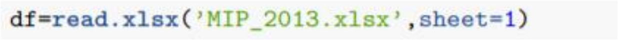
Community Innovation

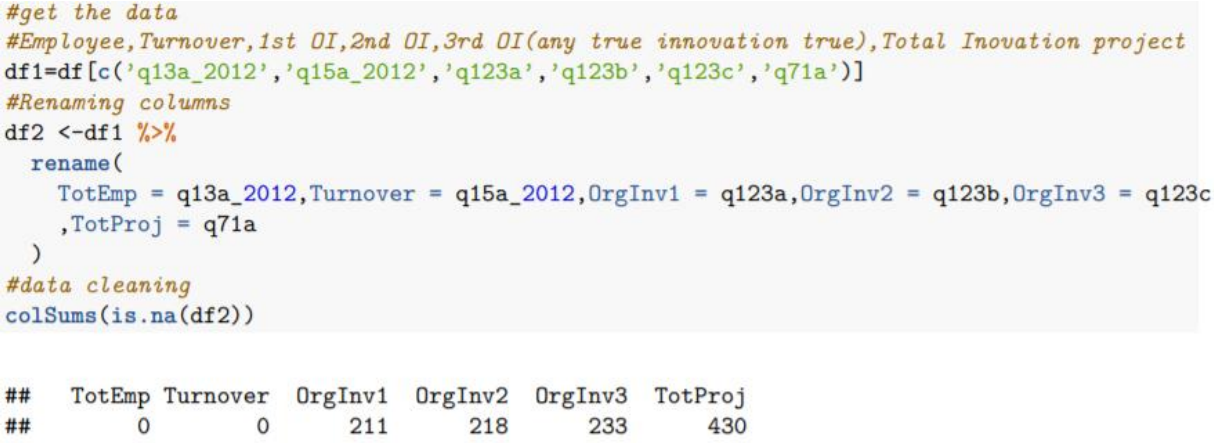
Loading and cleaning data

First, before loading the data we need to load all the packages which will be needed to perform further analysis.
After that data is loaded.



```
df=read.xlsx(’MIP_2013.xlsx’,sheet=1)
```

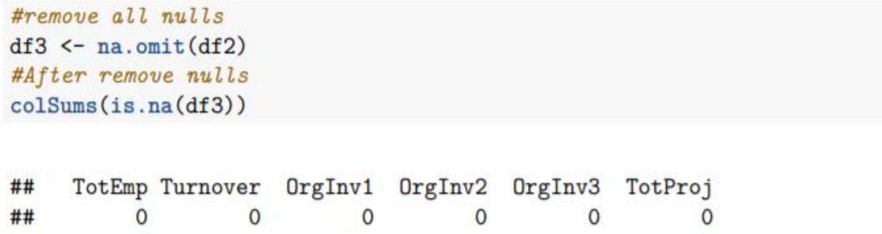
Now as we see there are many columns which are not needed to answer the question--Does organisational innovation lead to higher sales?
So, after going through the data and provided questionnaire, few columns are selected which are 'q13a_2012','q15a_2012','q123a','q123b','q123c','q71a', with each one represent Total Employee,Turnover,Organization innovation type 1,Organization innovation type 2,Organization innovation type 3(any true innovation true),Total Innovative projects.
Renaming is done accordingly.



```
#get the data
#Employee,Turnover,1st OI,2nd OI,3rd OI(any true innovation true),Total Inovation project
df1=df[c(’q13a_2012’,’q15a_2012’,’q123a’,’q123b’,’q123c’,’q71a’)]
#Renaming columns
df2 <-df1 %>%
  rename(
    TotEmp = q13a_2012,Turnover = q15a_2012,OrgInv1 = q123a,OrgInv2 = q123b,OrgInv3 = q123c
    ,TotProj = q71a
  )
#data cleaning
colSums(is.na(df2))
```

```
##    TotEmp Turnover  OrgInv1  OrgInv2  OrgInv3  TotProj
##         0        0      211      218      233      430
```

