Evelyn Galvan
Victor Florintsev
Kelvin Nguyen

1. Introduction

Part a: Description of Data

For this data we used 2 data sources:

Crime: https://query.data.world/s/bim6ij5fibnosxv26222gona3emzz4

UCR Crime Data: 17 columns and 2,929 rows

- Data about crime rates in various cities across the country
- Spans from 1975-2015

Temperature: https://query.data.world/s/gsdqkv5dndjfcl5z7vtdn2o5nopfes

City_data.csv: 4 columns, 71,311 rows

Average temperature data from year 1849 to 2013 by city.

We merged and filtered out any inconsistent date to get one single data set with both crime and temperature.

Part b: Questions/Tasks Posed

- Do heat waves correlate to a higher rate of crime?
- Do geography and climate zone correlate to crime?
- How much is crime expected to increase using future projections and our model?

2. Questions and Models

Part a: Question Answered

We narrowed the question down to "***How much is crime expected to increase using future projections and our model?***" The models we chose to use are Multiple Linear Regression and Neural Network (with a sprinkle of trees for funsies)

Part b:

Model 1: Linear Regression

I. Onecity.data$Homs_per_100k = -10.210 + Onecity.data$avg_temp * 3.653

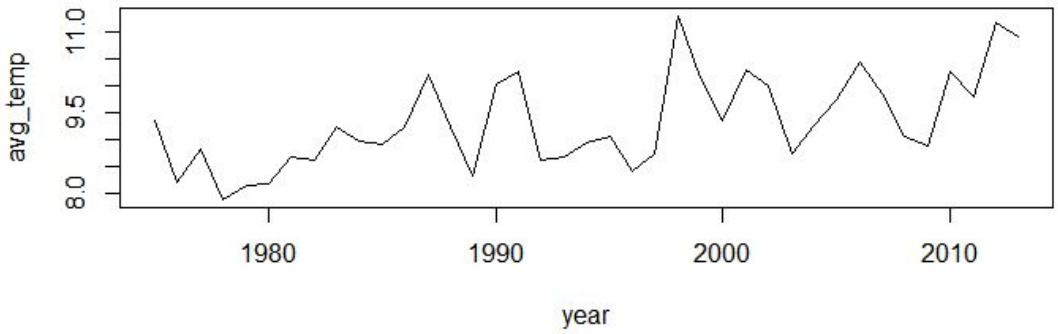
Reasoning:

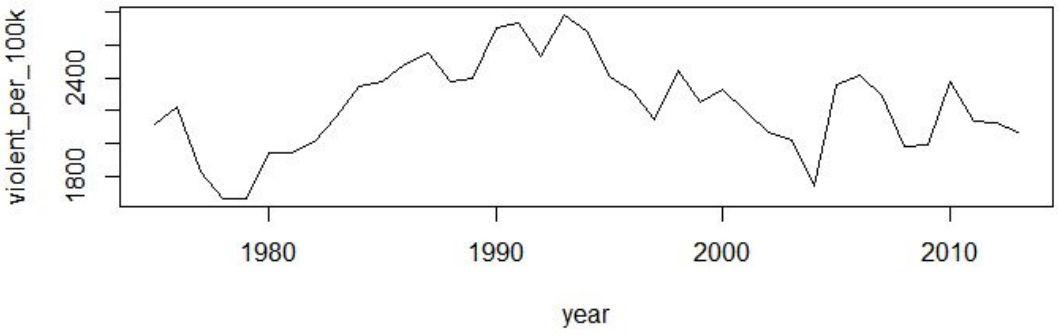
We will start off with a linear regression model, flipped predictors with the response as heat to evaluate which types of crime are actually correlated and get our head around how the data is correlated. This linear regression will be

bootstrapped. Then we will try to predict how much crime can be expected with the change in temperature in future years.

