Linear Regression Crime Prediction:

Dataset: This set includes data on 50 small US cities involving crime, and police expenditures.
Problem Statement: Using a random number seed, take a random sample of 12 cases from the master data and build a Simple Linear Regression Model in R and answer the following questions.

Analysis

1. Calculate and report the correlation coefficient between the two variables. Report and interpret the p values for the correlation coefficients.
2. Create a scatterplot of the data. Show police funding on the x-axis and reported crimes per 100,000 on the y-axis. x-axis is scaled between 0 and 100 and the y-axis is scaled between 300 and 2000.
3. Conduct a simple linear regression on the data with reported crimes per 100,000 as the dependent variable and police funding as the independent variable. As a part of this:
    a. Report the beta coefficients and associated p values and confidence intervals from your model.
    b. Assess model's conformance with the LINE assumptions of regression.
    c. For a given small city spending $41 per resident on police protection use your model to predict crime rate per 100,000 residents. Include a 95% prediction interval and an interpretation of both the prediction and the accompanying interval. If looked at this interval, what would it potentially indicate about model fit?
4. New York City's budgeted police expenditures in 2020 are $10.9 billion. Its estimated population for the same year is approximately 8,550,000. Give two reasons why it would be wrong to use this model to predict New York's crime rate per 100,000 residents.

I. Preprocessing

```
#Author: Suryateja Chalapati

#Importing required libraries
rm(list=ls())
library(rio)
library(moments)
library(dplyr)
library(MASS)
```

```
#Setting the working directory and importing the dataset
setwd("C:/Users/surya/Downloads")

df = import("Crime Data.xlsx", sheet = "Sheet1")
colnames(df)=tolower(make.names(colnames(df)))
attach(df)

#Setting seed and data sampling
set.seed(36991670)
data_sample = data.frame(df[sample(1:nrow(df), 12, replace = FALSE),])
data_sample <- data_sample %>% rename(crimes = reported.crimes.per.million,
funding=police.funding.dols.per.resident)
attach(data_sample)
```

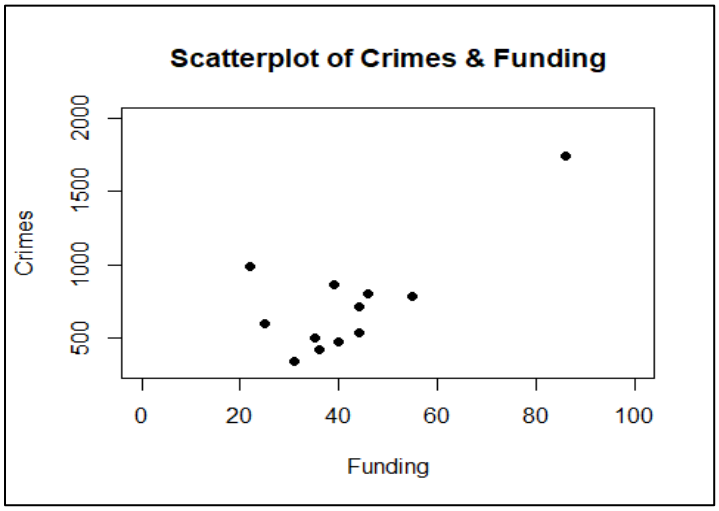
II. Analysis

```
#Analysis_1
cor(data_sample$funding,data_sample$crimes)

## [1] 0.7314597
```

- The correlation coefficient (R) is 0.73. This tells us that there is a strong uphill positive linear relationship. P-value is 0.007, which tells us there is a noticeable relationship between the variables.

```r
#Analysis_2
#Scatter plot
plot(funding, crimes, main="Scatterplot of Crimes & Funding",
     xlab="Funding", ylab="Crimes", pch=19, xlim=c(0,100),ylim=c(300,2000))
```



