DATA 607 FINAL PROJECT

Shoshana Farber

THE DATA

VISION ZERO
Building a Safer City

NYC OpenData

Car collision data for the five boroughs

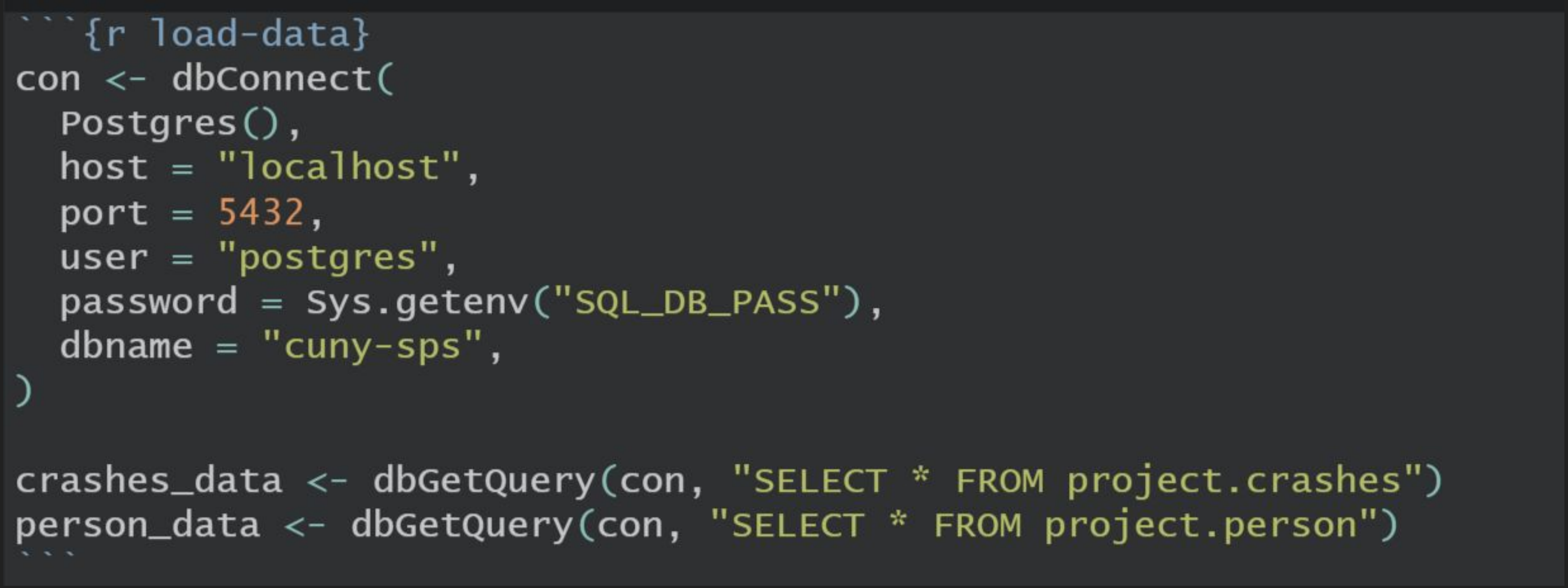
- Crashes
- Person

Data is collected by the NYPD

Source: https://www.nyc.gov/content/visionzero/pages/open-data

Loading the Data

Data stored in PostgreSQL server



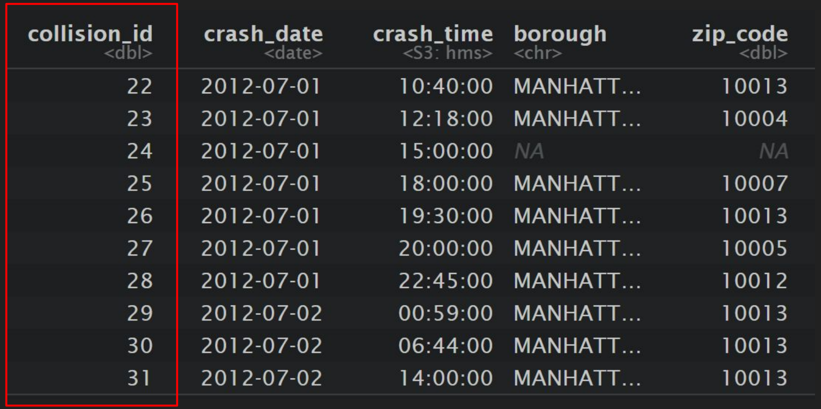
~~~~
```{r load-data}
con <- dbConnect(
  Postgres(),
  host = "localhost",
  port = 5432,
  user = "postgres",
  password = Sys.getenv("SQL_DB_PASS"),
  dbname = "cuny-sps",
)

crashes_data <- dbGetQuery(con, "SELECT * FROM project.crashes")
person_data <- dbGetQuery(con, "SELECT * FROM project.person")
```
~~~~

