Research Methods

Analysing Survey Data with R

Andy Gillett

1

Task

The HR manager of your company has asked you to investigate answers to several HR related questions and write a short report. You have access to a survey of the employees that has recently been carried out (jss.xlsx). The questions you need to investigate are:

1. What is the age distribution of the workforce?
2. What proportion of employees belongs to each ethnic group?
3. What is the average income?
4. How is number of years worked related to salary, if at all.
5. How different are the average salaries of the different skill categories?
6. Is there a significant difference between the proportion of males and females who attended the firm's meeting last month?

2

Relevant details of variables on the Job Satisfaction Survey file (jss.sav)	
ID	Employee Number
ETHNICGP	Ethnic group Missing Values: 0 Value Label 0 not reported 1 white 2 south asian 3 east asian 4 black 5 other
GENDER	Gender Value Label 1 male 2 female
INCOME	Annual income in pounds (to nearest hundred) Missing Values: 0
AGE	Age in years Missing Values: 0
YEARS	Years worked
ATTEND	Attended meeting Value Label 1. yes 2. no
SKILL	Rated skill level Value Label 1. unskilled 2. semi-skilled 3. fairly-skilled 4. highly skilled

3

Data

id	ethnicgp	gender	income	age	years	attend	skill
21	White	Male	5300	43	16	No	Unskilled
12	Black	Male	5500	19	2	Yes	Unskilled
36	Black	Male	5900	19	1	No	Unskilled
28	White	Female	6200	31	9	No	Unskilled
32	South Asia	Female	6200	28	10	Yes	Unskilled
63	South Asia	Female	6400	26	8	Yes	Unskilled

save as "jss.csv"

58	White	Female	6900	48	28	No	Semi-Skilled
26	White	Male	7000	21	1	Yes	Fairly Skilled
43	Black	Female	7000	27	5	No	Fairly Skilled
62	White	Male	7100	21	2	Yes	Fairly Skilled
11	White	Male	7100	29	4	No	Fairly Skilled
66	East Asian	Female	7100	30	8	Yes	Fairly Skilled
31	White	Male	7200	54	12	No	Fairly Skilled
2	Black	Male	7300	26	5	Yes	Fairly Skilled
49	White	Female	7300	42	10	No	Fairly Skilled
67	White	Male	7300	29	8	No	Fairly Skilled
54	South Asia	Female	7400	26	8	No	Fairly Skilled

4

Import Data

```
> setwd("C:/Users/andyg/Desktop/R")
```

```
> jss <- read.csv("jss.csv", header = TRUE,
colClasses = c("numeric","factor","factor",
"numeric", "numeric", "numeric", "factor",
"factor"))
```


5

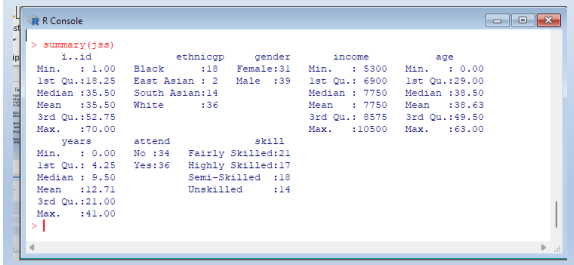
Import Data

```
> setwd("C:/Users/andyg/Desktop/R")
> jss <- read.csv("jss.csv", header = TRUE, colClasses = c("numeric","factor","f
> jss
   X..id    ethnicgp gender income age years attend          skill
1     21       White   Male   5300  43    16     No      Unskilled
2     12       Black   Male   5500  19     2    Yes      Unskilled
3     36       Black   Male   5900  19     1     No      Unskilled
4     28       White Female   6200  31     9     No      Unskilled
5     32 South Asian Female   6200  28    10    Yes      Unskilled
6     63 South Asian Female   6400  26     8    Yes      Unskilled
7     52       White Female   6500  31     2     No      Unskilled
8     45       White   Male   6600   0    10    Yes      Unskilled
9     46 South Asian   Male   6700  18     1    Yes      Unskilled
10    18       White Female   6700  18     1    Yes      Unskilled
11    40       Black   Male   6700  31    15     No      Unskilled
12    29       White   Male   6800  31    12    Yes      Unskilled
13    42 South Asian   Male   6800  23     3    Yes      Unskilled
14    48       White Female   6800  29    10    Yes      Unskilled
15    64       White Female   6800  46     0     No   Semi-Skilled
16    10       Black Female   6900  27     6    Yes   Semi-Skilled
17    68 South Asian   Male   6900  45     9     No   Semi-Skilled
18    70       White   Male   6900  47    22    Yes   Semi-Skilled
19    58       White Female   6900  48    28     No   Semi-Skilled
20    26       White   Male   7000  21     1    Yes Fairly Skilled
21    43       Black Female   7000  27     5     No Fairly Skilled
```


