Final project Data 607

Heleine

2023-12-12

This project examines inequities in traffic crashes in terms of motorists versus non motorists, i.e.,pedestrians, bicyclists and motorcyclists against motorists. Using and combining available data, the analysis will explore the level of road casualties in the above mentioned categories and will identify its leading causes. Although the focus is New York City, the data used in this project come from various sources both local, state, federal and international, including the The New York Times, NYPD Traffic Data, Vision Zero , NHTSA, Bureau Of Transportation Statistics, WHO -Global Status on Road Safety Report 2018

Getting Started: Loading libraries

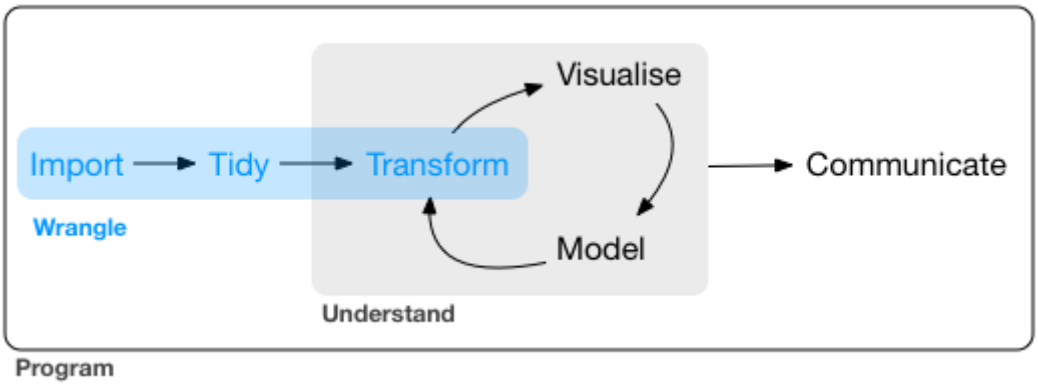


Figure 1: Data Wrangling Model

Importing the Data

```
df1 <- read_csv("https://raw.githubusercontent.com/Heleinef/Data-Science-Master_Heleine/main/Vehicle%20C
```

```
## Rows: 36 Columns: 15
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): GeoCode, GeoCodeLabel
## dbl (13): Year, Number_of_Motor_Vehicle_Collisions, Vehicles_or_Motorists_In...
##
## 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.
```

```
spec(df1)
```

```
## cols(
##   Year = col_double(),
##   GeoCode = col_character(),
```

```
##   GeoCodeLabel = col_character(),
##   Number_of_Motor_Vehicle_Collisions = col_double(),
##   Vehicles_or_Motorists_Involved = col_double(),
##   Injury_or_Fatal_Collisions = col_double(),
##   MotoristsInjured = col_double(),
##   MotoristsKilled = col_double(),
##   PassengInjured = col_double(),
##   PassengKilled = col_double(),
##   CyclistsInjured = col_double(),
##   CyclistsKilled = col_double(),
##   PedestrInjured = col_double(),
##   PedestrKilled = col_double(),
##   Bicycle = col_double()
## )
```

```
VehecileReportStatisticsCitywide <- df1
```

```
df2 <- read_csv("https://raw.githubusercontent.com/Heleinef/Data-Science-Master_Heleine/main/Collisions%
```

```
## Rows: 120 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (4): GeoCode, GeoCodeLabel, ContributingFactorCode, ContributingFactorDe...
## dbl (2): Year, Number_of_Vehicles
##
## 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.
```

```
spec(df2)
```

```
## cols(
##   Year = col_double(),
##   GeoCode = col_character(),
##   GeoCodeLabel = col_character(),
##   ContributingFactorCode = col_character(),
##   ContributingFactorDescription = col_character(),
##   Number_of_Vehicles = col_double()
## )
```

```
CollisionsContributingFactors <- df2
```

Let's take a quick peek at df1

```
glimpse(df1)
```

```
## Rows: 36
## Columns: 15
## $ Year                              <dbl> 2014, 2014, 2014, 2014, 2014, 2014,~
## $ GeoCode                           <chr> "C", "M", "B", "K", "Q", "S", "C", ~
## $ GeoCodeLabel                      <chr> "CITYWIDE", "MANHATTAN", "BRONX", "~
## $ Number_of_Motor_Vehicle_Collisions <dbl> 17720, 4026, 2455, 4960, 5195, 1084~
## $ Vehicles_or_Motorists_Involved    <dbl> 34721, 7672, 4816, 9725, 10367, 214~
## $ Injury_or_Fatal_Collisions        <dbl> 3249, 522, 556, 1077, 895, 199, 391~
## $ MotoristsInjured                  <dbl> 1522, 155, 283, 479, 471, 134, 2453~
## $ MotoristsKilled                   <dbl> 8, 1, 3, 1, 3, 0, 6, 0, 3, 1, 2, 0,~
## $ PassengInjured                    <dbl> 1677, 174, 331, 586, 485, 101, 1525~
## $ PassengKilled                     <dbl> 4, 0, 3, 0, 1, 0, 2, 0, 1, 0, 1, 0,~
```

```
## $ CyclistsInjured                    <dbl> 483, 119, 68, 182, 103, 11, 452, 11~
## $ CyclistsKilled                     <dbl> 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0,~
## $ PedestrInjured                     <dbl> 751, 174, 117, 263, 171, 26, 778, 1~
## $ PedestrKilled                      <dbl> 13, 5, 3, 2, 3, 0, 8, 0, 1, 4, 3, 0~
## $ Bicycle                            <dbl> 645, 194, 76, 241, 121, 13, 644, 19~
```

```
dim(df1)
```

```
## [1] 36 15
```

```
str(df1)
```

```
## spc_tbl_ [36 x 15] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Year                              : num [1:36] 2014 2014 2014 2014 2014 ...
##  $ GeoCode                           : chr [1:36] "C" "M" "B" "K" ...
##  $ GeoCodeLabel                      : chr [1:36] "CITYWIDE" "MANHATTAN" "BRONX" "BROOKLYN" ...
##  $ Number_of_Motor_Vehicle_Collisions: num [1:36] 17720 4026 2455 4960 5195 ...
##  $ Vehicles_or_Motorists_Involved    : num [1:36] 34721 7672 4816 9725 10367 ...
##  $ Injury_or_Fatal_Collisions        : num [1:36] 3249 522 556 1077 895 ...
##  $ MotoristsInjured                  : num [1:36] 1522 155 283 479 471 ...
##  $ MotoristsKilled                   : num [1:36] 8 1 3 1 3 0 6 0 3 1 ...
##  $ PassengInjured                    : num [1:36] 1677 174 331 586 485 ...
##  $ PassengKilled                     : num [1:36] 4 0 3 0 1 0 2 0 1 0 ...
##  $ CyclistsInjured                   : num [1:36] 483 119 68 182 103 11 452 117 48 199 ...
##  $ CyclistsKilled                    : num [1:36] 1 0 0 1 0 0 1 0 0 0 ...
##  $ PedestrInjured                    : num [1:36] 751 174 117 263 171 26 778 156 146 255 ...
##  $ PedestrKilled                     : num [1:36] 13 5 3 2 3 0 8 0 1 4 ...
##  $ Bicycle                           : num [1:36] 645 194 76 241 121 13 644 199 71 259 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Year = col_double(),
##   ..   GeoCode = col_character(),
##   ..   GeoCodeLabel = col_character(),
##   ..   Number_of_Motor_Vehicle_Collisions = col_double(),
##   ..   Vehicles_or_Motorists_Involved = col_double(),
##   ..   Injury_or_Fatal_Collisions = col_double(),
##   ..   MotoristsInjured = col_double(),
##   ..   MotoristsKilled = col_double(),
##   ..   PassengInjured = col_double(),
##   ..   PassengKilled = col_double(),
##   ..   CyclistsInjured = col_double(),
##   ..   CyclistsKilled = col_double(),
##   ..   PedestrInjured = col_double(),
##   ..   PedestrKilled = col_double(),
##   ..   Bicycle = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Let's take a quick peek at df2

```
glimpse(df2)
```

```
## Rows: 120
## Columns: 6
## $ Year                          <dbl> 2023, 2023, 2023, 2023, 2023, 2023, 2023~
## $ GeoCode                       <chr> "C", "C", "C", "C", "C", "C", "C", "C", ~
## $ GeoCodeLabel                  <chr> "CITYWIDE", "CITYWIDE", "CITYWIDE", "CIT~
```

```
## $ ContributingFactorCode        <chr> "28", "02", "03", "22", "04", "05", "06"~
## $ ContributingFactorDescription <chr> "AGGRESSIVE DRIVING/ROAD RAGE", "ALCOHOL~
## $ Number_of_Vehicles            <dbl> 89, 161, 226, 4, 2410, 204, 8, 113, 13, ~
```

```
dim(df2)
```

```
## [1] 120   6
```

```
str(df2)
```

```
## spc_tbl_ [120 x 6] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Year                          : num [1:120] 2023 2023 2023 2023 2023 ...
##  $ GeoCode                       : chr [1:120] "C" "C" "C" "C" ...
##  $ GeoCodeLabel                  : chr [1:120] "CITYWIDE" "CITYWIDE" "CITYWIDE" "CITYWIDE" ...
##  $ ContributingFactorCode        : chr [1:120] "28" "02" "03" "22" ...
##  $ ContributingFactorDescription: chr [1:120] "AGGRESSIVE DRIVING/ROAD RAGE" "ALCOHOL INVOLVEMENT" "
##  $ Number_of_Vehicles            : num [1:120] 89 161 226 4 2410 204 8 113 13 612 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Year = col_double(),
##   ..   GeoCode = col_character(),
##   ..   GeoCodeLabel = col_character(),
##   ..   ContributingFactorCode = col_character(),
##   ..   ContributingFactorDescription = col_character(),
##   ..   Number_of_Vehicles = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Data Tidying and Data Transformation

```
# Merging df1 and df2 into one single data frame
data <- df1 %>% inner_join(df2, by = "Year")
```

```
## Warning in inner_join(., df2, by = "Year"): Detected an unexpected many-to-many relationship between
## i Row 1 of `x` matches multiple rows in `y`.
## i Row 90 of `y` matches multiple rows in `x`.
## i If a many-to-many relationship is expected, set `relationship =
##   "many-to-many"` to silence this warning.
```

```
data
```

```
## # A tibble: 720 x 20
##     Year GeoCode.x GeoCodeLabel.x Number_of_Motor_Vehic~1 Vehicles_or_Motorist~2
##    <dbl> <chr>     <chr>                            <dbl>                  <dbl>
##  1  2014 C         CITYWIDE                         17720                  34721
##  2  2014 C         CITYWIDE                         17720                  34721
##  3  2014 C         CITYWIDE                         17720                  34721
##  4  2014 C         CITYWIDE                         17720                  34721
##  5  2014 C         CITYWIDE                         17720                  34721
##  6  2014 C         CITYWIDE                         17720                  34721
##  7  2014 C         CITYWIDE                         17720                  34721
##  8  2014 C         CITYWIDE                         17720                  34721
##  9  2014 C         CITYWIDE                         17720                  34721
## 10  2014 C         CITYWIDE                         17720                  34721
## # i 710 more rows
## # i abbreviated names: 1: Number_of_Motor_Vehicle_Collisions,
## #   2: Vehicles_or_Motorists_Involved
```

```
## # i 15 more variables: Injury_or_Fatal_Collisions <dbl>,
## #   MotoristsInjured <dbl>, MotoristsKilled <dbl>, PassengInjured <dbl>,
## #   PassengKilled <dbl>, CyclistsInjured <dbl>, CyclistsKilled <dbl>,
## #   PedestrInjured <dbl>, PedestrKilled <dbl>, Bicycle <dbl>, ...
```

Let's take a peek at the new data frame

```
glimpse(data)
```

```
## Rows: 720
## Columns: 20
## $ Year                                <dbl> 2014, 2014, 2014, 2014, 2014, 2014,~
## $ GeoCode.x                           <chr> "C", "C", "C", "C", "C", "C", "C", ~
## $ GeoCodeLabel.x                      <chr> "CITYWIDE", "CITYWIDE", "CITYWIDE",~
## $ Number_of_Motor_Vehicle_Collisions <dbl> 17720, 17720, 17720, 17720, 17720, ~
## $ Vehicles_or_Motorists_Involved      <dbl> 34721, 34721, 34721, 34721, 34721, ~
## $ Injury_or_Fatal_Collisions          <dbl> 3249, 3249, 3249, 3249, 3249, 3249,~
## $ MotoristsInjured                    <dbl> 1522, 1522, 1522, 1522, 1522, 1522,~
## $ MotoristsKilled                     <dbl> 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8,~
## $ PassengInjured                      <dbl> 1677, 1677, 1677, 1677, 1677, 1677,~
## $ PassengKilled                       <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,~
## $ CyclistsInjured                     <dbl> 483, 483, 483, 483, 483, 483, 483, ~
## $ CyclistsKilled                      <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ PedestrInjured                      <dbl> 751, 751, 751, 751, 751, 751, 751, ~
## $ PedestrKilled                       <dbl> 13, 13, 13, 13, 13, 13, 13, 13, 13,~
## $ Bicycle                             <dbl> 645, 645, 645, 645, 645, 645, 645, ~
## $ GeoCode.y                           <chr> "C", "C", "C", "C", "C", "C", "C", ~
## $ GeoCodeLabel.y                      <chr> "CITYWIDE", "CITYWIDE", "CITYWIDE",~
## $ ContributingFactorCode              <chr> "28", "02", "03", "22", "23", "04",~
## $ ContributingFactorDescription       <chr> "AGGRESSIVE DRIVING/ROAD RAGE", "AL~
## $ Number_of_Vehicles                  <dbl> 92, 233, 731, 9, 2, 3269, 322, 21, ~
```

```
names(data)
```

```
##  [1] "Year"                                "GeoCode.x"
##  [3] "GeoCodeLabel.x"                      "Number_of_Motor_Vehicle_Collisions"
##  [5] "Vehicles_or_Motorists_Involved"      "Injury_or_Fatal_Collisions"
##  [7] "MotoristsInjured"                    "MotoristsKilled"
##  [9] "PassengInjured"                      "PassengKilled"
## [11] "CyclistsInjured"                     "CyclistsKilled"
## [13] "PedestrInjured"                      "PedestrKilled"
## [15] "Bicycle"                             "GeoCode.y"
## [17] "GeoCodeLabel.y"                      "ContributingFactorCode"
## [19] "ContributingFactorDescription"       "Number_of_Vehicles"
```

