PA 446 Homework 4

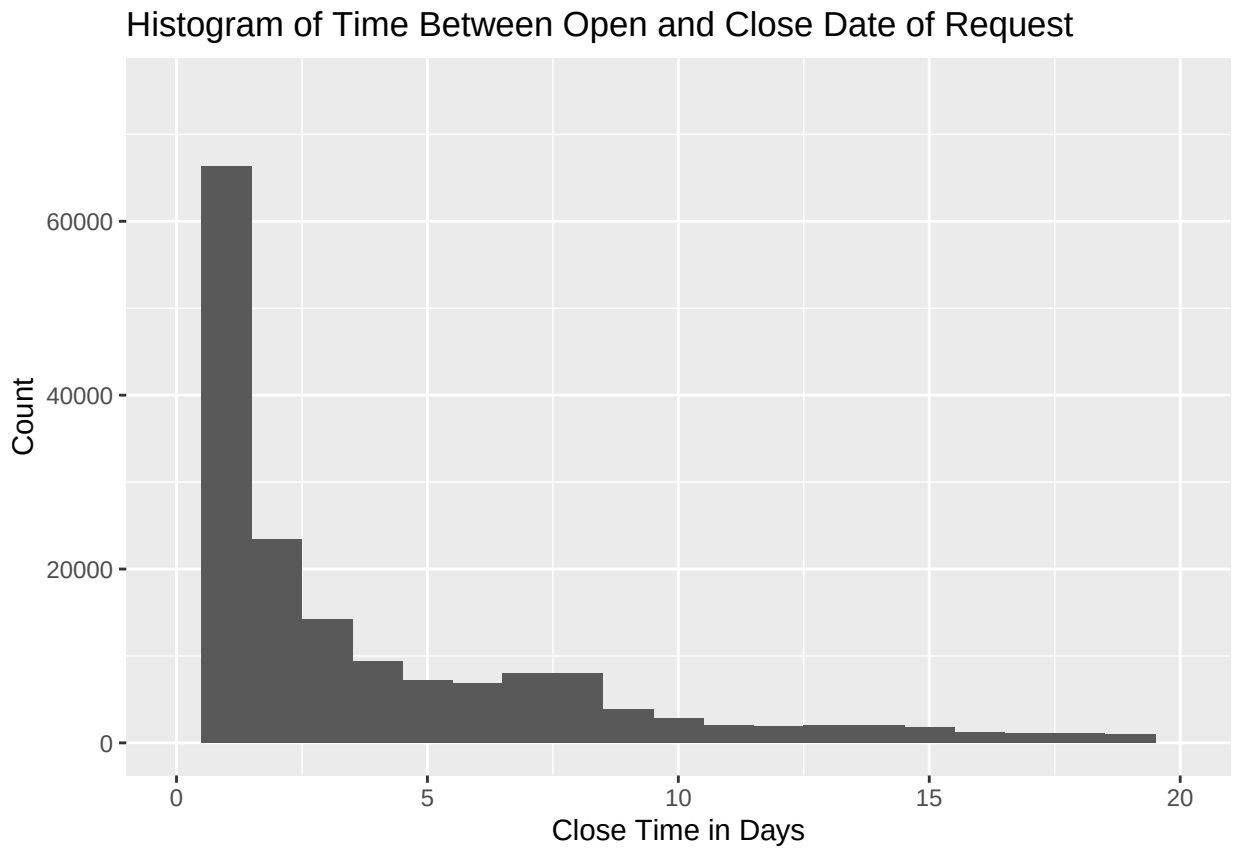
Alexis Kwan, kwan1, UID: 655727984

1. How long does it take to complete a request city-wide (hence known as close time)? Please provide 3 metrics to answer this question.
   - Why did you choose the 3 metrics?
   - What does the 3 metrics say collectively? For example, if one of your metrics is the mean, what might that number miss? What other metrics might you want to pull to supplement what the mean misses.