- Based on the scatterplot above, there seems to be a linear relationship but that's because of a single odd point far out. That might pull the regression line towards itself. But overall there is a linear relationship.

```r
#Analysis_3
#Simple Regression
lin_reg=lm(crimes~funding,data=data_sample)
summary(data_sample$crimes)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   341.0   499.0   659.0   732.7   820.5  1740.0
```

```r
summary(lin_reg)
```

```
##
## Call:
## lm(formula = crimes ~ funding, data = data_sample)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -225.81 -211.95  -82.55  156.25  582.79
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   45.603    216.640   0.211  0.83750
## funding       16.391      4.832   3.392  0.00686 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 266.3 on 10 degrees of freedom
## Multiple R-squared:  0.535,  Adjusted R-squared:  0.4885
## F-statistic: 11.51 on 1 and 10 DF,  p-value: 0.006861
```

```r
confint(lin_reg)
```

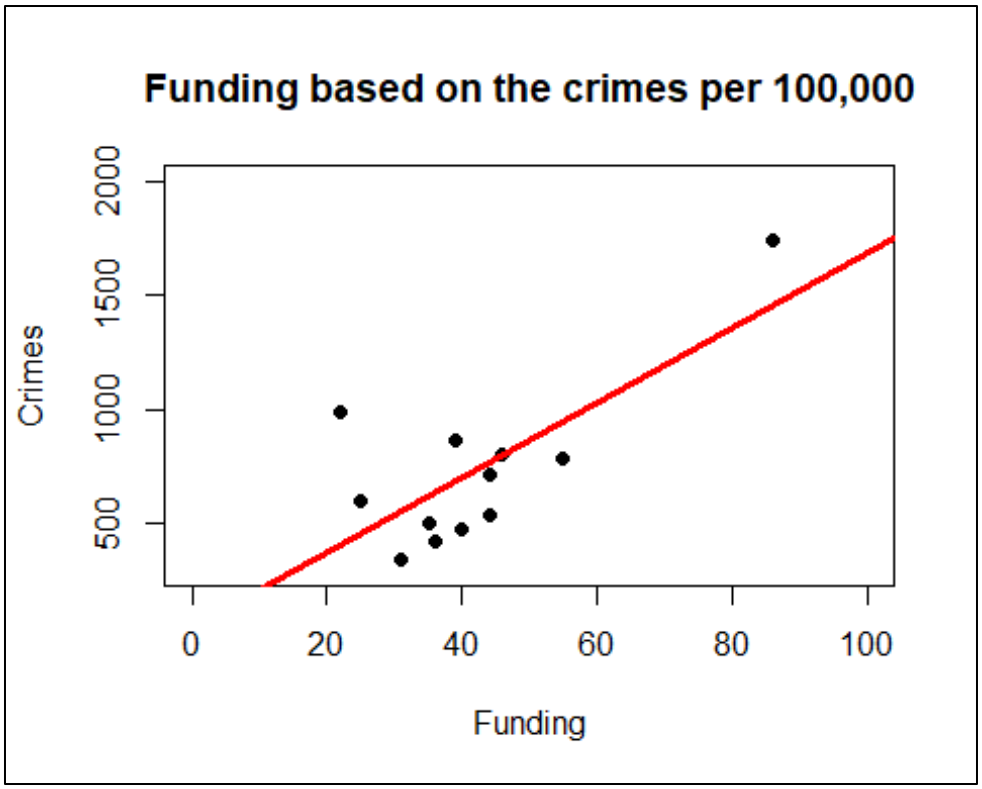
```
##                   2.5 %    97.5 %
## (Intercept) -437.100492 528.30605
## funding        5.624719  27.15765
```

From the Liner Regression analysis:

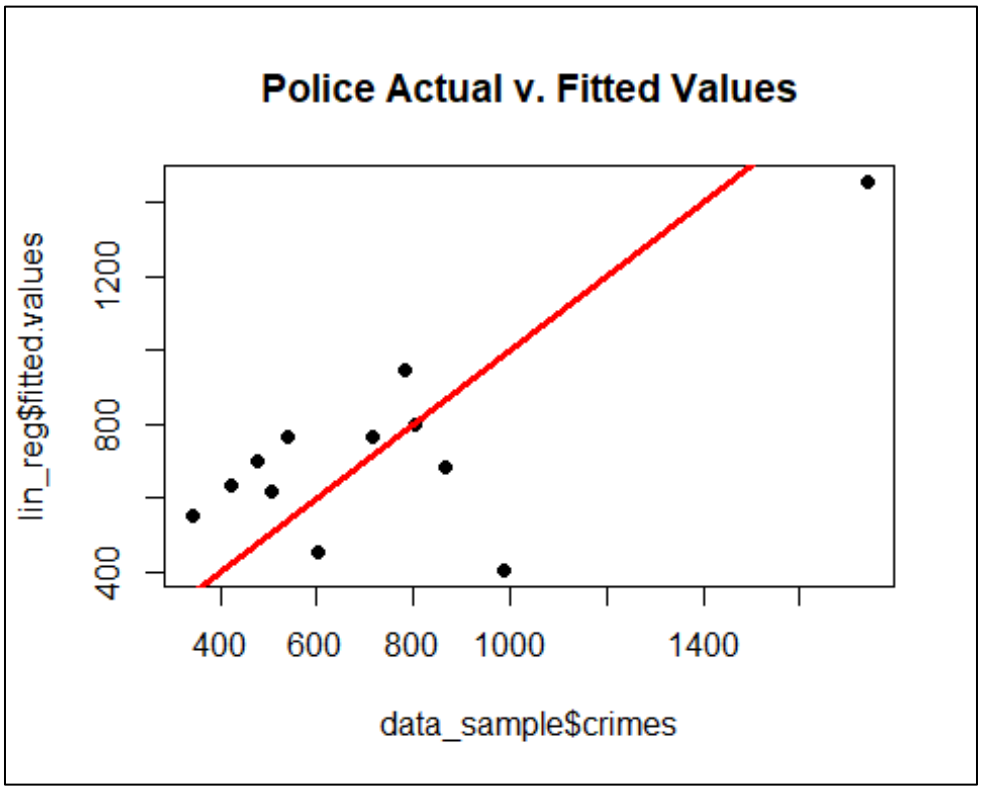
a. The estimated B0 – Coefficient [Intercept] = 45.06 and B1 – Coefficient [Slope] = 16.3. For the intercept the p-value is 0.83. This states we failed to reject the null, so the B0 could be zero or could be any value. For slope intercept p-value is 0.006. Here we can reject the null and conclude the B1 is a positive value.
b. The Regression Equation is **y [Crimes] = 45.603 + 16.391*Funding**.
c. The R-sq is 0.53, that means X explains about 53% of variation in Y.
d. There is a direct proportionality between the two variables.
e. The CI of Intercept is [-437.1, 528], which is a broad range so we cannot rely on this estimate. The CI of Slope is [5.6, 27.1], still a broad range. Even though we reject the bull hypothesis here, we cannot rely on the estimates here.

```
#L.I.N.E Analysis

plot(data_sample$funding,data_sample$crimes,
     pch=19,main="Funding based on the crimes per 1,000,000",xlim=c(0,100),ylim=c(300,2000),xl
ab="Funding",ylab="Crimes")
abline(lin_reg,col="red",lwd=3)
```

