Project 3

Parham & Banafshe

2022-07-15

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

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

```
data <- read_csv("C:\\Users\\Parham\\Desktop\\projectData.csv")
```

```
## Rows: 1624 Columns: 16

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (13): Incident.Number, Reported.Date, Nearest.Populated.Centre, Province...
## dbl  (3): Latitude, Longitude, Year
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
data <- data%>%
  rename(SubstanceRelease = `Substance release`)
glimpse(data)
```

```
## Rows: 1,624
## Columns: 16
## $ Incident.Number                <chr> "INC2007-097", "INC2008-001", "INC200~
## $ Reported.Date                  <chr> "01/02/2008", "01/02/2008", "01/23/20~
## $ Nearest.Populated.Centre       <chr> "Grande Prairie", "Cromer", "Cromer",~
## $ Province                       <chr> "Alberta", "Manitoba", "Manitoba", "B~
## $ Company                        <chr> "Alliance Pipeline Ltd.", "Enbridge P~
## $ Status                         <chr> "Closed", "Closed", "Closed", "Closed~
## $ Latitude                       <dbl> 54.84000, 49.73135, 49.73135, 58.0120~
## $ Longitude                      <dbl> -118.65000, -101.23557, -101.23557, -~
```

```
## $ Approximate.Volume.Released..m3. <chr> "Not Provided", "8.0000", "100.0000",~
## $ Substance                         <chr> "Natural Gas - Sweet", "Crude Oil - S~
## $ Release.Type                      <chr> "Gas", "Liquid", "Liquid", "Gas", "Mi~
## $ Significant                       <chr> "No", "No", "No", "No", "Yes", "No", ~
## $ Year                              <dbl> 2008, 2008, 2008, 2008, 2008, 2008, 2~
## $ What.Happened                     <chr> "Corrosion and Cracking", "Corrosion ~
## $ Why.It.Happened                   <chr> "Maintenance", "Maintenance", "Mainte~
## $ SubstanceRelease                  <chr> "Yes", "Yes", "Yes", "Yes", "Yes", "Y~
```

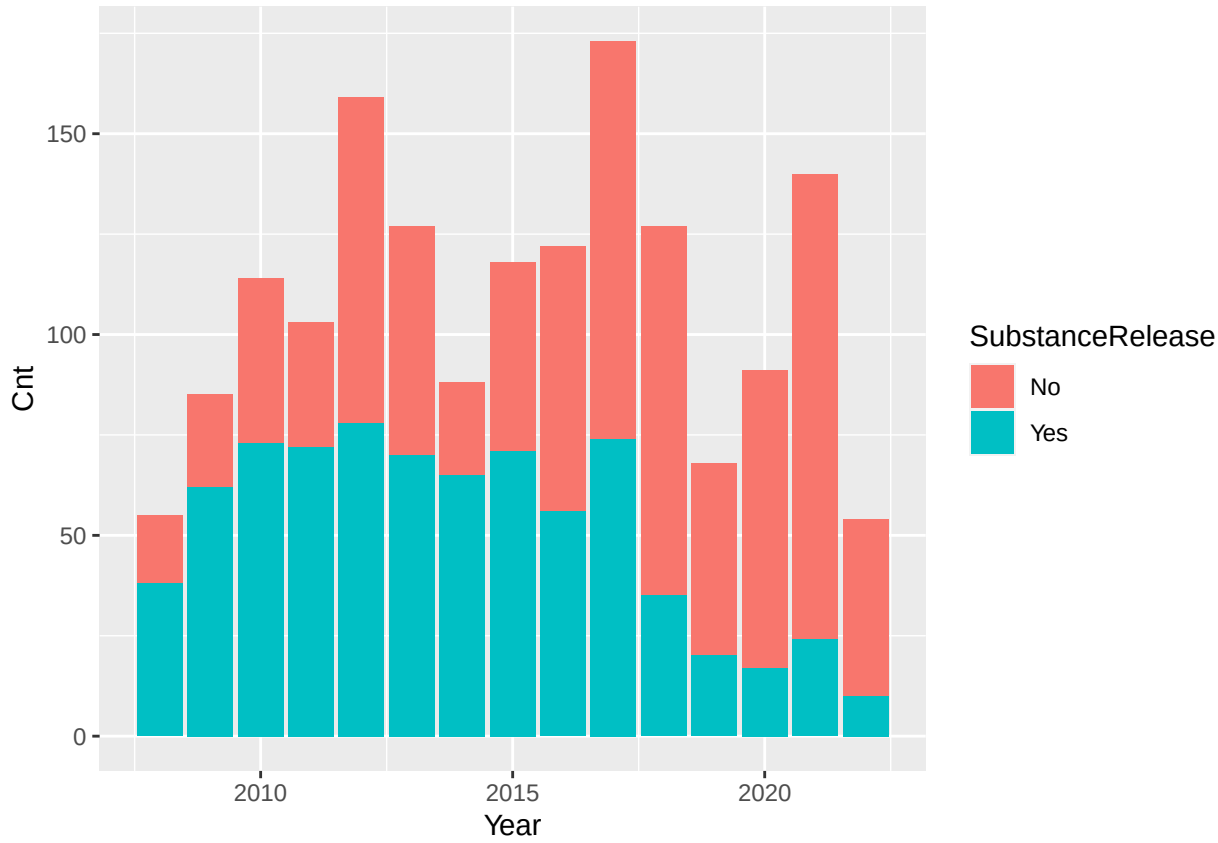
```
t<-data%>%
  group_by(Year,SubstanceRelease)%>%
  summarize(Cnt = n())
```