Suppose for a moment that we could believe that a mean statistic would represent the close time well. When we calculate that we get:

```
## # A tibble: 1 x 1
##   close_time_mean
##   <Period>
## 1 6d 13H 5M 25.7231442822376S
```

Six days sounds like a long time. However would that make sense if the data looks as it does below?

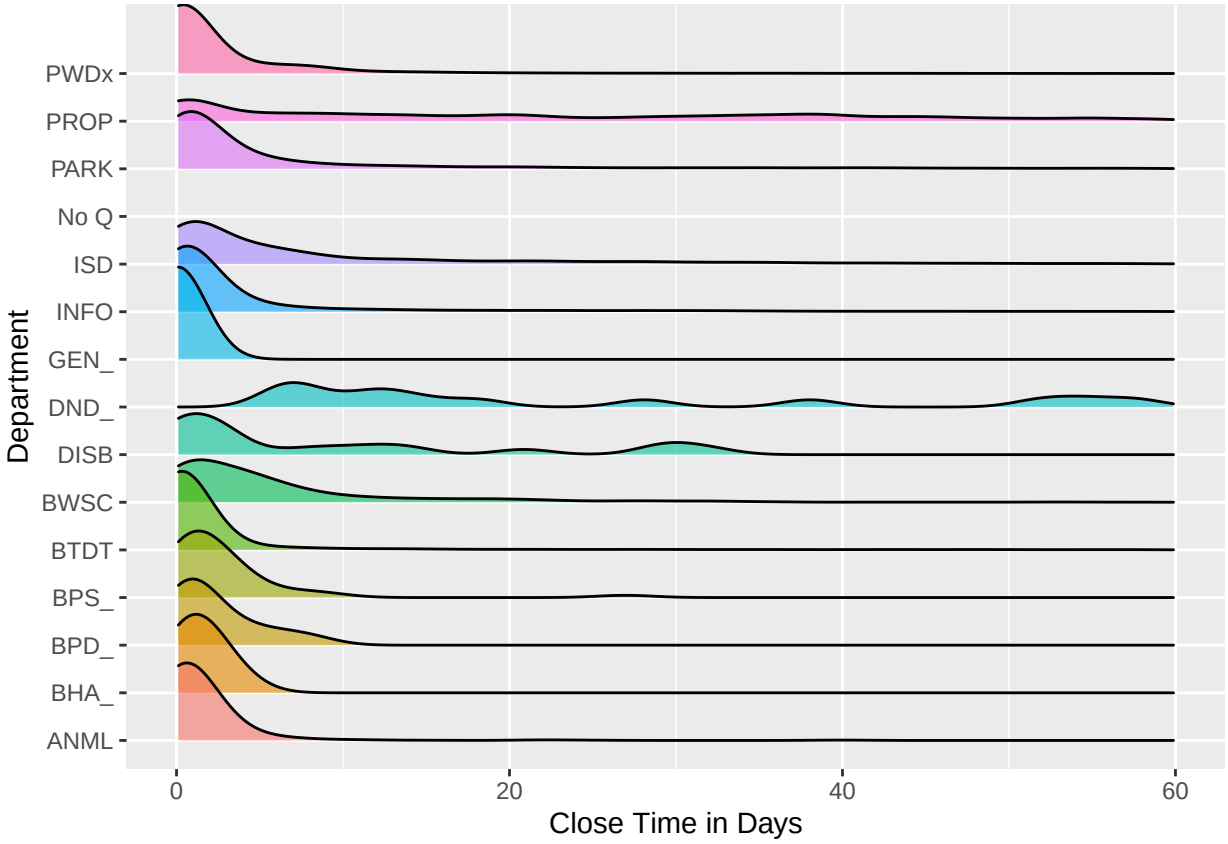


We see that the distribution for the time between when a request is opened and closed is very skewed. So if we were to measure or represent how long it would take to complete a request, we should at least use a

statistic that is robust to outliers and skew. If we take a median we get a median close time of about 8.5 hours.

```
## # A tibble: 1 x 1
##   close_time_median
##   <Period>
## 1 8H 40M 41S
```

However, if we break the close time down by the department for which the request is for, we see the distributions aren't so simple. In fact it appears that for some departments the distribution may be multi-modal, and there appears to be clear differences between departments. This likely reflects the difference in nature of the requests submitted to each department.



