Assignment No. 11.2

This assignment is based on the dataset named bank.additional, which has the following characteristics

> str(bank.additional)

'data.frame': 4119 obs. of 21 variables:

$ age : int 30 39 25 38 47 32 32 41 31 35 ...

$ job : chr "blue-collar" "services" "services" "services" ...

$ marital : chr "married" "single" "married" "married" ...

$ education : chr "basic.9y" "high.school" "high.school" "basic.9y" ...

$ default : chr "no" "no" "no" "no" ...

$ housing : chr "yes" "no" "yes" "unknown" ...

$ loan : chr "no" "no" "no" "unknown" ...

$ contact : chr "cellular" "telephone" "telephone" "telephone" ...

$ month : chr "may" "may" "jun" "jun" ...

$ day\_of\_week : chr "fri" "fri" "wed" "fri" ...

$ duration : int 487 346 227 17 58 128 290 44 68 170 ...

$ campaign : int 2 4 1 3 1 3 4 2 1 1 ...

$ pdays : int 999 999 999 999 999 999 999 999 999 999 ...

$ previous : int 0 0 0 0 0 2 0 0 1 0 ...

$ poutcome : chr "nonexistent" "nonexistent" "nonexistent" "nonexistent" ...

$ emp.var.rate : num -1.8 1.1 1.4 1.4 -0.1 -1.1 -1.1 -0.1 -0.1 1.1 ...

$ cons.price.idx: num 92.9 94 94.5 94.5 93.2 ...

$ cons.conf.idx : num -46.2 -36.4 -41.8 -41.8 -42 -37.5 -37.5 -42 -42 -36.4 ...

$ euribor3m : num 1.31 4.86 4.96 4.96 4.19 ...

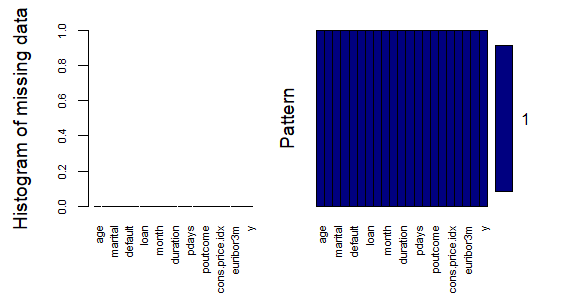
$ nr.employed : num 5099 5191 5228 5228 5196 ...

$ y : chr "no" "no" "no" "no" ...

The following is the data definitions of each of the variables

Input variables:  
# bank client data:  
1 - age (numeric)  
2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')  
3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)  
4 - education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')  
5 - default: has credit in default? (categorical: 'no','yes','unknown')  
6 - housing: has housing loan? (categorical: 'no','yes','unknown')  
7 - loan: has personal loan? (categorical: 'no','yes','unknown')  
# related with the last contact of the current campaign:  
8 - contact: contact communication type (categorical: 'cellular','telephone')   
9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')  
10 - day\_of\_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')  
11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.  
# other attributes:  
12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)  
13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)  
14 - previous: number of contacts performed before this campaign and for this client (numeric)  
15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')  
# social and economic context attributes  
16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)  
17 - cons.price.idx: consumer price index - monthly indicator (numeric)   
18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)   
19 - euribor3m: euribor 3 month rate - daily indicator (numeric)  
20 - nr.employed: number of employees - quarterly indicator (numeric)  
Output variable (desired target):  
21 - y - has the client subscribed a term deposit? (binary: 'yes','no')

> missdataplot <- aggr(bank.additional, col=c('navyblue','red'), numbers=TRUE, sortVars=TRUE, labels=names(data), cex.axis=.7, gap=3, ylab=c("Histogram of missing data","Pattern"))



The assignment requires the answers to the following questions

1. Is there any association between job and default?
2. Is there a significant difference in the duration of last call between people having housing loan or not?
3. Is there is any association between the consumer price index and consumer?
4. Is the employment variation rate consistent across job types?
5. Is the employment variation rate same across education?
6. Which group is more confident?

