CAPSTONE Project _TTC DELAY ANALYSIS AND PREDICTION

2024-02-17

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

#INITIAL ANALYSIS OF TTC DELAY DATASET:

```
# Load the TTC Delay data readxl package:
library(readxl)

# Read the Excel file
data2 <- read_excel("ttc-bus-delay-data-2021-23.xlsx")
```

```
## Warning: Expecting numeric in C1928 / R1928C3: got 'RAD 600'

## Warning: Expecting numeric in C2031 / R2031C3: got 'RAD'

## Warning: Expecting numeric in C2425 / R2425C3: got 'RAD'

## Warning: Expecting numeric in C3854 / R3854C3: got 'SHUTTLE'

## Warning: Expecting numeric in C4060 / R4060C3: got 'A236'

## Warning: Expecting numeric in C4156 / R4156C3: got 'SHUTTLE'

## Warning: Expecting numeric in C4526 / R4526C3: got 'RAD'

## Warning: Expecting numeric in C4596 / R4596C3: got 'RAD'

## Warning: Expecting numeric in C4914 / R4914C3: got 'LINE 1'

## Warning: Expecting numeric in C4976 / R4976C3: got 'OTC'

## Warning: Expecting numeric in C5819 / R5819C3: got 'LINE 1'

## Warning: Expecting numeric in C5825 / R5825C3: got 'LINE 1'
```

```
## Warning: Expecting numeric in C5827 / R5827C3: got 'LINE 1'

## Warning: Expecting numeric in C5967 / R5967C3: got 'RAD'

## Warning: Expecting numeric in C6036 / R6036C3: got 'RSEM'

## Warning: Expecting numeric in C6037 / R6037C3: got 'OTC'

## Warning: Expecting numeric in C6379 / R6379C3: got 'RAD 21'

## Warning: Expecting numeric in C6430 / R6430C3: got 'RAD'

## Warning: Expecting numeric in C7314 / R7314C3: got 'RAD'

## Warning: Expecting numeric in C7881 / R7881C3: got 'LINE 1'

## Warning: Expecting numeric in C8051 / R8051C3: got 'LIINE 1'

## Warning: Expecting numeric in C8069 / R8069C3: got 'OTC'

## Warning: Expecting numeric in C8318 / R8318C3: got 'LINE 1'

## Warning: Expecting numeric in C8320 / R8320C3: got 'LINE 1'

## Warning: Expecting numeric in C8850 / R8850C3: got 'RAD'

## Warning: Expecting numeric in C8875 / R8875C3: got 'RAD'

## Warning: Expecting numeric in C9024 / R9024C3: got 'LINE 1'

## Warning: Expecting numeric in C9028 / R9028C3: got 'LINE 1'

## Warning: Expecting numeric in C9062 / R9062C3: got 'FLEET'

## Warning: Expecting numeric in C9487 / R9487C3: got 'YU'

## Warning: Expecting numeric in C9602 / R9602C3: got 'RAD 600'

## Warning: Expecting numeric in C9784 / R9784C3: got 'BD'

## Warning: Expecting numeric in C9884 / R9884C3: got 'NON'

## Warning: Expecting numeric in C10027 / R10027C3: got 'LINE 1'

## Warning: Expecting numeric in C10264 / R10264C3: got 'SHP'

## Warning: Expecting numeric in C10602 / R10602C3: got 'SHUTTLE'
```

```
## Warning: Expecting numeric in C10603 / R10603C3: got 'LINE 1'

## Warning: Expecting numeric in C10679 / R10679C3: got 'LINE 1'

## Warning: Expecting numeric in C10910 / R10910C3: got 'LINE 1'

## Warning: Expecting numeric in C11213 / R11213C3: got 'YU'

## Warning: Expecting numeric in C11471 / R11471C3: got 'RAD'

## Warning: Expecting numeric in C11627 / R11627C3: got 'BD'

## Warning: Expecting numeric in C11738 / R11738C3: got 'HILLCREST'

## Warning: Expecting numeric in C11914 / R11914C3: got 'BD'

## Warning: Expecting numeric in C12460 / R12460C3: got 'SHUTTLE'

## Warning: Expecting numeric in C12601 / R12601C3: got 'RAD'

## Warning: Expecting numeric in C13454 / R13454C3: got 'RAD'

## Warning: Expecting numeric in C14616 / R14616C3: got 'RAD'

## Warning: Expecting numeric in C14948 / R14948C3: got 'RAD'

## Warning: Expecting numeric in C14965 / R14965C3: got 'SHUTTLE'

## Warning: Expecting numeric in C15206 / R15206C3: got 'RAD'

## Warning: Expecting numeric in C15221 / R15221C3: got 'RAD'

## Warning: Expecting numeric in C15351 / R15351C3: got 'SHUTTLE'

## Warning: Expecting numeric in C17143 / R17143C3: got 'RAD'

## Warning: Expecting numeric in C17899 / R17899C3: got 'SHUTTLE'

## Warning: Expecting numeric in C17925 / R17925C3: got 'SHUTTLE'

## Warning: Expecting numeric in C17942 / R17942C3: got 'SHUTTLE'

## Warning: Expecting numeric in C18170 / R18170C3: got 'RAD'

## Warning: Expecting numeric in C18244 / R18244C3: got 'OTC'

## Warning: Expecting numeric in C18449 / R18449C3: got 'SHUTTLE'
```

```
## Warning: Expecting numeric in C18562 / R18562C3: got 'OTC'

## Warning: Expecting numeric in C18662 / R18662C3: got 'OTC'

## Warning: Expecting numeric in C18854 / R18854C3: got 'RAD'

## Warning: Expecting numeric in C19542 / R19542C3: got 'RAD'

## Warning: Expecting numeric in C20005 / R20005C3: got 'SHUTTLE'

## Warning: Expecting numeric in C20006 / R20006C3: got 'SHUTTLE'

## Warning: Expecting numeric in C20011 / R20011C3: got 'SHUTTLE'

## Warning: Expecting numeric in C20031 / R20031C3: got 'SHUTTLE'

## Warning: Expecting numeric in C20584 / R20584C3: got 'SHUTTLE'

## Warning: Expecting numeric in C20585 / R20585C3: got 'SHUTTLE'

## Warning: Expecting numeric in C20621 / R20621C3: got 'LINE 2'

## Warning: Expecting numeric in C20645 / R20645C3: got 'SHUTTLE BUS - LINE 2'

## Warning: Expecting numeric in C20651 / R20651C3: got 'LINE 2'

## Warning: Expecting numeric in C20652 / R20652C3: got 'LINE 2'

## Warning: Expecting numeric in C20655 / R20655C3: got 'LINE 2'

## Warning: Expecting numeric in C20773 / R20773C3: got 'LINE 2'

## Warning: Expecting numeric in C20776 / R20776C3: got 'LINE 2'

## Warning: Expecting numeric in C20820 / R20820C3: got 'LINE 2'

## Warning: Expecting numeric in C20922 / R20922C3: got 'RAD'

## Warning: Expecting numeric in C21125 / R21125C3: got 'LINE 1'

## Warning: Expecting numeric in C21482 / R21482C3: got 'BROADVIEW'

## Warning: Expecting numeric in C21493 / R21493C3: got 'SHUTTLE'

## Warning: Expecting numeric in C21565 / R21565C3: got 'LINE 2'

## Warning: Expecting numeric in C21569 / R21569C3: got 'LINE 2'
```

```
## Warning: Expecting numeric in C21581 / R21581C3: got 'LINE 2'

## Warning: Expecting numeric in C21634 / R21634C3: got 'LINE 2'

## Warning: Expecting numeric in C21654 / R21654C3: got 'LINE 2'

## Warning: Expecting numeric in C21672 / R21672C3: got 'LINE 2'

## Warning: Expecting numeric in C21718 / R21718C3: got 'LINE 2'

## Warning: Expecting numeric in C24039 / R24039C3: got 'RAD / 501'

## Warning: Expecting numeric in C24190 / R24190C3: got 'LINE 2'

## Warning: Expecting numeric in C24555 / R24555C3: got 'BD'

## Warning: Expecting numeric in C24660 / R24660C3: got 'SHUTTLE'

## Warning: Expecting numeric in C25252 / R25252C3: got 'LINE 1'

## Warning: Expecting numeric in C25257 / R25257C3: got 'LINE 1'

## Warning: Expecting numeric in C25279 / R25279C3: got 'RAD'

## Warning: Expecting numeric in C25319 / R25319C3: got 'LINE 1'

## Warning: Expecting numeric in C25843 / R25843C3: got 'BD'

## Warning: Expecting numeric in C26696 / R26696C3: got 'LINE 1'

## Warning: Expecting numeric in C26794 / R26794C3: got 'LINE 1'

## Warning: Expecting numeric in C28778 / R28778C3: got 'YU'

## Warning: Expecting numeric in C28964 / R28964C3: got 'LINE 3'

## Warning: Expecting numeric in C29060 / R29060C3: got 'FLEET'

## Warning: Expecting numeric in C29085 / R29085C3: got 'SHUTTLE'

## Warning: Expecting numeric in C29400 / R29400C3: got 'LINE 3'

## Warning: Expecting numeric in C29424 / R29424C3: got 'SRT'

## Warning: Expecting numeric in C29439 / R29439C3: got 'SRT'

## Warning: Expecting numeric in C29520 / R29520C3: got 'LINE 3'
```

```
## Warning: Expecting numeric in C29586 / R29586C3: got 'SRT SHUTTE LINE 3'

## Warning: Expecting numeric in C29659 / R29659C3: got 'YU'

## Warning: Expecting numeric in C30185 / R30185C3: got 'YU'

## Warning: Expecting numeric in C30358 / R30358C3: got 'LINE 3'

## Warning: Expecting numeric in C30502 / R30502C3: got 'CARIBANA'

## Warning: Expecting numeric in C30540 / R30540C3: got 'SHUTTLE'

## Warning: Expecting numeric in C30707 / R30707C3: got 'RAD'

## Warning: Expecting numeric in C30813 / R30813C3: got 'LINE 3'

## Warning: Expecting numeric in C30835 / R30835C3: got 'BD'

## Warning: Expecting numeric in C30948 / R30948C3: got 'LINE 3'

## Warning: Expecting numeric in C31114 / R31114C3: got 'LINE 3'

## Warning: Expecting numeric in C31179 / R31179C3: got 'SHUTTLE'

## Warning: Expecting numeric in C31202 / R31202C3: got 'RAD'

## Warning: Expecting numeric in C31353 / R31353C3: got 'LINE 3'

## Warning: Expecting numeric in C31548 / R31548C3: got 'SHUTTLE'

## Warning: Expecting numeric in C31585 / R31585C3: got 'A'

## Warning: Expecting numeric in C32092 / R32092C3: got 'LINE 1'

## Warning: Expecting numeric in C32191 / R32191C3: got 'YU'

## Warning: Expecting numeric in C32371 / R32371C3: got 'LINE 3'

## Warning: Expecting numeric in C32401 / R32401C3: got 'SRT'

## Warning: Expecting numeric in C32512 / R32512C3: got 'LINE 3'

## Warning: Expecting numeric in C32654 / R32654C3: got 'RAD'

## Warning: Expecting numeric in C32873 / R32873C3: got 'BD'

## Warning: Expecting numeric in C32910 / R32910C3: got 'LINE 3'
```

```
## Warning: Expecting numeric in C33120 / R33120C3: got 'SHUTTLE'

## Warning: Expecting numeric in C33133 / R33133C3: got 'LINE 3'

## Warning: Expecting numeric in C33662 / R33662C3: got 'SRT'

## Warning: Expecting numeric in C34210 / R34210C3: got 'RAD'

## Warning: Expecting numeric in C34230 / R34230C3: got 'SHUTTLE'

## Warning: Expecting numeric in C35410 / R35410C3: got 'LINE 3'

## Warning: Expecting numeric in C35501 / R35501C3: got 'BD'

## Warning: Expecting numeric in C36186 / R36186C3: got 'RAD 600'

## Warning: Expecting numeric in C37090 / R37090C3: got 'LINE 1'

## Warning: Expecting numeric in C37466 / R37466C3: got 'SHUTTLE'

## Warning: Expecting numeric in C37493 / R37493C3: got 'LINE 3'

## Warning: Expecting numeric in C37737 / R37737C3: got 'SHUTTLE'

## Warning: Expecting numeric in C37904 / R37904C3: got 'LINE 3'

## Warning: Expecting numeric in C38367 / R38367C3: got 'BD'

## Warning: Expecting numeric in C39204 / R39204C3: got 'OTC'

## Warning: Expecting numeric in C39512 / R39512C3: got 'LINE 1'

## Warning: Expecting numeric in C39516 / R39516C3: got 'LINE 1'

## Warning: Expecting numeric in C39638 / R39638C3: got 'YU'

## Warning: Expecting numeric in C39698 / R39698C3: got 'SHUTTLE'

## Warning: Expecting numeric in C39868 / R39868C3: got 'LINE 1'

## Warning: Expecting numeric in C39875 / R39875C3: got 'LINE 1'

## Warning: Expecting numeric in C39884 / R39884C3: got 'LINE 1'

## Warning: Expecting numeric in C39986 / R39986C3: got 'LINE1'

## Warning: Expecting numeric in C40021 / R40021C3: got 'LINE 1'
```

```
## Warning: Expecting numeric in C40035 / R40035C3: got 'YU'

## Warning: Expecting numeric in C40069 / R40069C3: got 'YU'

## Warning: Expecting numeric in C40131 / R40131C3: got '600 - ROUTE LINE 1'

## Warning: Expecting numeric in C41468 / R41468C3: got 'RAD'

## Warning: Expecting numeric in C43668 / R43668C3: got 'YU'

## Warning: Expecting numeric in C44189 / R44189C3: got 'OTC'

## Warning: Expecting numeric in C44946 / R44946C3: got 'RAD'

## Warning: Expecting numeric in C45270 / R45270C3: got 'YU'

## Warning: Expecting numeric in C45613 / R45613C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C45698 / R45698C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C45795 / R45795C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C46126 / R46126C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C46143 / R46143C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C46167 / R46167C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C46240 / R46240C3: got 'LINE 1'

## Warning: Expecting numeric in C46248 / R46248C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C46303 / R46303C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C46357 / R46357C3: got 'LINE 1'

## Warning: Expecting numeric in C46358 / R46358C3: got 'RAD'

## Warning: Expecting numeric in C46377 / R46377C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C46430 / R46430C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C46439 / R46439C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C46485 / R46485C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C46648 / R46648C3: got '927 HIGHWAY 27'
```

```
## Warning: Expecting numeric in C46649 / R46649C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C46775 / R46775C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C46809 / R46809C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C46953 / R46953C3: got '600 (75'

## Warning: Expecting numeric in C47089 / R47089C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C47258 / R47258C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C47279 / R47279C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C47466 / R47466C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C47468 / R47468C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C47518 / R47518C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C47641 / R47641C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C47729 / R47729C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C47781 / R47781C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C47831 / R47831C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C47861 / R47861C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C48227 / R48227C3: got 'LINE 1 SHUTTLE - 600'

## Warning: Expecting numeric in C48233 / R48233C3: got 'SHUTTLE'

## Warning: Expecting numeric in C48524 / R48524C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C49108 / R49108C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C49180 / R49180C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C49345 / R49345C3: got 'YU'

## Warning: Expecting numeric in C49490 / R49490C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C49544 / R49544C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C49748 / R49748C3: got 'BD'
```

```
## Warning: Expecting numeric in C50162 / R50162C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C50263 / R50263C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C50363 / R50363C3: got '600 - ROUTE 301'

## Warning: Expecting numeric in C50419 / R50419C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C50474 / R50474C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C50603 / R50603C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C50722 / R50722C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C50809 / R50809C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C50864 / R50864C3: got '52 LAWRENCE WEST - 600'

## Warning: Expecting numeric in C50876 / R50876C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C50882 / R50882C3: got 'CHANGEOF'

## Warning: Expecting numeric in C55401 / R55401C3: got 'A242'

## Warning: Expecting numeric in C56989 / R56989C3: got 'OTC'

## Warning: Expecting numeric in C57011 / R57011C3: got 'RAD'

## Warning: Expecting numeric in C57196 / R57196C3: got 'OTC'

## Warning: Expecting numeric in C58240 / R58240C3: got 'OTC'

## Warning: Expecting numeric in C58335 / R58335C3: got 'OTC'

## Warning: Expecting numeric in C58853 / R58853C3: got 'RAD'

## Warning: Expecting numeric in C58993 / R58993C3: got 'RAD'

## Warning: Expecting numeric in C58996 / R58996C3: got 'OTC'

## Warning: Expecting numeric in C59095 / R59095C3: got 'RAD'

## Warning: Expecting numeric in C59215 / R59215C3: got 'RAD'

## Warning: Expecting numeric in C60790 / R60790C3: got 'RAD'

## Warning: Expecting numeric in C61424 / R61424C3: got 'RAD'
```

```
## Warning: Expecting numeric in C61847 / R61847C3: got 'RAD'

## Warning: Expecting numeric in C62778 / R62778C3: got 'OTC'

## Warning: Expecting numeric in C64856 / R64856C3: got 'OTC'

## Warning: Expecting numeric in C65111 / R65111C3: got '600 RAD (LINE 1'

## Warning: Expecting numeric in C69041 / R69041C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C69268 / R69268C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C69300 / R69300C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C69584 / R69584C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C69710 / R69710C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C69726 / R69726C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C69886 / R69886C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C70123 / R70123C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C70133 / R70133C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C70657 / R70657C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C71012 / R71012C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C71068 / R71068C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C71145 / R71145C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C71214 / R71214C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C71240 / R71240C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C71308 / R71308C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C71337 / R71337C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C71653 / R71653C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C71744 / R71744C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C71856 / R71856C3: got 'RAD'
```

```
## Warning: Expecting numeric in C71857 / R71857C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C72120 / R72120C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C72295 / R72295C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C72339 / R72339C3: got 'RAD'

## Warning: Expecting numeric in C72356 / R72356C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C72364 / R72364C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C72890 / R72890C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C73037 / R73037C3: got '939 FINCH EXPRESS / 39'

## Warning: Expecting numeric in C73080 / R73080C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C73151 / R73151C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C73371 / R73371C3: got '927 HIGHWAY 27'

## Warning: Expecting numeric in C91126 / R91126C3: got 'RAD'

## Warning: Expecting numeric in C91663 / R91663C3: got 'OTC'

## Warning: Expecting numeric in C93093 / R93093C3: got 'OTC'

## Warning: Expecting numeric in C105290 / R105290C3: got 'LINE 2'

## Warning: Expecting numeric in C111627 / R111627C3: got 'BD'

## Warning: Expecting numeric in C112153 / R112153C3: got '96 WILSON/996 AND 165'

## Warning: Expecting numeric in C117705 / R117705C3: got 'DON MILLS'

## Warning: Expecting numeric in C119690 / R119690C3: got 'LEASIDE'

## Warning: Expecting numeric in C121486 / R121486C3: got 'BD'

## Warning: Expecting numeric in C125181 / R125181C3: got 'RAD'

## Warning: Expecting numeric in C126226 / R126226C3: got 'TEST CAR'

## Warning: Expecting numeric in C127253 / R127253C3: got 'RAD'

## Warning: Expecting numeric in C128122 / R128122C3: got 'RAD'
```

```
## Warning: Expecting numeric in C129172 / R129172C3: got 'RAD'

## Warning: Expecting numeric in C129331 / R129331C3: got 'RAD'

## Warning: Expecting numeric in C129459 / R129459C3: got 'RAD'

## Warning: Expecting numeric in C130302 / R130302C3: got 'RAD'

## Warning: Expecting numeric in C130307 / R130307C3: got 'RAD'

## Warning: Expecting numeric in C132887 / R132887C3: got 'RAD'

## Warning: Expecting numeric in C132905 / R132905C3: got 'RAD'

## Warning: Expecting numeric in C133466 / R133466C3: got 'RAD'

## Warning: Expecting numeric in C135260 / R135260C3: got 'RAD'

## Warning: Expecting numeric in C136558 / R136558C3: got 'RAD'

## Warning: Expecting numeric in C136561 / R136561C3: got 'RAD'

## Warning: Expecting numeric in C137044 / R137044C3: got 'RAD'
```

```
# View the data (optional)
head(data2)
```

```
## # A tibble: 6 x 11
##   Date                Mode of Transportati~1 Route Time  Day   Location Incident
##   <dttm>              <chr>                  <dbl> <chr> <chr> <chr>    <chr>
## 1 2023-01-01 00:00:00 Bus                       91 02:30 Sund~ WOODBIN~ Diversi~
## 2 2023-01-01 00:00:00 Bus                       69 02:34 Sund~ WARDEN ~ Security
## 3 2023-01-01 00:00:00 Bus                       35 03:06 Sund~ JANE ST~ Cleaning
## 4 2023-01-01 00:00:00 Bus                      900 03:14 Sund~ KIPLING~ Security
## 5 2023-01-01 00:00:00 Bus                       85 03:43 Sund~ MEADOWA~ Security
## 6 2023-01-01 00:00:00 Bus                       40 03:47 Sund~ KIPLING~ Emergen~
## # i abbreviated name: 1: `Mode of Transportation`
## # i 4 more variables: `Min Delay` <dbl>, `Min Gap` <dbl>, Direction <chr>,
## #   Vehicle <dbl>
```

```
str(data2)
```

```
## tibble [151,889 x 11] (S3: tbl_df/tbl/data.frame)
##  $ Date                  : POSIXct[1:151889], format: "2023-01-01" "2023-01-01" ...
##  $ Mode of Transportation: chr [1:151889] "Bus" "Bus" "Bus" "Bus" ...
##  $ Route                 : num [1:151889] 91 69 35 900 85 40 336 52 24 36 ...
##  $ Time                  : chr [1:151889] "02:30" "02:34" "03:06" "03:14" ...
##  $ Day                   : chr [1:151889] "Sunday" "Sunday" "Sunday" "Sunday" ...
##  $ Location              : chr [1:151889] "WOODBINE AND MORTIMER" "WARDEN STATION" "JANE STATION" "K
##  $ Incident              : chr [1:151889] "Diversion" "Security" "Cleaning" "Security" ...
##  $ Min Delay             : num [1:151889] 81 22 30 17 1 0 138 30 20 334 ...
##  $ Min Gap               : num [1:151889] 111 44 60 17 1 0 168 60 40 344 ...
##  $ Direction             : chr [1:151889] "W" "S" "N" "N" ...
##  $ Vehicle               : num [1:151889] 8772 8407 1051 3334 1559 ...
```

```
summary(data2)
```

```
##       Date                          Mode of Transportation     Route
##  Min.   :2021-01-01 00:00:00.00   Length:151889          Min.   :     1.0
##  1st Qu.:2021-12-09 00:00:00.00   Class :character       1st Qu.:    37.0
##  Median :2022-08-03 00:00:00.00   Mode  :character       Median :    72.0
##  Mean   :2022-07-21 21:51:46.49                         Mean   :   197.4
##  3rd Qu.:2023-04-06 00:00:00.00                         3rd Qu.:   122.0
##  Max.   :2023-11-30 00:00:00.00                         Max.   :898630.0
##                                                         NA's   :1440
##      Time               Day              Location           Incident
##  Length:151889      Length:151889      Length:151889      Length:151889
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    Min Delay        Min Gap       Direction            Vehicle
##  Min.   :  0.0   Min.   :  0.0   Length:151889      Min.   :    0
##  1st Qu.:  9.0   1st Qu.: 17.0   Class :character   1st Qu.: 3110
##  Median : 11.0   Median : 22.0   Mode  :character   Median : 7261
##  Mean   : 19.9   Mean   : 32.5                      Mean   : 5467
##  3rd Qu.: 20.0   3rd Qu.: 38.0                      3rd Qu.: 8549
##  Max.   :999.0   Max.   :999.0                      Max.   :99035
##
```