While median is robust to outliers, it may not capture enough information on the width of the distribution. So what if used multiple statistics instead of just one?

```
## # A tibble: 3 x 2
##   quantile close_time_hours
##   <chr>               <dbl>
## 1 50%                     3
## 2 75%                    14
## 3 90%                    20
```

Based on these quantiles we can say that there is a 75% probability that your request will be finished in 14 hours and 90% that it will be completed in 20 hours. So most likely it will be finished in a day at most, across all departments. This seems to capture the width of the distributions better.

2. Assume that the primary goal of your analysis is to ensure that all city departments have as short of a close time as possible.

- What are the 3 departments that the City should focus on? Note the language here. Yes it is vague. It will be up to you to define what "should focus on" is and then conduct your analysis.
- What are the 3 departments that have done well so far?

```
## # A tibble: 15 x 2
##    department close_time_median
##    <chr>      <Period>
##  1 PROP       38d 20H 18M 47S
##  2 DND_       28d 2H 37M 43S
##  3 ISD        3d 20H 59M 5S
##  4 BWSC       3d 13H 7M 47S
##  5 DISB       2d 23H 0M 45.5S
##  6 PARK       1d 18H 49M 24.5S
##  7 BPS_       1d 10H 45M 54S
##  8 BPD_       1d 0H 12M 58S
##  9 BHA_       23H 56M 48S
## 10 INFO       19H 19M 16S
## 11 ANML       14H 22M 27.5S
## 12 PWDx       10H 8M 43S
## 13 BTDT       4H 1M 38.5S
## 14 GEN_       35M 21S
## 15 No Q       10M 10S
```

There are several ways to minimize close time. We could minimize the largest contributors, just based on department median close time, to the overall close time across departments. From the list of median close times shown above it would appear then that the departments that the City should focus on should be Property Management, Inspectional Services, and the Water and Sewer Commission. By the same measure, the best departments are the Transportation Department, Public Works and Animal Control, which I assume is ANML since it's not defined in the data dictionary. I ignore both DND_ and GEN_ since they are not defined in the data dictionary.

```
## # A tibble: 15 x 2
##    department      n
##    <chr>       <int>
##  1 PWDx       220950
##  2 BTDT       108470
##  3 ISD         25482
##  4 PARK        22638
##  5 GEN_        20952
##  6 INFO         9068
##  7 PROP         4533
##  8 BWSC          847
##  9 ANML          532
## 10 BHA_           55
## 11 BPS_           42
## 12 BPD_           22
## 13 DISB           18
## 14 DND_           17
## 15 No Q            1
```

However, given the most requests are for PWDx, BTDT, and ISD, it might make more sense to focus on those departments instead, since they make up the largest proportion of the overall close time.

3. In words, describe 2 confounding factors that can impact the close time of a ticket. You must be able to calculate these confounding factors from the dataset.

If we suppose that the department is the main reason for the close time of the requests, we can easily see

how the reason for the request would be a confounding factor. Departments are usually centered around certain objectives and the reasons will be directly related to those reasons. Neighborhood could also be a confounding factor for more subtle reasons. Neighborhoods that have been divested from are far from service or administration centers may require more time for requests.

4. In words, describe a confounding factor that can impact the close time of a ticket. This factor will come from an outside dataset.
    - Describe how you might be able to find this dataset.
    - Describe exactly how you will join this outside dataset to the 311 ticket data

In terms what I was alluding to earlier with factoring neighborhood into close time, we can understand it as a proxy for population characteristics like race and income (because of historic divestment, segregation, etc.). Based on this assumption we can then use Census data, specifically income and race data from the American Community Survey (table S1903). With this information we would essentially be exploring the question, "does close time vary by the makeup of an area of the city?" To join this dataset with our current one, there are several paths but the simplest way would be to join via zip code, which will then give us median incomes for each race for every zip code in our dataset.