The Nature of Data

Introduction

Given a Dataset in which each row of the file is an observation relating to one person from Australia. The file contains the variables:

• City: The city in which the person lives.
• Movie: A movie that the person most recently watched.
• Age: The age of the person.
• Rating: The rating (out of 10) that the person gave the movie.

```
project<-read.csv("project2018S.csv")
```

```
str(project)
```

```
'data.frame':     1000 obs. of  4 variables:
 $ City  : Factor w/ 6 levels "Adelaide","Brisbane",..: 6 4 4 4 5 2 6 6 ...
 $ Movie : Factor w/ 3 levels "The Cat with Two Tales",..: 2 1 1 1 3 1  ...
 $ Age   : int  24 27 15 30 46 14 33 23 27 36 ...
 $ Rating: int  8 8 8 7 8 8 9 7 8 9 ...
```

```
dim(project)
[1] 1000    4
```

```
head(project)
```

	City	Movie	Age	Rating
1	Sydney	Undergoal	24	8
2	Melbourne	The Cat with Two Tales	27	8
3	Melbourne	The Cat with Two Tales	15	8
4	Melbourne	The Cat with Two Tales	30	7
5	Perth	Washing Dishes 3	46	8
6	Brisbane	The Cat with Two Tales	14	8

1. Representative Sample

```
project<-read.csv("project2018S.csv")
City =c("Adelaide", "Brisbane", "Hobart", "Melbourne", "Perth", "Sydney")
Population=c(1313927,2408223,226884,4850740,2050138,5131326)
Propotion=c(0.082, 0.151, 0.014, 0.304, 0.128, 0.321)
OzPopulation= data.frame(City,Population,Proportion)
OzPopulation
```

```
City  Population  Proportion
1 Adelaide  1313927     0.082
2 Brisbane  2408223     0.151
3 Hobart    226884      0.014
4 Melbourne 4850740     0.304
5 Perth     2050138     0.128
6 Sydney    5131326     0.321
```

```
OzSample.count=table(project$City) #given count from the dataset
OzSample.count
```

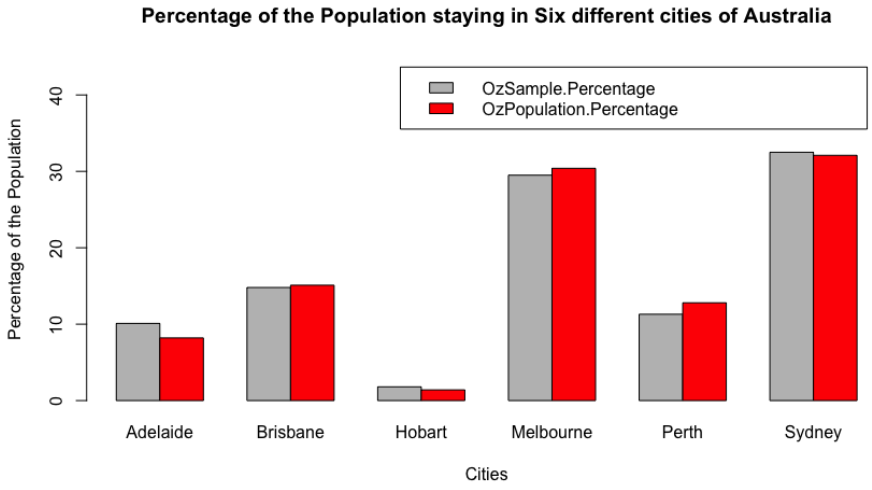
```
      Adelaide Brisbane    Hobart Melbourne   Perth    Sydney
      101       148        18       295       113       325
```

```
#proportion and %age calculation for given dataset
OzSample.Propotion =OzSample.count/1000
OzSample.Percentage =OzSample.Propotion*100
```

```
#proportion and %age calculation for given population
OzPopulation$Propotion =c(0.082, 0.151, 0.014, 0.304, 0.128, 0.321)
OzPopulation.Percentage=OzPopulation$Propotion*100
```

Plot:

```
plot1=rbind(OzSample.Percentage,OzPopulation.Percentage)
barplot(plot1, xlab="Cities",
        ylab="Percentage of the Population",
        main="Percentage of the Population staying in Six different cities
of Australia",
        beside= TRUE, col=c("grey", "red"),ylim=c(0,45),legend= TRUE)
```



Observation from the bar-Plot:

The above bar-plot shows the Percentage of the population staying in six different cities of Australia. The cities are shown on the x-axis and the Percentage of the population along the y-axis. The bars with grey colour show the percentage of population from the sample dataset and the bars with red colour displays the population percentages of the population given. The perusal of graph shows that the population spread of the 1000 observation from the Sample and the Australian Population is nearly the same in various cities of Australia.

