cor-SR

sr

```
library(readxl)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

1. load data tried load the original xls, didn't work, don't know why. saved the last worksheet to an independent xlsx

```
baseDF <- read_excel("./data/Book1.xlsx")
```

2. basic cleaning

```
#only work on my varialbes
srDF <- select(baseDF, Ward, Borough, `Unemployment rate 2009`:`Crime rate - 2013`, `GCSE point scores
#drop the NA row
srDF <- srDF[-1,]
#drop last few rows with words
nrows <- dim(srDF)[1]
srDF <- srDF[1:(nrows-4),]
srDF[,3:27] <- sapply(srDF[3:27],as.numeric)
#calculate avg across years
avg_names <- c("avg_unemployment", "avg_crime","avg_GCSE", "avg_schoolAbsence", "avg_dependentChild")
for (i in 0:4) {
  srDF[,avg_names[i+1]] <- rowMeans(srDF[,(i*5+3):(i*5+7)])
}
```

3. calculate correlation

```
corMat <- data.frame(cor(srDF[,28:32]))
corMat
```

```
##                    avg_unemployment  avg_crime   avg_GCSE
## avg_unemployment          1.0000000  0.4904946 -0.7004422
## avg_crime                 0.4904946  1.0000000 -0.4248357
## avg_GCSE                 -0.7004422 -0.4248357  1.0000000
## avg_schoolAbsence         0.6516553  0.4553845 -0.7400082
## avg_dependentChild        0.8385792  0.4915934 -0.7792106
##                    avg_schoolAbsence avg_dependentChild
## avg_unemployment           0.6516553          0.8385792
## avg_crime                  0.4553845          0.4915934
## avg_GCSE                  -0.7400082         -0.7792106
## avg_schoolAbsence          1.0000000          0.7351977
```

```
## avg_dependentChild         0.7351977          1.0000000
```

findings: 1. crime shows moderate correlation with other 4 variables (why? need split down & find evidence) 2. other 4 show high correlation in between try with log: more significant improvement

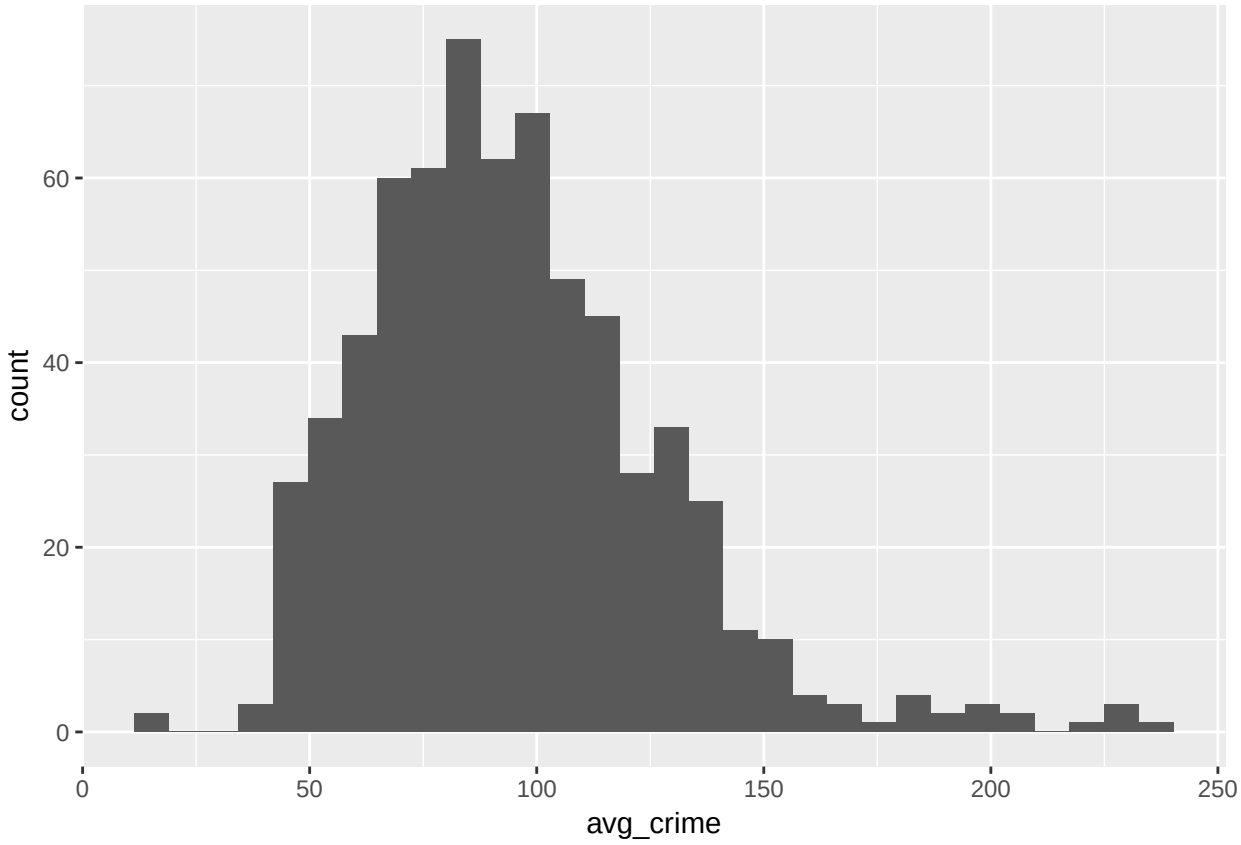
```
corMatLog <- data.frame(cor(log10(srDF[,28:32])))
corMatLog
```

