Report

R-code:

 # Fitting a multiple linear regression model to the data, considering all variables

```
1. data = read.csv("crime.csv")
2. str(data)
```

Output:

```
> str(data)
'data.frame':   50 obs. of  9 variables:
 $ state        : Factor w/ 50 levels "Alabama","Alaska",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ murder.rate  : num  7.4 4.3 7 6.3 6.1 3.1 2.9 3.2 5.6 8 ...
 $ poverty      : num  14.7 8.4 13.5 15.8 14 8.5 7.7 9.9 12 12.5 ...
 $ high.school  : num  77.5 90.4 85.1 81.7 81.2 89.7 88.2 86.1 84 82.6 ...
 $ college      : num  20.4 28.1 24.6 18.4 27.5 34.6 31.6 24 22.8 23.1 ...
 $ single.parent: num  26 23.2 23.5 24.7 21.8 20.8 22.9 25.6 26.5 25.5 ...
 $ unemployed   : num  4.6 6.6 3.9 4.4 4.9 2.7 2.3 4 3.6 3.7 ...
 $ metropolitan : num  70.2 41.6 87.9 49 96.7 84 95.6 81.4 93 69.1 ...
 $ region       : Factor w/ 4 levels "North Central",..: 3 4 4 3 4 4 2 3 3 3 ...
```

R-code:

```
3. attach(data)
4. table(region)
```

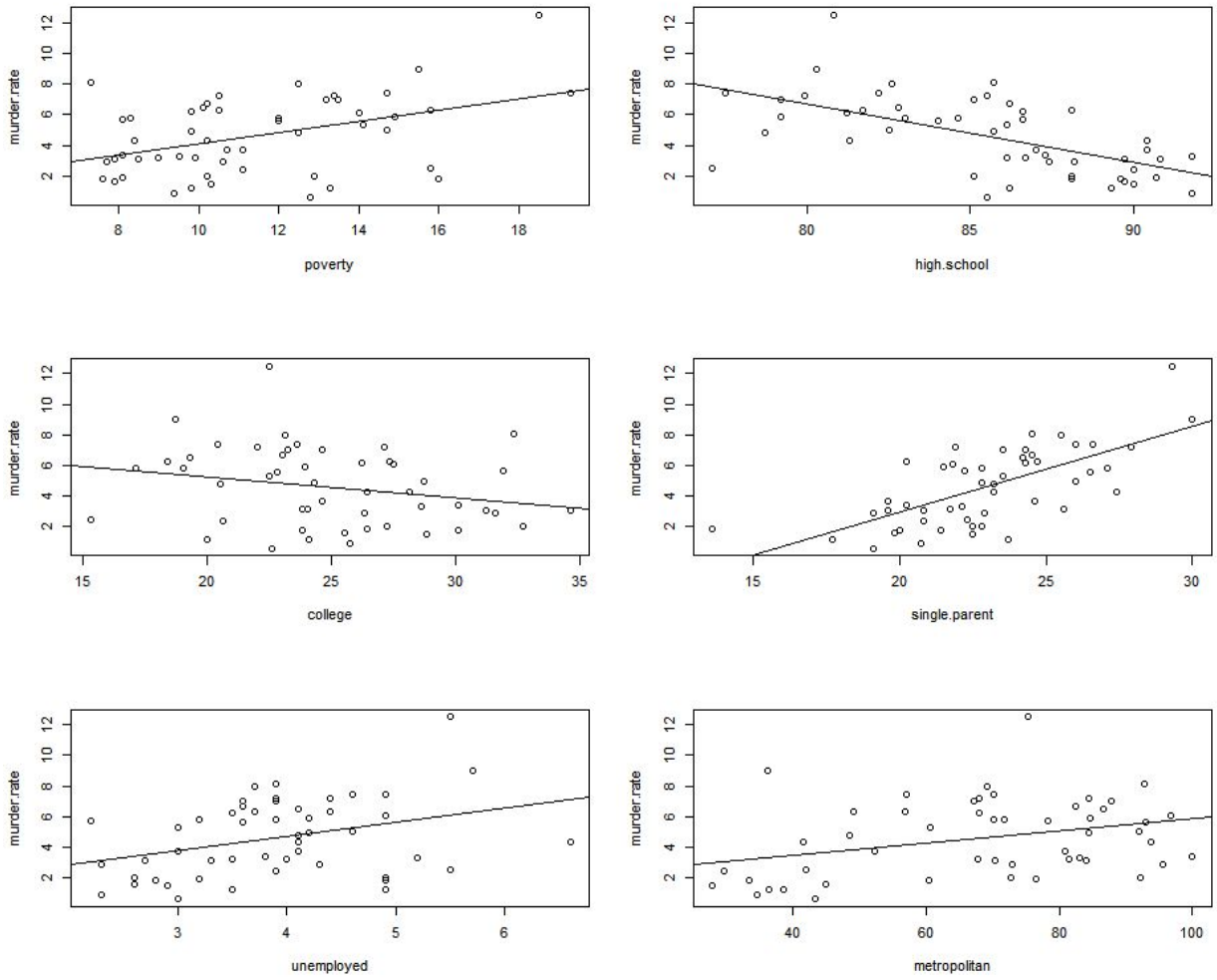
```
> table(region)
region
North Central     Northeast         South          West
           12             9            16            13
```