```
## `summarise()` has grouped output by 'Year'. You can override using the
## `.groups` argument.
```

```
t
```

```
## # A tibble: 30 x 3
## # Groups:   Year [15]
##     Year SubstanceRelease   Cnt
##    <dbl> <chr>            <int>
##  1  2008 No                  17
##  2  2008 Yes                 38
##  3  2009 No                  23
##  4  2009 Yes                 62
##  5  2010 No                  41
##  6  2010 Yes                 73
##  7  2011 No                  31
##  8  2011 Yes                 72
##  9  2012 No                  81
## 10  2012 Yes                 78
## # ... with 20 more rows
```

```
t%>%
  ggplot(aes(x=Year, y=Cnt,fill=SubstanceRelease)) +
  geom_bar(stat="identity")
```




```
t<-pivot_wider(
  t,
  names_from = SubstanceRelease,
  values_from = `Cnt`,
)
t
```

```
## # A tibble: 15 x 3
## # Groups:   Year [15]
##     Year    No   Yes
##    <dbl> <int> <int>
##  1  2008    17    38
##  2  2009    23    62
##  3  2010    41    73
##  4  2011    31    72
##  5  2012    81    78
##  6  2013    57    70
##  7  2014    23    65
##  8  2015    47    71
##  9  2016    66    56
## 10  2017    99    74
## 11  2018    92    35
## 12  2019    48    20
## 13  2020    74    17
## 14  2021   116    24
## 15  2022    44    10
```

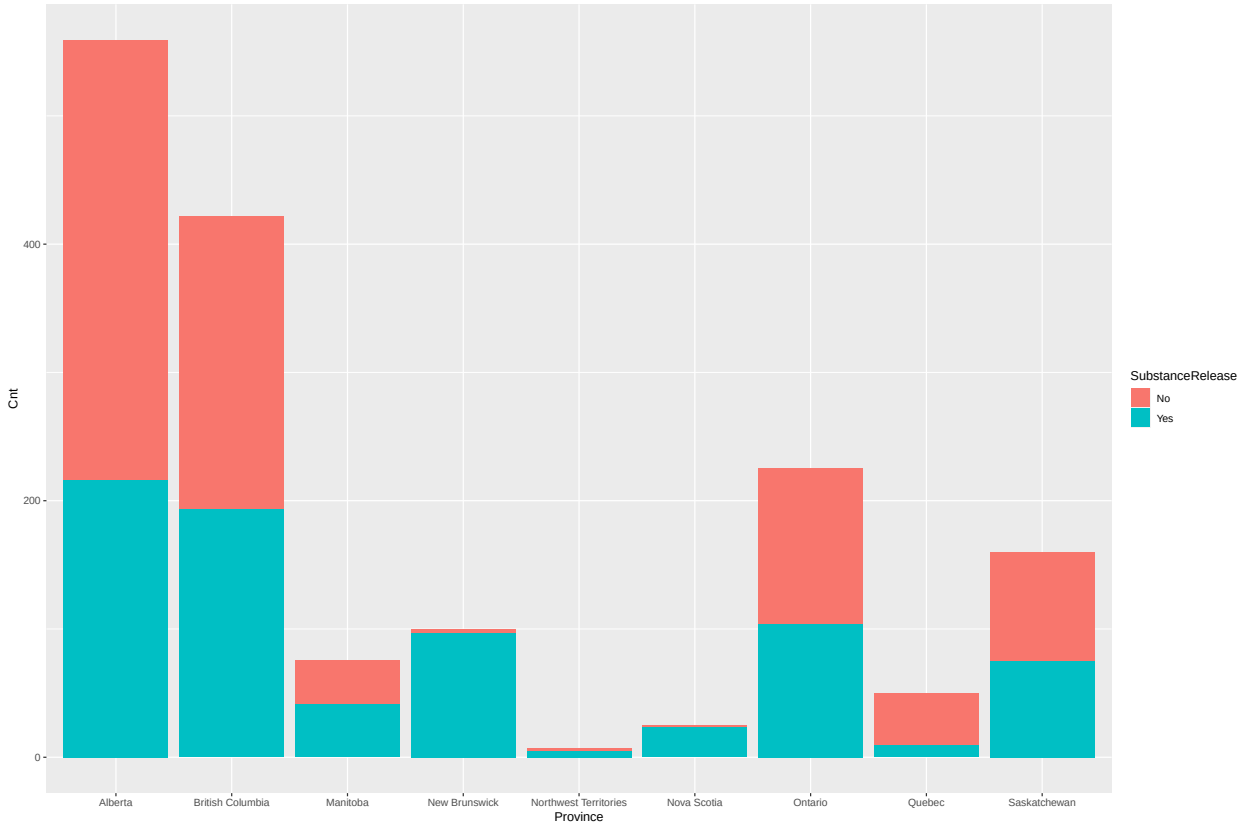
```r
t2<-data%>%
  group_by(Province,SubstanceRelease)%>%
  summarize(Cnt = n())%>%
  arrange(desc(Cnt))
```