Reasons/Advantages over other methods:

Our response variable, avg_tmp is a quantitative (continuous) variable. We're using this method, over others to determine which set of predictors are correlated with avg_tmp. Multiple linear regression is a VERY easy to interpret model (compared to random forests and neural networks), starting with this will give us a good idea of what types of crime are correlated and which types may not be (so we can throw them out from the predictor in the next model)






```
Call:
lm(formula = rape_per_100k ~ avg_temp + year, data = crime_temp)

Residuals:
    Min      1Q  Median      3Q     Max
-64.909 -18.793  -3.762  14.347 119.501

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1787.33275  111.53693   16.02   <2e-16 ***
avg_temp      -1.72222    0.15704  -10.97   <2e-16 ***
year          -0.85230    0.05603  -15.21   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26.6 on 1791 degrees of freedom
Multiple R-squared:  0.1771,    Adjusted R-squared:  0.1762
F-statistic: 192.7 on 2 and 1791 DF,  p-value: < 2.2e-16
```

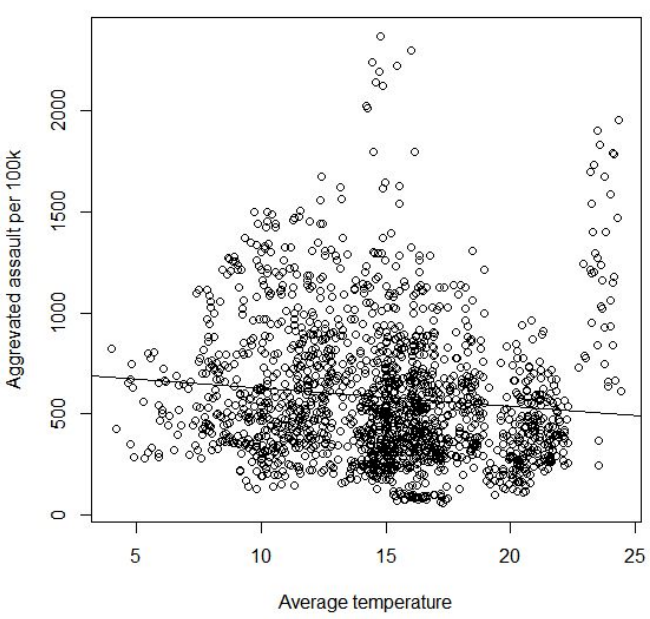
