# HW1

## Homework 1

## Intro/Background

We are using an NIBRS-compliant dataset submitted to the FBI by the Los Angeles Police Department (LAPD). From the dataset website: >This dataset reflects incidents of crime in the City of Los Angeles dating back to 2020. This data is transcribed from original crime reports that are typed on paper and therefore there may be some inaccuracies within the data. Some location fields with missing data are noted as (0°, 0°). Address fields are only provided to the nearest hundred block in order to maintain privacy. This data is as accurate as the data in the database. Please note questions or concerns in the comments."

### Examining the Data

```
spc_tbl_ [986,500 x 28] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ DR_NO         : num [1:986500] 1.90e+08 2.00e+08 2.00e+08 2.01e+08 2.21e+08 ...
 $ Date Rptd     : chr [1:986500] "03/01/2020 12:00:00 AM" "02/09/2020 12:00:00 AM" "11/11/20
 $ DATE OCC      : chr [1:986500] "03/01/2020 12:00:00 AM" "02/08/2020 12:00:00 AM" "11/04/20
 $ TIME OCC      : chr [1:986500] "2130" "1800" "1700" "2037" ...
 $ AREA          : chr [1:986500] "07" "01" "03" "09" ...
 $ AREA NAME     : chr [1:986500] "Wilshire" "Central" "Southwest" "Van Nuys" ...
 $ Rpt Dist No   : chr [1:986500] "0784" "0182" "0356" "0964" ...
 $ Part 1-2      : num [1:986500] 1 1 1 1 2 2 2 2 2 2 ...
 $ Crm Cd        : num [1:986500] 510 330 480 343 354 354 354 354 354 624 ...
 $ Crm Cd Desc   : chr [1:986500] "VEHICLE - STOLEN" "BURGLARY FROM VEHICLE" "BIKE - STOLEN"
 $ Mocodes       : chr [1:986500] NA "1822 1402 0344" "0344 1251" "0325 1501" ...
 $ Vict Age      : num [1:986500] 0 47 19 19 28 41 25 27 24 26 ...
 $ Vict Sex      : chr [1:986500] "M" "M" "X" "M" ...
 $ Vict Descent  : chr [1:986500] "O" "O" "X" "O" ...
 $ Premis Cd     : num [1:986500] 101 128 502 405 102 501 502 248 750 502 ...
$ Premis Desc   : chr [1:986500] "STREET" "BUS STOP/LAYOVER (ALSO QUERY 124)" "MULTI-UNIT DW
$ Weapon Used Cd: num [1:986500] NA NA NA NA NA NA NA NA NA 400 ...
$ Weapon Desc   : chr [1:986500] NA NA NA NA ...
$ Status        : chr [1:986500] "AA" "IC" "IC" "IC" ...
$ Status Desc   : chr [1:986500] "Adult Arrest" "Invest Cont" "Invest Cont" "Invest Cont" .
$ Crm Cd 1      : num [1:986500] 510 330 480 343 354 354 354 354 354 624 ...
$ Crm Cd 2      : num [1:986500] 998 998 NA NA NA NA NA NA NA NA ...
$ Crm Cd 3      : num [1:986500] NA NA NA NA NA NA NA NA NA NA ...
$ Crm Cd 4      : logi [1:986500] NA NA NA NA NA NA ...
$ LOCATION      : chr [1:986500] "1900 S  LONGWOOD                     AV" "1000 S  FLOWER
$ Cross Street  : chr [1:986500] NA NA NA NA ...
$ LAT           : num [1:986500] 34 34 34 34.2 34.1 ...
$ LON           : num [1:986500] -118 -118 -118 -118 -118 ...
- attr(*, "spec")=
 .. cols(
 ..   DR_NO = col_double(),
 ..   `Date Rptd` = col_character(),
 ..   `DATE OCC` = col_character(),
 ..   `TIME OCC` = col_character(),
 ..   AREA = col_character(),
 ..   `AREA NAME` = col_character(),
 ..   `Rpt Dist No` = col_character(),
 ..   `Part 1-2` = col_double(),
 ..   `Crm Cd` = col_double(),
 ..   `Crm Cd Desc` = col_character(),
 ..   Mocodes = col_character(),
 ..   `Vict Age` = col_double(),
 ..   `Vict Sex` = col_character(),
 ..   `Vict Descent` = col_character(),
 ..   `Premis Cd` = col_double(),
 ..   `Premis Desc` = col_character(),
 ..   `Weapon Used Cd` = col_double(),
 ..   `Weapon Desc` = col_character(),
 ..   Status = col_character(),
 ..   `Status Desc` = col_character(),
 ..   `Crm Cd 1` = col_double(),
 ..   `Crm Cd 2` = col_double(),
 ..   `Crm Cd 3` = col_double(),
 ..   `Crm Cd 4` = col_logical(),
 ..   LOCATION = col_character(),
 ..   `Cross Street` = col_character(),
 ..   LAT = col_double(),
 ..   LON = col_double()
```

```
  .. )
 - attr(*, "problems")=<externalptr>
```

