SDS 322E Project - EDA Report

10-21-2024

Description - EDA Report 2

Title and Introduction

We would like to know how construction affects traffic incident rates. We use construction permits as a proxy for the actual number of construction sites in a zip code. Our motivation is to show high risk construction projects, which may deserve a greater degree of road safety preparation. This study (https://www.sciencedirect.com/science/article/pii/S235214652100819X), by Mangones, et al. (2021), has shown that construction projects can increase traffic incidents when there is use of excavation of more than half a meter. We wish to localize this experiment to Austin and understand how the different types of construction projects affect traffic incident rates.

Name - Issued Construction Permits (https://data.austintexas.gov/Building-and-Development/Issued-Construction-Permits/3syk-w9eu/about_data)

Description - This data set contains “Building, Electrical, Mechanical, and Plumbing Permits and Driveway/ Sidewalk Permits issued by the City of Austin. Includes relevant details such as issue date, location, council district, expiration date, description of work, square footage, valuation, and units.” (City of Austin open data portal)

Rows & Columns - There are 897369 rows and 68 columns.

Unique Rows - A single row represents a construction permit.

Main Variables of Interest - Construction Permits

- **Permit.Type.Desc** -> Description of the Permit Type
- **Permit.Class** -> “Sub Type of the permit”, it will be re-categorized as Residential or Commercial permits.
- **Issued.Date** -> Date on which the permit was issued
- **Status.Current** -> Current status of permit
- **Number.Of.Floors** -> How many floors property has
- **Original.Zip** -> Zip code of the property associated with the permit

Dataset 2

Name - Real-Time Incidents (https://data.austintexas.gov/Transportation-and-Mobility/Real-Time-Traffic-Incident-Reports/dx9v-zd7x/about_data)

Description - “This data set contains various traffic incidents from the Austin-Travis County traffic reports collected from the various Public Safety agencies through a data feed from the Combined Transportation, Emergency, and Communications Center (CTECC).” (City of Austin open data portal)

Rows & Columns - There are 356150 rows and 22 columns.

Unique Rows - A single row represents a traffic incident.

Main Variables of Interest - Traffic Incidents:

- **Published_Date** -> The date the report was published
- **Issue_Reported** -> The reported issue, based on the selection by reporting agency
- **Zip code** -> Processed from lat/lon using ArcGIS reverse-geocoding

Expecations

Trends and Relationships

We expect that Commercial Projects cause an increase in Traffic Incident rates. We expect that there will be more traffic incidents during the months of mid spring when there is more rain. We also expect that there will be a difference in Traffic Incidents rates based upon the proximity of zip codes to the Austin city center.

Research Question - Dan

Is there a difference in impact of Residential or Commercial construction projects on Traffic Incidents? We will eventually determine this causal relationship through the rejection of the null hypothesis which states that the impact of Commercial Permits and Residential Permits on Traffic Incidents are equivalent.

Research Question - Tigris

How do spatial concentrations of Traffic Incidents change throughout the year?

Install & Call Libraries

```
#install.packages("ggmap")
#install.packages
#install.packages("tidygeocoder")
#install.packages("ggplot2")
#install.packages("sf")
#install.packages("tigris")
library("sf")
```

```
## Linking to GEOS 3.12.1, GDAL 3.8.4, PROJ 9.3.1; sf_use_s2() is TRUE
```

```
library("tigris")
```

```
## To enable caching of data, set `options(tigris_use_cache = TRUE)`
## in your R script or .Rprofile.
```

```
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
```

```
library("ggplot2")
library("ggmap")
```

```
## ℹ Google's Terms of Service: <https://mapsplatform.google.com>
##   Stadia Maps' Terms of Service: <https://stadiamaps.com/terms-of-service/>
##   OpenStreetMap's Tile Usage Policy: <https://operations.osmfoundation.org/policies/tiles/
>
## ℹ Please cite ggmap if you use it! Use `citation("ggmap")` for details.
```

```
library("tidyverse")
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ lubridate 1.9.3     ✓ tibble    3.2.1
## ✓ purrr     1.0.2     ✓ tidyr     1.3.1
## ✓ readr     2.1.5
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to be
come errors
```

```
library("zipcodeR")
library("tidygeocoder")
```

```
##
## Attaching package: 'tidygeocoder'
##
## The following object is masked from 'package:ggmap':
##
##     geocode
```