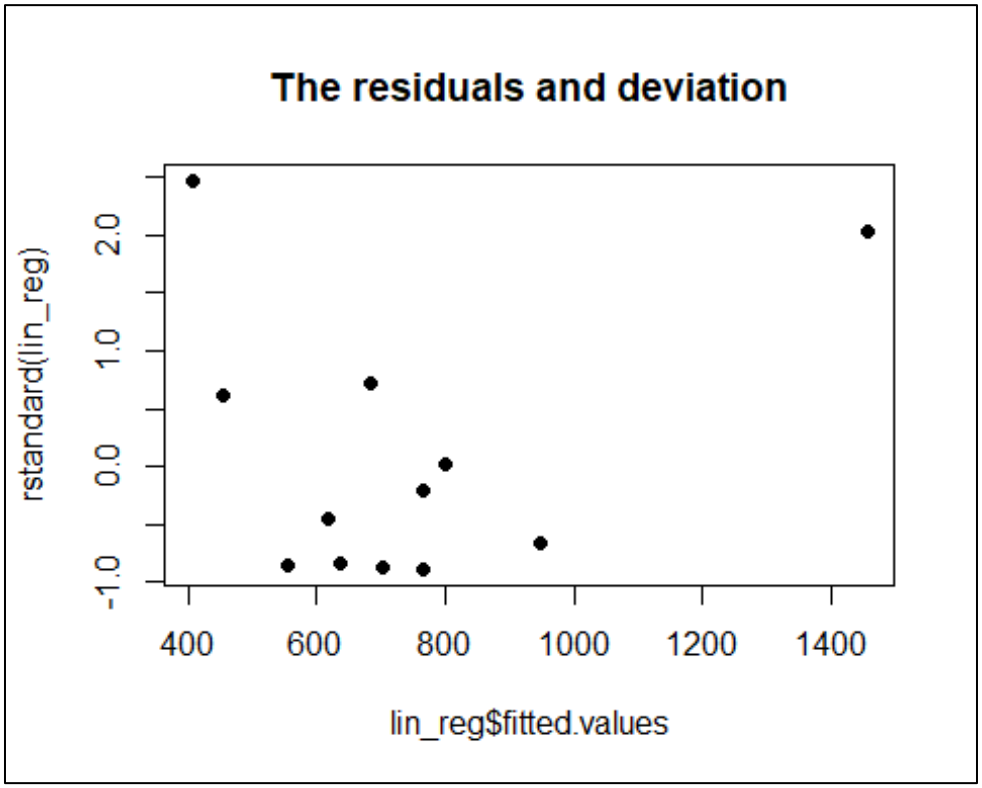



```
#Linearity
plot(data_sample$crimes,lin_reg$fitted.values,pch=19,main="Police Actual v. Fitted Values")
abline(0,1,col="red",lwd=3)
```



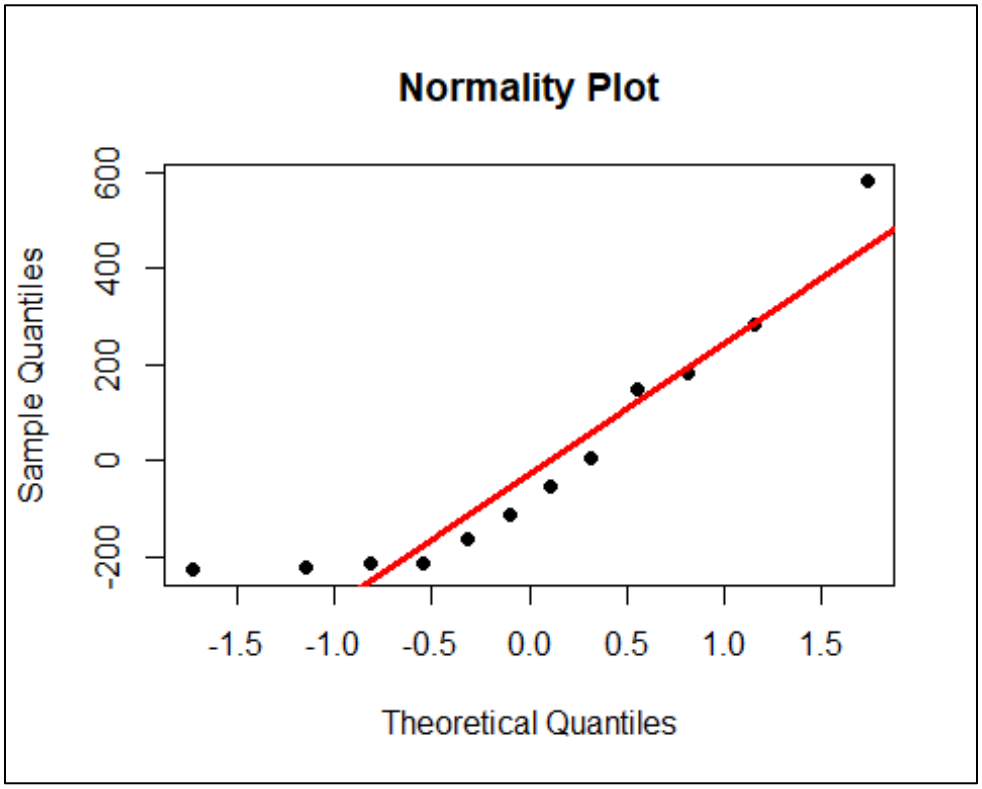
- Data follows linearity but a single outlier pulled the regression line towards itself. There is a positive liner relationship.

```
#Independence
plot(lin_reg$fitted.values,rstandard(lin_reg),pch=19,main="The residuals and deviation")
```