- This is an official explanation of the columns from the LAPD

## Cleaning the Data

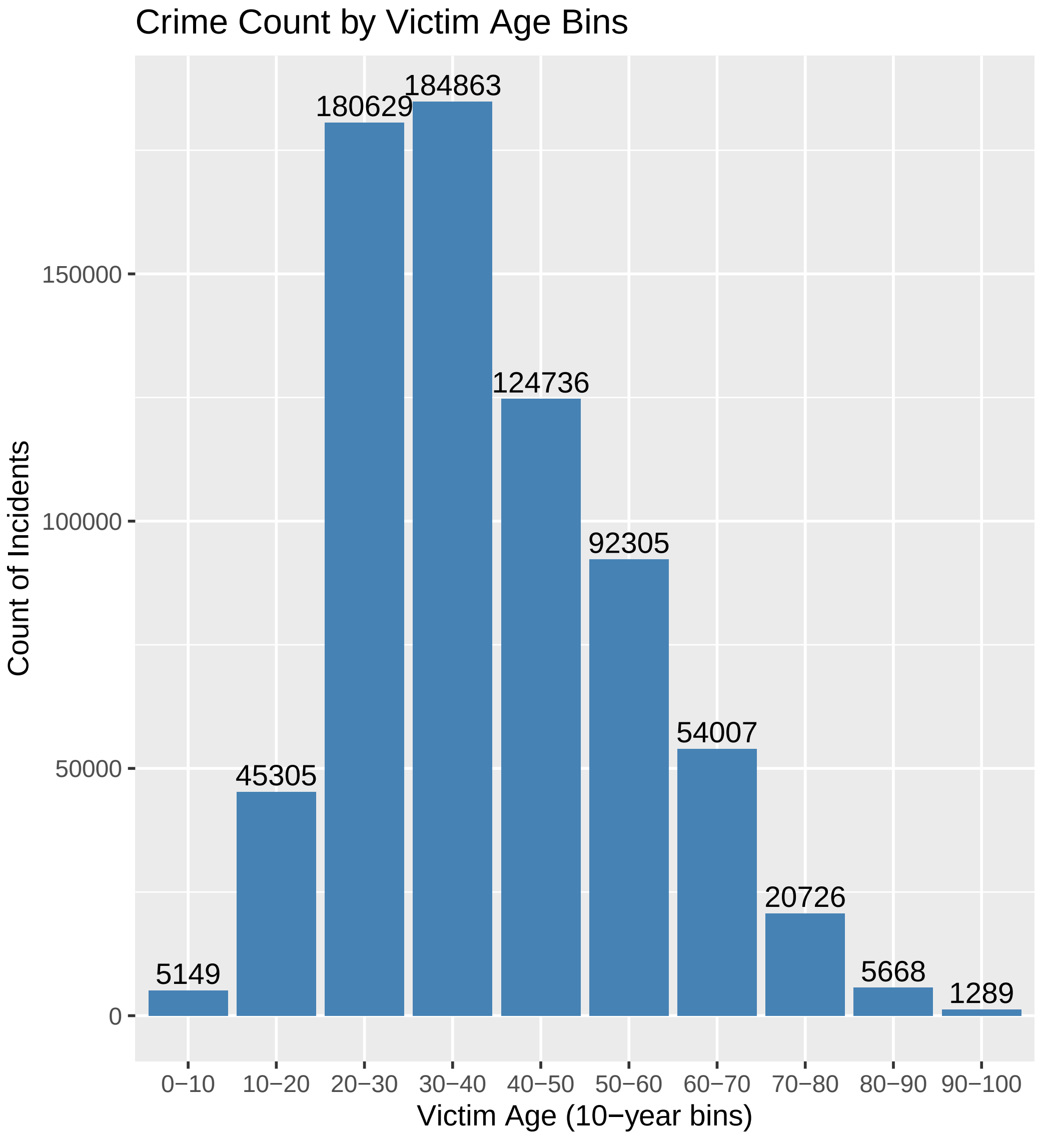
Initially the data has 986,500 rows corresponding to nearly 1 million individual crimes. However, upon cleaning the data, namely by filtering out crimes where the victim age is less than 0, if the victim sex is listed as "X" or "H", as well as dropping rows where certain columns may be missing, we are left with a dataset of 714,677 rows. Many of the aforementioned rows that have been dropped correspond to crimes that do not have a specific victim such as fraud, etc. We also rename the columns, replacing whitespace with underscores to avoid string issues.

