Assignment Two

 # Create RMSA function

```
calculate_RMSE<-function(actual, prediction)
{

  rmsa<-sqrt(mean((actual-prediction)^2))
  return(rmsa)
}
```

 # 1 Encode categorical variables to factors using the as.factor command

```
bikesharingbyhour$season <- as.factor(bikesharingbyhour$season)
bikesharingbyhour$yr <- as.factor(bikesharingbyhour$yr)
bikesharingbyhour$mnth <- as.factor(bikesharingbyhour$mnth)
bikesharingbyhour$hr <- as.factor(bikesharingbyhour$hr)
bikesharingbyhour$holiday <- as.factor(bikesharingbyhour$holiday)
bikesharingbyhour$weekday <- as.factor(bikesharingbyhour$weekday )
bikesharingbyhour$workingday <- as.factor(bikesharingbyhour$workingday)
bikesharingbyhour$weathersit <- as.factor(bikesharingbyhour$weathersit)
```

 # 2 Split the dataset into a training and test sets as we did in lecture. Use a 80%-20% split.

```
index <- sample(1:nrow(bikesharingbyhour), 0.80*nrow(bikesharingbyhour))
tr_df <- bikesharingbyhour[index,]
te_df <- bikesharingbyhour[-index,]
```

 # 3 Run a linear regression using the lm() command treating the cnt variable as the response variable.

```
# remove casual and registered variables from training dataset since both are direct sum of the response variable

tr_df <- subset(tr_df , select=-c(casual,registered))
reg_model <- lm(cnt~. -instant - dteday, data=tr_df)
```

#4 Now use the summary() function on the linear model you created above. Which predictor variables have the lowest p-values, hence, are the most statistically significant? How would you interpret the results?

```
summary(reg_model)
```

Output

```
Call:
lm(formula = cnt ~ . - instant - dteday, data = tr_df)

Residuals:
    Min      1Q  Median      3Q     Max 
-391.26  -61.22   -7.87   51.61  440.90 
```

```
Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -85.195      7.476 -11.395  < 2e-16 ***
season2       37.654      5.434   6.930 4.41e-12 ***
season3       31.887      6.454   4.940 7.89e-07 ***
season4       65.281      5.491  11.890  < 2e-16 ***
yr1           85.679      1.752  48.905  < 2e-16 ***
mnth2          3.305      4.401   0.751 0.452692
mnth3         12.737      4.951   2.573 0.010105 *
mnth4          3.839      7.316   0.525 0.599814
mnth5         17.686      7.822   2.261 0.023773 *
mnth6          4.600      8.055   0.571 0.567986
mnth7        -15.492      9.055  -1.711 0.087133 .
mnth8          5.156      8.815   0.585 0.558601
mnth9         30.349      7.842   3.870 0.000109 ***
mnth10        16.076      7.284   2.207 0.027326 *
mnth11        -9.220      7.011  -1.315 0.188534
mnth12        -6.803      5.578  -1.220 0.222673
hr1          -15.004      6.015  -2.494 0.012635 *
hr2          -23.711      6.028  -3.933 8.42e-05 ***
hr3          -36.343      6.073  -5.984 2.23e-09 ***
hr4          -40.214      6.101  -6.591 4.53e-11 ***
hr5          -19.597      6.020  -3.255 0.001136 **
hr6           36.179      5.980   6.050 1.49e-09 ***
hr7          170.927      6.022  28.381  < 2e-16 ***
hr8          302.746      6.048  50.058  < 2e-16 ***
hr9          164.835      6.012  27.416  < 2e-16 ***
hr10         109.995      6.001  18.328  < 2e-16 ***
hr11         136.338      6.105  22.333  < 2e-16 ***
hr12         176.589      6.143  28.745  < 2e-16 ***
hr13         166.974      6.191  26.969  < 2e-16 ***
hr14         150.621      6.198  24.303  < 2e-16 ***
hr15         163.867      6.241  26.258  < 2e-16 ***
hr16         224.825      6.262  35.900  < 2e-16 ***
hr17         379.984      6.161  61.675  < 2e-16 ***
hr18         345.237      6.128  56.339  < 2e-16 ***
hr19         241.614      6.123  39.459  < 2e-16 ***
hr20         159.538      6.097  26.167  < 2e-16 ***
hr21         109.938      6.053  18.164  < 2e-16 ***
hr22          71.590      6.046  11.841  < 2e-16 ***
hr23          33.770      6.015   5.614 2.01e-08 ***
holiday1     -28.185      5.432  -5.188 2.15e-07 ***
weekday1       9.741      3.335   2.921 0.003498 **
weekday2       9.929      3.252   3.053 0.002267 **
weekday3      11.560      3.249   3.559 0.000374 ***
weekday4      10.971      3.240   3.386 0.000712 ***
weekday5      17.264      3.246   5.318 1.07e-07 ***
weekday6      15.559      3.219   4.834 1.35e-06 ***
workingday1       NA         NA      NA       NA
weathersit2  -10.978      2.164  -5.072 3.98e-07 ***
weathersit3  -68.103      3.646 -18.679  < 2e-16 ***
weathersit4  -63.493     59.078  -1.075 0.282517
temp         123.064     32.310   3.809 0.000140 ***
atemp        126.201     33.452   3.773 0.000162 ***
hum          -81.880      6.221 -13.162  < 2e-16 ***
windspeed    -27.631      7.873  -3.510 0.000450 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 102 on 13850 degrees of freedom
```

Multiple R-squared: 0.6853, Adjusted R-squared: 0.6841
F-statistic: 579.9 on 52 and 13850 DF, p-value: < 2.2e-16

Hour, Season and weathersit , hum variables has got lowest p-values which is less than 2e-16

 # 5 Using the calculate_RMSE function you created in the first part of this assignment, calculate the RMSE against the test set corresponding to the linear model. How accurate is your model?

```
# prediction on the test set using our model
# remove casual and registered variables from test dataset since both are direct sum of the response variable

te_df <- subset(te_df , select=-c(casual,registered))
preds <- predict(reg_model, newdata = te_df)
rmsa <- calculate_RMSE(te_df$cnt, preds)
print(rmsa)
```

RMSA= 100.9349, when I compare the actual and predicted value it look like that the model did not accurate, there is significant difference between the actual and predicted values which resulted larger RMSA value.