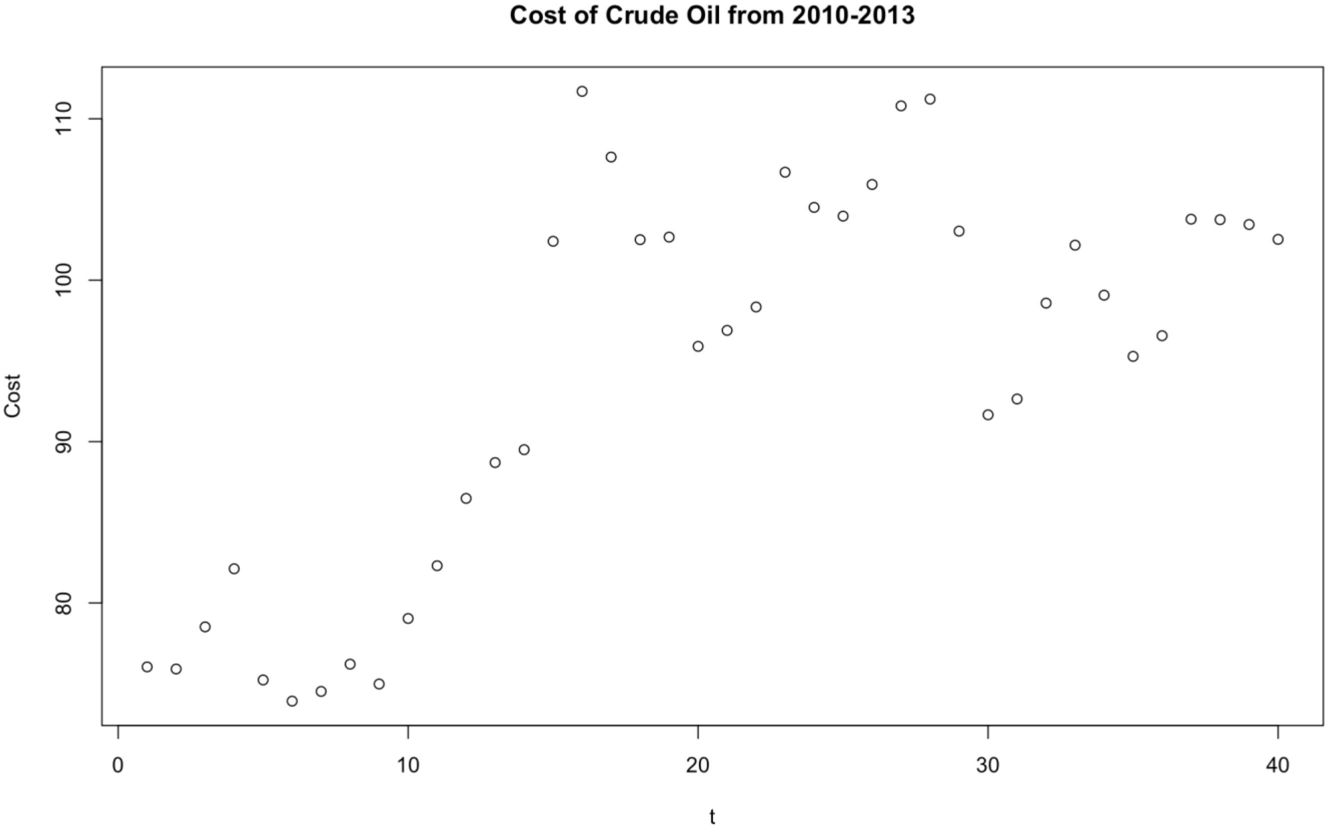
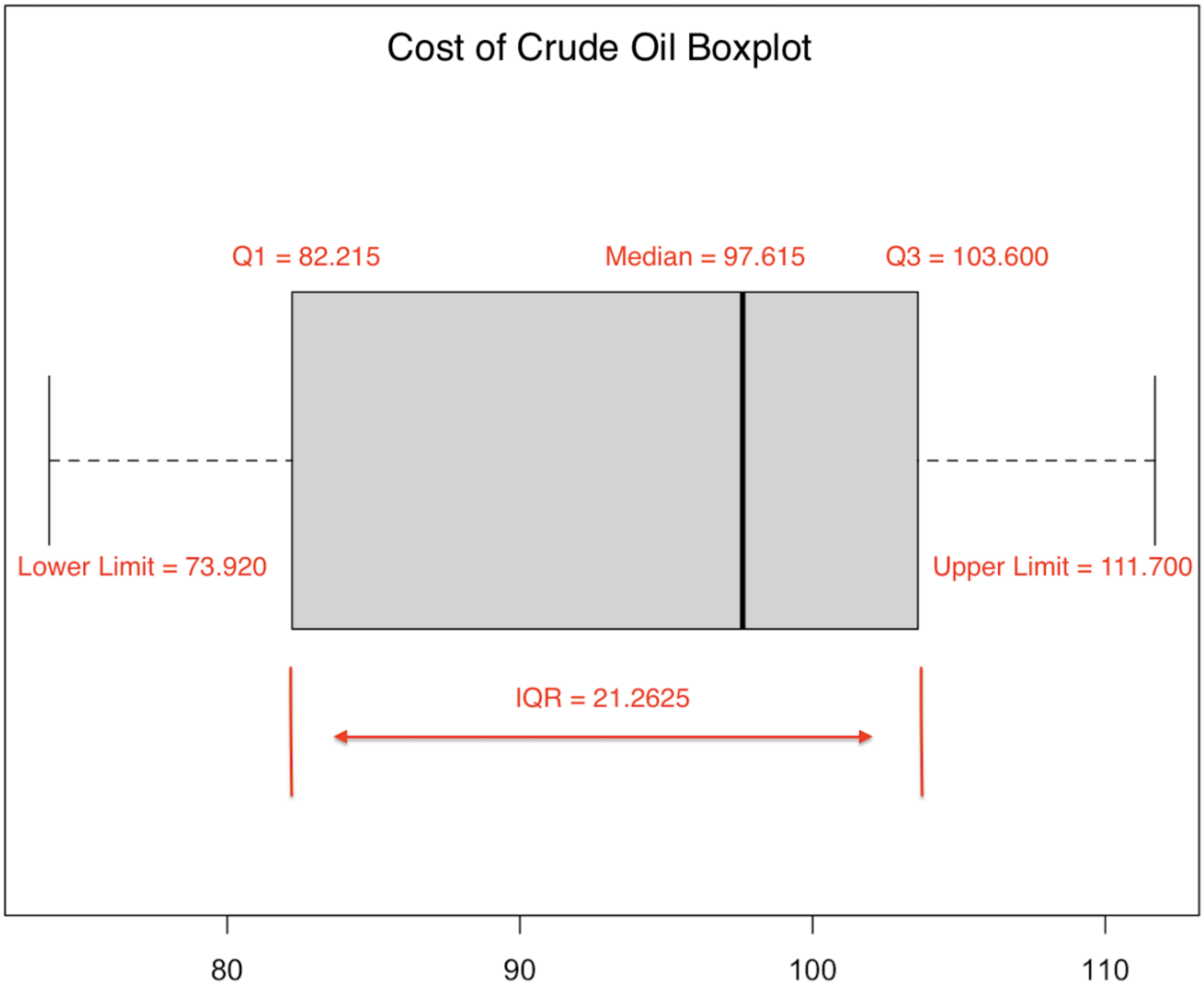
What is the goal in your project?
The goal of our project is to examine and understand the price of crude oil over time. Identify any patterns that exist, see if there is a relationship between future and past crude oil prices, and predict with a level of confidence what the price of crude oil will be for Jan 2014 using different statistical methods to see which method provides us with the most accurate prediction.