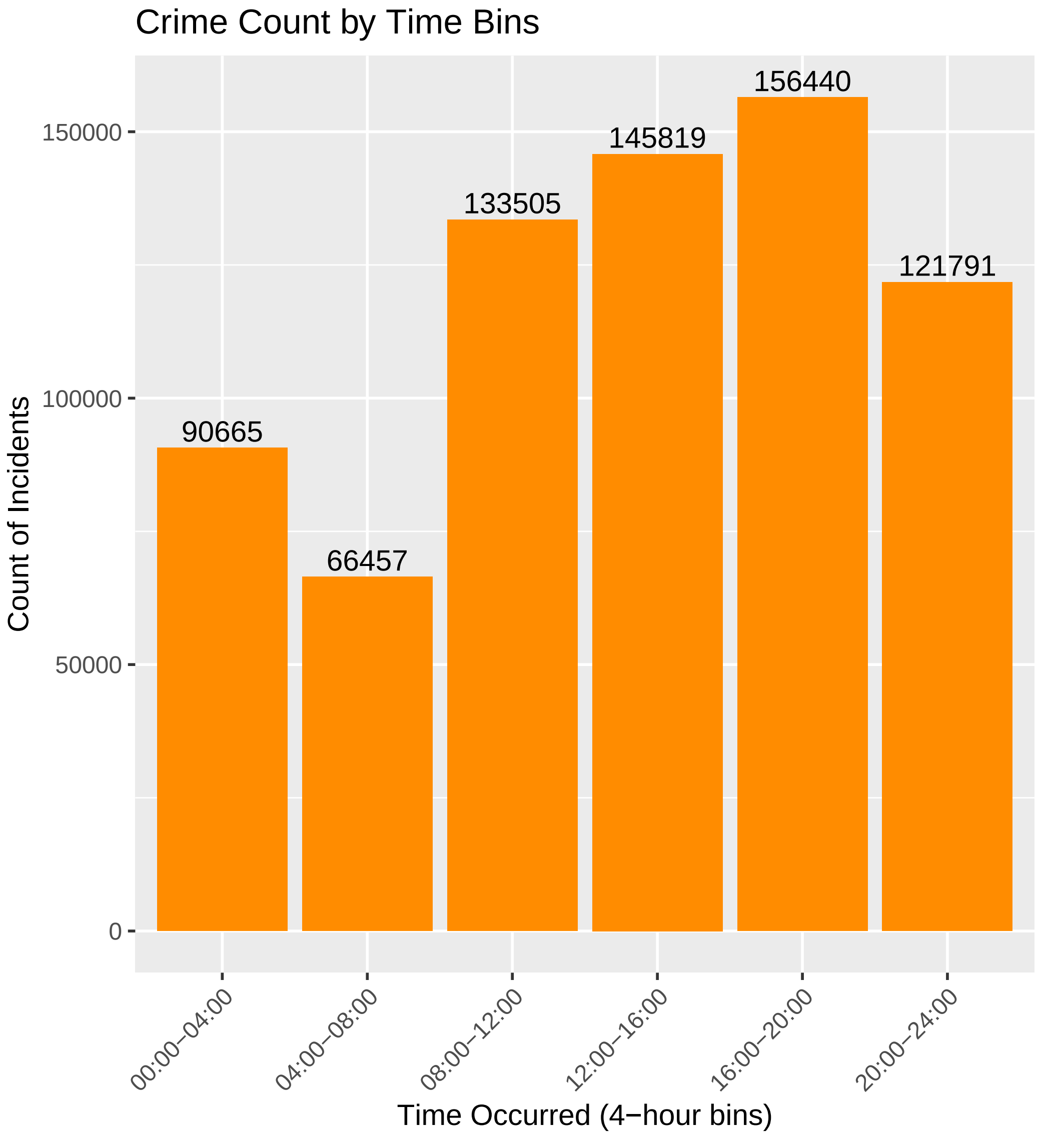
## Question 1 EDA

The first question we try to answer is whether the age of the victim (`Vict Age`) has a causal effect on the time a crime occurred (`Time Occ`).

We create bins to sort the age of victims into groups of 10 years each, so for example 1-10, 11-20, etc. We also use bins to sort the time that the crime occured into 4 hour blocks.