#ASSIGNING CORRECT DATA TYPE:

```
data2$Day <- as.factor(data2$Day)
data2$Incident <- as.factor(data2$Incident)
data2$Direction <- as.factor(data2$Direction)
str(data2)
```

```
## tibble [151,889 x 11] (S3: tbl_df/tbl/data.frame)
##  $ Date                  : POSIXct[1:151889], format: "2023-01-01" "2023-01-01" ...
##  $ Mode of Transportation: chr [1:151889] "Bus" "Bus" "Bus" "Bus" ...
##  $ Route                 : num [1:151889] 91 69 35 900 85 40 336 52 24 36 ...
##  $ Time                  : chr [1:151889] "02:30" "02:34" "03:06" "03:14" ...
##  $ Day                   : Factor w/ 7 levels "Friday","Monday",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ Location              : chr [1:151889] "WOODBINE AND MORTIMER" "WARDEN STATION" "JANE STATION" "K
##  $ Incident              : Factor w/ 12 levels "Cleaning","Collision",..: 3 11 1 11 11 4 3 4 1 3 ...
##  $ Min Delay             : num [1:151889] 81 22 30 17 1 0 138 30 20 334 ...
##  $ Min Gap               : num [1:151889] 111 44 60 17 1 0 168 60 40 344 ...
##  $ Direction             : Factor w/ 5 levels "B","E","N","S",..: 5 4 3 3 3 5 3 2 5 5 ...
##  $ Vehicle               : num [1:151889] 8772 8407 1051 3334 1559 ...
```

#IDENTIFYING PRESENCE OF 'MISSING VALUES' IN THE TTC DATASET:

```
# Identifying Presence of Missing Data
missing_values <- colSums(is.na(data2))
```

```r
# Print the count of missing values for each column
print(missing_values)
```

```
##                    Date Mode of Transportation                   Route
##                       0                       0                    1440
##                    Time                     Day                Location
##                       0                       0                       0
##                Incident              Min Delay                 Min Gap
##                       0                       0                       0
##               Direction                 Vehicle
##                    3735                       0
```

#HANDLING MISSING VALUES:

```r
#Subsetting rows where route is not provided
data2 <- subset(data2, !is.na(Route))
library(DescTools)
```

```
## Warning: package 'DescTools' was built under R version 4.3.1
```

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

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

```r
# Replacing Direction attibutes with NA values based on route and its corresponding mode of direction
data2_filled <- data2 %>%
  group_by(Route) %>%
  mutate(Direction = ifelse(is.na(Direction), Mode(as.integer(Direction), na.rm = TRUE), Direction)) %>%
  ungroup()
missing_values <- colSums(is.na(data2_filled))

# Print the count of missing values for each column
print(missing_values)
```