```
##                     avg_unemployment  avg_crime   avg_GCSE
## avg_unemployment          1.0000000  0.5533784 -0.7496962
## avg_crime                 0.5533784  1.0000000 -0.4583187
## avg_GCSE                 -0.7496962 -0.4583187  1.0000000
## avg_schoolAbsence         0.7172121  0.5243094 -0.7457549
## avg_dependentChild        0.8857623  0.5143908 -0.7837564
##                     avg_schoolAbsence avg_dependentChild
## avg_unemployment            0.7172121          0.8857623
## avg_crime                   0.5243094          0.5143908
## avg_GCSE                   -0.7457549         -0.7837564
## avg_schoolAbsence           1.0000000          0.7742086
## avg_dependentChild          0.7742086          1.0000000
```

4. plotting univariable:

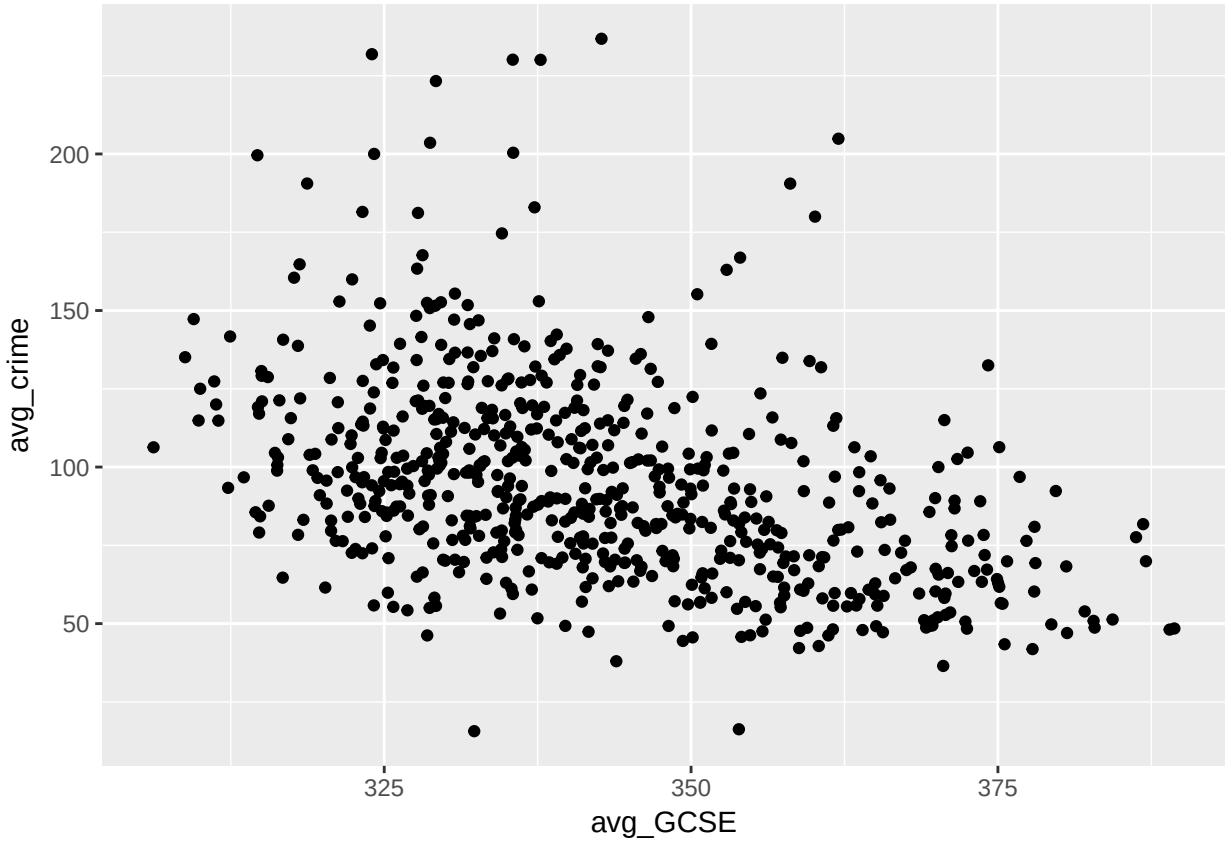
```
ggplot(srDF,aes(x=avg_crime))+geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