We then graph the counts of both the age and time bins, giving a general idea of the count of crimes with respect to both age and time.

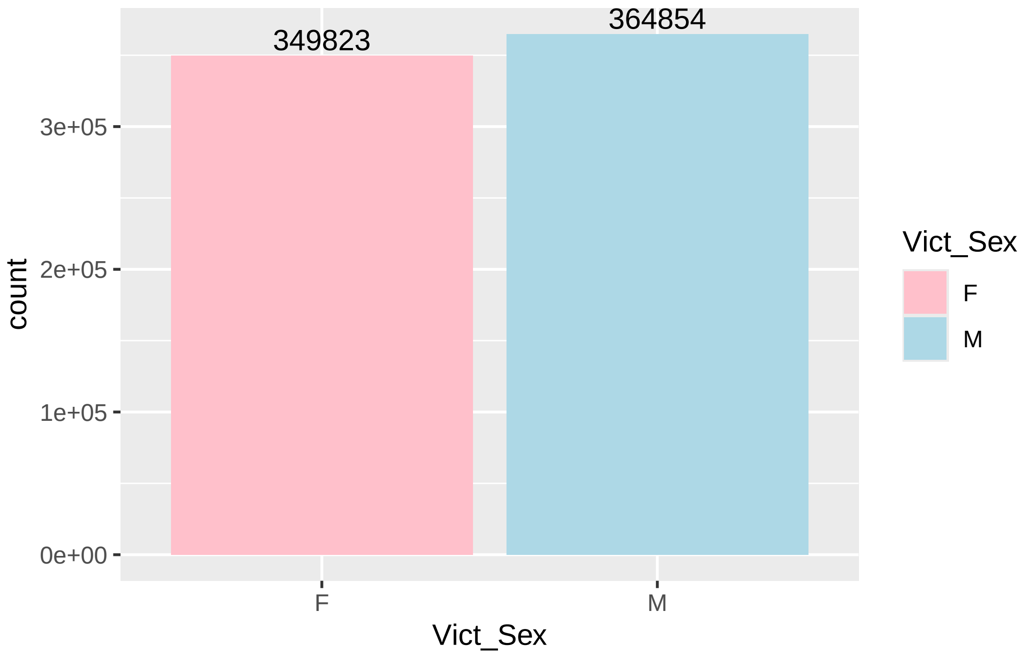

Crime Count by Victim Age Bins
Count of Incidents
150000
100000
50000
0
5149
45305
180629
184863
124736
92305
54007
20726
5668
1289
0−10
10−20
20−30
30−40
40−50
50−60
60−70
70−80
80−90
90−100
Victim Age (10−year bins)