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 11121.781   2640.022   4.213 2.65e-05 ***
avg_temp      -23.677      3.717  -6.370 2.40e-10 ***
year           -4.824      1.326  -3.637 0.000283 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 629.5 on 1791 degrees of freedom
Multiple R-squared:  0.03165,   Adjusted R-squared:  0.03057
F-statistic: 29.27 on 2 and 1791 DF,  p-value: 3.098e-13
```

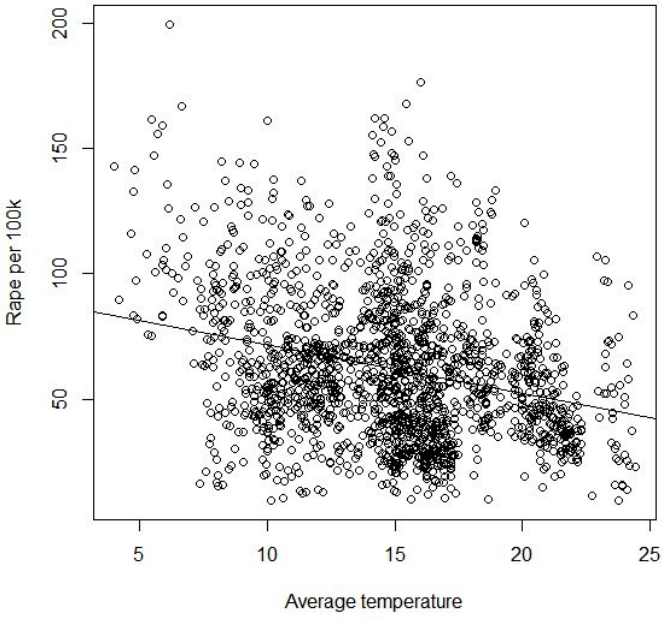
Violent_per_100k

The other crime types followed similar outcomes.

Reasoning: This shows that in general, crime has a negative relationship with year and average temperature.

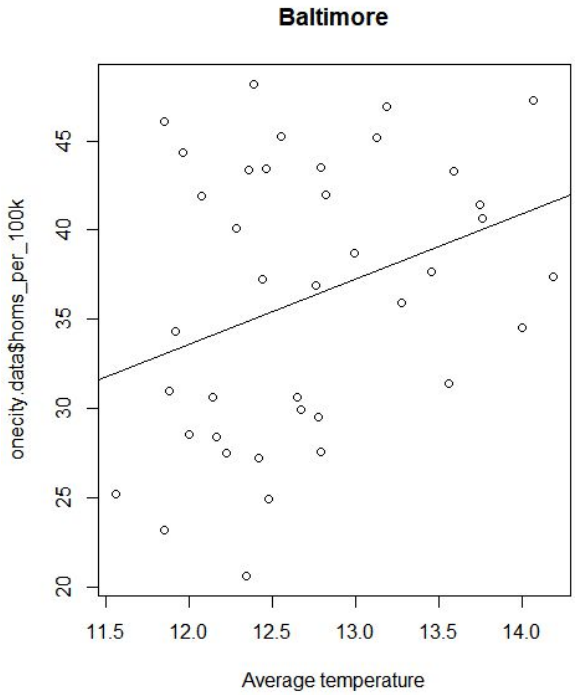
Specifically when we ran the linear model for the entire United States.





When we added city as a predictor variable, we were able to easily spot the cities that had higher crime.

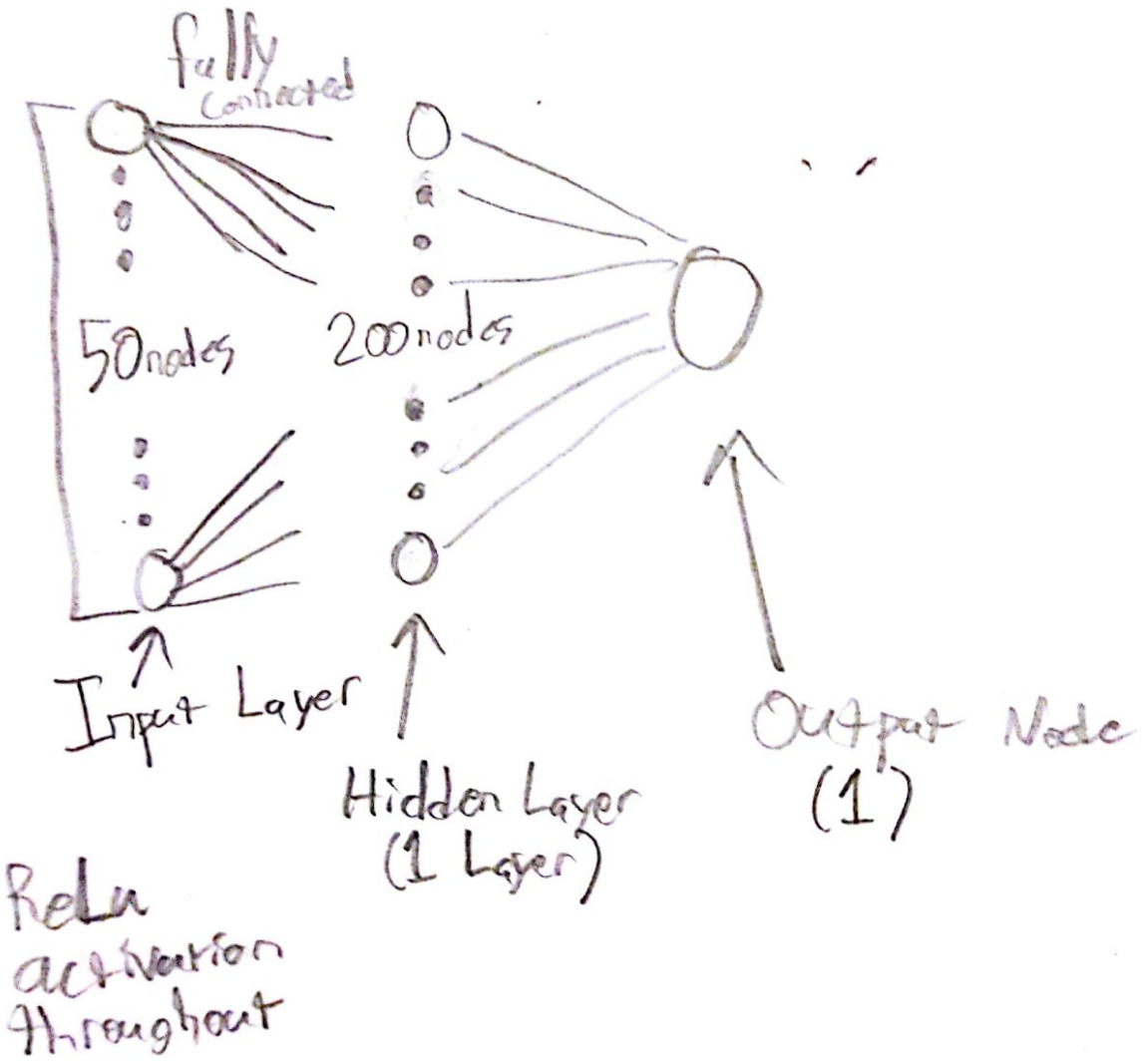
We plotted the cities with highest significance, and found that even when we take into consideration the cities with the highest crimes and a positive relationship between crime and heat, the data points are so spread out that we do not think we can generalize this relationship for the entire dataset.