Crashes

- 1.988M observations - each row represents a single collision

- `contributing_factor_vehicle_x`
- `num_persons_injured`
- `num_persons_killed`

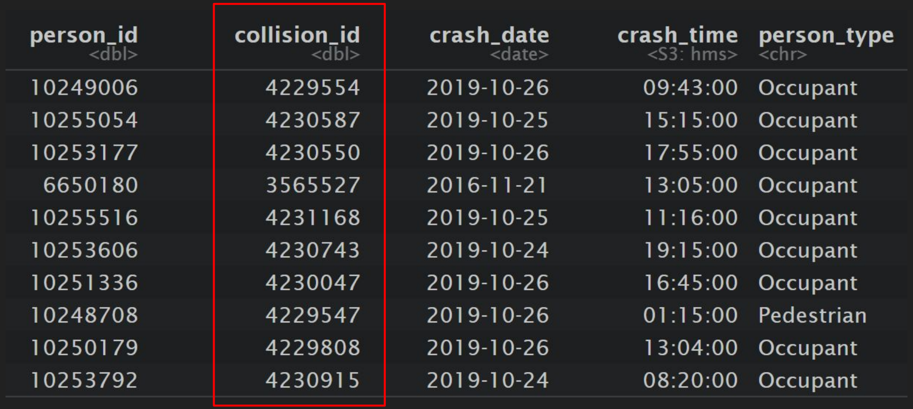


collision_id	crash_date	crash_time	borough	zip_code
<dbl>	<date>	<S3: hms>	<chr>	<dbl>
22	2012-07-01	10:40:00	MANHATT...	10013
23	2012-07-01	12:18:00	MANHATT...	10004
24	2012-07-01	15:00:00	NA	NA
25	2012-07-01	18:00:00	MANHATT...	10007
26	2012-07-01	19:30:00	MANHATT...	10013
27	2012-07-01	20:00:00	MANHATT...	10005
28	2012-07-01	22:45:00	MANHATT...	10012
29	2012-07-02	00:59:00	MANHATT...	10013
30	2012-07-02	06:44:00	MANHATT...	10013
31	2012-07-02	14:00:00	MANHATT...	10013

