$`Reported Date` <- as.Date(cleaned_data$`Reported Date`, format="%Y-%m-%d")

# Extract the year from the 'Reported Date'
cleaned_data$Year <- format(cleaned_data$`Reported Date`, "%Y")

# Group by year and count the number of incidents per year
incident_trend <- table(cleaned_data$Year)
incident_trend <- as.data.frame(incident_trend)
names(incident_trend) <- c("Year", "Incidents")

# Plot the trend of incidents over time
```

```
ggplot(data = incident_trend, aes(x = Year, y = Incidents, group = 1)) +
geom_line(color = "steelblue", size = 1) +
geom_point(color = "maroon", size = 2) +
labs(title = "Trend of Pipeline Incidents Over Time", x = "Year", y = "Number of Incidents") +
theme_minimal() +
theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

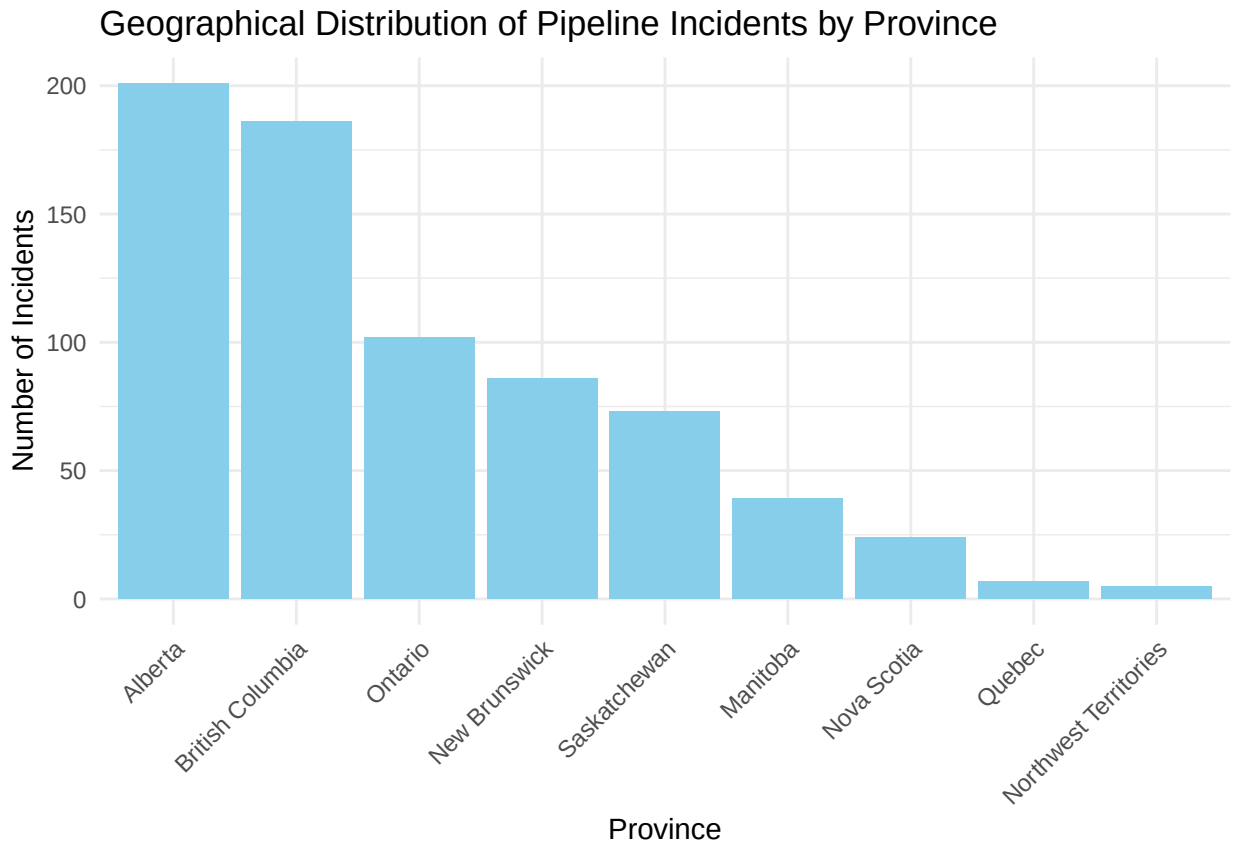


The line chart displays the number of pipeline incidents for each year from 2008 to 2022.We can easily analyze the trend of pipeline incidents from this chart. In 2008, the number incidents was near 40, which increases sharply to 60 within one year and increase steadily for the next year. From 2010 the trend goes up and down till 2017 and after that year it decreases consistently over the years which goes to zero in 2020.Changes in regulations or enforcement might lead to fluctuations in the number of reported incidents.