ASSOCIATION BETWEEN JOB AND DEFAULT

> JobDefault <- table(bank$job, bank$default)

> addmargins(JobDefault)

no yes Sum

admin. 472 6 478

blue-collar 932 14 946

entrepreneur 161 7 168

housemaid 110 2 112

management 955 14 969

retired 227 3 230

self-employed 179 4 183

services 410 7 417

student 83 1 84

technician 753 15 768

unemployed 125 3 128

unknown 38 0 38

Sum 4445 76 4521

> chisq.test(JobDefault)

Pearson's Chi-squared test

data: JobDefault

X-squared = 9.3064, df = 11, p-value = 0.5936

Warning message:

In chisq.test(JobDefault) : Chi-squared approximation may be incorrect

The test p-value clearly shows that we CANNOT REJECT the NULL hypothesis that “There is no association between Job and Default. However the warning message suggests that chi square test might be inappropriate, due to some of the row-column counts are less than 5 in the observed contingency table. Assuming the same would prevail in the expected contingency table such rows are deleted and retested again to see whether there is an association between job and default. If we go by the literal meaning of job as employment under someone then retired, self-employed, student, unemployed, and unknown rows can be removed to observe the difference in the chi-square test significance. This is attempted as follows.

> JobDefault <- JobDefault[-c(6,7,9,11,12),]

> JobDefault

no yes

admin. 472 6

blue-collar 932 14

entrepreneur 161 7

housemaid 110 2

management 955 14

services 410 7

technician 753 15

> chisq.test(JobDefault)

Pearson's Chi-squared test

data: JobDefault

X-squared = 7.6974, df = 6, p-value = 0.2611

Warning message:

In chisq.test(JobDefault) : Chi-squared approximation may be incorrect

So the test again CANNOT REJECT the NULL Hypothesis that “There is no association between the Job and Default”.

DIFFERENCE IN DURATION OF LAST CALL BETWEEN PEOPLE WITH HOUSING LOAN AND NOT

> library(sqldf)

> library(dplyr)

> hyesno <- bank.additional %>% filter (housing == "yes"|housing == "no") %>% select(housing,duration)

> head(hyesno)

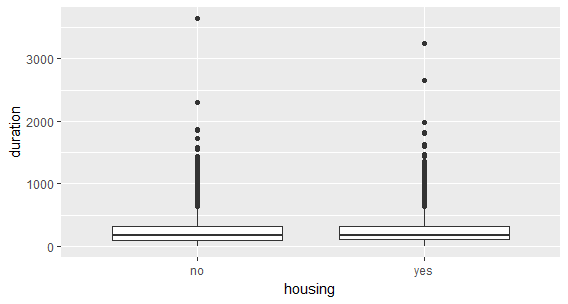
> summary(hyesno %>% filter(housing == "yes") %>% .$duration)

> summary(hyesno %>% filter(housing == "no") %>% .$duration)

> summary(hyesno$duration)

> library(ggplot2)

> ggplot(hyesno,aes(housing, duration)) + geom\_boxplot()

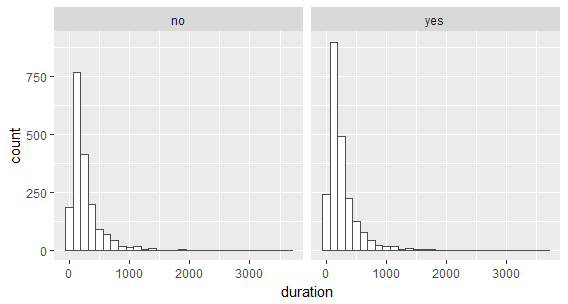


> p1 <- ggplot(hyesno, aes(duration)) +

geom\_histogram(fill = "white", color = "grey30") +

facet\_wrap(~ housing)

> p1



> t.test(duration~housing, data=hyesno)