Analysis and Result:

```
### calculating chisq distance
e.count= OzPopulation$Propotion*1000     #considering the same scale and
calulating the no of individuals city-vise

chi2=sum((OzSample.count-e.count)^2/e.count)
chi2
```

```
[1] 7.679003
```

Assuming the following Hypothesis

Null Hypothesis, H_0= The percentages of the sample population matches to the Population of Australia or there is no difference between population proportion and sample proportion.

Alternative Hypothesis, H_1=There is a difference between the population and sample proportions.

In order to test the hypothesis, chi-square test is applied.

```
chisq.test(OzSample.count, p=OzPopulation.Percentage, rescale.p=TRUE,
simulate.p.value = TRUE, B=1000)
```

```
      Chi-squared test for given probabilities with simulated p-value
(based on
      1000 replicates)
```

```
data:  OzSample.count
X-squared = 7.679, df = NA, p-value = 0.1716
```

The results show that the p-value (0.1716) is greater than 0.05. Therefore, we fail to reject null hypothesis that there is no difference between population proportion and sample proportion in this study.

Conclusion:

The graphical and statistical (chi-square test) method given above shows that no statistically significant difference between the sample proportion and population proportion. From chi square test we got the p-value (0.1716) which is not significant thus we fail to reject the null hypothesis and can conclude the sample is representative of the population.

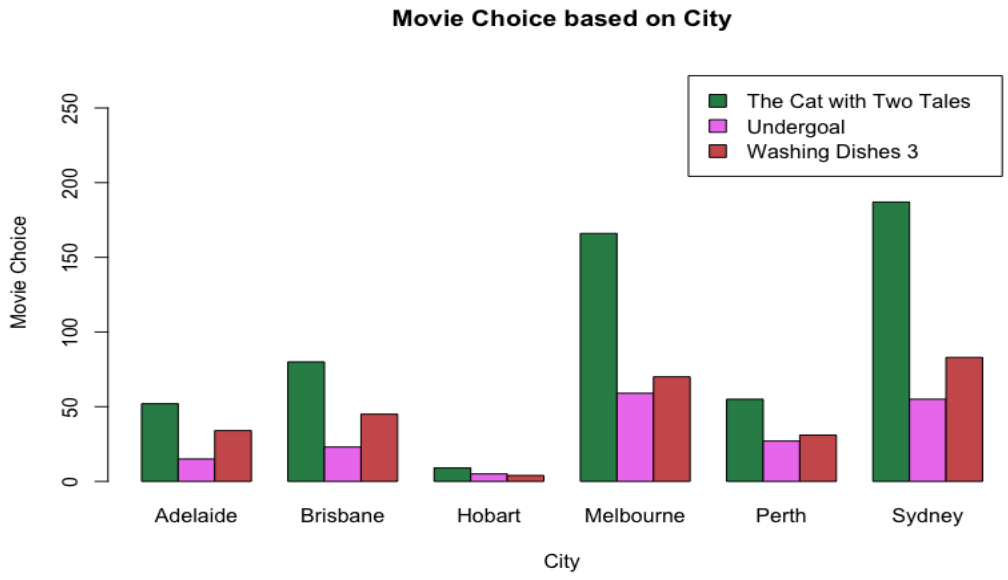
2. Chosen Movie

```
project<-read.csv("project2018S.csv")
movie_city=table(project$Movie,project$City)
movie_city
```

```
                          Adelaide  Brisbane Hobart Melbourne Perth Sydney
  The Cat with Two Tales        52        80      9       166    55    187
  Undergoal                     15        23      5        59    27     55
  Washing Dishes 3              34        45      4        70    31     83
```