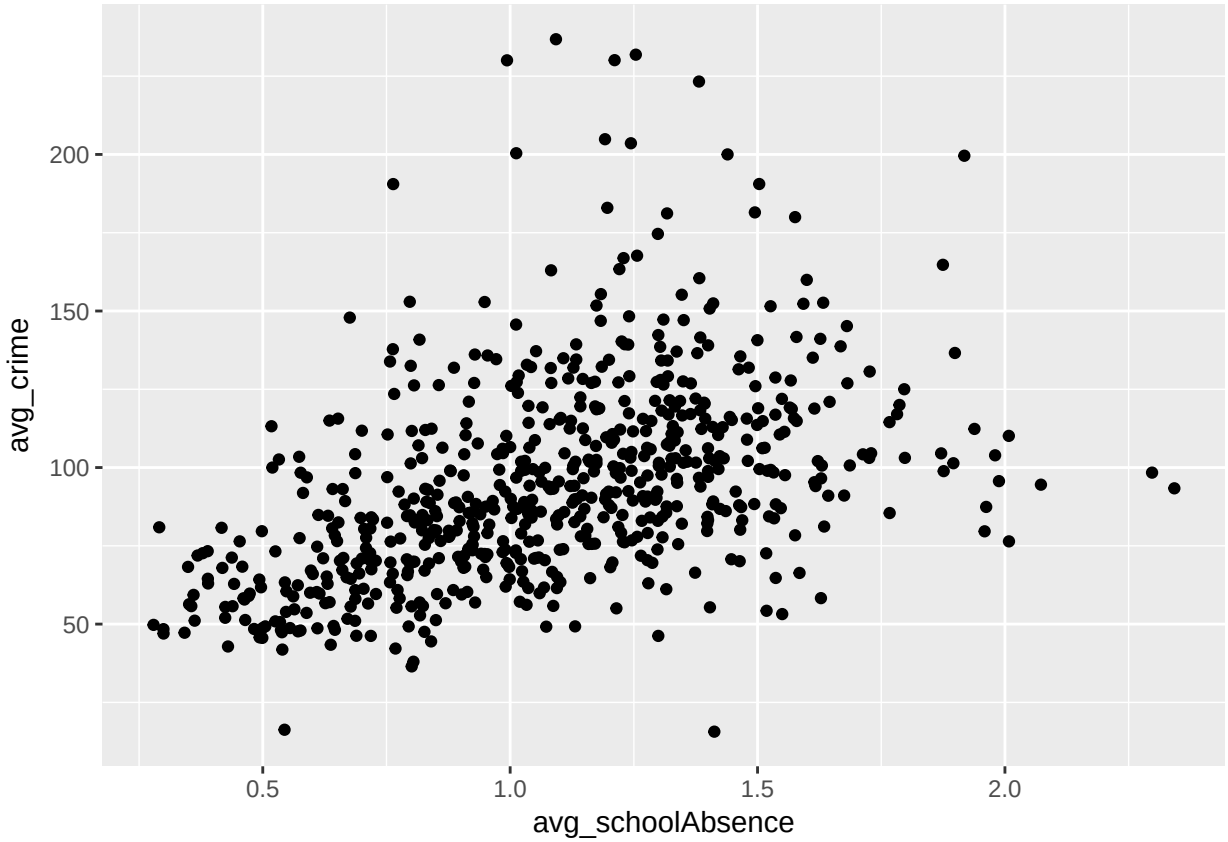


crime rate with positive skewness (in which case log transformation makes sense)