Geographical Analysis

```
library(readxl)
library(ggplot2)
library(dplyr)
```

```r
# Group by province and count the number of incidents per province
incident_geo <- as.data.frame(table(cleaned_data$Province))
names(incident_geo) <- c("Province", "Incidents")


# Plot the geographical distribution of incidents
ggplot(data = incident_geo, aes(x = reorder(Province, -Incidents), y = Incidents)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Geographical Distribution of Pipeline Incidents by Province",
       x = "Province", y = "Number of Incidents") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



This bar chart shows the distribution of pipeline incidents across different provinces. It highlights that "Alberta" and "British Columbia" provinces have experienced more incidents compared to others.

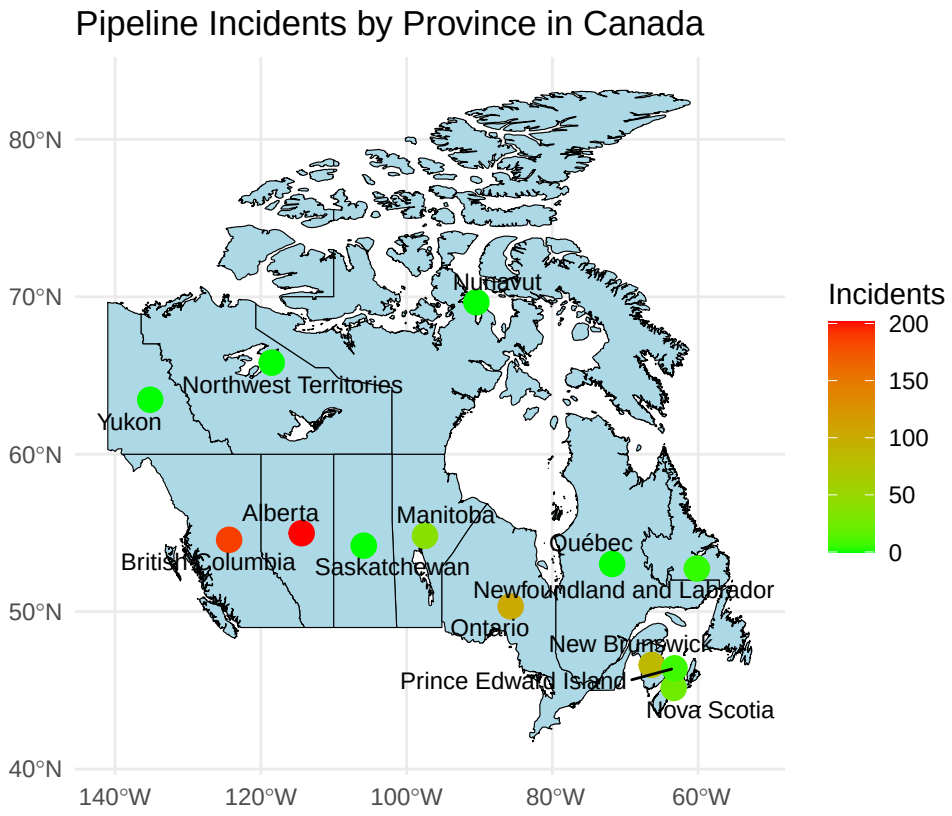
Pipeline Incidents by Province in Canada

```r
library(ggplot2)
library(rnaturalearth)
library(sf)
library(ggrepel)

# Group by province and count the number of incidents per province
incident_geo <- as.data.frame(table(cleaned_data$Province))
```

```r
names(incident_geo) <- c("Province", "Incidents")

# Get the map of Canada
canada_map <- ne_states(country = "canada", returnclass = "sf")

# Province name matching (modify if necessary)
province_name_map <- c(
  "Alberta" = "Alberta",
  "British Columbia" = "British Columbia",
  "Manitoba" = "Manitoba",
  "New Brunswick" = "New Brunswick",
  "Newfoundland and Labrador" = "Newfoundland and Labrador",
  "Nova Scotia" = "Nova Scotia",
  "Ontario" = "Ontario",
  "Prince Edward Island" = "Prince Edward Island",
  "Quebec" = "Quebec",
  "Saskatchewan" = "Saskatchewan"
)

# Merge the incidents data with the Canada map
incident_geo$Province <- province_name_map[incident_geo$Province]
canada_map <- merge(canada_map, incident_geo, by.x = "name", by.y = "Province", all.x = TRUE)