Welch Two Sample t-test

data: duration by housing

t = 0.78474, df = 3826.1, p-value = 0.4327

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-9.57093 22.34590

sample estimates:

mean in group no mean in group yes

260.5856 254.1982

The above test shows clearly that the p-value is greater than 0.05 or 5% significance level, hence WE CANNOT REJECT THE NULL hypothesis that “There is significant difference between the last call duration of people with housing loan and without housing loan”

> sum(is.na(hyesno$duration))

0

> wilcox.test(duration~housing, data=hyesno)

Wilcoxon rank sum test with continuity correction

data: duration by housing

W = 2007300, p-value = 0.8403

alternative hypothesis: true location shift is not equal to 0

Since the data on duration grouped by housing loan shows a skewed distribution and the normality assumption of t-test is not satisfied, we also perform Wilcoxon test relaxing the assumption of normality. Even in this test WE CANNOT REJECT THE NULL HYPOTHESIS “There is significant difference between the last call duration of people with housing loan and without housing loan”

ASSOCIATION BETWEEN CONSUMER PRICE INDEX AND CONSUMER CONFIDENCE INDEX

Since both the variables happen to numerical continuous variable, CORRELATION can be calculated and the association can be ascertained.

> sum(is.na(bank.additional$cons.price.idx))

[1] 0

> summary(bank.additional$cons.price.idx)

Min. 1st Qu. Median Mean 3rd Qu. Max.

92.20 93.08 93.75 93.58 93.99 94.77

> summary(bank.additional$cons.conf.idx)

Min. 1st Qu. Median Mean 3rd Qu. Max.

-50.8 -42.7 -41.8 -40.5 -36.4 -26.9

> library(Hmisc)

> rcorr(bank.additional$cons.conf.idx, bank.additional$cons.price.idx)

x y

x 1.00 0.05

y 0.05 1.00

n= 4119

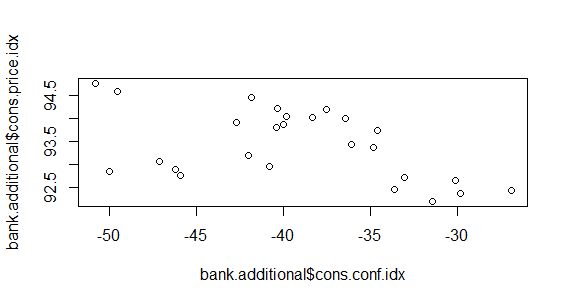
P

x y

x 0.0033

y 0.0033

> plot(bank.additional$cons.conf.idx,bank.additional$cons.price.idx)



The pearson’s correlation coefficient and the scatter plot clearly shows that there is very weak association between consumer price index and consumer confidence index. The correlation coefficient is 0.05 with p value of 0.0033

CONSISTENCY OF EMPLOYMENT VARIATION RATE ACROSS JOB TYPES

This requires an ANOVA test, since the number of groups are more than two hence the relevant data are extracted as follows

> jobemp <- bank.additional %>% select(job,emp.var.rate)

> jobemp$job <- gsub('blue-collar','bluecollar',jobemp$job)

> jobemp$job <- gsub('self-employed','selfemployed',jobemp$job)

> jobemp$job <- as.factor(jobemp$job)

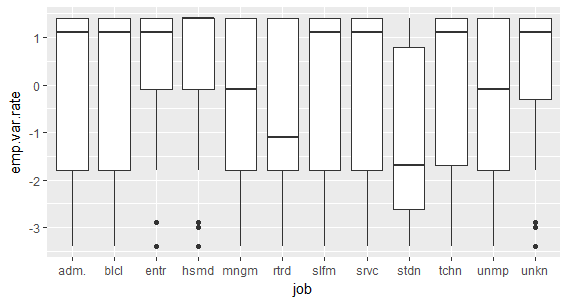
> str(jobemp)

'data.frame': 4119 obs. of 2 variables:

$ job : Factor w/ 12 levels "admin.","bluecollar",..: 2 8 8 8 1 8 1 3 8 2 ...