Methods

Inputting Data

```
# This code reads the Construction Permit data set into a R dataframe
construction_permits <- read.csv("~/My Documents/UT Course Files/UT Fall 2024/SDS 322E/SDS Pr
oject Files/Original Data/Issued_Construction_Permits_20241006.csv", header=TRUE)

head(construction_permits)
```

```
##   Permit.Type  Permit.Type.Desc     Permit.Num Permit.Class.Mapped
## 1          PP   Plumbing Permit 2023-141107 PP         Residential
## 2          PP   Plumbing Permit 2023-141108 PP         Residential
## 3          EP Electrical Permit 2023-143763 EP         Residential
## 4          BP   Building Permit 2023-161233 BP         Residential
## 5          BP   Building Permit 2023-162899 BP         Residential
## 6          BP   Building Permit 2023-162900 BP         Residential
##                            Permit.Class           Work.Class Condominium
## 1           R- 101 Single Family Houses                  New          No
## 2            R- 102 Secondary Apartment                  New          No
## 3 R- 329 Res Structures Other Than Bldg                  New          No
## 4         R- 434 Addition & Alterations Addition and Remodel          No
## 5           R- 101 Single Family Houses                  New          No
## 6            R- 102 Secondary Apartment                  New          No
##                Project.Name
## 1    3009 GARWOOD ST BLDG 1
## 2    3009 GARWOOD ST BLDG 2
## 3        6217 CARRINGTON DR
## 4          2805 BRIDLE PATH
## 5 1601 CANTERBURY ST Bldg 1
## 6 1601 CANTERBURY ST Bldg 2
##
Description
## 1 New 3-story 3 bedroom 3.5 bathroom principal single family residence with attached carpo
rt covered front porch rear covered deck uncovered wood deck and uncovered 2nd floor balcony.
## 2                                                 New 2-story 2 bedroom 2.5 bathroom seco
ndary dwelling unit with attached garage covered front porch and 2nd floor uncovered balcony.
## 3
New pool and Spa
## 4                                                                                 Garag
e conversion and interior remodel of existing 2 story SFR. create mudroom office and bathroom
## 5                                                                          Expedited Revi
ew - New Construction of a 2-Story Single Family Res. [4bed 3bath] with attached 1-car Garage
## 6                                                                       Expedited Review -
New Construction of a 2-Story Secondary Apartment [2bed 2.5bath] with attached 1-car Garage.
##      TCAD.ID                                  Property.Legal.Description
## 1 0204130709 LOT 9 BLK 1 OLT 27 DIV A BRASS G M SUBD PLUS 1/2 ADJ VAC ALLEY
## 2 0204130709 LOT 9 BLK 1 OLT 27 DIV A BRASS G M SUBD PLUS 1/2 ADJ VAC ALLEY
## 3 0418401008                    LOT 65 BLK A CIRCLE C RANCH PHS B SEC 20-A
## 4 0115061003                                    LOT 50 WESTENFIELD NO 1
## 5 0202070201                           LOT 1 BLK 5 OLT 47 DIV O RIVERSIDE
## 6 0202070201                           LOT 1 BLK 5 OLT 47 DIV O RIVERSIDE
##   Applied.Date Issued.Date Day.Issued Calendar.Year.Issued Fiscal.Year.Issued
## 1     8/4/2023    1/1/2024     MONDAY                 2024                2024
## 2     8/4/2023    1/1/2024     MONDAY                 2024                2024
## 3   10/17/2023    1/1/2024     MONDAY                 2024                2024
## 4   11/29/2023    1/1/2024     MONDAY                 2024                2024
## 5    9/18/2023    1/1/2024     MONDAY                 2024                2024
## 6    9/18/2023    1/1/2024     MONDAY                 2024                2024
##   Issued.In.Last.30.Days Issuance.Method Status.Current Status.Date
## 1                     No   Permit Center          Final   7/29/2024
```

```
## 2                    No   Permit Center          Final    8/1/2024
## 3                    No   Permit Center          Final    3/1/2024
## 4                    No   Permit Center         Active   6/24/2024
## 5                    No   Permit Center         Active   8/29/2024
## 6                    No   Permit Center         Active   8/29/2024
##   Expires.Date Completed.Date Total.Existing.Bldg.SQFT Remodel.Repair.SQFT
## 1    7/29/2024      7/29/2024                      NA                  NA
## 2     8/1/2024       8/1/2024                      NA                  NA
## 3     3/1/2024       3/1/2024                      NA                  NA
## 4   12/21/2024                                     NA                 500
## 5    2/18/2025                                     NA                  NA
## 6    2/18/2025                                     NA                  NA
##   Total.New.Add.SQFT Total.Valuation.Remodel Total.Job.Valuation
## 1               2862                     NA                  NA
## 2               1322                     NA                  NA
## 3                638                     NA                  NA
## 4                362                      0                   1
## 5               2847                     NA                   0
## 6               1750                     NA                   0
##   Number.Of.Floors Housing.Units Building.Valuation Building.Valuation.Remodel
## 1                3             1                  NA                         NA
## 2                2             1                  NA                         NA
## 3                1             1                  NA                         NA
## 4                1             1                  NA                          0
## 5                2             1                  NA                         NA
## 6                2             1                  NA                         NA
##   Electrical.Valuation Electrical.Valuation.Remodel Mechanical.Valuation
## 1                   NA                           NA                   NA
## 2                   NA                           NA                   NA
## 3                   NA                           NA                   NA
## 4                   NA                            0                   NA
## 5                   NA                           NA                   NA
## 6                   NA                           NA                   NA
##   Mechanical.Valuation.Remodel Plumbing.Valuation Plumbing.Valuation.Remodel
## 1                           NA                 NA                         NA
## 2                           NA                 NA                         NA
## 3                           NA                 NA                         NA
## 4                            0                 NA                          0
## 5                           NA                 NA                         NA
## 6                           NA                 NA                         NA
##   MedGas.Valuation MedGas.Valuation.Remodel        Original.Address.1
## 1               NA                       NA    3009 GARWOOD ST BLDG 1
## 2               NA                       NA    3009 GARWOOD ST BLDG 2
## 3               NA                       NA        6217 CARRINGTON DR
## 4               NA                       NA          2805 BRIDLE PATH
## 5               NA                       NA 1601 CANTERBURY ST BLDG 1
## 6               NA                       NA 1601 CANTERBURY ST BLDG 2
##   Original.City Original.State Original.Zip Council.District
## 1        AUSTIN             TX        78702                3
## 2        AUSTIN             TX        78702                3
## 3        AUSTIN             TX        78749                8
```

```
## 4        AUSTIN             TX        78703               10
## 5        AUSTIN             TX        78702                3
## 6        AUSTIN             TX        78702                3
##          Jurisdiction
## 1 AUSTIN FULL PURPOSE
## 2 AUSTIN FULL PURPOSE
## 3 AUSTIN FULL PURPOSE
## 4 AUSTIN FULL PURPOSE
## 5 AUSTIN FULL PURPOSE
## 6 AUSTIN FULL PURPOSE
##
Link
## 1 https://abc.austintexas.gov/web/permit/public-search-other?t_detail=1&t_selected_folderr
sn=13233784
## 2 https://abc.austintexas.gov/web/permit/public-search-other?t_detail=1&t_selected_folderr
sn=13233787
## 3 https://abc.austintexas.gov/web/permit/public-search-other?t_detail=1&t_selected_folderr
sn=13236776
## 4 https://abc.austintexas.gov/web/permit/public-search-other?t_detail=1&t_selected_folderr
sn=13256573
## 5 https://abc.austintexas.gov/web/permit/public-search-other?t_detail=1&t_selected_folderr
sn=13258544
## 6 https://abc.austintexas.gov/web/permit/public-search-other?t_detail=1&t_selected_folderr
sn=13258545
##   Project.ID Master.Permit.Num Latitude Longitude                    Location
## 1   13233784          13233776 30.26010 -97.70677 (30.26010298, -97.70676573)
## 2   13233787          13233777 30.26010 -97.70677 (30.26010298, -97.70676573)
## 3   13236776          13236775 30.20362 -97.88271  (30.2036226, -97.88271161)
## 4   13256573          13247138 30.29306 -97.77269 (30.29305796, -97.77269138)
## 5   13258544          13211262 30.25642 -97.72943 (30.25641539, -97.72943283)
## 6   13258545          13211262 30.25642 -97.72943 (30.25641539, -97.72943283)
##        Contractor.Trade             Contractor.Company.Name
## 1   Plumbing Contractor                    Loredos Plumbing
## 2   Plumbing Contractor                    Loredos Plumbing
## 3 Electrical Contractor                    Economy Electric
## 4    General Contractor  Wilmington-Gordon Inc.****MAIN****
## 5    General Contractor Guardian Custom Builders****MAIN***
## 6    General Contractor Guardian Custom Builders****MAIN***
##   Contractor.Full.Name Contractor.Phone Contractor.Address.1
## 1      Reynaldo Loredo       7372636253
## 2      Reynaldo Loredo       7372636253
## 3       Jerry Brinkley       5128454717
## 4       Kenneth Burger       5124547070
## 5      Jeffrey R Grier       2107105222
## 6      Jeffrey R Grier       2107105222
##       Contractor.Address.2 Contractor.City Contractor.Zip Applicant.Full.Name
## 1   13001  Amaryllis TRAIL           ELGIN     78621-____
## 2   13001  Amaryllis TRAIL           ELGIN     78621-____
## 3           1308-B  Kramer          Austin          78758
## 4       1209 W 49th STREET          Austin          78756      Kenneth Burger
## 5 777  SHADY LANE  Suite 8          AUSTIN          78702     Jeffrey R Grier
```

```
## 6 777  SHADY LANE  Suite 8          AUSTIN          78702     Jeffrey R Grier
##                Applicant.Organization Applicant.Phone Applicant.Address.1
## 1
## 2
## 3
## 4  Wilmington-Gordon Inc.****MAIN****      5124547070
## 5 Guardian Custom Builders****MAIN***      2107105222
## 6 Guardian Custom Builders****MAIN***      2107105222
##        Applicant.Address.2 Applicant.City Applicant.Zip
## 1
## 2
## 3
## 4       1209 W 49th STREET         Austin         78756
## 5 777  SHADY LANE  Suite 8         AUSTIN         78702
## 6 777  SHADY LANE  Suite 8         AUSTIN         78702
##   Certificate.Of.Occupancy Total.Lot.SQFT
## 1                       No             NA
## 2                       No             NA
## 3                       No             NA
## 4                       No          12256
## 5                      Yes           8680
## 6                      Yes           8680
```

```
as.data.frame(colnames(construction_permits))
```

```
##    colnames(construction_permits)
## 1                     Permit.Type
## 2                Permit.Type.Desc
## 3                      Permit.Num
## 4             Permit.Class.Mapped
## 5                    Permit.Class
## 6                      Work.Class
## 7                     Condominium
## 8                    Project.Name
## 9                     Description
## 10                        TCAD.ID
## 11    Property.Legal.Description
## 12                   Applied.Date
## 13                    Issued.Date
## 14                     Day.Issued
## 15           Calendar.Year.Issued
## 16             Fiscal.Year.Issued
## 17         Issued.In.Last.30.Days
## 18                Issuance.Method
## 19                 Status.Current
## 20                    Status.Date
## 21                   Expires.Date
## 22                 Completed.Date
## 23       Total.Existing.Bldg.SQFT
## 24            Remodel.Repair.SQFT
## 25             Total.New.Add.SQFT
## 26        Total.Valuation.Remodel
## 27            Total.Job.Valuation
## 28               Number.Of.Floors
## 29                  Housing.Units
## 30             Building.Valuation
## 31     Building.Valuation.Remodel
## 32           Electrical.Valuation
## 33   Electrical.Valuation.Remodel
## 34           Mechanical.Valuation
## 35   Mechanical.Valuation.Remodel
## 36             Plumbing.Valuation
## 37     Plumbing.Valuation.Remodel
## 38               MedGas.Valuation
## 39       MedGas.Valuation.Remodel
## 40             Original.Address.1
## 41                  Original.City
## 42                 Original.State
## 43                   Original.Zip
## 44               Council.District
## 45                   Jurisdiction
## 46                           Link
## 47                     Project.ID
## 48              Master.Permit.Num
## 49                       Latitude
## 50                      Longitude
```

```
## 51                      Location
## 52              Contractor.Trade
## 53        Contractor.Company.Name
## 54           Contractor.Full.Name
## 55               Contractor.Phone
## 56           Contractor.Address.1
## 57           Contractor.Address.2
## 58                Contractor.City
## 59                 Contractor.Zip
## 60            Applicant.Full.Name
## 61         Applicant.Organization
## 62                Applicant.Phone
## 63            Applicant.Address.1
## 64            Applicant.Address.2
## 65                 Applicant.City
## 66                  Applicant.Zip
## 67       Certificate.Of.Occupancy
## 68                 Total.Lot.SQFT
```

```
dim(construction_permits)
```

```
## [1] 897369     68
```

Note - ArcGIS Reverse Geocoding Process

In the Traffic Incidents data set there was not a column for Zip codes. Because there was available data for latitude and longitude, we decided to reverse-geocode the lat/lon values in a column of zip codes. We initially were attempting to utilize Maps API services to perform this task but it was cost prohibitive. We opted to use a Geographic Information System software named ArcGIS to produce the Zipcode data leveraging a TIGER/Line Shapefile provided by the US Census Bureau to reverse geocode the lat/lon figures through ArcGIS's geoferencing capabilities.

```
# This code reads the Traffic Incidents data set into a R dataframe
traffic_indcidents <- read.csv("~/My Documents/UT Course Files/UT Fall 2024/SDS 322E/SDS Proj
ect Files/Original Data/joined_traffic_zipcode_csv.csv")

head(traffic_indcidents)
```

```
##   Join_Count TARGET_FID                                Traffic_Report_ID
## 1          0          1 121E4F6B2D93D3F508359C8700406A1B992733AF_1659735434
## 2          0          2 B852035718A45A6479B38C26FA96B28B0C9A8A56_1661547901
## 3          0          3 B7CA5DF711D07BA6D03B6EF004402A6594C77CD6_1662264962
## 4          1          4 5F5898E4726001663BA5A126B313B03B1AED3F07_1663168459
## 5          1          5 3791C82875F2B544CEFA8FBE35109575B1431A30_1663163459
## 6          1          6 00F36866326DA3B8DDE0D960226DCD6AFB5AF127_1663074132
##                  Published_Date    Issue_Reported                    Location
## 1 08/05/2022 09:37:14 PM +0000 TRFC HAZD/ DEBRIS                  POINT (0 0)
## 2 08/26/2022 09:05:01 PM +0000 TRFC HAZD/ DEBRIS                  POINT (0 0)
## 3 09/04/2022 04:16:02 AM +0000         COLLISION                  POINT (0 0)
## 4 09/14/2022 03:14:19 PM +0000      Crash Urgent POINT (-97.711561 30.307396)
## 5 09/14/2022 01:50:59 PM +0000 TRFC HAZD/ DEBRIS POINT (-97.820007 30.233228)
## 6 09/13/2022 01:02:12 PM +0000     Crash Service POINT (-97.780078 30.439546)
##   Latitude Longitude                    Address   Status
## 1  0.00000   0.00000                        tra ARCHIVED
## 2  0.00000   0.00000               900 S FM 973 ARCHIVED
## 3  0.00000   0.00000         12009 W US 290 HWY ARCHIVED
## 4 30.30740 -97.71156            4900 N Ih 35 Nb ARCHIVED
## 5 30.23323 -97.82001  4953-4973 W Us 290 Hwy Eb ARCHIVED
## 6 30.43955 -97.78008 13096 N Us 183 Hwy Svrd Sb ARCHIVED
##                    Status_Date Agency ZCTA5CE20 GEOID20      GEOIDFQ20
## 1 08/05/2022 09:50:03 PM +0000     NA        NA      NA
## 2 08/26/2022 09:35:03 PM +0000     NA        NA      NA
## 3 09/04/2022 05:40:03 AM +0000     NA        NA      NA
## 4 09/14/2022 03:30:04 PM +0000     NA     78723   78723 860Z200US78723
## 5 09/14/2022 02:10:02 PM +0000     NA     78745   78745 860Z200US78745
## 6 09/13/2022 01:25:04 PM +0000     NA     78750   78750 860Z200US78750
##   CLASSFP20 MTFCC20 FUNCSTAT20  ALAND20 AWATER20 INTPTLAT20 INTPTLON20
## 1                                   NA       NA         NA         NA
## 2                                   NA       NA         NA         NA
## 3                                   NA       NA         NA         NA
## 4        B5   G6350          S 18281811        0   30.30424  -97.68575
## 5        B5   G6350          S 35576949        0   30.20742  -97.79824
## 6        B5   G6350          S 30557605        0   30.44108  -97.78667
```

```
as.data.frame(colnames(traffic_indcidents))
```

```
##    colnames(traffic_indcidents)
## 1                    Join_Count
## 2                    TARGET_FID
## 3             Traffic_Report_ID
## 4                Published_Date
## 5                Issue_Reported
## 6                      Location
## 7                      Latitude
## 8                     Longitude
## 9                       Address
## 10                       Status
## 11                  Status_Date
## 12                       Agency
## 13                   ZCTA5CE20
## 14                     GEOID20
## 15                   GEOIDFQ20
## 16                   CLASSFP20
## 17                     MTFCC20
## 18                  FUNCSTAT20
## 19                     ALAND20
## 20                    AWATER20
## 21                  INTPTLAT20
## 22                  INTPTLON20
```

```
dim(traffic_indcidents)
```

```
## [1] 356150     22
```