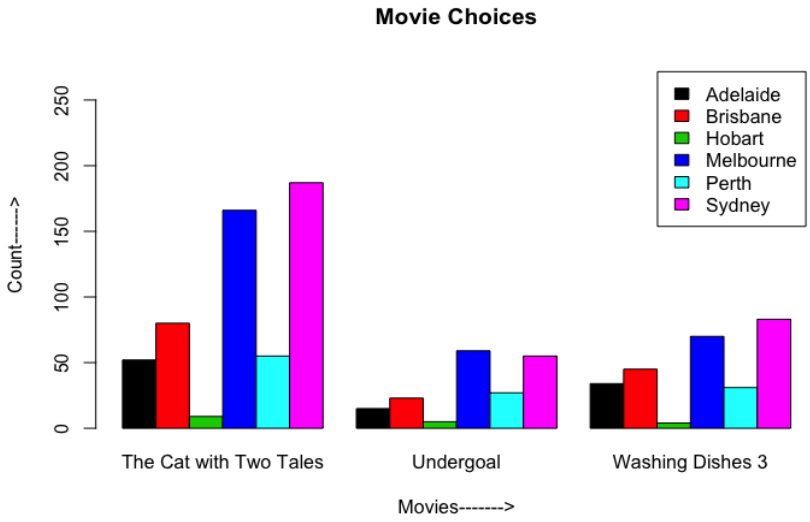
Plot 1:

```
barplot(movie_city,beside = TRUE,
      col=c ("sea green", "violet","indian red"),
      xlab="City",
      ylab="Movie Choice",
      main="Movie Choice based on City", ylim=c (0,280), legend=TRUE)
```



Plot 2:

```
barplot(city_movie,beside = TRUE, col=1:6,
                        xlab="Movies------->",
                        ylab="Count------>",
                  main="Movie Choices", ylim=c (0,280), legend=T)
```



Observation from the plot:

Plot1 and Plot 2 describes the Movie Choices in different cities. From the above bar-plots the Movie "The Cat with two tales" is preferred in each of the six cities in comparison to the "Undergoal" and "Washing Dishes 3". On the other hand, the choice preferences of "Undergoal" and "Washing Dishes 3" seem to almost similar. Yet, by looking at both the bar-plots it is very hard to say that whether city plays any role in movie choices.

Analysis and Results:

In order to authenticate the results of bar-plot given above, the statistical measure of chi-square applied to test the following hypothesis.
Null Hypothesis, H_0: The movie preference is independent of cities.

Alternative Hypothesis, H_1: The movie preference is dependent on cities.

The chi sq. approximation and chi-sq. simulation on the data are performed in order to analyse the data:

```
chisq.test(movie_city)
```

```
Pearson's Chi-squared test
data:  movie_city
X-squared = 10.375, df = 10, p-value = 0.4082
```

```
chisq.test(movie_city,simulate.p.value = TRUE,B=2000)
#p value is high, so we can't say that variables are dependent.
```

```
Pearson's Chi-squared test with simulated p-value (based on 2000
      replicates)

data:  movie_city
X-squared = 10.375, df = NA, p-value = 0.4023
```

Above results clearly show that the p-value (0.4023) is greater than 0.05. Therefore, we fail to reject the null hypothesis and can say that there is no significant difference between the city and the preference of the chosen movie.

Conclusion

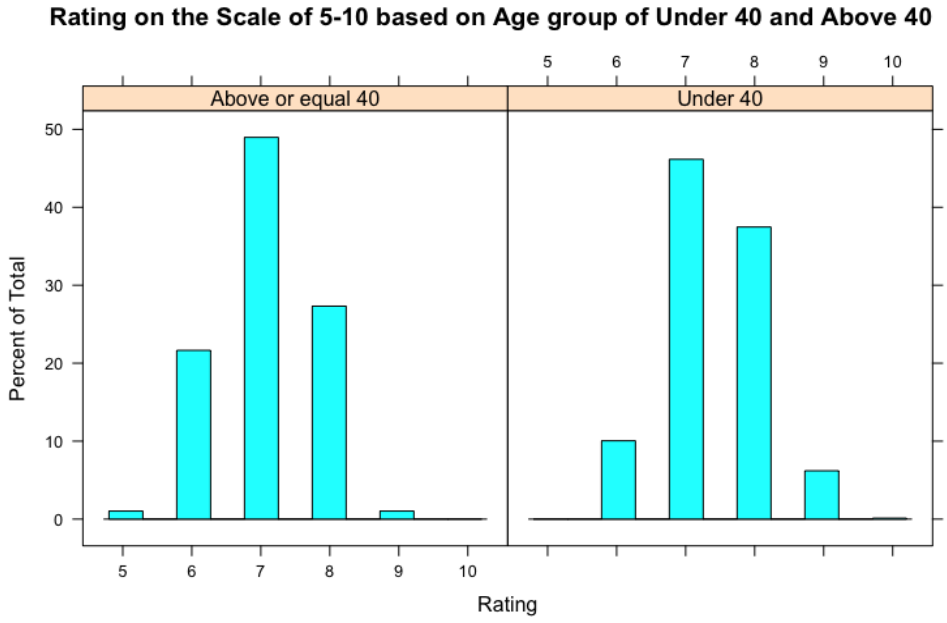
Since from the chisq approximation, the p value is higher than 0.05, it can be concluded that the movie preference is independent of the cities.

