Table of Contents

1.0 Case Study Description

The case study chosen is on lung capacity for people of age 3 to 19 years old. The dataset has 755 observations and 6 variables. The variables are LungCap, Age, Height, Smoke, Gender and Caesarean. The variables are better described in a table in the next unit.

We can said that lung capacity of a 15 years old child is greater than an 8 years old child which means lung capacity increases proportionally with age. But far is this true. Also we are not sure that the lung capacity of a person depends on the height of a person. That the higher is the height, greater is the lung capacity. Also does smoking and gender affect the lung capacity of a person?

Therefore supervised machine learning technique is applied on the dataset to determine to what extend these factors influence the lung capacity of a person. 2 models are formulated to perform Linear Regression to know which factor really influence the Lung Capacity. This is done in Unit 4.0.

Unsupervised machine learning technique is used to perform clustering on the dataset. Clustering is an important activity that enables grouping of data based on different clusters. Different models have been tested to know if K-Means clustering has worked correctly or not on the respectively model. This can be seen in Unit 5.0.

2.0 Fields of the dataset and its description

Field Name	Description	Data Type
LungCap	The lung capacity of the person (litres)	double
Age	How old is the person (years)	int
Height	The height of the person (inches)	double
Smoke	If the person smokes (yes) or doesn't smoke (no)	char
Gender	If the person is a male or female	char
Caesarean	If they're born by caesarean, 'yes' or 'no'	char

3.0 Setting R environment and load dataset

The package SparkR has been used as it has an in-built Machine Learning library that consists of a number of algorithms that run in memory.

- The package sparklyr alongside others required packages are installed and loaded in R.

```
library(sparklyr)
library(dplyr)
library(ggplot2)
```

- Connect to the local instance of Spark and remote Spark clusters. spark_connect function is used to connect to a local instance of Spark.

```
sc<-spark_connect(master="local")
```

- Upload dataset from source and edit dataset

```
#load dataset
lungCap<-read.csv("C:/Users/galaxy/Desktop/LungCapData.csv")
```

LungCap	Age	Height	Smoke	Gender	Caesarean
6.475	6	62.1	no	male	no
10.125	18	74.7	yes	female	no
9.550	16	69.7	no	female	yes
11.125	14	71.0	no	male	no
4.800	5	56.9	no	male	no
6.225	11	58.7	no	female	no
4.950	8	63.3	no	male	yes

As we can see, the values for 'Smoke' and 'Caesarean' are represented as 'yes' and 'no'. Thus, they are converted to '1' and '0' values respectively for better processing of data.

- Convert column 'Smoke' and 'Caesarean' values to 1 and 0.

```
lungCap$Smoke<- ifelse(lungCap$Smoke == "yes", 1, 0)
lungCap$Caesarean<- ifelse(lungCap$Caesarean == "yes", 1, 0)
```

LungCap	Age	Height	Smoke	Gender	Caesarean
6.475	6	62.1	0	male	0
10.125	18	74.7	1	female	0
9.550	16	69.7	0	female	1
11.125	14	71.0	0	male	0
4.800	5	56.9	0	male	0
6.225	11	58.7	0	female	0
4.950	8	63.3	0	male	1

- Load data to Spark

```
lungCap_tbl<-copy_to(sc,lungCap)
```

- View the data using head() function.

```
> head(lungCap_tbl)
# Source: spark<?> [?? x 6]
  LungCap   Age Height Smoke Gender Caesarean
    <dbl> <int>  <dbl> <dbl> <chr>      <dbl>
1    6.48     6   62.1     0 male           0
2   10.1     18   74.7     1 female         0
3    9.55    16   69.7     0 female         1
4   11.1     14   71       0 male           0
5    4.8      5   56.9     0 male           0
6    6.22    11   58.7     0 female         0
```

4.0 Applying supervised Machine Learning techniques

4.1 Linear Regression

Linear regression is one of the most commonly used predictive modelling techniques. It is used to determine the extent to which there is a linear relationship between a dependent variable and one or more independent variables. The dependent variable (response variable) must be measured on a continuous measurement scale (In our case, LungCap is used as dependent variable) and the independent variable (predictor variable) can be measured on either a categorical (e.g. male versus female) or continuous measurement scale.

Different linear models are formulated to know which factor really influence the Lung Capacity.

