# Activity Sheet

## Learning outcomes:

After completing this exercise, you should be able to understand and perform below tasks.

1. Building Regression models
   Regression model using linear regression technique
2. Validating the model results and optimizing the model
3. Handling multicollinearity and dimensionality reduction
4. Evaluation of error metrics
5. Applying the models on un-seen data
   a. Splitting data into train and test data sets
   b. Comparing the error metrics
6. Interpretation of the results

**Linear Regression with one variable**

1. Read the BigMac data into R
2. Change the 2nd and 3rd column names of this data set
3. Build linear regression and interpret the results
4. Find the confidence and prediction limits value for the predicted value at 0.95 significance level
   #predict(model, data, interval="confidence",level=0.95)

5. Plot the fitted line with confidence limits

```
plot(data$Price,data$NetHourlyWage)
#points of confidence interval
points(data$Price,Pred$fit,type="l", col="red", lwd=2)
points(data$Price,Pred$lwr,pch="-", col="red", lwd=4)
points(data$Price,Pred$upr,pch="-", col="red", lwd=4)
#points of prediction interval
points(data$Price,Pred_pred$lwr,pch="o", col="green", lwd=4)
points(data$Price,Pred_pred$upr,pch="o", col="green", lwd=4)
```

6. Interpret the summary of the regression model

**Problem Statement:**

An online gaming portal wants to understand their customer patterns based on their transactional behavior. For this, they have constructed a customer level data based on the details they are tracking. The customer database consists of demographic and transactional information for each customer.

The objectives of today's activity are

- Building a regression model to predict the customer revenue based on other factors

**Steps:**

1. Read the data 'CustomerData.csv' into R.

2. Understand the structure of the data and pre-process
   a. Drop the attribute 'CustomerID'
   b. Convert 'City' as factor variable

3. Split the data into train and test data sets
   ```
   rows=seq(1,nrow(data),1)
   set.seed(123)
   trainRows=sample(rows,(70*nrow(data))/100)
   train = data[trainRows,]
   test = data[-trainRows,]
   ```

4. Build linear regression and interpret the results
   #Input attributes by selection City, NoOfChildren, Tenure, NoOfUnitsPurchased and predict the total revenue generated.
   #Input all attributes into model

5. Error metrics evaluation on train data and test data
   ```
   library(DMwR)
   #Error verification on train data-
   regr.eval(data$TotalRevenueGenerated, model$fitted.values)
   #Error verification on test data
   Pred<-predict(LinReg, test) #target variable should be excluded while giving test to predict function
   regr.eval(test$TotalRevenueGenerated, Pred)
   ```

6. Experiment with multiple combinations of independent attributes in the function of the model and check the results

7. Perform multicollinearity check

```
#Multicollinearity check
library(car)
vif(model)
# Stepwise Regression
library(MASS)
step <- stepAIC(model, direction="both")
```

Identify the best attributes and update the model and observe the results