Week 3 homework

```r
#Loading packages
library('GGally')
library('stargazer')
library('tidyverse')
```

```r
#Loading data

#temperature data file
tempData <- read.table("https://d37djvu3ytnwxt.cloudfront.net/assets/courseware/v1/592f3be3e90d2bdfe6a69

#Crime data file
crimeData <- read.table("http://www.statsci.org/data/general/uscrime.txt", header = TRUE)

#Set seed for reproducibility
set.seed(156)

# stargazer(tempData)
# stargazer(crimeData)
```

Summary of Datasets for Week 3 Homework

Table 1: Summary of Temperature Data

Statistic	N	Mean	St. Dev.	Min	Max
M	47	13.857	1.257	11.900	17.700
So	47	0.340	0.479	0	1
Ed	47	10.564	1.119	8.700	12.200
Po1	47	8.500	2.972	4.500	16.600
Po2	47	8.023	2.796	4.100	15.700
LF	47	0.561	0.040	0.480	0.641
M.F	47	98.302	2.947	93.400	107.100
Pop	47	36.617	38.071	3	168
NW	47	10.113	10.283	0.200	42.300
U1	47	0.095	0.018	0.070	0.142
U2	47	3.398	0.845	2.000	5.800
Wealth	47	5,253.830	964.909	2,880	6,890
Ineq	47	19.400	3.990	12.600	27.600
Prob	47	0.047	0.023	0.007	0.120
Time	47	26.598	7.087	12.200	44.000
Crime	47	905.085	386.763	342	1,993