Cleaning Data

Removing unneeded columns

```
# Selecting only the useful columns for Traffic Indcidents Dataset
ti_df = traffic_indcidents[,c('Published_Date',
                              'Issue_Reported',
                              'ZCTA5CE20')]

# renaming the zipcode column
ti_df <- ti_df |>
  rename(zipcode = ZCTA5CE20)
```

```
# Selecting only the useful columns for Construction permits
cp_df = construction_permits[,c('Permit.Class',
                                'Permit.Type.Desc',
                                'Issued.Date',
                                'Status.Current',
                                'Number.Of.Floors',
                                'Original.Zip')]
```

```
# if starting character equals R, we isolate it to represent Residential
# if starting character equals C, we isolate it to represent Commercial
cp_df$Permit.Class.Simple <- ifelse(cp_df$Permit.Class == "", "",
                                    substr(cp_df$Permit.Class, 1, 1))

cp_df$Permit.Class.Simple <- ifelse(cp_df$Permit.Class.Simple == "S", NA,
                                    cp_df$Permit.Class.Simple)

cp_df$Permit.Class.Simple <- ifelse(cp_df$Permit.Class.Simple == "", NA,
                                    cp_df$Permit.Class.Simple)

unique(cp_df$Permit.Class.Simple)
```

```
## [1] "R" "C" NA
```

NA Values

Count NA Values

```
# counts the number of rows with missing values
missing_rows_count_ti <- sum(!complete.cases(ti_df))
missing_rows_count_cp <- sum(!complete.cases(cp_df))


sprintf("There are % s rows with missing values in the traffic incidents dataset.", missing_r
ows_count_ti)
```

```
## [1] "There are 817 rows with missing values in the traffic incidents dataset."
```

```
sprintf("There are % s rows with missing values in the construction permits dataset.", missin
g_rows_count_cp)
```

```
## [1] "There are 191411 rows with missing values in the construction permits dataset."
```

Drop NA Values

```
# dropping NA values from the Traffic Incidents data set & saving the data frame
ti_df_noNA <- na.omit(ti_df)
```

```
# assign rows with description type of "Driveway / Sidewalks" to 0
cp_df$Number.Of.Floors <- ifelse(is.na(cp_df$Number.Of.Floors) & cp_df$Permit.Type.Desc == "D
riveway / Sidewalks", 0, cp_df$Number.Of.Floors)

# dropping NA values from the Construction Permits data set & saving the data frame
cp_df_noNA <- na.omit(cp_df)
```

Rows and Columns of resulting data set after cleaning

```
# Rows and Columns of the cleaned Traffic Incident data set
dim(ti_df_noNA)
```

```
## [1] 355333      3
```

```
# Rows and Columns of the cleaned Construction Permit data set
dim(cp_df_noNA)
```

```
## [1] 737781      7
```

In the Traffic Incident data set we started with 356,150 rows and 22 columns. In the Construction Permits data set we started with 897,369 rows and 68 columns. In the Traffic Incident data set we ended with 355,333 and 3 columns. In the Construction Permits data set we ended with 737,781 and 7 columns.

We selected certain rows based on their relevance to our area of investigation, which reduced the number of columns. We utilized ArcGIS Pro to create a zipcode column for traffic incidents based on the existing latitude and longitude columns for each row.

We also dropped any rows that contained missing values across the columns of interest. And prior to that, we converted the number of floors for rows with the ‘Driveway/Sidewalk’ permit type description and missing floor values to be 0 as we can safely assume that sidewalks and driveways are on at ground level.

The resulting data tidy in each data set as each column represents a characteristic, each row represents an individual permit or incident, and each cell represents an individual value.

To match the data sets we need to group by zip codes to relate the two datasets togehter. We do this by using the groupby() method.

Results

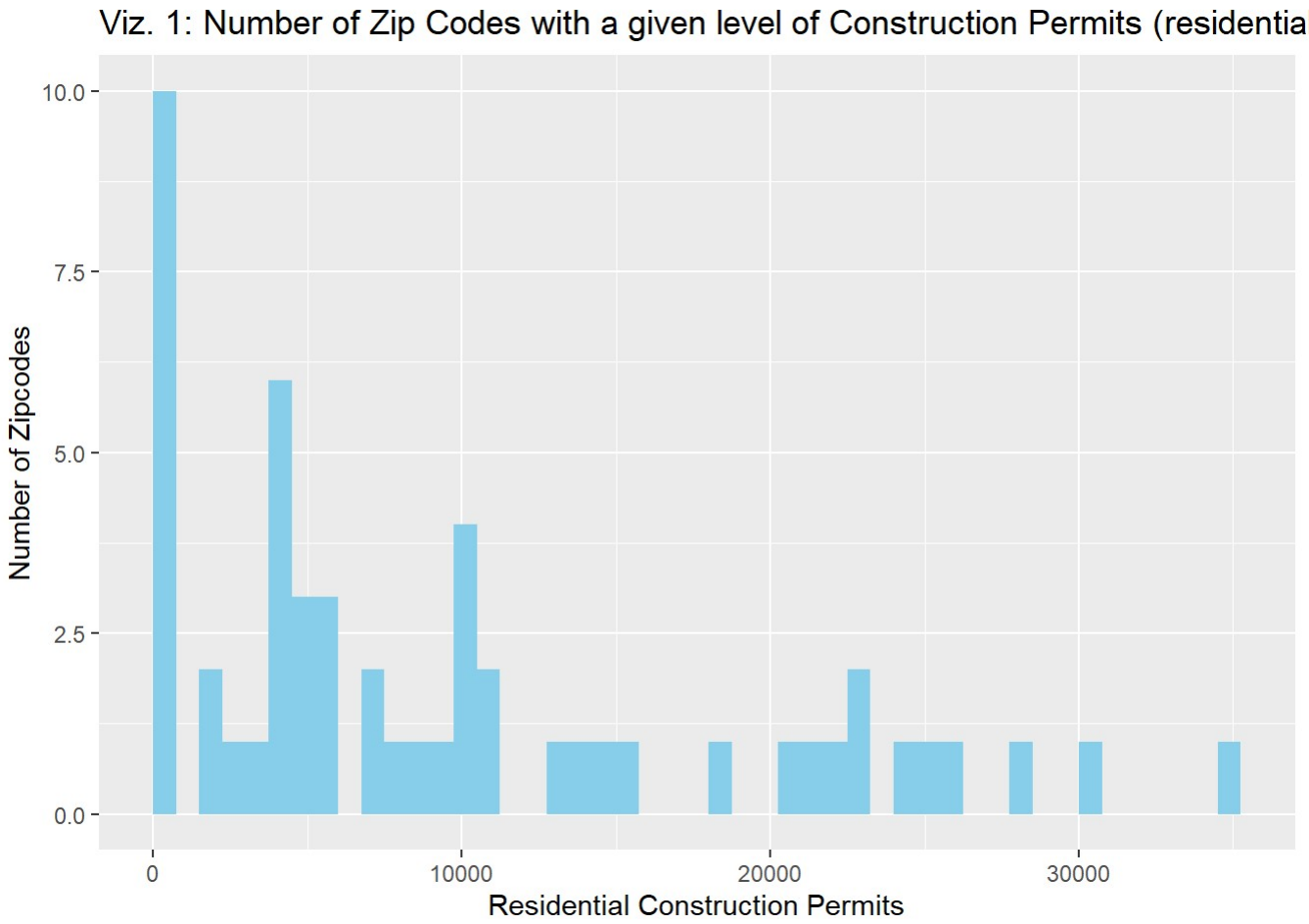
Research Question 1 - EDA ; Visualization #1 & #2

Is there a difference in impact of Residential or Commercial construction projects on Traffic Incidents?

```
# find the number of zip codes
zip_code_counts <- cp_df_noNA |>
  count(Original.Zip)
```

Zip Code Counts - Summary Statistics

```
zip_count_r <- cp_df_noNA[cp_df_noNA$Permit.Class.Simple == "R",] |>
  count(Original.Zip)

zip_count_c <- cp_df_noNA[cp_df_noNA$Permit.Class.Simple == "C",] |>
  count(Original.Zip)

# plot distribution of residential permits and commercial permits across the zipcode
ggplot(zip_count_r, aes(x = n)) +
  geom_histogram(fill = "skyblue", binwidth = 750, center = 375) +
  labs(title = "Viz. 1: Number of Zip Codes with a given level of Construction Permits (resid
ential).",
       y = "Number of Zipcodes",
       x = 'Residential Construction Permits')
```