Let's add and mutate some of the data frame variables for analysis convenience

```
# Adding and renaming a few new variables and changing some
data_new <-data %>%
  mutate(Contributing_Factor = ContributingFactorDescription, GeoCodeLabel = GeoCodeLabel.x, non_motoris
  rename(Motorists_Involved =Vehicles_or_Motorists_Involved)
data_new
```

```
## # A tibble: 720 x 24
##     Year GeoCode.x GeoCodeLabel.x Number_of_Motor_Vehicle_C~1 Motorists_Involved
##    <dbl> <chr>     <chr>                                <dbl>              <dbl>
##  1  2014 C         CITYWIDE                             17720              34721
```

```
##  2  2014 C         CITYWIDE                                17720                34721
##  3  2014 C         CITYWIDE                                17720                34721
##  4  2014 C         CITYWIDE                                17720                34721
##  5  2014 C         CITYWIDE                                17720                34721
##  6  2014 C         CITYWIDE                                17720                34721
##  7  2014 C         CITYWIDE                                17720                34721
##  8  2014 C         CITYWIDE                                17720                34721
##  9  2014 C         CITYWIDE                                17720                34721
## 10  2014 C         CITYWIDE                                17720                34721
## # i 710 more rows
## # i abbreviated name: 1: Number_of_Motor_Vehicle_Collisions
## # i 19 more variables: Injury_or_Fatal_Collisions <dbl>,
## #   MotoristsInjured <dbl>, MotoristsKilled <dbl>, PassengInjured <dbl>,
## #   PassengKilled <dbl>, CyclistsInjured <dbl>, CyclistsKilled <dbl>,
## #   PedestrInjured <dbl>, PedestrKilled <dbl>, Bicycle <dbl>, GeoCode.y <chr>,
## #   GeoCodeLabel.y <chr>, ContributingFactorCode <chr>, ...
```

Data Analysis:

Descriptive statistics

```
# Summary statistics
summary(data_new)
```

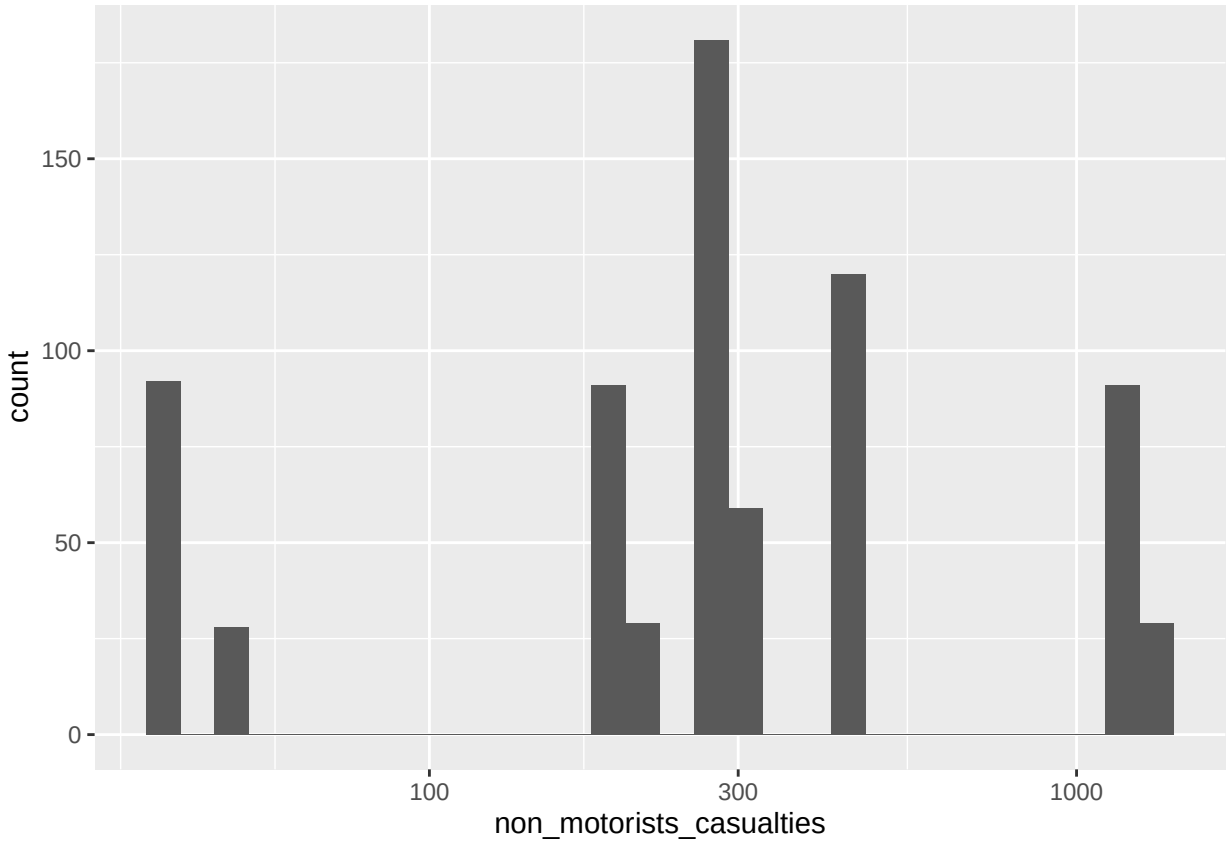
```
##       Year       GeoCode.x         GeoCodeLabel.x
##  Min.   :2014   Length:720         Length:720
##  1st Qu.:2014   Class :character   Class :character
##  Median :2019   Mode  :character   Mode  :character
##  Mean   :2019
##  3rd Qu.:2020
##  Max.   :2023
##  Number_of_Motor_Vehicle_Collisions Motorists_Involved
##  Min.   :  440                      Min.   :  881
##  1st Qu.: 1395                      1st Qu.: 2485
##  Median : 2886                      Median : 5737
##  Mean   : 4463                      Mean   : 8741
##  3rd Qu.: 5195                      3rd Qu.:10367
##  Max.   :17720                      Max.   :34721
##  Injury_or_Fatal_Collisions MotoristsInjured MotoristsKilled PassengInjured
##  Min.   : 139               Min.   : 103     Min.   : 0.00   Min.   :  65
##  1st Qu.: 548               1st Qu.: 228     1st Qu.: 0.75   1st Qu.: 179
##  Median : 788               Median : 457     Median : 2.00   Median : 359
##  Mean   :1175               Mean   : 671     Mean   : 2.82   Mean   : 493
##  3rd Qu.:1172               3rd Qu.: 726     3rd Qu.: 3.00   3rd Qu.: 487
##  Max.   :3919               Max.   :2453     Max.   :14.00   Max.   :1677
##  PassengKilled   CyclistsInjured CyclistsKilled  PedestrInjured PedestrKilled
##  Min.   :0.000   Min.   :  8     Min.   :0.000   Min.   : 20    Min.   : 0.00
##  1st Qu.:0.000   1st Qu.: 66     1st Qu.:0.000   1st Qu.:117    1st Qu.: 1.00
##  Median :1.000   Median :117     Median :0.000   Median :171    Median : 3.00
##  Mean   :0.994   Mean   :178     Mean   :0.892   Mean   :235    Mean   : 3.34
##  3rd Qu.:1.000   3rd Qu.:199     3rd Qu.:1.000   3rd Qu.:255    3rd Qu.: 4.25
##  Max.   :4.000   Max.   :693     Max.   :6.000   Max.   :778    Max.   :13.00
##     Bicycle       GeoCode.y         GeoCodeLabel.y     ContributingFactorCode
##  Min.   : 10.0   Length:720         Length:720         Length:720
```

```
##  1st Qu.: 74.8   Class :character   Class :character   Class :character
##  Median :168.5   Mode  :character   Mode  :character   Mode  :character
##  Mean   :219.6
##  3rd Qu.:259.0
##  Max.   :719.0
##  ContributingFactorDescription Number_of_Vehicles Contributing_Factor
##  Length:720                    Min.   :   1       Length:720
##  Class :character              1st Qu.:  11       Class :character
##  Mode  :character              Median :  84       Mode  :character
##                                Mean   : 326
##                                3rd Qu.: 316
##                                Max.   :5721
##  GeoCodeLabel       non_motorists_casualties motorists_casualties
##  Length:720         Min.   :  37             Min.   : 174
##  Class :character   1st Qu.: 195             1st Qu.: 398
##  Mode  :character   Median : 277             Median : 864
##                     Mean   : 417             Mean   :1168
##                     3rd Qu.: 456             3rd Qu.:1166
##                     Max.   :1270             Max.   :3986
```

```
# Histogram of all non motorists killed or injured
ggplot(data_new, aes(x = non_motorists_casualties)) +
  geom_histogram(bindwidth = 0.3) + scale_x_log10()
```

```
## Warning in geom_histogram(bindwidth = 0.3): Ignoring unknown parameters:
## `bindwidth`
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



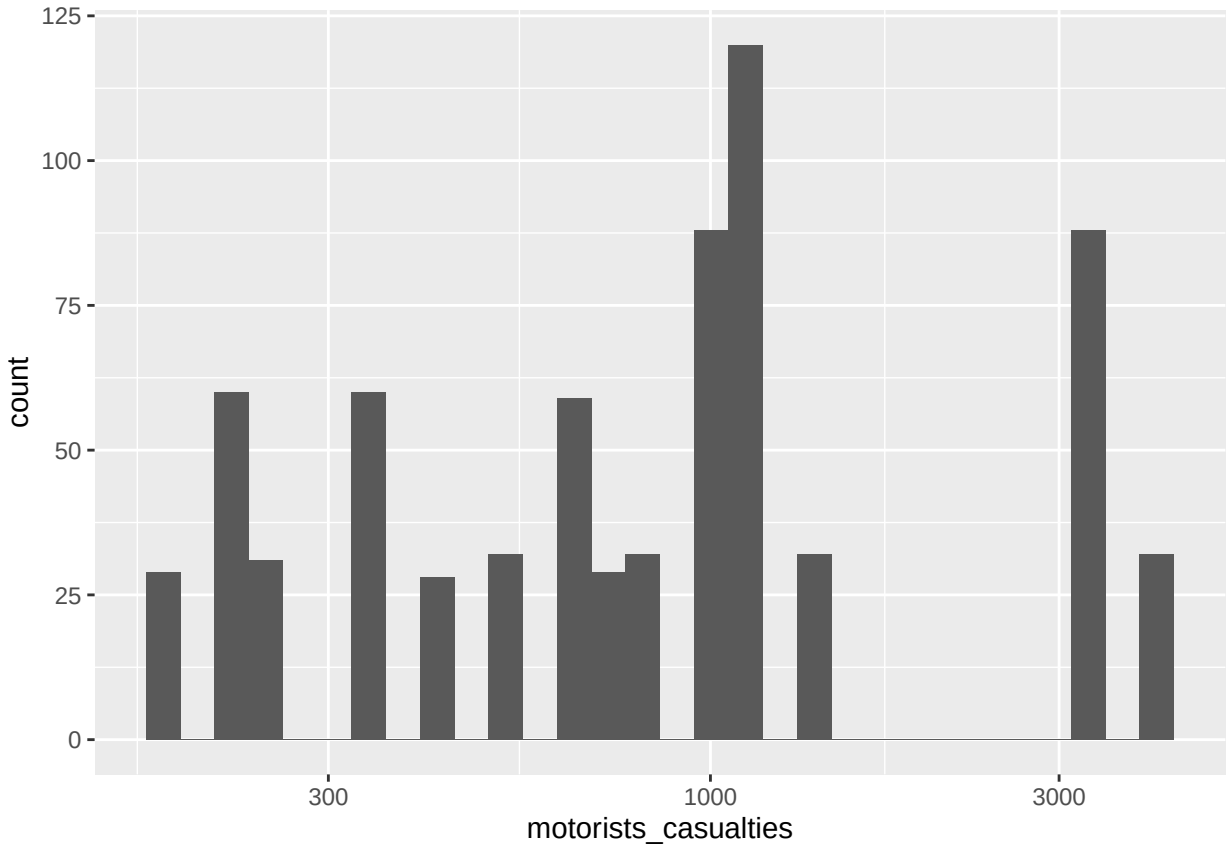
```
xlab("Non_motorists_casualties")
```

