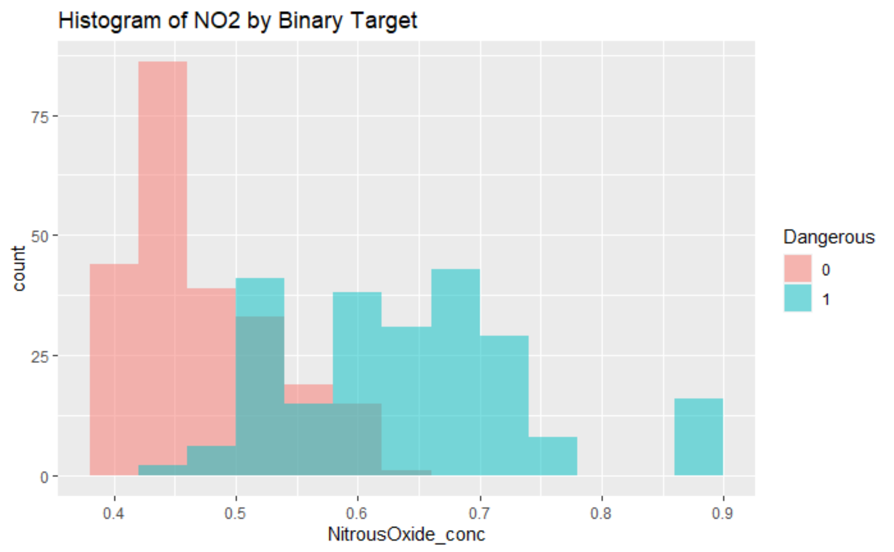
DATA 621 HW3

Keeno Glanville

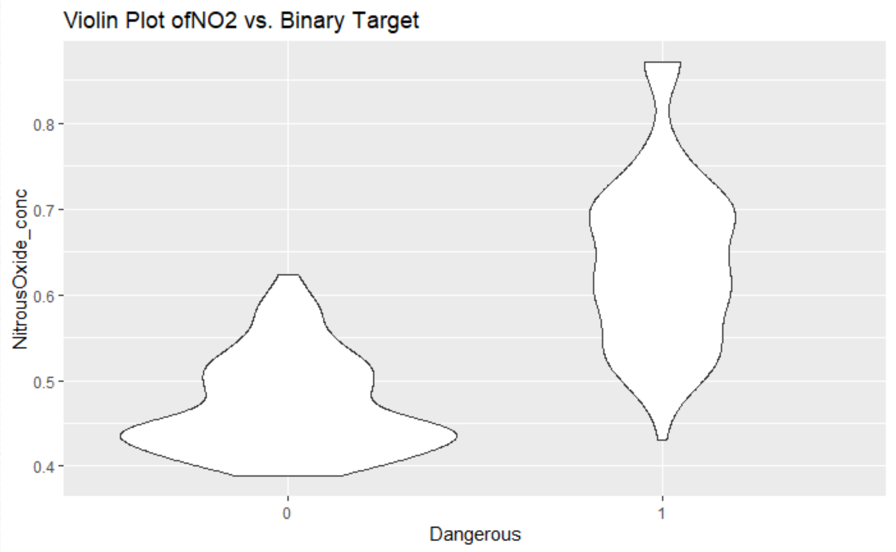
DATA EXPLORATION

I first explored the data through applying a point biserial correlation to the target variable. With this I was able to deduce the most influential variables on our measured outcomes. We then plot the most influential variable which was Nitrous Oxide Concentration. This was quite surprising.

Histogram



Violin Plot




```
     Residential_zone_Large Industrial_zone NitrousOxide_conc  Rooms_avg OwnerOccupiedUnits dis_to_emplyoymentcenter       tax   ptratio
[1,]             -0.4316818       0.6048507         0.7261062 -0.1525533          0.6301062               -0.6186731 0.6111133 0.2508489
        lstat       medv
[1,] 0.469127 -0.2705507
```

DATA PREPARATION

Preparing the data was quite simple. It involved really renaming the columns to fit my understanding as well as correctly labelling each column with its type (factor, numerical, etc.)

```
# DATA PREPARATION

```{r}
train <- trainraw%>%
  rename(Residential_zone_Large = zn)%>%
  rename(Industrial_zone = indus )%>%
  rename(Charles_River_border = chas)%>%
  rename(NitrousOxide_conc = nox)%>%
  rename(Rooms_avg = rm)%>%
  rename(OwnerOccupiedUnits= age)%>%
  rename(Highway_Index = rad)%>%
  rename(dis_to_emplyoymentcenter=dis)%>%
  rename(Dangerous = target)%>%
  mutate(Charles_River_border= factor(Charles_River_border))%>%
  mutate(Highway_Index= factor(Highway_Index))%>%
  mutate(Dangerous= factor(Dangerous))

test <- testraw%>%
  rename(Residential_zone_Large = zn)%>%
  rename(Industrial_zone = indus )%>%
  rename(Charles_River_border = chas)%>%
  rename(NitrousOxide_conc = nox)%>%
  rename(Rooms_avg = rm)%>%
  rename(OwnerOccupiedUnits= age)%>%
  rename(Highway_Index = rad )%>%
  rename(dis_to_emplyoymentcenter=dis)%>%
  mutate(Charles_River_border= factor(Charles_River_border))%>%
  mutate(Highway_Index= factor(Highway_Index))

```
```