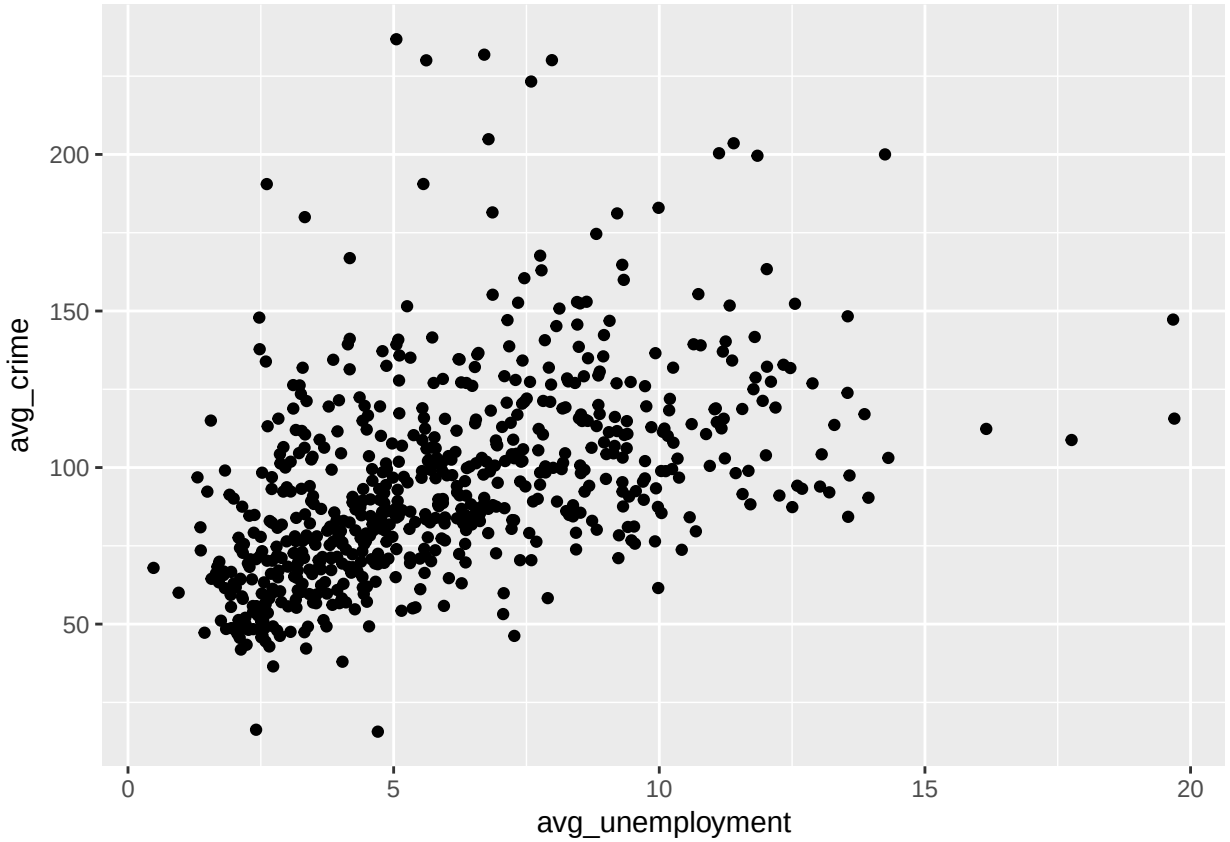
```
ggplot(srDF,aes(x=avg_GCSE,y=avg_crime))+geom_point()
```