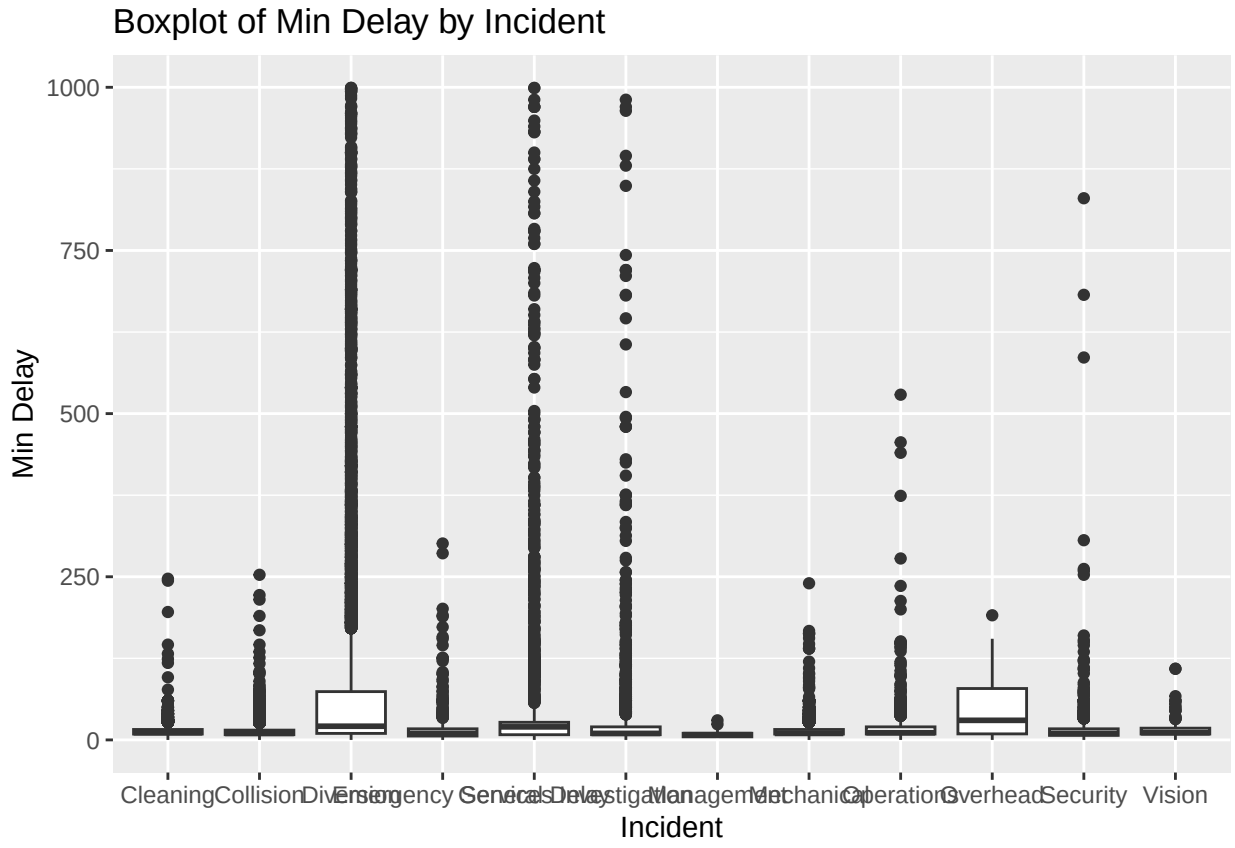
```
##                    Date Mode of Transportation                   Route
##                       0                       0                       0
##                    Time                     Day                Location
##                       0                       0                       0
##                Incident              Min Delay                 Min Gap
##                       0                       0                       0
##               Direction                 Vehicle
##                       0                       0
```

#IDENTIFYING PRESENCE OF OUTLIERS IN DEPENDENT VARIABLE 'MIN DELAY' ACROSS INCIDENT TYPES:

```
# Load the ggplot2 package
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.1
```

```
# OUTLIER ANALYSIS: Boxplot to clearly see the interquartile range
ggplot(data2_filled, aes(x = Incident, y = `Min Delay`)) +
  geom_boxplot(coef = 1.5) +  # Adjust the coef parameter to control the length of the whiskers
  labs(x = "Incident", y = "Min Delay", title = "Boxplot of Min Delay by Incident")
```



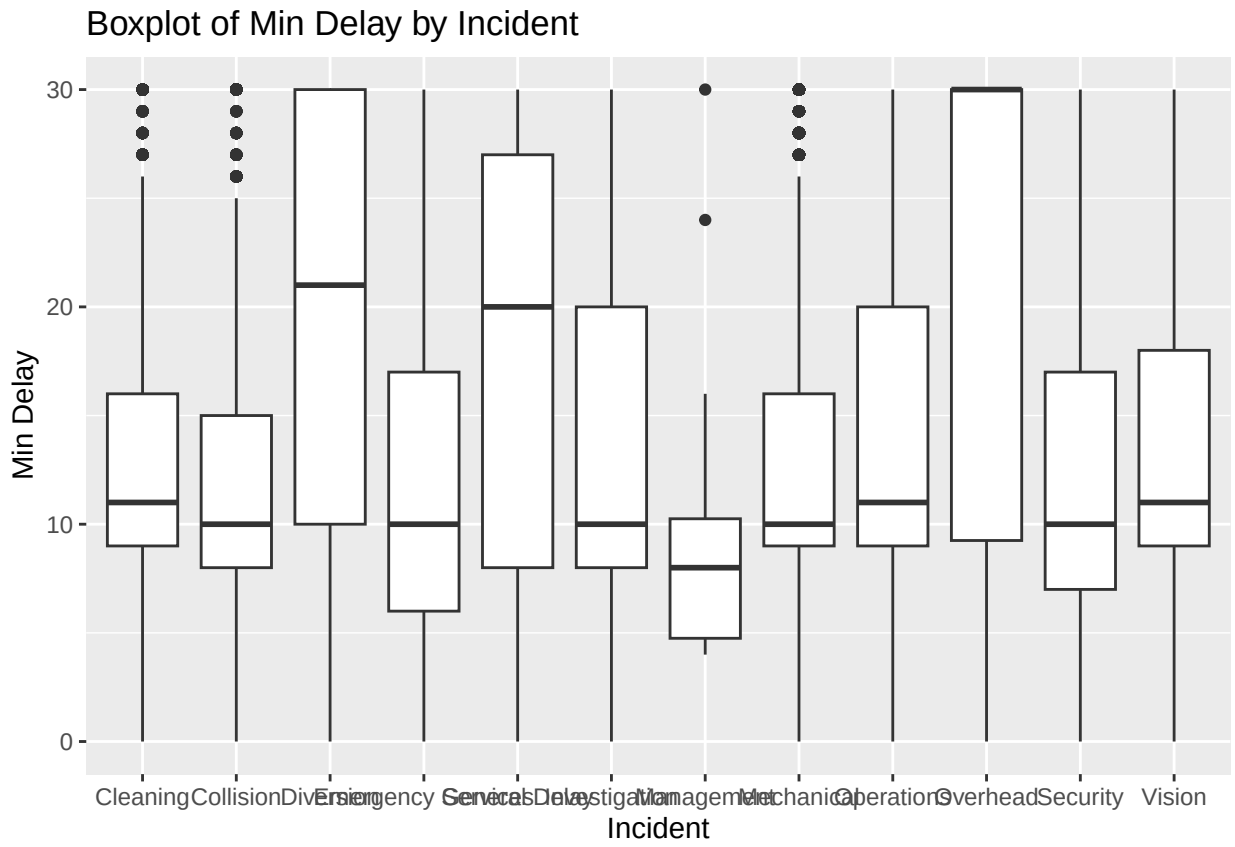
#PERFORMING 'WINSORIZATION' TO FIX OUTLIERS IN DEPENDENT VARIABLE 'MIN DELAY':

```
# Perform 'Winsorization' to Fix Outlier Issues: capping the outliers at a certain percentile. For exam
library(DescTools)

# Winsorize the 'Min Delay' column at the 2ND and 94th percentiles for the entire data
data2_filled$Min_Delay_Winsorized <- Winsorize(data2_filled$`Min Delay`, probs = c(0.02, 0.94))

# Check the results
summary(data2_filled$Min_Delay_Winsorized)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    9.00   11.00   13.93   20.00   30.00
```

BOXPLOT OF 'MIN DELAY' DATA ACROSS INCIDENT TYPES POST WINSORIZATION

```
library(ggplot2)
# OUTLIER ANALYSIS: Boxplot to clearly see the interquartile range
ggplot(data2_filled, aes(x = Incident, y = `Min_Delay_Winsorized`)) +
  geom_boxplot(coef = 1.5) +  # Adjust the coef parameter to control the length of the whiskers
  labs(x = "Incident", y = "Min Delay", title = "Boxplot of Min Delay by Incident")
```



#GROUPING 'MIN DELAY' DEPENDENT VARIABLE TO CHECK DATA IMBALANCE

```
# GROUPING 'MIN DELAY' DATA FOR FUTURE ANALYSIS

data2_filled_updated <- data2_filled %>%
  mutate(Delay_Severity = case_when(
    `Min_Delay_Winsorized` >= 0 & `Min_Delay_Winsorized` < 5 ~ "<5 Min",
    `Min_Delay_Winsorized` >= 5 & `Min_Delay_Winsorized` <= 10 ~ "5-10 Min",
    `Min_Delay_Winsorized` >10 & `Min_Delay_Winsorized` <= 15 ~ "11-15 Min",
    `Min_Delay_Winsorized` >15 & `Min_Delay_Winsorized` <= 20 ~ "16-20 Min",
    `Min_Delay_Winsorized` > 20 ~ ">20 Min",
    TRUE ~ "On Time"  # Handle cases where Min Delay is negative or other values
  ))
data2_filled_updated$Delay_Severity <- as.factor(data2_filled_updated$Delay_Severity)

# CHECK FOR DATA IMBALANCE
```

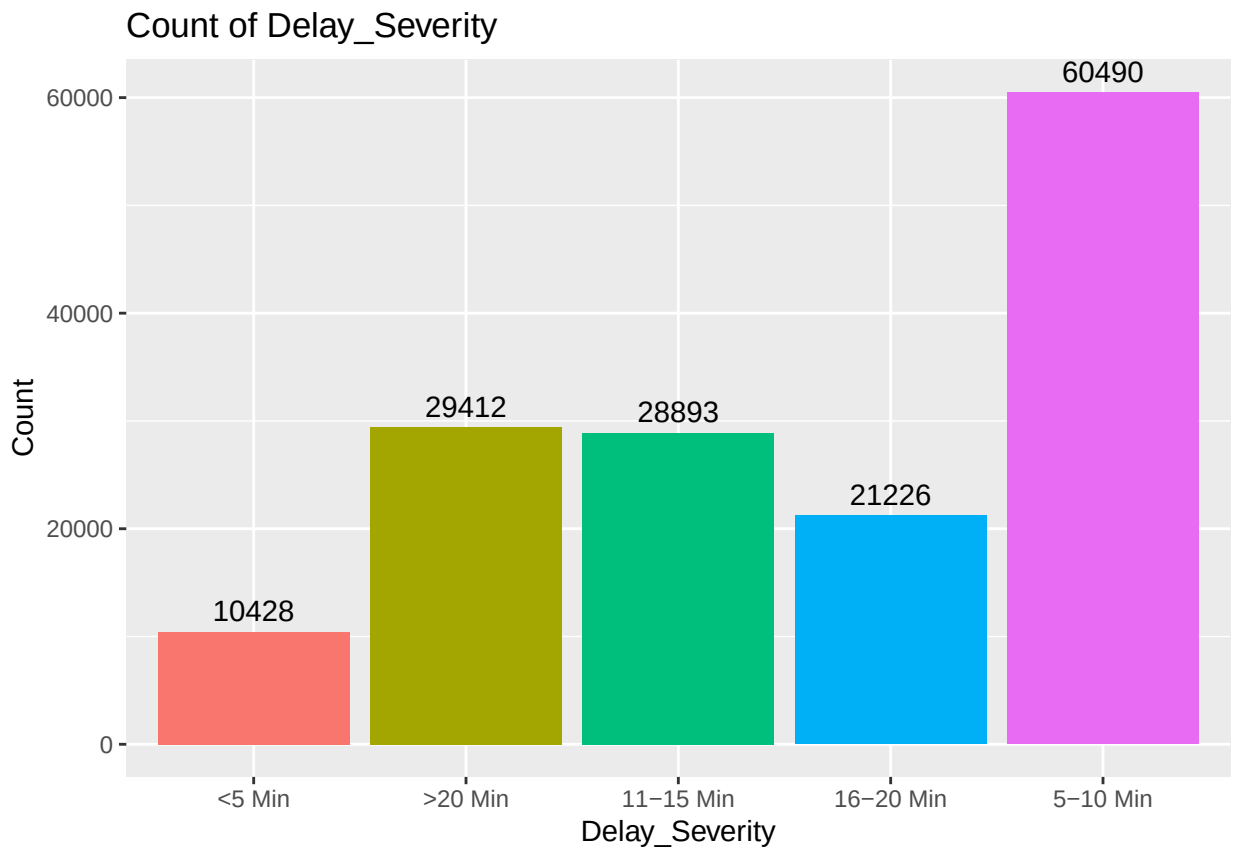
```r
library(dplyr)

# Count the frequency of each level in the 'Incident' column and calculate percentage
Delay_Severity_balance <- data2_filled_updated %>%
  count(Delay_Severity) %>%
  mutate(Percentage = n / sum(n) * 100) %>%
  arrange(desc(n))

# Print the counts and percentages to check for imbalance
print(Delay_Severity_balance)
```

```
## # A tibble: 5 x 3
##   Delay_Severity     n Percentage
##   <fct>          <int>      <dbl>
## 1 5-10 Min       60490      40.2
## 2 >20 Min        29412      19.5
## 3 11-15 Min      28893      19.2
## 4 16-20 Min      21226      14.1
## 5 <5 Min         10428       6.93
```