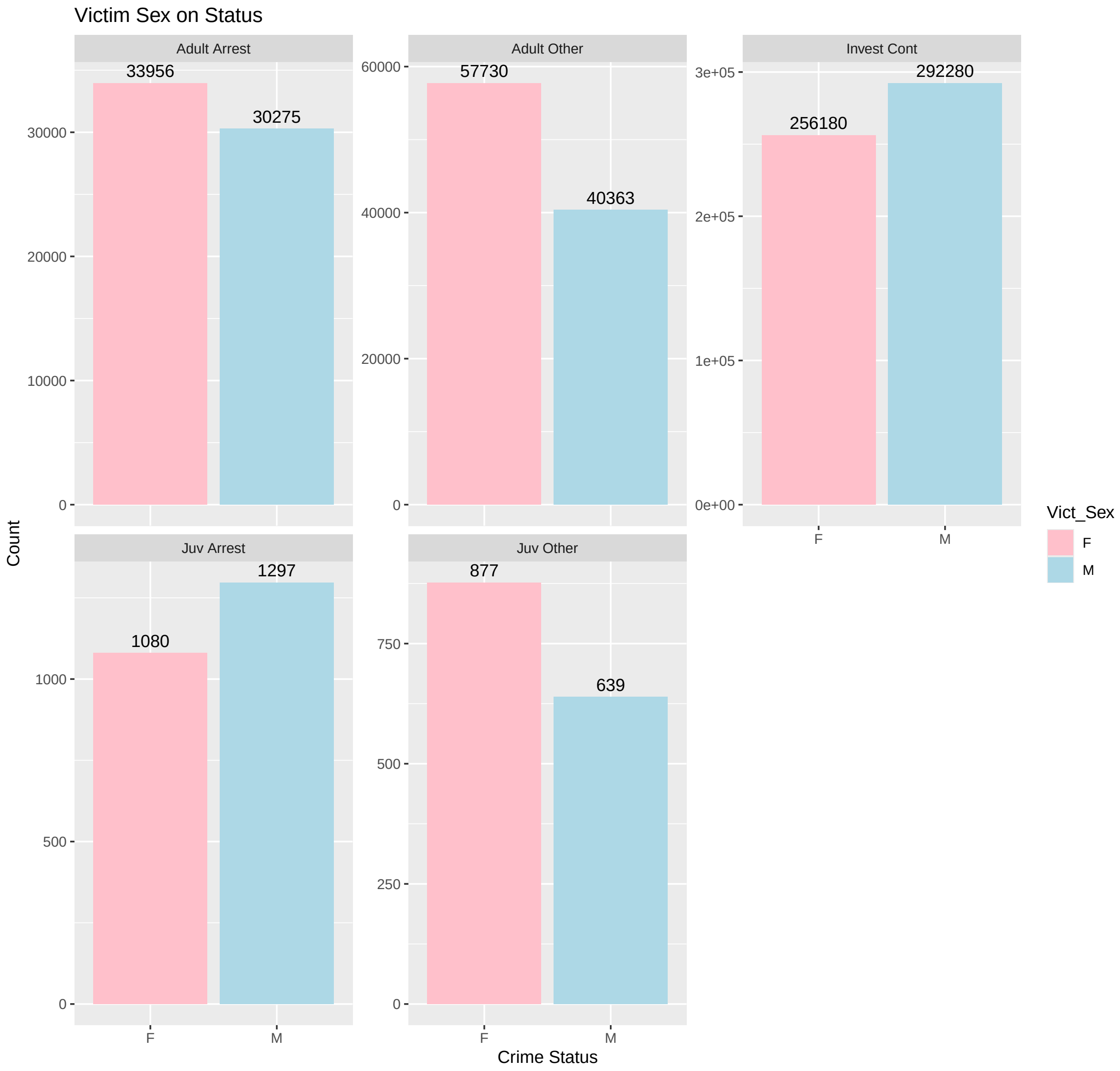


At a general glance, it seems that most victim ages are centered around individuals in their 20s to 40s. There does seem to be some crime for older victims around 40-60. In terms of time bins, there doesn't seem to be as much concentration, with a decent distribution of crime occurring throughout the day. Crime incidents peak between 16:00-20:00 (4 PM to 8 PM), suggesting higher activity during the late afternoon and early evening. Also, there are fewer incidents between 04:00-08:00 (4 AM to 8 AM), which aligns with when people are typically asleep. This distribution shows that crimes are more likely to occur during times when people are active and outside (afternoon and evening).

### Question 2 EDA

The second question we attempt to answer is whether the sex of the victim (`Vict Sex`) can affect the `Status` of the crime.





Overall it seems that there are slightly more male victims of crime vs female victims. That being said, it is kind of interesting that when it comes to juvenile arrests vs adult arrests, there are more male victims for the former, whereas for the latter, adult arrests, there do seem to be more male arrests. The juvenile and adult other categories basically act as a "catch-all", accounting for situations such as charges being dropped, citations, and other situations that did not lead to a conviction. It is also important to note there is a significant increase in disparity when comparing Male and Female victims when it comes to investigation continuing.

# Analysis

## Question 1

To answer whether the age of the victim affects the time that the crime occurred, we decided to utilize a GAM without propensity score matching given that both victim age and time_in_minutes (from midnight), a custom column we made to be easier to work with, are continuous values. Overall, propensity score matching would be much less impactful when not dealing with categorical values. The GAM model includes the victim age as a smooth term to map non-linear relationships and adjusts for victim race, crime code, area, and victim sex as additional covariate variables.

## Question 2

Since we are dealing with a binary value, namely the victim being male or female, we decided to utilize propensity score matching in order to compare outcomes based on some covariate values. We used logistic regression via a GLM to model whether a victim as male or female based on the covariates. Then we we used nearest neighbor matching to match similar propensity scores in order to create a matched dataset that reduces bias. The covariates in question we decided to consider are the victim age bins, victim race, the crime (code), and the general area the crime occurred in. We then used a multinomal logistic regression (given that there are more than 2 possible crime status outcomes) in order to model the (crime) Status variable based on victim sex. We did have to subset the data using a subset of 100,000 rows due to both memory and performance constraints.

```
Attaching package: 'nnet'

The following object is masked from 'package:mgcv':

    multinom

Warning: Fewer control units than treated units; not all treated units will get
a match.

# weights:  15 (8 variable)
initial  value 138379.471711
iter  10 value 61804.315448
iter  20 value 61323.709804
final  value 61320.455229
converged
```