```
## $x
## [1] "Non_motorists_casualties"
##
## attr(,"class")
## [1] "labels"
```

```
# Histogram of all motorists killed or injured
ggplot(data_new, aes(x = motorists_casualties)) +
  geom_histogram(bindwidth = 0.3) + scale_x_log10()
```

```
## Warning in geom_histogram(bindwidth = 0.3): Ignoring unknown parameters:
## `bindwidth`
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

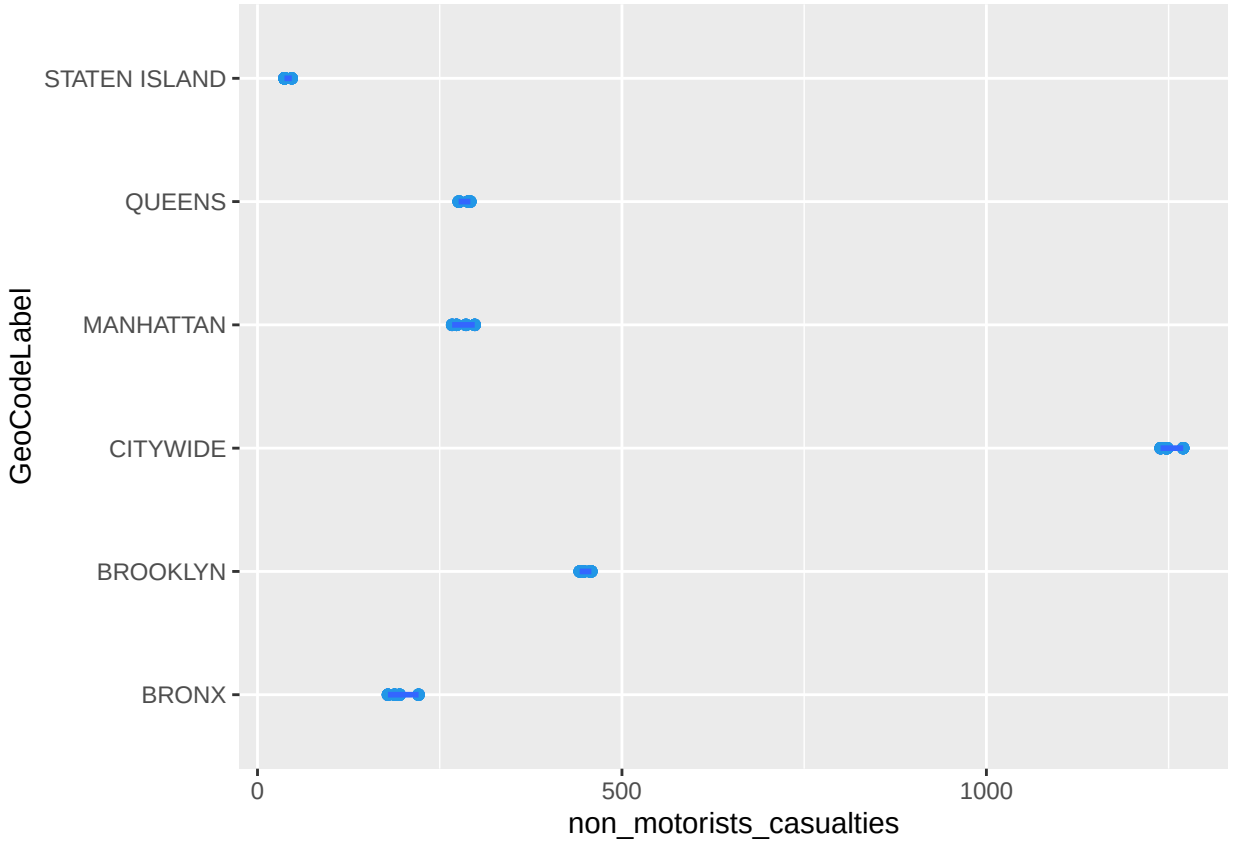



```
xlab("Motorists_casualties")
```

```
## $x
## [1] "Motorists_casualties"
##
## attr(,"class")
## [1] "labels"
```

```
# scatter plot of non - motorist casualties
ggplot(data = data_new, aes( x = non_motorists_casualties, y = GeoCodeLabel)) +
  geom_point(color = 4, alpha = 0.3) +
  stat_smooth(method = "lm", se = FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

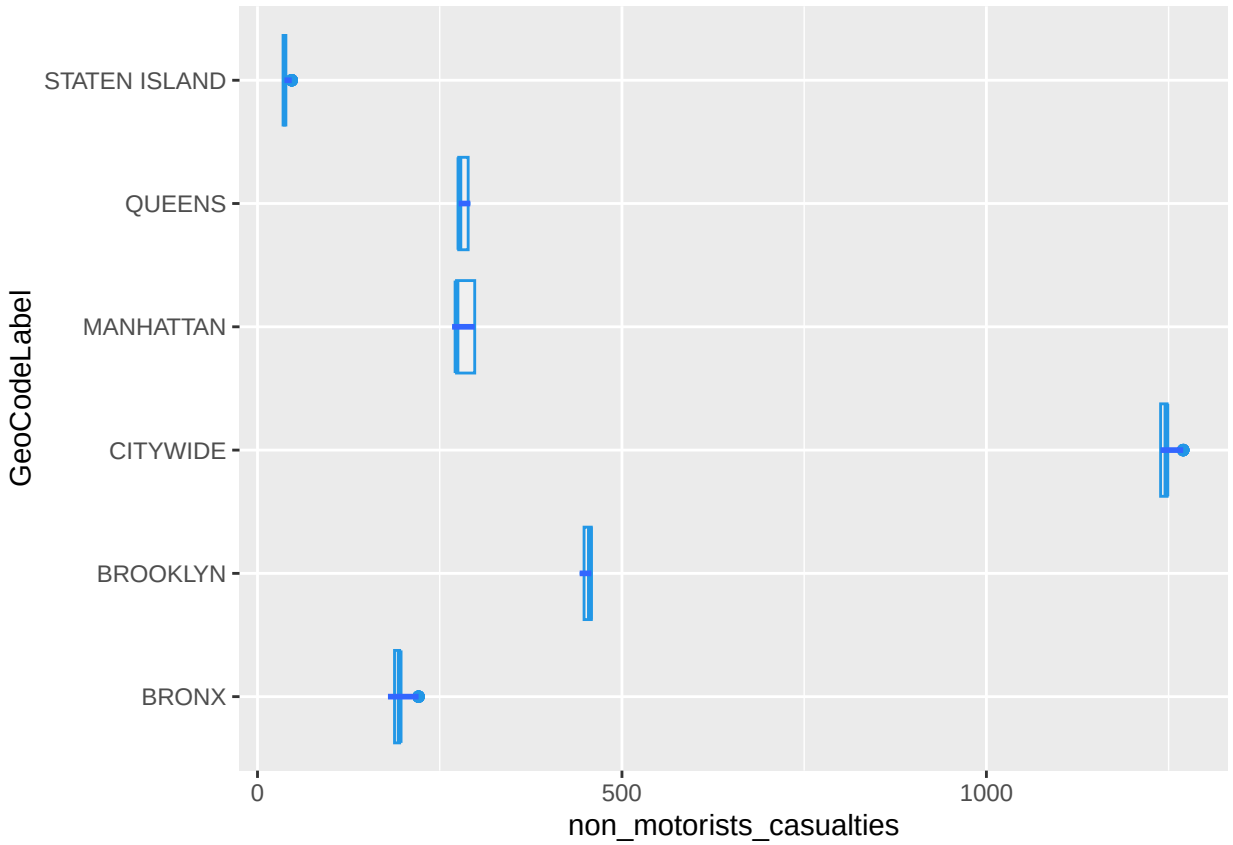



```
  labs(
    title = ("Scatter Plot of Non - Motorists Casuaties per Borough"))
```

```
## $title
## [1] "Scatter Plot of Non - Motorists Casuaties per Borough"
##
## attr(,"class")
## [1] "labels"
```

```
# scatter plot of non - motorist casualties
ggplot(data = data_new, aes( x = non_motorists_casualties, y = GeoCodeLabel)) +
  geom_boxplot(color = 4, alpha = 0.3) +
  stat_smooth(method = "lm", se = FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

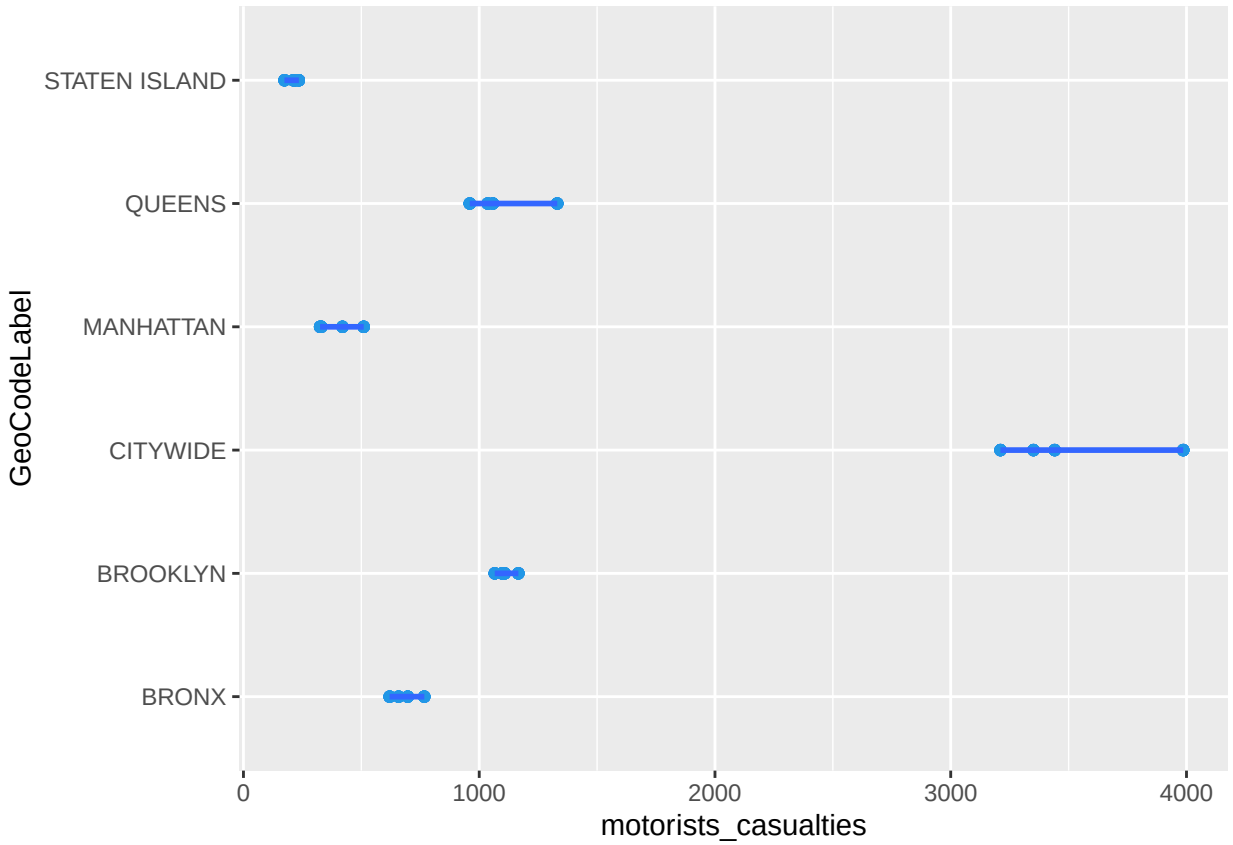



```
  labs(
    title = ("BoxPlot of Non - Motorists Casuaties per Borough"))
```

```
## $title
## [1] "BoxPlot of Non - Motorists Casuaties per Borough"
##
## attr(,"class")
## [1] "labels"
```

```
# scatter plot of motorist casualties
ggplot(data = data_new, aes( x = motorists_casualties, y = GeoCodeLabel)) +
  geom_point(color = 4, alpha = 0.3) +
  stat_smooth(method = "lm", se = FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

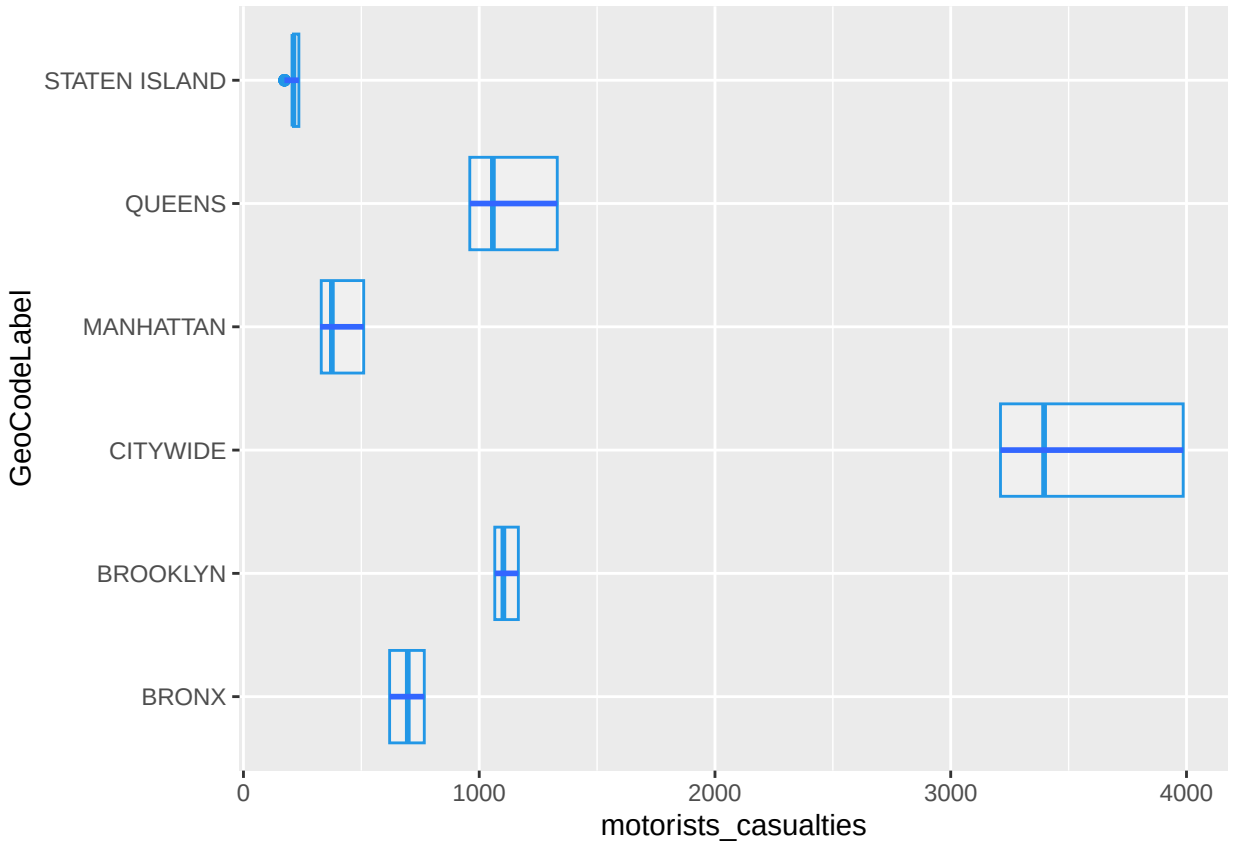