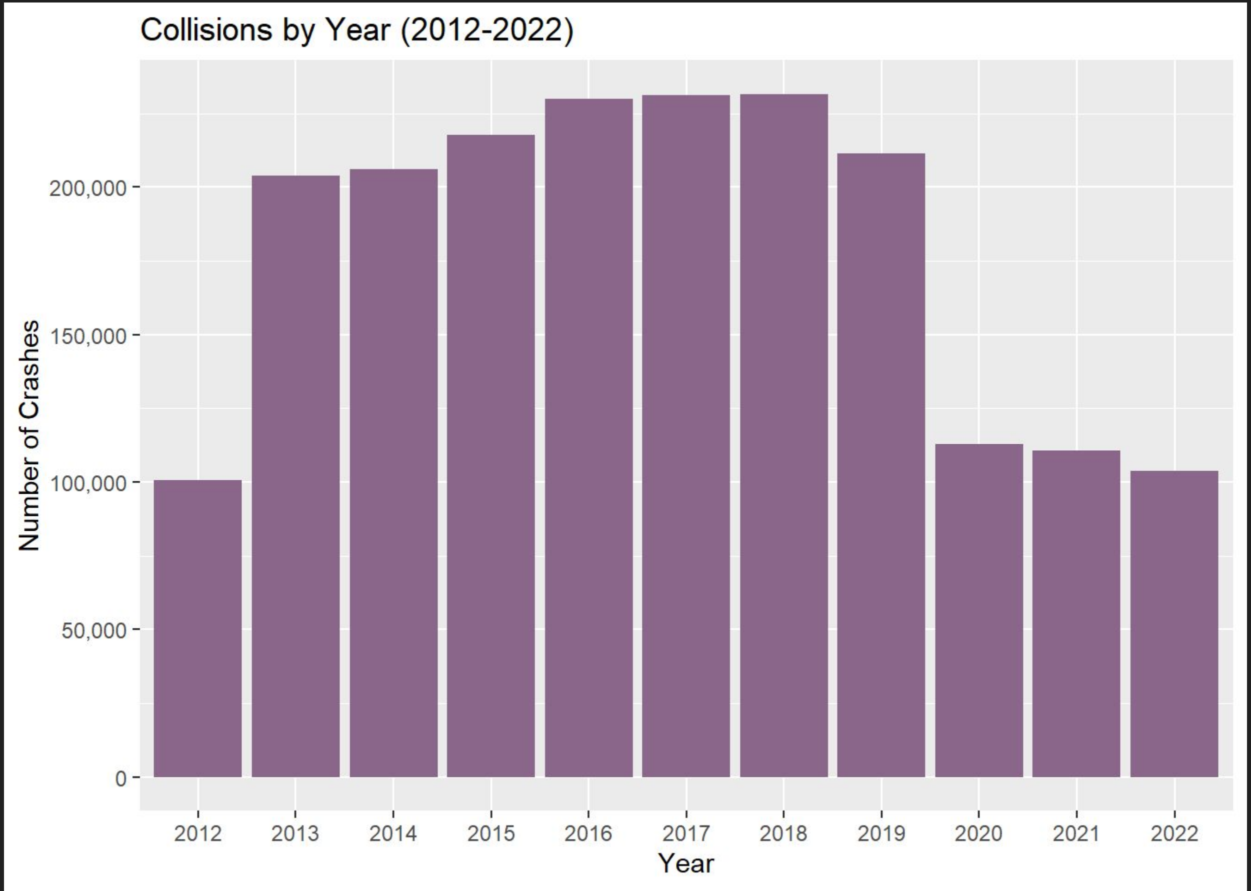
Person

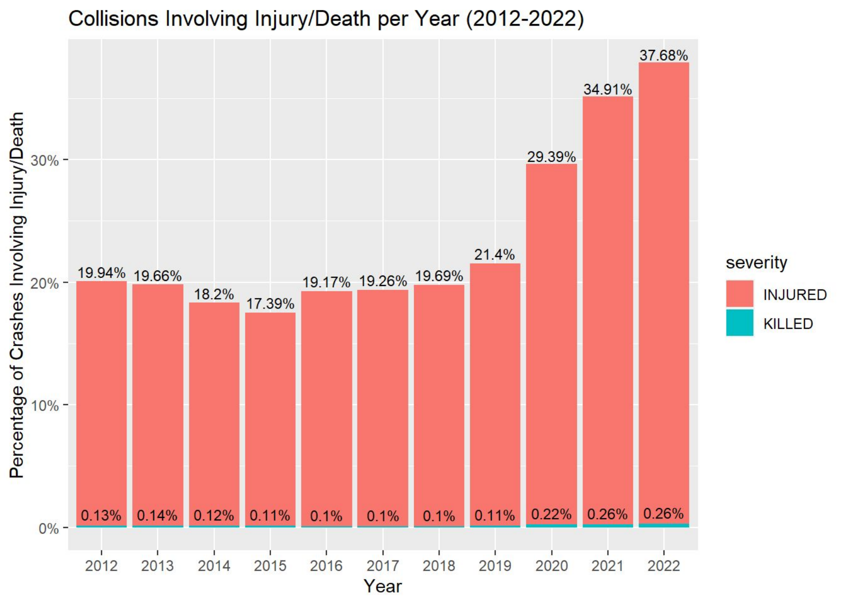
- 5.01M observations - each row represents a person involved in a collision

- `person_type`
- `position_in_vehicle`
- `person_age`
- `person_sex`

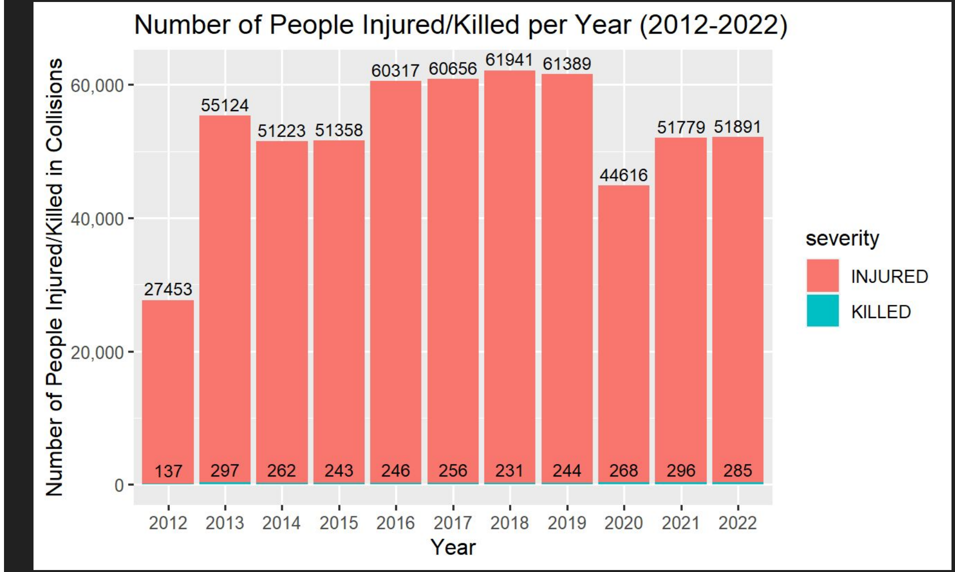


person_id	collision_id	crash_date	crash_time	person_type
<dbl>	<dbl>	<date>	<S3: hms>	<chr>
10249006	4229554	2019-10-26	09:43:00	Occupant
10255054	4230587	2019-10-25	15:15:00	Occupant
10253177	4230550	2019-10-26	17:55:00	Occupant
6650180	3565527	2016-11-21	13:05:00	Occupant
10255516	4231168	2019-10-25	11:16:00	Occupant
10253606	4230743	2019-10-24	19:15:00	Occupant
10251336	4230047	2019-10-26	16:45:00	Occupant
10248708	4229547	2019-10-26	01:15:00	Pedestrian
10250179	4229808	2019-10-26	13:04:00	Occupant
10253792	4230915	2019-10-24	08:20:00	Occupant


Collisions by Year (2012-2022)
Number of Crashes
200,000
150,000
100,000
50,000
0
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
Year



Collisions Involving Injury/Death per Year (2012-2022)

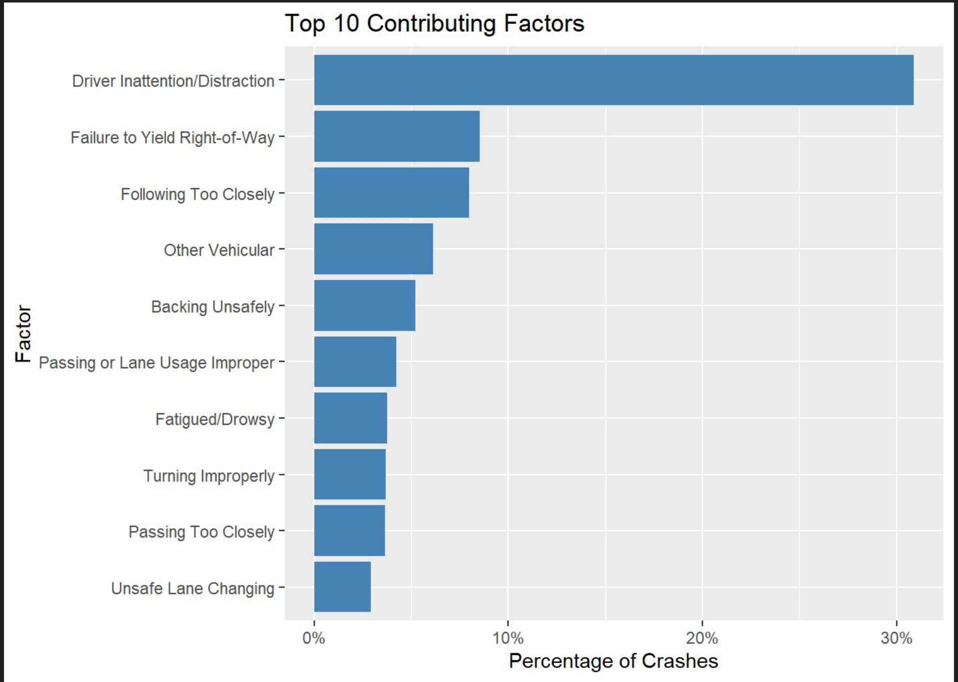


Number of People Injured/Killed per Year (2012-2022)