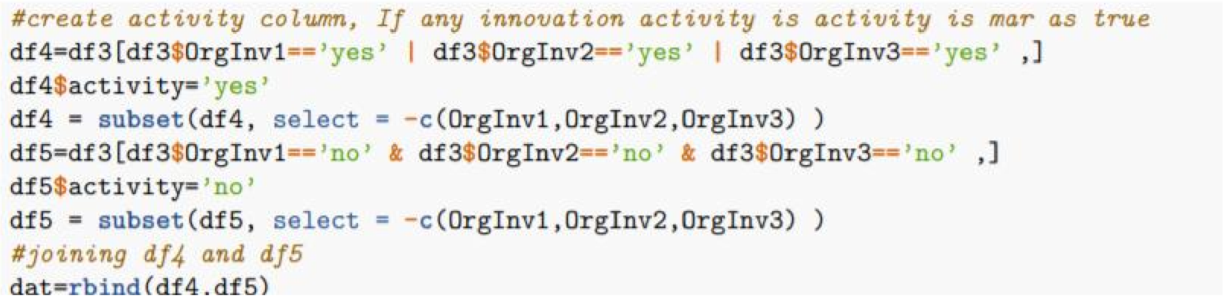
The first step toward data cleaning is getting rid of null values as the machine cannot process Nulls.
This can be done using **omit().**



```
#remove all nulls
df3 <- na.omit(df2)
#After remove nulls
colSums(is.na(df3))
```

```
##    TotEmp Turnover  OrgInv1  OrgInv2  OrgInv3  TotProj
##         0        0        0        0        0        0
```

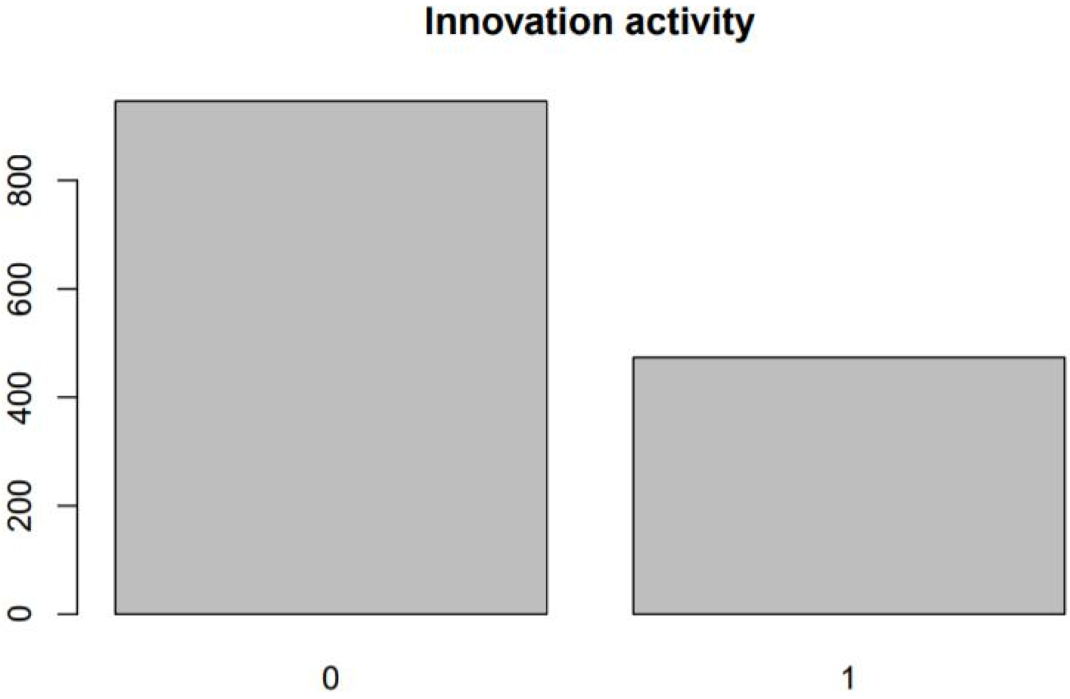
Furthermore, manipulation is done like creating an activity column and if any one Organization innovation type is ‘yes’, output should be ‘yes’ and if all the Types are ‘no’, output will be ‘no’ in the activity column.