# Replace NAs with 0 incidents
canada_map$Incidents[is.na(canada_map$Incidents)] <- 0

# Calculate centroids for province names
canada_map_centroids <- st_centroid(canada_map)
```

```
## Warning: st_centroid assumes attributes are constant over geometries
```

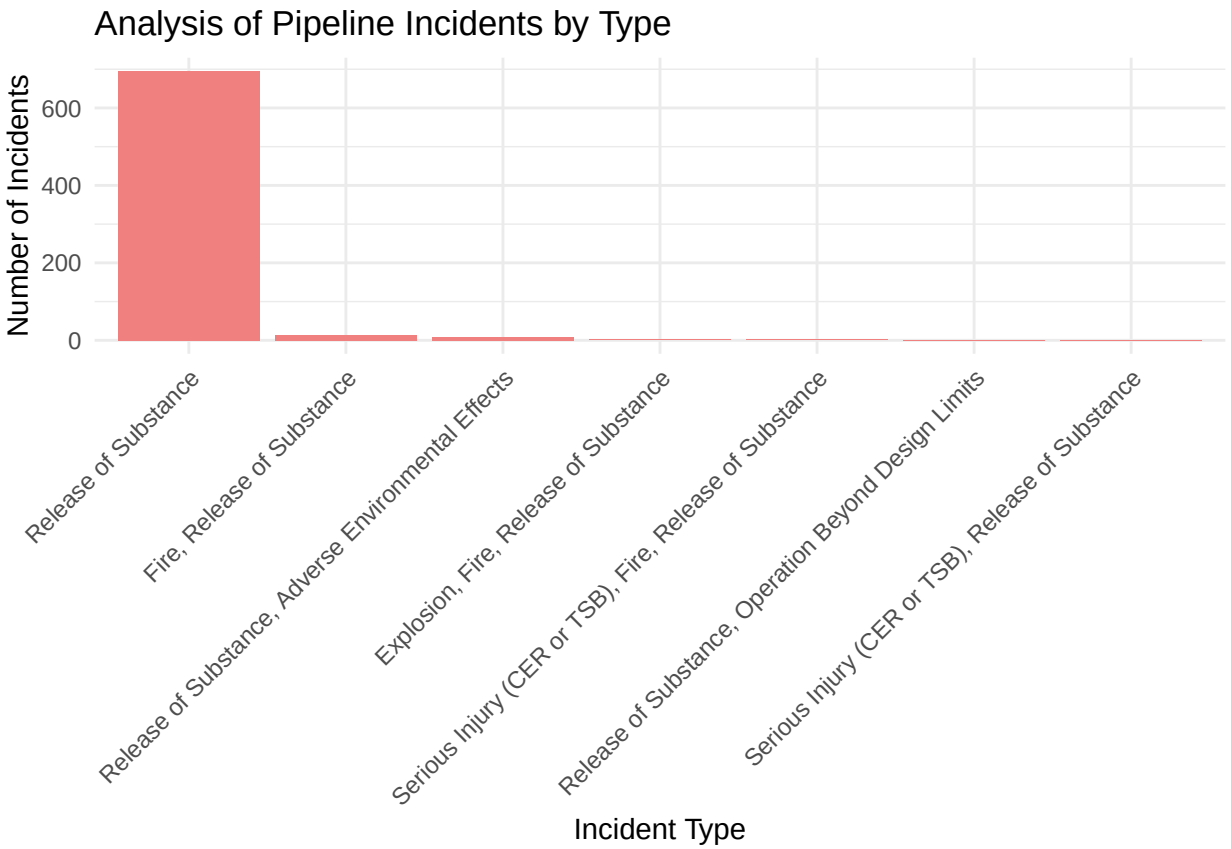
```r
# Convert centroids to a data frame with coordinates for scatter plot
centroid_coords <- st_coordinates(canada_map_centroids)
centroid_df <- data.frame(Province = canada_map$name, Incidents = canada_map$Incidents,
                          X = centroid_coords[, 1], Y = centroid_coords[, 2])

# Plot the map with scatter points and labels using geom_text_repel to avoid overlaps
ggplot() +
  geom_sf(data = canada_map, fill = "lightblue", color = "black") +  # Background map
  geom_point(data = centroid_df, aes(x = X, y = Y, color = Incidents), size = 4) +  # Scatter plot poin
  geom_text_repel(data = centroid_df, aes(x = X, y = Y, label = Province), size = 3, color = "black") +
  scale_color_gradient(low = "green", high = "red", name = "Incidents") +  # Color gradient based on in
  labs(title = "Pipeline Incidents by Province in Canada", x = "", y = "") +
  theme_minimal()
```



This heatmap provides the clear picture about the frequency of incidents according to the province which may shows that the province that has higher population densities have more pipeline infrastructure, leading to a higher likelihood of incidents.

Incident type analysis

```r
# Load necessary libraries
library(readxl)
library(ggplot2)
library(dplyr)

# Group by 'Incident Types' and count the number of incidents for each type
incident_type_analysis <- as.data.frame(table(cleaned_data$`Incident Types`))
names(incident_type_analysis) <- c("Incident_Type", "Incidents")

# Plot the analysis of incidents by type
ggplot(data = incident_type_analysis, aes(x = reorder(Incident_Type, -Incidents), y = Incidents)) +
  geom_bar(stat = "identity", fill = "lightcoral") +
  labs(title = "Analysis of Pipeline Incidents by Type", x = "Incident Type", y = "Number of Incidents"
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



This bar chart categorizes pipeline incidents based on their types. It provides insights into the most common types of incidents occurring in pipelines and in this case Release of Substances is the most common.

Incidents by Company and Province

```
# Load necessary libraries
library(readxl)
library(ggplot2)
library(dplyr)
library(plotly)
```

```
## Warning: package 'plotly' was built under R version 4.3.3
```