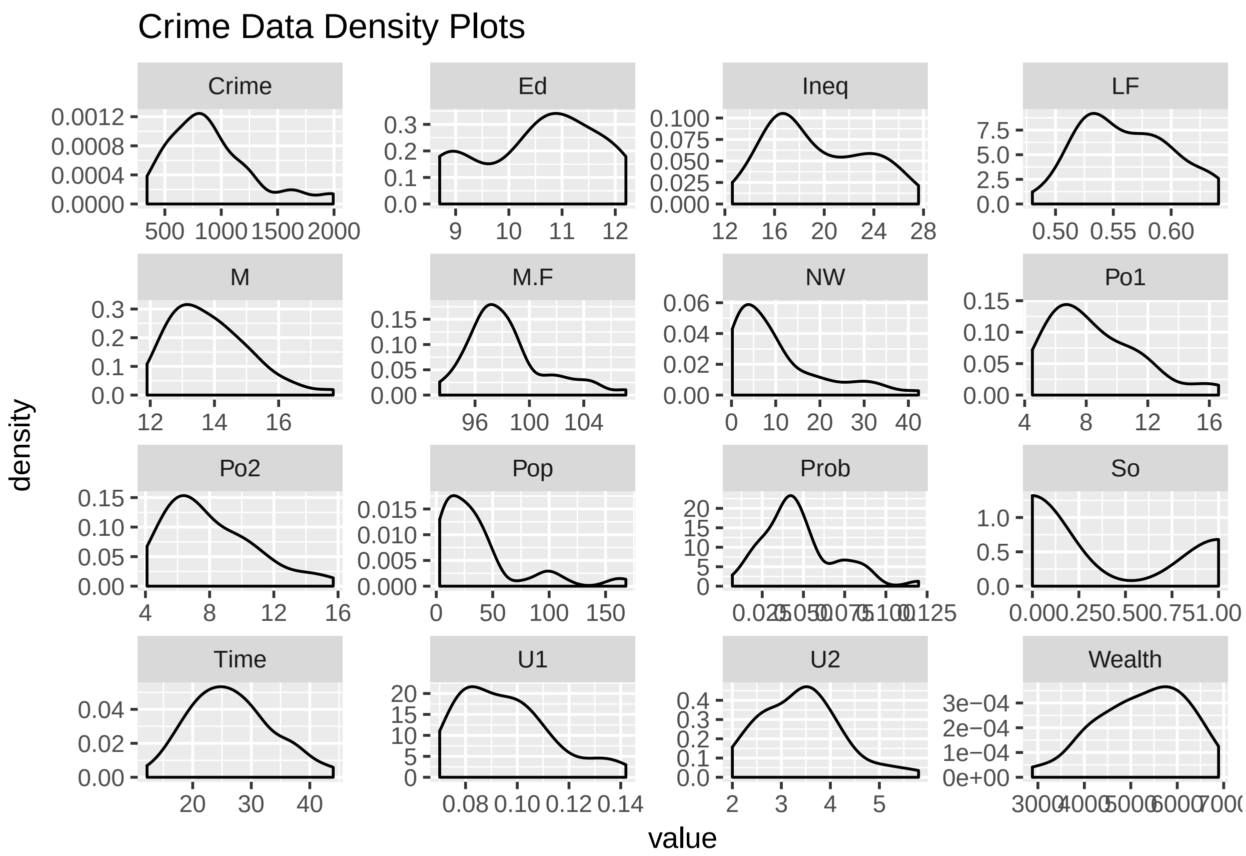
Table 2: Summary of Crime Data

Statistic	N	Mean	St. Dev.	Min	Max
X1996	123	83.715	8.548	60	99
X1997	123	81.675	9.319	55	95
X1998	123	84.260	6.409	63	95
X1999	123	83.358	9.723	57	99
X2000	123	84.033	9.519	55	101
X2001	123	81.553	8.225	51	93
X2002	123	83.585	9.426	57	97
X2003	123	81.480	7.018	57	91
X2004	123	81.764	6.663	62	95
X2005	123	83.358	7.733	54	94
X2006	123	83.049	9.794	53	98
X2007	123	85.398	9.033	59	104
X2008	123	82.512	8.733	50	95
X2009	123	80.992	9.013	51	95
X2010	123	87.211	7.445	67	97
X2011	123	85.276	9.931	59	99
X2012	123	84.650	9.252	56	105
X2013	123	81.667	7.727	56	92
X2014	123	83.943	6.591	63	95
X2015	123	83.301	8.709	56	97

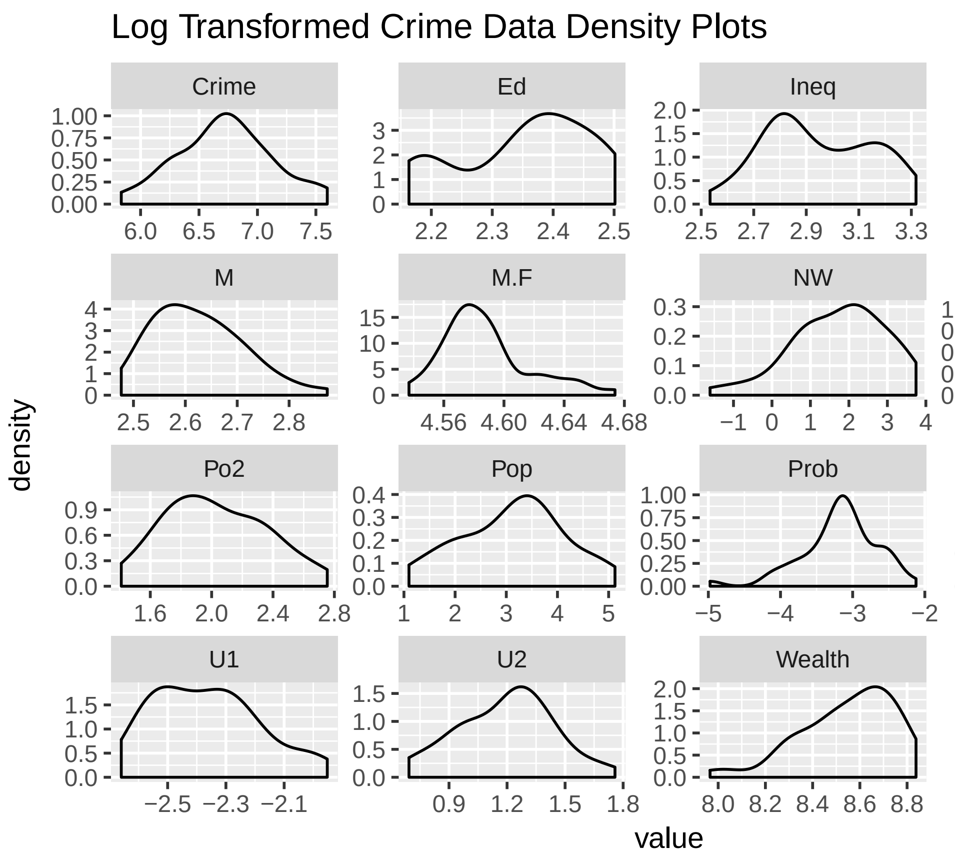
Crime Data Exploration and Transformation

linear models work best with gaussian distributions. Many of the predictors are skewed so I log transformed them to better fit a gaussian distribution.

```
crimeData %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_density() +
    labs(title = "Crime Data Density Plots")
```



```
#log transformation to better fit a gaussian distrubution - did not transform column 'SO' as it is logi
log(crimeData[,c(1,3:16)]) %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_density() +
    labs(title = "Log Transformed Crime Data Density Plots")
```



```
#Building New Log dataset
logCrimeData <- log(crimeData[,c(1,3:16)])
logCrimeData$So <- crimeData$So
```

