Exercises

1. Download boston dataset from eLearn@USM

In [91]:
```
boston_data <- read.csv("boston.data.csv", sep=",",  stringsAsFactors=TRUE)
```

1. Load the datase

In [92]:
```
summary(boston_data)
```

```
      CRIM                ZN              INDUS            CHAS        
 Min.   :0.00000   Min.   :  0.0   Min.   : 0.000   Min.   :0.0000  
 1st Qu.:0.04944   1st Qu.:  0.0   1st Qu.: 3.440   1st Qu.:0.0000  
 Median :0.14466   Median :  0.0   Median : 6.960   Median :0.0000  
 Mean   :1.26920   Mean   : 13.3   Mean   : 9.205   Mean   :0.1408  
 3rd Qu.:0.81962   3rd Qu.: 18.1   3rd Qu.:18.100   3rd Qu.:0.0000  
 Max.   :9.96654   Max.   :100.0   Max.   :27.740   Max.   :1.0000  

      NOX               RM              AGE              DIS        
 Min.   :0.385   Min.   :  3.561   Min.   :  1.137   Min.   : 1.130  
 1st Qu.:0.449   1st Qu.:  5.962   1st Qu.: 32.000   1st Qu.: 2.431  
 Median :0.538   Median :  6.322   Median : 65.250   Median : 3.926  
 Mean   :1.101   Mean   : 15.680   Mean   : 58.745   Mean   : 6.173  
 3rd Qu.:0.647   3rd Qu.:  6.949   3rd Qu.: 89.975   3rd Qu.: 6.332  
 Max.   :7.313   Max.   :100.000   Max.   :100.000   Max.   :24.000  

      RAD              TAX            PTRATIO             B        
 Min.   :  1.00   Min.   : 20.2   Min.   :  2.60   Min.   :  0.32  
 1st Qu.:  4.00   1st Qu.:254.0   1st Qu.: 17.00   1st Qu.:365.00  
 Median :  5.00   Median :307.0   Median : 18.90   Median :390.66  
 Mean   : 78.06   Mean   :339.3   Mean   : 42.62   Mean   :332.79  
 3rd Qu.: 24.00   3rd Qu.:403.0   3rd Qu.: 20.20   3rd Qu.:395.62  
 Max.   :666.00   Max.   :711.0   Max.   :396.90   Max.   :396.90  

     LSTAT            MEDV      
 Min.   : 1.730   Min.   : 6.30  
 1st Qu.: 6.878   1st Qu.:18.50  
 Median :10.380   Median :21.95  
 Mean   :11.538   Mean   :23.75  
 3rd Qu.:15.015   3rd Qu.:26.60  
 Max.   :34.410   Max.   :50.00  
                  NA's   :54     
```

1. Detect any missing value in the dataset

   - Answer: from the summary of the data we can see there are missing values on `MEDV` column