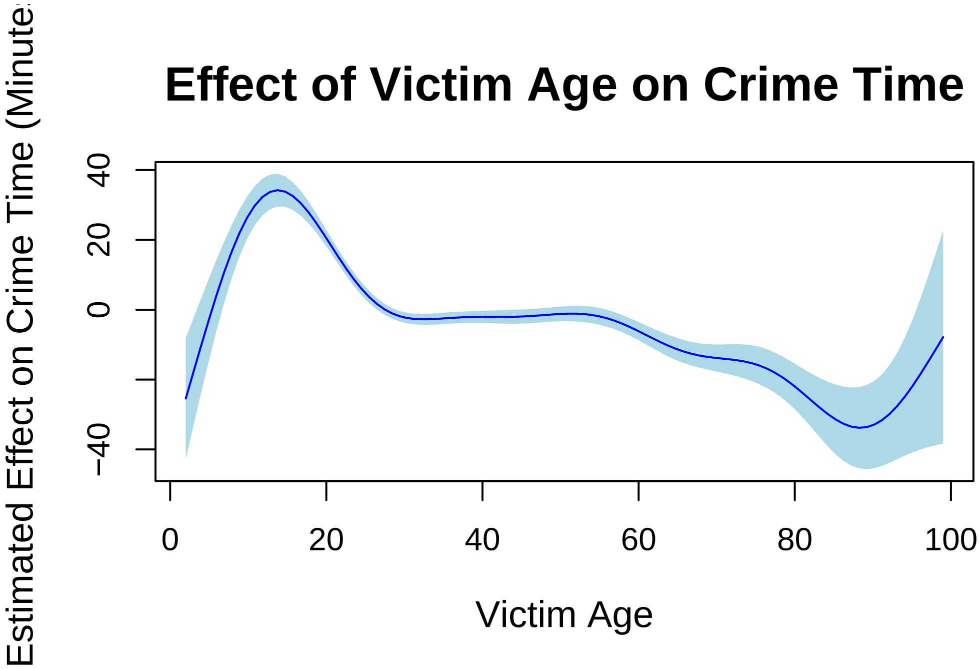
## Results

### Question 1

```
Family: gaussian
Link function: identity

Formula:
time_in_minutes ~ s(Vict_Age) + Vict_Descent + Crm_Cd + AREA +
    Vict_Sex

Parametric coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   697.861155 390.787899   1.786  0.07414 .
Vict_DescentA  99.512448 390.789978   0.255  0.79900
Vict_DescentB  69.920909 390.782203   0.179  0.85800
Vict_DescentC 141.796865 390.826325   0.363  0.71674
Vict_DescentD  52.041603 393.019874   0.132  0.89466
Vict_DescentF 109.922520 390.823952   0.281  0.77851
Vict_DescentG  99.327016 393.484588   0.252  0.80071
Vict_DescentH  95.595144 390.780959   0.245  0.80675
Vict_DescentI 109.045807 390.981327   0.279  0.78032
Vict_DescentJ 123.469075 390.911206   0.316  0.75212
Vict_DescentK 112.785260 390.815920   0.289  0.77290
Vict_DescentL -41.866035 393.485207  -0.106  0.91527
Vict_DescentO  95.866290 390.782971   0.245  0.80621
Vict_DescentP  62.101193 391.487792   0.159  0.87396
Vict_DescentS  -0.463656 394.669671  -0.001  0.99906
Vict_DescentU  89.225271 391.889119   0.228  0.81990
Vict_DescentV 129.444915 390.955407   0.331  0.74057
Vict_DescentW  91.565147 390.781187   0.234  0.81474
Vict_DescentX 106.120102 390.808402   0.272  0.78598
Vict_DescentZ 129.232219 391.146099   0.330  0.74110
Crm_Cd          0.039953   0.002123  18.822  < 2e-16 ***
AREA02        -13.765731   2.795953  -4.923 8.51e-07 ***
AREA03         -6.525577   2.613954  -2.496  0.01254 *
AREA04        -21.482145   3.108248  -6.911 4.81e-12 ***
AREA05        -14.083778   2.951541  -4.772 1.83e-06 ***
AREA06        -33.041899   2.654532 -12.447  < 2e-16 ***
AREA07          2.545129   2.756906   0.923  0.35591
AREA08         -8.188952   2.812554  -2.912  0.00360 **
```

```
AREA09        -15.213012   2.783530  -5.465 4.62e-08 ***
AREA10         -6.830247   2.853967  -2.393  0.01670 *
AREA11         -2.265231   2.904133  -0.780  0.43539
AREA12        -21.574632   2.563240  -8.417  < 2e-16 ***
AREA13        -31.161424   2.788388 -11.175  < 2e-16 ***
AREA14         -1.923801   2.613924  -0.736  0.46174
AREA15        -14.789071   2.720726  -5.436 5.46e-08 ***
AREA16        -18.423980   3.075898  -5.990 2.10e-09 ***
AREA17         -7.896502   2.907800  -2.716  0.00662 **
AREA18        -18.620459   2.726709  -6.829 8.56e-12 ***
AREA19        -16.789215   2.893354  -5.803 6.53e-09 ***
AREA20        -18.705208   2.732635  -6.845 7.65e-12 ***
AREA21          0.996650   2.816598   0.354  0.72345
Vict_SexM      13.541144   0.942379  14.369  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
              edf Ref.df     F p-value
s(Vict_Age) 8.781   8.98 46.99  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.00283   Deviance explained = 0.29%
GCV = 1.5271e+05  Scale est. = 1.527e+05  n = 714677
```