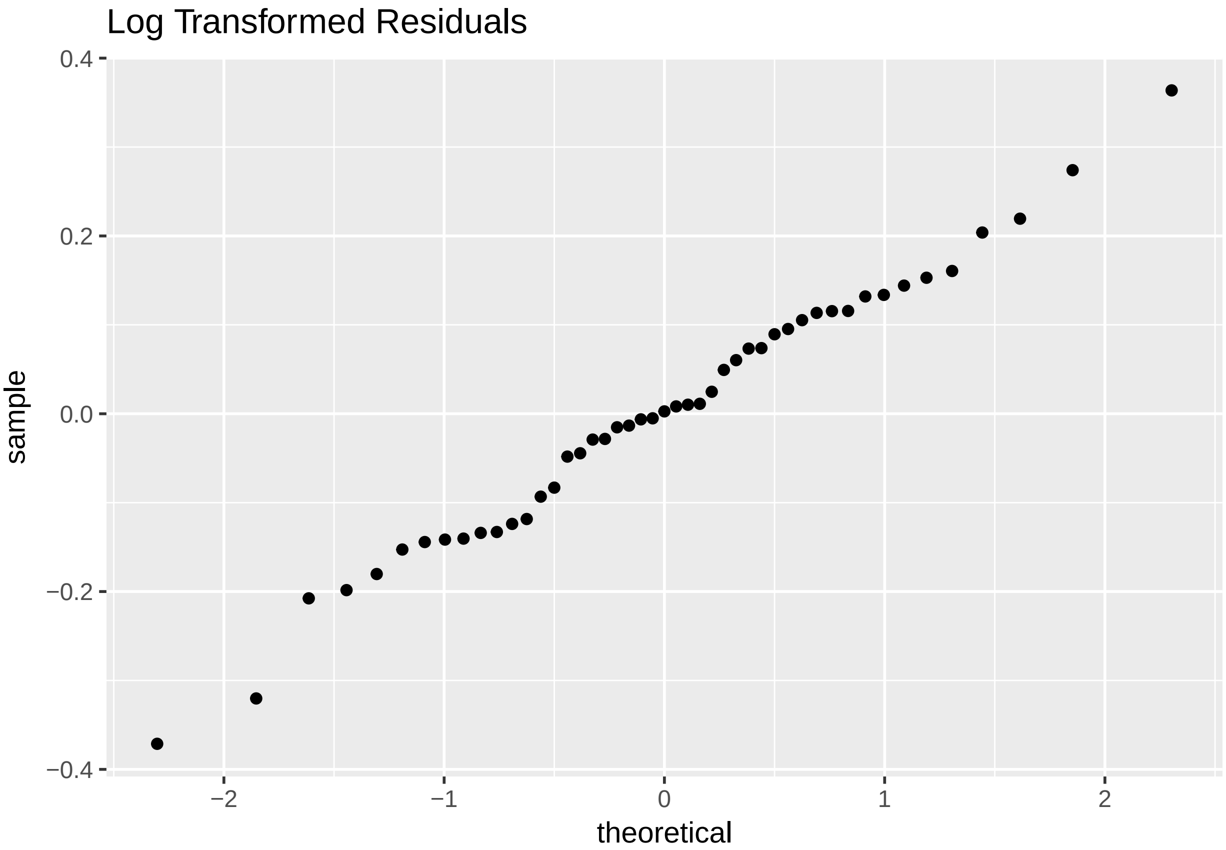
Crime Data Linear Model

After settling on a log transformed dataset, I will now build the model.