1. **Examine descriptive statistics to summarize and understand our data.**
    a. Scatterplot– are there any trends? Is it normally distributed? Shape of graph? Outliers?
    b. Do we have any bias/sampling errors?
    c. Find the mean median mode what does this tell us.



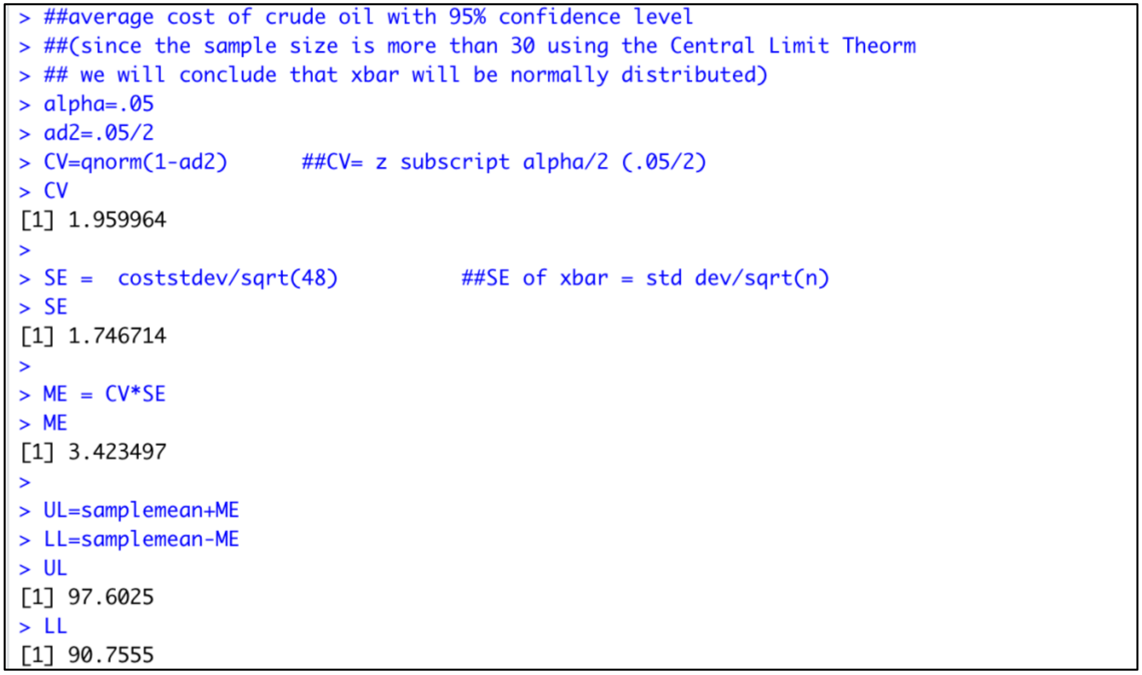
Mean: 94.179
Median: 97.615
Mode: No reoccurring prices
Range: 37.78
IQR = 21.2625
Variance = 146.4485
Standard Deviation = 12.10159

Coefficient of Variation = 12.84956



2. Interval estimation with 95% and 99% confidence level to give us parameters of our goal.

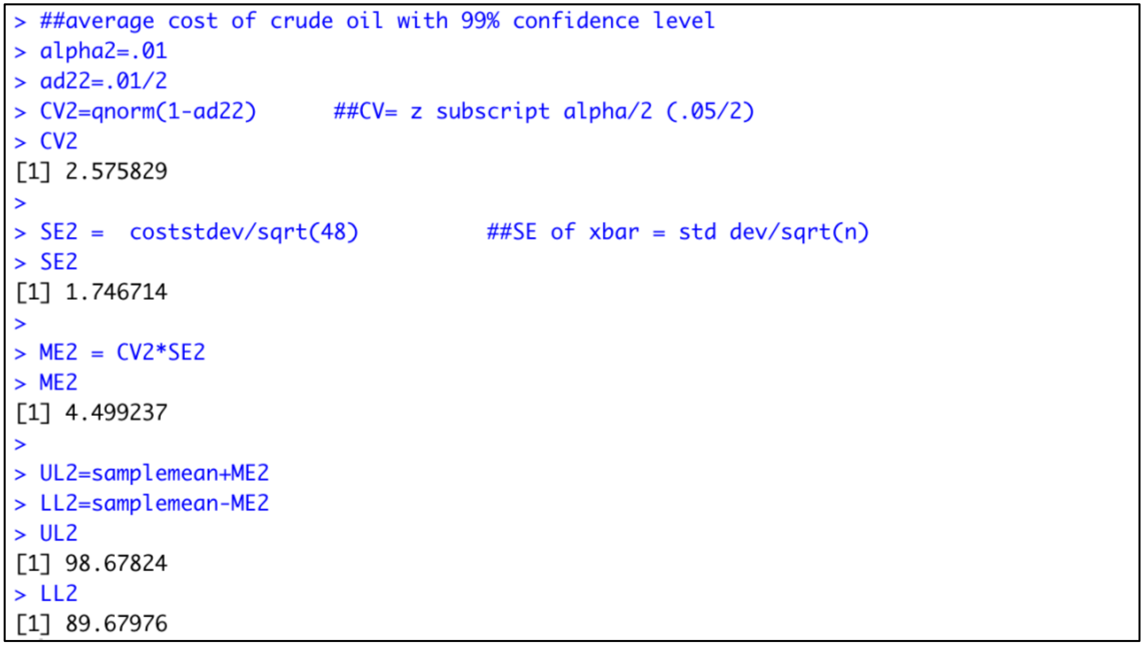
a. average cost of crude oil with 95% confidence level



```
> ##average cost of crude oil with 95% confidence level
> ##(since the sample size is more than 30 using the Central Limit Theorm
> ## we will conclude that xbar will be normally distributed)
> alpha=.05
> ad2=.05/2
> CV=qnorm(1-ad2)      ##CV= z subscript alpha/2 (.05/2)
> CV
[1] 1.959964
>
> SE =  coststdev/sqrt(48)          ##SE of xbar = std dev/sqrt(n)
> SE
[1] 1.746714
>
> ME = CV*SE
> ME
[1] 3.423497
>
> UL=samplemean+ME
> LL=samplemean-ME
> UL
[1] 97.6025
> LL
[1] 90.7555
```