```
## `summarise()` has grouped output by 'Province'. You can override using the
## `.groups` argument.
```

```r
t2
```

```
## # A tibble: 18 x 3
## # Groups:   Province [9]
##    Province              SubstanceRelease   Cnt
##    <chr>                 <chr>            <int>
##  1 Alberta               No                 343
##  2 British Columbia      No                 229
##  3 Alberta               Yes                216
##  4 British Columbia      Yes                193
##  5 Ontario               No                 121
##  6 Ontario               Yes                104
##  7 New Brunswick         Yes                 97
##  8 Saskatchewan          No                  85
##  9 Saskatchewan          Yes                 75
## 10 Manitoba              Yes                 42
## 11 Quebec                No                  41
## 12 Manitoba              No                  34
## 13 Nova Scotia           Yes                 24
## 14 Quebec                Yes                  9
## 15 Northwest Territories Yes                  5
## 16 New Brunswick         No                   3
## 17 Northwest Territories No                   2
## 18 Nova Scotia           No                   1
```

```r
t2%>%
  ggplot(aes(x=Province, y=Cnt,fill=SubstanceRelease)) +
  geom_bar(stat="identity")
```



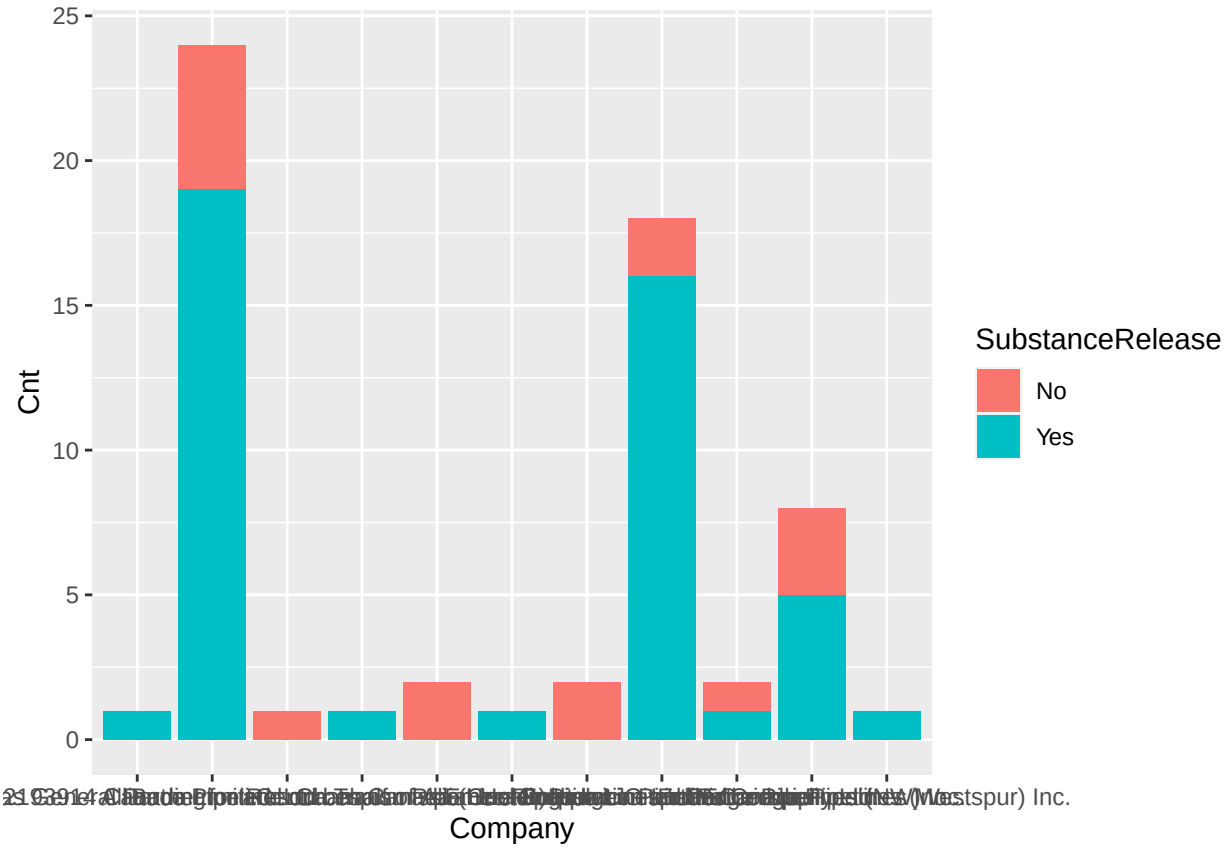

```
t3<-data%>%
  group_by(Company,SubstanceRelease)%>%
  summarize(Cnt = n())
```