$ emp.var.rate: num -1.8 1.1 1.4 1.4 -0.1 -1.1 -1.1 -0.1 -0.1 1.1 ...

> ggplot(jobemp,aes(job,emp.var.rate)) + geom\_boxplot() + scale\_x\_discrete(labels =abbreviate)



The above box plot clearly shows that the median employment variation rate is definitely not consistent across various job types. If we consider the entire dataset with all the various categories of job types considered in the survey, retired, student, unemployed, and unknown seem to be meaningless to be compared or even tested for a hypothesis since they are not essentially employed currently. So we perform the analysis with them and without them.

Entire Data Set

> result <- aov(emp.var.rate~job,data=jobemp)

> summary(result)

Df Sum Sq Mean Sq F value Pr(>F)

job 11 280 25.432 10.68 <2e-16 \*\*\*

Residuals 4107 9782 2.382

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

The p-value clearly has a value almost at 0% significance level or 100% confidence level, and the F statistic conveys that the mean employment variation rate across the various job types are not consistent .

Typical ANOVA result do not convey which two group differ the most and where the inconsistency among the various job types. Hence the TukeyHSD is performed to identify such pairs. The following result again confirms the fact that student, retired, self-employed, unemployed are creating the distortion.

> TukeyHSD(result)

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = emp.var.rate ~ job, data = jobemp)

$`job`

diff lwr upr p adj

retired-admin. -0.641582933 -1.06416616 -0.21899971 0.0000460

student-admin. -1.254145377 -1.83356792 -0.67472283 0.0000000

retired-bluecollar -0.815640844 -1.24251354 -0.38876814 0.0000001

student-bluecollar -1.428203289 -2.01076161 -0.84564496 0.0000000

retired-entrepreneur -0.843633995 -1.41414474 -0.27312325 0.0000877

student-entrepreneur -1.456196440 -2.15091683 -0.76147605 0.0000000

retired-housemaid -1.006484118 -1.62690737 -0.38606087 0.0000078

student-housemaid -1.619046563 -2.35530664 -0.88278648 0.0000000

retired-management -0.530429124 -1.01210556 -0.04875269 0.0167638

student-management -1.142991569 -1.76682335 -0.51915979 0.0000002

selfemployed-retired 0.666284004 0.10630340 1.22626461 0.0057286

services-retired 0.647804960 0.18067309 1.11493683 0.0003691

technician-retired 0.828890381 0.39269481 1.26508595 0.0000001

student-selfemployed -1.278846449 -1.96494569 -0.59274720 0.0000001

student-services -1.260367405 -1.87303867 -0.64769614 0.0000000

technician-student 1.441452826 0.85202897 2.03087668 0.0000000

unemployed-student 1.010700945 0.27585864 1.74554324 0.0004379

unknown-student 1.497811132 0.51620477 2.47941749 0.0000405

If we remove the respective irrelevant variables then the following results emerge.

> jobemp2 <- jobemp %>% filter(job != "student" & job != "selfemployed" & job != "retired" & job != "unemployed") %>% select(job,emp.var.rate)

> table(jobemp2$job)

admin. bluecollar entrepreneur housemaid management

1012 884 148 110 324

retired selfemployed services student technician

0 0 393 0 691

unemployed unknown

0 39

> result <- aov(emp.var.rate~job,data=jobemp2);

> summary(result)

Df Sum Sq Mean Sq F value Pr(>F)

job 7 48 6.848 2.951 0.00438 \*\*

Residuals 3593 8337 2.320

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

The results show that the employment variation rate is NOT CONSISTENT ACROSS JOB TYPES.