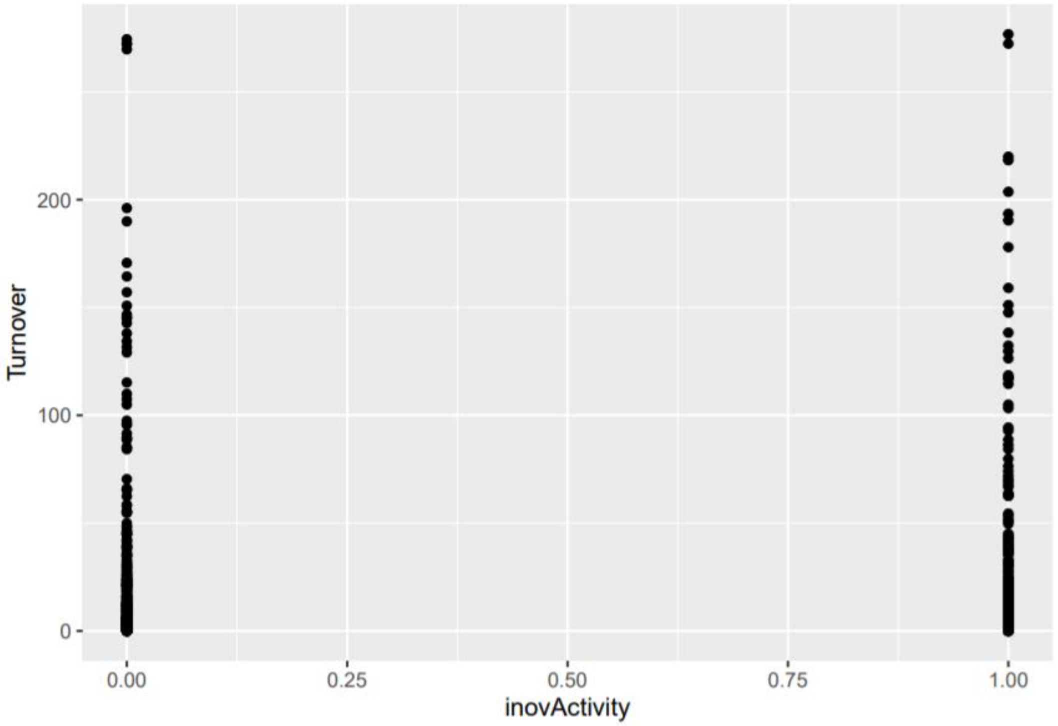
```
#create activity column, If any innovation activity is activity is mar as true
df4=df3[df3$OrgInv1==’yes’ | df3$OrgInv2==’yes’ | df3$OrgInv3==’yes’ ,]
df4$activity=’yes’
df4 = subset(df4, select = -c(OrgInv1,OrgInv2,OrgInv3) )
df5=df3[df3$OrgInv1==’no’ & df3$OrgInv2==’no’ & df3$OrgInv3==’no’ ,]
df5$activity=’no’
df5 = subset(df5, select = -c(OrgInv1,OrgInv2,OrgInv3) )
#joining df4 and df5
dat=rbind(df4,df5)
```

Total Employee cannot be a decimal number, so it will be converted to integer using round.

Data manipulation and visualization

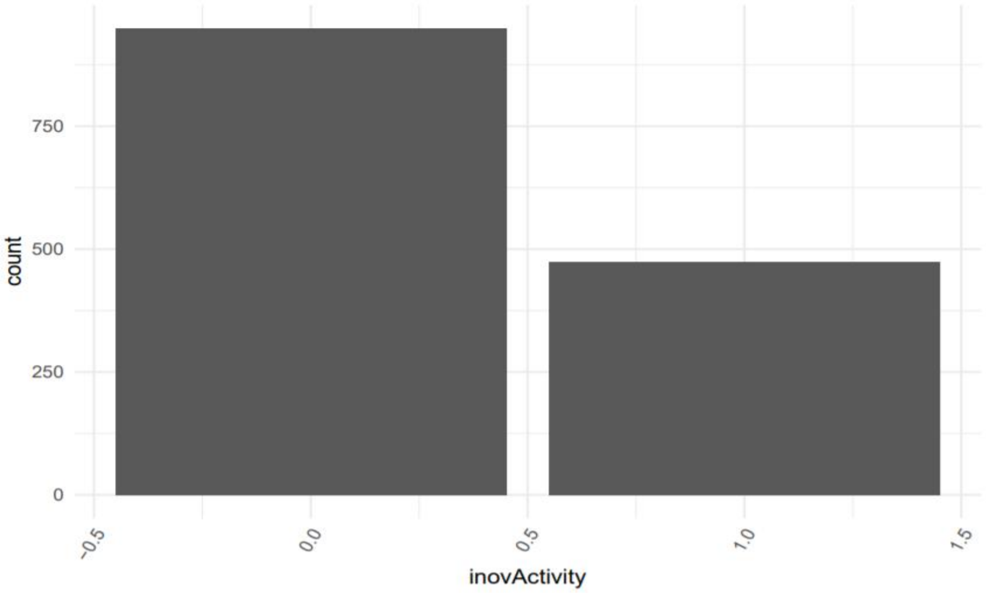


Here 0 represents no Organizational innovation activity and 1 as there is.
Total number of 0 cases → 949
Total number of 1 cases → 474
Next by making use of ggplot, plot between innovation activity and Turnover.