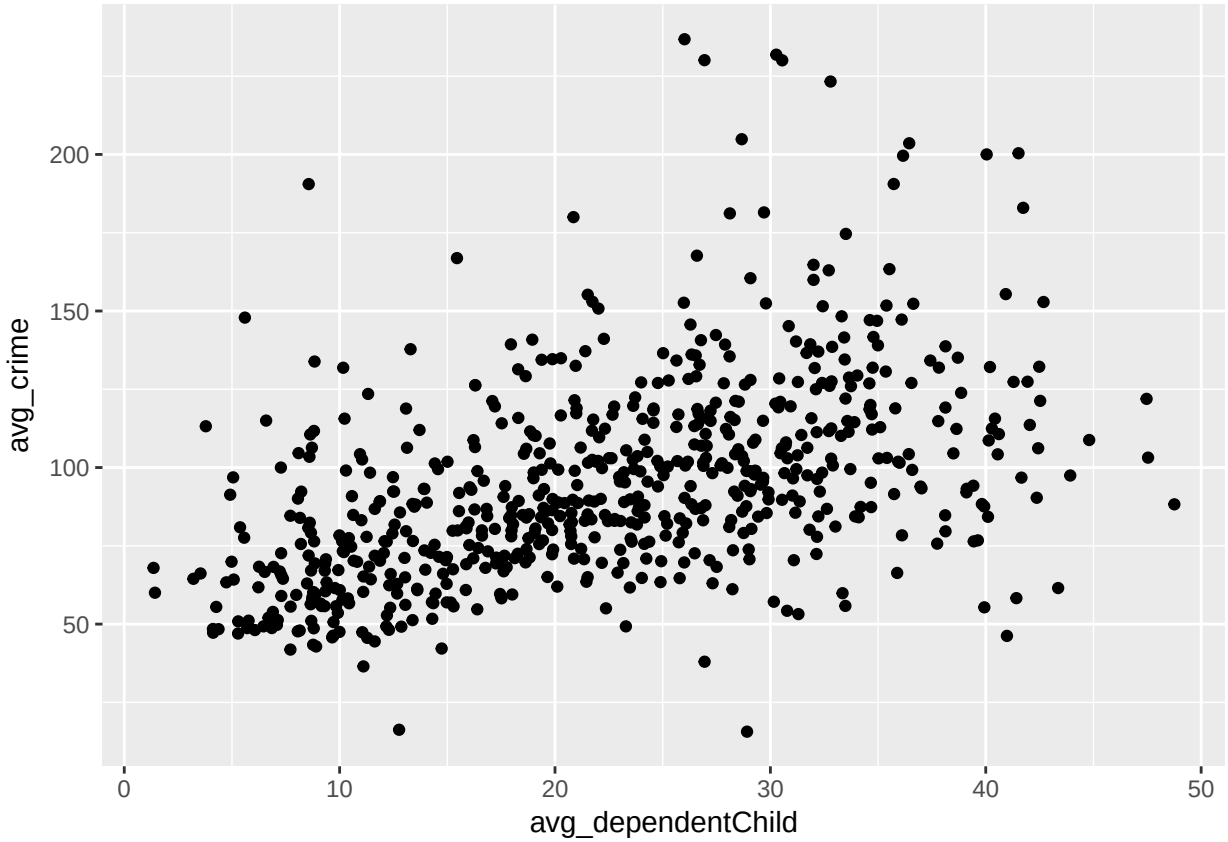

```
ggplot(srDF,aes(x=avg_schoolAbsence,y=avg_crime))+geom_point()
```