Main Contributing Factors

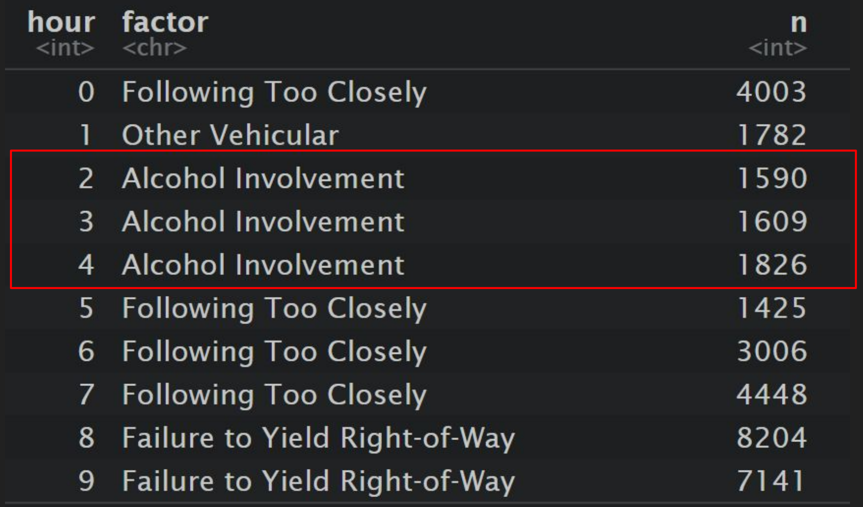
```
contributing_factors_by_crash <- crashes |>
  select(collision_id, crash_date, crash_time, borough,
c(contributing_factor_vehicle_1:contributing_factor_vehicl
e_5))

contributing_factors <- contributing_factors_by_crash |>
  pivot_longer(col =
c(contributing_factor_vehicle_1:contributing_factor_vehicl
e_5),
               names_to = "vehicle",
               values_to = "factor") |>
  mutate(factor = case_when(str_detect(factor, "Cell
Phone") ~ "Cell Phone",
                            str_detect(factor, "Drugs") ~ "Drugs",
                            str_detect(factor, "Ill") ~ "Illness",
                            str_detect(factor, "Uninvolved Vehicle")
~ "Reaction to Uninvolved Vehicle",
                            TRUE~factor)) |>
  filter(!is.na(factor), !factor %in% c("Unspecified", "1",
"80"))
```



Main Contributing Factors by Hour

~~~
```{r}
contributing_factors |>
  filter(!str_detect(factor, "Inattention/")) |>
  mutate(hour = hour(crash_time)) |>
  group_by(hour) |>
  count(factor) |>
  filter(n == max(n))
```
~~~



hour <int>	factor <chr>	n <int>
0	Following Too Closely	4003
1	Other Vehicular	1782
2	Alcohol Involvement	1590
3	Alcohol Involvement	1609
4	Alcohol Involvement	1826
5	Following Too Closely	1425
6	Following Too Closely	3006
7	Following Too Closely	4448
8	Failure to Yield Right-of-Way	8204
9	Failure to Yield Right-of-Way	7141

