Data Science as a Field Project

J. Kolb

3/28/2022

Importing Data

We begin by going to the website https://catalog.data.gov/dataset and finding the dataset titled "NYPD Shooting Incident Data (Historic)". We then import the CSV file into a data set called 'NYPD_data'. We'll also use the tidyverse lobby to take advantage of the read_CSV function. Our analysis is currently focused on the date and time of the incident, so we only import those columns. The data set also includes fields for the location of the incident and the demographic information of the victims and perpetrators. While those might be of interest in the future, we are keeping our scope narrow and focused only on seasonality for the moment.

```
## Warning: package 'tidyverse' was built under R version 4.1.3
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.3     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.0     v forcats 0.5.1
```

```
## Warning: package 'ggplot2' was built under R version 4.1.3
```

```
## Warning: package 'forcats' was built under R version 4.1.3
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
## Warning: package 'lubridate' was built under R version 4.1.3
```

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

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

```
## Rows: 23585 Columns: 3
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): OCCUR_DATE
## dbl  (1): INCIDENT_KEY
## time (1): OCCUR_TIME

##
## 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.
```

Tidying Data

The data doesn't have any null or missing values in these columns, so that isn't an issue with our current data. However, we would like to tidy up the data by categorizing them into seasons (Spring, Summer, Autumn, and Winter) and time of day (Morning, Afternoon, Evening, and Night). The seasons will be based on the meteorological seasons and the time of day will be based on midnight and noon.

```
NYPD_data = NYPD_data %>%
  mutate(
    OCCUR_DATE = as.Date(OCCUR_DATE,"%m/%d/%Y"),
    SEASON = case_when(month(OCCUR_DATE) %in% 6:8 ~ "Summer",
                       month(OCCUR_DATE) %in% 9:11 ~ "Autumn",
                       month(OCCUR_DATE) %in% c(1,2,12) ~ "Winter",
                       month(OCCUR_DATE) %in% 3:5 ~ "Spring"),
    DAY_PART = case_when(hour(OCCUR_TIME) %in% 0:5 ~ "Night",
                         hour(OCCUR_TIME) %in% 6:11 ~ "Morning",
                         hour(OCCUR_TIME) %in% 12:17 ~ "Afternoon",
                         hour(OCCUR_TIME) %in% 18:23 ~ "Evening")
  )
```

Here is the final result of our changes summarized.

```
summary(NYPD_data)
```

```
##   INCIDENT_KEY         OCCUR_DATE          OCCUR_TIME          SEASON
##  Min.   :  9953245   Min.   :2006-01-01   Length:23585      Length:23585
##  1st Qu.: 55322804   1st Qu.:2008-12-31   Class1:hms        Class :character
##  Median : 83435362   Median :2012-02-27   Class2:difftime   Mode  :character
##  Mean   :102280741   Mean   :2012-10-05   Mode  :numeric
##  3rd Qu.:150911774   3rd Qu.:2016-03-02
##  Max.   :230611229   Max.   :2020-12-31
##    DAY_PART
##  Length:23585
##  Class :character
##  Mode  :character
##
##
##
```

Visualization

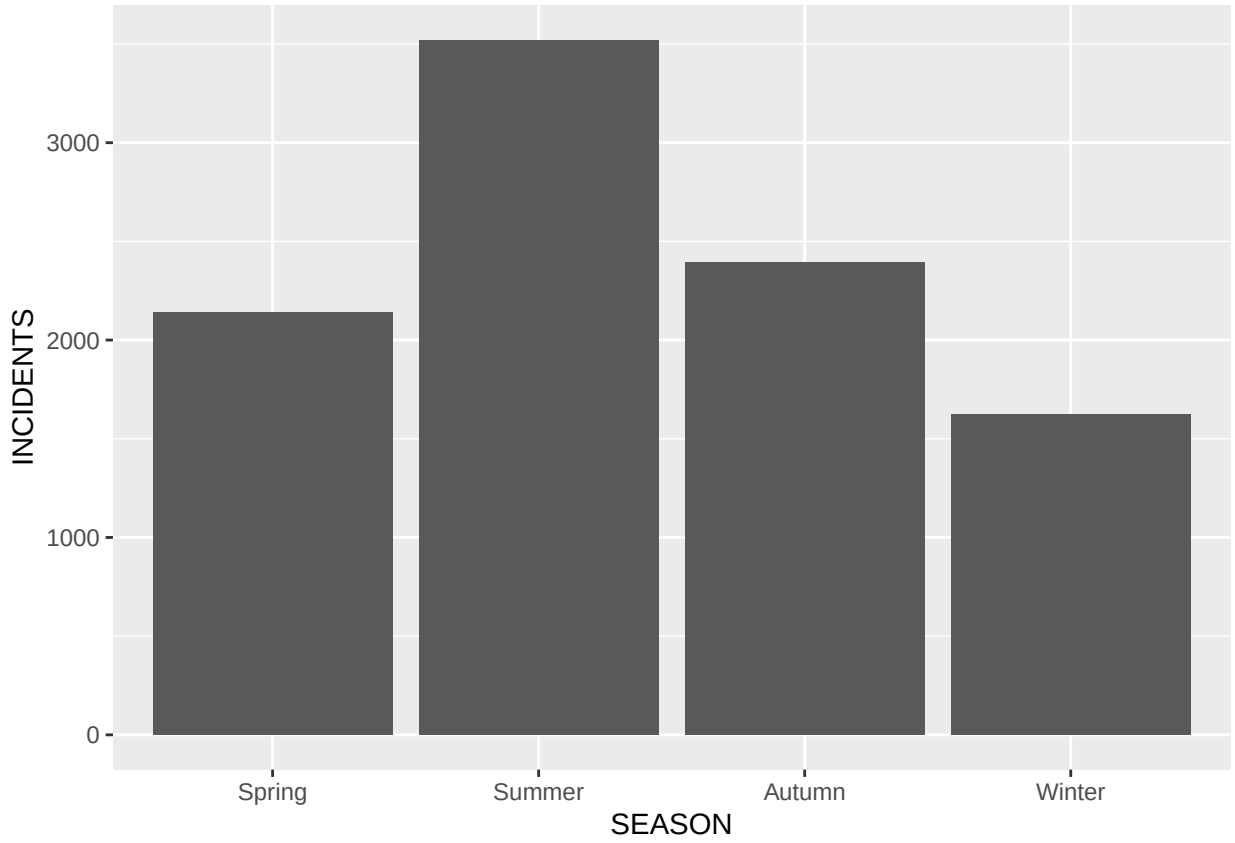
Next, we would like to take a look at the data visually and provide some analysis. Our questions at the moment are "Does the season affect the prevalence of shooting incidents" and "Does the season affect when