We are 95% confident that the average cost of crude oil is between $90.76 and $97.60

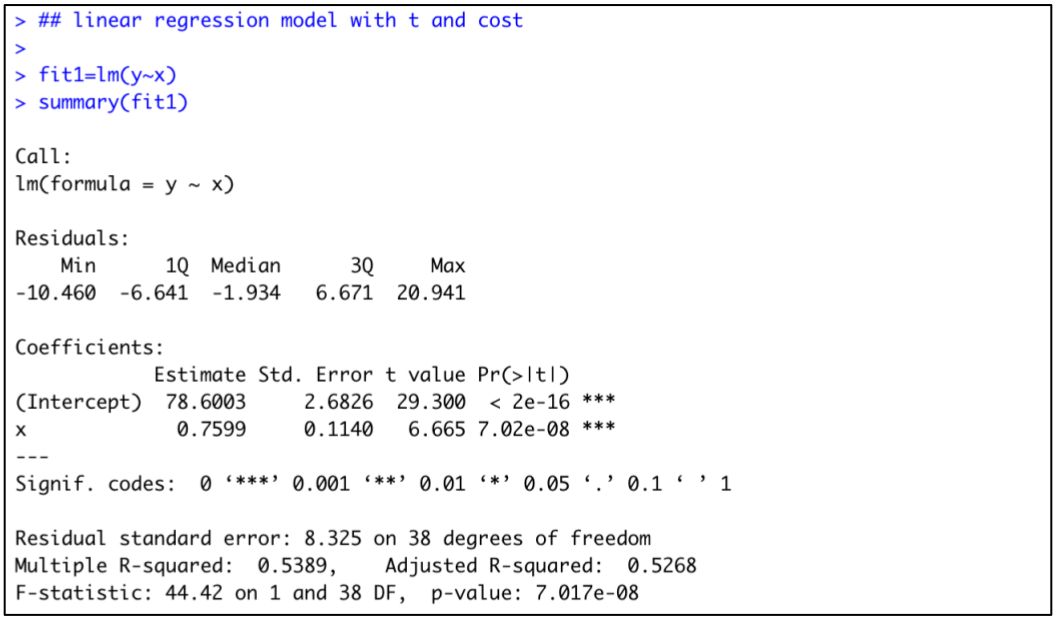
b. average cost of crude oil with 99% confidence level



```
> ##average cost of crude oil with 99% confidence level
> alpha2=.01
> ad22=.01/2
> CV2=qnorm(1-ad22)      ##CV= z subscript alpha/2 (.05/2)
> CV2
[1] 2.575829
>
> SE2 =  coststdev/sqrt(48)          ##SE of xbar = std dev/sqrt(n)
> SE2
[1] 1.746714
>
> ME2 = CV2*SE2
> ME2
[1] 4.499237
>
> UL2=samplemean+ME2
> LL2=samplemean-ME2
> UL2
[1] 98.67824
> LL2
[1] 89.67976
```