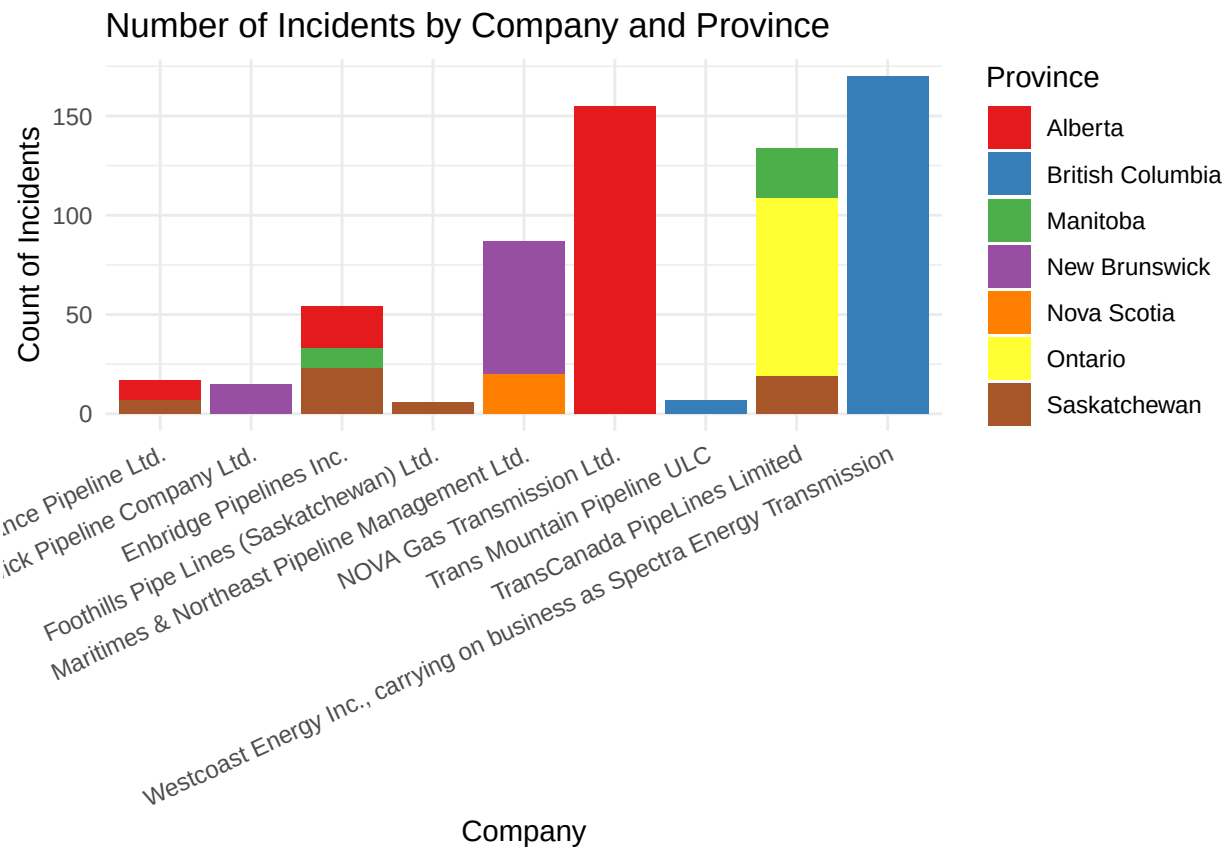
```
##
## Attaching package: 'plotly'
```

```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following object is masked from 'package:graphics':
##
##     layout
```

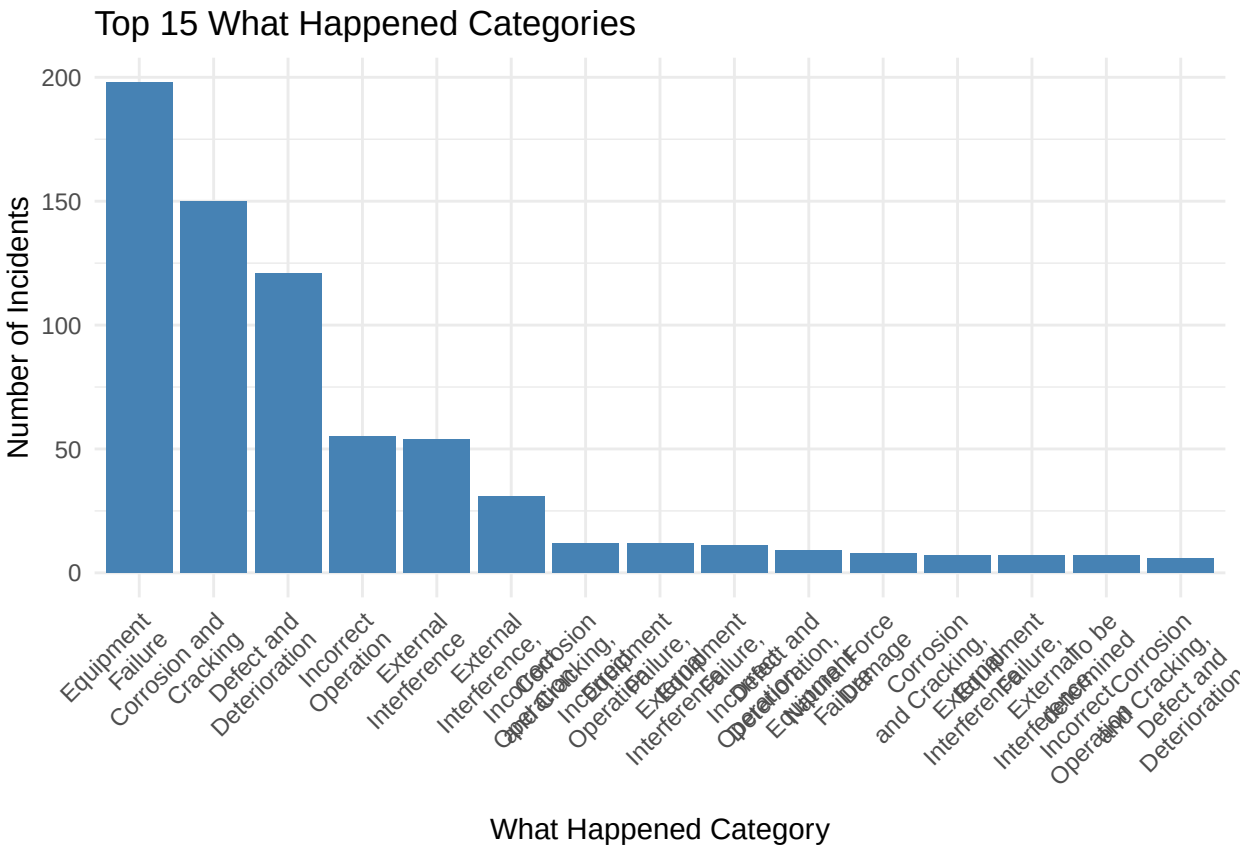
```r
# Summarize the number of incidents by Company and Province
cleaned_summary <- cleaned_data %>%
  group_by(Company, Province) %>%
  summarise(Incident_Count = n(), .groups = 'drop')

# Filter to keep only the top N categories (e.g., top 15)
top_n <- 15  # Change this to the desired number of top categories
top_categories <- cleaned_summary %>%
  top_n(top_n, Incident_Count) %>%
  arrange(desc(Incident_Count))


# Stacked bar plot of incidents by Company and Province
ggplot(top_categories, aes(x = Company, y = Incident_Count, fill = Province)) +
  geom_bar(stat = "identity") +  # Default is 'stack', so no need to specify position
  ggtitle("Number of Incidents by Company and Province") +
  xlab("Company") +
  ylab("Count of Incidents") +
  theme_minimal() +
  scale_fill_brewer(palette = "Set1")+
  theme(axis.text.x = element_text(angle = 25, hjust = 1))
```