Given that the p-value for the smooth term of victim age is extremely low, at p < 2e-16, there is a strong non-linear relationship between victim age and the time that a crime occurred.



The solid blue line represents the fitted smooth function of victim age on the time a crime occurred. The shaded light blue area around the line represents a default 95% confidence interval, with a larger area corresponding to greater uncertainty at the respective age range, and vice-versa. In general, this plot shows how the estimated time of day for a crime tends to occur earlier or later depending on the victim's age, with upward trends indicating crimes that occur later in the day and downward trends indicating crimes that occur earlier in the day for the respective age. (basically we look at the slope)

For example, when we look at younger victims, especially those closer to the age of 20, it would appear that crimes involving them tend to occur much later in the day. However, when we get to the 20-30 age, it appears that the line approaches to 0, and from 30-40, the line remains relatively flat throughout the age bracket, meaning that for those 20-30, crime occurs earlier in the day whereas for those 30-40, there isn't really any correlation to their age and the time a crime occurs. For victims aged 40-90, it does seem like there is a gradual downward trend, indicating that crimes with older victims occur earlier in the day. Interestingly enough, for those aged 90-100, the line trends up again, indicating that crime related to them occur later in the day. However, the respective confidence interval area is also quite high, meaning that the estimate is not as certain.

Overall, the age of a victim plays a significant role in determining when the crime in question occurs.

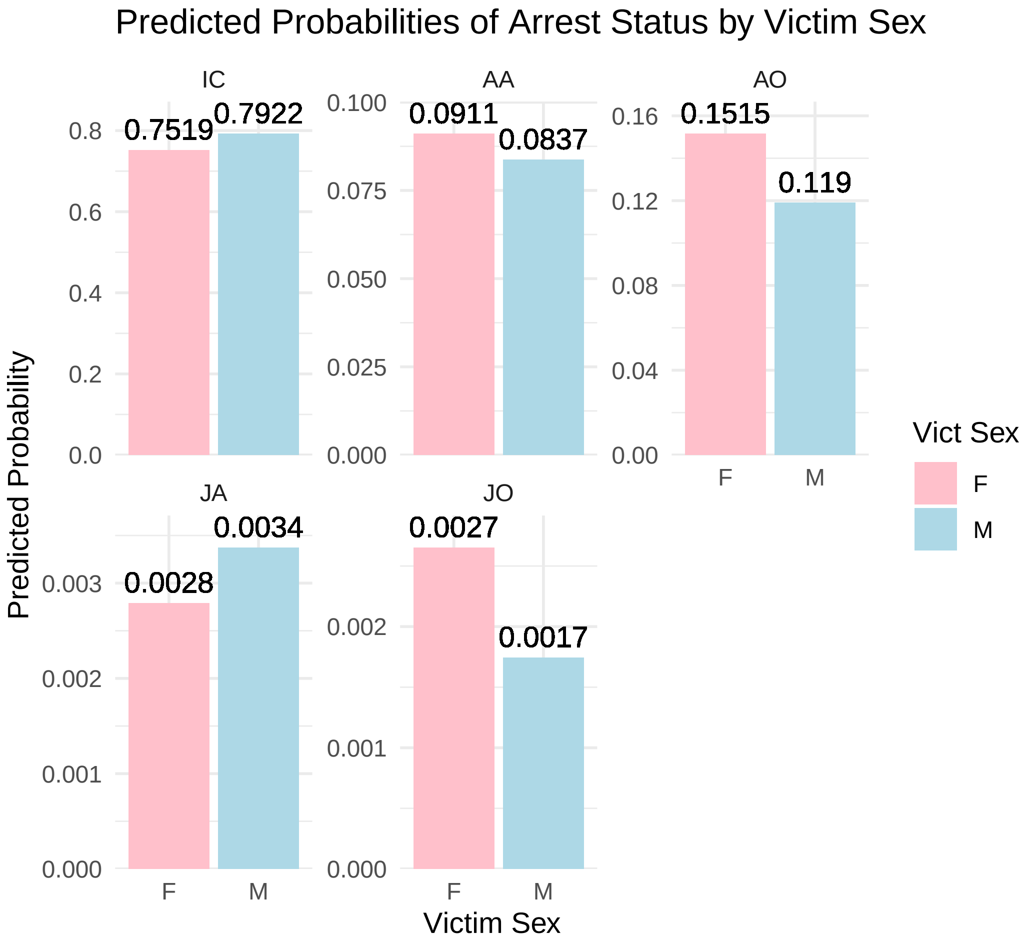
## Question 2

```
Call:
nnet::multinom(formula = Status ~ Vict_Sex, data = matched_data_df)

Coefficients:
   (Intercept)  Vict_SexM
AA   -2.111058 -0.1366001
AO   -1.601764 -0.2940942
JA   -5.596149  0.1370788
JO   -5.647445 -0.4708383

Std. Errors:
   (Intercept)  Vict_SexM
AA  0.01692226 0.02436510
AO  0.01358038 0.02023098
JA  0.09145700 0.12365431
JO  0.09382435 0.14888064

Residual Deviance: 122640.9
AIC: 122656.9
```