```
fit <- lm(onecity.data$homs_per_100k ~ onecity.data$avg_temp)
onecity.data = crime_temp[crime_temp$city == "Baltimore",]
plot(onecity.data$homs_per_100k ~ onecity.data$avg_temp, main="Baltimore",xlab="Average temperature")
abline(fit)
```

After running 10 fold cross validation on a least squares regression fit with the response variable being homicides per 100,000 people, and the predictor being average temperature. We came up with a coefficient of -0.01004356 for average temperature. This low coefficient highlights the fact that this dataset does not support our hypothesis that homicides are affected by the average temperature of a city. The 10 fold cross validation was done with a shuffled dataset, and a 80:20 split of a training:validation ratio.

Model 2: Artificial Neural Network (ANN)



Reasoning:

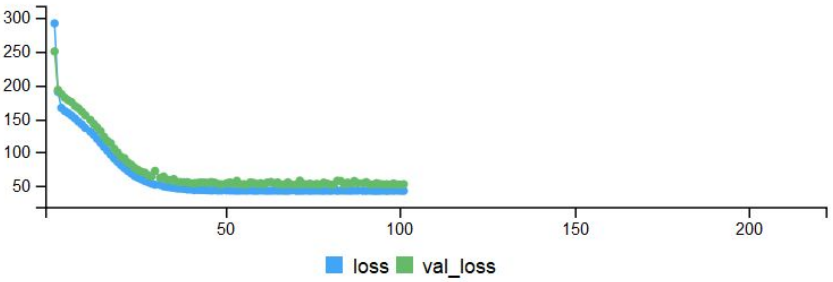
We will train our NN with

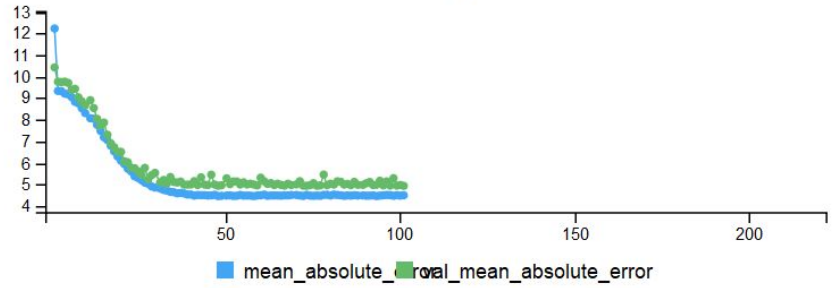
inputs: city, temp_delta (abs. value of temp_diff change), temp_diff (), avg_temp &

response variable: homs_per_100k

Reasons/Advantages over other methods:

- ➢ Ability to learn from initial inputs and their relationships. Can learn hidden relationships in the data
- ➢ Ability to infer unseen relationships on unseen data and eventually generalize and predict on that unseen data
- ➢ Can better model data with high volatility and non-constant variance





MSE: 4.871394

Given how difficult it was to discrena relationship from linear models, we were not expecting linear networks to be at all accurate. But to our surprise, we obtained a mean absolute error even through ten-fold cross-validation of only 4.87. For the inputs, we essentially used city one-hot encoded and average temperature for the year. Average temperature for the year was also transformed into the temperature difference from the average and the absolute difference from the average to give the neural network more to work with.