```
## `summarise()` has grouped output by 'Company'. You can override using the
## `.groups` argument.
```

```
t3
```

```
## # A tibble: 74 x 3
## # Groups:   Company [50]
##    Company                                                SubstanceRelease   Cnt
##    <chr>                                                  <chr>            <int>
##  1 2193914 Canada Limited                                 Yes                  1
##  2 Alliance Pipeline Ltd.                                 No                   5
##  3 Alliance Pipeline Ltd.                                 Yes                 19
##  4 Alliance Pipeline Ltd., as General Partner for and on~ No                   1
##  5 Burlington Resources Canada (Hunter) Ltd.              Yes                  1
##  6 Centra Transmission Holdings Inc.                      No                   2
##  7 Champion Pipe Line Corporation Limited                 Yes                  1
##  8 Cochin Pipe Lines Ltd.                                 No                   2
##  9 Emera Brunswick Pipeline Company Ltd.                  No                   2
## 10 Emera Brunswick Pipeline Company Ltd.                  Yes                 16
## # ... with 64 more rows
```

```
t3[1:15,]%>%
  ggplot(aes(x=Company, y=Cnt,fill=SubstanceRelease)) +
  geom_bar(stat="identity")
```



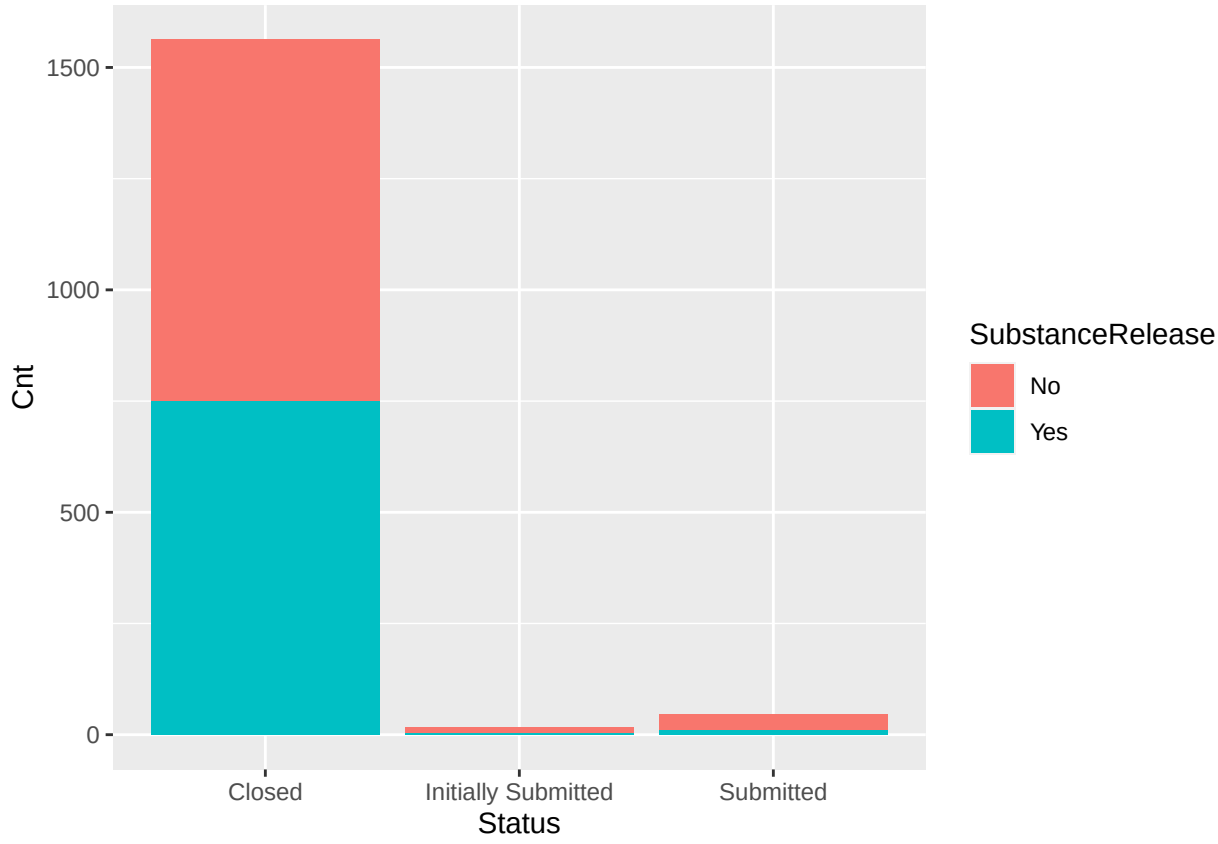

```
t4<-data%>%
  group_by(Status,SubstanceRelease)%>%
  summarize(Cnt = n())%>%
  arrange(desc(Cnt))
```