```
  labs(
    title = ("Scatter Plot of Motorists Casuaties per Borough "))
```

```
## $title
## [1] "Scatter Plot of Motorists Casuaties per Borough "
##
## attr(,"class")
## [1] "labels"
```

```
# Box plot of motorists casualties
ggplot(data = data_new, aes( x = motorists_casualties, y = GeoCodeLabel)) +
  geom_boxplot(color = 4, alpha = 0.3) +
  stat_smooth(method = "lm", se = FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

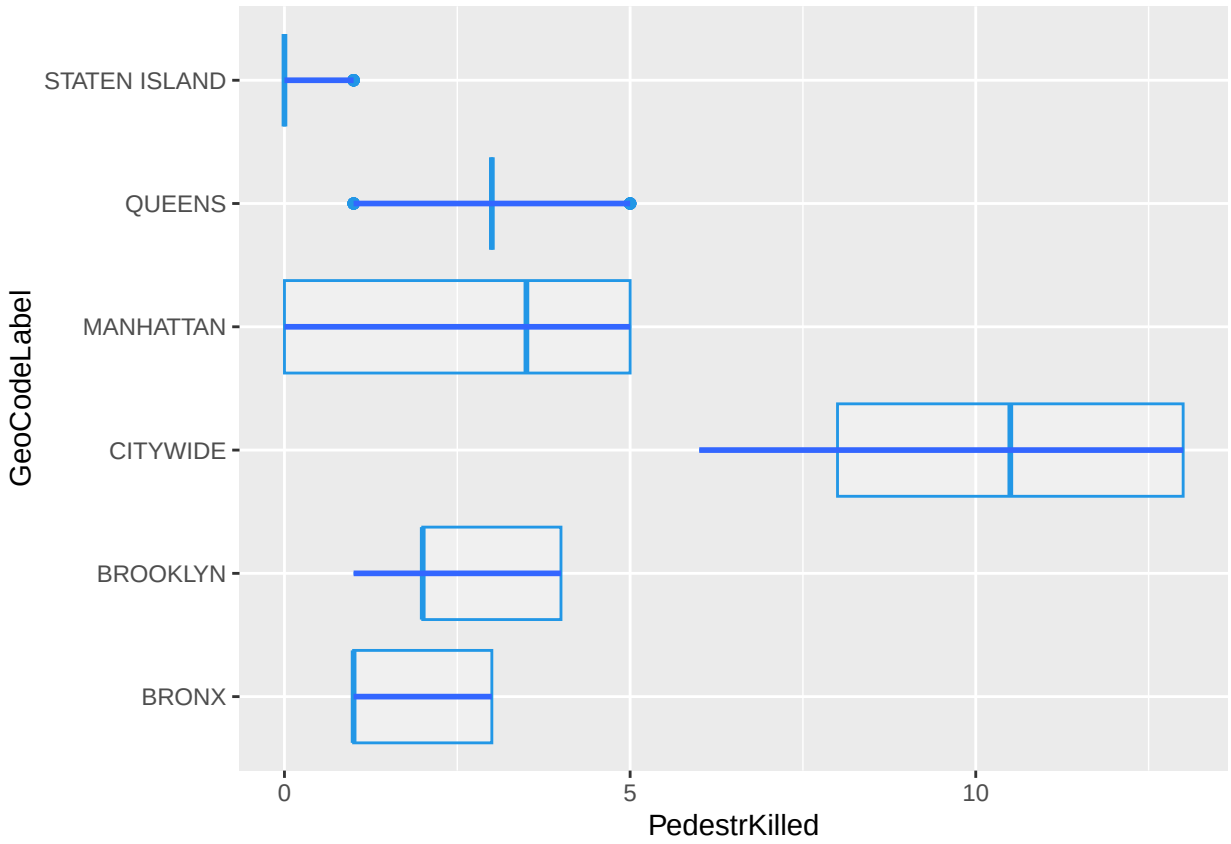



```
  labs(
    title = ("Box Plot of Motorists Casuaties per Borough "))
```

```
## $title
## [1] "Box Plot of Motorists Casuaties per Borough "
##
## attr(,"class")
## [1] "labels"
```

```
ggplot(data = data_new, aes( x = PedestrKilled, y = GeoCodeLabel)) +
  geom_boxplot(color = 4, alpha = 0.3) +
  stat_smooth(method = "lm", se = FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

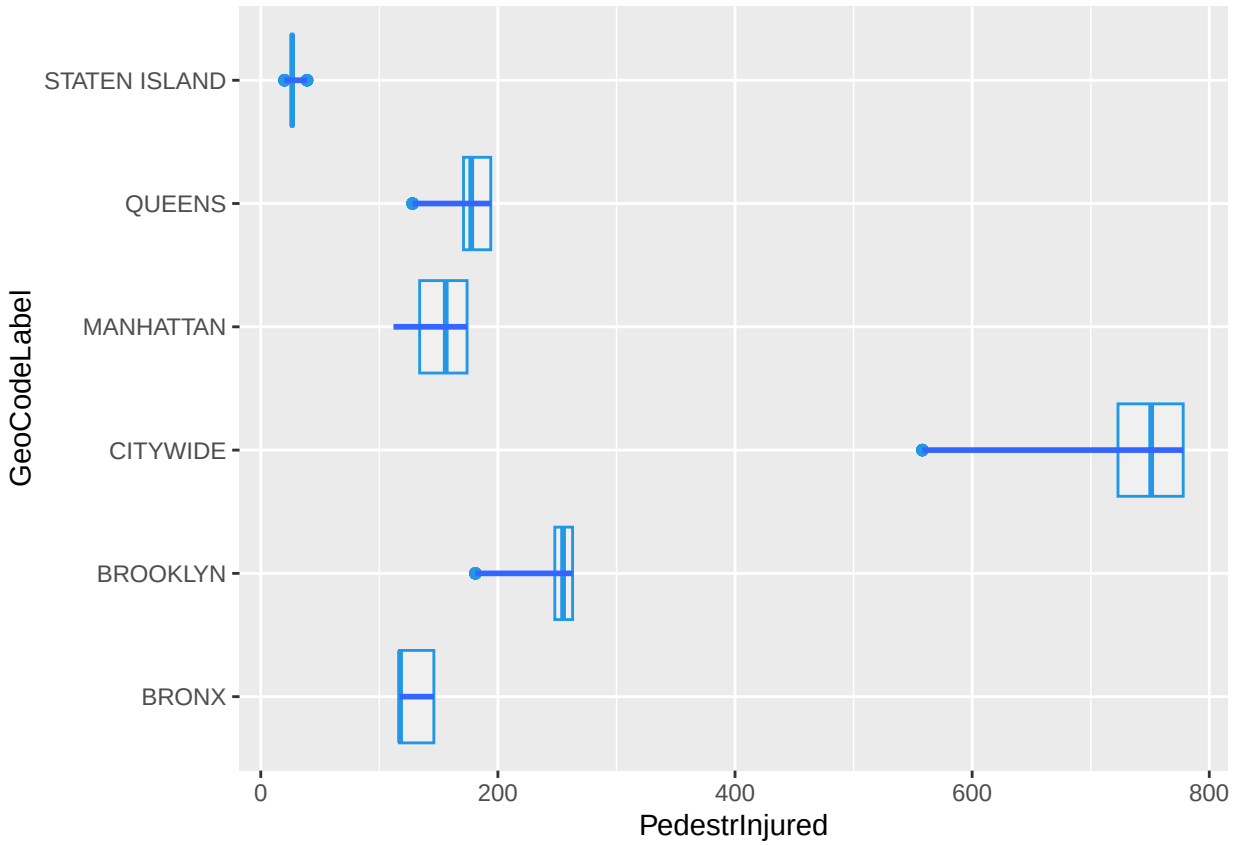



```
  labs(
    title = ("Box Plot of Pedestrian Kiled per Borough "))
```

```
## $title
## [1] "Box Plot of Pedestrian Kiled per Borough "
##
## attr(,"class")
## [1] "labels"
```

```
ggplot(data = data_new, aes( x = PedestrInjured, y = GeoCodeLabel)) +
  geom_boxplot(color = 4, alpha = 0.3) +
  stat_smooth(method = "lm", se = FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