In terms of the coefficients, they represent the log odds for each crime status relative to the reference category, which in this case, we used a female victim with the crime status being "IC" or investigation continuing. For example, when the victim is male, then the log odds decrease by 0.1366 compared to if the victim is female for an adult arrest. Also, for juvenile arrests, then if the victim is male, the log odds increases by 0.1371.

```
   (Intercept) Vict_SexM
AA 0.121109770 0.8723190
AO 0.201540666 0.7452063
JA 0.003712131 1.1469186
JO 0.003526515 0.6244785
```

Here we perform exponentiation on the coefficients to get the odds ratios in order to get values that are "easier" to interpret. For example, for an adult arrest, a value of 0.87 for a male victim means that an adult arrest for a male victim is 13% less likely compared to a female victim. For 0.745 for an AO, a male victim will be roughly 25.5% less likely to have an adult other status to a crime vs a female victim. Also, for juvenile crimes, a male victim will be 14.7% more likely to lead to a juvenile arrest vs a female victim. In addition, a male victim will be 37.6% more likely to have a juvenile other status.



Here we've decided to graph the respective predicted probability of both a male and female victim of a crime being in each possible status. For example, lets look at the top middle graph, represented by "AA" or adult arrests. The female bar has a value of 0.0911, and the male bar has a value of 0.0837. This means that 9.11% of male victim cases are expected to have an adult arrested status, while 8.37% of female cases are expected to have an adult arrested status.

Overall it does seem that the sex of the victim does have some effect on the status of the crime. For example, if there is a female victim, there is a greater chance of the crime being an adult arrest, adult other, or juvenile other. Conversely, if the victim is male, then there is a greater chance of the crime being an investigation continuing or juvenile arrest.

## Discussion

Regarding question 1 (does victim age affect crime time), its kind of interesting how in general, crime for older victims tends to occur earlier in the day, whereas for younger victims, crime occurs later in the day. A possible reason for this is that younger folks are either in school or working during the day, whereas older folks could be out and about during the day. In addition, younger victims are more likely to be out during the evening, whereas older victims could be at home, sleeping early, etc.

Regarding question 2 (does victim sex affect crime status), in general it seems that for female victims, its much more likely for a crime to be "resolved" especially due to the disparity in male and female victims when considering how many more investigation continuing cases are comprised of male victims. It is important to note that at least for juvenile cases, male victims were more likely to lead to a juvenile arrest. Part of this could be related to the fact that there may be more public pressure to investigate and close cases that relate to violence against women.

In general, the dataset does seem somewhat limited. For example, there is no information about both the race and gender of the actual perpetrator of the crime. In addition, it would be great if there was a greater level of granularity when it came to describing crimes. For example, including whether a crime as violent or non-violent, a sex-related crime, a felony or misdemeanor, etc would all be helpful in seeing more interesting trends. We think this is especially important as crime data can be interpreted in many ways and lead to serious real-world implications, such as additional funding, an increase in policing, and an overall varying approach towards how to even approach policing, such as being stricter on crime vs a more harm-reduction, community based approach.