shooting incidents occur". We plot both questions below.

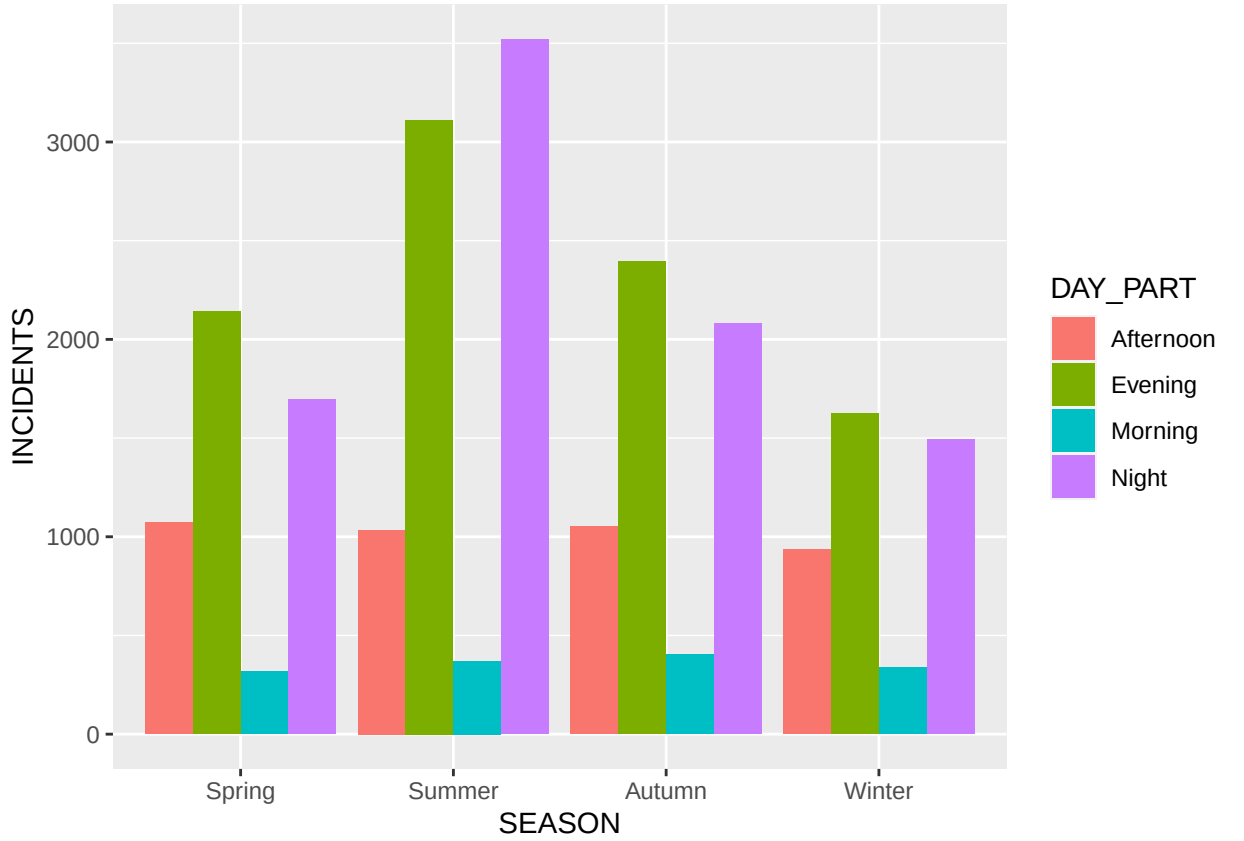
```
library(dplyr)
library(ggplot2)
Incident_by_time = NYPD_data %>%
  group_by(SEASON, DAY_PART) %>%
  summarize(INCIDENTS = n())
```

```
## `summarise()` has grouped output by 'SEASON'. You can override using the `.groups` argument.
```

```
ggplot(Incident_by_time, aes(y=INCIDENTS, x=SEASON)) +
    geom_bar(position="dodge", stat="identity") +
  scale_x_discrete(limits=c('Spring','Summer','Autumn','Winter'))
```




```
ggplot(Incident_by_time, aes(fill=DAY_PART, y=INCIDENTS, x=SEASON)) +
    geom_bar(position="dodge", stat="identity") +
  scale_x_discrete(limits=c('Spring','Summer','Autumn','Winter'))
```