1. Detect any outlier in column DIS using boxplot

In [120]:
```
library(ggplot2)
library(scales)
library(mgcv)
```

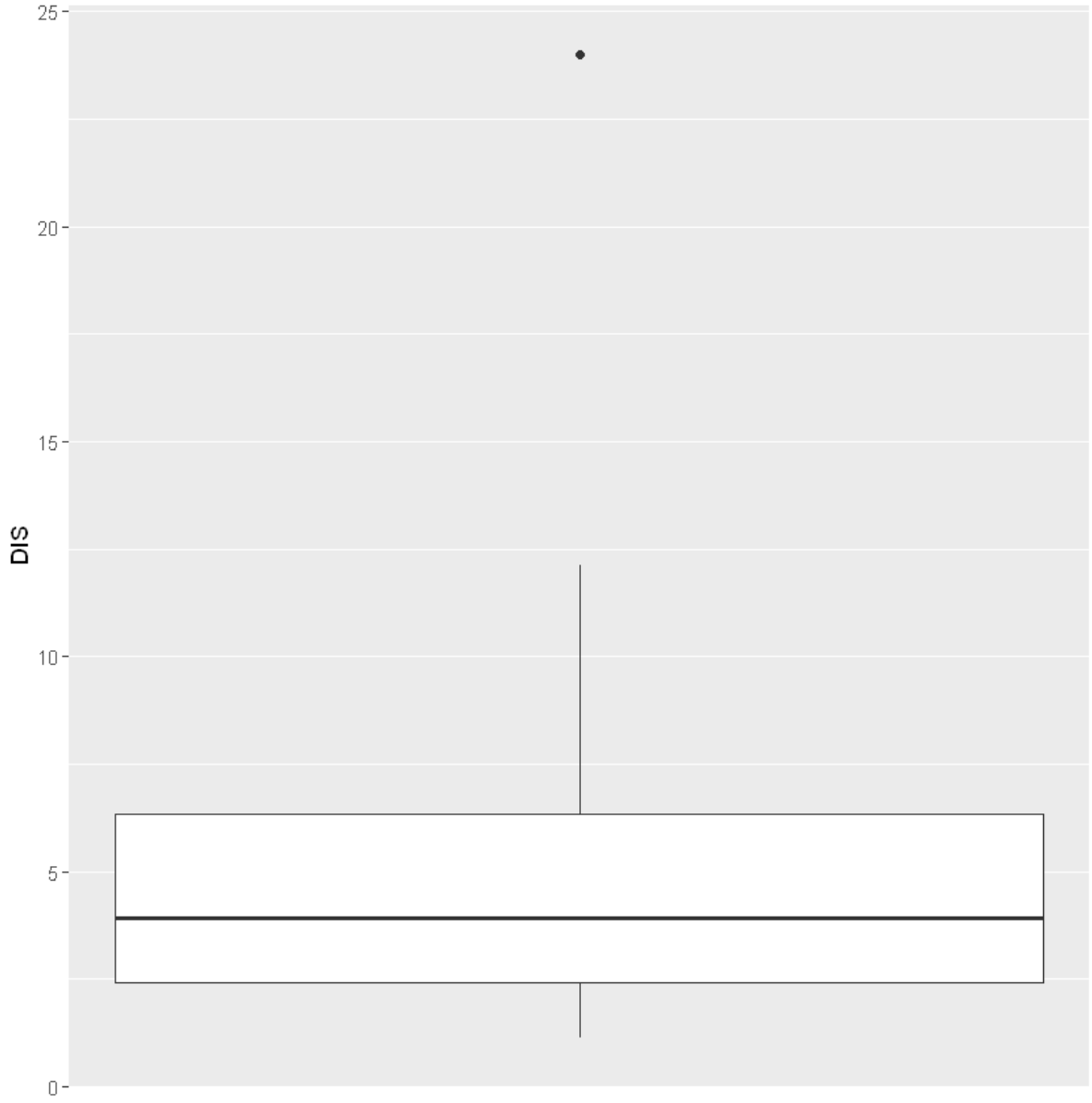
```
Error in library(mgcv): there is no package called 'mgcv'
Traceback:

1. library(mgcv)
```

In [99]:
```
class(boston_data$DIS_df)
```

'data.frame'

In [115]:
```
ggplot(boston_data, aes(x=1, y=DIS)) +
  geom_boxplot() +
  scale_x_continuous(breaks = NULL) + #removes the tick markers from the x axis
  theme(axis.title.x = element_blank())
```



Answer: As shown by the boxplot above, there is one outlier in DIS column

1. Write a programming function for IQR rule. The function should accept first quartile and third quartile as arguments and return the lower and upper bounds as a vector. Use the function to detect any outlier in column LSTAT

$$IQR = Q3 - Q1$$

$Q3 : 75^{th}\ percentile$

$Q1 : 25^{th}\ percentile$

In [64]:
```
# defining and calculating Q1
Q1 <- quantile(boston_data$LSTAT, 0.25)
Q1
```

25%: 6.8775

In [65]:
```
# defining and calculating Q3
Q3 <- quantile(boston_data$LSTAT, 0.75)
Q3
```

75%: 15.015

$$Upper\ bound\ =\ Q3 + (1.5\ x\ IQR)$$
$$Lower\ bound\ =\ Q1 - (1.5\ x\ IQR)$$

In [56]:
```
iqr_bound <- function(Q1, Q3){
    iqr <- (Q3 - Q1)
    low_bound <- (Q1 - (1.5*iqr))
    up_bound <- (Q3 + (1.5*iqr))
cat(paste("iqr = ", iqr, "\nlower-bound = ", low_bound,
          "\nupper-bound = ", up_bound))
}
```