This stacked bar chart displays the number of pipeline incidents for each company, further broken down by province. It helps identify which companies and provinces have higher incident counts.

Number of Incidents for top 15 What happened Categories

```r
library(dplyr)
library(ggplot2)

# Count occurrences of each "What happened category"
category_counts <- cleaned_data %>%
  group_by(`What happened category`) %>%
  summarise(Incident_Count = n()) %>%
  ungroup()

library(stringr)

# Use str_wrap to wrap long category names to a fixed width
category_counts$`What happened category` <- str_wrap(category_counts$`What happened category`, width = 


# Filter to keep only the top N categories (e.g., top 5)
top_n <- 15  # Change this to the desired number of top categories
top_categories <- category_counts %>%
  top_n(top_n, Incident_Count) %>%
  arrange(desc(Incident_Count))

# Create a histogram for the top categories
ggplot(top_categories, aes(x = reorder(`What happened category`, -Incident_Count), y = Incident_Count)) 
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(title = paste("Top", top_n, "What Happened Categories"),
       x = "What Happened Category", y = "Number of Incidents") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate x-axis labels for better readabili
```

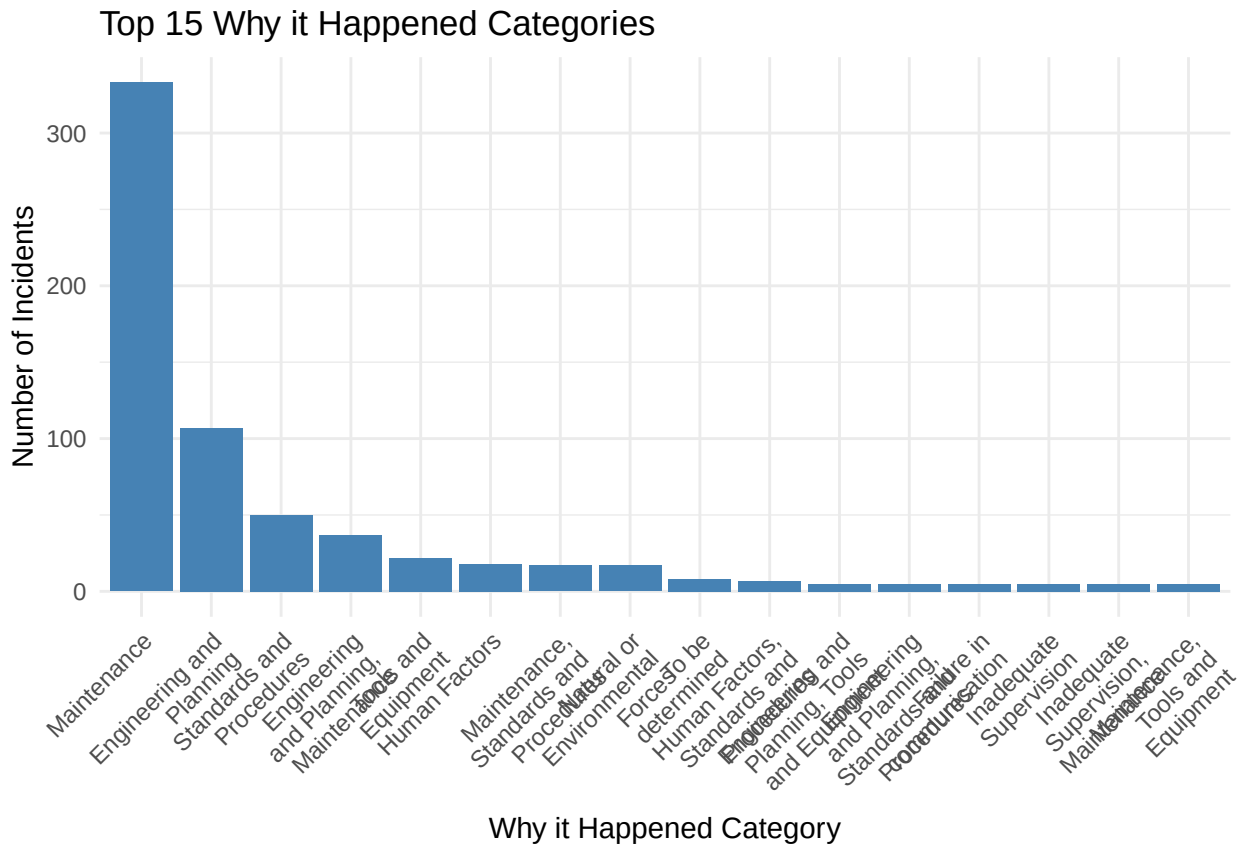


By categorizing incidents based on what happened, we can identify the most frequent scenarios leading to pipeline failures. The bar chart shows top 15 what happened categories where "Equipment Failure" , "Corrosion and Cracking" and "Defect and Deterioration" are the most common categories, it indicates that these are prevalent issues in pipeline operations compare to others.