	Residential_zone_Large	Industrial_zone	Charles_River_border	NitrousOxide_conc	Rooms_avg	OwnerOccupiedUnits	dis_to_emplyoymentcenter	Highway_Index	tax	ptratio	lstat	medv	Dangerous
1	0.0	19.58	0	0.6050	7.929	96.2	2.0459	5	403	14.7	3.70	50.0	1
2	0.0	19.58	1	0.8710	5.403	100.0	1.3216	5	403	14.7	26.82	13.4	1
3	0.0	18.10	0	0.7400	6.485	100.0	1.9784	24	666	20.2	18.85	15.4	1
4	30.0	4.93	0	0.4280	6.393	7.8	7.0355	6	300	16.6	5.19	23.7	0
5	0.0	2.46	0	0.4880	7.155	92.2	2.7006	3	193	17.8	4.82	37.9	0
6	0.0	8.56	0	0.5200	6.781	71.3	2.8561	5	384	20.9	7.67	26.5	0
7	0.0	18.10	0	0.6930	5.453	100.0	1.4896	24	666	20.2	30.59	5.0	1
8	0.0	18.10	0	0.6930	4.519	100.0	1.6582	24	666	20.2	36.98	7.0	1
9	0.0	5.19	0	0.5150	6.316	38.1	6.4584	5	224	20.2	5.68	22.2	0
10	80.0	3.64	0	0.3920	5.876	19.1	9.2203	1	315	16.4	9.25	20.9	0
11	22.0	5.86	0	0.4310	6.438	8.9	7.3967	7	330	19.1	3.59	24.8	0
12	0.0	12.83	0	0.4370	6.286	45.0	4.5026	5	398	18.7	8.94	21.4	0
13	0.0	18.10	0	0.5320	7.061	77.0	3.4106	24	666	20.2	7.01	25.0	1
14	22.0	5.86	0	0.4310	8.259	8.4	8.9067	7	330	19.1	3.54	42.8	1
15	0.0	2.46	0	0.4880	6.153	68.8	3.2797	3	193	17.8	13.15	29.6	0
16	0.0	2.18	0	0.4580	6.430	58.7	6.0622	3	222	18.7	5.21	28.7	0
17	100.0	1.32	0	0.4110	6.816	40.5	8.3248	5	256	15.1	3.95	31.6	0
18	20.0	3.97	0	0.6470	5.560	62.8	1.9865	5	264	13.0	10.45	22.8	1
19	0.0	18.10	0	0.6790	5.896	95.4	1.9096	24	666	20.2	24.39	8.3	1
20	0.0	18.10	0	0.6710	6.545	99.1	1.5192	24	666	20.2	21.08	10.9	1
21	0.0	3.24	0	0.4600	6.144	32.2	5.8736	4	430	16.9	9.09	19.8	0
22	0.0	6.20	1	0.5070	6.726	66.5	3.6519	8	307	17.4	8.05	29.0	1
23	0.0	2.89	0	0.4450	7.416	62.5	3.4952	2	276	18.0	6.19	33.2	0
24	18.0	2.31	0	0.5380	6.575	65.2	4.0900	1	296	15.3	4.98	24.0	0
25	0.0	9.90	0	0.5440	6.382	67.2	3.5325	4	304	18.4	10.36	23.1	1

MODEL BUILDING

The model utilized was a logistic regression which proved to initially show an overfit to the data. This could be due to the insufficient observations within the dataset. The regression was initially done factoring all variables, then removal of categorical variables, finally a scaled and one-hot encoding model.