It can be seen that trends in both cases are not drastically different. In the case of innovation activity as 1, the mean is a little bit higher but overall it's the same if the difference in sample size is taken into account.

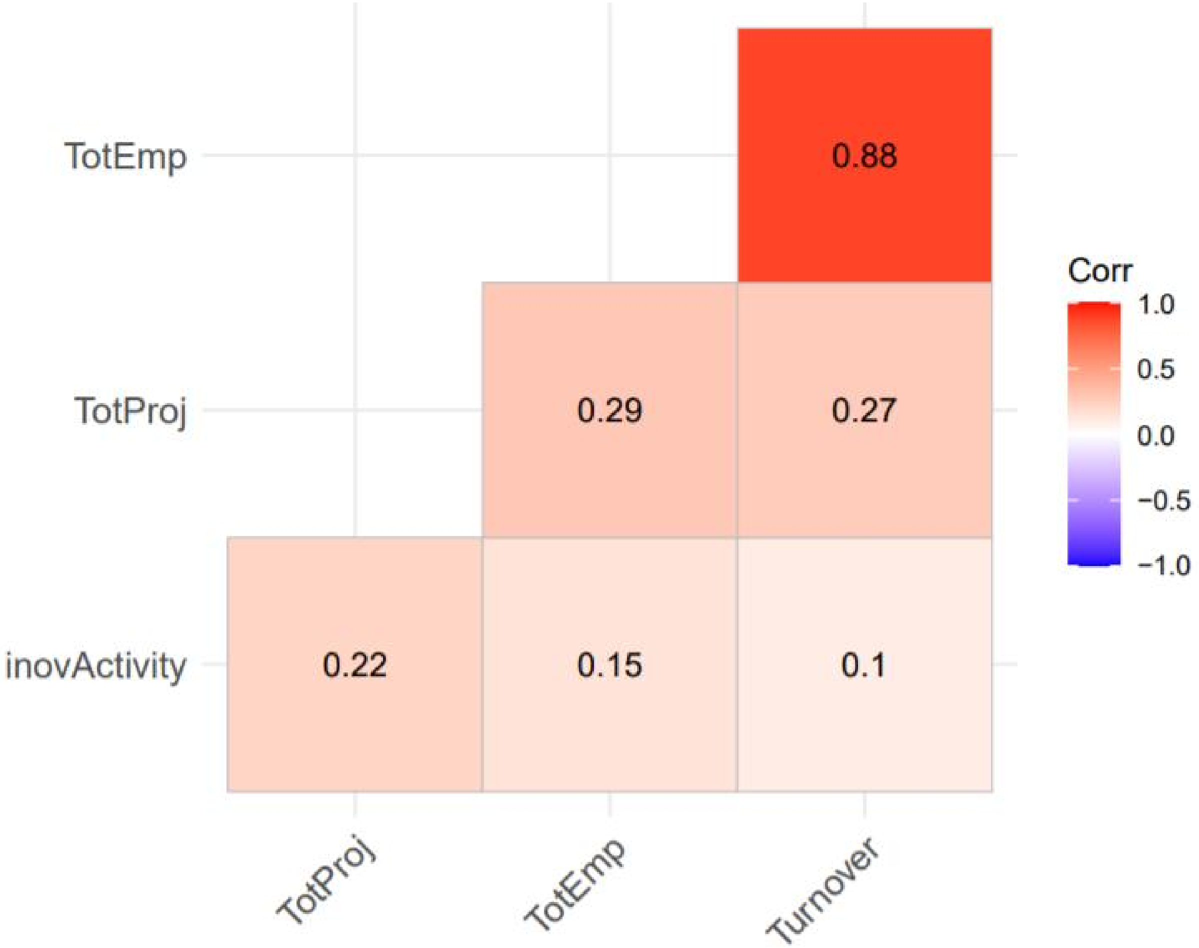
This can be also tested using two sample t-tests, where we will get the average mean for group 0 as 10 and average mean for group 1 as 17 at p-value<0.05.



Count of occurrences of group 0 and group 1.

We can answer our question which was -Does organisational innovation lead to higher sales?