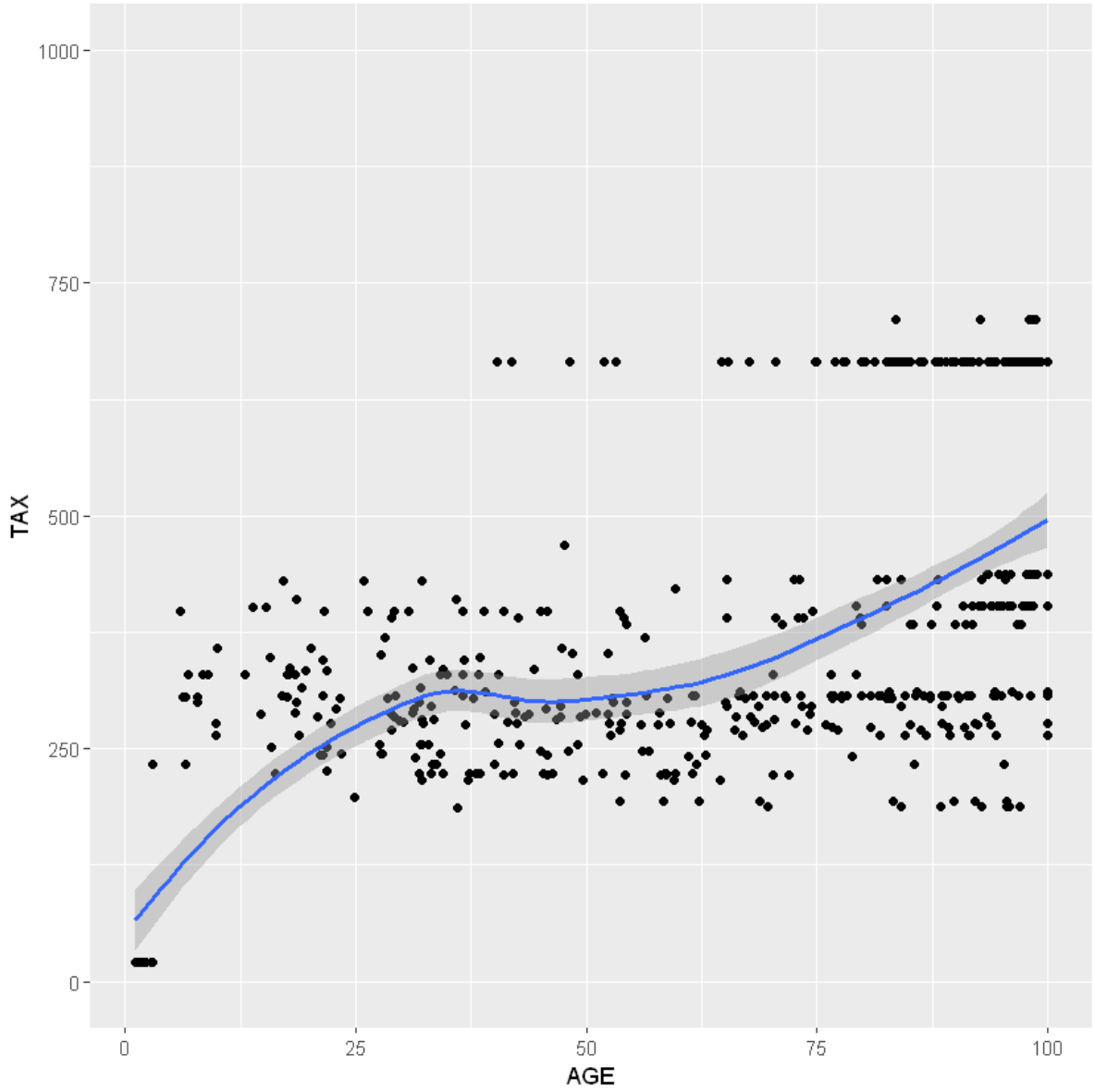
In [55]:
```
iqr_bound(Q1, Q3)
```

```
iqr =  8.1375
lower-bound =  -5.32875
upper-bound =  27.22125
```

1. Examine the relationship between attribute AGE and attribute TAX. Determine if it is a positive or negative correlation or no correlation

In [134]:
```
ggplot(boston_data, aes(x=AGE, y=TAX)) + geom_point() + ylim(0, 1000) + geom_smooth()
```

```
`geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Answer : There is no clear correlation between attribute AGE and TAX

1. Visualize the relationship between AGE and TAX and fit a linear line through the data. Observe the slope of the linear line

In [130]:
```
ggplot(boston_data, aes(x=AGE, y=TAX)) + geom_point() + ylim(0,1000) + stat_smooth(method='lm')
```

```
`geom_smooth()` using formula 'y ~ x'
```