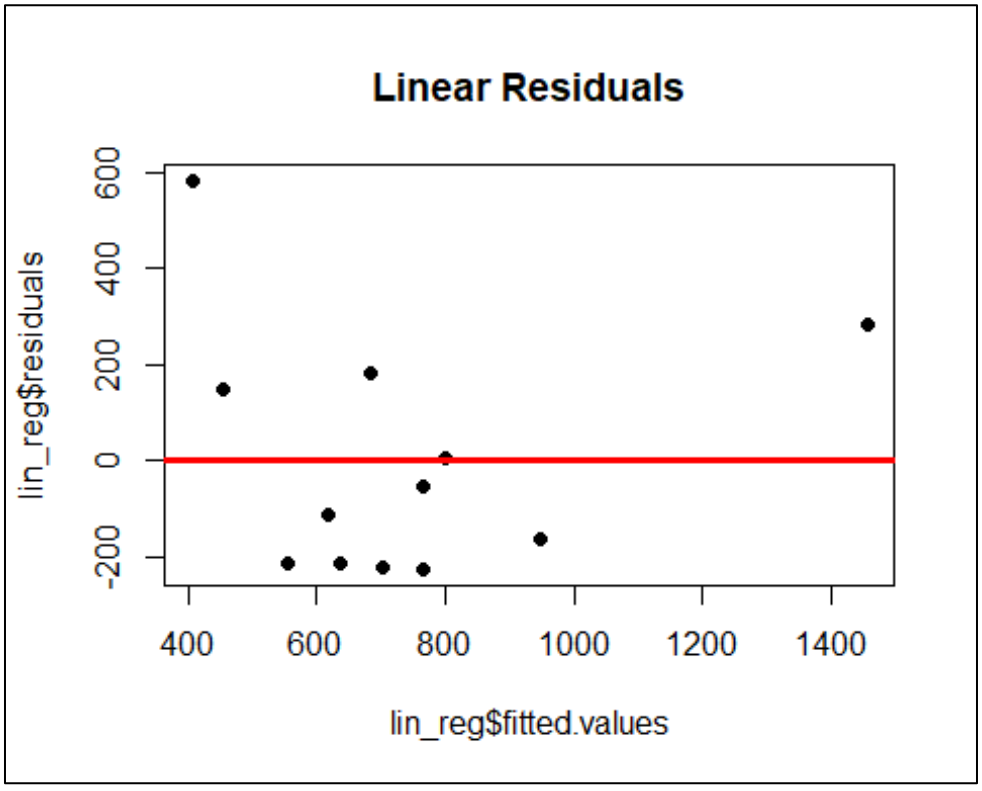
- The fitted vales with their residuals show no pattern and is random. So, it qualifies the Independence test.

```r
#Normality
qqnorm(lin_reg$residuals,pch=19,main="Normality Plot")
qqline(lin_reg$residuals,col="red",lwd=3)
```



- From the plot the data almost follows a normal distribution but there are few values out in the tails.

```r
#Equality of Variances
plot(lin_reg$fitted.values,lin_reg$residuals,pch=19,main="Linear Residuals")
abline(0,0,col="red",lwd=3)
```



- We have an abrupt scattering across the plot. But, we do not have any pattern. So, it satisfies the equality of variance.

```
#Price prediction using Regression Equation
predi=data.frame(funding = 41)
predict(lin_reg, predi, interval = "prediction")
```

```
##        fit      lwr      upr
## 1 717.6414 100.0146 1335.268
```

- Now based on the value, the estimated y value comes out to 717 cases per million, with wide range on the upper and lower limit. The fit is not usable just because of the wide CI range.

```
#Analysis_4
```

- New York City's budgeted police expenditures in 2020 are $10.9 billion. Its estimated population for the same year is approximately 8,550,000. Reasons that this model doesn't fit:
    1. The value of the intercept could be zero or any other value, stating we cannot use the estimated intercept, which is a constant for our predictions.
    2. With a sample size of just 12, we cannot estimate the crimes per million. Its just a very small dataset to predict something this crucial. We also are missing a couple more variables for our estimations.