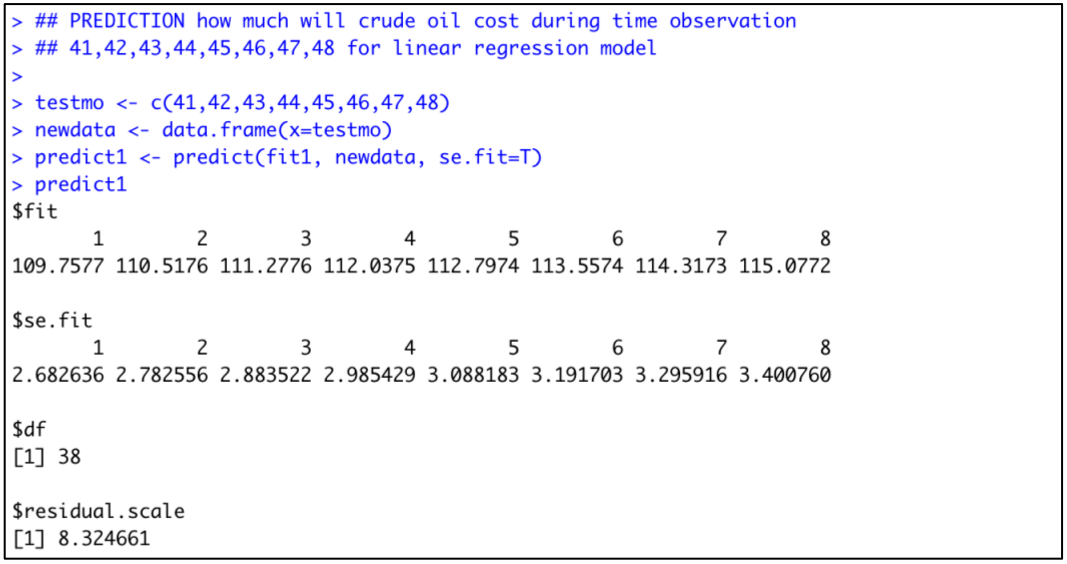
We are 99% confident that the average cost of crude oil is between $89.68 and $98.68

3. Time series plot – pattern? (Horizontal, trend, seasonal?) – with this what form of analysis should we use to examine data
   There is an increasing trend till t=16 and then it shows a trend of increasing and decreasing of the cost.
4. Compute linear trend equation to forecast the crude cost for Jan 2014.
   a.



```
> ## linear regression model with t and cost
>
> fit1=lm(y~x)
> summary(fit1)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-10.460  -6.641  -1.934   6.671  20.941 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  78.6003     2.6826  29.300  < 2e-16 ***
x             0.7599     0.1140   6.665 7.02e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.325 on 38 degrees of freedom
Multiple R-squared:  0.5389,	Adjusted R-squared:  0.5268 
F-statistic: 44.42 on 1 and 38 DF,  p-value: 7.017e-08
```

   b. What is the linear trend equation? Is the model a good fit? What is the goodness of fit for this model?
   Cost ($ per Barrel) = 78.600 + 0.7599 t
   c. What is the cost of crude oil for May 2013 using the linear trend equation?