#BAR GRAPH OF 'DELAY_SEVERITY' DEPENDENT VARIABLE TO CHECK DATA IMBALANCE

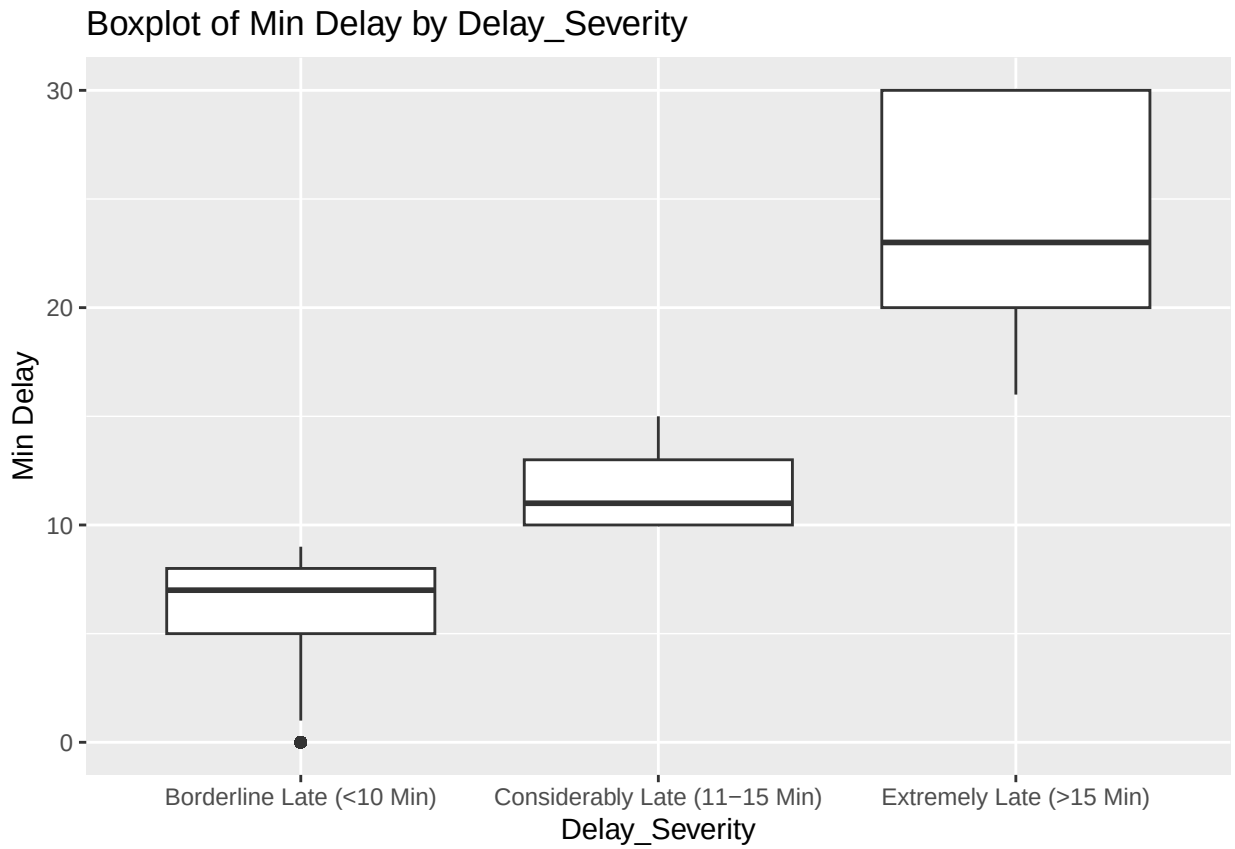
```r
#COUNT OF INCIDENTS
library(ggplot2)

# Create a bar graph with data labels and different bar colors
ggplot(data2_filled_updated, aes(x = Delay_Severity, fill = Delay_Severity)) +
  geom_bar() +
  geom_text(stat = 'count', aes(label = ..count..), vjust = -0.5) +  # Add data labels
  scale_fill_discrete(name = "Delay_Severity") +  # Customize legend title
  labs(x = "Delay_Severity", y = "Count", title = "Count of Delay_Severity") +
  theme(legend.position = "none")  # Hide legend (optional)
```

```
## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(count)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



#REDUCING NUMBER OF CLASS/LEVELS WITHIN TARGET VARIABLE FROM 5 TO 3 TO ADDRESS DATA IMBALANCE:

```
#REGROUPING 'DELAY_SEVERITY' DATA FOR EASE OF CLASSIFICATION

data2_filled_updated1 <- data2_filled_updated %>%
  mutate(Delay_Severity = case_when(
    `Min_Delay_Winsorized` >= 0 & `Min_Delay_Winsorized` < 10 ~ "Borderline Late (<10 Min)",
    `Min_Delay_Winsorized` >=10 & `Min_Delay_Winsorized` <= 15 ~ "Considerably Late (11-15 Min)",
    `Min_Delay_Winsorized` > 15 ~ "Extremely Late (>15 Min)",
    TRUE ~ "On Time"  # Handle cases where Min Delay is negative or other values
  ))
data2_filled_updated1$Delay_Severity <- as.factor(data2_filled_updated1$Delay_Severity)

# Create a customized boxplot to clearly see the interquartile range
ggplot(data2_filled_updated1, aes(x = Delay_Severity, y = `Min_Delay_Winsorized`)) +
  geom_boxplot(coef = 1.5) +  # Adjust the coef parameter to control the length of the whiskers
  labs(x = "Delay_Severity", y = "Min Delay", title = "Boxplot of Min Delay by Delay_Severity")
```



#DISTRIBUTION OF DELAY INCIDENCE ACROSS 3 CLASSES OF 'DELAY_SEVERITY' TARGET VARIABLE:

```
# CHECK FOR DATA IMBALANCE
library(dplyr)

# Count the frequency of each level in the 'Incident' column and calculate percentage
Delay_Severity_balance1 <- data2_filled_updated1 %>%
  count(Delay_Severity) %>%
  mutate(Percentage = n / sum(n) * 100) %>%
  arrange(desc(n))

# Print the counts and percentages to check for imbalance
print(Delay_Severity_balance1)
```

```
## # A tibble: 3 x 3
##   Delay_Severity                    n Percentage
##   <fct>                         <int>      <dbl>
## 1 Considerably Late (11-15 Min) 53115       35.3
## 2 Extremely Late (>15 Min)      50638       33.7
## 3 Borderline Late (<10 Min)     46696       31.0
```

#ADDRESSING CATEGORICAL VARIABLES WITH EXCESSIVE LABELS:

#'INCIDENT' VARAIBELE WITH 12 LABELS

```r
# Load necessary library
library(dplyr)

# Count incidents for each type and calculate the percentage of the grand total
incident_type_counts <- data2_filled_updated1 %>%
  group_by(Incident) %>%
  summarise(IncidentCount = n(), .groups = 'drop') %>%
  mutate(TotalIncidents = sum(IncidentCount), # Calculate the total number of incidents
         PercentageOfTotal = IncidentCount / TotalIncidents * 100) %>%
  arrange(desc(IncidentCount)) # Arrange by descending order of incident counts

# Print the results
print(incident_type_counts)
```

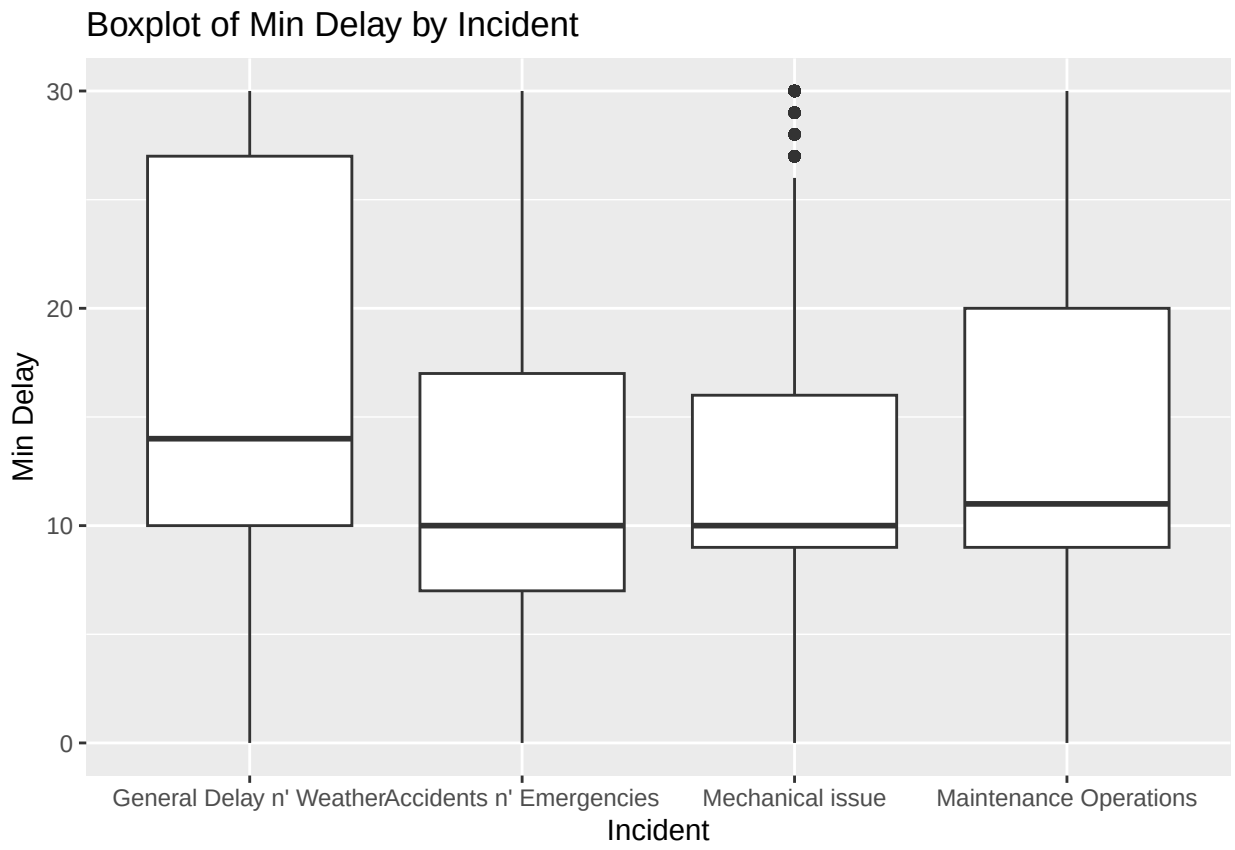
```
## # A tibble: 12 x 4
##    Incident           IncidentCount TotalIncidents PercentageOfTotal
##    <fct>                      <int>          <int>             <dbl>
##  1 Mechanical                 44113         150449           29.3
##  2 Operations                 39322         150449           26.1
##  3 Diversion                  14799         150449            9.84
##  4 Cleaning                   13052         150449            8.68
##  5 Security                   10209         150449            6.79
##  6 Collision                   9029         150449            6.00
##  7 General Delay               7599         150449            5.05
##  8 Emergency Services          6736         150449            4.48
##  9 Investigation               3510         150449            2.33
## 10 Vision                      2034         150449            1.35
## 11 Management                    28         150449            0.0186
## 12 Overhead                      18         150449            0.0120
```

REDUCING 'INCIDENT' LABELS FROM 12 TO 4:

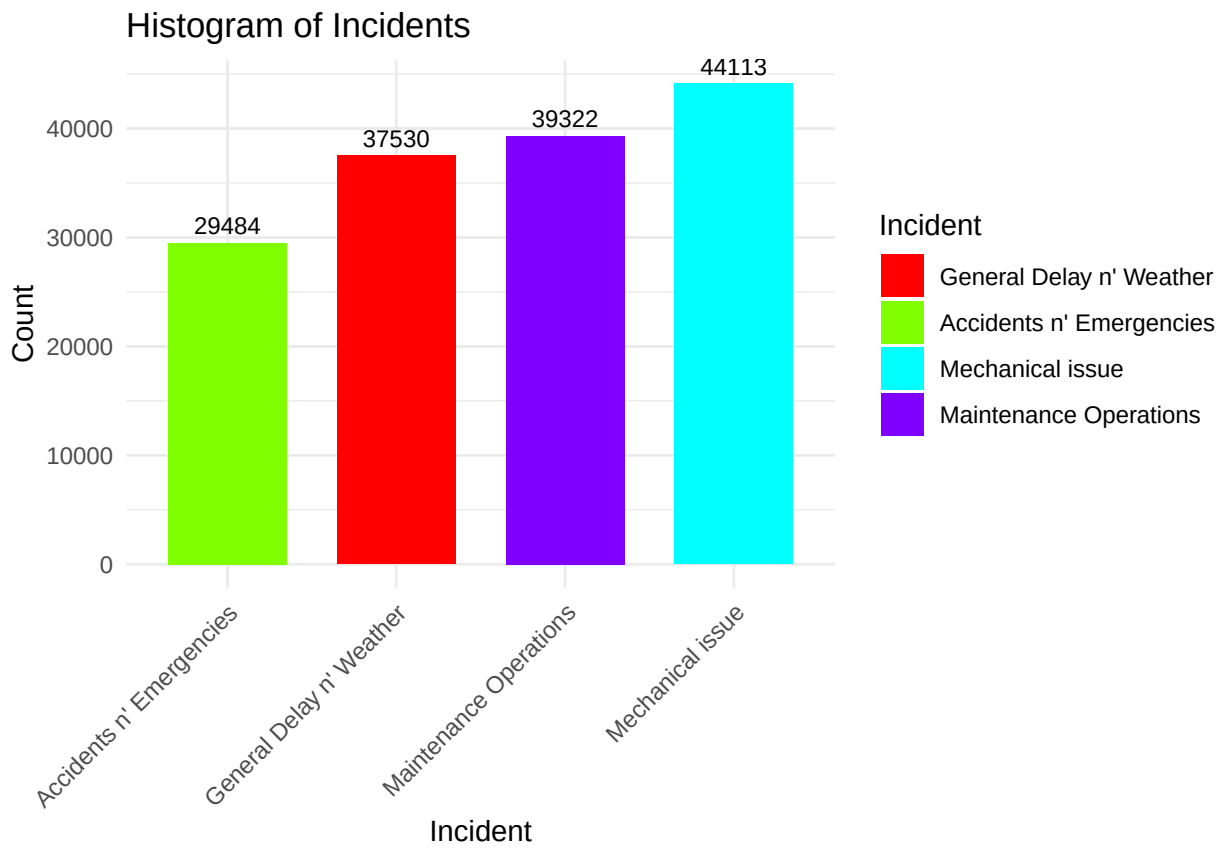
```r
data2_filled_updated2 <- data2_filled_updated1 %>%
  mutate(Incident = recode(Incident,
                           "Cleaning" = "General Delay n' Weather",
                           "Collision" = "Accidents n' Emergencies",
                           "Diversion" = "General Delay n' Weather",
                           "Emergency Services" = "Accidents n' Emergencies",
                           "General Delay" = "General Delay n' Weather",
                           "Investigation" = "Accidents n' Emergencies",
                           "Management" = "General Delay n' Weather",
                           "Mechanical" = "Mechanical issue",
                           "Operations" = "Maintenance Operations",
                           "Overhead" = "General Delay n' Weather",
                           "Security" = "Accidents n' Emergencies",
                           "Vision" = "General Delay n' Weather"
                           ))
```

