R Notebook

```
library(readr)
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──
```

```
## ✔ ggplot2 3.3.5     ✔ dplyr   1.0.8
## ✔ tibble  3.1.6     ✔ stringr 1.4.0
## ✔ tidyr   1.2.0     ✔ forcats 0.5.1
## ✔ purrr   0.3.4
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library(plotrix)
```

```
crime <- read_csv('Downloads/Crime.csv')
```

```
## Rows: 361027 Columns: 20
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (8): OFFENSE_CODE, OFFENSE_TYPE_ID, OFFENSE_CATEGORY_ID, FIRST_OCCURREN...
## dbl (12): incident_id, offense_id, OFFENSE_CODE_EXTENSION, GEO_X, GEO_Y, GEO...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
dim(crime)
```

```
## [1] 361027     20
```

```
str(crime)
```

```
## spec_tbl_df [361,027 × 20] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ incident_id          : num [1:361027] 2.02e+10 2.02e+07 2.02e+07 2.02e+07 2.02e+07 ...
##  $ offense_id           : num [1:361027] 2.02e+16 2.02e+13 2.02e+13 2.02e+13 2.02e+13 ...
##  $ OFFENSE_CODE         : chr [1:361027] "2999" "2999" "2999" "2999" ...
##  $ OFFENSE_CODE_EXTENSION: num [1:361027] 0 0 0 0 0 0 0 0 0 0 ...
##  $ OFFENSE_TYPE_ID      : chr [1:361027] "criminal-mischief-other" "criminal-mischief-other" "criminal-mischief-other" "criminal-mischief-other" ...
##  $ OFFENSE_CATEGORY_ID  : chr [1:361027] "public-disorder" "public-disorder" "public-disorder" "public-disorder" ...
##  $ FIRST_OCCURRENCE_DATE : chr [1:361027] "1/4/2022 11:30:00 AM" "1/3/2022 6:45:00 AM" "1/3/2022 1:00:00 AM" "1/3/2022 7:47:00 PM" ...
##  $ LAST_OCCURRENCE_DATE  : chr [1:361027] "1/4/2022 12:00:00 PM" NA NA NA ...
##  $ REPORTED_DATE         : chr [1:361027] "1/4/2022 8:36:00 PM" "1/3/2022 11:01:00 AM" "1/3/2022 6:11:00 AM" "1/3/2022 9:12:00 PM" ...
##  $ INCIDENT_ADDRESS      : chr [1:361027] "128 S CANOSA CT" "650 15TH ST" "919 E COLFAX AVE" "2345 W ALAMEDA AVE" ...
##  $ GEO_X                 : num [1:361027] 3135366 3142454 3147484 3136478 3169237 ...
##  $ GEO_Y                 : num [1:361027] 1685410 1696151 1694898 1684414 1705800 ...
##  $ GEO_LON               : num [1:361027] -105 -105 -105 -105 -105 ...
##  $ GEO_LAT               : num [1:361027] 39.7 39.7 39.7 39.7 39.8 ...
##  $ DISTRICT_ID           : num [1:361027] 4 6 6 4 5 6 3 6 3 3 ...
##  $ PRECINCT_ID           : num [1:361027] 411 611 621 411 512 621 312 623 311 123 ...
##  $ NEIGHBORHOOD_ID       : chr [1:361027] "valverde" "cbd" "north-capitol-hill" "valverde" ...
##  $ IS_CRIME              : num [1:361027] 1 1 1 1 1 1 1 1 1 1 ...
##  $ IS_TRAFFIC            : num [1:361027] 0 0 0 0 0 0 0 0 0 0 ...
##  $ VICTIM_COUNT          : num [1:361027] 1 1 1 1 1 1 1 1 1 1 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   incident_id = col_double(),
##   ..   offense_id = col_double(),
##   ..   OFFENSE_CODE = col_character(),
##   ..   OFFENSE_CODE_EXTENSION = col_double(),
##   ..   OFFENSE_TYPE_ID = col_character(),
##   ..   OFFENSE_CATEGORY_ID = col_character(),
##   ..   FIRST_OCCURRENCE_DATE = col_character(),
##   ..   LAST_OCCURRENCE_DATE = col_character(),
##   ..   REPORTED_DATE = col_character(),
##   ..   INCIDENT_ADDRESS = col_character(),
##   ..   GEO_X = col_double(),
##   ..   GEO_Y = col_double(),
##   ..   GEO_LON = col_double(),
##   ..   GEO_LAT = col_double(),
##   ..   DISTRICT_ID = col_double(),
##   ..   PRECINCT_ID = col_double(),
##   ..   NEIGHBORHOOD_ID = col_character(),
##   ..   IS_CRIME = col_double(),
##   ..   IS_TRAFFIC = col_double(),
##   ..   VICTIM_COUNT = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
summary(crime)
```

```
##   incident_id         offense_id        OFFENSE_CODE
##  Min.   :2.020e+04   Min.   :2.020e+10   Length:361027
##  1st Qu.:2.018e+09   1st Qu.:2.018e+15   Class :character
##  Median :2.020e+09   Median :2.020e+15   Mode  :character
##  Mean   :5.677e+09   Mean   :5.677e+15
##  3rd Qu.:2.022e+09   3rd Qu.:2.022e+15
##  Max.   :2.021e+12   Max.   :2.021e+18
##
##  OFFENSE_CODE_EXTENSION OFFENSE_TYPE_ID    OFFENSE_CATEGORY_ID
##  Min.   :0.0000         Length:361027      Length:361027
##  1st Qu.:0.0000         Class :character   Class :character
##  Median :0.0000         Mode  :character   Mode  :character
##  Mean   :0.2633
##  3rd Qu.:0.0000
##  Max.   :5.0000
##
##  FIRST_OCCURRENCE_DATE LAST_OCCURRENCE_DATE REPORTED_DATE
##  Length:361027         Length:361027       Length:361027
##  Class :character      Class :character    Class :character
##  Mode  :character      Mode  :character    Mode  :character
##
##
##
##
##  INCIDENT_ADDRESS       GEO_X             GEO_Y            GEO_LON
##  Length:361027      Min.   :      1   Min.   :      1   Min.   :-115.5
##  Class :character   1st Qu.: 3139841   1st Qu.: 1683183   1st Qu.:-105.0
##  Mode  :character   Median : 3146086   Median : 1694802   Median :-105.0
##                     Mean   : 3156584   Mean   : 1693516   Mean   :-104.9
##                     3rd Qu.: 3164305   3rd Qu.: 1701690   3rd Qu.:-104.9
##                     Max.   :40674766   Max.   :10890452   Max.   :   0.0
##                     NA's   :4738       NA's   :4738       NA's   :5321
##     GEO_LAT       DISTRICT_ID    PRECINCT_ID    NEIGHBORHOOD_ID       IS_CRIME
##  Min.   : 0.00   Min.   :1.00   Min.   :111.0   Length:361027      Min.   :1
##  1st Qu.:39.71   1st Qu.:2.00   1st Qu.:222.0   Class :character   1st Qu.:1
##  Median :39.74   Median :3.00   Median :324.0   Mode  :character   Median :1
##  Mean   :39.73   Mean   :3.65   Mean   :382.9                      Mean   :1
##  3rd Qu.:39.76   3rd Qu.:5.00   3rd Qu.:523.0                      3rd Qu.:1
##  Max.   :39.90   Max.   :7.00   Max.   :759.0                      Max.   :1
##  NA's   :5321    NA's   :585    NA's   :585
##    IS_TRAFFIC   VICTIM_COUNT
##  Min.   :0    Min.   : 1.000
##  1st Qu.:0    1st Qu.: 1.000
##  Median :0    Median : 1.000
##  Mean   :0    Mean   : 1.019
##  3rd Qu.:0    3rd Qu.: 1.000
##  Max.   :0    Max.   :32.000
##
```

```
#Remove variables not in use for analysis
crime <- select(crime, c('incident_id', 'DISTRICT_ID'))
head(crime)
```