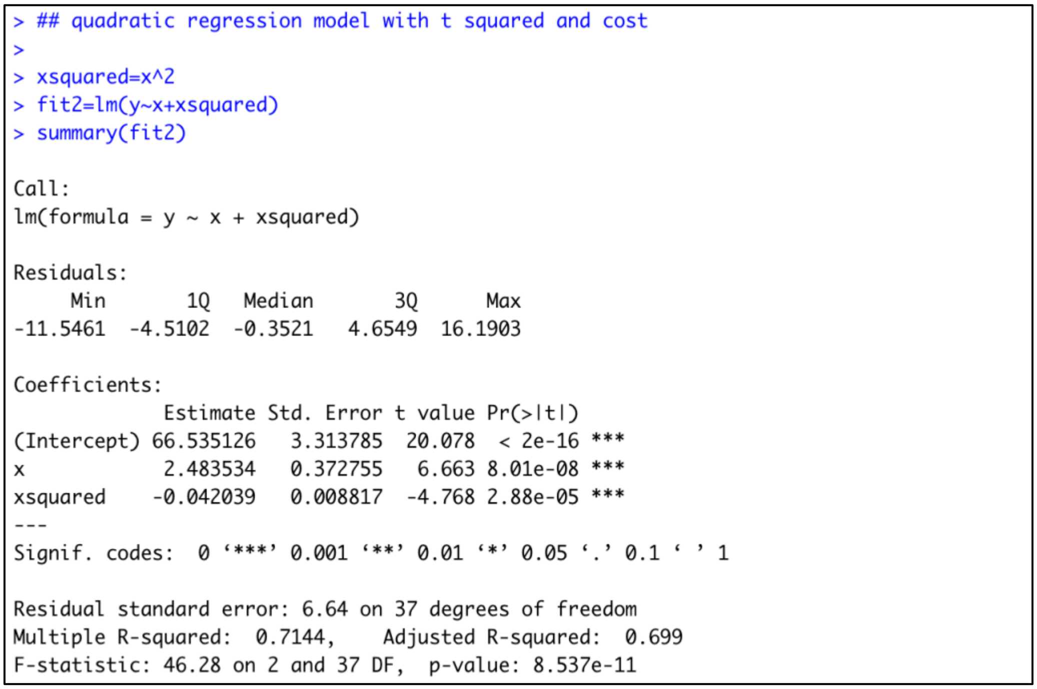


```
> ## PREDICTION how much will crude oil cost during time observation
> ## 41,42,43,44,45,46,47,48 for linear regression model
>
> testmo <- c(41,42,43,44,45,46,47,48)
> newdata <- data.frame(x=testmo)
> predict1 <- predict(fit1, newdata, se.fit=T)
> predict1
$fit
       1        2        3        4        5        6        7        8 
109.7577 110.5176 111.2776 112.0375 112.7974 113.5574 114.3173 115.0772 

$se.fit
       1        2        3        4        5        6        7        8 
2.682636 2.782556 2.883522 2.985429 3.088183 3.191703 3.295916 3.400760 

$df
[1] 38

$residual.scale
[1] 8.324661
```

   Cost for May 2013 = 78.600 + 0.7599 (41) = $109.76
   d. How does this compare to the average cost found in number 2?
   This result is outside both our 95% and 99% confidence intervals by nearly a full standard deviation.

5. Compute quadratic trend equation for time series to forecast the crude cost for Jan 2014 using dummy variables.

   a.

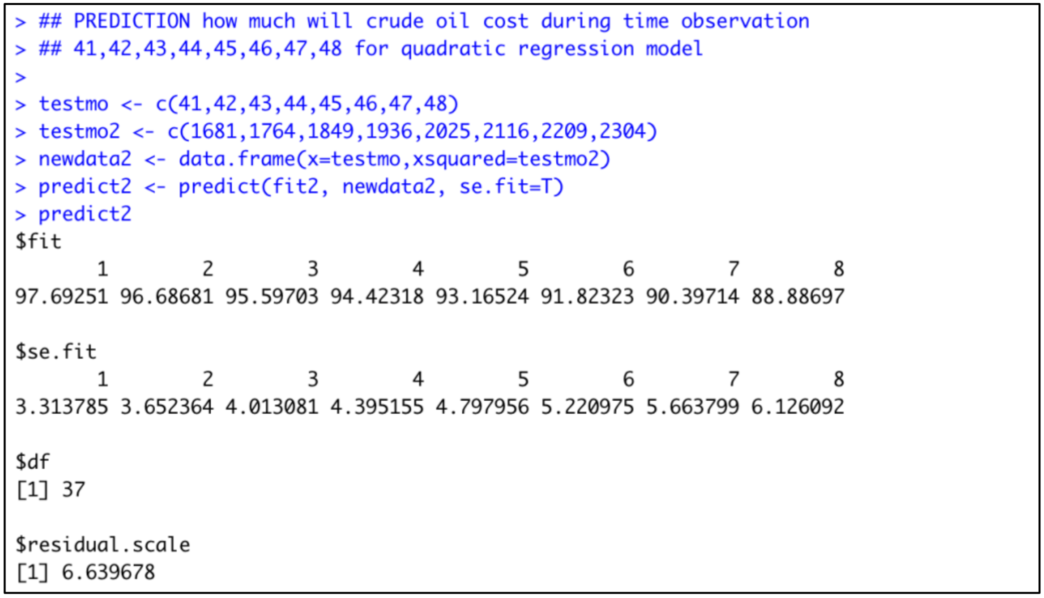


```
> ## quadratic regression model with t squared and cost
>
> xsquared=x^2
> fit2=lm(y~x+xsquared)
> summary(fit2)

Call:
lm(formula = y ~ x + xsquared)

Residuals:
     Min       1Q   Median       3Q      Max 
-11.5461  -4.5102  -0.3521   4.6549  16.1903 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 66.535126   3.313785  20.078  < 2e-16 ***
x            2.483534   0.372755   6.663 8.01e-08 ***
xsquared    -0.042039   0.008817  -4.768 2.88e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.64 on 37 degrees of freedom
Multiple R-squared:  0.7144,	Adjusted R-squared:  0.699 
F-statistic: 46.28 on 2 and 37 DF,  p-value: 8.537e-11
```