4.1.1 Model 1: If all the factors influence the lung capacity?

The following code is used to formulate model 1.

```
#linear Reggression
#model1
lm_model1<-lungCap_tbl%>%select(LungCap,Age,Height,Smoke,Gender,Caesarean)
          %>% ml_linear_regression(LungCap~Age+Height+Smoke+Gender+Caesarean)
```

After running the above codes, the following results are obtained:

```
> summary(lm_model1)
Deviance Residuals:
     Min       1Q   Median       3Q      Max
-3.33879 -0.72001  0.04444  0.70930  3.01724

Coefficients:
(Intercept)         Age      Height       Smoke Gender_male   Caesarean
-11.3224856   0.1605296   0.2641128  -0.6095592   0.3870117  -0.2142182

R-Squared: 0.8542
Root Mean Squared Error: 1.016
```

Interpretation of Results:-

These values displayed for each factor represent the percentage influence that these predictor variables have on the response variable. For example: Age has a value 0.1605296 which shows that as age increase by 1 year, the lung capacity increases by 16.1%. The value of Smoke is -0.6095592 which show if a person smoked, this decreases the lung capacity by 60.9% maintaining others factors into considerations (ex the person is a male and caesarean is 'yes'). If a child is born by caesarean, this decreases its lung capacity by 21.4%.

Predictions:-

Prediction is used to estimate a particular value based on the model formulated. In this case, we want to predict the LungCap using the influencing factors as derived in the linear model. To do the prediction, we have partitioned our dataset as 50% of the dataset for the training and 50% for the test.

The following code is applied for the partitioning:

```
partitions<-lungCap_tbl%>%sdf_partition(training=0.5,test=0.5)
```

To formulate the prediction based on linear model 1, the following codes are used:

```
preds<-sdf_predict(partitions$test,lm_model1)
```

Result:

A column prediction is appended to the original table. This column shows the predicted value of the lung capacity generated from the model applied. It can be seen that for nearly all the values, the predicted LungCap is close to the real LungCap, which indicates that linear regression is an appropriate model that can be used for this type of data.

LungCap	Age	Height	Smoke	Gender	Caesarean	prediction
1.450	3	45.3	0	female	0	1.123413
1.025	3	47.0	0	female	0	1.572405
3.025	6	47.4	0	female	0	2.159639
1.925	5	48.0	0	male	0	2.544589
1.900	8	48.1	0	male	1	2.838370
2.875	7	48.2	0	male	0	2.918470
1.125	4	48.7	0	female	0	2.181926
1.950	7	48.8	0	male	0	3.076938
1.325	5	48.9	0	female	0	2.395278