6

Summary

```
> summary(jss)
```



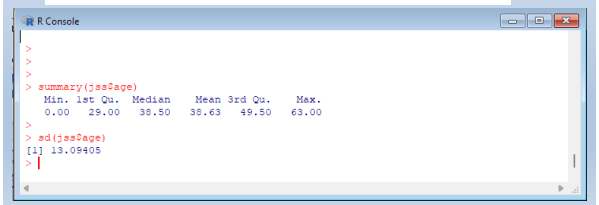
7

7

Task 1

What is the age distribution of the workforce?

```
> summary(jss$age)

> sd(jss$age)
```

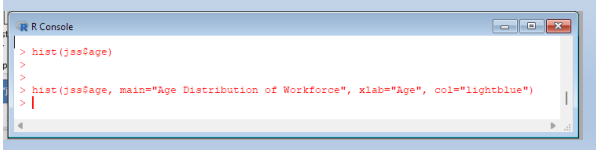


8

8

Task 1

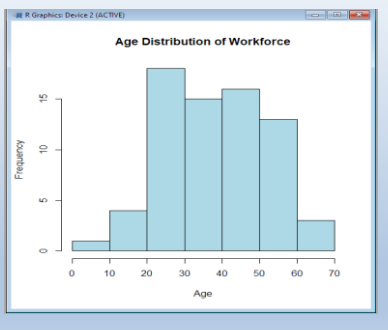
```
> hist(jss$age)

> hist(jss$age, main="Age Distribution of Workforce", xlab="Age", col="lightblue")
```



9

9

Task 1



File -> Save As

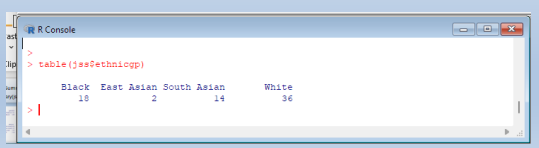
10

10

Task 2

What proportion of employees belongs to each ethnic group?

```
> table(jss$ethnicgp)
```



11

11

Task 2

```
> prop <- table(jss$ethnicgp)
```

Proportions

```
> prop.table(prop)
```

To produce percentages rounded to whole numbers

```
> round(100*prop.table(prop),digits=0)
```

12

12

Task 2

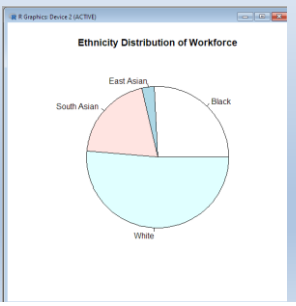
```
> table(jss$ethnicgp)

      Black  East Asian South Asian       White
         18           2          14          36
> prop <- table(jss$ethnicgp)
>
> prop

      Black  East Asian South Asian       White
         18           2          14          36
> prop.table(prop)

      Black  East Asian South Asian       White
 0.25714286  0.02857143  0.20000000  0.51428571
> round(100*prop.table(prop),digits=0)

      Black  East Asian South Asian       White
         26           3          20          51
>
>
>
```

13

13

Task 2

```
> pie(prop, main="Ethnicity Distribution of Workforce")
```



14

14

Task 3

What is the average income?

```
> summary(jss$income)
> mean(jss$income)
```

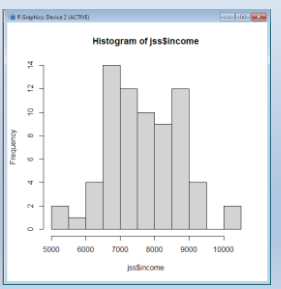
```
>
>
> summary(jss$income)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   5300    6900    7750    7750    8575   10500
> mean(jss$income)
[1] 7750
>
```

15

15

Task 3

```
> hist(jss$income)
```



16

16

Task 3

```
> tapply(jss$income, jss$gender, mean)
> tapply(jss$income, jss$ethnicgp, mean)
```

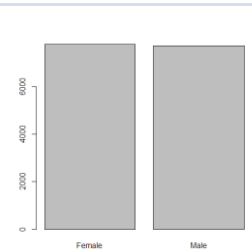
```
> tapply(jss$income, jss$skill, mean)
Fairly Skilled Highly Skilled   Semi-Skilled      Unskilled
      7695.238       9129.412       7588.889       6364.286
> tapply(jss$income, jss$gender, mean)
  Female     Male
7777.419 7728.205
> tapply(jss$income, jss$ethnicgp, mean)
      Black  East Asian South Asian       White
   7761.111    8050.000    7728.571    7736.111
>
```

17

17

Task 3

```
> barplot(tapply(jss$income, jss$gender, mean))
```

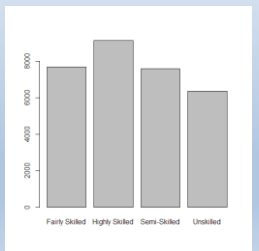


18

18

Task 3

```
> barplot(tapply(jss$income, jss$skill, mean))
```



19

19

Task 4

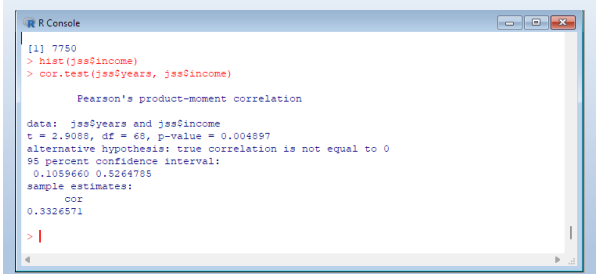
How is number of years worked related to salary, if at all?

```
> cor.test(jss$years, jss$income)
```