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 75  1
         1  4 60

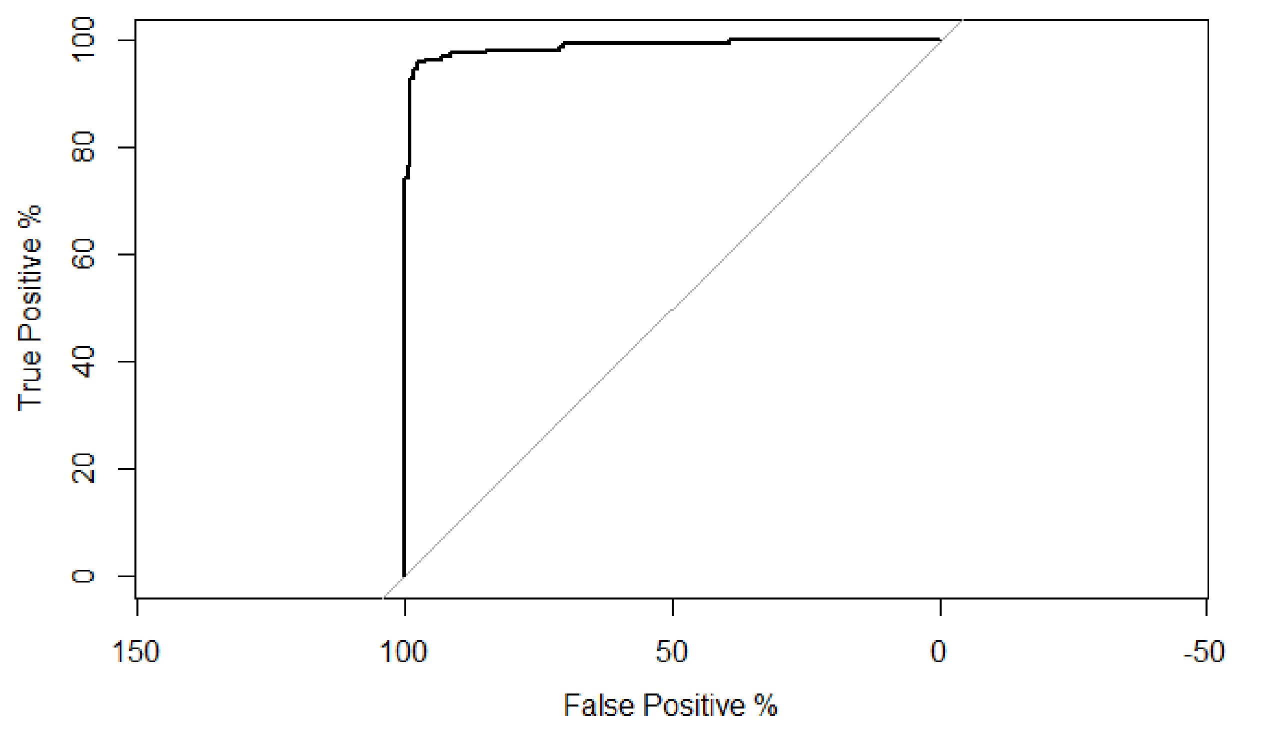
               Accuracy : 0.9643
                 95% CI : (0.9186, 0.9883)
    No Information Rate : 0.5643
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9278

 Mcnemar's Test P-Value : 0.3711

            Sensitivity : 0.9494
            Specificity : 0.9836
         Pos Pred Value : 0.9868
         Neg Pred Value : 0.9375
             Prevalence : 0.5643
         Detection Rate : 0.5357
   Detection Prevalence : 0.5429
      Balanced Accuracy : 0.9665

       'Positive' Class : 0
```



SELECTING MODEL

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 71  1
         1  8 60

               Accuracy : 0.9357
                 95% CI : (0.8815, 0.9702)
    No Information Rate : 0.5643
    P-Value [Acc > NIR] : <2e-16

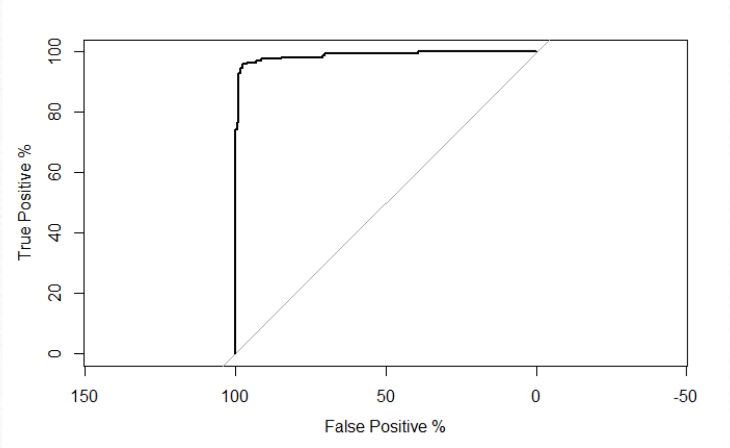
                  Kappa : 0.871

 Mcnemar's Test P-Value : 0.0455

            Sensitivity : 0.8987
            Specificity : 0.9836
         Pos Pred Value : 0.9861
         Neg Pred Value : 0.8824
             Prevalence : 0.5643
         Detection Rate : 0.5071
   Detection Prevalence : 0.5143
      Balanced Accuracy : 0.9412

       'Positive' Class : 0
```

I wanted to chose a model that was accurate but didn't seem to be overfitted. The way I accomplished this was by scaling the numerical variables and one-hot encoding the categorical variables. The test set provided didn't have values I could use for the prediction ROC and AUC curve so I instead subset the training data.



Appendix

https://github.com/kglan/MSDS/blob/03829122632aef9e839d0130ffa23c8cf47ea680/DATA621/HW3/HW3.Rmd

https://github.com/kglan/MSDS/blob/03829122632aef9e839d0130ffa23c8cf47ea680/DATA621/HW3/HW3.pdf