As we can clearly see, there is an effect on how shooting incidents occur based on the season. Additionally, while the absolute number of shooting incidents that occur in the morning and afternoon remain about the same, the number of incidents in the evening and night both in absolute terms and as a percentage of total incidents changes based on the season. In particular, the number of shooting incidents that happen after midnight dramatically increases in the summer season.

Analysis

However, we would like to rely on more than intuition to analyzing our data. As such, I'll utilize Pearson's chi-squared test to evaluate likely it is that this difference in category arose by chance. We compare the actual results against what we'd expect if the two categories were independent. The resulting statistic and p-value will let us test our null hypothesis, that Season and Time of Day are independent.

```
library(dplyr)
library(tidyr)
Incident_wide = pivot_wider(Incident_by_time, id_cols = SEASON, names_from = DAY_PART, values_from = IN
Incidents = select(ungroup(Incident_wide),c('Night','Evening','Afternoon','Morning'))
rownames(Incidents) = c('Autumn','Spring','Summer','Winter')
```

```
## Warning: Setting row names on a tibble is deprecated.
```

```
chisq <- chisq.test(Incidents)
chisq
```

```
##
```

```
##  Pearson's Chi-squared test
##
## data:  Incidents
## X-squared = 372.35, df = 9, p-value < 2.2e-16
```

```
chisq$observed
```

```
##        Night Evening Afternoon Morning
## Autumn  2080    2396      1053     404
## Spring  1694    2140      1071     318
## Summer  3520    3112      1034     371
## Winter  1493    1626       937     336
```

```
round(chisq$expected,2)
```

```
##          Night Evening Afternoon Morning
## Autumn 2210.44 2332.95   1030.13  359.48
## Spring 1945.92 2053.77    906.86  316.46
## Summer 2994.32 3160.28   1395.44  486.96
## Winter 1636.32 1727.00    762.57  266.11
```

```
round(chisq$residuals, 3)
```

```
##         Night Evening Afternoon Morning
## Autumn -2.774   1.305     0.713   2.348
## Spring -5.711   1.903     5.451   0.087
## Summer  9.607  -0.859    -9.676  -5.255
## Winter -3.543  -2.430     6.317   4.284
```

Looking at the results, we can say with near absolute certainty that this distribution did not occur by chance. The X-Squared value for this distribution is 372, with a p-value that is effectively 0. We can see the reason for this by comparing the expected and observed values. If time of day and season were independent, we'd expect to observe about 2,994 incidents on Summer Nights versus the actual observation of 3,520.

Furthermore, we can split apart which cells are contributing the most to the X-Squared value by calculating the residuals. A large positive value indicates a strong positive correlation, while a large negative value indicates a negative correlation. As a result we can see a strong positive correlation between summer shooting incidents at night, and between afternoon shooting incidents in the winter and spring.

Conclusion

Based on our analysis, we can conclude that season and time of day are not independent categories. We can identify several strong positive and negative correlations. Finally, we can recommend several shifts of shooting prevention resources based on the time of year and time of day.

Bias and Other Issues

There are two main issues with biases in this analysis. First, there is the inherent bias of the data. This data is collected and provided by the NYPD. In the past, the NYPD has been accused of racial bias and unfair treatment of minority groups. This history also fed into my personal biases - my distrust of the NYPD and

the data collect caused me to view the provided data around the racial identity of victims and perpetrators with animus. As a result, I chose to investigate the more 'objective' question of when incidents occurred, rather than who was involved in the occurrence.

There are also two obvious issues with the analysis so far. First, meteorological seasons are only an approximation of what I suspect are the true underlying causes - more pleasant weather or school being out of session. For the former, a fruitful branch of study would be investigating the connection between precipitation and temperature with shooting incidents. For the latter, analyzing incidents and their relation to the academic calendar for NYC public schools and universities could also be helpful.

In addition, categorizing incidents into 6 hour day parts may be obscuring some detail. For instance, if the vast majority of summer incidents occur around midnight, the arbitrary division of evening and night at that point is obscuring that reality. A more granular view of when incidents occur could help resolve this issue.