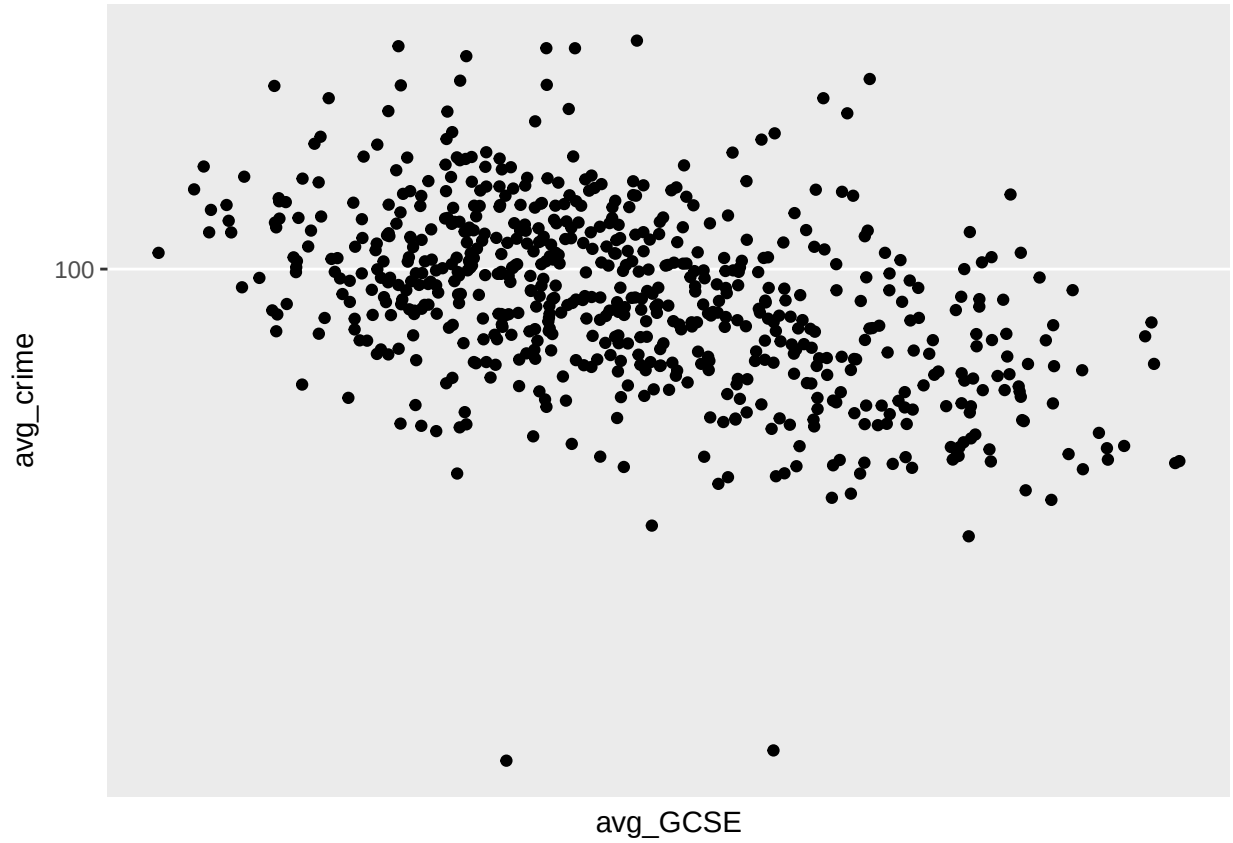

```
ggplot(srDF,aes(x=avg_unemployment,y=avg_crime))+geom_point()
```



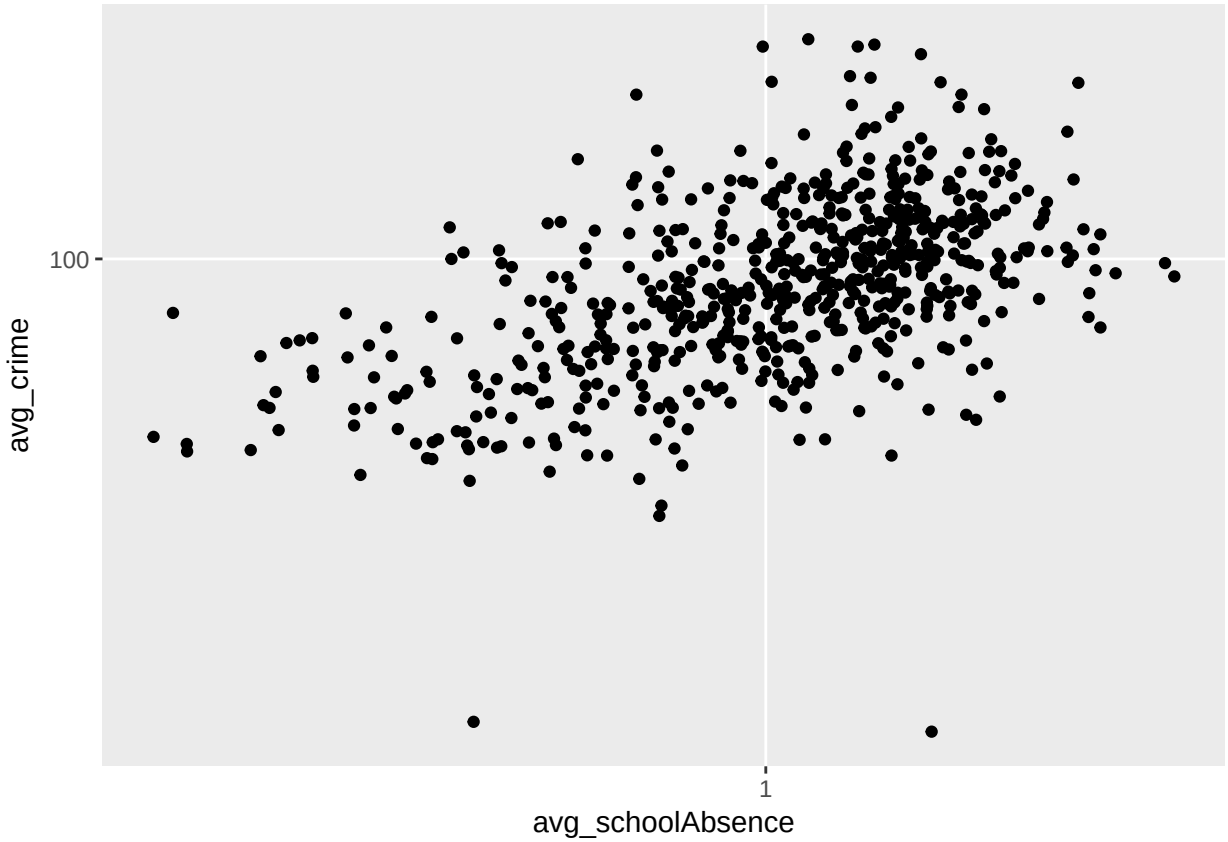

```
ggplot(srDF,aes(x=avg_dependentChild,y=avg_crime))+geom_point()
```