3. Harsh Raters

```
project<-read.csv("project2018S.csv")
Age1= ifelse(project$Age< 40,"Under 40","Above or equal 40")
project1= data.frame(project,Age1)
```

Plot:

```
library(lattice)
histogram(~Rating|Age1, data=project1, main="Rating on the Scale of 5-10
based on Age group of Under 40 and Above 40")
```



Observation from the plot:

The bar-plot shows that rating grouped under the category of "Above or equal 40" and "Under 40" (Age). It broadly shows that the people above the Age of 40 tend to rate the movies between 6-8 but

people with age group less than 40 tend to rate movies between 7 and 8. Yet, we cannot conclude that the older people rate lower in comparison to younger people.

Analysis and Results:

In order to show that the rating differs with age group significantly, let us statistically study the difference between the two means for the groups. Assuming the following hypothesis

Null hypothesis H_0: There is no difference between means of movie rating between people below 40 and people above or equal 40.

Alternative hypothesis H_1: There is a difference in means of movie rating.

In order to test the difference between two means of two samples, we carry the t-test, assuming equal variances and the pooled variance.

```
t.test(Rating~Age1, data=project1, var.equal=TRUE)
```

```
      Two Sample t-test

data:  Rating by Age1
t = -5.7025, df = 998, p-value = 1.555e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.4641042 -0.2264640
sample estimates:
mean in group Above or equal 40          mean in group Under 40
                       7.056701                        7.401985
```

The p-value is 1.555e-08 which is 0 (approx.) which is lower than 0.05 thus rejecting the null hypothesis. The p-value implies that there is difference in the means of the rating between the two age groups.

Conclusion:

The results are evident to prove that the mean rating for the movies varies for different age groups i.e. for people above and equal 40 and people under 40.

4. Best Movie

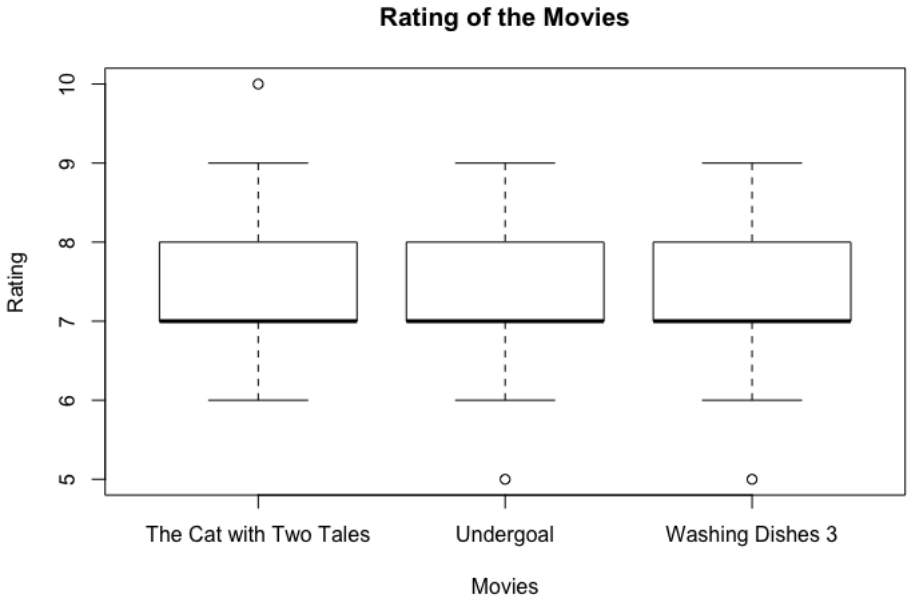
```
project<-read.csv("project2018S.csv")
movie<-project[,-c(1,3)]
movie_data<-table(movie)
movie_data
```

```
                       Rating
      Movie             5   6   7   8   9  10
The Cat with Two Tales   0  48 245 212  43   1
Undergoal1               1  18 100  58  7    0
```

```
  Washing Dishes3        1  57  122  85   2   0
```