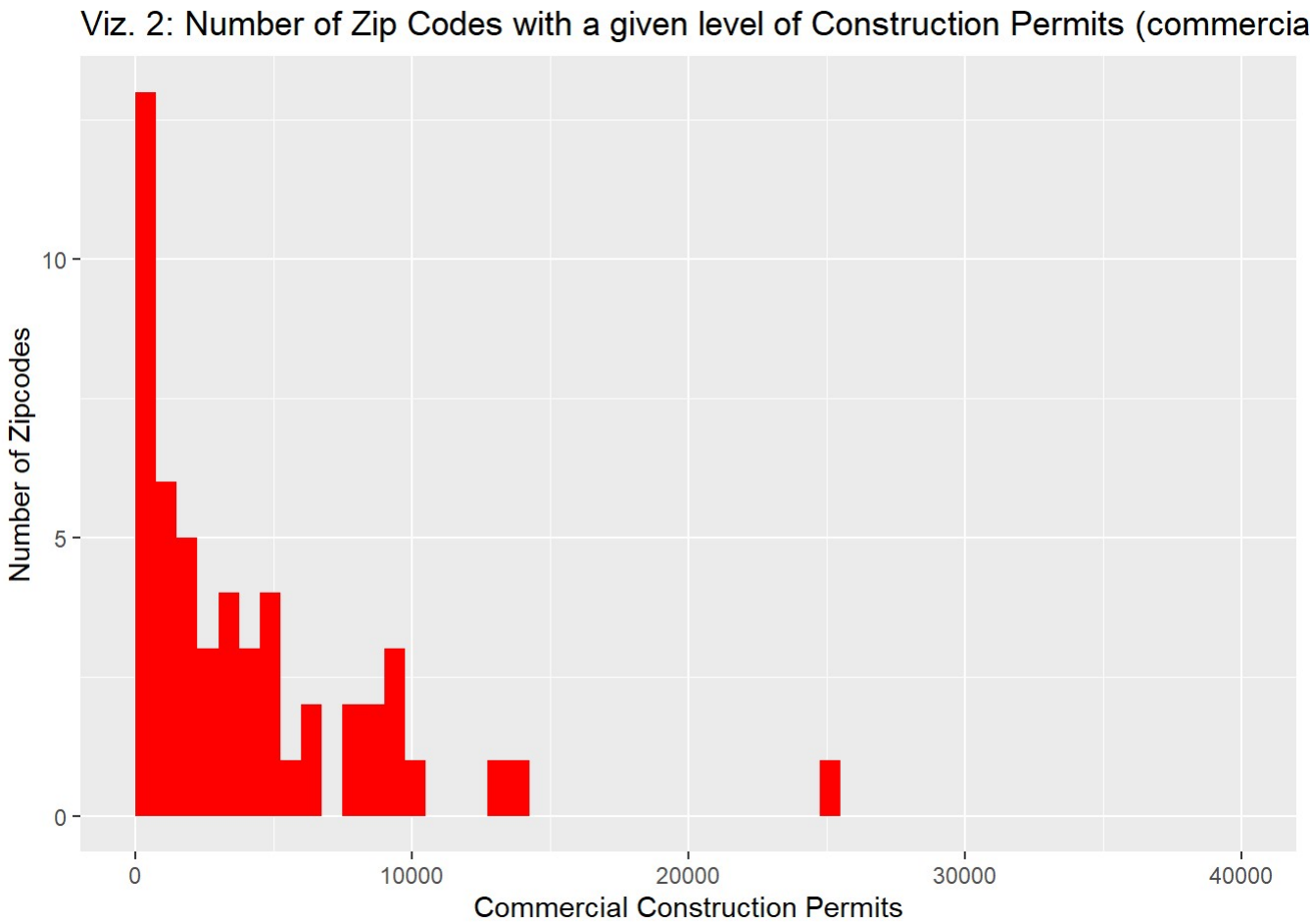

```
summary(zip_count_r)
```

```
##   Original.Zip         n
##  Min.   :78610   Min.   :    2
##  1st Qu.:78704   1st Qu.: 3487
##  Median :78730   Median : 7043
##  Mean   :78717   Mean   : 9876
##  3rd Qu.:78745   3rd Qu.:14316
##  Max.   :78759   Max.   :35104
```

There are more zipcodes with fewer residential construction permits. The data is also positively skewed. The mean number of permits per zipcode is 9,876 whilst the median is 7,043.

```
ggplot(zip_count_c, aes(x = n)) +
  geom_histogram(fill = "red", binwidth = 750, center = 375) +
  xlim(0, 40000) +
  labs(title = "Viz. 2: Number of Zip Codes with a given level of Construction Permits (comme
rcial).",
       y = "Number of Zipcodes",
       x = 'Commercial Construction Permits')
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_bar()`).
```



```
summary(zip_count_c)
```

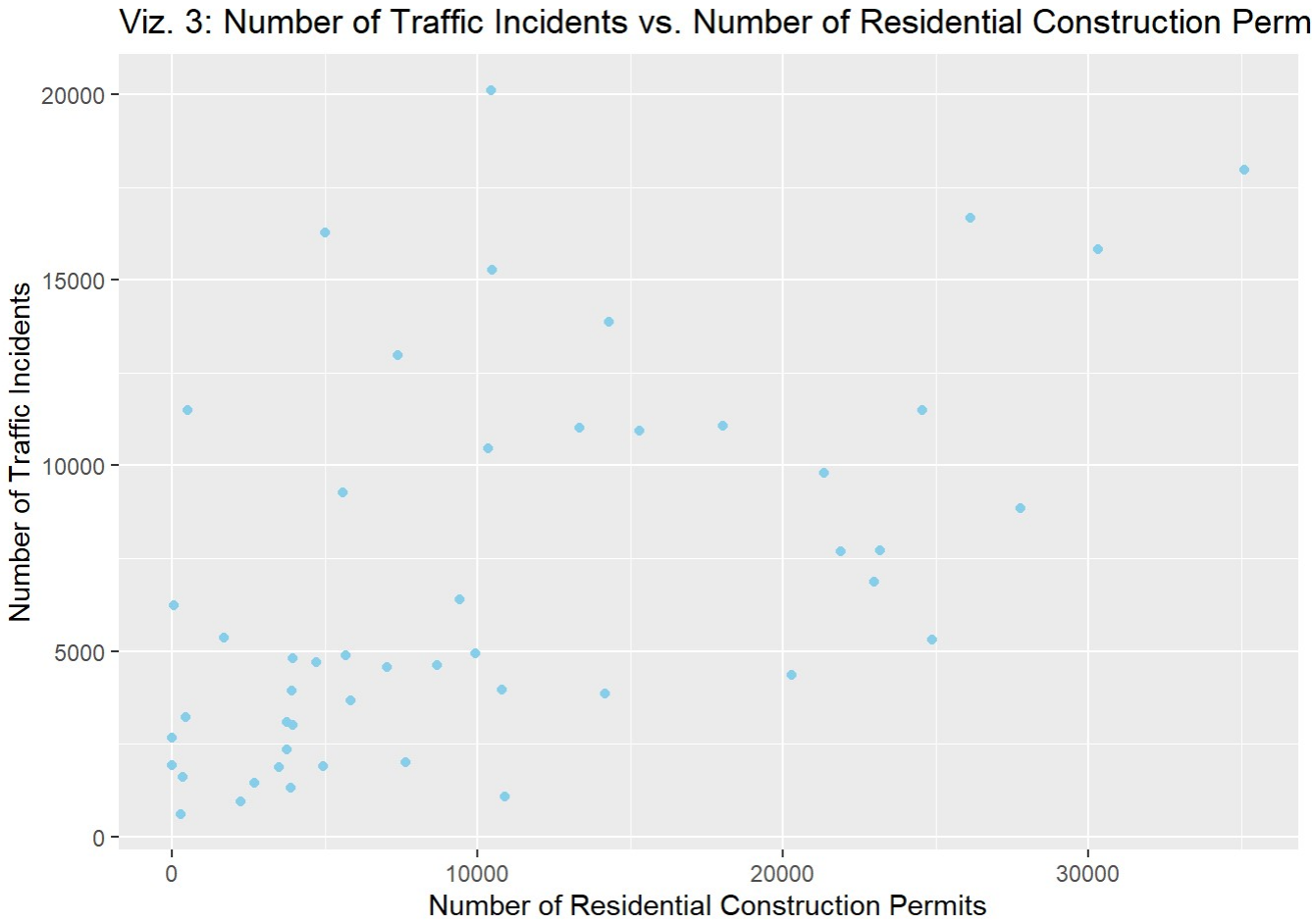
```
##   Original.Zip         n          
##  Min.   :78610   Min.   :    3  
##  1st Qu.:78710   1st Qu.:  778  
##  Median :78731   Median : 2576  
##  Mean   :78719   Mean   : 4122  
##  3rd Qu.:78745   3rd Qu.: 5712  
##  Max.   :78759   Max.   :25153
```

There are more zipcodes with fewer commercial construction permits. The data is also positively skewed. The mean number of permits per zipcode is 4,122 whilst the median is 2,576.

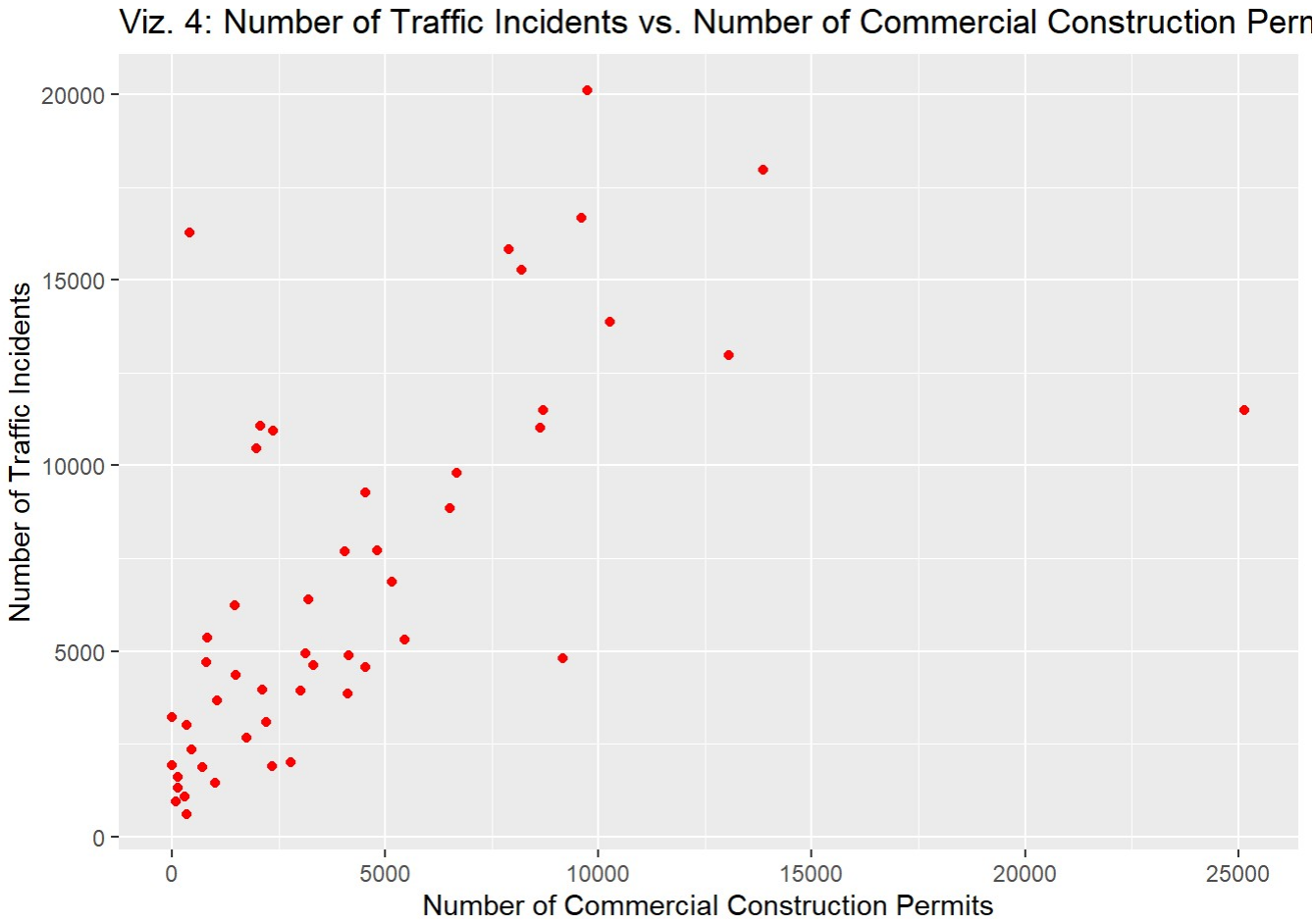
```
zip_code_counts_TI <- ti_df_noNA |> count(zipcode)

summary(zip_code_counts_TI)
```

```
##     zipcode            n          
##  Min.   :76527   Min.   :    1.0  
##  1st Qu.:78641   1st Qu.:  131.5  
##  Median :78705   Median : 1917.0  
##  Mean   :78570   Mean   : 4281.1  
##  3rd Qu.:78738   3rd Qu.: 6305.5  
##  Max.   :78957   Max.   :20116.0
```

```
permit_join <- full_join(zip_count_c, zip_count_r, by = c("Original.Zip"))

permit_join <- permit_join |> rename(C_count = n.x, R_count = n.y)

total_join <- full_join(permit_join, zip_code_counts_TI, by = c("Original.Zip" = "zipcode"))
|> rename(TI_count = n)

total_join |>
  drop_na() |>
  ggplot() +
  geom_point(aes(x = R_count, y = TI_count), color = "skyblue") +
  labs(title = "Viz. 3: Number of Traffic Incidents vs. Number of Residential Construction Pe
rmits", y = "Number of Traffic Incidents", x = 'Number of Residential Construction Permits')
```




```
total_join |>
  drop_na() |>
  ggplot() +
  geom_point(aes(x = C_count, y = TI_count), color = "red") +
  labs(title = "Viz. 4: Number of Traffic Incidents vs. Number of Commercial Construction Per
mits", y = "Number of Traffic Incidents", x = 'Number of Commercial Construction Permits')
```




```
summary(total_join)
```

```
##   Original.Zip      C_count         R_count         TI_count
##  Min.   :76527   Min.   :    3   Min.   :    2   Min.   :    1.0
##  1st Qu.:78641   1st Qu.:  778   1st Qu.: 3487   1st Qu.:  131.5
##  Median :78705   Median : 2576   Median : 7043   Median : 1917.0
##  Mean   :78570   Mean   : 4122   Mean   : 9876   Mean   : 4281.1
##  3rd Qu.:78738   3rd Qu.: 5712   3rd Qu.:14316   3rd Qu.: 6305.5
##  Max.   :78957   Max.   :25153   Max.   :35104   Max.   :20116.0
##                  NA's   :31      NA's   :30
```