incident_id <dbl>	DISTRICT_ID <dbl>
20226000193	4
20223319	6
20223093	6
20224000	4
20223956	5
20223903	6

6 rows

```
#check for null values
crime[crime == "?"] <- NA
colSums(is.na(crime))
```

```
## incident_id DISTRICT_ID
##           0         585
```

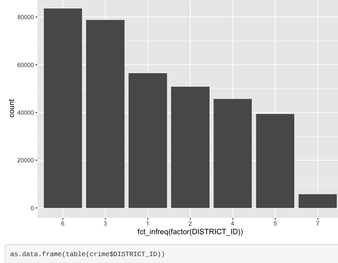
```
crime <- na.omit(crime)
```

```
#check if null values were omitted
colSums(is.na(crime))
```

```
## incident_id DISTRICT_ID
##           0           0
```

```
write.csv(crime, "crime_capstone.csv", row.names = TRUE)
```

```
graph <- ggplot(filter(crime), aes(fct_infreq(factor(DISTRICT_ID)))) + geom_bar()
graph
```




```
as.data.frame(table(crime$DISTRICT_ID))
```

Var1 <fct>	Freq <int>
1	56478
2	50802
3	78771
4	45707
5	39354
6	83586
7	5744

7 rows

```
slices <- c(56478, 50802, 78771, 45707, 39354, 83586, 5744)
labels <- c("D1", "D2", "D3", "D4", "D5", "D6", "D7")
percent <- round(slices/sum(slices)*100)
labels <- paste(labels, percent)
labels <- paste(labels, "%", sep="")
pie(slices, labels = labels, main = "Pie Chart of Crime", col=rainbow(length(labels)))
```