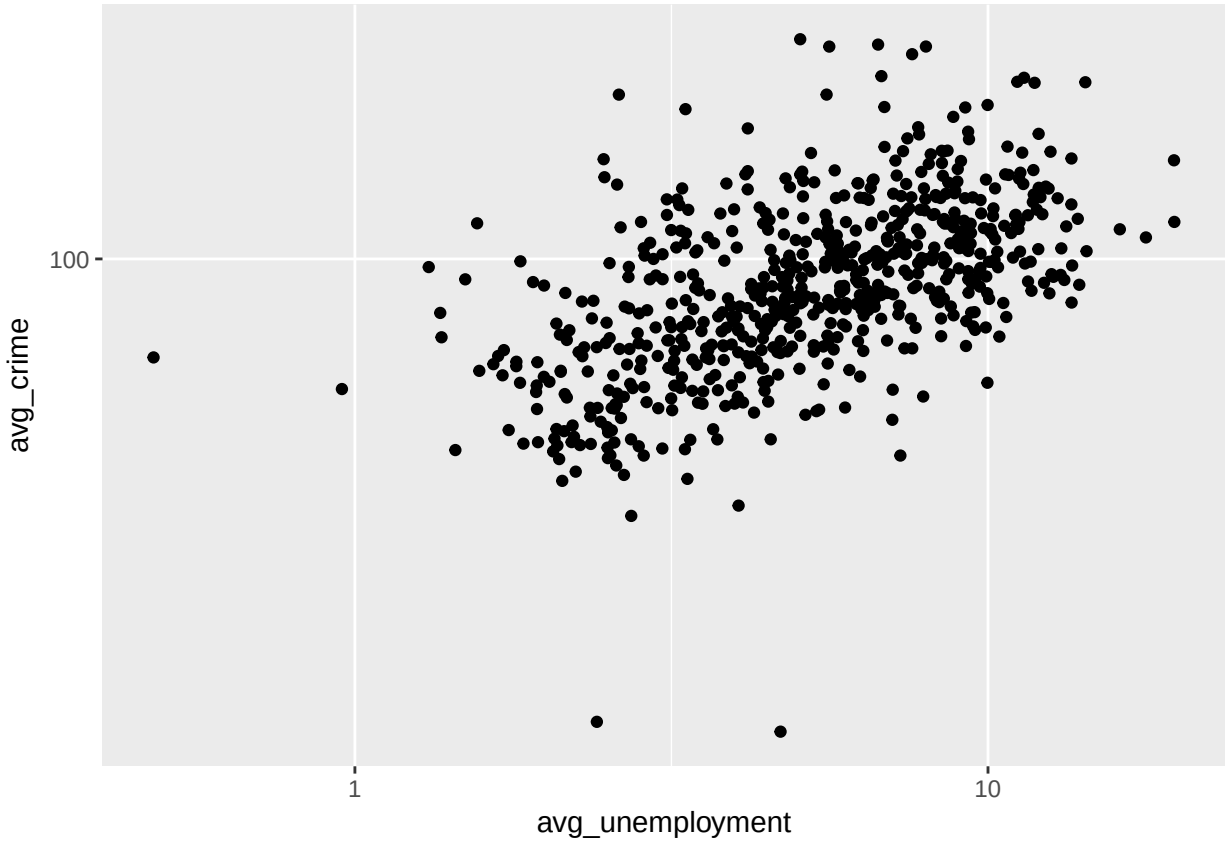

```
ggplot(srDF,aes(x=avg_GCSE,y=avg_crime))+geom_point()+scale_x_log10() + scale_y_log10()
```