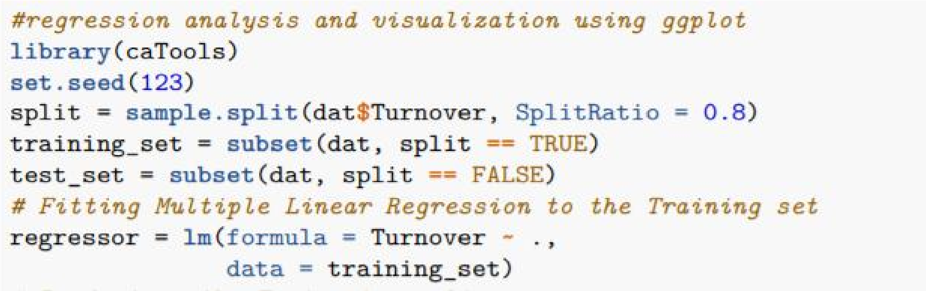
The answer is no as from the above visualization we cannot see that there is strong correlation between Turnover and Innovation Activity.
To make it obvious we can see correlation plots.



Here we see that Turnover is positively correlated with all other variables.
There is a strong correlation between Number of employees and Turnover but at the same time the correlation between Turnover and innovation activity is 0.1, which is insignificant .
So, Total sale of an organization is best explained by how many employees there are and there is negligible relationship between Total sale of an organization and Innovation Activity.

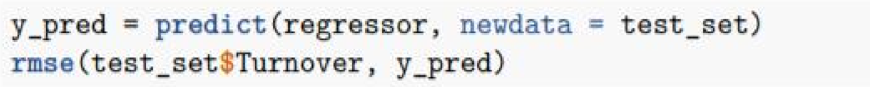
Machine Learning Regression

Now we can perform Multiple regression where we can use Turnover as our response variable. In this case we will have Total employee, Total Projects, Innovation Activity as Independent variable and Turnover as dependent variable, dataframe used is named as **dat.**
Before moving forward we need to split data as training and test sets. While building the model, the model is built on a training set and test set is used for testing the predictions.



```
#regression analysis and visualization using ggplot
library(caTools)
set.seed(123)
split = sample.split(dat$Turnover, SplitRatio = 0.8)
training_set = subset(dat, split == TRUE)
test_set = subset(dat, split == FALSE)
# Fitting Multiple Linear Regression to the Training set
regressor = lm(formula = Turnover ~ .,
               data = training_set)
```

Where the training set consists of 80% of total data and the remaining data is used in the while testing.
Here to build our regression model, we have to make use of the caTools lm function.
Next step will be the prediction of the result and checking the Root mean square error.



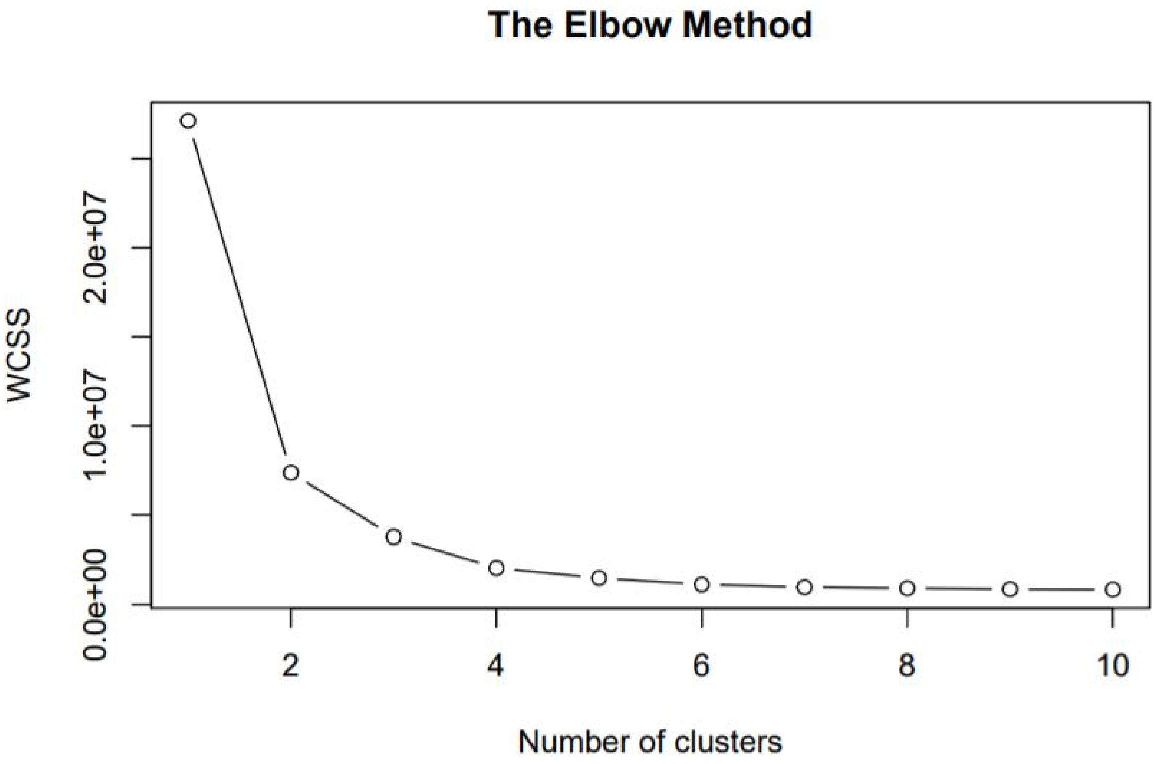
```
y_pred = predict(regressor, newdata = test_set)
rmse(test_set$Turnover, y_pred)
```

