Data Analytics

Internship Project

Time Frame : 1 month ~ 30 –32 hrs

Index

- Project 1 : Iris DataSet
- Project 2 : Titanic DataSet
- Project 3 : BigMart Sales Dataset

Project 1 : Iris Dataset

- **Perform Data Analysis on Iris dataset.**

Steps :

1. Import iris.csv file from folder iris_dataset. *– use read.csv()*
2. Exploratory Analysis.

   A. Explore / Print first 3 Records from Dataset. *– use head()*

   B. Find Dimension of Dataset. *– use dim()*

   C. Find Names , Class of features in the Dataset. *– use names(), class()*

   D. Find missing values (if any) & make the data consistent by removing it. *– use is.na(), na.omit()*

   E. Find Structure of Data. *– use str()*

Project 1 : Iris Dataset

Steps continued :

F. Find mean, median, quartile, max, min data for every feature.

– use summary()

G. Plot a Boxplot Graph, Pie chart respective to their Species.

– use boxplot(), pie(table(),)

H. Subset tuples based on their Species in different R-Object.

– use subset()

I. Plot a BoxPlot Graph for Individual R-Object. *– use boxplot()*

J. Plot a Histogram on feature Petal lengths of iris dataset .

– use hist()

Project 1 : Iris Dataset

Steps continued :

K. Plot a Histogram for Petal Lengths of **Different Species on different Graph.** *– use hist() & subset()*

L. Find correlation between multiple features also plot a scatter plot for correlation. *– use corr()*

4. Classify Data based on iris Species and plot a Decision Tree.

– use rpart,rattle package

– use rpart(), fancyRpartPlot()

# -- Draw a conclusion from above and Create a Report

Project 2 : Titanic Dataset

- ## Analysis of Survivors

The RMS Titanic was a British passenger liner that *sank* in the North Atlantic Ocean in the early morning of April 15, 1912 after colliding with an iceberg during her *maiden voyage* from Southampton to New York City. The ship contained 2,224 passengers and crew, out of which *1,500 died* in the unfortunate incident.

Perform a statistical analysis of the fatalities on the ship using the Titanic dataset.

The main question that we are addressing here is whether there is a **statistically significance** *relation between the death of the person and their passenger class, age, sex and/or port where they embarked their journey*.

Project 2 : Titanic Dataset

- **Perform an Data Analysis on Titanic dataset.**

Steps :

1. Import train.csv file from Titanic_dataset. *– use read.csv()*

2. Factors and Levels

   A. Find number of Passengers according to their Group Class: 1st , 2nd , 3rd *– use as.factor(),summary()*

   B. Find number of Passengers according to their Group Sex: Male, Female. *– use summary()*

   c. Find stats of Passengers Age. *– use summary()*

   Rectify if Age is less than one, is value fractional ?

Project 2 : Titanic Dataset

Steps continued :

D. Find number of Passengers according to their Group Embarked:
Place where the passenger embarked their journey. One of *Cherbourg, Queenstown or Southampton.*

– use summary()

3. Response Variables

A. Validate number of passengers who survived / Not Survived

– use as.factor(),summary()

4. Exploratory Data Analysis:

A. Explore / Print first n Records from Dataset.

– use head()

Project 2 : Titanic Dataset

Steps continued :

B. Find mean, median, quartile, max, min data for every feature.

– use summary()

C. For the purposes of this study, we work with only four input variables and one response variable.

Input variables : Passenger Class, Sex, Age, and Port of Embarkment.

Response variable : Survived.