All rows with zip codes that are not in the traffic incident data set is removed. There is one zipcode missing from the commercial permits that is in the residential permits. There is 53 points in the residential scatter plot. There are 52 points in the commercial data set where each point is a zipcode. There appears to be a positive correlation between traffic incidents and both residential and commercial construction permit numbers. However, the correlation with residential construction permits is much stronger. We already discussed the commercial and resident permit statistics, but here, let us bring your attention to the traffic incidents. The mean number of traffic incidents per zipcode is 4,281 with a median of 1,917, again indicating positive skew.

Research Question 2 - EDA ; Visualization #3 & #4

How do spatial concentrations of Traffic Incidents change throughout the year?

Visualization 1

```
ti_df_formatted <- ti_df_noNA
```

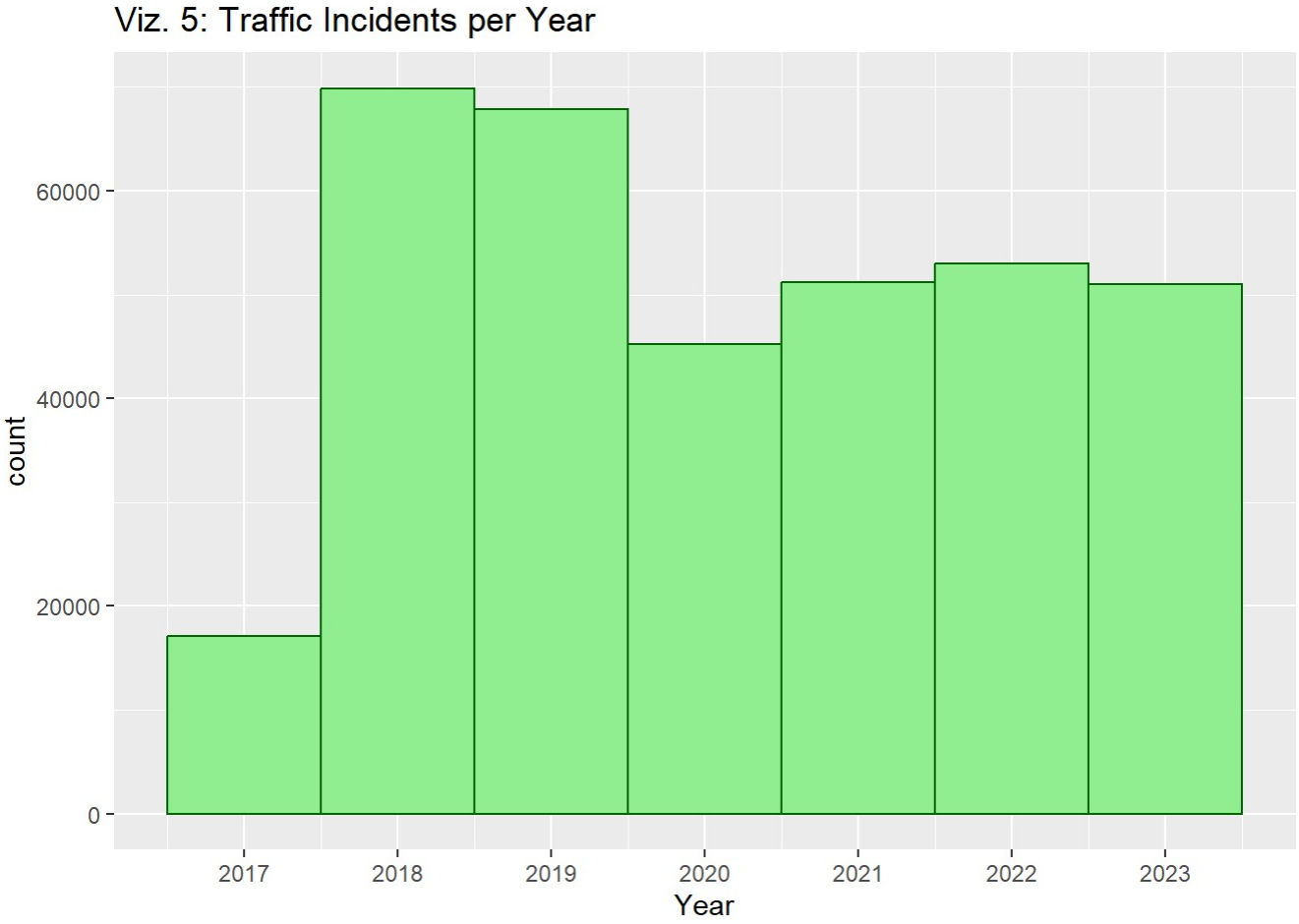
```
ti_df_formatted$Published_Date <- as.POSIXct(ti_df_formatted$Published_Date, format = "%m/%d/
%Y %H:%M:%OS")

# Converting 'Issued.Date' to Date to ensure correct format
#ti_df_noNA$Published_Date <- as.Date(ti_df_noNA$Published_Date)


# Extracting the year from 'Issued.Date'
ti_df_formatted$Year <- as.integer(format(ti_df_formatted$Published_Date, "%Y"))

ti_df_formatted <- na.omit(ti_df_formatted) |>
  filter(Year < 2024)

incidents_summary <- ti_df_formatted |>
  group_by(`zipcode`) |>
    summarize(incidents_count = n())

# showing the number of incidents per year
years_count <- ti_df_formatted |>
  count(Year)

years_count
```

```
##   Year     n
## 1 2017 17111
## 2 2018 69887
## 3 2019 67890
## 4 2020 45220
## 5 2021 51224
## 6 2022 52954
## 7 2023 50957
```

```
summary(ti_df_formatted)
```

```
##  Published_Date                  Issue_Reported       zipcode
##  Min.   :2017-09-26 04:11:00.00   Length:355243      Min.   :76527
##  1st Qu.:2019-01-11 11:23:28.50   Class :character   1st Qu.:78704
##  Median :2020-06-23 10:38:01.00   Mode  :character   Median :78732
##  Mean   :2020-08-28 20:14:42.58                      Mean   :78720
##  3rd Qu.:2022-04-23 10:56:32.00                      3rd Qu.:78748
##  Max.   :2023-12-31 12:44:32.00                      Max.   :78957
##       Year
##  Min.   :2017
##  1st Qu.:2019
##  Median :2020
##  Mean   :2020
##  3rd Qu.:2022
##  Max.   :2023
```

```
# Show the traffic incidents per year
ti_df_formatted |>
  ggplot(aes(x = Year)) +
  geom_histogram(binwidth = 1, center = 0, color = "darkgreen", fill = "lightgreen") +
  scale_x_continuous(breaks = seq(2017, 2023, by = 1), labels = seq(2017, 2023, by = 1)) +
  labs(title = "Viz. 5: Traffic Incidents per Year")
```