The RMSE for this mode is 17.759. Which is good.

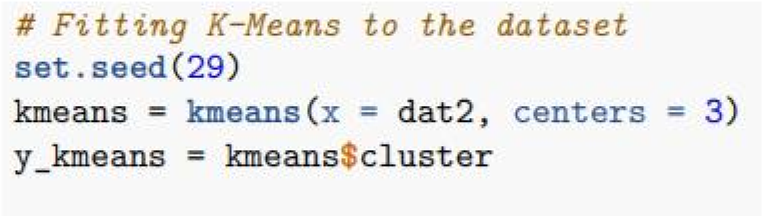
K-Mean Cluster

The K-mean Clustering is an unsupervised learning process that generates a dependent variable from a given independent variable.

This algorithm's main goal is to reduce within-cluster variation. The first step is to determine the best number of clusters, or K-value, which may be done using the Elbow Method or, more precisely, the Within Cluster Sum of Square method.
WCSS : The basic idea is to keep the total as low as possible. Let's imagine we have N points in a dataset and we utilise K-values that are equal to N (centre=N). Now, in this scenario, WCSS becomes 0 since each point in the dataset functions as a centroid, and so the distance will be 0 when employing the WCSS method, giving us perfect clusters. However, because all we have are clusters, this is worthless. As a result, there are optimum spots for K value/centre that we can find using the Elbow approach.