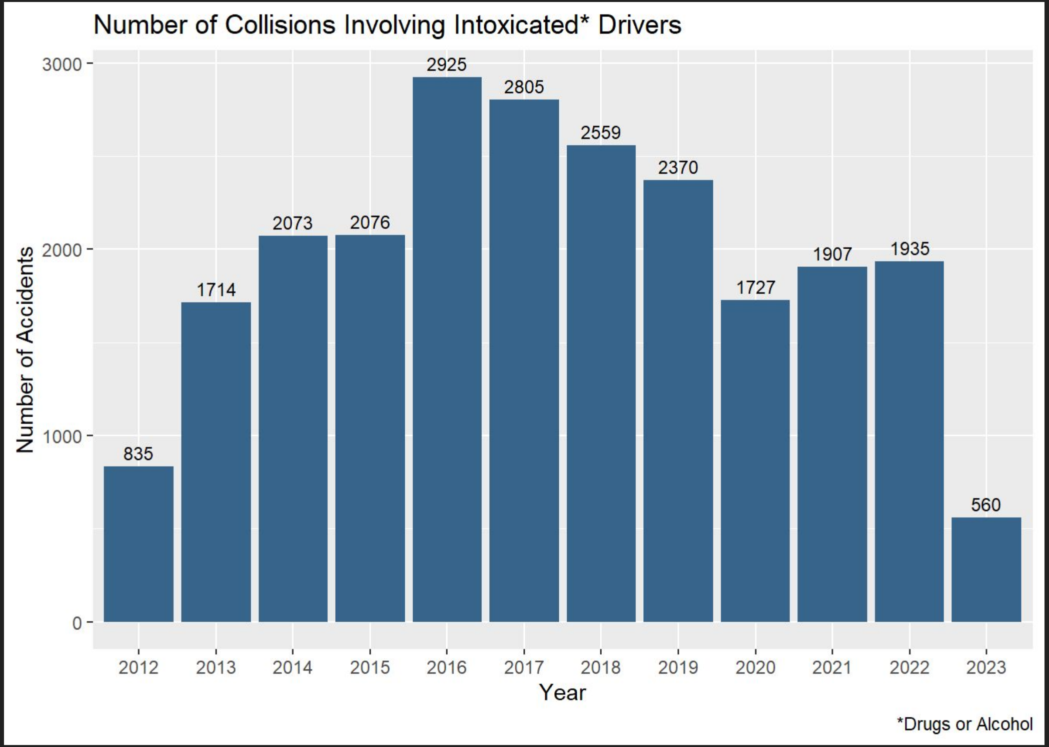
Collisions by Intoxicated Drivers

- Total number of recorded collisions due to drug/alcohol intoxication: **23,486**

- **7,439** resulted in serious injury

- **117** resulted in deaths

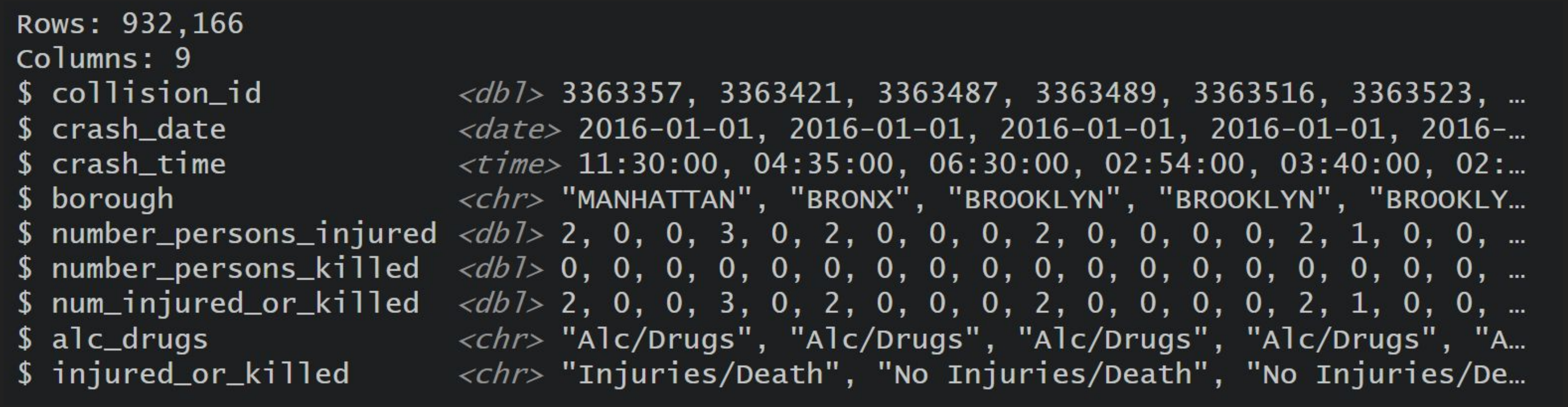
* Data for 2023 is incomplete - spans from January to April 2023



STATISTICAL ANALYSIS

Is there a significant relationship between alcohol/drug involvement and serious injuries or deaths in collisions?

alc_drug_involvement



```
Rows: 932,166
Columns: 9
$ collision_id           <dbl> 3363357, 3363421, 3363487, 3363489, 3363516, 3363523, …
$ crash_date             <date> 2016-01-01, 2016-01-01, 2016-01-01, 2016-01-01, 2016-…
$ crash_time             <time> 11:30:00, 04:35:00, 06:30:00, 02:54:00, 03:40:00, 02:…
$ borough                <chr> "MANHATTAN", "BRONX", "BROOKLYN", "BROOKLYN", "BROOKLY…
$ number_persons_injured <dbl> 2, 0, 0, 3, 0, 2, 0, 0, 0, 2, 0, 0, 0, 0, 2, 1, 0, 0, …
$ number_persons_killed  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
$ num_injured_or_killed  <dbl> 2, 0, 0, 3, 0, 2, 0, 0, 0, 2, 0, 0, 0, 0, 2, 1, 0, 0, …
$ alc_drugs              <chr> "Alc/Drugs", "Alc/Drugs", "Alc/Drugs", "Alc/Drugs", "A…
$ injured_or_killed      <chr> "Injuries/Death", "No Injuries/Death", "No Injuries/De…
```

Question: Is there a significant relationship between alcohol/drug involvement and serious injuries or deaths in collisions?

H_0 : There is no significant relationship between alcohol/drug use and serious injuries or deaths in collisions.

H_A: There is a significant relationship between alcohol/drug use and serious injuries or deaths in collisions.

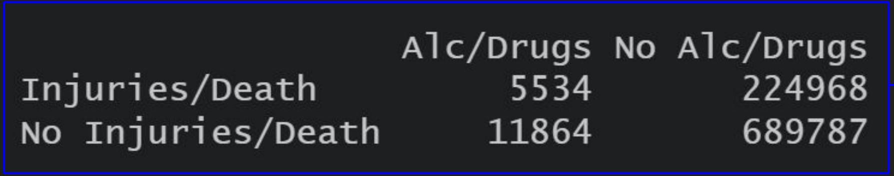
$\alpha = 0.05$

Variables (categorical):

- `alc_drugs` - whether or not there was alcohol/drug involvement (“Alc/Drugs”, “No Alc/Drugs”)
- `injured_or_killed` - whether or not there was serious injury or death resulting from this collision (“Injuries/Death”, “No Injuries/Death”)