Observation:
From above we can see the distribution of the **region** variable, it has 4 categories/factors.

```
5. drawPlot = function(x,string){
6.   plot(x,murder.rate,xlab=string)
7.   abline(lm(murder.rate~x))
8. }
9.    # Relationship between murder.rate and each predictor one-by-one.
10.  par(mfrow=c(3,2))
11.  drawPlot(poverty,"poverty")
```

```
12.  drawPlot(high.school,"high.school")
13.  drawPlot(college,"college")
14.  drawPlot(single.parent,"single.parent")
15.  drawPlot(unemployed,"unemployed")
16.  drawPlot(metropolitan,"metropolitan")
```

#Below is the plot of murder.rate and all other variables



 # From above plots, we can see a positive trend in plots for murder.rate vs poverty,single.parent, unemployed and metropolitan.

 # Whereas, plots between murder.rate vs high.school and college show a negative trend

```
# Fitting a multiple linear regression model to the data,
with all predictors together
fit.1 = lm(murder.rate ~ poverty + high.school +college +
unemployed + single.parent + metropolitan + region)
summary(fit.1)
```

Output

```
19.  par(mfrow=c(1,1))
20.  #Model diagnostics for the model
21.  #Residual plot
22.  plot(fitted(fit.1),resid(fit.1))
23.  abline(h=0)
```

 # From the above residual plot, there is no trend and no change in vertical scatter, which verifies the assumption that errors have mean zero and constant variance

```
24.  #QQ normal plot for residuals
25.  qqnorm(resid(fit.1))
26.  qqline(resid(fit.1))
```

#As we can see from the above Normal QQ-plot that our normality assumption is satisfied, so we do not need to do any kind of transformations.

```
27.  #Time series plot for residuals
28.  plot.ts(resid(fit.1))
29.  abline(h=0)
```

 # From the above time series plot, we can say that there is no dependency over time as it seems random, so we can verify our assumption that errors are independent when the data are collected over time.

Conclusion:

- From the observation from the above model diagnostic, we can say that our fitted model is a good representation of the data and follows the standard assumptions for linear models.

(b)

R-code:

```
1. #Remove poverty, high.school, college, unemployed, metropolitan
2. fit.2 = update(fit.1, . ~ . - poverty - high.school - college - unemployed - metropolitan)
3. summary(fit.2)
```

Output:

Observation:

- Here, we have created model with predictor **single.parent** and **region**. This model does not looks good based on R^2 value which is 0.5695. Even though, P values of predictor are near to zero, so predictors are good for this model.

R-code:

```
1. fit.3=update(fit.2, . ~ . + poverty + high.school  )
2. summary(fit.3)
```

Output:

Observation:

- Here, We have added poverty and high.school in fit.2 model. This model has higher R^2 value than fit.2 model.
- But P value for poverty is more than cut off 0.05, so we need to accept the null hypothesis, which says murder.rate doesn't depend on poverty.
- We observe that even increasing R^2 by adding predictors, but both of them are not good fit for model. So we need to update our model to make it more better.

R-code:

```
1. fit.4=update(fit.3, . ~ . - poverty + college +unemployed )
2. summary(fit.4)
```

Output:

Observation:

- Here, We have removed useless predictor from fit.3 model and added college and unemployed in fit.3 model. This model has higher R^2 value than fit.2 model.