   b. What is the quadratic trend equation? Is the model a good fit? What is the goodness of fit for this model?

      Cost ($ per Barrel) = 66.54 + 2.484 (t) – 0.04204 $(t)^2$

   c. What is the cost of crude oil for May 2013 using the quadratic trend equation?



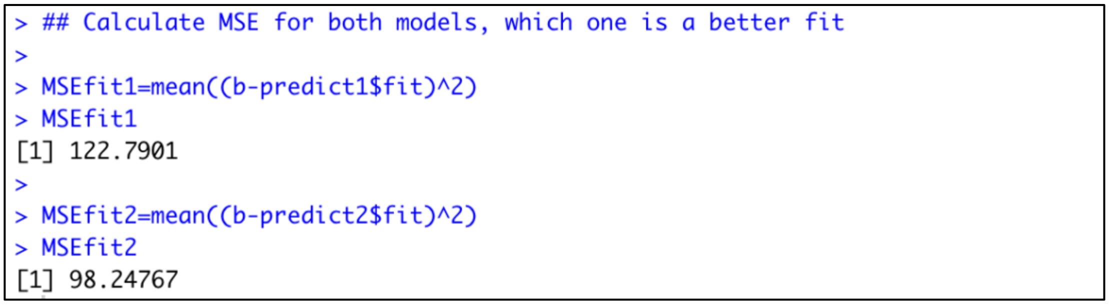
```
> ## PREDICTION how much will crude oil cost during time observation
> ## 41,42,43,44,45,46,47,48 for quadratic regression model
>
> testmo <- c(41,42,43,44,45,46,47,48)
> testmo2 <- c(1681,1764,1849,1936,2025,2116,2209,2304)
> newdata2 <- data.frame(x=testmo,xsquared=testmo2)
> predict2 <- predict(fit2, newdata2, se.fit=T)
> predict2
$fit
       1        2        3        4        5        6        7        8 
97.69251 96.68681 95.59703 94.42318 93.16524 91.82323 90.39714 88.88697 

$se.fit
       1        2        3        4        5        6        7        8 
3.313785 3.652364 4.013081 4.395155 4.797956 5.220975 5.663799 6.126092 

$df
[1] 37

$residual.scale
[1] 6.639678
```

      Cost for May 2013 = 66.54 + 2.484 (41) – 0.04204 $(41)^2$ = $97.69

   d. How does this compare to the linear trend equation? What observations can we make?

This value falls within our 95 and 99% estimation intervals as opposed to our linear model result, which is a value nearly a full standard deviation above that range.