χ^2 Test of Independence

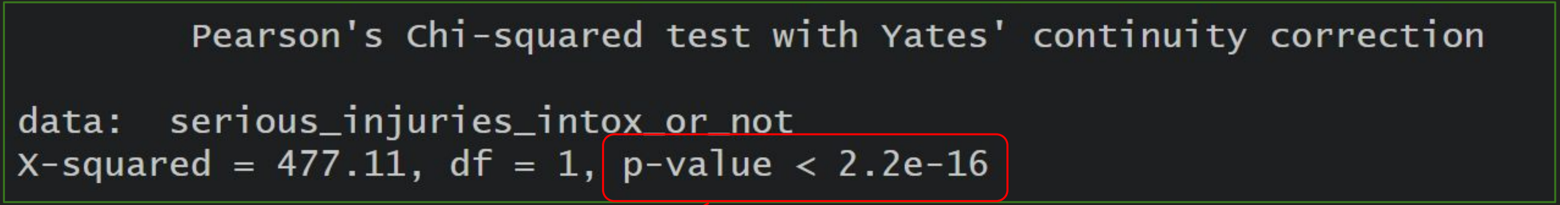
Contingency Table:



```
                   Alc/Drugs No Alc/Drugs
Injuries/Death          5534       224968
No Injuries/Death      11864       689787
```

Number of observed collisions involving or not involving death based on driver intoxication status

Results of `chisq.test()`:

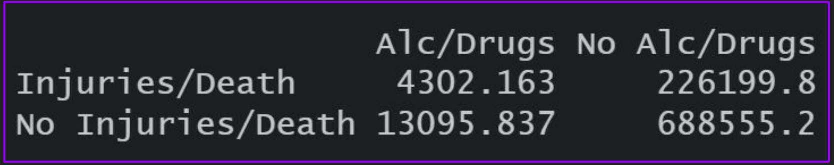


```
        Pearson's Chi-squared test with Yates' continuity correction

data:  serious_injuries_intox_or_not
X-squared = 477.11, df = 1, p-value < 2.2e-16
```

SIGNIFICANT

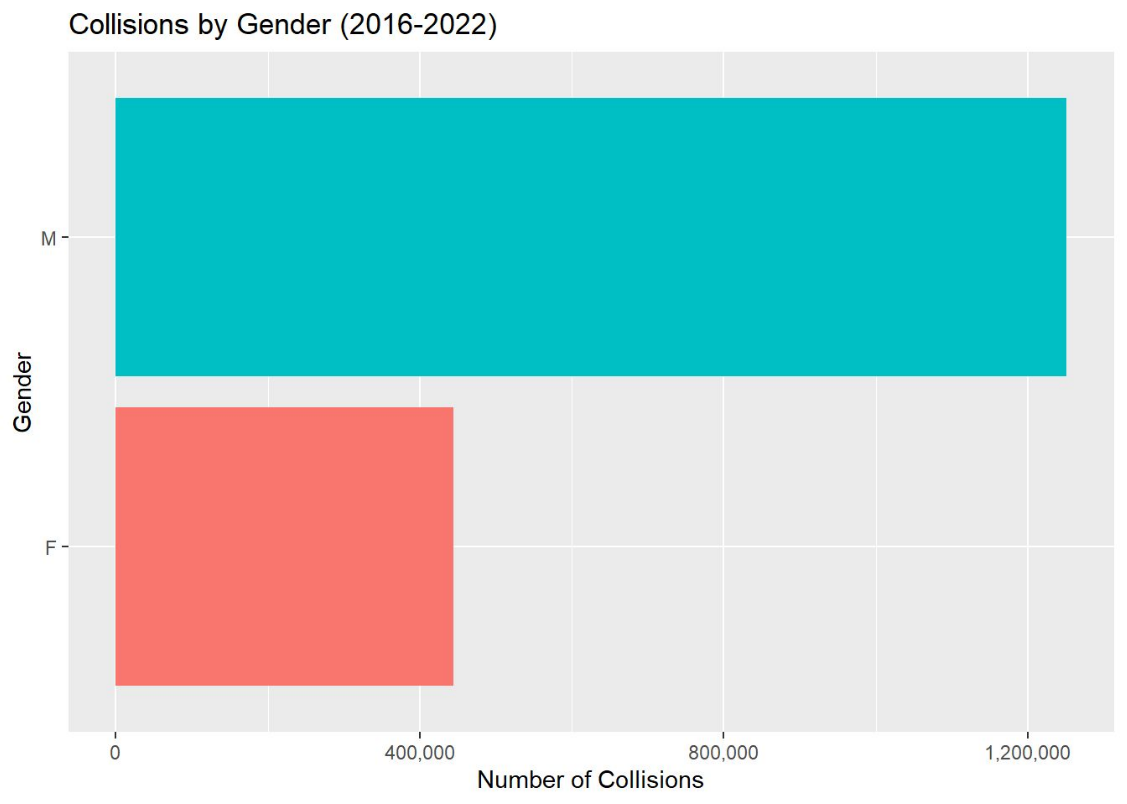
Expected Values:

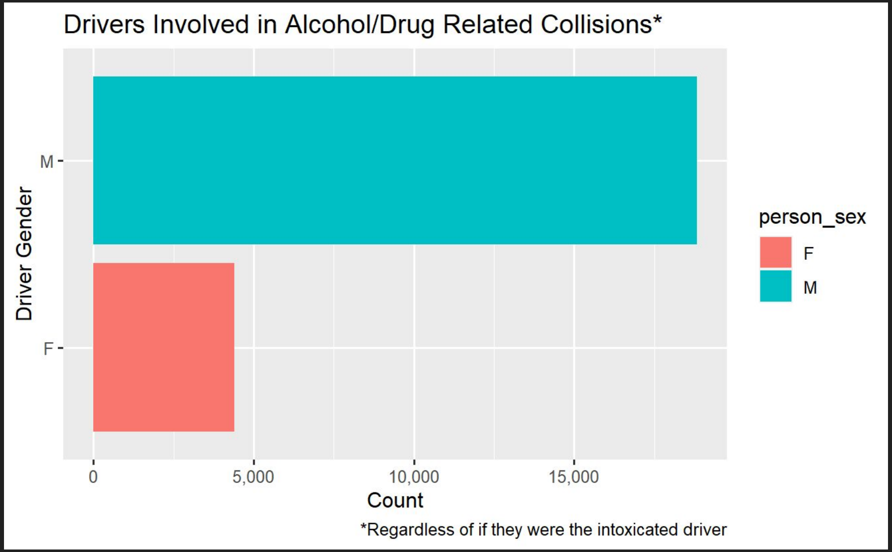


```
                  Alc/Drugs No Alc/Drugs
Injuries/Death     4302.163     226199.8
No Injuries/Death 13095.837     688555.2
```