```
Call:
lm(formula = homs_per_100k ~ avg_temp + year + city, data = crime_temp)

Residuals:
    Min      1Q  Median      3Q     Max
-25.295  -3.013  -0.342   2.712  48.031

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       409.48615   30.13783  13.587  < 2e-16 ***
avg_temp            0.36710    0.26462   1.387 0.165535
year               -0.20228    0.01589 -12.731  < 2e-16 ***
cityArlington      -6.83422    1.78794  -3.822 0.000137 ***
cityAtlanta        24.32779    1.69089  14.388  < 2e-16 ***
cityAustin         -5.99776    2.74221  -2.187 0.028860 *
cityBaltimore      25.40771    1.46760  17.312  < 2e-16 ***
cityBoston          3.79498    1.67474   2.266 0.023572 *
cityChicago        13.58172    1.45273   9.349  < 2e-16 ***
cityColumbus        1.41130    1.64317   0.859 0.390518
cityDallas         11.99045    2.33072   5.145 2.98e-07 ***
cityDenver          3.04788    1.55153   1.964 0.049637 *
cityDetroit        37.63638    1.56018  24.123  < 2e-16 ***
cityEl Paso        -6.85822    1.89080  -3.627 0.000295 ***
cityFort Worth      6.64624    2.33072   2.852 0.004401 **
cityFresno          1.95619    1.90283   1.028 0.304072
cityHouston         7.54373    2.80136   2.693 0.007152 **
cityIndianapolis    5.16316    1.44054   3.584 0.000347 ***
cityKansas City    14.09361    1.45921   9.658  < 2e-16 ***
cityLas Vegas       0.85571    2.19971   0.389 0.697317
cityLong Beach      2.52802    1.96013   1.290 0.197318
cityLos Angeles     7.21656    1.91761   3.763 0.000173 ***
cityMemphis         9.07045    1.96386   4.619 4.14e-06 ***
cityMesa          -10.18087    2.99819  -3.396 0.000700 ***
cityMiami          13.68666    3.49969   3.911 9.55e-05 ***
cityMilwaukee       5.85536    1.45273   4.031 5.80e-05 ***
cityMinneapolis     3.72585    2.06856   1.801 0.071846 .
cityNashville       3.61533    1.67707   2.156 0.031240 *
cityNew Orleans    35.39654    2.81086  12.593  < 2e-16 ***
cityNew York City   6.11685    1.47057   4.159 3.34e-05 ***
cityOakland        17.90418    1.68419  10.631  < 2e-16 ***
cityOklahoma City   1.95846    1.80957   1.082 0.279278
cityOmaha          -2.10673    1.45292  -1.450 0.147239
cityPhiladelphia   11.83896    1.46785   8.066 1.34e-15 ***
cityPhoenix        -1.28892    2.99819  -0.430 0.667322
```

Part c: Prediction Comparison

i.

```
> totalAccuracy = totalAccuracy / n
> totalAccuracy
[1] 4.996666
>
> totalAccuracy_linear = totalAccuracy_linear / n
> totalAccuracy_linear
[1] 162.5126
```

ii. Interpretation

The neural network clearly outperformed the linear model.

Part d: Inference; fitting the model on the full data set

i. Output:

```
> totalAccuracy
[1] 4.573646
>
> totalAccuracy_linear
[1] 142.0309
```

ii. Interpretation

From these results, we can see that both of these models are overfitted because they have lower errors than when they were cross validated. That demonstrates the utility of cross validation to us. Additionally, we surmise that year and city predictors affect the response the most.

E. Conclusion

In predictive performance, the neural network outperformed the linear regression. This is clear in the mse's.
Though for interpretation purposes, the linear regression model is easier to interpret. We could not conclude a relationship from the linear regression, but we could clearly see which predictors were correlated: cities, year. We could clearly see that average temperature did not play a big role in predicting homicides among other cime types.
The results for the linear regression model were counter-intuitive to what we thought would occur. Initially we thought

that heat may induce higher crime rates. We analyzed different linear models for all cities and for individual cities. We realized that we cannot generalize a conclusion for a relationship between crime rate and heat (waves) for a linear model.