Plot:

```
boxplot (Rating~Movie, data= project,
                  xlab="Movies",
                  ylab="Rating",
            main="Rating of the Movies")
```



Observation from the plot:

The above plot shows the rating for the movies namely "The Cat with Two Tales", "Undergoal", "Washing Dishes 3". Along the y-axis is the rating for the movie and along the x-axis is the movies. The graph doesn't say much about the difference because for every movie the 1st quartile and median is at level 7 and the 3rd quartile is at level 8. The interesting fact is that, there are outliers for the three movies. For movie "The Cat with Two Tales", the outlier is at level 10 whereas for the movie "Undergoal" and "Washing Dishes 3" the outlier is at level 5.

Analysis and Results:

In order to see whether the movies are equally preferred a chisq test is performed. Assuming the following hypothesis

Null Hypothesis H_0- movies are equally preferred,
Alternative Hypothesis H_1- movies are not equally preferred.

```
chisq.test(movie_data, simulate.p.value = TRUE, B=2000)
```

```
      Pearson's Chi-squared test with simulated p-value (based on 2000
replicates)
```

```
data:  movie_data
X-squared = 52.155, df = NA, p-value = 0.0004998
```

Since p-value is less than 0.05, the null hypothesis is rejected, and it can be said that the movies are not equally preferred.

In order to check, whether all mean ratings are equal for each movie, one-way ANOVA (aov) is run.

```
movie.r= aov(Rating~Movie, data= project)
summary(movie.r)
```

```
Df Sum Sq Mean Sq F value   Pr(>F)
Movie         2   22.4   11.22   19.68 4.16e-09 ***
Residuals   997  568.3    0.57
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The F-value is 19.68 and p-value is very low, and it is statistically significant at 95% Confidence level. Hence, it points to the fact we reject the null hypothesis(H_0) and accept the alternative hypothesis (H_1). From the above test it is concluded that the mean rating for different movies is different.
However, it is still not answered which movie is preferred over which. For this TukeyHSD test was performed to see where the actual difference lies.

```
TukeyHSD(movie.r)
```

```
Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Rating ~ Movie, data = project)