> TukeyHSD(result)

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = emp.var.rate ~ job, data = jobemp2)

$`job`

diff lwr upr p adj

bluecollar-admin. 0.174057911 -0.03861160 0.38672742 0.2033306

entrepreneur-admin. 0.202051063 -0.20449554 0.60859767 0.8038459

housemaid-admin. 0.364901186 -0.09887910 0.82868148 0.2483630

management-admin. -0.111153809 -0.40603218 0.18372457 0.9472119

services-admin. 0.006222028 -0.26834861 0.28079266 1.0000000

technician-admin. 0.187307448 -0.04066402 0.41527892 0.1991284

unknown-admin. 0.243665755 -0.51017785 0.99750936 0.9772627

entrepreneur-bluecollar 0.027993152 -0.38229154 0.43827785 0.9999992

housemaid-bluecollar 0.190843274 -0.27621726 0.65790381 0.9202227

management-bluecollar -0.285211720 -0.58522280 0.01479936 0.0762978

services-bluecollar -0.167835884 -0.44791163 0.11223986 0.6082244

technician-bluecollar 0.013249537 -0.22132323 0.24782230 0.9999998

unknown-bluecollar 0.069607843 -0.68625827 0.82547395 0.9999935

housemaid-entrepreneur 0.162850123 -0.41869727 0.74439751 0.9901734

management-entrepreneur -0.313204872 -0.77152623 0.14511649 0.4331197

services-entrepreneur -0.195829035 -0.64135596 0.24969789 0.8863716

technician-entrepreneur -0.014743615 -0.43316470 0.40367747 1.0000000

unknown-entrepreneur 0.041614692 -0.78988119 0.87311057 0.9999999

management-housemaid -0.476054994 -0.98582993 0.03371994 0.0875169

services-housemaid -0.358679158 -0.85698254 0.13962422 0.3622924

technician-housemaid -0.177593738 -0.65181754 0.29663007 0.9489965

unknown-housemaid -0.121235431 -0.98216366 0.73969280 0.9998829

services-management 0.117375836 -0.22927580 0.46402747 0.9704722

technician-management 0.298461257 -0.01258426 0.60950678 0.0708703

unknown-management 0.354819563 -0.42816097 1.13780009 0.8690756

technician-services 0.181085420 -0.11077943 0.47295027 0.5637332

unknown-services 0.237443727 -0.53811690 1.01300435 0.9833568

unknown-technician 0.056358306 -0.70395495 0.81667156 0.9999986

Though the ANOVA results show that the employment variation rate is NOT consistent across job types, Truly speaking the inconsistency arises due to the difference between the Management professionals and blue-collared workers, technicians, and house maids.

CONSISTENCY OF EMPLOYMENT VARIATION RATE ACROSS EDUCATION

> empedu <- bank.additional %>% select(education,emp.var.rate)

> head(empedu)

education emp.var.rate

1 basic.9y -1.8

2 high.school 1.1

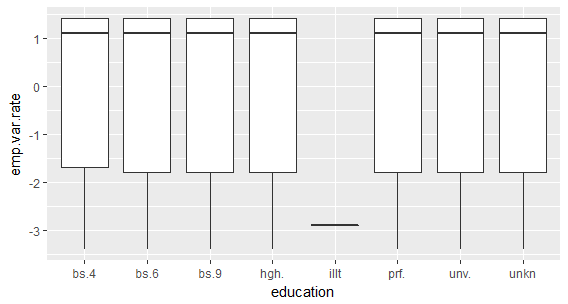
3 high.school 1.4

4 basic.9y 1.4

5 university.degree -0.1

6 university.degree -1.1

> ggplot(empedu,aes(education,emp.var.rate)) + geom\_boxplot() + scale\_x\_discrete(labels =abbreviate)



> result <- aov(emp.var.rate~education,data=empedu);

> summary(result);

Df Sum Sq Mean Sq F value Pr(>F)

education 7 67 9.523 3.917 0.000289 \*\*\*

Residuals 4111 9995 2.431

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> TukeyHSD(result)

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = emp.var.rate ~ education, data = empedu)

$`education`

upr p adj

high.school-basic.4y -0.01795116 0.0274062

university.degree-basic.4y -0.03737029 0.0126800

The ANOVA results clearly REJECT THE NULL Hypothesis that “The employment variation rate is consistent across education levels”. The Employment variation rate differs significantly between High-school and basic 4 years education levels and University.degree and basic4.years education.