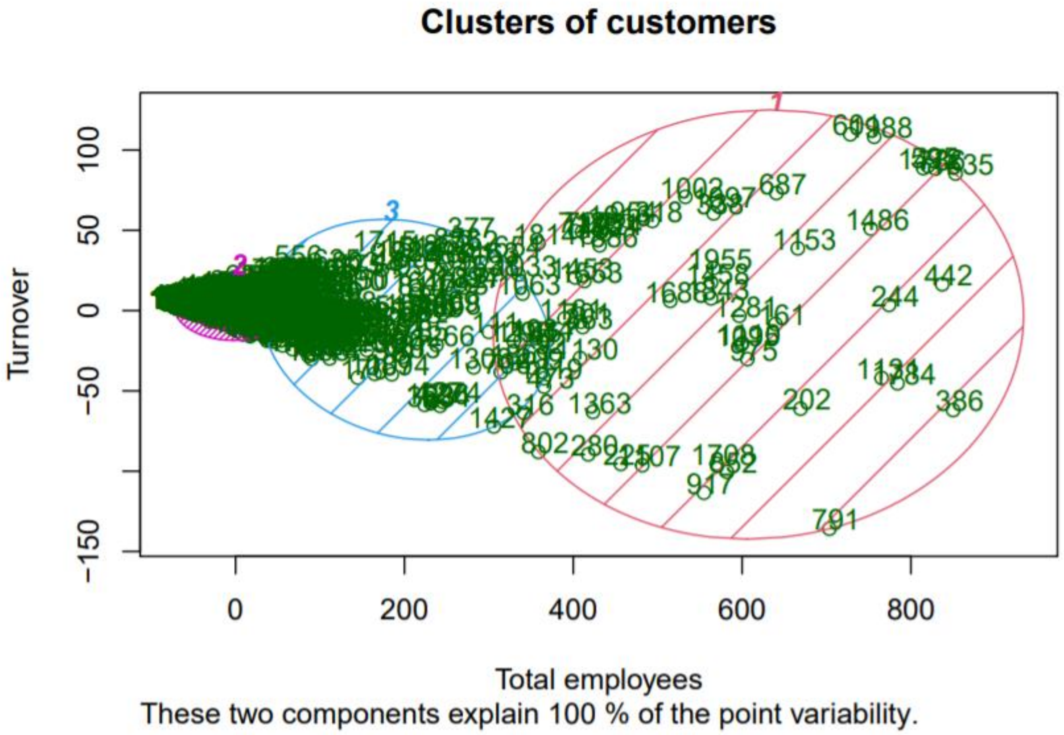


Using the elbow method, we got that the optimal cluster should be 3.
Using that we fit K-mean to the datasets, where the centre is 3.



```
# Fitting K-Means to the dataset
set.seed(29)
kmeans = kmeans(x = dat2, centers = 3)
y_kmeans = kmeans$cluster
```

To see how clustering is done, we make use of cluster library and plotted cluster plot.



Here from the cluster plot, we see that we got 3 clusters.

Now The first one, in the red circle, represents when the number of employees is around 600, in that case Turnover is very high.

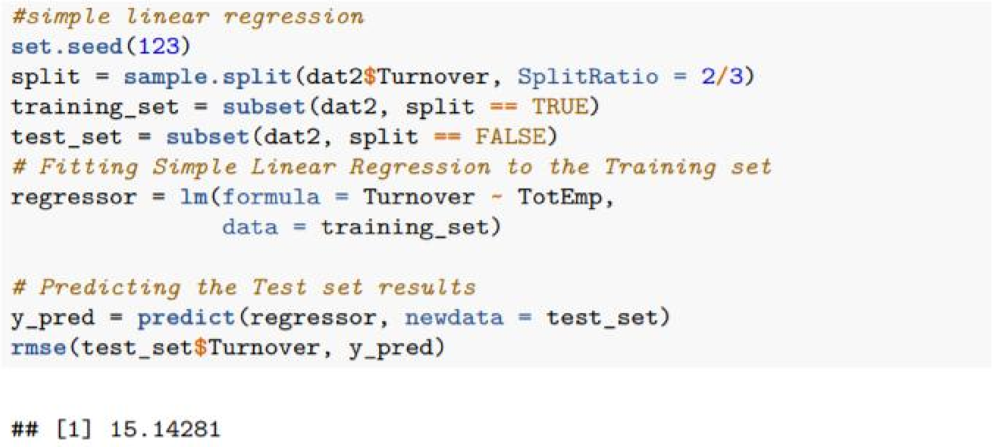
For the second one, with employees around 200, overall Turnover sales is moderate and for the last one, where the majority of population lies, where number of employees are least, turnover is also least of all of them.

So, we can say that Total sales and Total employees follow a linear nature.

Let’s build a simple linear regression model to see the relationship between Turnover and Total employees.

Simple linear regression model

We will make use of dat2, which have two columns named Turnover and Total employees to build a simple regression model.