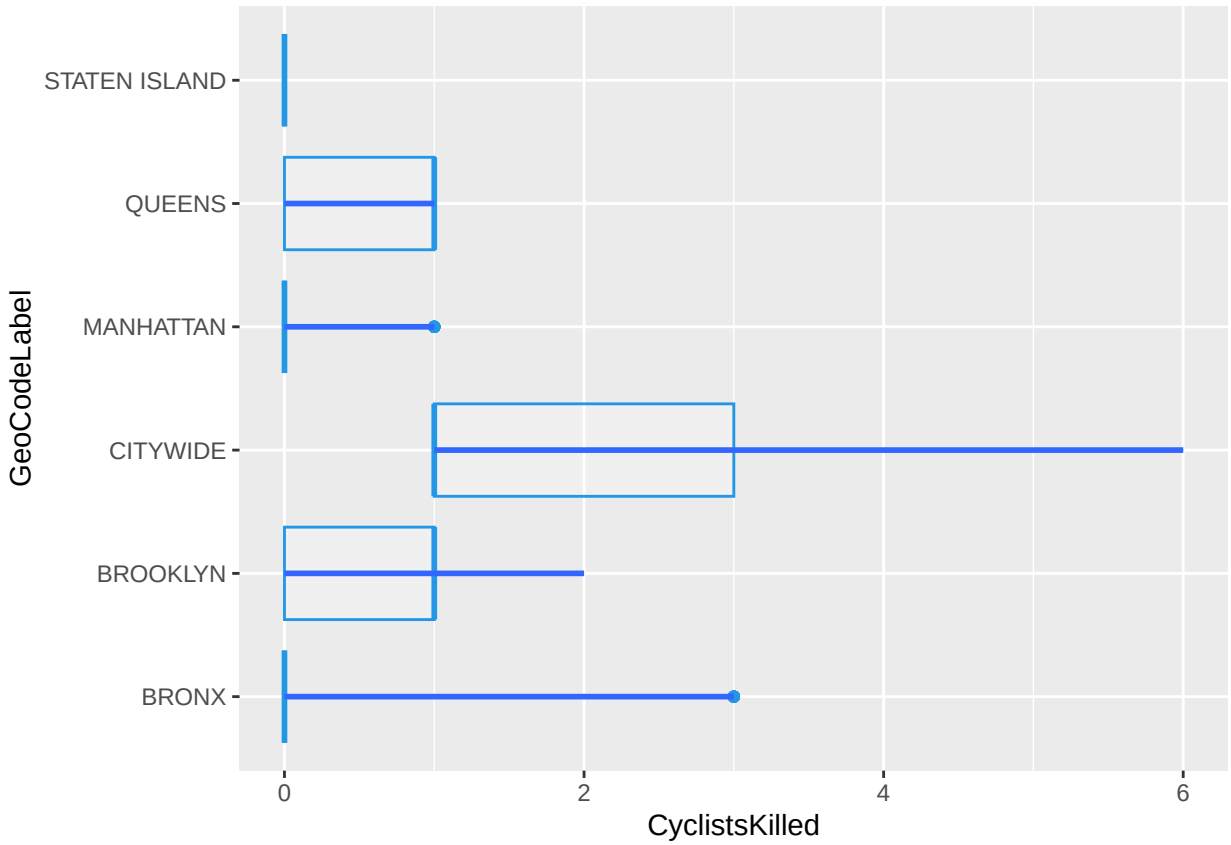



```
  labs(
    title = ("Box Plot of Pedestrian Injured per Borough "))
```

```
## $title
## [1] "Box Plot of Pedestrian Injured per Borough "
##
## attr(,"class")
## [1] "labels"
```

```
ggplot(data = data_new, aes( x = CyclistsKilled, y = GeoCodeLabel)) +
  geom_boxplot(color = 4, alpha = 0.3) +
  stat_smooth(method = "lm", se = FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

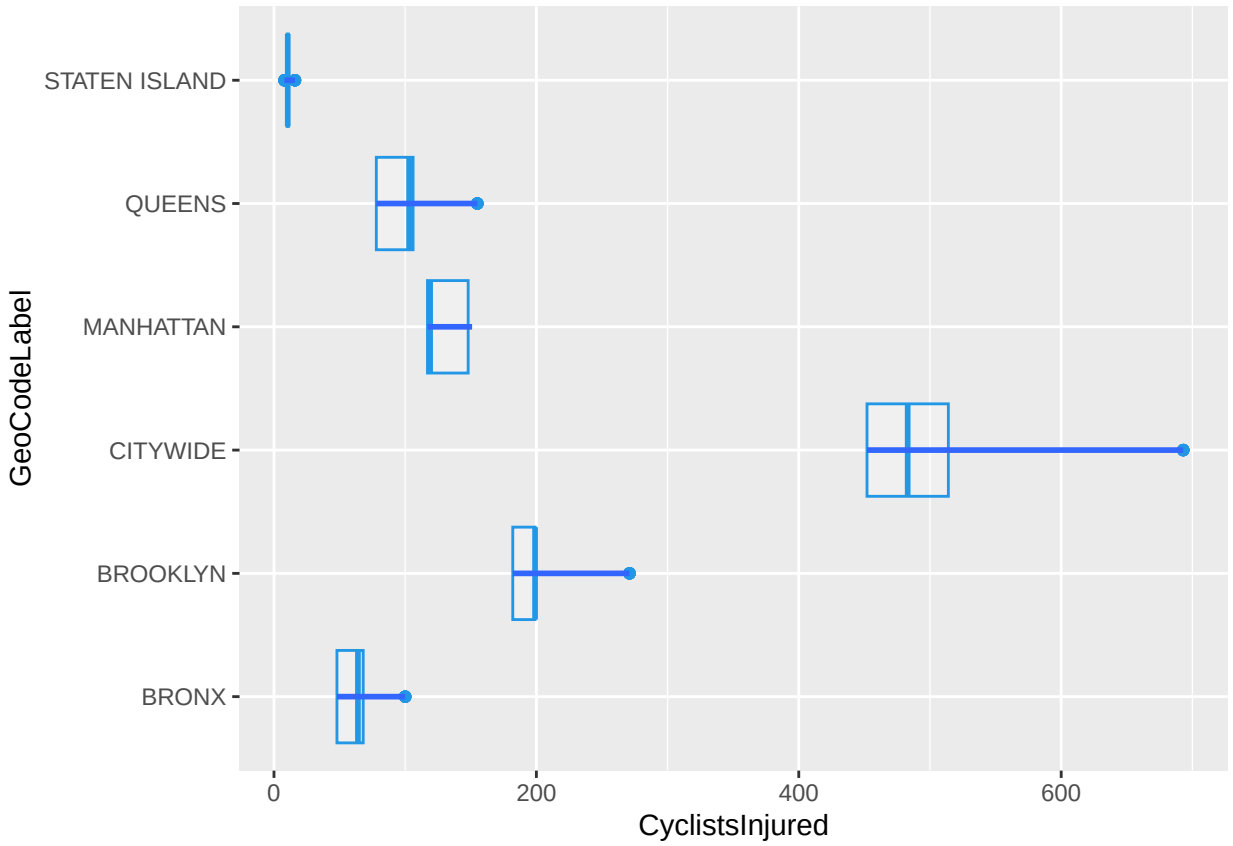



```
  labs(
    title = ("Box Plot of Cyclists Killed per Borough "))
```

```
## $title
## [1] "Box Plot of Cyclists Killed per Borough "
##
## attr(,"class")
## [1] "labels"
```

```
ggplot(data = data_new, aes( x = CyclistsInjured, y = GeoCodeLabel)) +
  geom_boxplot(color = 4, alpha = 0.3) +
  stat_smooth(method = "lm", se = FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



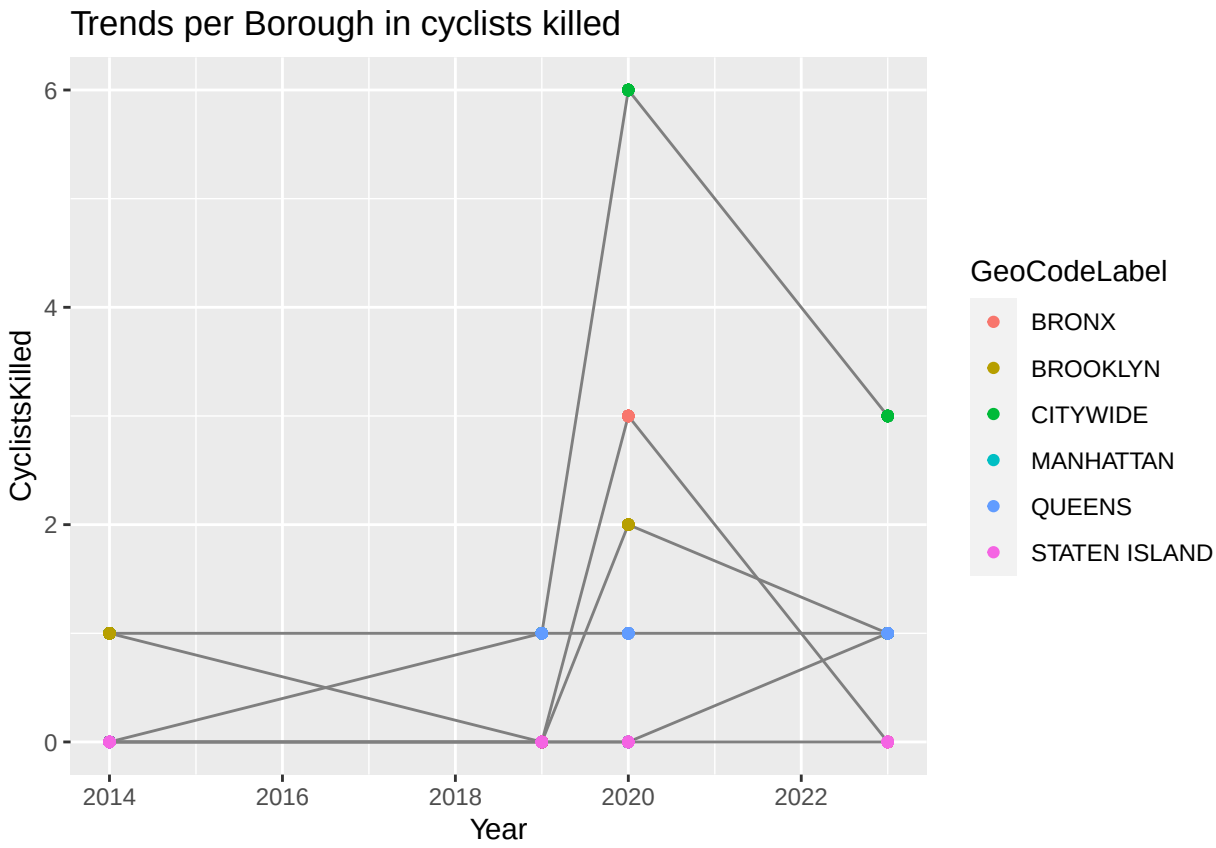

```
  labs(
    title = ("Box Plot of Cyclists Injured per Borough "))
```

```
## $title
## [1] "Box Plot of Cyclists Injured per Borough "
##
## attr(,"class")
## [1] "labels"
```

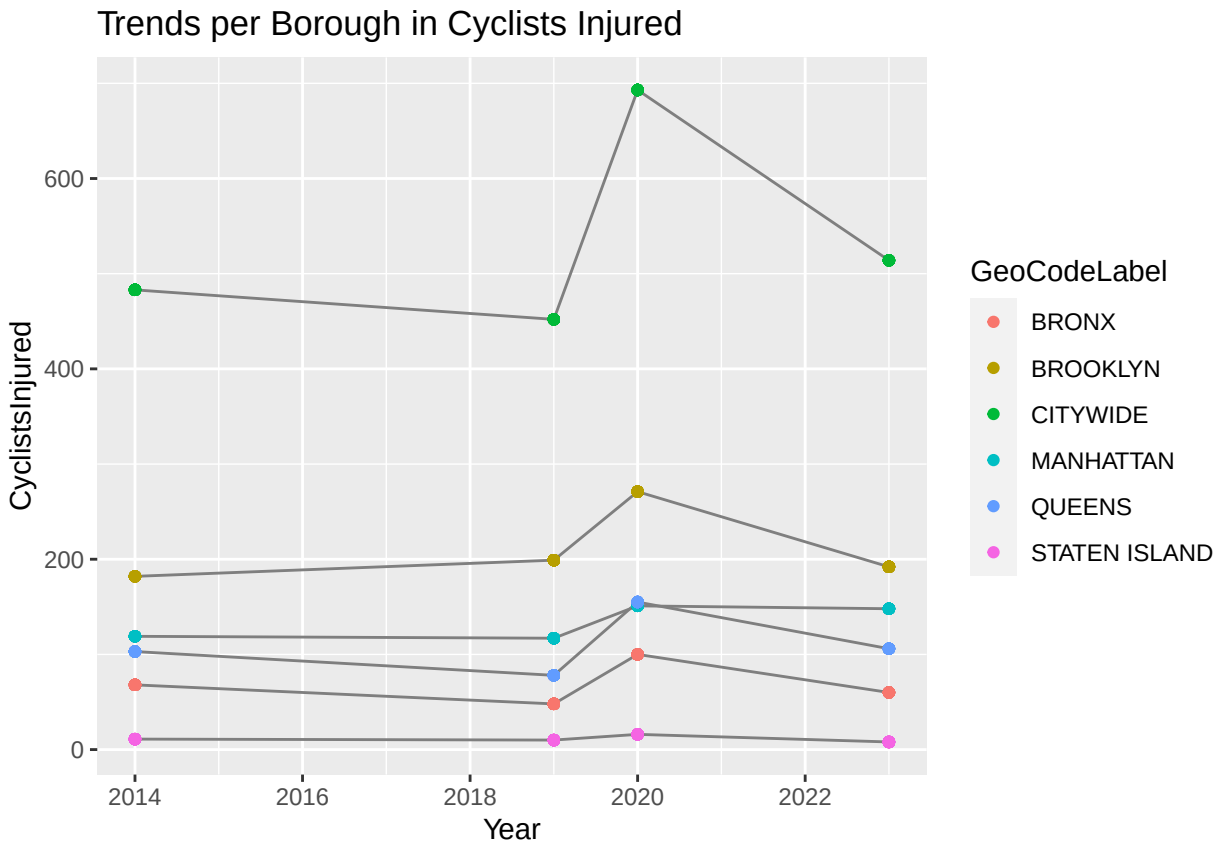
Data Visualization:

Visualizing Changes in casualties Over Time since 2014

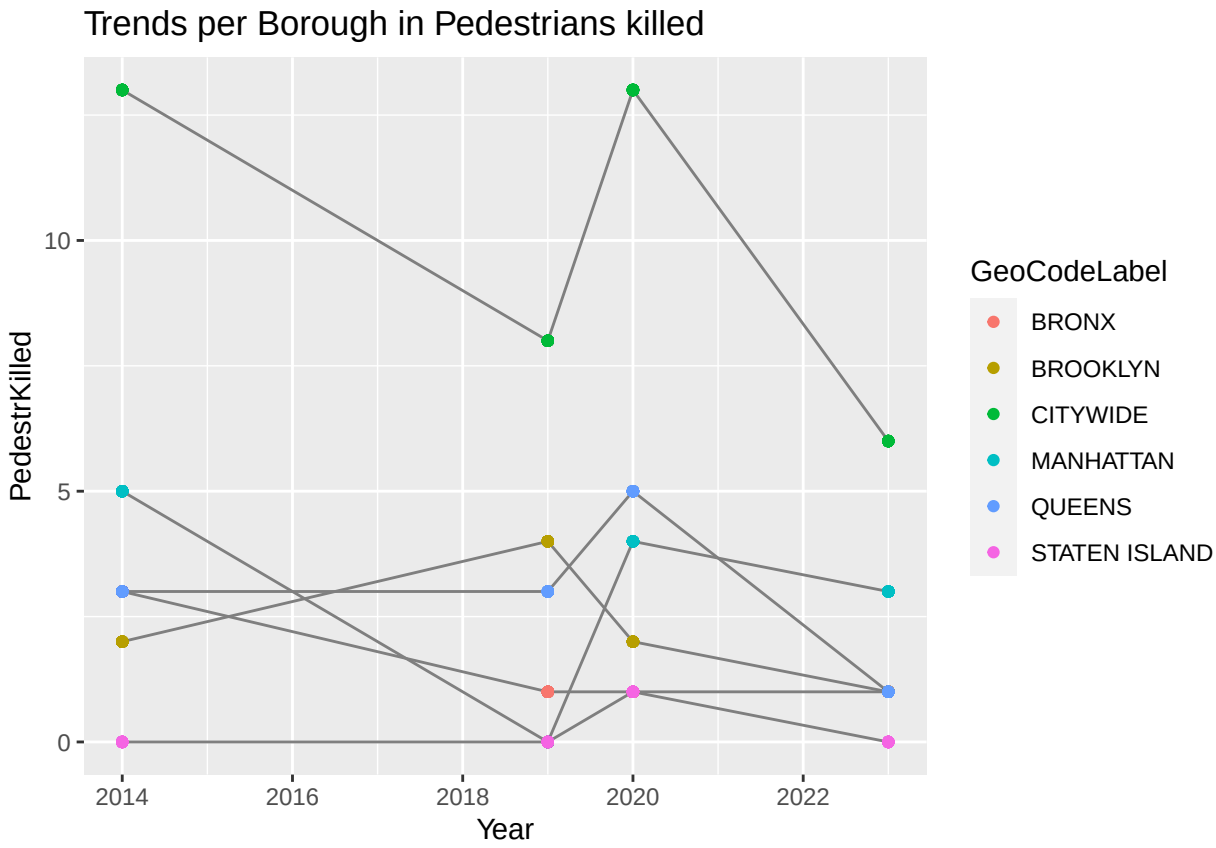
```
 ## Trends per boroughs in cyclists killed in NYC
ggplot(data_new, aes(x = Year, y = CyclistsKilled)) +
  geom_line(aes(group = GeoCodeLabel), colour = "grey50") +
  geom_point(aes(colour = GeoCodeLabel )) +
  labs(
    title = ("Trends per Borough in cyclists killed"))
```



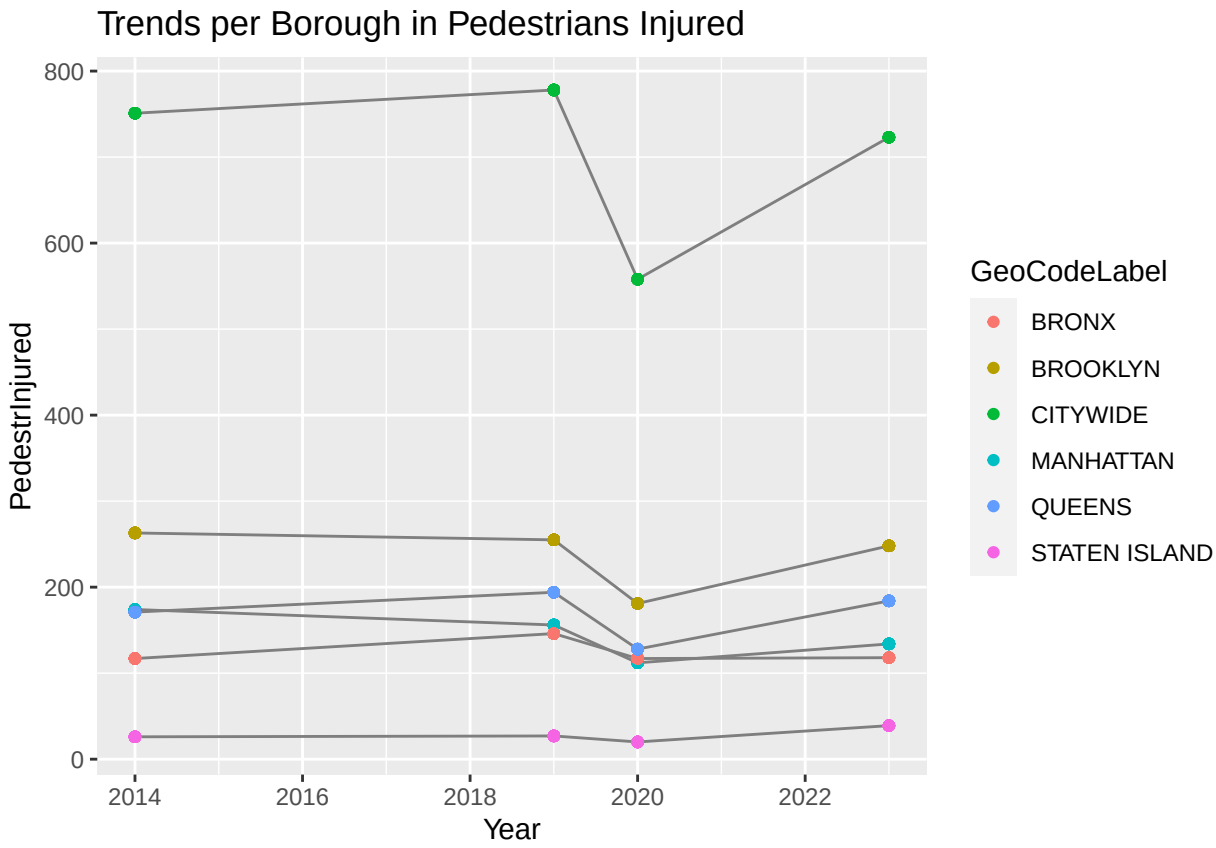

```
 ## Trends per boroughs in Cyclists Injured
ggplot(data_new, aes(x = Year, y = CyclistsInjured)) +
  geom_line(aes(group = GeoCodeLabel), colour = "grey50") +
  geom_point(aes(colour = GeoCodeLabel )) +
 labs(
    title = ("Trends per Borough in Cyclists Injured"))
```




```
## Trends per borough in pedestrians killed
ggplot(data_new, aes(x = Year, y = PedestrKilled)) +
  geom_line(aes(group = GeoCodeLabel), colour = "grey50") +
  geom_point(aes(colour = GeoCodeLabel )) +
 labs(
    title = ("Trends per Borough in Pedestrians killed"))
```



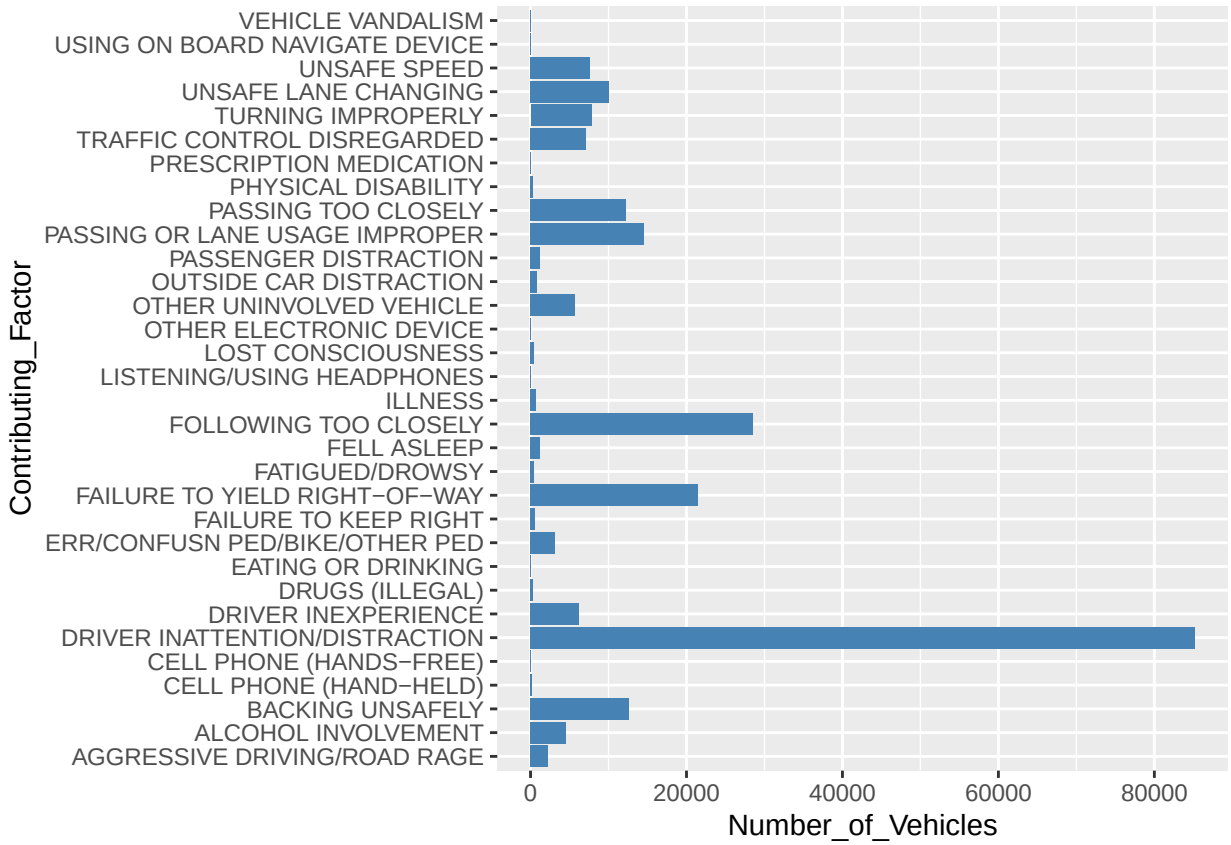

```
## Trends per borough in Pedestrians Injured
ggplot(data_new, aes(x = Year, y = PedestrInjured)) +
  geom_line(aes(group = GeoCodeLabel), colour = "grey50") +
  geom_point(aes(colour = GeoCodeLabel )) +
 labs(
    title = ("Trends per Borough in Pedestrians Injured"))
```



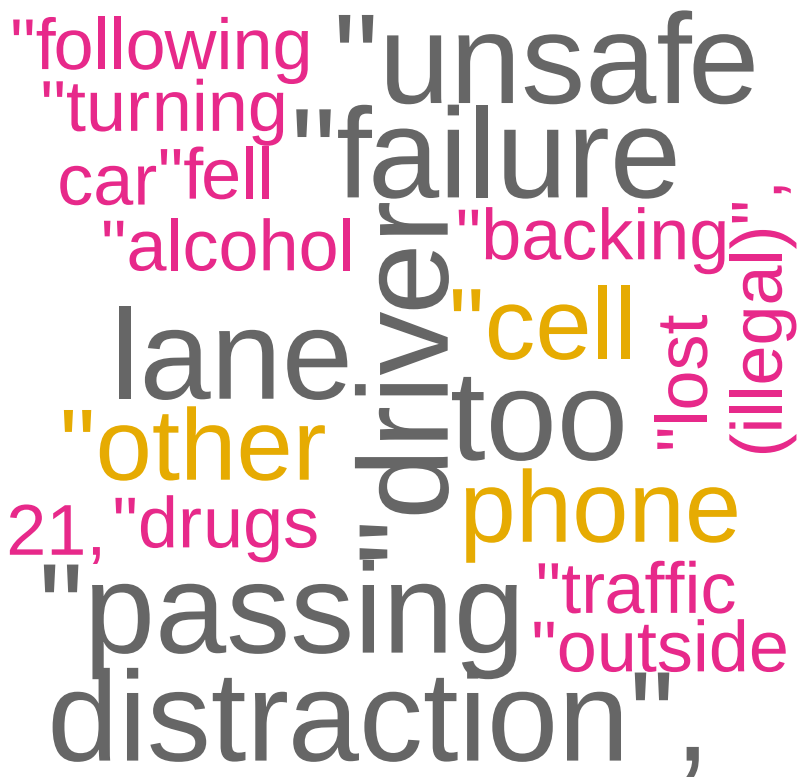
Visualizing the contributing factors to road collisions in NYC

```
# Visualizing collisions factors in a bar chart
data2 <- as.data.frame(contributing_factors)
 ggplot(data2, aes(x = Contributing_Factor, y = Number_of_Vehicles
)) +
geom_bar(stat = "identity", freq = 500, fill = "steelblue")+ coord_flip()
```