There are 26 incidents in the year of 2024, we will remove these outliers since there are not very many.

```
# Fetch ZIP code shapefiles for Texas
zipcodes <- zctas(year = 2010, state = "TX") |>
  select(ZCTA5CE10, geometry) |>
  mutate(ZCTA5CE10 = as.numeric(ZCTA5CE10))
```

```
## ZCTAs can take several minutes to download.  To cache the data and avoid re-downloading in
future R sessions, set `options(tigris_use_cache = TRUE)`
```

```
##   |                                                                      |
|   0%  |                                                                      |
|   1%  |                                                                      |=
|   1%  |                                                                      |=
|   2%  |                                                                      |==
|   2%  |                                                                      |==
|   3%  |                                                                      |===
|   4%  |                                                                      |===
|   5%  |                                                                      |====
|   5%  |                                                                      |====
|   6%  |                                                                      |=====
|   7%  |                                                                      |=====
|   8%  |                                                                      |=====
=                                                                  |   8%  |
|======                                                                |   9%  |
|=======                                                               |   9%  |
|=======                                                               |  10%  |
|=======                                                               |  11%  |
|========                                                              |  11%  |
|========                                                              |  12%  |
|=========                                                             |  12%  |
|=========                                                             |  13%  |
|==========                                                            |  14%  |
|==========                                                            |  15%  |
|===========                                                           |  15%  |
|===========                                                           |  16%  |
|============                                                          |  16%  |
|============                                                          |  17%  |
|============                                                          |  18%  |
|=============                                                         |  18%  |
|=============                                                         |  19%  |
|==============                                                        |  19%  |
|==============                                                        |  20%  |
|==============                                                        |  21%  |
|===============                                                       |  21%  |
|===============                                                       |  22%  |
|================                                                      |  22%  |
|================                                                      |  23%  |
|=================                                                     |  24%  |
|=================                                                     |  25%  |
|==================                                                    |  25%  |
|==================                                                    |  26%  |
|===================                                                   |  27%  |
|===================                                                   |  28%  |
|====================                                                  |  28%  |
|====================                                                  |  29%  |
|=====================                                                 |  30%  |
|=====================                                                 |  31%  |
|======================                                                |  31%  |
|======================                                                |  32%  |
|=======================                                               |  32%  |
```

```
|======================                                                |  33%  |
|=======================                                               |  34%  |
|=======================                                               |  35%  |
|========================                                              |  35%  |
|========================                                              |  36%  |
|=========================                                             |  37%  |
|=========================                                             |  38%  |
|==========================                                            |  38%  |
|==========================                                            |  39%  |
|===========================                                           |  39%  |
|===========================                                           |  40%  |
|============================                                          |  41%  |
|============================                                          |  42%  |
|=============================                                         |  42%  |
|=============================                                         |  43%  |
|==============================                                        |  44%  |
|==============================                                        |  45%  |
|===============================                                       |  45%  |
|===============================                                       |  46%  |
|================================                                      |  46%  |
|================================                                      |  47%  |
|================================                                      |  48%  |
|=================================                                     |  48%  |
|=================================                                     |  49%  |
|==================================                                    |  49%  |
|==================================                                    |  50%  |
|==================================                                    |  51%  |
|===================================                                   |  51%  |
|===================================                                   |  52%  |
|====================================                                  |  52%  |
|====================================                                  |  53%  |
|=====================================                                 |  54%  |
|=====================================                                 |  55%  |
|======================================                                |  56%  |
|======================================                                |  57%  |
|======================================                                |  58%  |
|=======================================                               |  58%  |
|=======================================                               |  59%  |
|========================================                              |  60%  |
|========================================                              |  61%  |
|=========================================                             |  61%  |
|=========================================                             |  62%  |
|==========================================                            |  62%  |
|==========================================                            |  63%  |
|==========================================                            |  64%  |
|===========================================                           |  64%  |
|===========================================                           |  65%  |
|============================================                          |  65%  |
|============================================                          |  66%  |
|=============================================                         |  67%  |
|=============================================                         |  68%  |
```

```
|=============================================                         |  68%  |
|=============================================                         |  69%  |
|==============================================                        |  69%  |
|==============================================                        |  70%  |
|==============================================                        |  71%  |
|===============================================                       |  71%  |
|===============================================                       |  72%  |
|================================================                      |  72%  |
|================================================                      |  73%  |
|================================================                      |  74%  |
|=================================================                     |  74%  |
|=================================================                     |  75%  |
|=================================================                     |  75%  |
|==================================================                    |  76%  |
|==================================================                    |  77%  |
|==================================================                    |  78%  |
|===================================================                   |  78%  |
|===================================================                   |  79%  |
|====================================================                  |  79%  |
|====================================================                  |  80%  |
|====================================================                  |  81%  |
|=====================================================                 |  82%  |
|=====================================================                 |  82%  |
|=====================================================                 |  83%  |
|======================================================                |  84%  |
|======================================================                |  85%  |
|======================================================                |  85%  |
|=======================================================               |  86%  |
|=======================================================               |  86%  |
|=======================================================               |  87%  |
|========================================================              |  88%  |
|========================================================              |  88%  |
|========================================================              |  89%  |
|=========================================================             |  89%  |
|=========================================================             |  90%  |
|=========================================================             |  91%  |
|==========================================================            |  91%  |
|==========================================================            |  92%  |
|==========================================================            |  92%  |
|===========================================================           |  93%  |
|============================================================          |  94%  |
|============================================================          |  95%  |
|============================================================          |  95%  |
|============================================================          |  96%  |
|=============================================================         |  96%  |
|=============================================================         |  97%  |
|=============================================================         |  98%  |
|==============================================================        |  98%  |
|==============================================================        |  99%  |
|==============================================================| 100%
```

```
ti_df_17_20 <- ti_df_formatted |>
  filter(Year <= 2020) |>
  group_by(`zipcode`) |>
  summarize(incidents_count = n())

summary(ti_df_17_20)
```

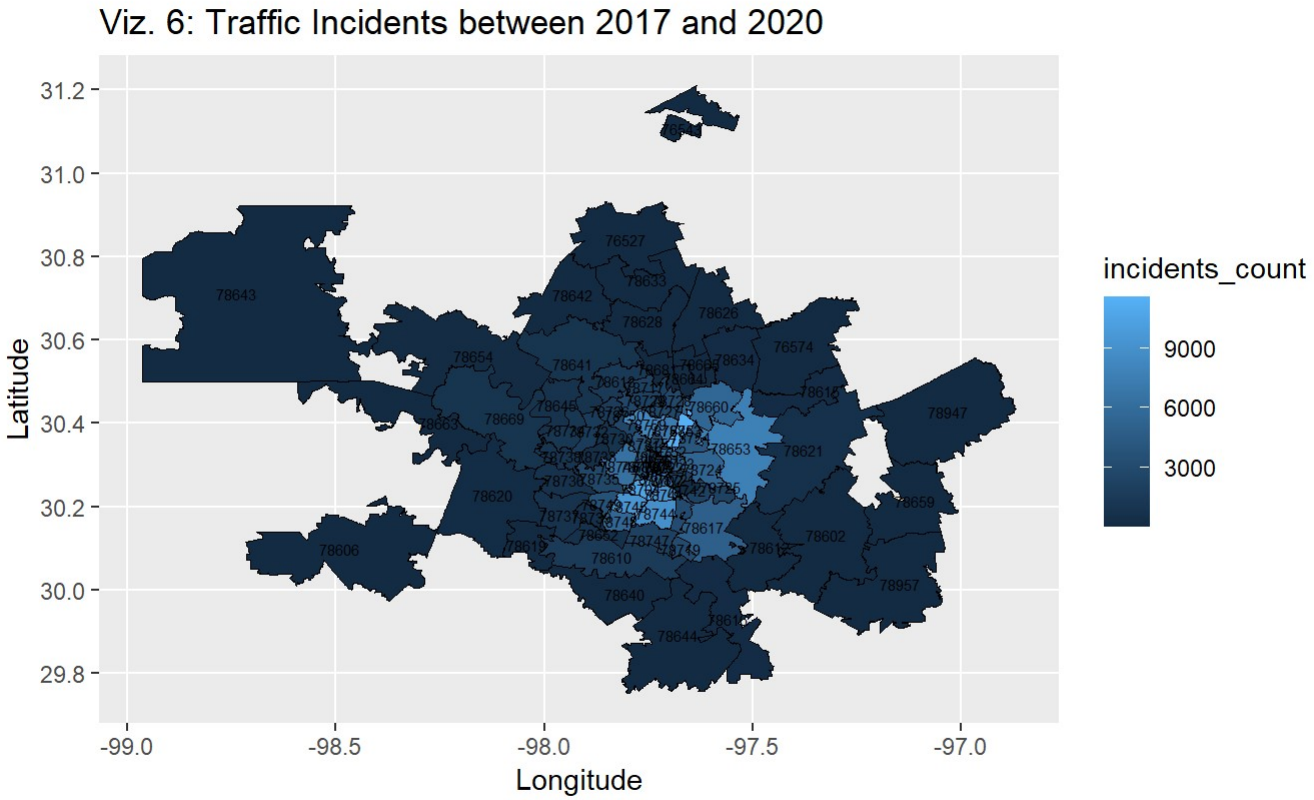
```
##     zipcode      incidents_count
##  Min.   :76527   Min.   :    1
##  1st Qu.:78642   1st Qu.:  109
##  Median :78715   Median : 1245
##  Mean   :78620   Mean   : 2501
##  3rd Qu.:78738   3rd Qu.: 3895
##  Max.   :78957   Max.   :11650
```

```
ti_df_21_23 <- ti_df_formatted |>
  filter(Year > 2020) |>
  group_by(`zipcode`) |>
  summarize(incidents_count = n())

summary(ti_df_21_23)
```

```
##     zipcode      incidents_count
##  Min.   :76527   Min.   :   1.0
##  1st Qu.:78637   1st Qu.: 151.5
##  Median :78705   Median :1123.0
##  Mean   :78558   Mean   :1963.7
##  3rd Qu.:78737   3rd Qu.:2794.5
##  Max.   :78759   Max.   :8544.0
```