$Movie
                                      diff         lwr         upr     p adj
Undergoal-The Cat with Two Tales      -0.1782292 -0.3291908 -0.0272676033 0.0156861
Washing Dishes3-The Cat with Two Tales-0.3484783 -0.4807031 -0.2162535639 0.0000000
Washing Dishes 3-Undergoal            -0.1702491 -0.3400474 -0.0004508672 0.0492200
```

From the TukeyHSD there are three different pairs ans all the different pairs for movie are statistically significant having p-values 0.02, 0.00 and 0.05 respectively.
The difference in Tukey is pairwise so difference in mean for “Undergoal” and “The Cat with two Tales” is -0.18 which means that “The Cat with two Tales” has a higher rating than “Undergoal” and similarly the difference in mean for “Washing Dishes 3” and “The Cat with two Tales” is -0.35 which means that “The Cat with two Tales” has a higher rating than “Washing Dishes 3”. Thus, concluding that “The Cat with two Tales” is preferred among the above three movies.

Conclusion:

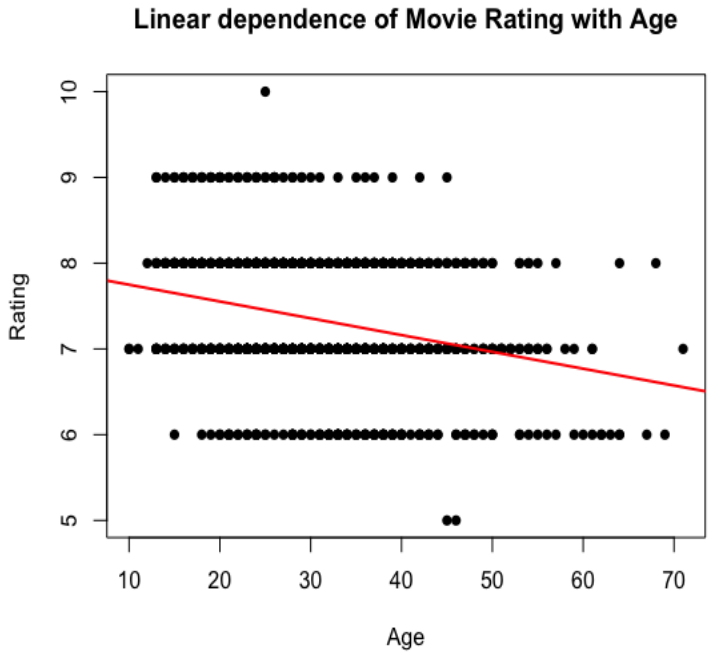
From the above tests, a conclusion can be drawn that the movies are not equally preferred. Among the three movies it’s evident that “The Cat with Two Tales” is preferred over the other two.

5. Age and Ratings

Simple linear regression model is applied taking Rating as a response variable and Age as the predictor.

Plot:

```
project<-read.csv("project2018S.csv")
MovieR_lm=lm(Rating~Age, data= project)
plot(x= project$Age, y= project$Rating, xlab= "Age",ylab= "Rating",
main="Linear dependence of Movie Rating with Age", pch=16)
abline(MovieR_lm, col="red",lwd=2)
```



Observation from the plot:

The above graph shows a Linear dependence of Movie Rating with Age. On y-axis we have taken a rating from scale 5 to 10, on the x-axis we have taken age ranging from 10 to 70.
It is evident from the plot that the trend line is negatively sloping showing that the rating of the movie and age are inversely related. It means that older people give low rating to movie and younger people more.

Analysis and Results of Simple Regression:

```
summary(MovieR_lm)
```

```
Call:
lm(formula = Rating ~ Age, data = project)

Residuals:
    Min      1Q    Median      3Q     Max
```

```
-2.0635 -0.4557   -0.1616     0.5836  2.5443

Coefficients:
            Estimate    Std. Error  t value     Pr(>|t|)
(Intercept) 7.945823    0.073595    107.966     <2e-16 ***
Age         -0.019607   0.002239    -8.756      <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7414 on 998 degrees of freedom
Multiple R-squared:  0.07134, Adjusted R-squared:  0.07041
F-statistic: 76.67 on 1 and 998 DF, p-value: < 2.2e-16
```

We have fitted simple regression by taking two variables; the response variable is Rating of the movie and predictor variable is age of the respondents.

Interpreting the slope of the Linear Regression. We run the Hypothesis Test (b(slope)=0)
Null Hypothesis H_0-b=0
Alternative Hypothesis H_1- b!=0

```
x=replicate(1000,{
  Age.Shuff = sample(project$Age)
  MovieR_lm1=lm(Rating~Age.Shuff, data= project)
  coef(MovieR_lm1)[2]
})
MovieR_lm=lm(Rating~Age, data= project)
slope= coef(MovieR_lm)[2]
pVal=mean(x>abs(slope)) + mean(x< -abs(slope))
pVal
[1] 0
```

As the p-Value is 0, it shows that age is statistically significant variable to determine Rating of the movie. As p-value is less than 0.05 , the null hypothesis is rejected. Thus concluding he population slope b≠0.

Prediction from the above model based on Age=32

```
predict(MovieR_lm, list(Age=32))
       1
7.318413

coef(MovieR_lm)

(Intercept)          Age
 7.94582266    -0.01960656
```

The Rating at the age 32 years is calculated to be 7.3184.

Conclusion

Regression coefficient represents the association of Rating of the movie and the Age of the respondent. The regression coefficient for age is -0.0196. It is proved that older people give low Rating to the movies. It can also be interpreted that with the increase in Age the rating for the movie decreases. The coefficient of age is statistically significant.