DEMOGRAPHICS

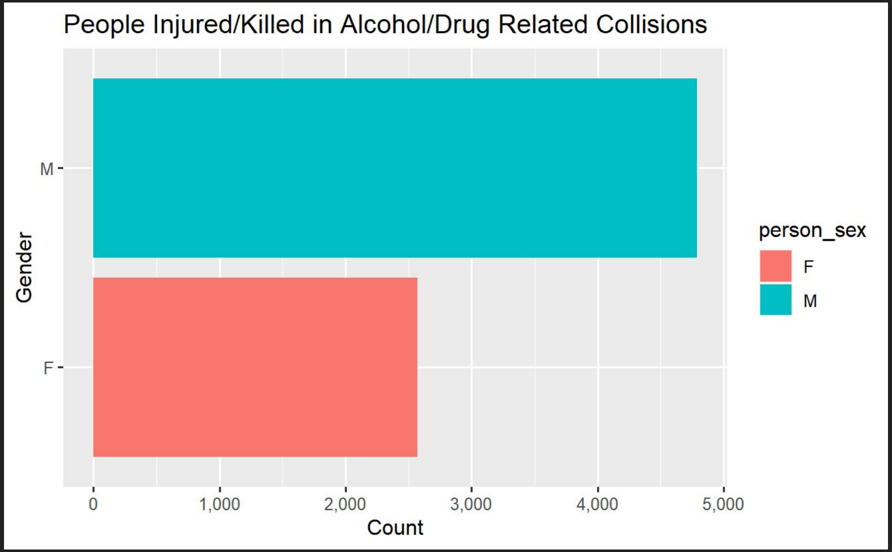
Male vs. Female Drivers





Drivers Involved in Alcohol/Drug Related Collisions*

*Regardless of if they were the intoxicated driver

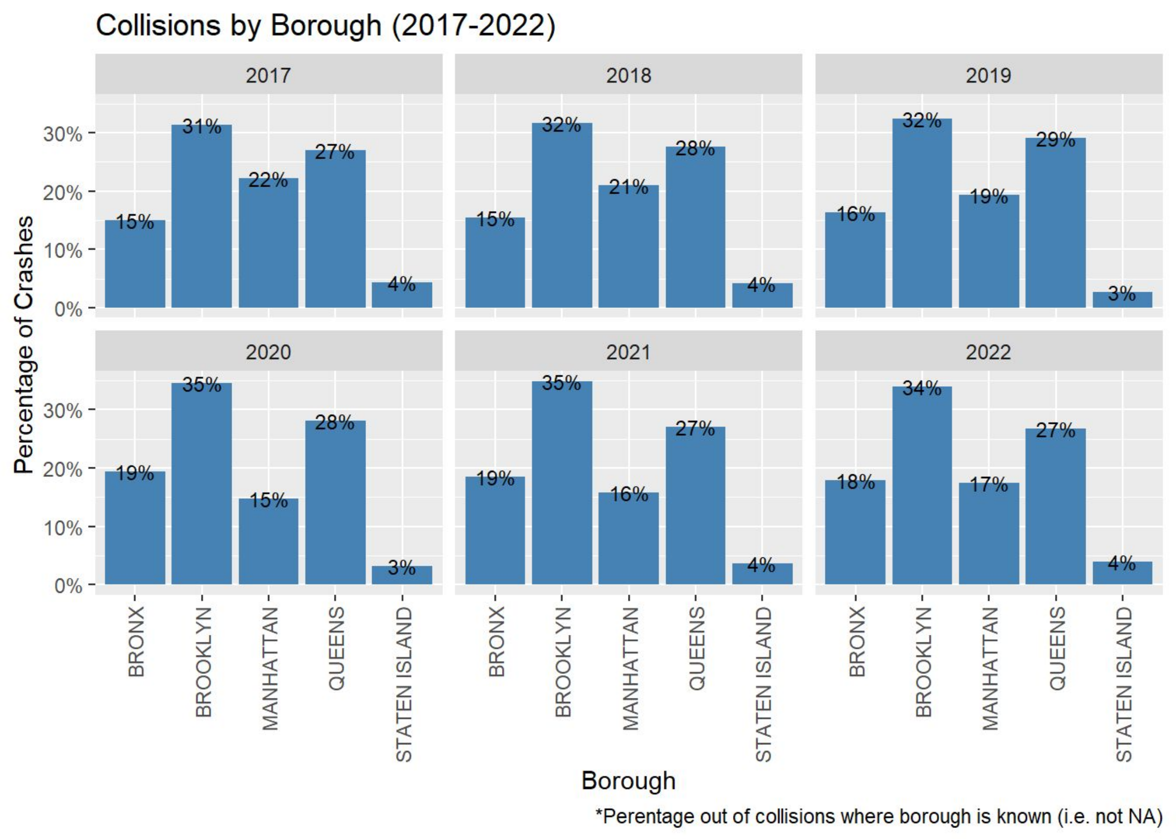


People Injured/Killed in Alcohol/Drug Related Collisions

```
person_alc_drugs <- alc_drug_involvement |>
  select(-crash_date, -crash_time, -borough) |>
  right_join(person, by = "collision_id") |>
  filter(!is.na(person_sex),
         alc_drugs == "Alc/Drugs",
         person_type == "Occupant",
         person_sex %in% c("M", "F"))
```

```
person_alc_drugs |>
  filter(person_injury %in% c("Killed", "Injured")) |>
  ggplot(aes(y = person_sex, fill = person_sex)) +
  geom_bar() +
  scale_x_continuous(label = scales::comma) +
  labs(title = "People Injured/Killed in Alcohol/Drug
Related Collisions", x = "Count", y = "Gender")
```

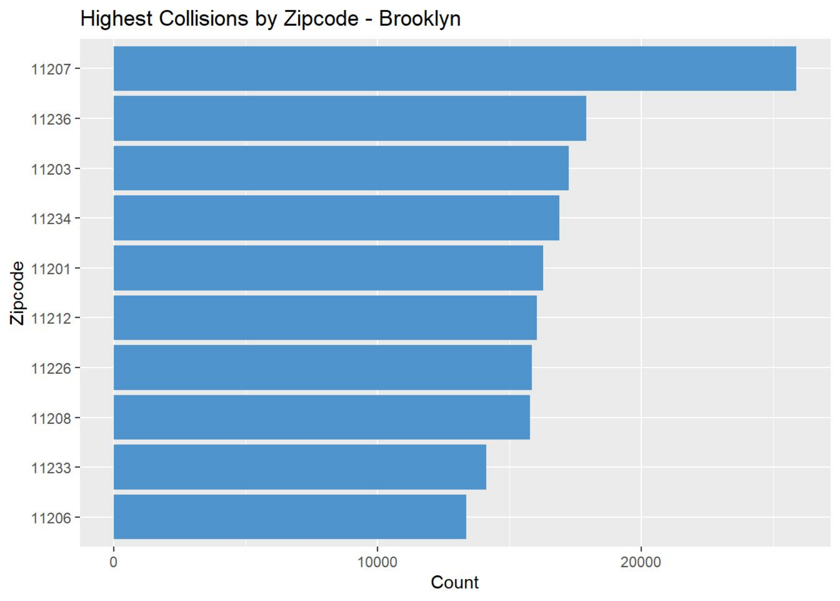
Collisions by Borough



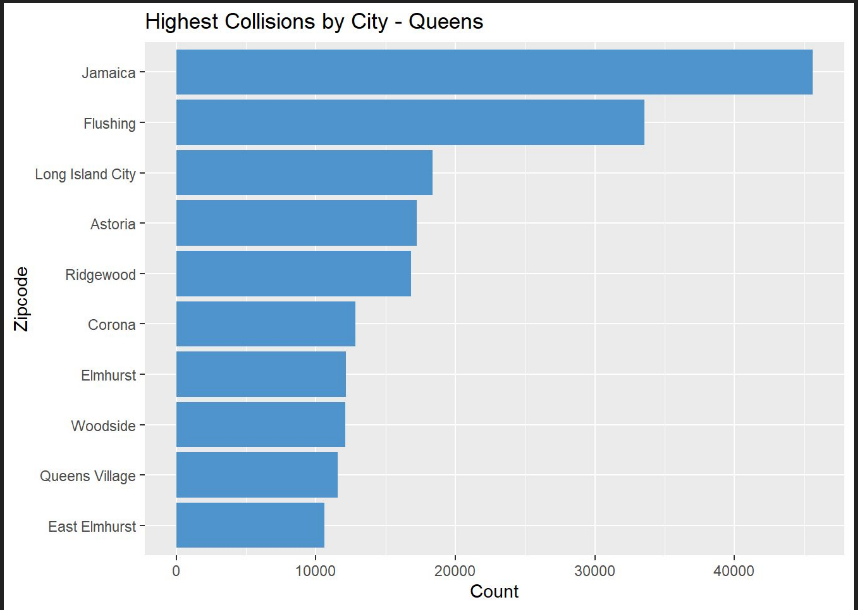
*Perentage out of collisions where borough is known (i.e. not NA)

Highest Collisions by City and Zipcode

Brooklyn - Zipcode



Queens - City



Conclusions

- Driver inattention/distraction is the leading cause for car collisions.
- Alcohol abuse is the second leading cause for collisions between the hours of 2-4 AM.
- There is a significant relationship between alcohol/drug involvement and collision severity (i.e. collisions resulting in injury/death).
- More male drivers are involved in motor vehicle collisions than female drivers.
- The highest percentage of collisions annually occurs in Brooklyn and Queens, accounting for 31-35% and 27-29% of collisions respectively.

Limitations

- Main limitation - **crashes** data set has columns for the contributing factor for each vehicle involved in a collision. However, the person data set does not indicate which vehicle number they were assigned to in a crash, so there is no way to match up the attributes of the driver to their contributing factor for the collision.