Exploratory Data Analysis

Statistical Analysis in R is performed by many inbuilt functions. Most of these functions are part of the base R package.

1. Basic Statistics Functions

mean()
Syntax mean(x,trim=0,na.rm=FALSE)
X - input vector
Trim is used to drop some observations from both end of the sorted vector and na.rm is used to remove the missing values from the input vector

```
x = c(10,2,30,4,5,6,70,8,9,10)
y=c(1,2,3,NA,4,5)
mean(x)
mean(x,trim = 0.2)
mean(y,na.rm=TRUE)
```

1.1 Mean of a column of a dataframe

```
df1 = data.frame(Name = c('Rahul','Amy', 'Anu','Aadhi','Ravi','Priya','Raks'),
gpa=c(9,9,8,7,6,7,8),
c_score=c(45,78,44,89,66,49,72),
python_score=c(56,52,45,88,33,90,47))
df1
mean(df1$c_score)
```

1.2 Median()

Middle most value in a dataseries is called the median. The median function() function is used.
Syntax median(x,na.rm=FALSE)

```
median(x)
median(y,na.rm=TRUE)
```

To determine the median of one column

```
median(df1$c_score)
```

To determine the median of all columns, use apply function

```
apply(df1,2,median)
```

1.3 Measure of Dispersion

Standard deviation is the most used measure of dispersion in statistics.The Standard Deviation is a measure of how spread out numbers are.Standard deviation is a measure of dispersement in statistics. “Dispersement” tells you how much your data is spread out. Specifically, it shows you how much your data is spread out around the mean or average.The higher the standard deviation, the wider the spread of values.The lower the standard deviation, the narrower the spread of values. It is calculated using sd(data).

sd(x)

To determine the standard deviation of one column of a dataset

sd(df1$c_score)

1.4 Minimum and Maximum values

To determine the minimum and maximum values, range() function is used.

range(x)

1.5 Quantiles

x=c(1,2,3,4,5,6,7,8,9,10)
quantile(x)
quantile(x,c(0.2,0.4))

1.6 Percentiles

To calculate the percentile we pass the data and the value of the required percentile.

quantile(x,probs=0.5)

1.7 Variance

To determine the variance of a dataset we use var() function.

var(x)

To determine variance of the column of a dataframe**.**

var(df1$c_score)

2. Linear, Logistic Regression and Making Prediction

Regression is a statistical measure used in finance, investing and other disciplines that attempts to determine the relationship between one dependent variable and one dependent variable. There are different types of regression like simple linear regression, multiple linear regression, logistic regression, poisson regression etc.

2.1 Linear Regression

Mathematically a linear relationship represents a straight line when plotted as a graph. The general mathematical equation for a linear regression is
y=ax+b, where is the independent variable, y is the dependent variable, a is the y intercept and b is the slope of the line.

To establish a Linear Regression in R are as follows

1. Collect the sample values of dependent and independent variables
   ```
   year=c(1997,1998,1999,2000,2001,2002)
   demand=c(50,60,50,80,72,90)
   ```
2. Create a relationship using lm() function.
   ```
   relation=lm( demand ~ year)
   ```
3. Find the coefficients from the model created
   ```
   print(relation)
   ```
4. Get the summary of the relation to know the average error in prediction
   ```
   print(summary(relation))
   ```
5. To predict the demand in future use the predict() function
   ```
   a=data.frame(year=2008)
   result=predict(relation,a)
   result
   ```

The regression can be visualized graphically as shown below

```
plot(year,demand,type = “o”, xlab = “year”, ylab=”demand”, main = “demand of a product”)
```

H0: There is no linear relationship between the variables
H1: There is a linear relationship between the variables.

2.1 Logistic Regression

The logistic regression is a regression model in which the response variable has categorical values such as True/False or 0/1. The general equation of logistic regression is as follows.

y = 1/(1+e^-(a+b1x1+b2x2+b3x3+...))

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. In logistic regression, the dependent variable is binary i.e it only contains data coded as 1 (TRUE) or 0 (False). To create a logistic regression, we use the glm() function.

Problem

The age,height and smoking status of 10 people are summarized in the table given below. Create a logistic regression model between smoking status ad three other variables age, height and weight. IN this problem, the dependent variable is smoking status and hence coded as 0 or 1.

Age	**30**	**25**	**32**	**40**	**70**	**60**	**38**	**54**	**63**	**57**
Height	**165**	**180**	**172**	**150**	**169**	**175**	**179**	**168**	**171**	**185**
Weight	**68**	**80**	**76**	**54**	**71**	**78**	**84**	**73**	**79**	**90**
Smoking Status	**0**	**1**	**1**	**1**	**0**	**0**	**1**	**0**	**1**	**0**

```
df1=read.csv("smoking.csv")
df1
model = glm(df1$smoking.status ~ df1$Age+df1$height+df1$weight,family=binomial, data = df1)
model
summary(model)
```

```
Call:
glm(formula = df1$smoking.status ~ df1$Age + df1$height + df1$weight,
    family = binomial, data = df1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9229  -0.6304   0.1857   0.8222   1.1715

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  95.2480    74.9908   1.270    0.204
df1$Age      -0.1824     0.1330  -1.372    0.170
df1$height   -0.9188     0.7457  -1.232    0.218
df1$weight    0.9338     0.7754   1.204    0.228

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 13.8629  on 9  degrees of freedom
Residual deviance:  8.7483  on 6  degrees of freedom
AIC: 16.748

Number of Fisher Scoring iterations: 6
```

Null Hypothesis: Ho: There is no statistically significant relationship between the dependent and independent variables
Alternate Hypothesis: H1: There is a significant relationship between the dependent variables and independent variables.

If the p-value is a smaller amount than 0.05, we reject the null hypothesis that there is no difference between the means and conclude that ..

```
df1=read.csv("smoking.csv")
df1
x=df1$smoking.status
y=df1$weight
z=df1$height
model = glm(x ~ y+z,family=binomial, data = df1)
model
summary(model)
predict(model, data.frame(z=165,y=56), type="response")
```