Number of Incidents for top 15 Why it happened Categories

```
library(dplyr)
library(ggplot2)

# Count occurrences of each "Why it happened category"
category_counts <- cleaned_data %>%
  group_by(`Why it happened category`) %>%
  summarise(Incident_Count = n()) %>%
  ungroup()

library(stringr)

# Use str_wrap to wrap long category names to a fixed width
category_counts$`Why it happened category` <- str_wrap(category_counts$`Why it happened category`, width


# Filter to keep only the top N categories (e.g., top 5)
top_n <- 15  # Change this to the desired number of top categories
top_categories <- category_counts %>%
```

```
  top_n(top_n, Incident_Count) %>%
  arrange(desc(Incident_Count))

# Create a histogram for the top categories
ggplot(top_categories, aes(x = reorder(`Why it happened category`, -Incident_Count), y = Incident_Count)
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(title = paste("Top", top_n, "Why it Happened Categories"),
       x = "Why it Happened Category", y = "Number of Incidents") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate x-axis labels for better readabili
```



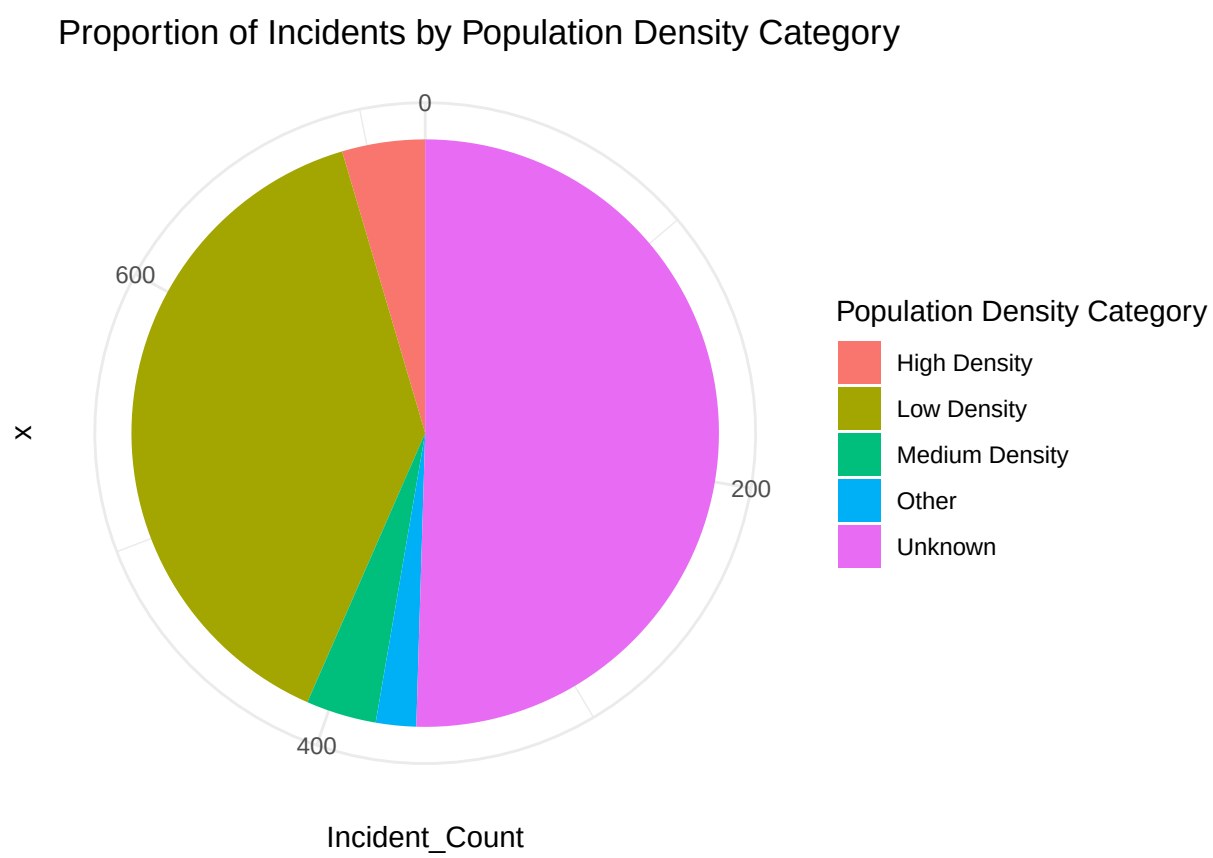
This bar chart displays the top 15 categories explaining why the pipeline incidents occurred. It helps understand the common causes behind these incidents.the bar chart shows that "maintenance" is the main cause which lead to the highest number of incidents ,more than 300 , where the second main cause is "Engineering and Planning".

Regulatory bodies can use this analysis to ensure compliance with safety standards.