- But P value for unemployed is more than cut off 0.05, so we need to accept the null hypothesis, which says murder.rate doesn't depend on unemployed. So we need to update our model to make it more better.

R-code:

```
1. fit.5=update(fit.4, .~. -college-unemployed +metropolitan )
2. summary(fit.5)
```

Output:

Observation:

- Here, we have updated fit.4 model by **removing** useless predictors **college** & **unemployed** and **added metropolitan**. This model has higher R^2 value than fit.3 model.

- But P value for high.school is more than cut off 0.05, so we need to accept the null hypothesis, which says murder.rate doesn't depend on high.school. So we need to update our model to make it more better.

R-code:

```
1. fit.without.high.school = update(fit.5, . ~ . - high.school)
2. summary(fit.without.high.school)
```

Observation:

- Here, We have updated fit.5 model by removing useless predictors high.school. This model has lower R^2 value than fit.5 model.
- But every predictors in this model is good fit for this model. Here one thing we observed that for region, there are two categories regionSouth and regionWest, which is not good fit for model.

R-code:

```
1. fit.without.region = update(fit.without.high.school, . ~ . -
   region)
2. summary(fit.without.region)
```

Observation:

- Here, We have updated fit.without.high.school model by removing **region** predictor. This model has lower R^2 value than fit.without.high.school model. But every predictors in this model is good fit for this model.

R-code:

```
#Now perform a partial F-test to check significance of
high.school
1. anova(fit.5, fit.without.high.school)
```

Output:

Observation:

- Here F test value for fit.5 and fit.without.high.school is same as T test value for high.school as 0.1161

- Null hypothesis, that high.school is not significant, is true. So we can say that high.school is not significant for model.

R-code:

```
anova(fit.without.region,fit.without.high.school)
```

Output:

Observation:

- Here F-test value for fit.without.region and fit.without.high.school is not same as T test value for region.
- Even F test value is significantly near to 0. Null hypothesis, that region is not significant, is false. So we can accept the full model which is fit.without.high.school.
- Thus our **final model is fit.without.high.school(murder.rate ~ single.parent + region + metropolitan)**

#Extra Credit Part:

#fit.without.high.school compare with automatic stepwise model selection procedures based on AIC.

 # Forward selection based on AIC

R-code:

```
fit.forward = step ( lm( murder.rate ~ 1, data=data),   scope = list( upper = ~poverty + high.school + college + single.parent + unemployed + metropolitan + region), direction = "forward")
```

Output:

 # Backward elimination based on AIC

R-code:

```
1. fit.backward = step(lm(murder.rate ~ poverty + high.school +
   college + single.parent + unemployed + metropolitan + region,
   data = data),  scope = list(lower = ~1), direction = "backward")
```

Output:

 # Both forward/backward

1. fit.both = step(lm(murder.rate ~ 1, data=data),scope = list(upper = ~poverty + high.school + college + single.parent + unemployed + metropolitan + region, lower = ~1),direction = "both")

Output:

R-code:

```
anova(fit.without.high.school, fit.both)
```

Output:

Observation:

- We see that direction = backward, direction = both and direction = forward pick the following model:**single.parent + region + metropolitan + high.school**
- By doing partial F-test to check which is better model from fit.without.high.school and fit.both, we observed that F value is significantly more than 0.05 cut off. So our generated model fit.without.high.school is better than fit.both.

#Perform model diagnostics:

 # Residual plot

```
1. plot(fitted(fit.without.high.school),abs(resid(fit.without.high
   .school)))
2. abline(h = 0)
```

 # Normal QQ plot

```
1. qqnorm(resid(fit.without.high.school))
2. qqline(resid(fit.without.high.school))
```

#Time series plot for residuals

```
1. plot.ts(resid(fit.without.high.school))
2. abline(h=0)
```

Observation:

This preliminary model passes the diagnostics. So we can take this as our final model.

(c) Prediction based on obtained model

R-code

```
1. single.parent.mean = mean(single.parent)
2. metro.mean = mean(metropolitan)
3. summary(region)
```

R-code

```
1. region.val = "South"
2. predict(fit.without.high.school,data.frame("single.parent"=
   single.parent.mean,"metropolitan"=metro.mean,"region"=region
   .val))
```

Output:

```
 1
5.428477
```

- murder rate of a state whose predictor values are set at the average in the data for single.parent and metropolitan predictor and the most frequent category for a region is South is 5.428477.