```
## `summarise()` has grouped output by 'Status'. You can override using the
## `.groups` argument.
```

```
t4
```

```
## # A tibble: 6 x 3
## # Groups:   Status [3]
##   Status              SubstanceRelease   Cnt
##   <chr>               <chr>            <int>
## 1 Closed              No                 812
## 2 Closed              Yes                750
## 3 Submitted           No                  34
## 4 Initially Submitted No                  13
## 5 Submitted           Yes                 11
## 6 Initially Submitted Yes                  4
```

```
t4%>%
  ggplot(aes(x=Status, y=Cnt,fill=SubstanceRelease)) +
  geom_bar(stat="identity")
```



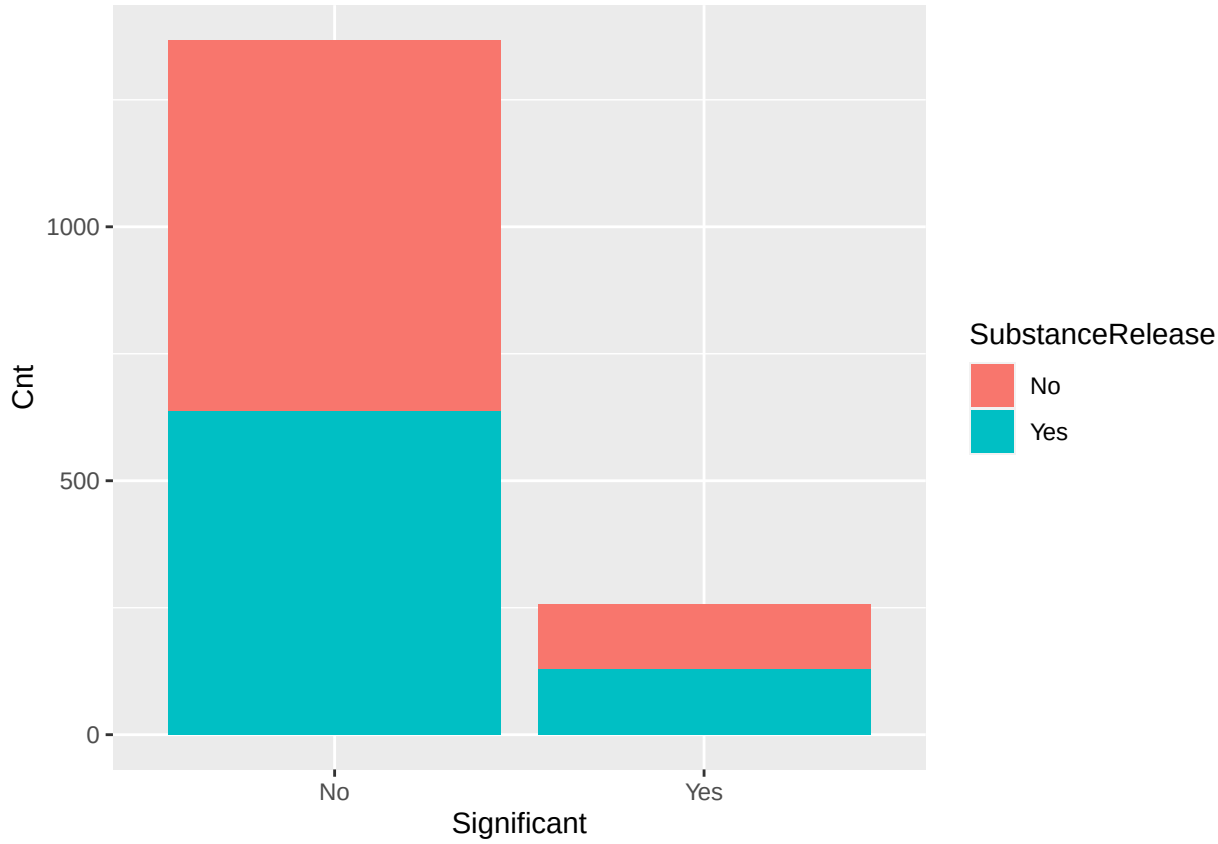

```
t5<-data%>%
  group_by(Significant,SubstanceRelease)%>%
  summarize(Cnt = n())%>%
  arrange(desc(Cnt))
```

```
## `summarise()` has grouped output by 'Significant'. You can override using the
## `.groups` argument.
```

```
t5
```

```
## # A tibble: 4 x 3
## # Groups:   Significant [2]
##   Significant SubstanceRelease   Cnt
##   <chr>       <chr>            <int>
## 1 No          No                 732
## 2 No          Yes                636
## 3 Yes         Yes                129
## 4 Yes         No                 127
```

```
t5%>%
  ggplot(aes(x=Significant, y=Cnt,fill=SubstanceRelease)) +
  geom_bar(stat="identity")
```



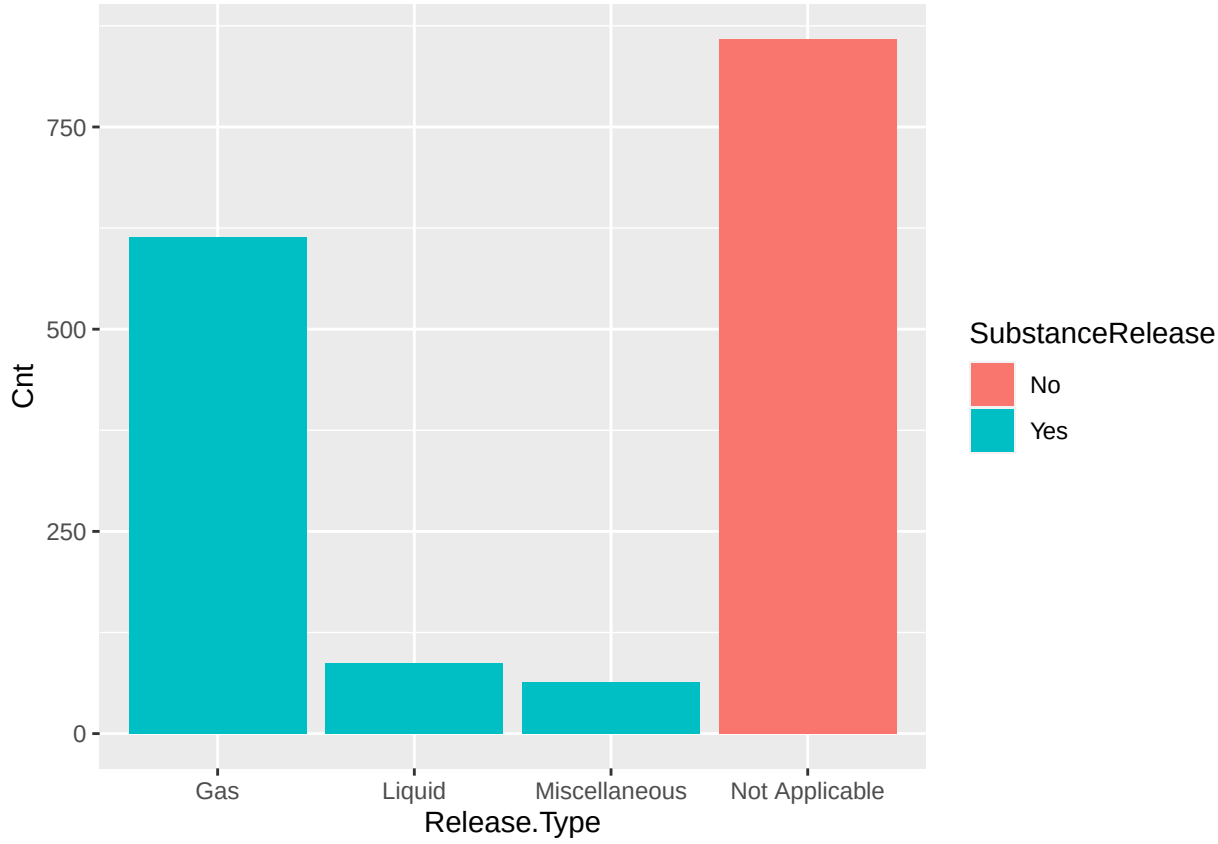

```
t6<-data%>%
  group_by(Release.Type,SubstanceRelease)%>%
  summarize(Cnt = n())%>%
  arrange(desc(Cnt))
```