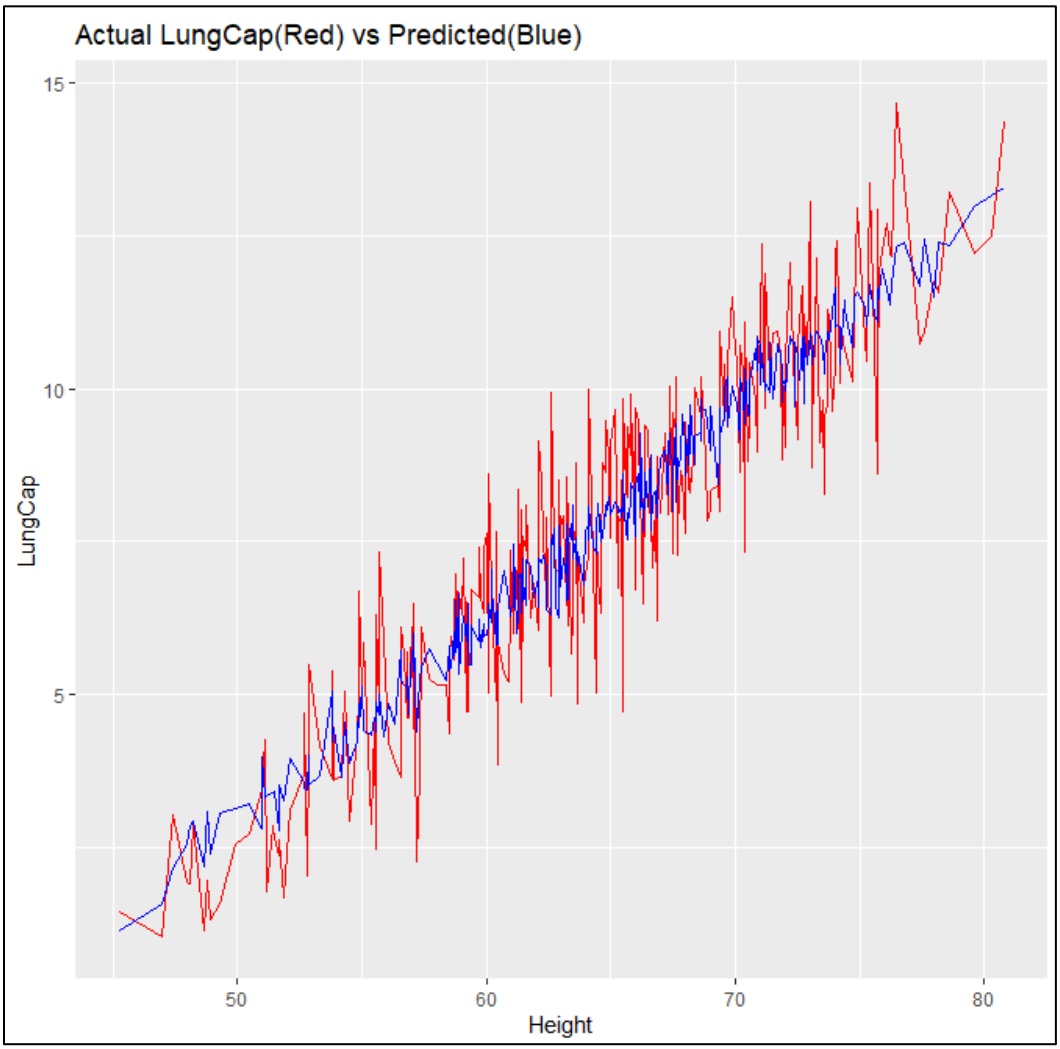
Graphical Representation:-

A graph is used to visualise the real LungCap values compared to the predicted values. We had seen that Height (26.4%) has a higher influence to LungCap compared to Age (16.1%). Thus Height values are used as the x-axis.

The following codes and library ggplot2 are used to plot the graph.

```
#plot Actual vs prediction
preds<-preds[order(preds$Height),]
v<-preds$LungCap
v1<-preds$Height
v2<-data.frame(v,v1)
w<-preds$prediction
w1<-data.frame(w,v1)

g=ggplot()+
  geom_line(data=v2,aes(x=v1,y=v),color="red")+
  geom_line(data=w1,aes(x=v1,y=w),color="blue")+xlab('Height')+ylab('LungCap')
    +ggtitle("Actual LungCap(Red) vs Predicted(Blue)")
g
```



4.1.2 Model 2: If Age and Smoke influence the lung capacity?

The following code is used to formulate model 2.

```
#model2
lm_model2<-lungCap_tbl%>%select(LungCap,Age,Smoke)%>% ml_linear_regression(LungCap~Age+Smoke)
```

After running the above codes, the following results are obtained:

```
> summary(lm_model2)
Deviance Residuals:
     Min       1Q   Median       3Q      Max
-4.85587 -1.02889 -0.03627  1.00834  4.19952

Coefficients:
(Intercept)         Age       Smoke
  1.0857247   0.5553959  -0.6485883

R-Squared: 0.6773
Root Mean Squared Error: 1.511
```

Interpretation of Results:-

Age has a value 0.5553959 which indicates that as age increased by 1 year, the LungCap increases by 55.5%. And if the person smoked, its lung capacity decreases by 64.9% and the Intercept represent the mean lung capacity.

Using the regression equation formula, we can calculate the regression for Smoker and non-Smoker and the result is as follows:-

Non-Smoker:

=1.08+0.555*Age-0.649*Smoke

=1.08+0.555*Age-0.649*0

=1.08+0.555*Age

Smoker:

=1.08+0.555*Age-0.649*Smoke

=1.08+0.555*Age-0.649*1

= (1.08-0.649) +0.555*Age

=0.431+0.555*Age

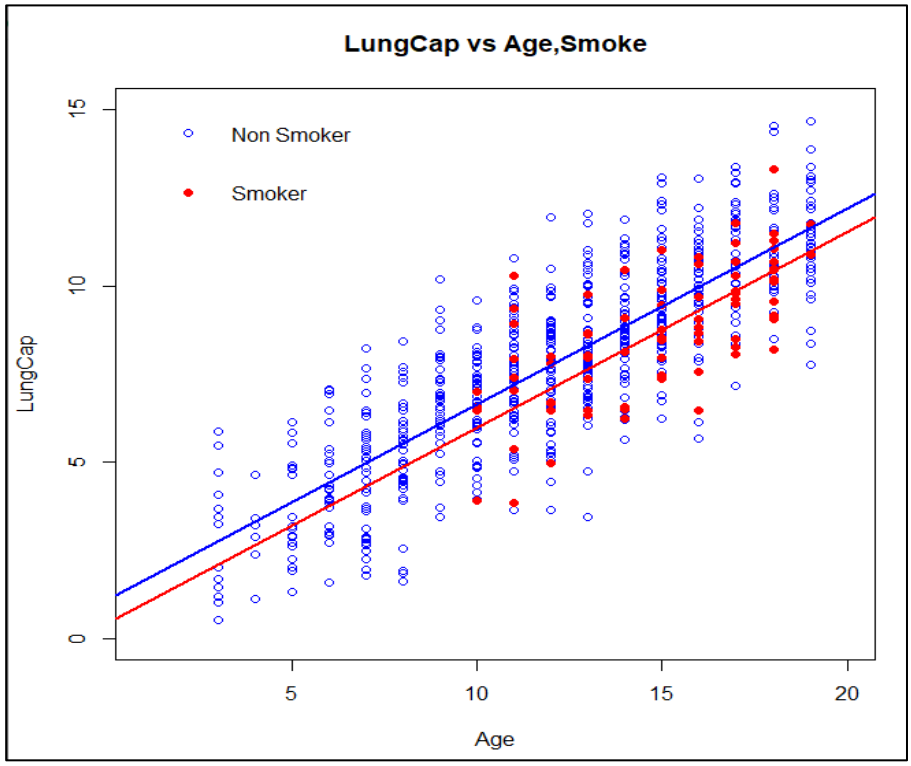
We conclude that if someone smokes, the mean lung capacity is 0.431 unlike if he does not smoke, the mean lung capacity is 1.08.

Graphical Representation:-

The following code is used to plot the graph representing how Age, Smoker, Non-Smoker influence the LungCap.

```
#plot graph LungCap vs Age,Smoke
plot(lungCap$Age[lungCap$Smoke=="0"],lungCap$LungCap[lungCap$Smoke=="0"],col="blue",xlim =c(1,20)
     ,ylim =c(0,15),xlab ="Age",ylab = "LungCap",main="LungCap vs Age,Smoke")
points(lungCap$Age[lungCap$Smoke=="1"],lungCap$LungCap[lungCap$Smoke=="1"],col="red",pch=16)
legend(1,16,legend = c("Non Smoker","Smoker"), col=c("blue","red"),pch = c(1,16),bty = "n")
abline(a=1.08,b=0.555,col="blue",lwd=2)
abline(a=0.431,b=0.555,col="red",lwd=2)
```

Result:



Conclude:- Age has an effect on LungCap, as Age increases LungCap increases.

- Smoke has an effect on LungCap, for a smoker the mean lung capacity decreases by 0.649 and this effect is assumed to be for all ages.
- Effect of Age is independent of Smoke that is the effect of smoking is not modified by age.

Predictions:-

In this case, we want to predict the LungCap using the influencing factors (Age and Smoke) as derived in the linear model 2. 50% of the dataset is used for the training and 50% for the test.

To formulate the prediction based on linear model 2, the following codes are used:

```
preds1<-sdf_predict(partitions$test,lm_model2)
```

Result:

We noticed that the predicted value is the same for a group age if the person does not smoke.

LungCap	Age	Height	Smoke	Gender	Caesarean	prediction
1.025	3	47.0	0	female	0	2.751912
1.125	4	48.7	0	female	0	3.307308
1.325	5	48.9	0	female	0	3.862704
1.450	3	45.3	0	female	0	2.751912
1.575	6	49.3	0	male	0	4.418100
1.675	3	51.9	0	male	0	2.751912
1.775	7	51.2	0	female	0	4.973496
1.900	8	48.1	0	male	1	5.528892
1.925	5	48.0	0	male	0	3.862704
1.950	7	48.8	0	male	0	4.973496
2.025	5	52.8	0	female	0	3.862704

But the predicted value changes if the person smokes.

6.100	10	57.0	0	male	0	6.639683
6.100	10	57.4	0	female	0	6.639683
6.575	10	63.2	1	male	1	5.991095
6.725	10	60.2	0	female	0	6.639683

Therefore as stated before, smoking has an effect on lung capacity.

5.0 Applying Unsupervised Machine learning techniques

5.1 K-Means Algorithm

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. It is an iterative algorithm that partition the dataset into K pre-defined distinct clusters where each data point belongs to only one group. It tries to make the inter-cluster data points as similar as possible while also keeping the clusters as far as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more similar the data points are within the same cluster (Imad Dabbura, 2018).