```r
# Create a customized boxplot to clearly see the interquartile range
ggplot(data2_filled_updated2, aes(x = Incident, y = `Min_Delay_Winsorized`)) +
```

```
  geom_boxplot(coef = 1.5) +  # Adjust the coef parameter to control the length of the whiskers
  labs(x = "Incident", y = "Min Delay", title = "Boxplot of Min Delay by Incident")
```




```
library(dplyr)
library(ggplot2)
library(forcats)  # Load the forcats package for fct_reorder

# Calculate the count of each incident
incident_counts <- data2_filled_updated2 %>%
  count(Incident) %>%
  arrange(desc(n))  # Arrange in descending order of frequency

# Create a ggplot with multiple colors, data labels, and ordered incident types
ggplot(incident_counts, aes(x = fct_reorder(Incident, n), y = n, fill = Incident)) +
  geom_bar(stat = "identity", width = 0.7) +
  geom_text(aes(label = n), vjust = -0.5, size = 3) +  # Add data labels
  labs(title = "Histogram of Incidents", x = "Incident", y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +  # Rotate x-axis labels
  scale_fill_manual(values = rainbow(length(incident_counts$Incident)))  # Use multicolor
```

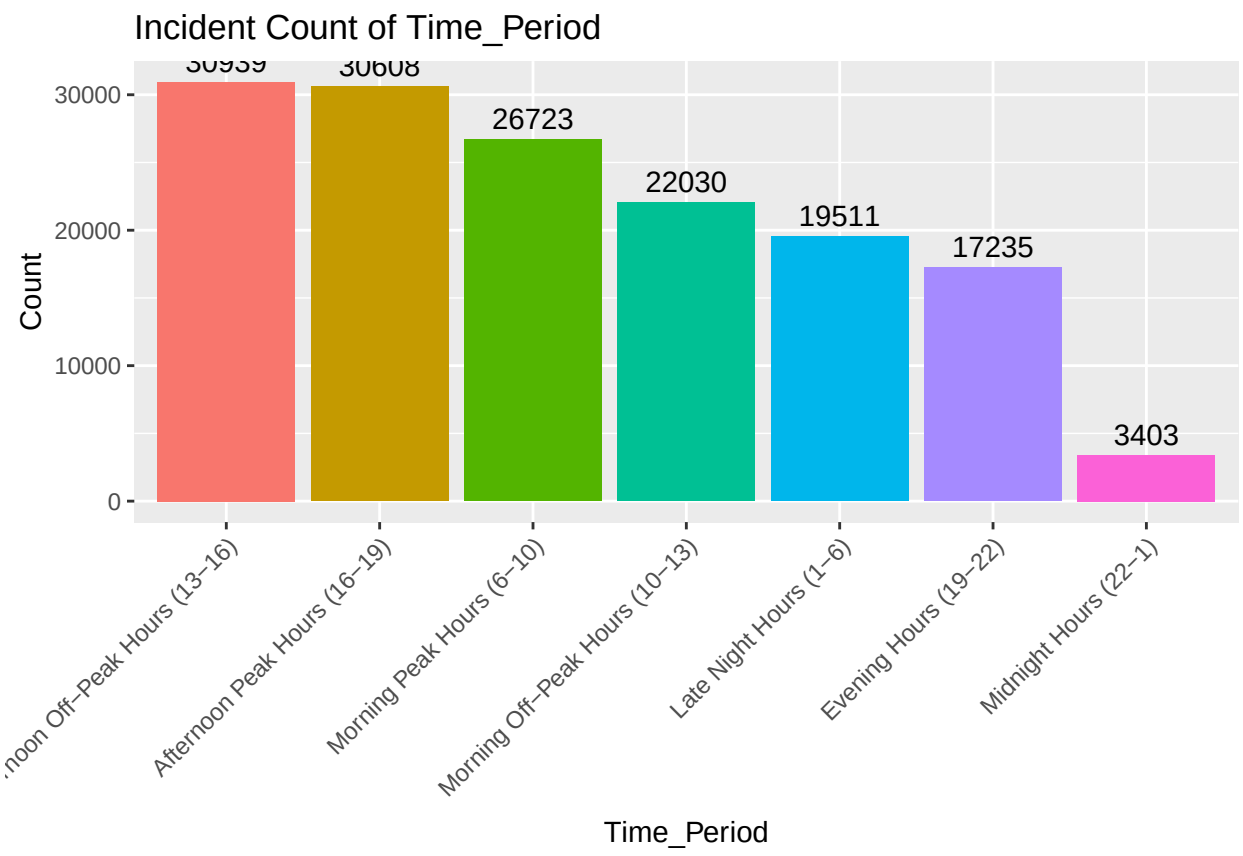


GROUPING 'TIME' VARIABLE ACROSS 4 TIME-PERIODS:

```
library(dplyr)

data2_filled_updated2 <- data2_filled_updated2 %>%
  mutate(Hour = as.integer(substr(Time, 1, 2)),
         Minute = as.integer(substr(Time, 4, 5)),
         Time_Period = case_when(
           (Hour == 6 & Minute >= 0) | (Hour > 6 & Hour < 10) ~ "Morning Peak Hours (6-10)",
           (Hour == 10 & Minute >= 0) | (Hour > 10 & Hour < 13) ~ "Morning Off-Peak Hours (10-13)",
           (Hour == 13 & Minute >= 0) | (Hour > 13 & Hour < 16) ~ "Afternoon Off-Peak Hours (13-16)",
           (Hour == 16 & Minute >= 0) | (Hour > 16 & Hour < 19) ~ "Afternoon Peak Hours (16-19)",
           (Hour == 19 & Minute >= 0) | (Hour > 19 & Hour < 22) ~ "Evening Hours (19-22)",
           (Hour == 23 & Minute >= 0) | (Hour > 23 & Hour < 1) ~ "Midnight Hours (22-1)",
           TRUE ~ "Late Night Hours (1-6)"
         ))
data2_filled_updated2 <- data2_filled_updated2 %>%
  select(-c(Hour, Minute))  # Remove the "Hour" and "Minute" columns

library(ggplot2)

# Reorder the levels of Time_Period based on count of incidents
data2_filled_updated2$Time_Period <- factor(data2_filled_updated2$Time_Period,
```

```
                                                levels = names(sort(table(data2_filled_updated2$Time_Period
```

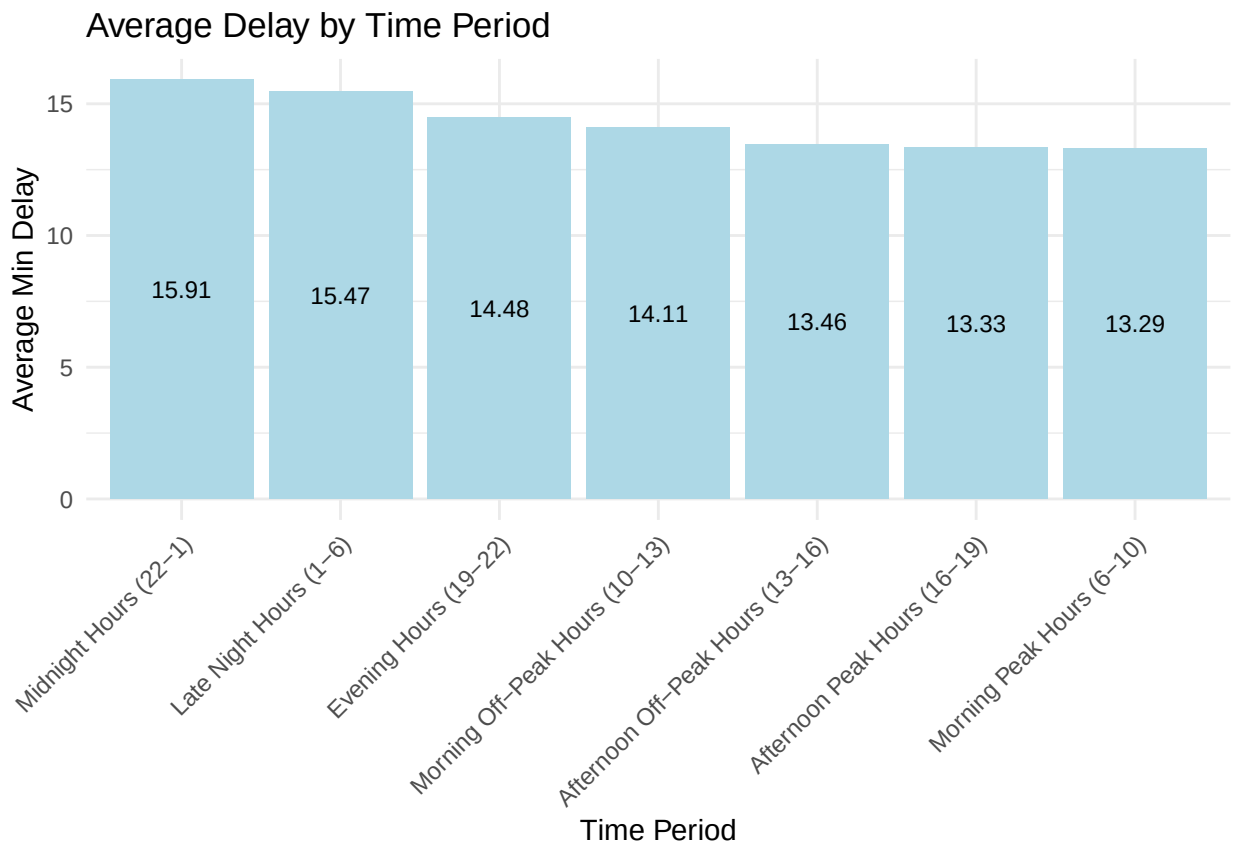
```
# Plotting with ggplot
ggplot(data2_filled_updated2, aes(x = Time_Period, fill = Time_Period)) +
  geom_bar() +
  geom_text(stat = 'count', aes(label = ..count..), vjust = -0.5) +
  scale_fill_discrete(name = "Time_Period") +
  labs(x = "Time_Period", y = "Count", title = "Incident Count of Time_Period") +
  theme(legend.position = "none", axis.text.x = element_text(angle = 45, hjust = 1))
```



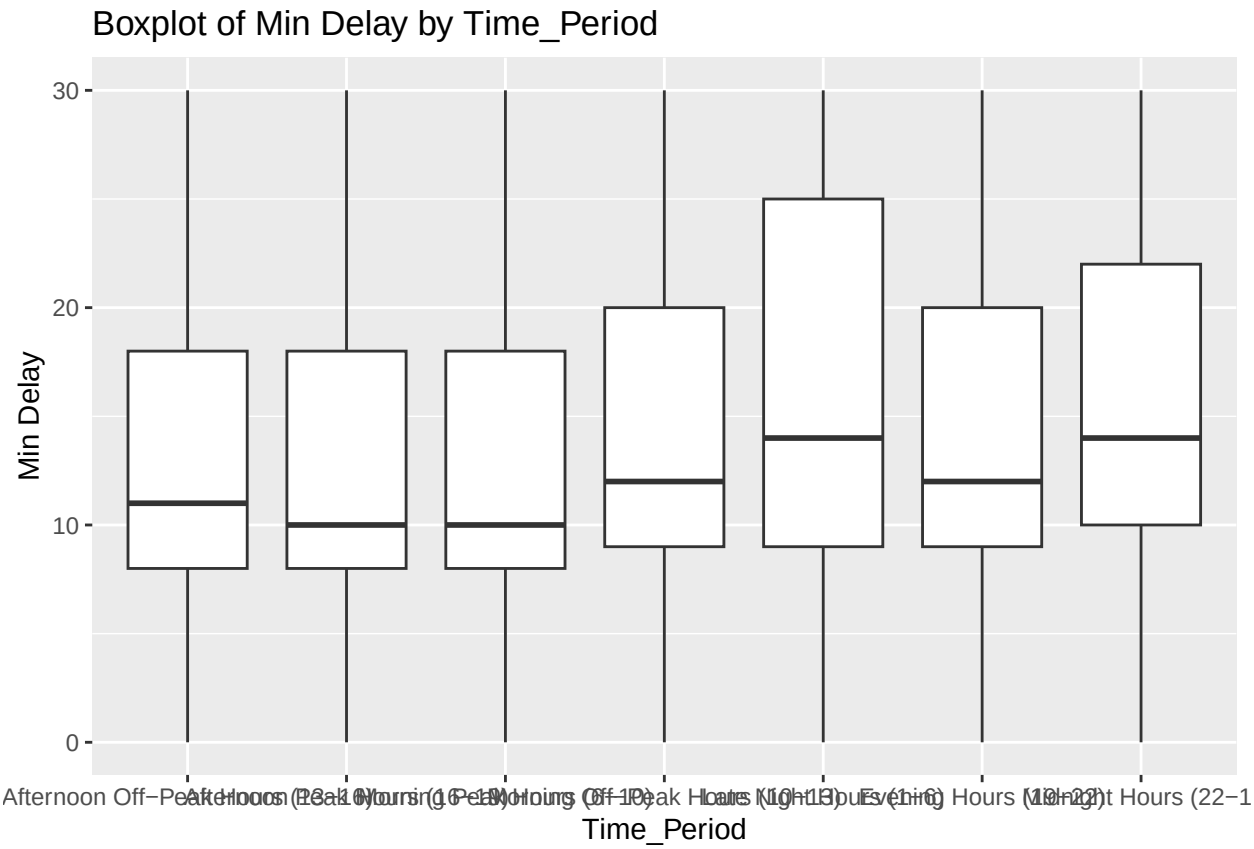