```
## `summarise()` has grouped output by 'Release.Type'. You can override using the
## `.groups` argument.
```

```
t6
```

```
## # A tibble: 4 x 3
## # Groups:   Release.Type [4]
##   Release.Type   SubstanceRelease   Cnt
##   <chr>          <chr>            <int>
## 1 Not Applicable No                 859
## 2 Gas            Yes                614
## 3 Liquid         Yes                 87
## 4 Miscellaneous  Yes                 64
```

```
t6%>%
  ggplot(aes(x=Release.Type, y=Cnt,fill=SubstanceRelease)) +
  geom_bar(stat="identity")
```




```
t7<-data%>%
  group_by(Substance,SubstanceRelease)%>%
  summarize(Cnt = n())%>%
  arrange(desc(Cnt))
```

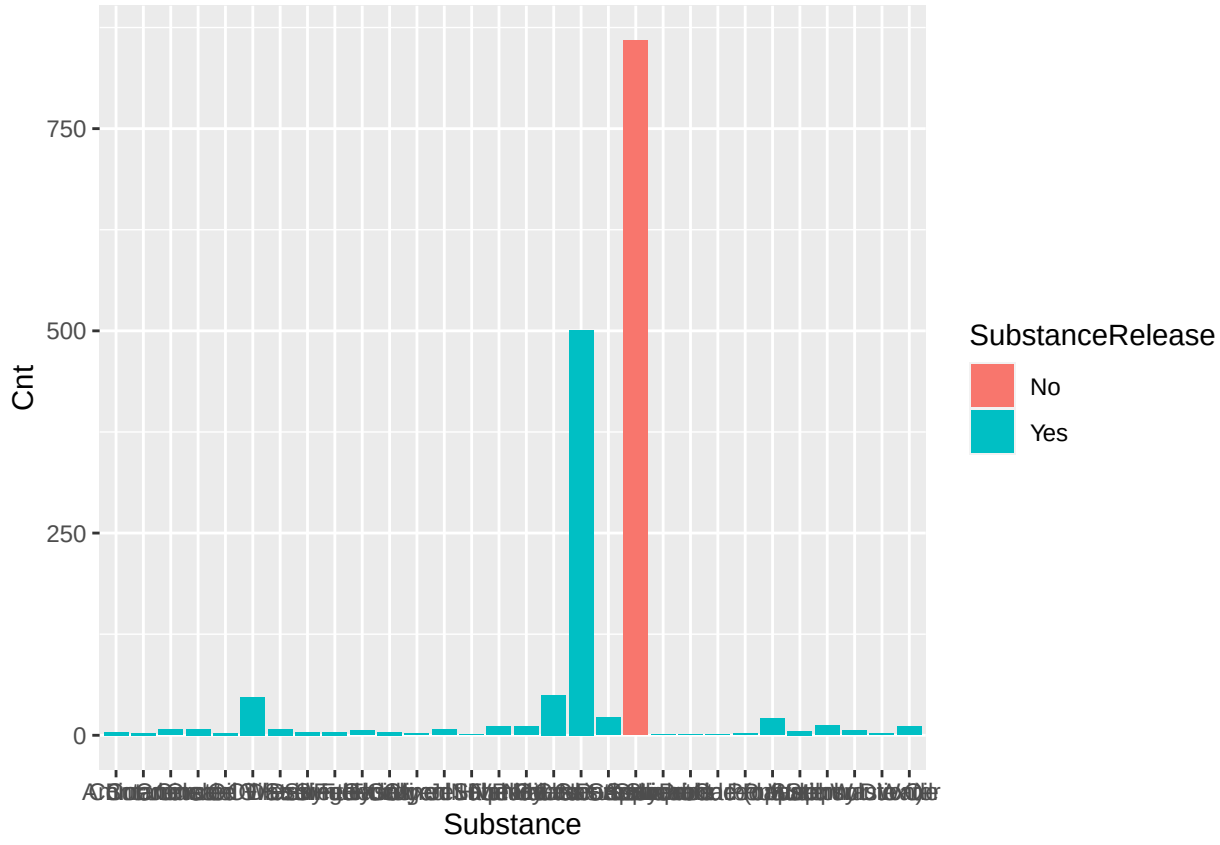
```
## `summarise()` has grouped output by 'Substance'. You can override using the
## `.groups` argument.
```

```
t7
```

```
## # A tibble: 30 x 3
## # Groups:   Substance [30]
##    Substance              SubstanceRelease   Cnt
##    <chr>                  <chr>            <int>
##  1 Not Applicable         No                 859
##  2 Natural Gas - Sweet    Yes                501
##  3 Natural Gas - Sour     Yes                 50
##  4 Crude Oil - Sweet      Yes                 47
##  5 Natural Gas Liquids    Yes                 22
##  6 Propane                Yes                 21
```

```
##  7 Sulphur                Yes                 12
##  8 Lube Oil               Yes                 11
##  9 Mixed HVP Hydrocarbons Yes                 11
## 10 Water                  Yes                 11
## # ... with 20 more rows
```

```
t7%>%
  ggplot(aes(x=Substance, y=Cnt,fill=SubstanceRelease)) +
  geom_bar(stat="identity")
```




```
table(data$SubstanceRelease,data$Release.Type)
```

```
##
##       Gas Liquid Miscellaneous Not Applicable
##   No    0      0             0            859
##   Yes 614     87            64              0
```

```
table(data$SubstanceRelease,data$Significant)
```

```
##
##        No Yes
##   No  732 127
##   Yes 636 129
```

```
chisq.test(table(data$SubstanceRelease,data$Significant))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(data$SubstanceRelease, data$Significant)
## X-squared = 1.1641, df = 1, p-value = 0.2806
```

#modeling with tain and test

```
data<-data%>%
  mutate(SubstanceRelease = ifelse(SubstanceRelease == "Yes",1,0),
         Significant = ifelse(Significant == "Yes",1,0))
n<-nrow(data)
n.train = trunc(0.7*n)
n.test = n - n.train
train = sample(1:n,n.train)
train.x = data[train,-16]
train.y = data[train,16]
test.x = data[-train,-16]
test.y = data[-train,16]

fit1<-glm(SubstanceRelease ~ Latitude + Longitude ,family = binomial(link="logit"),data=cbind(train.x,t
fit1
```

```
##
## Call:  glm(formula = SubstanceRelease ~ Latitude + Longitude, family = binomial(link = "logit"),
##     data = cbind(train.x, train.y))
##
## Coefficients:
## (Intercept)     Latitude    Longitude
##    -5.57499      0.22618      0.05905
##
## Degrees of Freedom: 1135 Total (i.e. Null);  1133 Residual
## Null Deviance:       1571
## Residual Deviance: 1487  AIC: 1493
```

```
summary(fit1)
```

```
##
## Call:
## glm(formula = SubstanceRelease ~ Latitude + Longitude, family = binomial(link = "logit"),
##     data = cbind(train.x, train.y))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8084  -1.0909  -0.6112   1.1779   1.9286
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.574987   0.978319  -5.699 1.21e-08 ***
## Latitude     0.226179   0.029682   7.620 2.54e-14 ***
```

```
## Longitude    0.059047   0.006926   8.526  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1571.0  on 1135  degrees of freedom
## Residual deviance: 1486.5  on 1133  degrees of freedom
## AIC: 1492.5
##
## Number of Fisher Scoring iterations: 4
```