5.1.1 Model 1

The same dataset is being used and 'Gender' is used as the response variable (class). The response variable denotes "Male" and "Female". The predictors "LungCap" and "Age" are used to perform the k-Means for model 1.

Steps:-

- Load the data to Spark

```
#K means algorithm
lungCap_tbl2<-copy_to(sc,lungCap)
```

- The following codes are used to formulate the k-means model 1.

```
kmeans_model<-lungCap_tbl2%>% ml_kmeans(formula= ~ LungCap+Age,k=2)
```

- View k-means result

The k-means is calculated and the cluster centers for each predictor variable are displayed. We can see that the centre of the 2 clusters have different values which mean they are not over-lapping and the clusters are distinct.

```
> kmeans_model
K-means clustering with 2 clusters

Cluster centers:
   LungCap       Age
1 9.559014 15.144231
2 5.580039  8.533981

Within Set Sum of Squared Errors =  6187.622
```

- Prediction is made on the model 1 using the following codes.

```
predicted<-ml_predict(kmeans_model,lungCap_tbl2)%>% collect
           table(predicted$Gender,predicted$prediction)
```

- View predicted result

```
          0   1
 female 209 149
 male   207 160
```

Interpretation of results

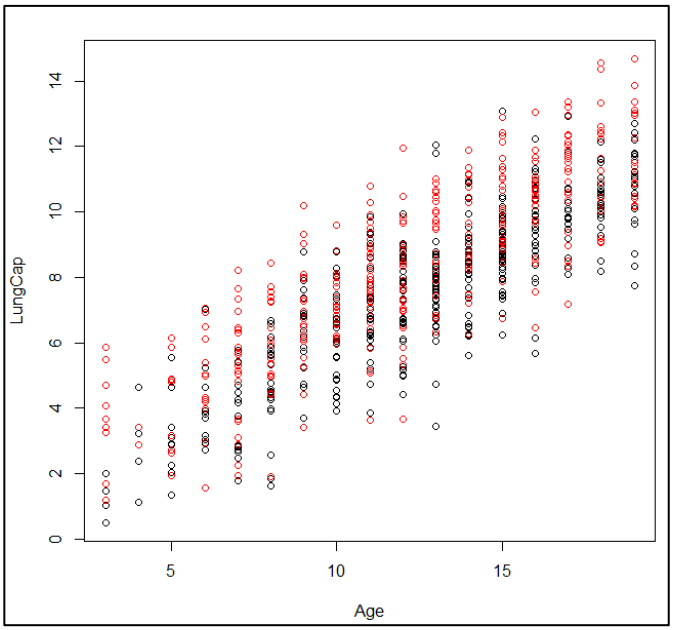
Out of 358 female, we can see that 209 cases are classified under 0 and 149 cases under 1. And out of 367 men, 207 cases are classified under 0 and 160 under 1. For both male and female, they are classified under first cluster which means k-means have not performed correctly for model 1.

Graphical Representation:- We are going to plot 2 graphs to compare the result of the actual dataset and after performing the k-means clustering.

The following codes are used to represent the actual data, Age in x-axis, LungCap in y-axis group by Gender.

```
plot(lungCap[c("Age","LungCap")],col=lungCap$Gender)
```

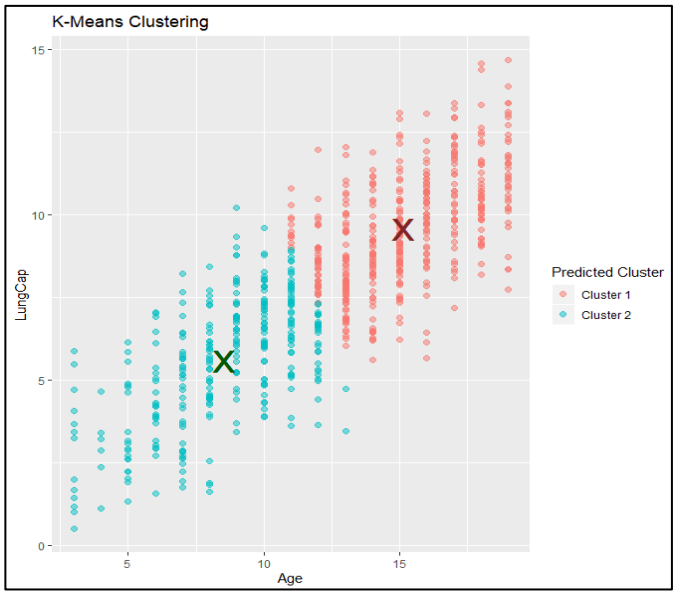
Result Real Data:



The following codes are used to represent the k-means results in a graph.

```
# plot cluster graph
ml_predict(kmeans_model) %>%
collect() %>%
ggplot(aes(Age, LungCap)) +
geom_point(aes(Age, LungCap, col = factor(prediction + 1)),size = 2, alpha = 0.5) +
              geom_point(data = kmeans_model$centers, aes(Age, LungCap),
                         col = scales::muted(c("red", "green")),pch = 'x', size = 12) +
              scale_color_discrete(name = "Predicted Cluster",labels = paste("Cluster", 1:2)) +
              labs(x = "Age",y = "LungCap",title = "K-Means Clustering")
```

Result K-Means data: We plot Age against LungCap and the two clusters are indicated by X. We can see that this k-means graph is different from the actual dataset graph.



5.1.2 Model 2

We have seen in the previous model using 'Gender' as the response variable has not working correctly for the K-Means algorithm. For model 2 we are using Height as the variable.

- First we convert the Height values into 4 different categories.

Category	Height value
A	<50
B	51-60
C	61-70
D	>70

- View the new dataset

```
> lungData
    LungCap Age Height Smoke Gender Caesarean
1     6.475   6      C     0   male         0
2    10.125  18      D     1 female         0
3     9.550  16      C     0 female         1
4    11.125  14      D     0   male         0
5     4.800   5      B     0   male         0
6     6.225  11      B     0 female         0
7     4.950   8      C     0   male         1
8     7.325  11      D     0   male         0
```

- Check if data is correctly categorised.

```
> lungCap$Height[1:10]
 [1] 62.1 74.7 69.7 71.0 56.9 58.7 63.3 70.4 70.5 59.2
> CatHeight[1:10]
 [1] C D C D B B C D D B
Levels: A B C D
```

We can see that the data is correctly categorised.

Perform K-Means Clustering

We used Height as the response variable which has 4 cases, 'A', 'B', 'C' and 'D'. The predictor LungCap, Age and Smoke are used for this model.

- The following codes are used to formulate the k-means model 2.

```
lungData2<-copy_to(sc,lungData)
kmeans_model3<-lungData2%>% ml_kmeans(formula= ~Age+Smoke+LungCap,k=4)
```

- View k-means result

  The k-means is calculated and the cluster centers for each predictor variable are displayed. We can see that the centre of the 4 clusters have different values which mean they are not over-lapping and the clusters are distinct.

```
K-means clustering with 4 clusters

Cluster centers:
        Age      Smoke   LungCap
1 13.609244 0.13025210  8.606092
2  6.090909 0.00000000  3.954603
3 17.297143 0.17714286 10.873000
4 10.125654 0.07853403  6.655759

Within Set Sum of Squared Errors =  2619.307
```

- Prediction is made on the model 2 using the following codes.

```
predicted<-ml_predict(kmeans_model3,lungData2)%>% collect
table(predicted$Height,predicted$prediction)
```

- View predicted result

```
     0   1   2   3
A    0  18   0   0
B    8  94   0  68
C  181   9  40 120
D   49   0 135   3
```

Interpretation of results

For category A, all the 18 cases are classified under 1 (2nd Cluster). 94 of category B cases have been classified under 1 and 76 cases have been misclassified. For category C, 181 cases have been classified under 0, 120 under 3, 40 under 2 and 9 under 1. And for category D, 135 cases have been correctly classified under 2 and 49 misclassified under 0 and 3 misclassified under 3. Overall we conclude that the performance is good and the error is not too high. Therefore the K-Means algorithm has classified correctly the data into the 4 clusters.

Graphical Representation:-

The following codes are used to represent the above results in a graph.

```
# plot cluster graph
ml_predict(kmeans_model3) %>%
  collect() %>%
  ggplot(aes(Age, LungCap)) +
  geom_point(aes(Age, LungCap, col = factor(prediction + 1)),size = 2, alpha = 0.5) +
  scale_color_discrete(name = "Predicted Cluster",labels = paste("Cluster", 1:4)) +
  labs(x = "Age",y = "LungCap",title = "K-Means Clustering Model 2")
```

Result: We can see that the clusters are correctly classified, there is no overlapping.