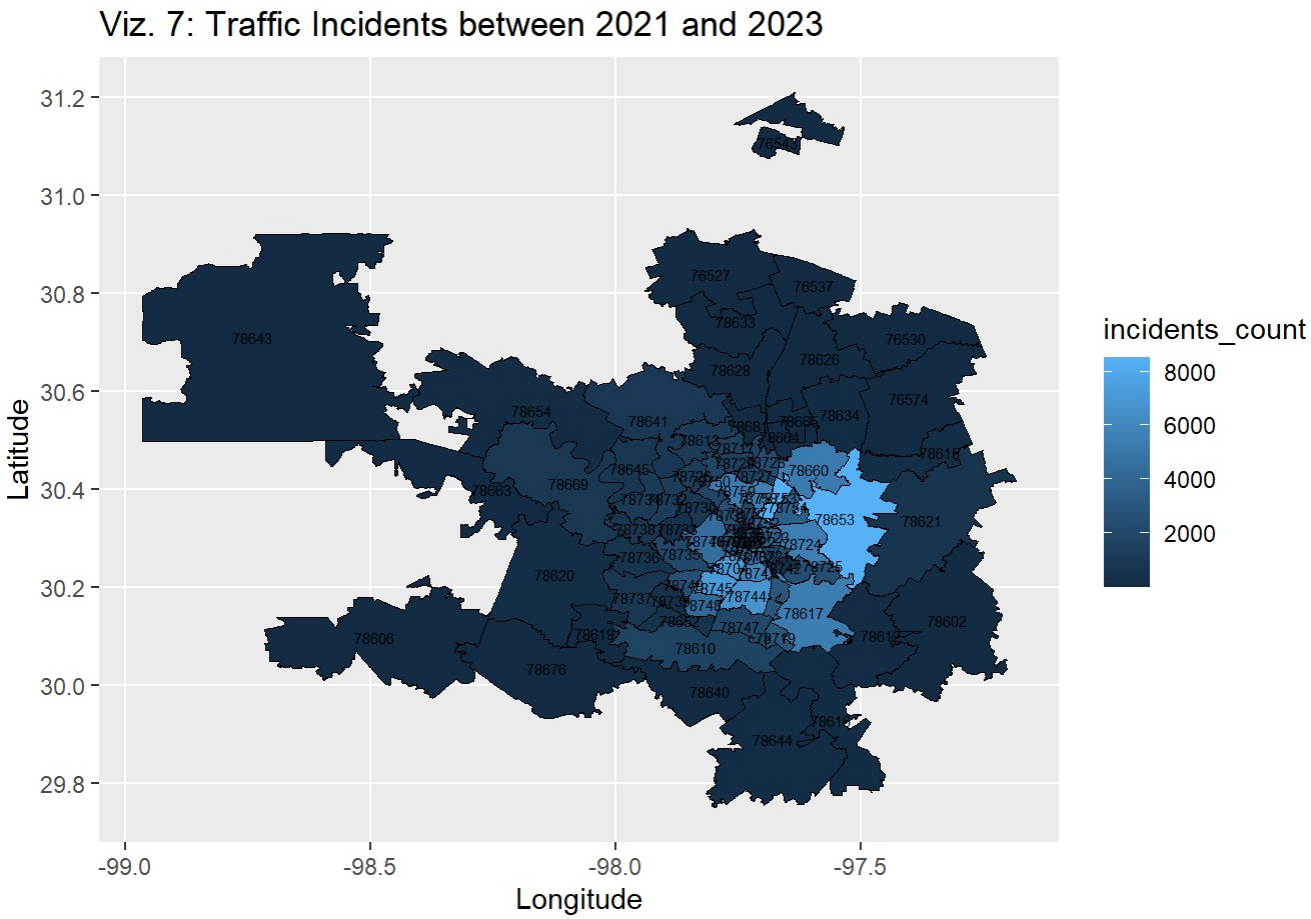
```
join_17_20 <- ti_df_17_20 |>
  inner_join(zipcodes, by = c("zipcode" = "ZCTA5CE10"))

ti_df_17_20 |>
  # Finds the elements that exists in both the shape file and the coyotes_summary data set
  inner_join(zipcodes, by = c("zipcode" = "ZCTA5CE10")) |>
  # ggplot sets the charting library
  ggplot() +
  #
  geom_sf(aes(geometry = geometry, fill = incidents_count), color = "black") +
  #
  geom_sf_text(aes(geometry = geometry, label = zipcode), size = 2, color = "black") +

  labs(title = "Viz. 6: Traffic Incidents between 2017 and 2020", x = "Longitude", y = "Latit
ude")
```

```
## Warning in st_point_on_surface.sfc(sf::st_zm(x)): st_point_on_surface may not
## give correct results for longitude/latitude data
```




```
ti_df_21_23 |>
  # Finds the elements that exists in both the shape file and the coyotes_summary data set
  inner_join(zipcodes, by = c("zipcode" = "ZCTA5CE10")) |>
  # ggplot sets the charting library
  ggplot() +
  #
  geom_sf(aes(geometry = geometry, fill = incidents_count), color = "black") +
  #
  geom_sf_text(aes(geometry = geometry, label = zipcode), size = 2, color = "black") +

  labs(title = "Viz. 7: Traffic Incidents between 2021 and 2023", x = "Longitude", y = "Latit
ude")
```

```
## Warning in st_point_on_surface.sfc(sf::st_zm(x)): st_point_on_surface may not
## give correct results for longitude/latitude data
```

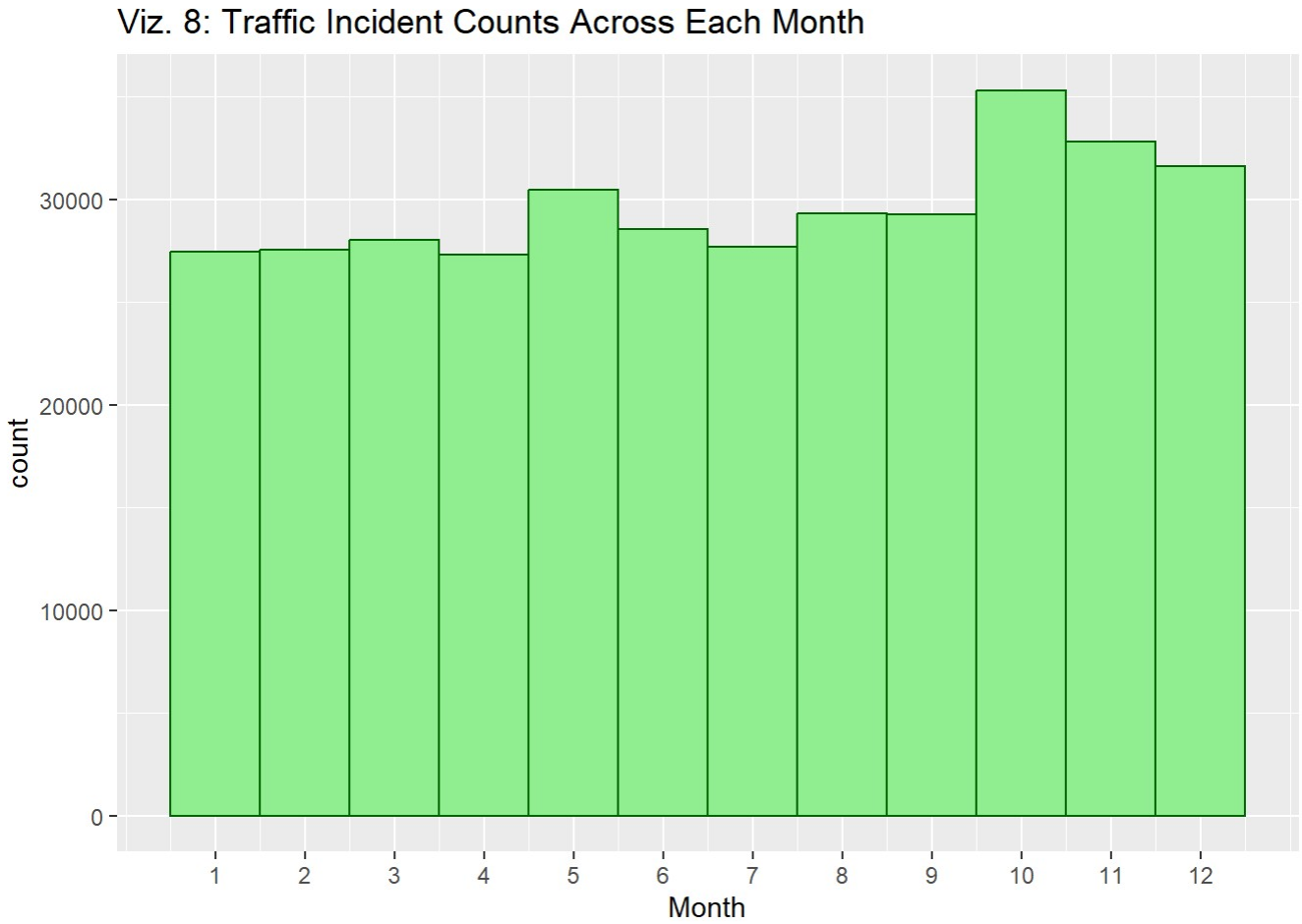


These visualizations agree with our hypothesis that traffic incidents would be more focused in the city center. There are more zipcodes in the dataset with the years of 2017 through 2020. Between the years of 2017 and 2020 the median and mean number of traffic incidents is 1,245 and 2,501, respectively. Between the years of 2021 and 2024 the median and mean number of traffic incidents is 1,123 and 1,963.7, respectively.

Visualization 2

How do trends in traffic incidents changes throughout the year (difference by month)

```
# Extracting the month from 'Issued.Date'
ti_df_formatted$Month <- as.integer(format(ti_df_formatted$Published_Date, "%m"))

ti_df_formatted |>
  ggplot(aes(x = Month)) +
  geom_histogram(binwidth = 1, center = 0, color = "darkgreen", fill = "lightgreen") +
  scale_x_continuous(breaks = seq(1, 12, by = 1), labels = seq(1, 12, by = 1)) +
  labs(title = "Viz. 8: Traffic Incident Counts Across Each Month")
```

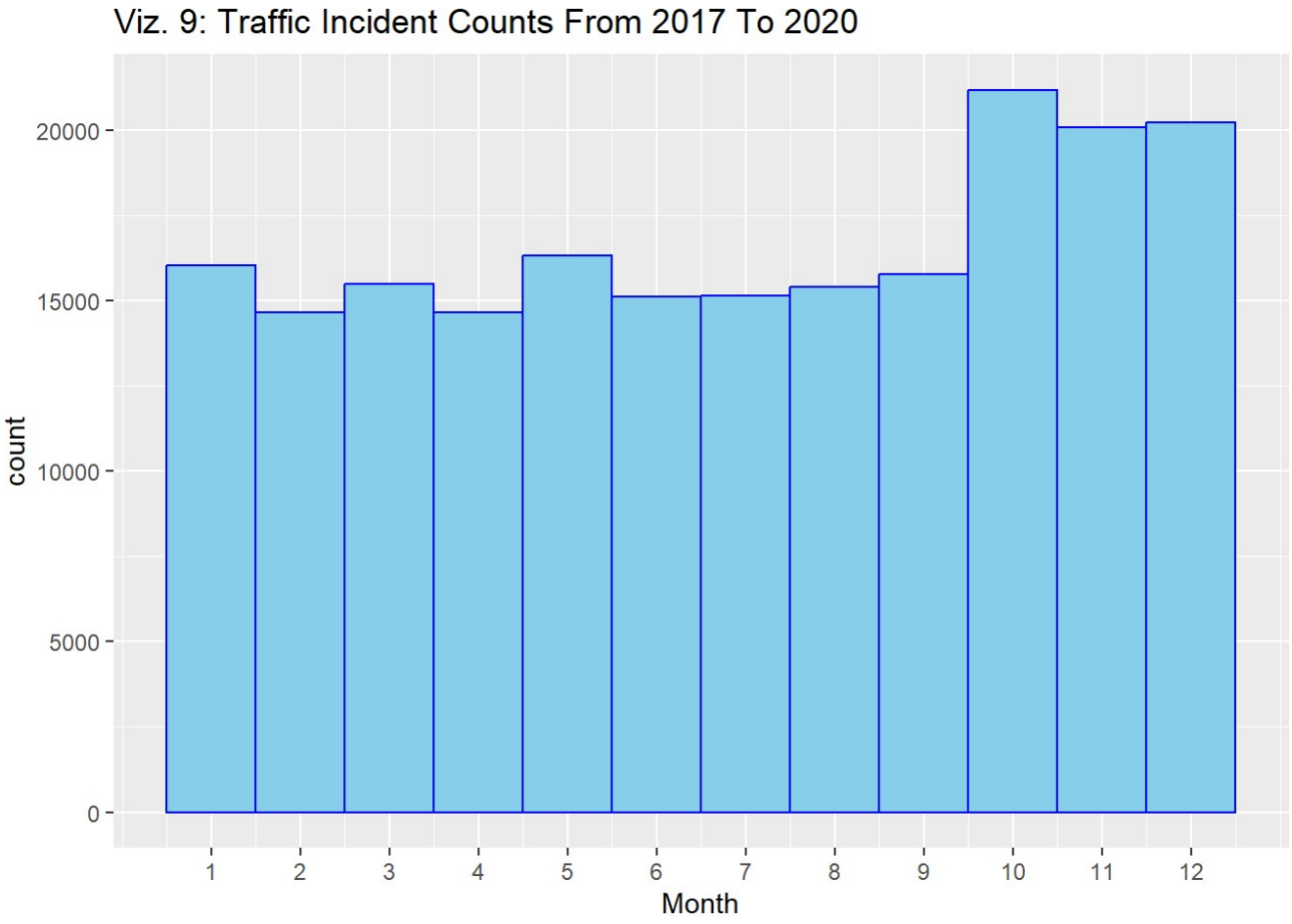



```
# displays the summary statistics for the
# number of traffic incidents during each month
ti_df_formatted |>
  count(Month) |>
  summary()
```

```
##      Month             n
##  Min.   : 1.00   Min.   :27278
##  1st Qu.: 3.75   1st Qu.:27638
##  Median : 6.50   Median :28895
##  Mean   : 6.50   Mean   :29604
##  3rd Qu.: 9.25   3rd Qu.:30753
##  Max.   :12.00   Max.   :35308
```