```
yhat<-round(predict.glm(fit1,newdata = test.x,type = "response"))
tb<-table(yhat,as.data.frame(test.y)[,1])
sum(diag(tb))/sum(tb)
```

```
## [1] 0.5614754
```

```
fit2<-glm(SubstanceRelease ~ Latitude + Longitude + Province,family = binomial(link="logit"),data=data)
fit2
```

```
##
## Call:  glm(formula = SubstanceRelease ~ Latitude + Longitude + Province,
##     family = binomial(link = "logit"), data = data)
##
## Coefficients:
##                   (Intercept)                       Latitude
##                    -10.539122                       0.177370
##                     Longitude       ProvinceBritish Columbia
##                      -0.004275                       0.208596
##              ProvinceManitoba          ProvinceNew Brunswick
##                      1.507722                       5.612166
## ProvinceNorthwest Territories            ProvinceNova Scotia
##                     -0.334621                       5.472283
##               ProvinceOntario                 ProvinceQuebec
##                      1.856941                       0.617251
##          ProvinceSaskatchewan
##                      0.992350
##
## Degrees of Freedom: 1623 Total (i.e. Null);  1613 Residual
## Null Deviance:       2246
## Residual Deviance: 1993  AIC: 2015
```

```
summary(fit2)
```

```
##
## Call:
## glm(formula = SubstanceRelease ~ Latitude + Longitude + Province,
##     family = binomial(link = "logit"), data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -2.7527  -1.0015  -0.7138   1.1513   1.8782
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -10.539122   2.067468  -5.098 3.44e-07 ***
## Latitude                       0.177370   0.026630   6.660 2.73e-11 ***
## Longitude                     -0.004275   0.019918  -0.215  0.83004
## ProvinceBritish Columbia       0.208596   0.185287   1.126  0.26025
## ProvinceManitoba               1.507722   0.371907   4.054 5.03e-05 ***
## ProvinceNew Brunswick          5.612166   1.066206   5.264 1.41e-07 ***
## ProvinceNorthwest Territories -0.334621   0.875228  -0.382  0.70222
## ProvinceNova Scotia            5.472283   1.395250   3.922 8.78e-05 ***
## ProvinceOntario                1.856941   0.602669   3.081  0.00206 **
## ProvinceQuebec                 0.617251   0.832450   0.741  0.45840
## ProvinceSaskatchewan           0.992350   0.244525   4.058 4.94e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2245.9  on 1623  degrees of freedom
## Residual deviance: 1993.3  on 1613  degrees of freedom
## AIC: 2015.3
##
## Number of Fisher Scoring iterations: 6
```

```
yhat<-round(predict.glm(fit2,newdata = test.x,type = "response"))
tb<-table(yhat,as.data.frame(test.y)[,1])
sum(diag(tb))/sum(tb)
```

```
## [1] 0.625
```

```
fit3<-glm(SubstanceRelease ~ Latitude + Longitude + Province + Significant,family = binomial(link="logi
fit3
```

```
##
## Call:  glm(formula = SubstanceRelease ~ Latitude + Longitude + Province +
##     Significant, family = binomial(link = "logit"), data = data)
##
## Coefficients:
##                   (Intercept)                       Latitude
##                    -10.545715                       0.174881
##                     Longitude       ProvinceBritish Columbia
##                      -0.005276                       0.163445
##               ProvinceManitoba          ProvinceNew Brunswick
##                       1.496511                       5.659437
## ProvinceNorthwest Territories             ProvinceNova Scotia
##                      -0.323705                       5.520712
##                ProvinceOntario                 ProvinceQuebec
##                       1.884083                       0.655016
##           ProvinceSaskatchewan                    Significant
##                       1.001589                       0.202180
##
```