```
data2_filled_updated2$Time_Period <- as.factor(data2_filled_updated2$Time_Period)
```

```
# Aggregate data by time period and calculate the mean Min Delay for each time period
data_by_time_period <- data2_filled_updated2 %>%
  group_by(Time_Period) %>%
  summarise(Avg_Min_Delay = mean(`Min_Delay_Winsorized`))

# Reorder levels of Time_Period according to Avg_Min_Delay in descending order
data_by_time_period <- data_by_time_period %>%
  arrange(desc(Avg_Min_Delay)) %>%
  mutate(Time_Period = factor(Time_Period, levels = Time_Period))

# Create bar plot
ggplot(data_by_time_period, aes(x = Time_Period, y = Avg_Min_Delay)) +
  geom_bar(stat = "identity", fill = "lightblue") +  # Bar plot with different color
```

```
  geom_text(aes(label = round(Avg_Min_Delay, 2)), position = position_stack(vjust = 0.5), color = "blac
  labs(title = "Average Delay by Time Period",
       x = "Time Period",
       y = "Average Min Delay") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate x-axis labels for better readabili
```




```
# Create a customized boxplot to clearly see the interquartile range
ggplot(data2_filled_updated2, aes(x = Time_Period, y = `Min_Delay_Winsorized`)) +
  geom_boxplot(coef = 1.5) +  # Adjust the coef parameter to control the length of the whiskers
  labs(x = "Time_Period", y = "Min Delay", title = "Boxplot of Min Delay by Time_Period")
```



MONTH-WISE DELAY INCIDENTS:

```
library(dplyr)
library(ggplot2)
library(lubridate)
```

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

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

```
# Extract the month name from the Date column
data2_filled_updated2$Month_Name <- month(data2_filled_updated2$Date, label = TRUE, abbr = FALSE)

# Aggregate count by month name
incidents_by_month_name <- data2_filled_updated2 %>%
  group_by(Month_Name) %>%
  summarise(Total_Incidents = n()) %>%
  mutate(Month_Name = reorder(Month_Name, -Total_Incidents)) # Reorder based on total incidents in desc
```

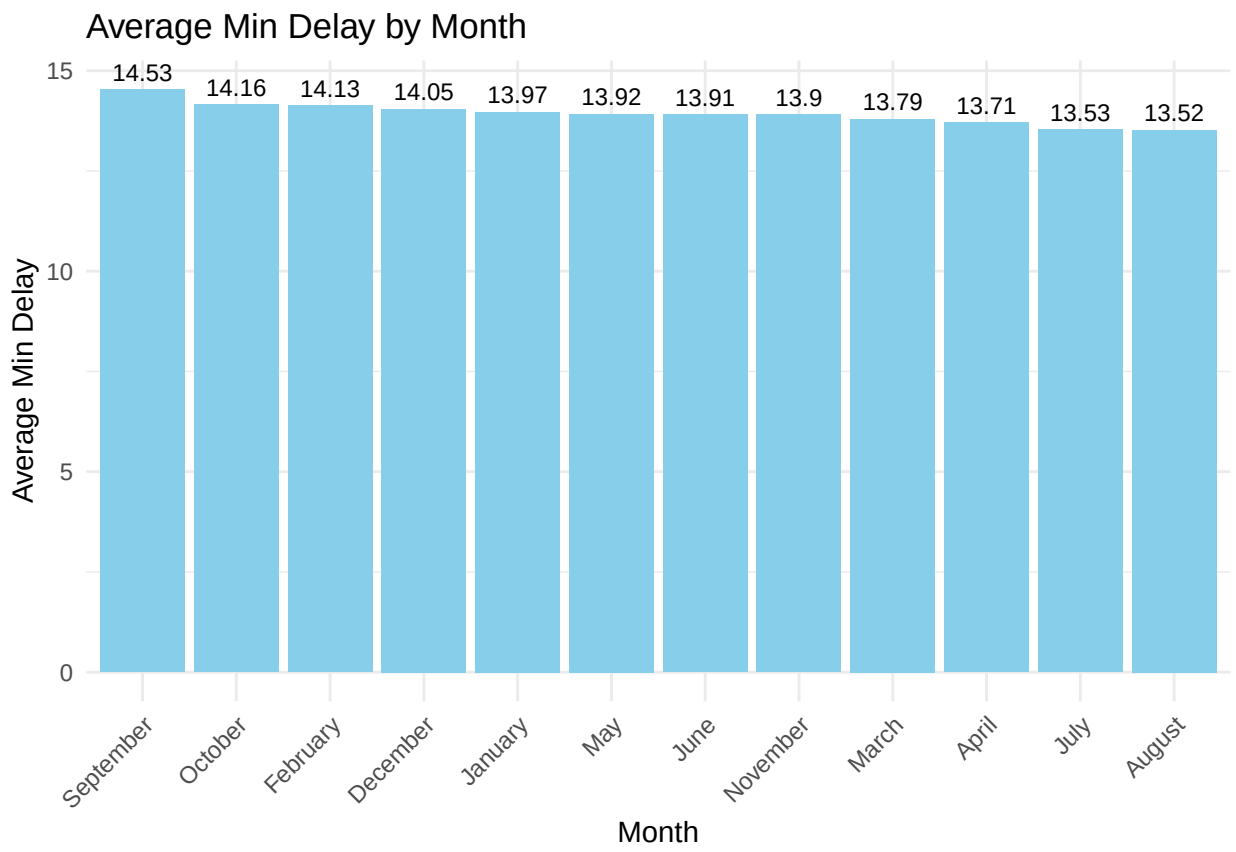
```r
# Extract month from the Date column and count incidents for each month
month_wise_incident_counts <- data2_filled_updated2 %>%
  mutate(Month = month(Date, label = TRUE)) %>%
  count(Month) %>%
  mutate(Percentage = n / sum(n) * 100) %>%
  arrange(Month)

# Print the month-wise counts and percentages
print(month_wise_incident_counts)
```

```
## # A tibble: 12 x 3
##    Month     n Percentage
##    <ord> <int>      <dbl>
##  1 Jan   12503       8.31
##  2 Feb   10885       7.24
##  3 Mar   12976       8.62
##  4 Apr   12355       8.21
##  5 May   13035       8.66
##  6 Jun   13847       9.20
##  7 Jul   12219       8.12
##  8 Aug   12418       8.25
##  9 Sep   13745       9.14
## 10 Oct   13851       9.21
## 11 Nov   12787       8.50
## 12 Dec    9828       6.53
```

AVERAGE 'MIN DELAY' ACROSS MONTHS:

```r
# MONTHWISE AVERAGE DELAY
# Load required libraries
library(dplyr)
library(ggplot2)

# Convert Date column to year-month format
data2_filled_updated2$Month <- format(data2_filled_updated2$Date, "%m")

# Aggregate data by month and calculate the mean Min Delay for each month
data_by_month <- data2_filled_updated2 %>%
  group_by(Month) %>%
  summarise(Avg_Min_Delay = mean(`Min_Delay_Winsorized`))

# Convert month numbers to month names
data_by_month$Month <- factor(month.name[as.numeric(data_by_month$Month)], levels = month.name)

# Reorder levels of Month according to Avg_Min_Delay in descending order
data_by_month <- data_by_month %>%
  arrange(desc(Avg_Min_Delay)) %>%
  mutate(Month = factor(Month, levels = Month))

# Create bar plot
```

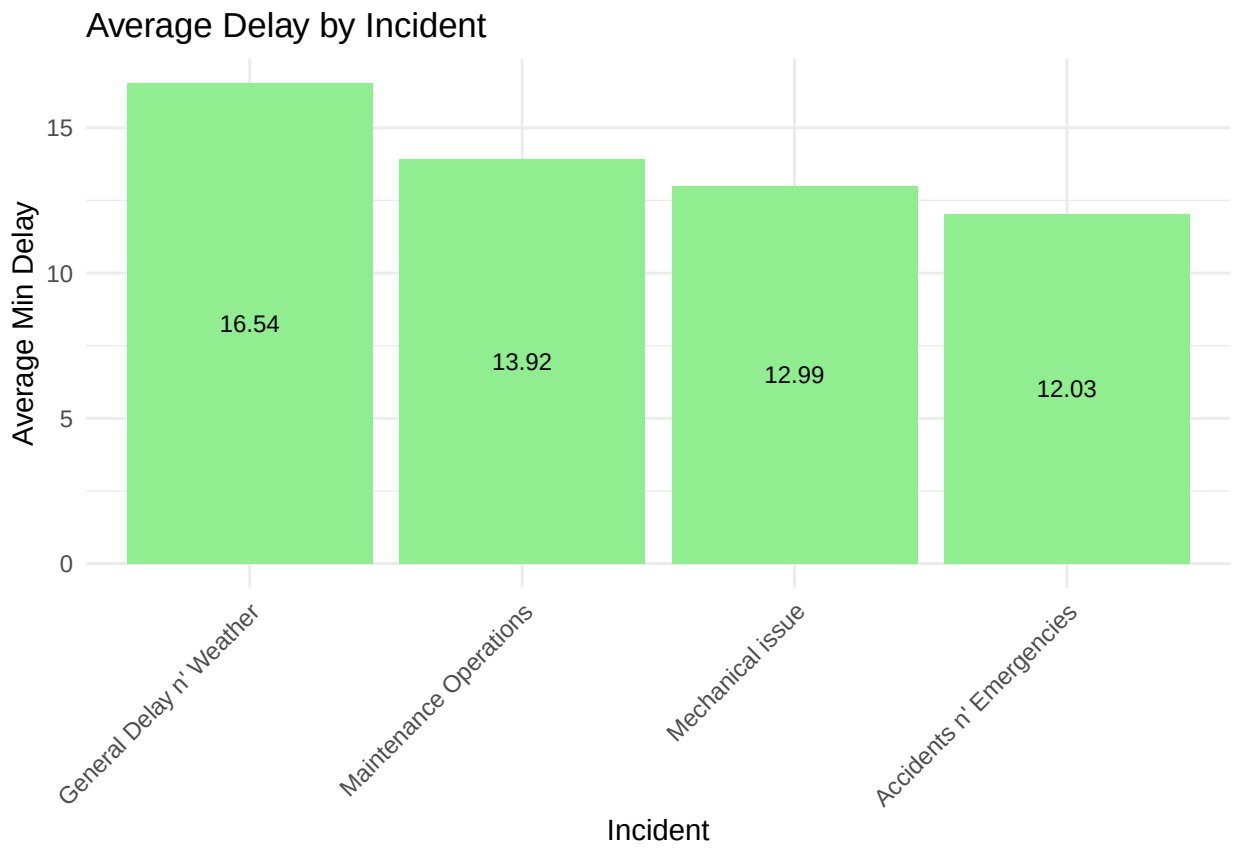
```r
ggplot(data_by_month, aes(x = Month, y = Avg_Min_Delay)) +
  geom_bar(stat = "identity", fill = "skyblue") +  # Bar plot
  geom_text(aes(label = round(Avg_Min_Delay, 2)), vjust = -0.5, color = "black", size = 3) +  # Add dat
  labs(title = "Average Min Delay by Month",
       x = "Month",
       y = "Average Min Delay") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate x-axis labels for better readabili
```



AVERAGE DELAY BY INCIDENT:

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

# Aggregate data by incident and calculate the mean Min Delay for each incident
data_by_incident <- data2_filled_updated2 %>%
  group_by(Incident) %>%
  summarise(Avg_Min_Delay = mean(`Min_Delay_Winsorized`))

# Reorder levels of Incident according to Avg_Min_Delay in descending order
data_by_incident <- data_by_incident %>%
  arrange(desc(Avg_Min_Delay)) %>%
  mutate(Incident = factor(Incident, levels = Incident))

# Create bar plot
ggplot(data_by_incident, aes(x = Incident, y = Avg_Min_Delay)) +
```