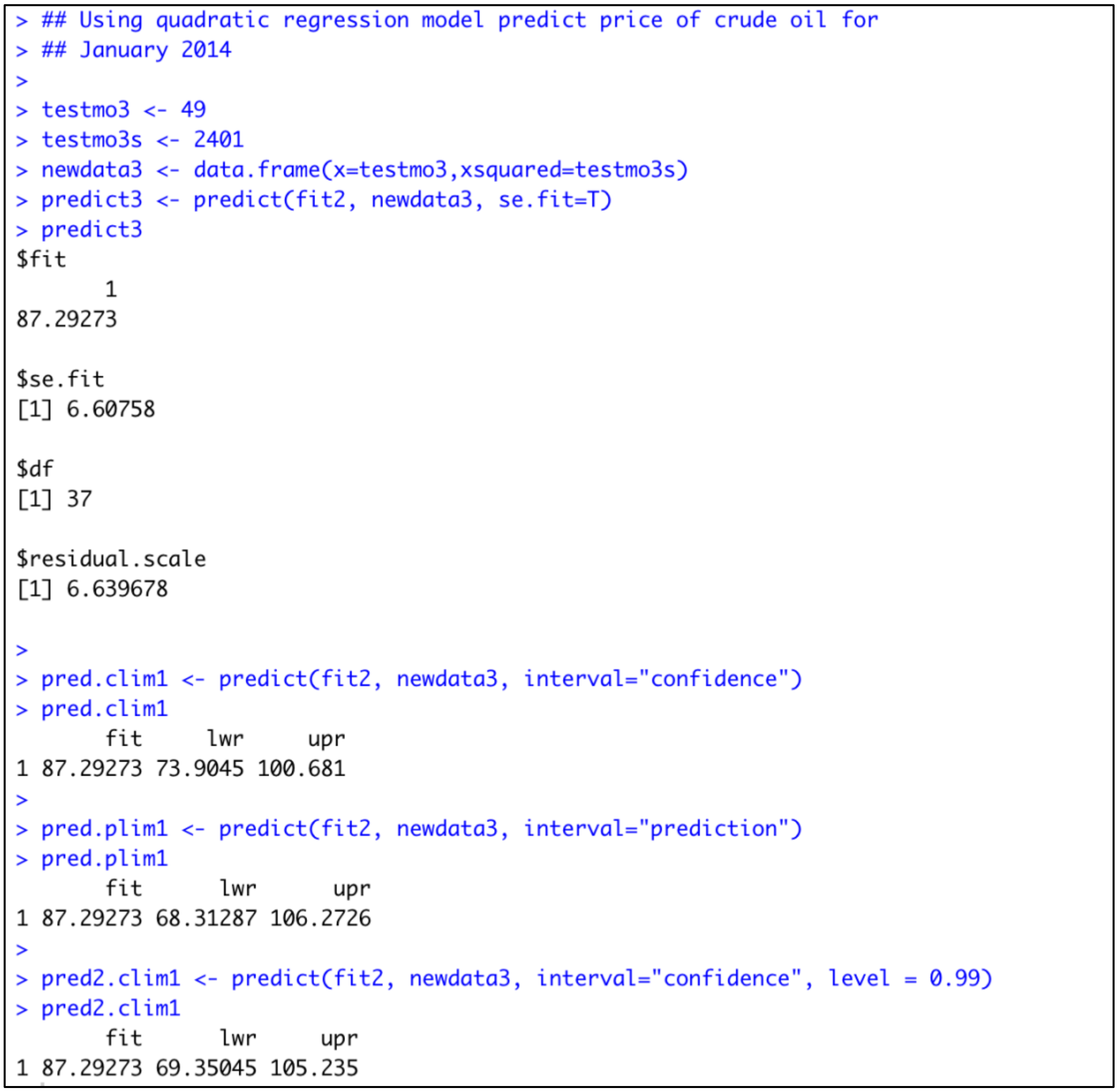
6. Use MSE to see which approach provides the most accurate forecast.
    a.



```
> ## Calculate MSE for both models, which one is a better fit
>
> MSEfit1=mean((b-predict1$fit)^2)
> MSEfit1
[1] 122.7901
>
> MSEfit2=mean((b-predict2$fit)^2)
> MSEfit2
[1] 98.24767
```

    b. Which model is a better fit for the given data?
    The quadratic model has a lower mean squared error and therefore fits the model better than the linear model.

7. Using the better of the two models predict the cost of crude oil for January 2014



```
> ## Using quadratic regression model predict price of crude oil for
> ## January 2014
>
> testmo3 <- 49
> testmo3s <- 2401
> newdata3 <- data.frame(x=testmo3,xsquared=testmo3s)
> predict3 <- predict(fit2, newdata3, se.fit=T)
> predict3
$fit
       1
87.29273

$se.fit
[1] 6.60758

$df
[1] 37

$residual.scale
[1] 6.639678

>
> pred.clim1 <- predict(fit2, newdata3, interval="confidence")
> pred.clim1
       fit     lwr     upr
1 87.29273 73.9045 100.681
>
> pred.plim1 <- predict(fit2, newdata3, interval="prediction")
> pred.plim1
       fit      lwr      upr
1 87.29273 68.31287 106.2726
>
> pred2.clim1 <- predict(fit2, newdata3, interval="confidence", level = 0.99)
> pred2.clim1
       fit      lwr     upr
1 87.29273 69.35045 105.235
```