```
## Degrees of Freedom: 1623 Total (i.e. Null);  1612 Residual
## Null Deviance:       2246
## Residual Deviance: 1991  AIC: 2015
```

```
summary(fit3)
```

```
##
## Call:
## glm(formula = SubstanceRelease ~ Latitude + Longitude + Province +
##     Significant, family = binomial(link = "logit"), data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8206  -1.0142  -0.7175   1.1464   1.8816
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -10.545715   2.068646  -5.098 3.43e-07 ***
## Latitude                       0.174881   0.026711   6.547 5.87e-11 ***
## Longitude                     -0.005276   0.019937  -0.265  0.79128
## ProvinceBritish Columbia       0.163445   0.188194   0.868  0.38512
## ProvinceManitoba               1.496511   0.372227   4.020 5.81e-05 ***
## ProvinceNew Brunswick          5.659437   1.067226   5.303 1.14e-07 ***
## ProvinceNorthwest Territories -0.323705   0.875160  -0.370  0.71147
## ProvinceNova Scotia            5.520712   1.395922   3.955 7.66e-05 ***
## ProvinceOntario                1.884083   0.603197   3.123  0.00179 **
## ProvinceQuebec                 0.655016   0.833231   0.786  0.43180
## ProvinceSaskatchewan           1.001589   0.244888   4.090 4.31e-05 ***
## Significant                    0.202180   0.147992   1.366  0.17189
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2245.9  on 1623  degrees of freedom
## Residual deviance: 1991.5  on 1612  degrees of freedom
## AIC: 2015.5
##
## Number of Fisher Scoring iterations: 6
```

```
yhat<-round(predict.glm(fit3,newdata = test.x,type = "response"))
tb<-table(yhat,as.data.frame(test.y)[,1])
sum(diag(tb))/sum(tb)
```

```
## [1] 0.6229508
```

```
fit4<-glm(SubstanceRelease ~ Release.Type,family = binomial(link="logit"),data=data)
```

```
## Warning: glm.fit: algorithm did not converge
```

```
fit4
```

```
##
## Call:  glm(formula = SubstanceRelease ~ Release.Type, family = binomial(link = "logit"),
##     data = data)
##
## Coefficients:
##                 (Intercept)          Release.TypeLiquid
##                   2.657e+01                    4.472e-06
##  Release.TypeMiscellaneous  Release.TypeNot Applicable
##                   4.664e-06                   -5.313e+01
##
## Degrees of Freedom: 1623 Total (i.e. Null);  1620 Residual
## Null Deviance:       2246
## Residual Deviance: 9.422e-09     AIC: 8
```

```
summary(fit4)
```

```
##
## Call:
## glm(formula = SubstanceRelease ~ Release.Type, family = binomial(link = "logit"),
##     data = data)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -2.409e-06  -2.409e-06  -2.409e-06   2.409e-06   2.409e-06
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 2.657e+01  1.437e+04   0.002    0.999
## Release.TypeLiquid          4.472e-06  4.080e+04   0.000    1.000
## Release.TypeMiscellaneous   4.664e-06  4.678e+04   0.000    1.000
## Release.TypeNot Applicable -5.313e+01  1.882e+04  -0.003    0.998
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2.2459e+03  on 1623  degrees of freedom
## Residual deviance: 9.4218e-09  on 1620  degrees of freedom
## AIC: 8
##
## Number of Fisher Scoring iterations: 25
```

```
yhat<-round(predict.glm(fit4,newdata = test.x,type = "response"))
tb<-table(yhat,as.data.frame(test.y)[,1])
sum(diag(tb))/sum(tb)
```

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