CONFIDENCE OF THE GROUP BASED ON THEIR JOB TYPE

> jobconf <- bank.additional %>% select(job,cons.conf.idx);

> head(jobconf)

job cons.conf.idx

1 blue-collar -46.2

2 services -36.4

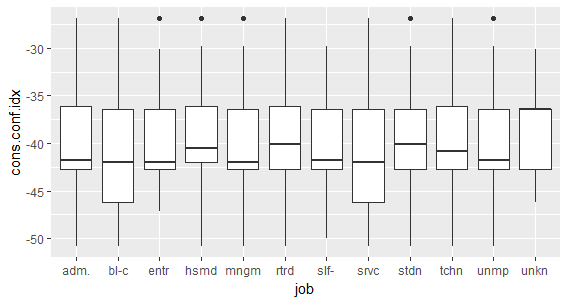
3 services -41.8

4 services -41.8

5 admin. -42.0

6 services -37.5

> ggplot(jobconf,aes(job,cons.conf.idx)) + geom\_boxplot() + scale\_x\_discrete(labels =abbreviate)



> result <- aov(cons.conf.idx~job,data=jobconf);

> summary(result);

Df Sum Sq Mean Sq F value Pr(>F)

job 11 2605 236.86 11.54 <2e-16 \*\*\*

Residuals 4107 84326 20.53

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> TukeyHSD(result)

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = cons.conf.idx ~ job, data = jobconf)

$`job`

diff lwr upr p adj

blue-collar-admin. -1.55676989 -2.23888378 -0.8746560 0.0000000

services-admin. -1.51733775 -2.39799263 -0.6366829 0.0000013

housemaid-blue-collar 2.24694776 0.74890286 3.7449927 0.0000622

management-blue-collar 0.97982655 0.01757428 1.9420788 0.0415076

retired-blue-collar 2.53715586 1.28381677 3.7904950 0.0000000

self-employed-blue-collar 1.24759384 -0.02876142 2.5239491 0.0624734

student-blue-collar 1.90563955 0.19519257 3.6160865 0.0144323

technician-blue-collar 1.67554531 0.92317918 2.4279114 0.0000000

unknown-blue-collar 2.59969834 0.17534159 5.0240551 0.0232461

housemaid-entrepreneur 1.51531941 -0.34992938 3.3805682 0.2487685

retired-entrepreneur 1.80552752 0.13045337 3.4806017 0.0218499

services-housemaid -2.20751561 -3.80576847 -0.6092628 0.0004014

retired-management 1.55732932 0.14308121 2.9715774 0.0167722

services-retired -2.49772372 -3.86926757 -1.1261799 0.0000002

student-services 1.86620741 0.06734595 3.6650689 0.0339578

technician-services 1.63611317 0.69998900 2.5722373 0.0000008

unknown-services 2.56026620 0.07274146 5.0477909 0.0370249

The ANOVA results show that there is significant difference between the groups by job type regarding their confidence about the economy reflected by their consumer confidence index. Further pairwise comparison shows that there are various pairs which show difference in their confidence. The blue-collar and the services sector groups are definitely forming the pairs which is creating the difference. That tells that they should have some opinion which is very different from others. The group wise mean analysis shown below suggests that the UNKNOWN AND RETIRED, are the most confident groups followed by the HOUSEMAID group. As the negative confidence index decreases in value (i.e. it reaches towards 0 and positive) it means that the confidence is increasing.

> group\_by(jobconf,job) %>% summarise(conf = mean(cons.conf.idx))

# A tibble: 12 x 2

job conf

*<chr>* *<dbl>*

1 admin. -40.0

2 blue-collar -41.6

3 entrepreneur -40.8

4 housemaid -39.3

5 management -40.6

6 retired -39.0

7 self-employed -40.3

8 services -41.5

9 student -39.7

10 technician -39.9

11 unemployed -40.5

12 unknown -39.0