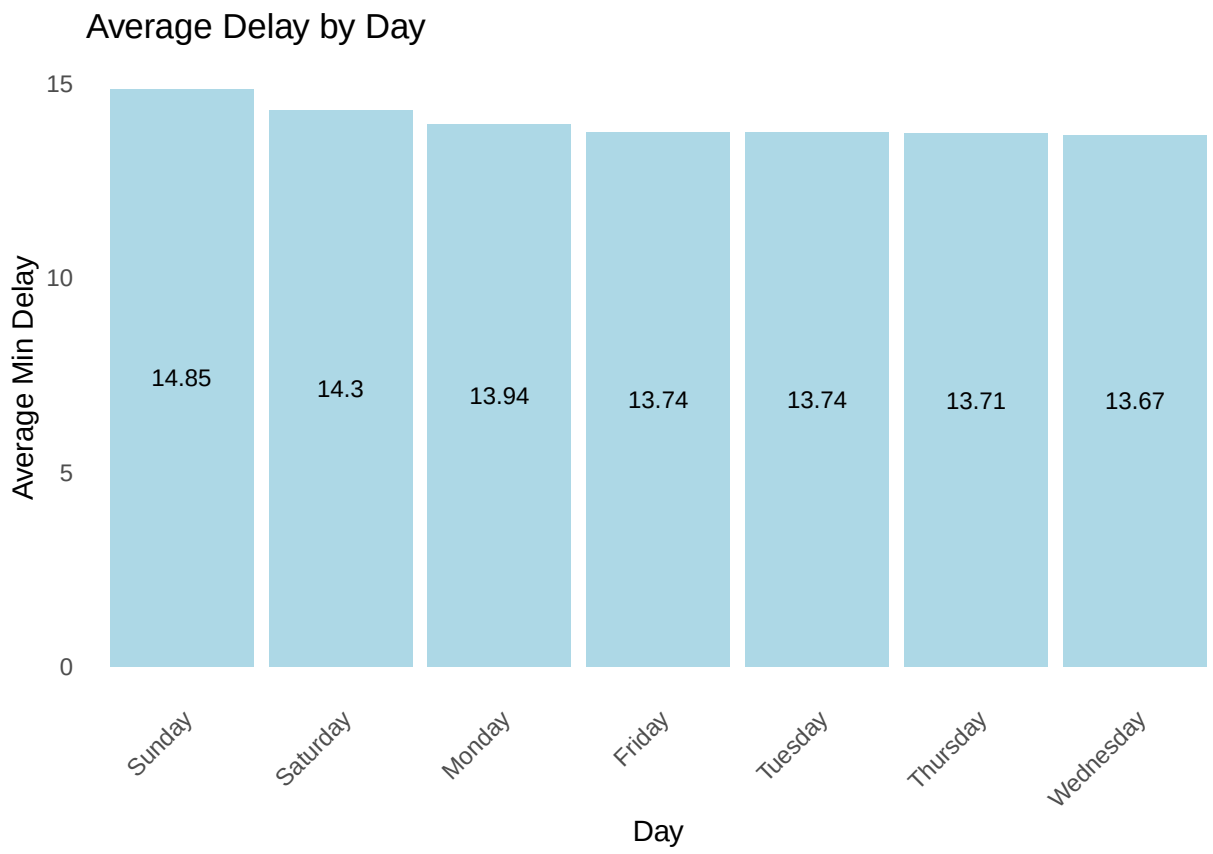
```
  geom_bar(stat = "identity", fill = "lightgreen") +  # Bar plot with different color
  geom_text(aes(label = round(Avg_Min_Delay, 2)), position = position_stack(vjust = 0.5), color = "blac
  labs(title = "Average Delay by Incident",
       x = "Incident",
       y = "Average Min Delay") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate x-axis labels for better readabili
```



AVERAGE DELAY BY DAY

```
# Aggregate data by day and calculate the mean Min Delay for each day
data_by_day <- data2_filled_updated2 %>%
  group_by(Day) %>%
  summarise(Avg_Min_Delay = mean(`Min_Delay_Winsorized`))

# Reorder levels of Day according to Avg_Min_Delay in descending order
data_by_day <- data_by_day %>%
  arrange(desc(Avg_Min_Delay)) %>%
  mutate(Day = factor(Day, levels = Day))

# Create bar plot
ggplot(data_by_day, aes(x = Day, y = Avg_Min_Delay)) +
  geom_bar(stat = "identity", fill = "lightblue") +  # Bar plot with different color
  geom_text(aes(label = round(Avg_Min_Delay, 2)), position = position_stack(vjust = 0.5), color = "blac
  labs(title = "Average Delay by Day",
       x = "Day",
```

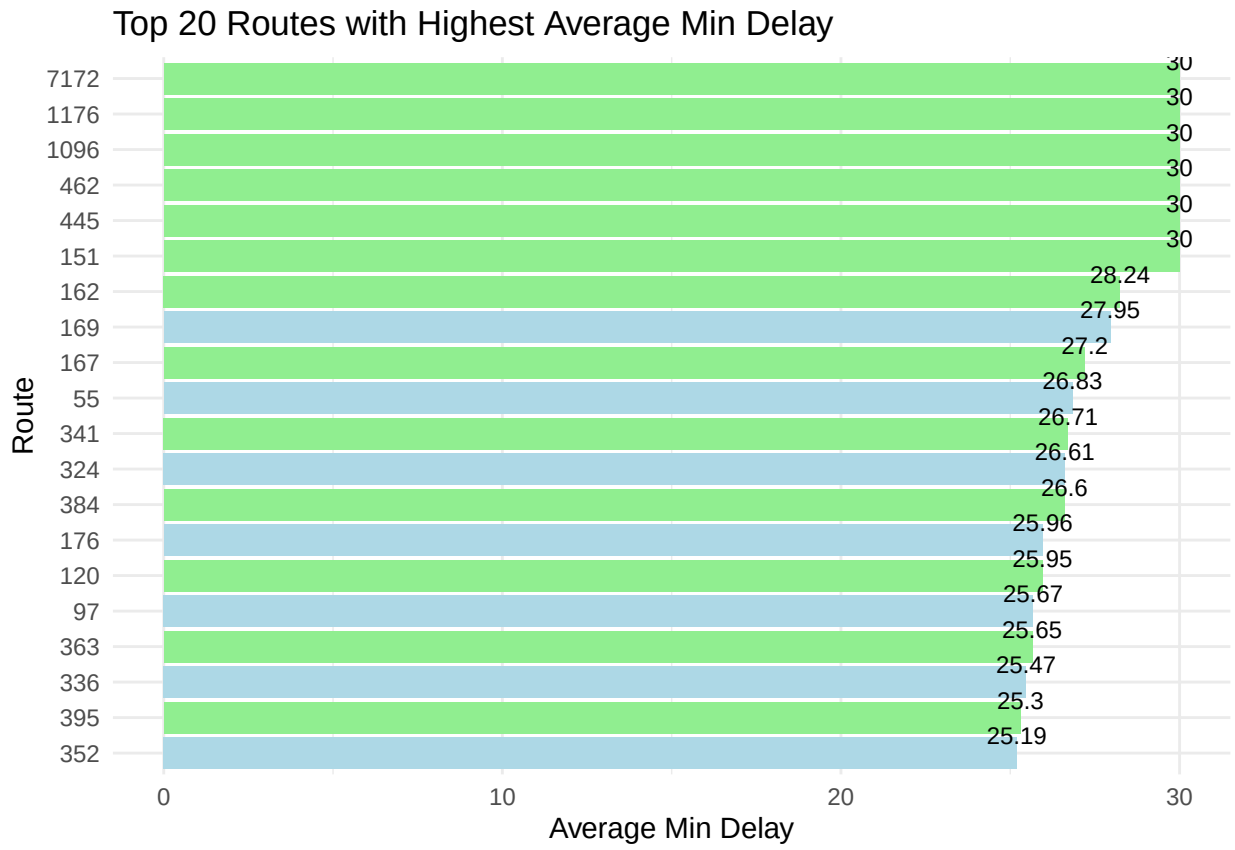
```
      y = "Average Min Delay") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1), panel.grid.major = element_blank(), panel.gri
```



TOP 20 ROUTES WITH HIGHEST DELAY INCIDENT

```
# Aggregate data by route and calculate the mean Min Delay for each route
data_by_route <- data2_filled_updated2 %>%
  group_by(Route) %>%
  summarise(Avg_Min_Delay = mean(`Min_Delay_Winsorized`))

# Reorder levels of Route according to Avg_Min_Delay in descending order
data_by_route <- data_by_route %>%
  arrange(desc(Avg_Min_Delay)) %>%
  mutate(Route = factor(Route, levels = Route))

# Select the top 25 routes with the highest average minimum delay
top_routes <- head(data_by_route, 20)

# Create bar plot for the top 20 routes
ggplot(top_routes, aes(x = reorder(Route, Avg_Min_Delay), y = Avg_Min_Delay)) +
  geom_bar(stat = "identity", fill = ifelse(rank(-top_routes$Avg_Min_Delay) %% 2 == 0, "lightblue", "lig
```

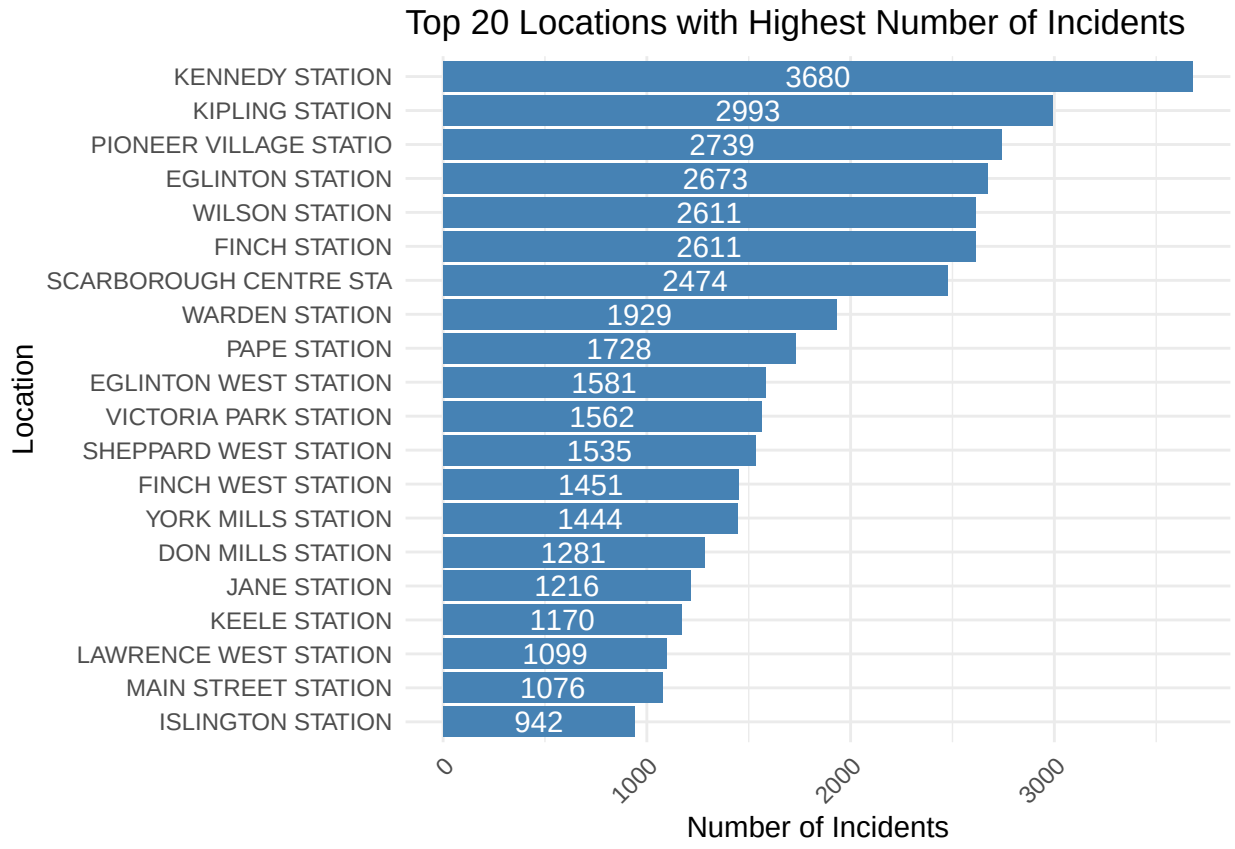
```
  geom_text(aes(label = round(Avg_Min_Delay, 2)), position = position_dodge(width = 0.9), vjust = -0.5,
  coord_flip() +  # Flip coordinates for horizontal bar chart
  labs(title = "Top 20 Routes with Highest Average Min Delay",
       x = "Route",
       y = "Average Min Delay") +
  theme_minimal()
```



TOP 20 LOCATION WITH HIGHEST DELAY INCIDENCE

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

# Aggregate data to count the number of incidents per location
incident_counts <- data2_filled_updated2 %>%
  group_by(Location) %>%
  summarise(Incidents = n()) %>%
  arrange(desc(Incidents)) %>%
  top_n(20, Incidents)

# Plot the top 20 locations with the highest number of incidents, including data labels
ggplot(incident_counts, aes(x = reorder(Location, Incidents), y = Incidents)) +
```

```
  geom_bar(stat = "identity", fill = "steelblue") +
  geom_text(aes(label = Incidents), position = position_stack(vjust = 0.5), color = "white") +
  coord_flip() + # Flip the coordinates to make it a horizontal bar plot
  labs(title = "Top 20 Locations with Highest Number of Incidents",
       x = "Location",
       y = "Number of Incidents") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) # Improve label readability
```