Proportion of population density according to incident number

```
library(dplyr)
library(ggplot2)


# Clean the Population Density data
```

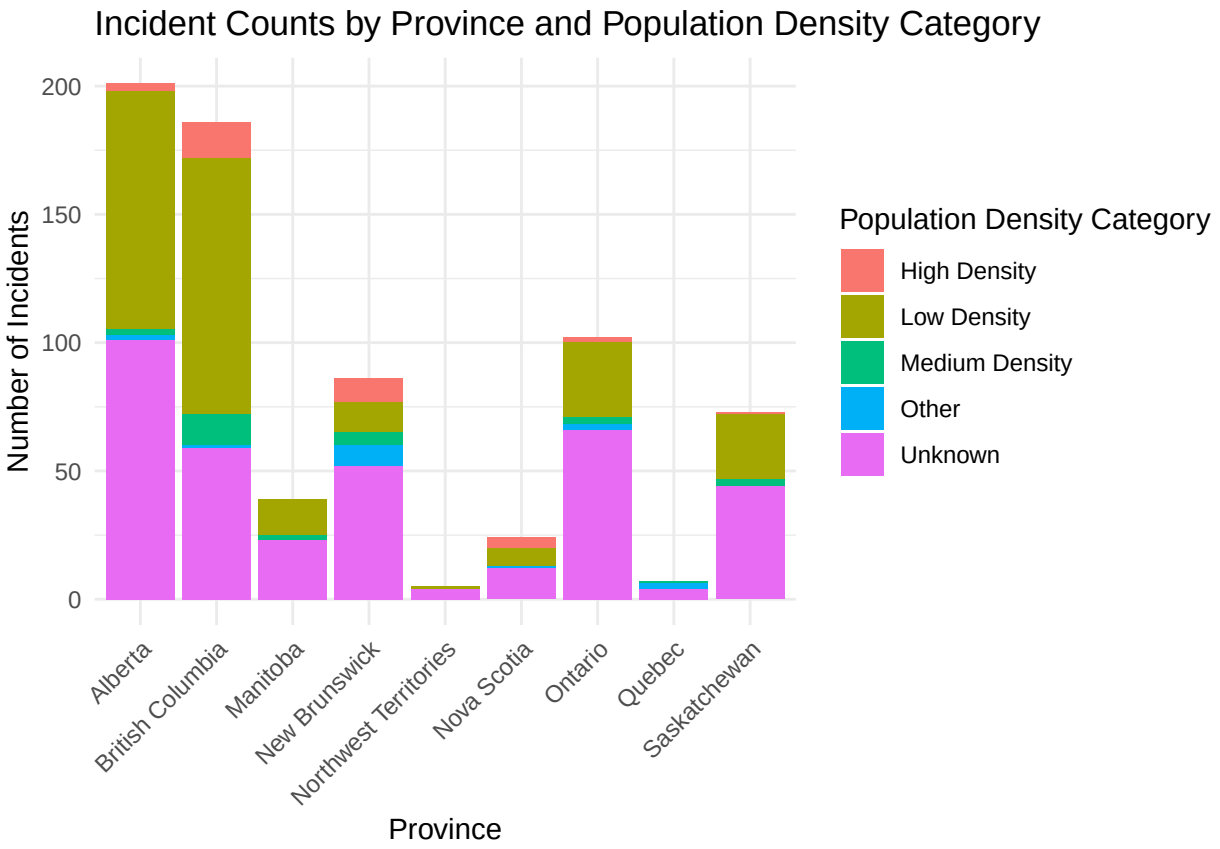
```r
cleaned_data <- cleaned_data %>%
  mutate(Population_Density_Cleaned = case_when(
    grepl("Unknown Population Density", `Population Density`) ~ "Unknown",
    grepl("10 or fewer dwelling units", `Population Density`) ~ "Low Density",
    grepl("11 to 45 dwelling units", `Population Density`) ~ "Medium Density",
    grepl("46 or more dwelling units", `Population Density`) ~ "High Density",
    TRUE ~ "Other"  # Optional for any other cases
  ))
# Count incidents per cleaned population density category
density_counts <- cleaned_data %>%
  group_by(Population_Density_Cleaned) %>%
  summarise(Incident_Count = n()) %>%
  ungroup()

# Create a pie chart for the proportion of incidents by cleaned population density category
ggplot(density_counts, aes(x = "", y = Incident_Count, fill = Population_Density_Cleaned)) +
  geom_col() +
  coord_polar(theta = "y") +
  labs(title = "Proportion of Incidents by Population Density Category",
       fill = "Population Density Category") +
  theme_minimal()
```



Analyzing the proportion of pipeline incidents by population density can provide valuable insights into how population density affects the frequency and nature of these incidents. This pie chart shows that incidents are more frequent in the low density region (rural area).

Relation between province and population Density with incident number

```r
library(dplyr)
library(ggplot2)

# Clean the Population Density data and summarize incident counts
cleaned_data <- cleaned_data %>%
  mutate(Population_Density_Cleaned = case_when(
    grepl("Unknown Population Density", `Population Density`) ~ "Unknown",
    grepl("10 or fewer dwelling units", `Population Density`) ~ "Low Density",
    grepl("11 to 45 dwelling units", `Population Density`) ~ "Medium Density",
    grepl("46 or more dwelling units", `Population Density`) ~ "High Density",
    TRUE ~ "Other"  # Optional for any other cases
  ))


# Count incidents by Province and Population Density Category
incident_counts_by_density <- cleaned_data %>%
  group_by(Province, Population_Density_Cleaned) %>%
  summarise(Incident_Count = n(), .groups = 'drop')

# Create a stacked bar chart for incident counts by province and population density category
ggplot(incident_counts_by_density, aes(x = Province, y = Incident_Count, fill = Population_Density_Clea
  geom_bar(stat = "identity") +
  labs(title = "Incident Counts by Province and Population Density Category",
       x = "Province",
       y = "Number of Incidents",
       fill = "Population Density Category") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate x-axis labels for better readabili
```



By grouping incidents by province and population density, stakeholders can gain a comprehensive understanding of the factors influencing pipeline safety in different regions.This chart shows that all provinces follow same pattern that low density region (i.e. rural area) has the biggest proportion other than unknown.