20

20

Task 4



21

21

Task 4

Scatterplot

```
> plot(jss$years, jss$income)
```

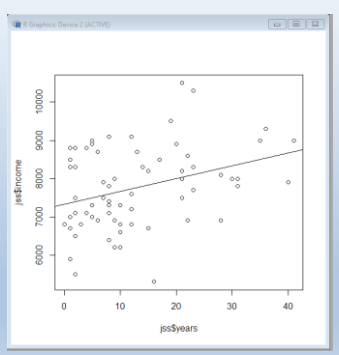
Add a line of best fit - Linear Regression

```
> abline (lm(jss$income ~ jss$years))
```

22

22

Task 4



23

23

Task 4

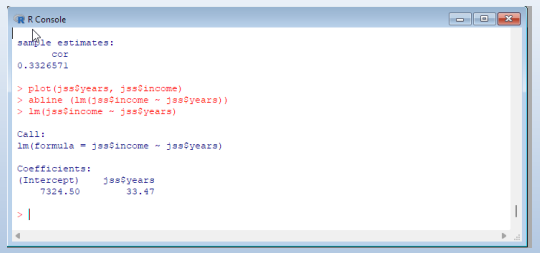
Linear Regression Equation

```
> lm(jss$income ~ jss$years)
```

24

24

Task 4



Relationship between years worked and salary is:

Salary = £7324.50 + 33.47 x Years Worked

25

25

Task 5

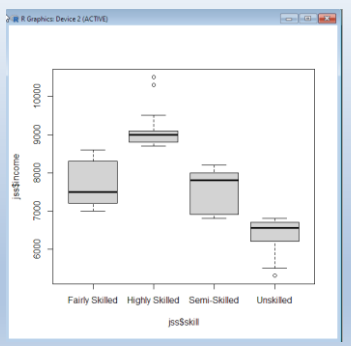
How different are the average salaries of the different skill categories?

```
> boxplot(jss$income ~ jss$skill)
```

26

26

Task 5

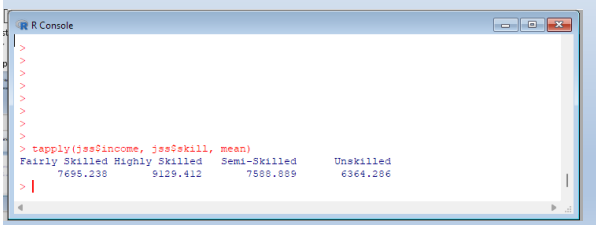


27

27

Task 5

```
> tapply(jss$income, jss$skill, mean)
```



28

28

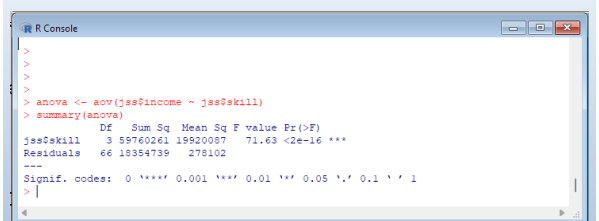
Task 5

```
> anova <- aov(jss$income ~ jss$skill)

> summary(anova)
```

29

29

Task 5

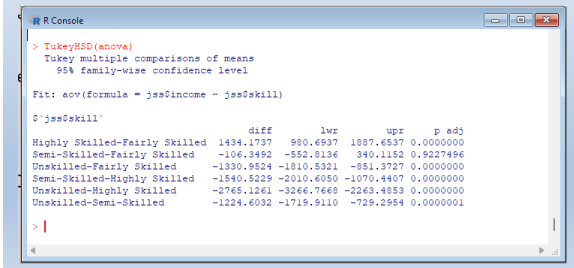


30

30

Task 5

> TukeyHSD(anova)



31

Task 6

Is there a significant difference between the proportion of males and females who attended the firm's meeting last month?

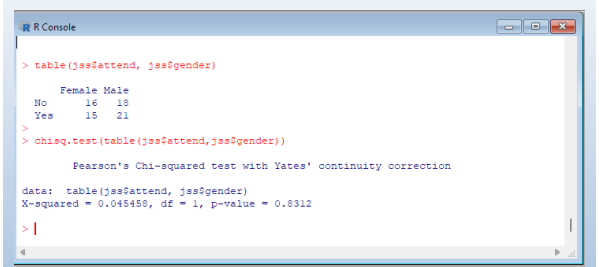
> table(jss$attend, jss$gender)

> chisq.test(table(jss$attend,jss$gender))

32

32

Task 6



21 out of 36 of men (55%) attended the meeting, whereas 14 out of 32 (44%) of women attended the meeting. So the percentage of men attending the meeting was higher than the percentage of women attending the meeting. However, Sig = 0.83. Therefore, there is no significant difference between M/F attendance at the meeting.

33

33