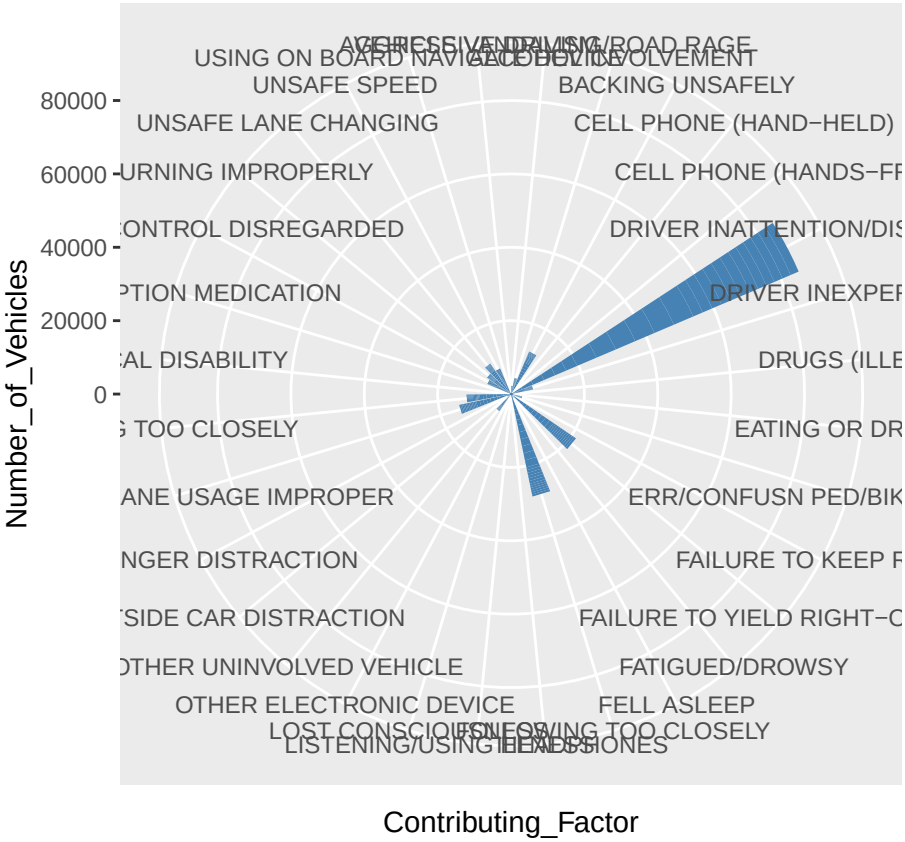
```
## Warning in geom_bar(stat = "identity", freq = 500, fill = "steelblue"):
## Ignoring unknown parameters: `freq`
```




```r
library(tm)
# Visualizing collisions factors as a wordcloud
set.seed(337)
wordcloud(data2, max.words = 2000, random.order = FALSE, min.freq = 20, colors = brewer.pal(8,"Dark2"))
```




```
data2 <- as.data.frame(contributing_factors)
 ggplot(data2, aes(x = Contributing_Factor, y = Number_of_Vehicles
)) +
geom_bar(stat = "identity", fill = "steelblue") +
coord_polar(theta = "x")
```



Narrowing things down

```
# Getting distinct values from data2
distinct_data2 <- distinct(data2,Contributing_Factor,Number_of_Vehicles)
```

```
# Identifying the main contributing factors in collisions in NYC
top_Contributing_Factor <- distinct_data2 %>%
  arrange(desc(Number_of_Vehicles)) %>%
  slice_head(n = 10)

 print(top_Contributing_Factor)
```

```
##                Contributing_Factor Number_of_Vehicles
## 1  DRIVER INATTENTION/DISTRACTION               5721
## 2  DRIVER INATTENTION/DISTRACTION               3269
## 3  DRIVER INATTENTION/DISTRACTION               2800
## 4  DRIVER INATTENTION/DISTRACTION               2410
## 5            FOLLOWING TOO CLOSELY              1959
## 6            FOLLOWING TOO CLOSELY              1447
## 7   FAILURE TO YIELD RIGHT-OF-WAY               1291
## 8   FAILURE TO YIELD RIGHT-OF-WAY               1046
## 9  PASSING OR LANE USAGE IMPROPER                977
## 10           PASSING TOO CLOSELY                 928
```

Visualizing the top contributing factors to collisions in NYC

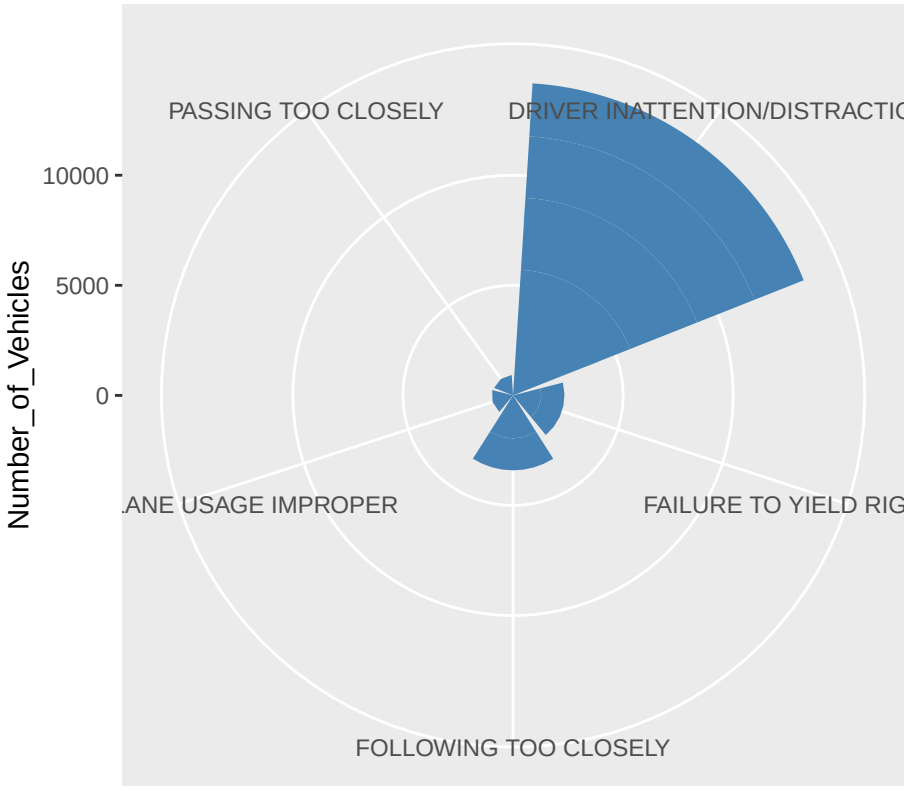
```
# Visualizing collisions main contributing factors in a bar chart
 ggplot(top_Contributing_Factor, aes(x = Contributing_Factor, y = Number_of_Vehicles
)) +
geom_bar(stat = "identity", freq = 500, fill = "steelblue")+ coord_flip()
```

```
## Warning in geom_bar(stat = "identity", freq = 500, fill = "steelblue"):
## Ignoring unknown parameters: `freq`
```


```
library(tm)
# Visualizing collisions main contributing factors as a wordcloud
set.seed(337)
wordcloud(top_Contributing_Factor, max.words = 1000, random.order = FALSE, min.freq = 20, colors = brewe
```

c(5721,
right−of−way",
1959, "passing 977,
yield too 1046,
closely")
improper", "driver 2410,
1291,
usage "failure c("driver
"following lane
928)
closely", 1447,
2800, 3269,

```
# Main contributing factors in road collisions in NYC
 ggplot(top_Contributing_Factor, aes(x = Contributing_Factor,y = Number_of_Vehicles
)) +
geom_bar(stat = "identity", fill = "steelblue") +
coord_polar(theta = "x")
```




```
collisions_2020 <- data_new %>%
  group_by(Contributing_Factor) %>%
  filter(Year == 2020)
```

```r
## distinct collisions factors for 2020
collisions_2020 <- distinct(collisions_2020)
```

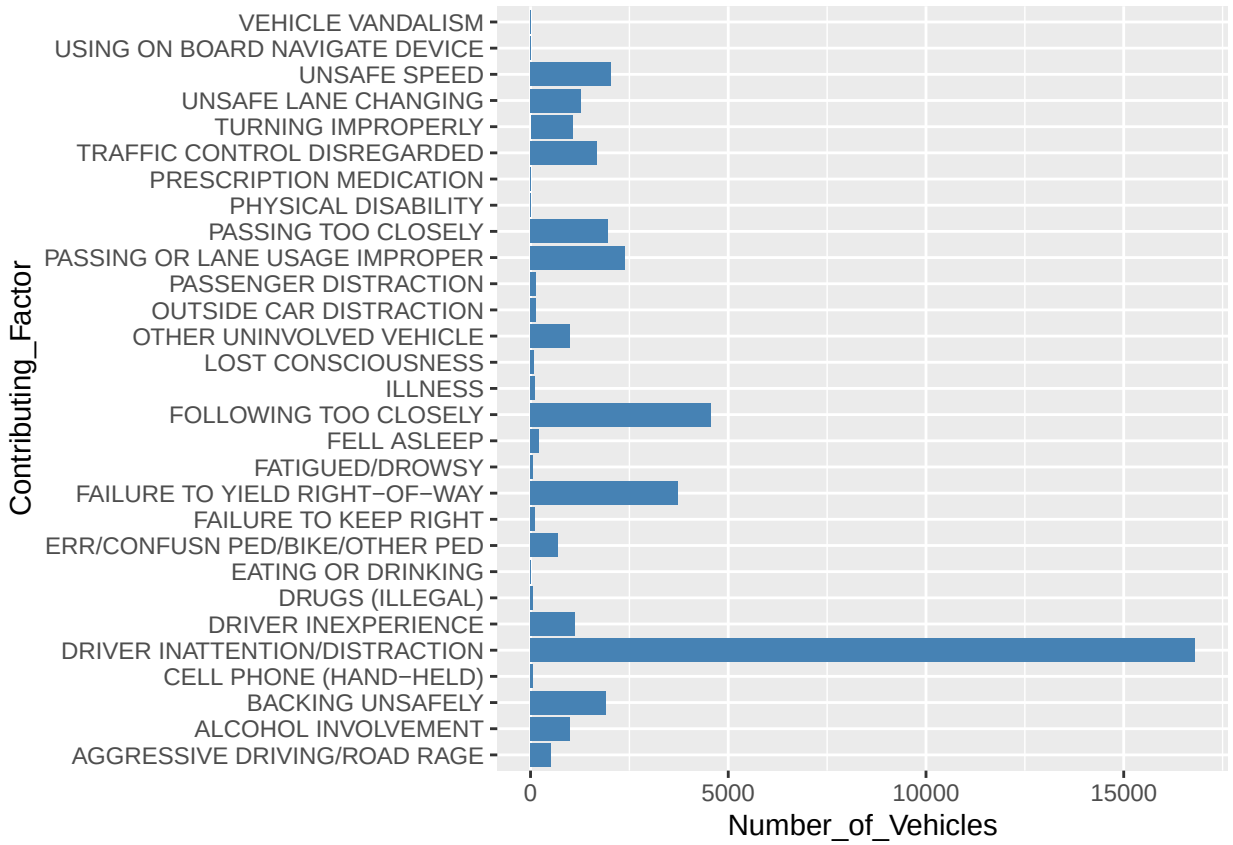
```r
# Main contributing factors of collisions in 2020
Main_Factor_2020 <- collisions_2020 %>%
  arrange(desc(Number_of_Vehicles)) %>%
  slice_head(n = 10)

 print(Main_Factor_2020)
```

```
## # A tibble: 174 x 24
## # Groups:   Contributing_Factor [29]
##     Year GeoCode.x GeoCodeLabel.x Number_of_Motor_Vehicle_C~1 Motorists_Involved
##    <dbl> <chr>     <chr>                                <dbl>              <dbl>
##  1  2020 C         CITYWIDE                              9429              18541
##  2  2020 M         MANHATTAN                             1383               2485
##  3  2020 B         BRONX                                 1868               3684
##  4  2020 K         BROOKLYN                              3126               6270
##  5  2020 Q         QUEENS                                2612               5211
##  6  2020 S         STATEN ISLAND                          440                891
##  7  2020 C         CITYWIDE                              9429              18541
##  8  2020 M         MANHATTAN                             1383               2485
##  9  2020 B         BRONX                                 1868               3684
## 10  2020 K         BROOKLYN                              3126               6270
## # i 164 more rows
## # i abbreviated name: 1: Number_of_Motor_Vehicle_Collisions
## # i 19 more variables: Injury_or_Fatal_Collisions <dbl>,
## #   MotoristsInjured <dbl>, MotoristsKilled <dbl>, PassengInjured <dbl>,
## #   PassengKilled <dbl>, CyclistsInjured <dbl>, CyclistsKilled <dbl>,
## #   PedestrInjured <dbl>, PedestrKilled <dbl>, Bicycle <dbl>, GeoCode.y <chr>,
## #   GeoCodeLabel.y <chr>, ContributingFactorCode <chr>, ...
```