Conclusion:

By incorporating these insights into emergency response planning, stakeholders can enhance the safety and preparedness of densely populated areas, reducing the impact of pipeline incidents on public health and safety.

Modelling Part

```
#Summarize data by Nearest Populated Center
full_summary <- cleaned_data %>%
group_by(`Nearest Populated Centre`) %>%
summarise(incident_count = n(),
Population_Density = first(`Population Density`),
Latitude = first(Latitude.x),
Longitude = first(Longitude.x)) %>%
ungroup()
full_summary_clean <- na.omit(full_summary)
full_summary[complete.cases(full_summary[, c("incident_count", "Nearest Populated Centre", "Population_
```

```
## # A tibble: 328 x 5
##    Nearest Populated Cent~1 incident_count Population_Density Latitude Longitude
##    <chr>                             <int> <chr>                    <dbl>     <dbl>
```

```
##  1 .5 km                                1 11 to 45 dwelling~     56.1    -121.
##  2 10.8 km south of Tumble~             1 10 or fewer dwell~     55.0    -121.
##  3 100 Mile House                       1 10 or fewer dwell~     51.6    -121.
##  4 15000                                2 11 to 45 dwelling~     47.4     -68.3
##  5 23 km SW of Chetwynd, BC             1 Unknown Populatio~     55.6    -122.
##  6 4000                                 1 11 to 45 dwelling~     56.8    -120.
##  7 9000                                 1 10 or fewer dwell~     45.9     -66.6
##  8 Abbotsford                           2 Unknown Populatio~     49.1    -122.
##  9 Agassiz                              1 10 or fewer dwell~     49.4    -121.
## 10 Airdrie                              1 Buildings greater~     51.4    -114.
## # i 318 more rows
## # i abbreviated name: 1: `Nearest Populated Centre`
```

```
# Convert Nearest Populated Center to factor if it isn't already

full_summary_clean$`Nearest Populated Centre` <- as.factor(full_summary_clean$`Nearest Populated Centre`
# Ensure other variables are numeric
full_summary_clean$Population_Density <- as.factor(full_summary_clean$Population_Density)
full_summary_clean$Latitude <- as.numeric(full_summary_clean$Latitude)
full_summary_clean$Longitude <- as.numeric(full_summary_clean$Longitude)

# Now apply the summarise again after cleaning
incident_rate_data <- full_summary_clean %>%
group_by(`Nearest Populated Centre`) %>%
summarise(
total_incidents = sum(incident_count),
Population_Density = mean(as.numeric(Population_Density), na.rm = TRUE), # Convert and handle NAs
Latitude = mean(as.numeric(Latitude), na.rm = TRUE), # Convert and handle NAs
Longitude = mean(as.numeric(Longitude), na.rm = TRUE) # Convert and handle NAs
)
str(incident_rate_data)
```

```
## tibble [328 x 5] (S3: tbl_df/tbl/data.frame)
##  $ Nearest Populated Centre: Factor w/ 328 levels ".5 km","10.8 km south of Tumbler Ridge, BC",..: 1
##  $ total_incidents         : int [1:328] 1 1 1 2 1 1 1 2 1 1 ...
##  $ Population_Density      : num [1:328] 2 1 1 2 5 2 1 5 1 4 ...
##  $ Latitude                : num [1:328] 56.1 55 51.6 47.4 55.6 ...
##  $ Longitude               : num [1:328] -120.7 -121 -121.3 -68.3 -121.9 ...
```

```
# Poisson regression model after cleaning
poisson_model_rate <- glm(total_incidents ~ Population_Density + Latitude + Longitude,
data = incident_rate_data, family = poisson())
#Model summary
summary(poisson_model_rate)
```

```
##
## Call:
## glm(formula = total_incidents ~ Population_Density + Latitude +
##     Longitude, family = poisson(), data = incident_rate_data)
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -2.934855   0.544413  -5.391 7.01e-08 ***
## Population_Density  0.172653   0.020705   8.339  < 2e-16 ***
## Latitude            0.081159   0.015758   5.150 2.60e-07 ***
## Longitude           0.009937   0.003723   2.669  0.00761 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 780.88  on 327  degrees of freedom
## Residual deviance: 684.98  on 324  degrees of freedom
## AIC: 1474.4
##
## Number of Fisher Scoring iterations: 6
```

Null deviance = 780.88: This is the deviance of the model without any predictors (just the intercept). Residual deviance = 684.97: This shows the deviance after including the predictors (Population_Density, Latitude, Longitude). Lower deviance values indicate a better fit. The residual deviance is relatively high compared to the null deviance, suggesting that while the model explains some variability in incident counts, there may be other unaccounted factors. AIC = 1474.4: The Akaike Information Criterion (AIC) is a measure of model quality. Lower AIC values indicate a better fit, this can be compared to other models to assess relative performance. Reducing the number of predictors might get rid of overfitting problem.

```r
# Calculate the dispersion statistic
dispersion_statistic <- sum(residuals(poisson_model_rate , type = "pearson")^2) / 327
dispersion_statistic
```

```
## [1] 4.003258
```

It is bigger than one, so there is overdispersion.