```
# Display the plot
ggsave("Top20_Locations_Incidents_with_Labels.png", width = 10, height = 8, dpi = 300)
```

WINSORIZATION OF 'MIN GAP' VARIABLE

```
# Perform 'Winsorization' to Fix Outlier Issues: capping the outliers at a certain percentile. For exam
library(DescTools)

# Winsorize the 'Min Delay' column at the 5th and 95th percentiles for the entire data
data2_filled_updated2$Min_Gap_Winsorized <- Winsorize(data2_filled_updated2$`Min Gap`, probs = c(0.02, 

# Check the results
summary(data2_filled_updated2$Min_Gap_Winsorized)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   17.00   22.00   27.16   38.00   60.00
```

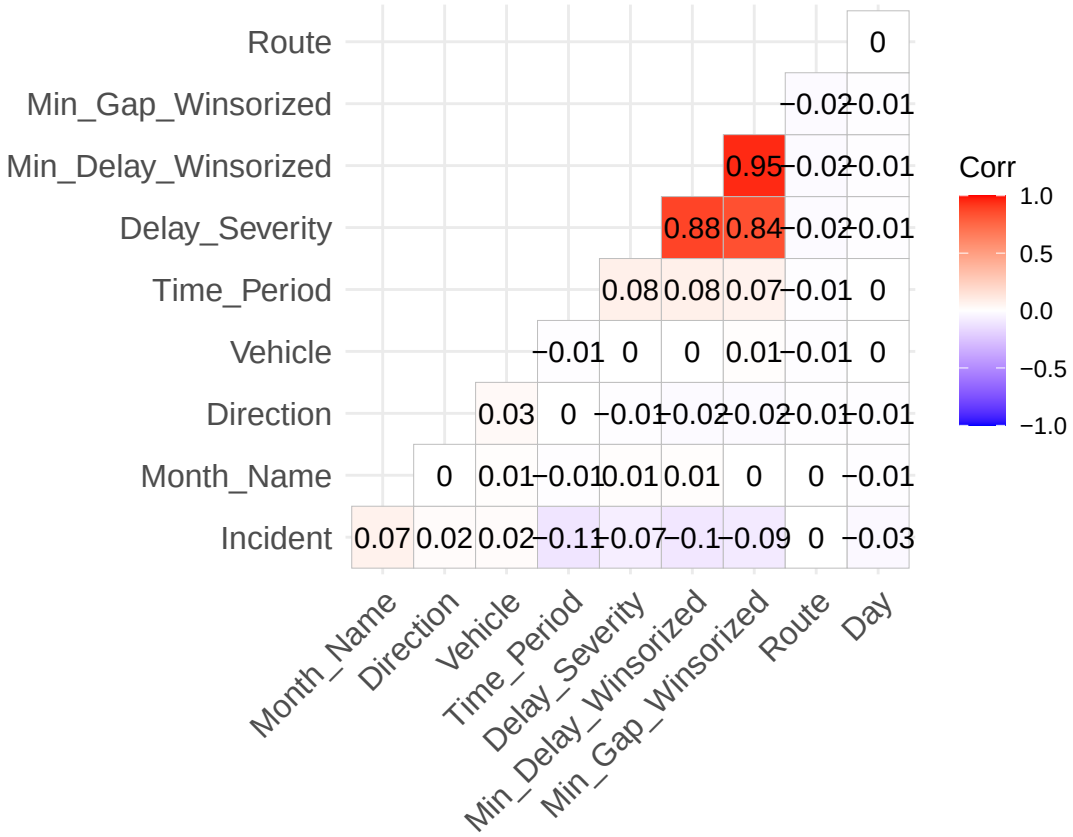
```
numeric_data1 <- data2_filled_updated2
str(numeric_data1)
```

```
## tibble [150,449 x 17] (S3: tbl_df/tbl/data.frame)
##  $ Date                   : POSIXct[1:150449], format: "2023-01-01" "2023-01-01" ...
##  $ Mode of Transportation: chr [1:150449] "Bus" "Bus" "Bus" "Bus" ...
##  $ Route                  : num [1:150449] 91 69 35 900 85 40 336 52 24 36 ...
##  $ Time                   : chr [1:150449] "02:30" "02:34" "03:06" "03:14" ...
##  $ Day                    : Factor w/ 7 levels "Friday","Monday",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ Location               : chr [1:150449] "WOODBINE AND MORTIMER" "WARDEN STATION" "JANE STATION" "K
##  $ Incident               : Factor w/ 4 levels "General Delay n' Weather",..: 1 2 1 2 2 2 1 2 1 1 ...
##  $ Min Delay              : num [1:150449] 81 22 30 17 1 0 138 30 20 334 ...
##  $ Min Gap                : num [1:150449] 111 44 60 17 1 0 168 60 40 344 ...
##  $ Direction              : int [1:150449] 5 4 3 3 3 5 3 2 5 5 ...
##  $ Vehicle                : num [1:150449] 8772 8407 1051 3334 1559 ...
##  $ Min_Delay_Winsorized   : num [1:150449] 30 22 30 17 1 0 30 30 20 30 ...
##  $ Delay_Severity         : Factor w/ 3 levels "Borderline Late (<10 Min)",..: 3 3 3 3 1 1 3 3 3 3 ..
##  $ Time_Period            : Factor w/ 7 levels "Afternoon Off-Peak Hours (13-16)",..: 5 5 5 5 5 5 5 5
##  $ Month_Name             : Ord.factor w/ 12 levels "January"<"February"<..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Month                  : chr [1:150449] "01" "01" "01" "01" ...
##  $ Min_Gap_Winsorized     : num [1:150449] 60 44 60 17 1 0 60 60 40 60 ...
```

CORRELATION ANALYSIS OF VARIABLES

```
library(ggcorrplot)
library(dplyr)
numeric_data1$Time_Period<-as.numeric(numeric_data1$Time_Period)
numeric_data1$Delay_Severity<-as.numeric(numeric_data1$Delay_Severity)
numeric_data1$Incident<-as.numeric(numeric_data1$Incident)
numeric_data1$Day<-as.numeric(numeric_data1$Day)
numeric_data1$Month_Name<-as.numeric(numeric_data1$Month_Name)
numeric_data1$Direction<-as.numeric(numeric_data1$Direction)

numeric_data1 <- numeric_data1 %>%
  select(-`Min Delay`, -`Min Gap`)
```

```
# Select numeric variables
numeric_data <- numeric_data1 %>%
  select_if(is.numeric)

# Calculate the correlation matrix
correlation_matrix <- cor(numeric_data)

# Plot the correlation matrix using ggcorrplot
ggcorrplot(correlation_matrix, hc.order = TRUE, type = "lower", lab = TRUE)
```




```
summary(lm (Min_Delay_Winsorized ~ Incident + Time_Period + Route + Day + Direction , data = data2_fille
```

```
##
## Call:
## lm(formula = Min_Delay_Winsorized ~ Incident + Time_Period +
##     Route + Day + Direction, data = data2_filled_updated2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.522  -5.304  -2.106   5.421  54.469
##
## Coefficients:
##                                            Estimate Std. Error t value
## (Intercept)                                1.628e+01  9.711e-02 167.664
## IncidentAccidents n' Emergencies          -4.666e+00  6.225e-02 -74.960
## IncidentMechanical issue                  -3.418e+00  5.613e-02 -60.898
## IncidentMaintenance Operations            -2.493e+00  5.790e-02 -43.059
## Time_PeriodAfternoon Peak Hours (16-19)   -1.124e-01  6.422e-02  -1.750
## Time_PeriodMorning Peak Hours (6-10)      -5.391e-02  6.678e-02  -0.807
## Time_PeriodMorning Off-Peak Hours (10-13)  4.848e-01  7.043e-02   6.884
## Time_PeriodLate Night Hours (1-6)          2.192e+00  7.327e-02  29.917
## Time_PeriodEvening Hours (19-22)           8.857e-01  7.596e-02  11.661
## Time_PeriodMidnight Hours (22-1)           2.329e+00  1.443e-01  16.141
## Route                                     -7.167e-05  8.793e-06  -8.151
## DayMonday                                  1.629e-01  7.484e-02   2.176
```

```
## DaySaturday                                  5.022e-01  7.676e-02   6.543
## DaySunday                                    1.029e+00  8.322e-02  12.369
## DayThursday                                 -8.497e-02  7.204e-02  -1.179
## DayTuesday                                  -7.248e-02  7.315e-02  -0.991
## DayWednesday                                -1.617e-01  7.192e-02  -2.248
## Direction                                   -1.114e-01  1.860e-02  -5.992
##                                             Pr(>|t|)
## (Intercept)                                  < 2e-16 ***
## IncidentAccidents n' Emergencies             < 2e-16 ***
## IncidentMechanical issue                     < 2e-16 ***
## IncidentMaintenance Operations               < 2e-16 ***
## Time_PeriodAfternoon Peak Hours (16-19)       0.0801 .
## Time_PeriodMorning Peak Hours (6-10)          0.4195
## Time_PeriodMorning Off-Peak Hours (10-13) 5.83e-12 ***
## Time_PeriodLate Night Hours (1-6)            < 2e-16 ***
## Time_PeriodEvening Hours (19-22)             < 2e-16 ***
## Time_PeriodMidnight Hours (22-1)             < 2e-16 ***
## Route                                     3.64e-16 ***
## DayMonday                                    0.0295 *
## DaySaturday                               6.07e-11 ***
## DaySunday                                    < 2e-16 ***
## DayThursday                                  0.2382
## DayTuesday                                   0.3218
## DayWednesday                                 0.0246 *
## Direction                                 2.08e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.965 on 150431 degrees of freedom
## Multiple R-squared:  0.05216,    Adjusted R-squared:  0.05205
## F-statistic: 486.9 on 17 and 150431 DF,  p-value: < 2.2e-16
```

```r
Cleaned_Data<-data2_filled_updated2
Cleaned_Data$Time_Period<- as.integer(Cleaned_Data$Time_Period)
Cleaned_Data$Incident<- as.numeric(Cleaned_Data$Incident)
Cleaned_Data$Day<- as.numeric(Cleaned_Data$Day)
str(Cleaned_Data)
```

```
## tibble [150,449 x 17] (S3: tbl_df/tbl/data.frame)
##  $ Date                  : POSIXct[1:150449], format: "2023-01-01" "2023-01-01" ...
##  $ Mode of Transportation: chr [1:150449] "Bus" "Bus" "Bus" "Bus" ...
##  $ Route                 : num [1:150449] 91 69 35 900 85 40 336 52 24 36 ...
##  $ Time                  : chr [1:150449] "02:30" "02:34" "03:06" "03:14" ...
##  $ Day                   : num [1:150449] 4 4 4 4 4 4 4 4 4 4 ...
##  $ Location              : chr [1:150449] "WOODBINE AND MORTIMER" "WARDEN STATION" "JANE STATION" "K
##  $ Incident              : num [1:150449] 1 2 1 2 2 2 1 2 1 1 ...
##  $ Min Delay             : num [1:150449] 81 22 30 17 1 0 138 30 20 334 ...
##  $ Min Gap               : num [1:150449] 111 44 60 17 1 0 168 60 40 344 ...
##  $ Direction             : int [1:150449] 5 4 3 3 3 5 3 2 5 5 ...
##  $ Vehicle               : num [1:150449] 8772 8407 1051 3334 1559 ...
##  $ Min_Delay_Winsorized  : num [1:150449] 30 22 30 17 1 0 30 30 20 30 ...
##  $ Delay_Severity        : Factor w/ 3 levels "Borderline Late (<10 Min)",..: 3 3 3 3 1 1 3 3 3 3 ..
##  $ Time_Period           : int [1:150449] 5 5 5 5 5 5 5 5 5 5 ...
##  $ Month_Name            : Ord.factor w/ 12 levels "January"<"February"<..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
##  $ Month                 : chr [1:150449] "01" "01" "01" "01" ...
##  $ Min_Gap_Winsorized    : num [1:150449] 60 44 60 17 1 0 60 60 40 60 ...
```

```
library(class)
library(gmodels)
```

```
## Warning: package 'gmodels' was built under R version 4.3.1
```

```
## Registered S3 method overwritten by 'gdata':
##   method         from
##   reorder.factor DescTools
```

```
round(prop.table(table(Cleaned_Data$Incident))*100,digits = 1)
```

```
##
##    1    2    3    4
## 24.9 19.6 29.3 26.1
```

```
normalize<-function(x){return((x-min(x))/(max(x)-min(x)))}
```

```
normalized<-as.data.frame(lapply(Cleaned_Data[c(3,5,7,10,11,14)],normalize))
```

```
Preprocessed_dataset<-cbind(Cleaned_Data$Delay_Severity,normalized)
str(Preprocessed_dataset)
```

```
## 'data.frame':    150449 obs. of  7 variables:
##  $ Cleaned_Data$Delay_Severity: Factor w/ 3 levels "Borderline Late (<10 Min)",..: 3 3 3 3 1 1 3 3 3
##  $ Route                      : num  1.00e-04 7.57e-05 3.78e-05 1.00e-03 9.35e-05 ...
##  $ Day                        : num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
##  $ Incident                   : num  0 0.333 0 0.333 0.333 ...
##  $ Direction                  : num  1 0.75 0.5 0.5 0.5 1 0.5 0.25 1 1 ...
##  $ Vehicle                    : num  0.0886 0.0849 0.0106 0.0337 0.0157 ...
##  $ Time_Period                : num  0.667 0.667 0.667 0.667 0.667 ...
```