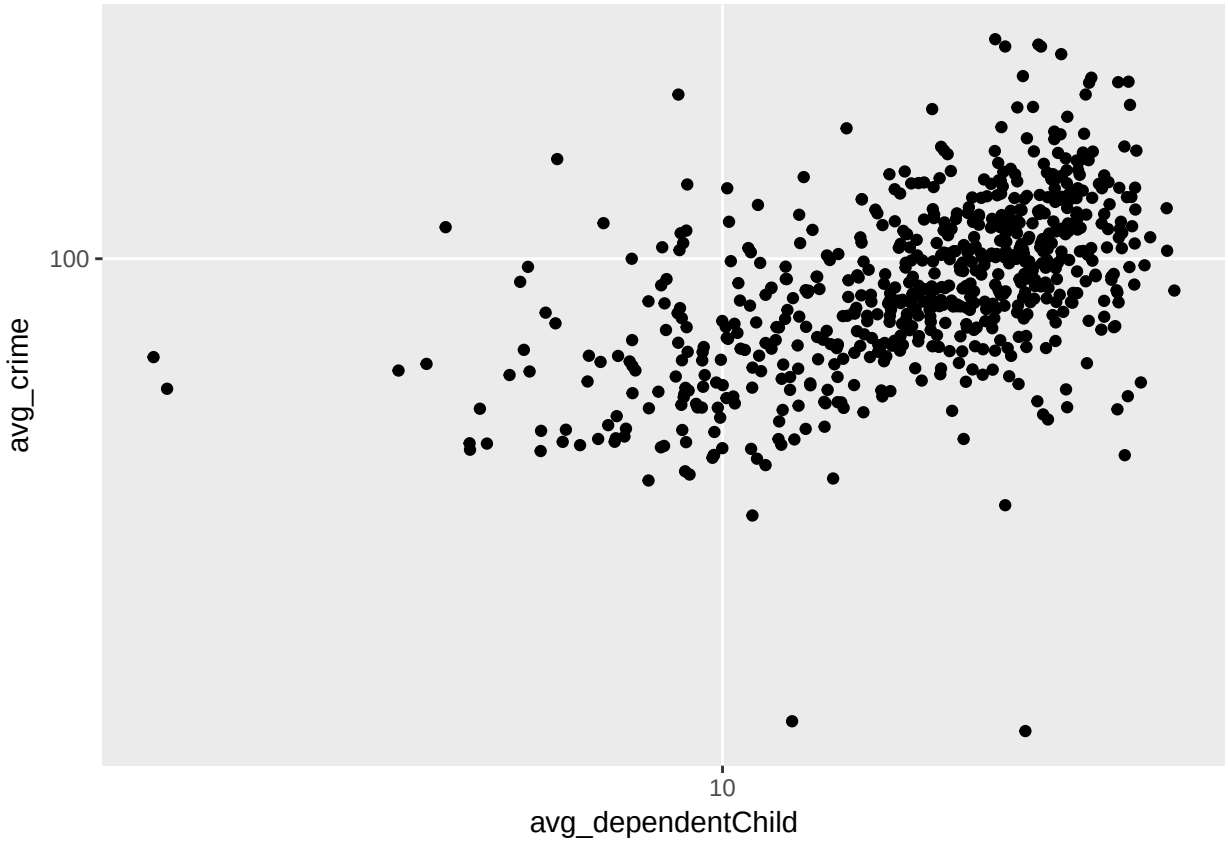

```
ggplot(srDF,aes(x=avg_schoolAbsence,y=avg_crime))+geom_point()+scale_x_log10() + scale_y_log10()
```




```
ggplot(srDF,aes(x=avg_unemployment,y=avg_crime))+geom_point()+scale_x_log10() + scale_y_log10()
```




```
ggplot(srDF,aes(x=avg_dependentChild,y=avg_crime))+geom_point()+scale_x_log10() + scale_y_log10()
```



plots with log transformation looks nicer

5. for discussion

a. do we need to dig deeper on why other variables are not HIGHLY correlated with crime rate? (e.g. split down by type of crimes, divide boroughs into groups) => takes time, may not find good answer
b. do we do log transformation? I think necessary for regression part