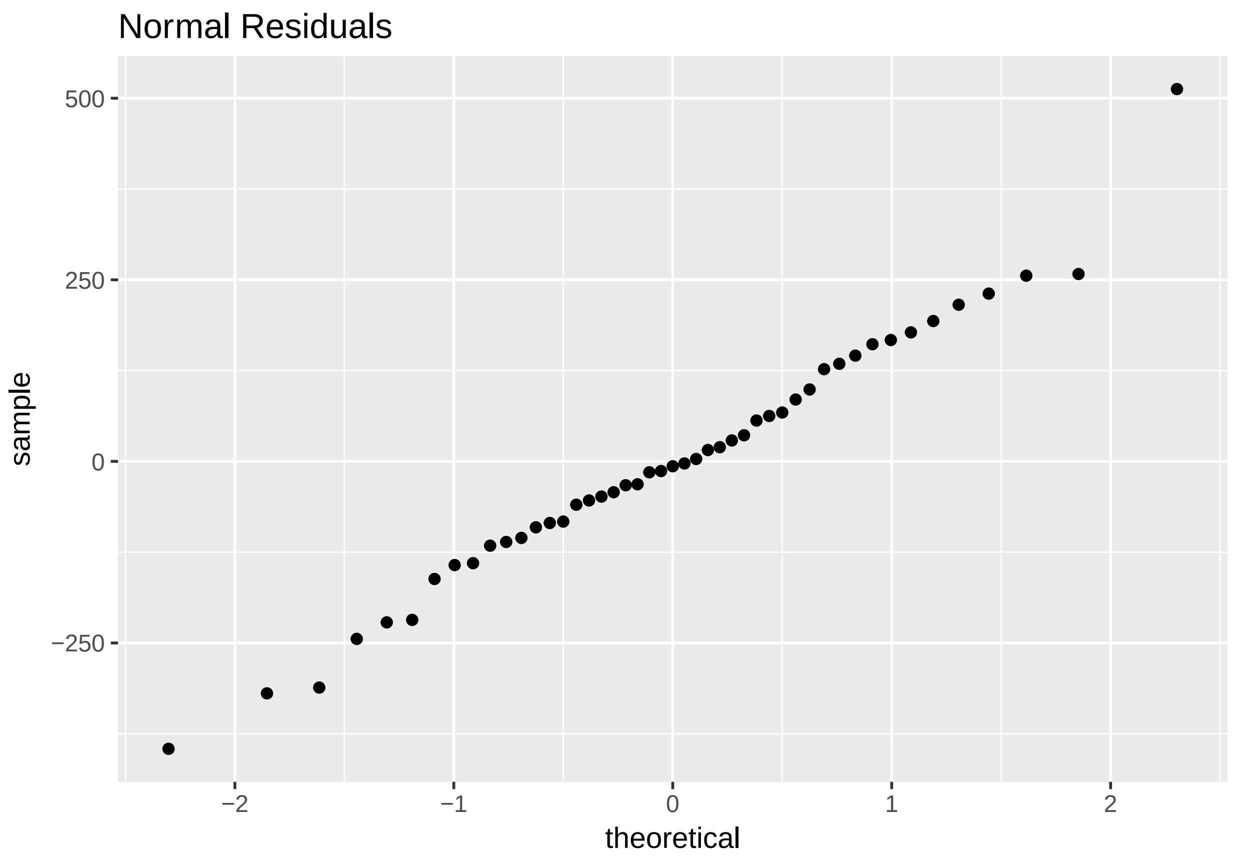
```
crimeModelLog <- lm(Crime ~ ., data = logCrimeData)
crimeModel <- lm(Crime ~ ., data = crimeData)

# Obtain predicted and residual values
logCrimeData$predicted <- predict(crimeModelLog)
logCrimeData$residuals <- residuals(crimeModelLog)

crimeData$predicted <- predict(crimeModel)
crimeData$residuals <- residuals(crimeModel)

#Creating residual df for plotitng
modelResiduals <- data.frame(data=(cbind(residuals(crimeModel),residuals(crimeModelLog))))
colnames(modelResiduals) <- c('Normal', 'Log')
```

plotting the residuals

```
#qqplots to determine if residuals are normally distributed. Log trasnformed model looks more normally
modelResiduals %>%
  ggplot(aes(sample=modelResiduals$Log)) +
  stat_qq() +
  labs(title = "Log Transformed Residuals")
```



```
modelResiduals %>%
  ggplot(aes(sample=modelResiduals$Normal)) +
  stat_qq() +
  labs(title = "Normal Residuals")
```


Normal Residuals
500
250
0
−250
sample
−2
−1
0
1
2
theoretical