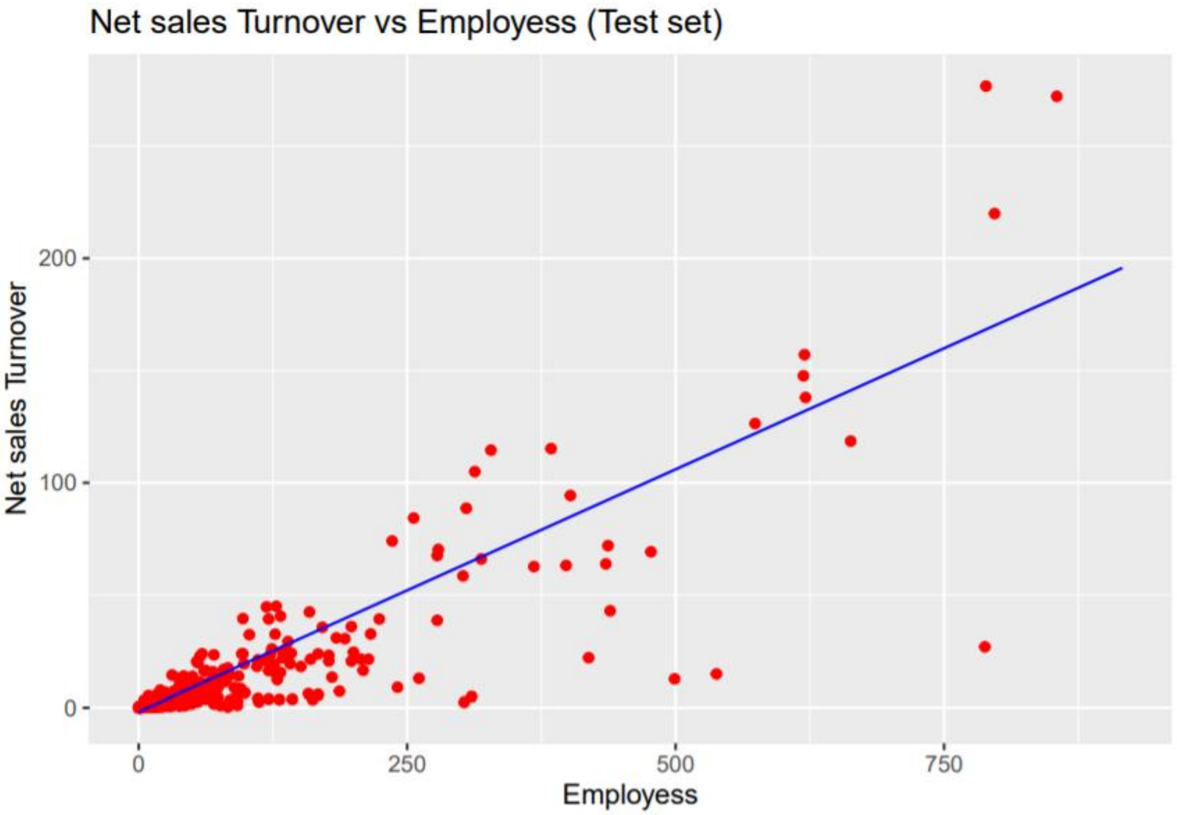


```
#simple linear regression
set.seed(123)
split = sample.split(dat2$Turnover, SplitRatio = 2/3)
training_set = subset(dat2, split == TRUE)
test_set = subset(dat2, split == FALSE)
# Fitting Simple Linear Regression to the Training set
regressor = lm(formula = Turnover ~ TotEmp,
               data = training_set)

# Predicting the Test set results
y_pred = predict(regressor, newdata = test_set)
rmse(test_set$Turnover, y_pred)
```

```
## [1] 15.14281
```

We can see we got RMSE 15.14, which is less than the 17.759 we got while building the Multiple regression model.
So, we can say that Total employee and Turnover are linearly proportional.
Let's see the plot.



Last but not the least, when we try to predict Turnover wrt Organization activity, we will get RMSE of 28.48, which clearly shows that there is no significant relationship between Turnover and Organization activity.

Summary

We loaded community innovation data in a dataframe df. There were a total of 2000 rows and a total of 211 columns. From that there are only a few columns which are necessary to answer the question Does organisational innovation lead to higher sales?
The columns are retrieved and renamed accordingly. The next step in data cleaning is to get rid of all null values and encode all categorical columns to numerical ones.A new column is created called '**innovActivity**' which represents whether Organizational innovation activity is performed or not.This column have two values 0 and 1, where 0 represent no Organizational innovation activity is done and 1 as Organizational innovation activity is performed.
The next step is data visualization which tells that there are a total of 949 samples of group 0 and 474 samples of group 1 in the '**innovActivity**' column.
Histogram and ggplot is used to view that visually.
To answer the question: Does organisational innovation lead to higher sales? We plotted a graph using ggplot where we saw that points on Y axis in both group 0 as well as 1 looks almost the same. Although the mean of group 1 was higher than that of mean group 0, there is no strong correlation between them. This theory is verified using correlation plot, which tells that sales/turnover is correlated to organisational innovation activity by a negligible factor of 0.1.
After that Using regression and clustering more insight is retrieved which tells Total sales is strongly correlated to number of employees and by using K-mean cluster, organization is grouped in 3 major groups.

References

1. Dr. Michael 07-2018, understanding K-mean Clustering in Machine learning Clustering, toward data science.
2. Chan YH. Biostatistics 201: Linear regression analysis. Age (years). Singapore Med J 2004;45:55-61.
3. Gaddis ML, Gaddis GM. Introduction to biostatistics: Part 6, correlation and regression. Ann Emerg Med 1990;19:1462-8.
4. Box, G.E.P., Hunter, W.G. and Hunter, J.S. (1978). Statistics For Experimenters: An Introduction to Design, Data Analysis and Model Building. New York: John Wiley and Sons, Inc.
5. Breusch, T. and Pagan, A. (1979). "A Simple Test for Heteroscedasticity and Random Coefficient Variation." Econometrica 47, 1287–1294.