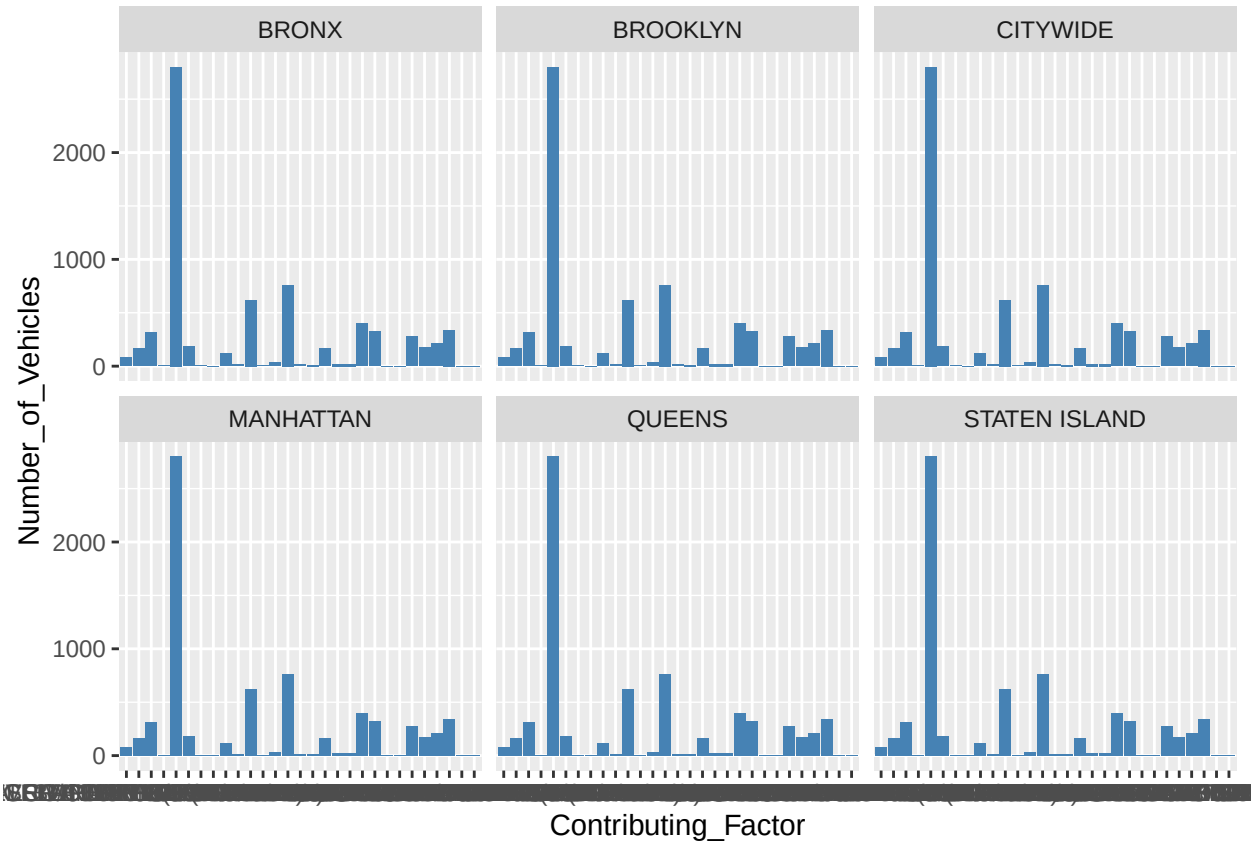
```r
ggplot(collisions_2020, aes(x = Contributing_Factor, y = Number_of_Vehicles
)) +
geom_bar(stat = "identity", freq = 500, fill = "steelblue")+ coord_flip()
```

```
## Warning in geom_bar(stat = "identity", freq = 500, fill = "steelblue"):
## Ignoring unknown parameters: `freq`
```




```
ggplot(collisions_2020 , aes(x = Contributing_Factor, y = Number_of_Vehicles
)) +
geom_bar(stat = "identity", freq = 500, fill = "steelblue")+ facet_wrap(~GeoCodeLabel)
```

```
## Warning in geom_bar(stat = "identity", freq = 500, fill = "steelblue"):
## Ignoring unknown parameters: `freq`
```



Calculating rates of non motorist casualties

```
# Non motorist casualties - rate per 1000

 data_new1<- data_new %>%
  mutate(rate = non_motorists_casualties / Number_of_Motor_Vehicle_Collisions * 1000)
data_new1
```

```
## # A tibble: 720 x 25
##     Year GeoCode.x GeoCodeLabel.x Number_of_Motor_Vehicle_C~1 Motorists_Involved
##    <dbl> <chr>     <chr>                                <dbl>              <dbl>
##  1  2014 C         CITYWIDE                             17720              34721
##  2  2014 C         CITYWIDE                             17720              34721
##  3  2014 C         CITYWIDE                             17720              34721
##  4  2014 C         CITYWIDE                             17720              34721
##  5  2014 C         CITYWIDE                             17720              34721
##  6  2014 C         CITYWIDE                             17720              34721
##  7  2014 C         CITYWIDE                             17720              34721
##  8  2014 C         CITYWIDE                             17720              34721
##  9  2014 C         CITYWIDE                             17720              34721
## 10  2014 C         CITYWIDE                             17720              34721
## # i 710 more rows
## # i abbreviated name: 1: Number_of_Motor_Vehicle_Collisions
## # i 20 more variables: Injury_or_Fatal_Collisions <dbl>,
## #   MotoristsInjured <dbl>, MotoristsKilled <dbl>, PassengInjured <dbl>,
## #   PassengKilled <dbl>, CyclistsInjured <dbl>, CyclistsKilled <dbl>,
## #   PedestrInjured <dbl>, PedestrKilled <dbl>, Bicycle <dbl>, GeoCode.y <chr>,
```

```
## #   GeoCodeLabel.y <chr>, ContributingFactorCode <chr>, ...
```

```
### 2020 rate of non-motorist casualties
rate1 <- data_new1 %>%
  filter(Year == 2020) %>%
  select(GeoCodeLabel.x, Number_of_Motor_Vehicle_Collisions, non_motorists_casualties, rate) %>%
  rename(GeoCodeLabel = GeoCodeLabel.x,Total_Collisions = Number_of_Motor_Vehicle_Collisions, rate_per_
distinct(rate1)
```

```
## # A tibble: 6 x 4
##   GeoCodeLabel  Total_Collisions non_motorists_casualties rate_per_1000
##   <chr>                   <dbl>                    <dbl>         <dbl>
## 1 CITYWIDE                 9429                     1270         135.
## 2 MANHATTAN                1383                      267         193.
## 3 BRONX                    1868                      221         118.
## 4 BROOKLYN                 3126                      456         146.
## 5 QUEENS                   2612                      289         111.
## 6 STATEN ISLAND             440                       37          84.1
```

```
rate1
```

```
## # A tibble: 174 x 4
##    GeoCodeLabel Total_Collisions non_motorists_casualties rate_per_1000
##    <chr>                   <dbl>                    <dbl>         <dbl>
##  1 CITYWIDE                 9429                     1270          135.
##  2 CITYWIDE                 9429                     1270          135.
##  3 CITYWIDE                 9429                     1270          135.
##  4 CITYWIDE                 9429                     1270          135.
##  5 CITYWIDE                 9429                     1270          135.
##  6 CITYWIDE                 9429                     1270          135.
##  7 CITYWIDE                 9429                     1270          135.
##  8 CITYWIDE                 9429                     1270          135.
##  9 CITYWIDE                 9429                     1270          135.
## 10 CITYWIDE                 9429                     1270          135.
## # i 164 more rows
```

```
### 2019 rate of non-motorist casualties
rate2<- data_new1 %>%
  filter(Year == 2019) %>%
  select(GeoCodeLabel.x, Number_of_Motor_Vehicle_Collisions, non_motorists_casualties, rate) %>%
  rename(GeoCodeLabel = GeoCodeLabel.x,Total_Collisions = Number_of_Motor_Vehicle_Collisions, rate_per_
distinct(rate2)
```

```
## # A tibble: 6 x 4
##   GeoCodeLabel  Total_Collisions non_motorists_casualties rate_per_1000
##   <chr>                   <dbl>                    <dbl>         <dbl>
## 1 CITYWIDE                17380                     1239          71.3
## 2 MANHATTAN                3470                      273          78.7
## 3 BRONX                    2886                      195          67.6
## 4 BROOKLYN                 5155                      458          88.8
## 5 QUEENS                   5323                      276          51.9
## 6 STATEN ISLAND             546                       37          67.8
```

```
rate2
```

```
## # A tibble: 192 x 4
##    GeoCodeLabel Total_Collisions non_motorists_casualties rate_per_1000
##    <chr>                   <dbl>                    <dbl>         <dbl>
```

```
##  1 CITYWIDE               17380                    1239          71.3
##  2 CITYWIDE               17380                    1239          71.3
##  3 CITYWIDE               17380                    1239          71.3
##  4 CITYWIDE               17380                    1239          71.3
##  5 CITYWIDE               17380                    1239          71.3
##  6 CITYWIDE               17380                    1239          71.3
##  7 CITYWIDE               17380                    1239          71.3
##  8 CITYWIDE               17380                    1239          71.3
##  9 CITYWIDE               17380                    1239          71.3
## 10 CITYWIDE               17380                    1239          71.3
## # i 182 more rows
```

Correlations: Evaluating the correlation between non - motorists casualties and the main contributing factors

Strong positive relationship between car collisions and passengers killed

```
(correlation <- cor( data_new$Number_of_Motor_Vehicle_Collisions,
data_new$PassengKilled))
```

```
## [1] 0.6401
```

Strong positive correlation between car collisions and death of passengers

```
# Correlation between car collisions and passenger death
(correlation <- cor( data_new$Number_of_Motor_Vehicle_Collisions,
data_new$PassengInjured))
```

```
## [1] 0.9358
```

A very strong and positive correlation between vehicle collisions and death of pedestrians

```
# Correlation between car collisions and pedestrian death
(correlation <- cor( data_new$Number_of_Motor_Vehicle_Collisions,
data_new$PedestrKilled))
```

```
## [1] 0.824
```

An even higher positive correlation between car collisions and pedestrians injuries

```
# Correlation between car collisions and pedestrian injuries
(correlation <- cor( data_new$Number_of_Motor_Vehicle_Collisions,
data_new$PedestrInjured))
```

```
## [1] 0.9332
```

Positive and strong correlation between car collisions and cyclists injured

```
# Correlation between car collisions and cyclist injuries
(correlation <- cor( data_new$Number_of_Motor_Vehicle_Collisions,
data_new$CyclistsInjured))
```

```
## [1] 0.7874
```

Positive and moderate correlation between car collisions and cyclists killed

```
(correlation <- cor( data_new$Number_of_Motor_Vehicle_Collisions,
data_new$CyclistsKilled))
```

```
## [1] 0.3282
```

Running Correlation tests

```
# Correlation test of injured pedestrians
(correlation <- cor.test( data_new$Number_of_Motor_Vehicle_Collisions,
data_new$PedestrInjured))
```

```
##
##  Pearson's product-moment correlation
##
## data:  data_new$Number_of_Motor_Vehicle_Collisions and data_new$PedestrInjured
## t = 70, df = 718, p-value <2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9231 0.9421
## sample estimates:
##    cor
## 0.9332
```

```
# correlation test for pedestrians killed
(correlation <- cor.test( data_new$Number_of_Motor_Vehicle_Collisions,
data_new$PedestrKilled))
```

```
##
##  Pearson's product-moment correlation
##
## data:  data_new$Number_of_Motor_Vehicle_Collisions and data_new$PedestrKilled
## t = 39, df = 718, p-value <2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7990 0.8461
## sample estimates:
##   cor
## 0.824
```

```
# correlation test for cyclists killed
(correlation <- cor.test( data_new$Number_of_Motor_Vehicle_Collisions,
data_new$CyclistsKilled))
```

```
##
##  Pearson's product-moment correlation
##
## data:  data_new$Number_of_Motor_Vehicle_Collisions and data_new$CyclistsKilled
## t = 9.3, df = 718, p-value <2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2614 0.3918
## sample estimates:
##    cor
## 0.3282
```

```
# Correlation test of passengers killed
(correlation <- cor.test( data_new$Number_of_Motor_Vehicle_Collisions,
data_new$PassengKilled))
```

```
##
##  Pearson's product-moment correlation
##
```

```
## data:  data_new$Number_of_Motor_Vehicle_Collisions and data_new$PassengKilled
## t = 22, df = 718, p-value <2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5949 0.6813
## sample estimates:
##    cor
## 0.6401
```

```
# Correlation test on passengers injured
(correlation <- cor.test( data_new$Number_of_Motor_Vehicle_Collisions,
data_new$PassengInjured))
```

```
##
##  Pearson's product-moment correlation
##
## data:  data_new$Number_of_Motor_Vehicle_Collisions and data_new$PassengInjured
## t = 71, df = 718, p-value <2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9261 0.9443
## sample estimates:
##    cor
## 0.9358
```

Main Findings

1. There were more non-motorists casualties in 2020 than in any years since 2014 , the year the city launched its vision zero initiative
2. The main causes of vehicles collisions are related to drivers inattention and following too closely
3. Brooklyn, Queens and Manhattan are the boroughs where it is dangerous to be a pedestrian or a cyclist. Staten Island is the safest borough for both pedestrians and cyclists
4. There is a Strong positive relationship between car collisions and passengers killed :0.6401476 5.There is a Strong positive correlation between car collisions and death of passengers : 0.9358199
5. There is a very strong and positive correlation between vehicle collisions and death of pedestrians:0.8239562
6. There is a very high positive correlation between car collisions and pedestrians injuries:0.9332216
7. Thers is positive and strong correlation between car collisions and cyclists injured:0.7873529
8. There is also a positive but moderate correlation between car collisions and cyclists killed:0.3281672