– use slicing

D. Perform data cleaning steps *– use na.omit(),rownames()*

E. Encode Data

Make Age as a categorical variable as follows:

# If age <= 18, then age = child

# If 18 < age <= 60, then age = adult

# If age > 60, then age = senior *– use if else condition*

F. Validate above 2 steps. *– use head()*

Project 2 : Titanic Dataset

Steps continued :

5. Data Analysis to perform

Computing main effects for all four factors

Validate computed effects if True

Draw Conclusion over above Analysis

A. Plot the barplot of all four input variables: *– use barplot(table(),)*

B. Convert the categorical dataframe into numeric dataframe. *– use as.integer()*

6. Statistical Analysis:

A. Number of survivors on an average from Class & Plot a scatter plot

– use mean() for every class

– use plot(),axis() to plot for average value

Project 2 : Titanic Dataset

Steps continued :

B. Number of survivors on an average from Gender & Plot a scatter plot

– use mean() for every gender

– use plot(),axis() to plot for average value

C. Number of survivors on an average from Every Port of Embarkment & Plot a scatter plot.

– use mean() for every Port of Embarkment

– use plot(),axis() to plot for average value

D. Validate above scatterplots using ANOVA (1 way Interaction)

– use aov( ) & anova()

# -- Draw a conclusion from above and Create a Report

Project 3 : BigMart SalesDataset

- **Perform an Data Analysis on BigMart Sales Dataset.**

Steps :

1. Import Train & Test DataSet from BigMart Dataset folder.
   – use read.csv()
2. Check dimensions (number of row & columns) & Structure in dataset. *– use dim()*
3. Find Missing Values in the dataset. *– use table(is.na())*
4. Find Missing Values according to Columns.
   – use colSums(is.na())
5. Find Summary of DataSet & Draw Conclusions from it.
   – use summary()

Project 3 : BigMart SalesDataset

Steps continued :

6. ScatterPlots

A. Plot a **ScatterPlot** using ggplot for Item_Visibility vs Item_Outlet_Sales & draw conclusion from which products visibility is more sales. *– use ggplot(train,aes(V,O)) + geom_point()*

B. Plot a **Barplot** using ggplot for Outlet_Identifier vs Item_Outlet_Sales & Draw conclusion who has contributed to majority of sales. *– use ggplot(train,aes(OI,OS)) + geom_bar()*

C. Plot a **Barplot** using ggplot for Item_Type vs Item_Outlet_Sales also draw conclusion which items are sold more. *– use ggplot(train,aes(IT,OS)) + geom_bar()*

Project 3 : BigMart SalesDataset

Steps continued :

D. Plot a **Boxplot** using ggplot for Item_Type vs Item_Outlet_Sales also draw conclusion which items are sold more.

– use ggplot(train,aes(IT,OS))

+ geom_boxplot()

7. Manipulating Dataset to make it consistent

A. Add Item_Outlet_Sales Column to test dataset which is'nt available & assign integer 1. Also Combine Both Train + Test Datasets.

– use test$Item_Outlet_Sales <- 1

– use rbind()

B. Impute missing value in Item_Weight using median because it is highly robust to Outliers. *– use median()*

```
# df$a[is.na(df$a)] <- median(df$a, na.rm = TRUE)
```

Project 3 : BigMart SalesDataset

Steps continued :

C. We saw item visibility has zero value also, which is practically not feasible. Impute median value where item_visibility 0.

– use ifelse

D. Rename level in Outlet_Size to *since mis-matched levels in variables needs to be corrected..*

levels(combi$Outlet_Size)[1] <- "Other“

E. Rename levels of Item_Fat_Content since value are “LF” / “low fat”, so make them consistent. *– use library(plyr)*

df$a <- revalue(df$a , c(“LF” = “Low Fat”, “reg” = “Regular”)

F. Create a new column 2013 – Year (For Prediction).

df$Year <- 2013 – df$Outlet_Establishment_Year

Project 3 : BigMart SalesDataset

Steps continued :

G. Drop variables not required in modelling i.e. Item_Identifier, Outlet_Identifier, Outlet_Establishment_Year as they aren't needed for prediction. *– use library(dplyr)*

# df <- select(df , -c(col , col ,col))

H. Divide data set into Train and Test.

# new_train <- df[1:nrow(train),]

I. Perform a Regression testing on training dataset

– use lm() on train data

J. Plot Summary and Predict sales for Testing Dataset.

– use summary(),predict()

# -- Draw a conclusion from above and Create a Report