For the number of traffic incidents across the months, there is a spike in accidents during October that decreases during January. This indicates an increase in traffic incidents in the winter months. The median number of traffic incidents across the months is 28,895 while the mean is 29,604. There is a Min of 27,278 and a Max of 35,308.

```
ti_df_formatted |>
  filter(Year <= 2020) |>
  ggplot(aes(x = Month)) +
  geom_histogram(binwidth = 1, center = 0, color = "blue", fill = "skyblue") +
  scale_x_continuous(breaks = seq(1, 12, by = 1), labels = seq(1, 12, by = 1)) +
  labs(title = "Viz. 9: Traffic Incident Counts From 2017 To 2020")
```



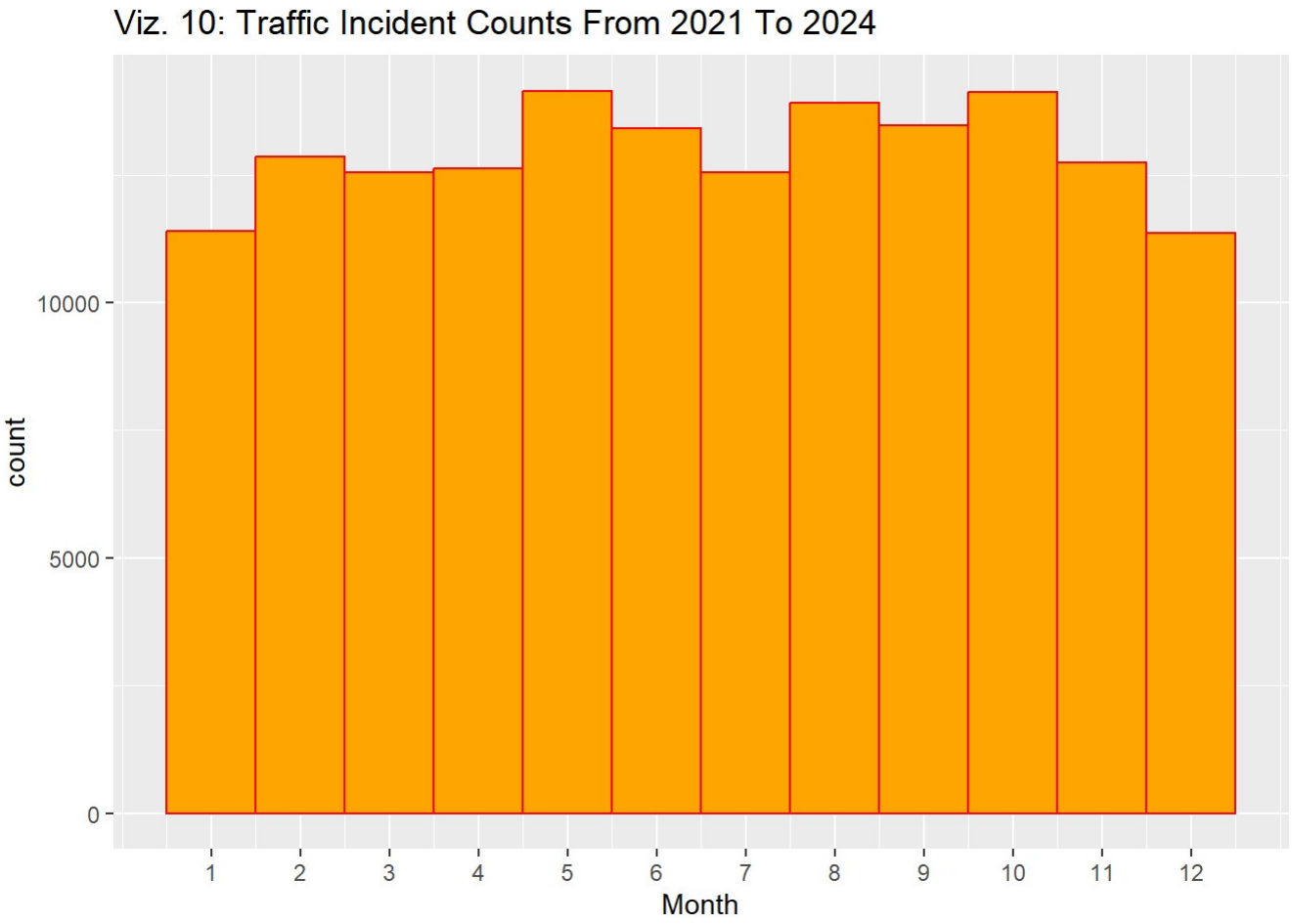

```
ti_df_formatted |>
  filter(Year <= 2020) |>
  count(Month) |>
  summary()
```

```
##      Month             n
##  Min.   : 1.00   Min.   :14650
##  1st Qu.: 3.75   1st Qu.:15132
##  Median : 6.50   Median :15628
##  Mean   : 6.50   Mean   :16676
##  3rd Qu.: 9.25   3rd Qu.:17261
##  Max.   :12.00   Max.   :21184
```

When we separate the the month against traffic incident count into two different segments based on year, we find that between the years of 2017 and 2020 the trend is similar to the total trend. There is a slightly more pronounced increase in traffic incidents in the winter months. The median and mean

number of traffic incidents are 15,628 and 16,676, respectively with a Min of 14,650 and a Max 21,184.

```
ti_df_formatted |>
  filter(Year > 2020) |>
  ggplot(aes(x = Month)) +
  geom_histogram(binwidth = 1, center = 0, color = "red", fill = "orange") +
  scale_x_continuous(breaks = seq(1, 12, by = 1), labels = seq(1, 12, by = 1)) +
  labs(title = "Viz. 10: Traffic Incident Counts From 2021 To 2024")
```




```
ti_df_formatted |>
  filter(Year > 2020) |>
  count(Month) |>
  summary()
```

```
##      Month             n
##  Min.   : 1.00   Min.   :11361
##  1st Qu.: 3.75   1st Qu.:12548
##  Median : 6.50   Median :12793
##  Mean   : 6.50   Mean   :12928
##  3rd Qu.: 9.25   3rd Qu.:13574
##  Max.   :12.00   Max.   :14150
```

Between the years of 2021 and 2023 the median and mean number of traffic incidents are 12,793 and

12,928, respectively with a Min of 11,361 and a Max 14,150. This shows that the "winter effect" is not as present in years closer to the most recent years in the data set.

Discussion

There is indeed a difference in impact between Residential construction project and Commercial construction projects on Traffic Incidents. Specifically looking at the scatterplots (Vis. 3 and 4), we see that there is a stronger positive correlation between the number of Commercial construction permits and the number of Traffic Incidents than between the number of Residential construction projects and the number of Traffic Incidents, though as we expected, both seem to show positive correlations.

The visualization match our expectations as shown in Visualization 6 & 7 as the number traffic incidents increase per zipcode as they are closer in proximity to the city center. It is important to note that the zipcodes that are closer to the city center are smaller in area and higher in density. Visualizations 9 & 10 show that in earlier years (Vis. 9) the number of traffic incidents increased in the winter months while in recent years (Vis. 10) the traffic incident counts across the month are more uniform. We would like to note that shape file that we used was for the year of 2010 while the data is for the years of 2017 through 2023. If these ZIP codes were changed in their spatial construction, the results of this study may differ, though we do expect these differences to be minor.

The biggest takeaways from this EDA Study would be that 1) In past years, there was an increase in traffic incidents during the winter months compared to more recent years, 2) The density of traffic incidents increases in the ZIP codes closer in proximity to the city center, 3) There are generally more zipcodes with fewer permits when looking at either residential or commercial construction permits, and 4) There is a stronger positive correlation between the number of commercial construction permits and the number of traffic incidents than there is between the number of residential construction permits and the number of traffic incidents.

Reflection, acknowledgements, and references

One of the biggest challenges was understanding how to get the data into the correct dataframes and format to answer our research questiosn. To accomplish this task we had to understand certain syntax rules in R, particularly considering grouping, counting, and using shapefiles to map geospatial data. We learned how to better clean and visualize our data in order to answer our research questions to a satisfactory degree. Furthermore, we came to realize how important it is to clean the data before beginning our analysis and how to put certain values in useable formats, e.g., dates.

The datasets are linked above, where one can find information about the data owners. We would like to extend our gratitude to our professor, Dr. Layla Guyot, and our UGCA, Vamsi Abena, for their help and guidance with our EDA. Also, thank you to the City of Austin for making this data publicly accessible so that we can carry out this project. We, Tigris and Dan, contributed to equal parts of the project with special individual contributions to our particularly research questions.

Links and References

Construction Permit Dataset (https://data.austintexas.gov/Building-and-Development/Issued-Construction-Permits/3syk-w9eu/about_data) Traffic Incident Dataset (https://data.austintexas.gov/Transportation-and-

Mobility/Real-Time-Traffic-Incident-Reports/dx9v-zd7x/about_data) Crash Rates During Non-Construction and Construction Periods (Mangones et al., 2021) (https://www.sciencedirect.com/science/article/pii/S235214652100819X)