Housing data analysis - Women Who Code workshop

Yournamehere

13/03/2018

Basics —-

```
# Arithmetic operations: R is a calculator
1 + 1
```

```
## [1] 2
```

```
10**2
```

```
## [1] 100
```

```
TRUE + TRUE
```

```
## [1] 2
```

Gotcha: R is not python (works only for numeric):

```
#"1" + "1"
```

Variables:

```
x <- 1 # the R "way" to do it ...
y = 2
print(x)
```

```
## [1] 1
```

```
print(y)
```

```
## [1] 2
```

More about why

Data types: vectors

- "character" (aka string), numeric and boolean

```
a <- c(1,2,5.3,6,-2,4) # numeric vector
b <- c("one","two","three") # character vector
c <- c(TRUE,TRUE,TRUE,FALSE,TRUE,FALSE) #logical vector
print(paste(c(class(a), class(b), class(c))))
```

```
## [1] "numeric"   "character" "logical"
```

Getting help and the combine function:

Data types: data frames (the main class for data analysis) —

Load the data in from csv.

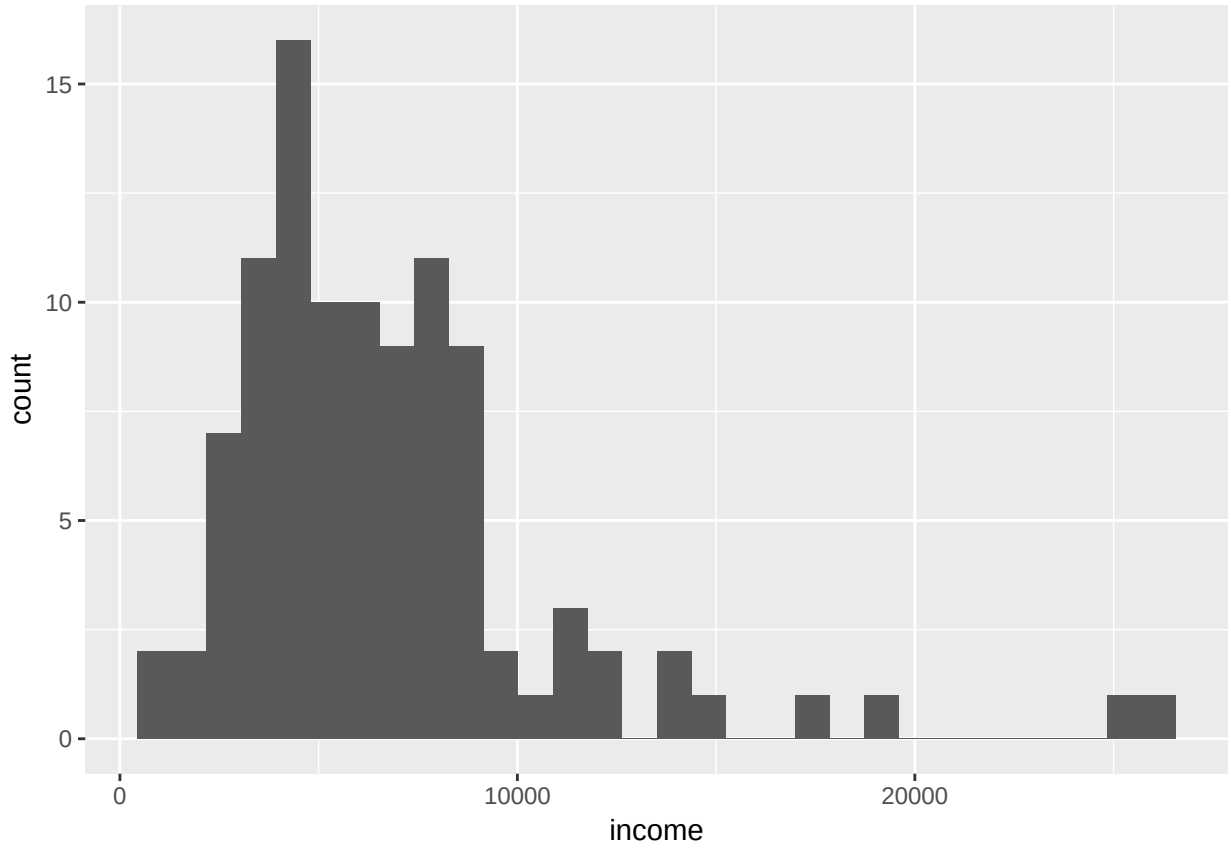
What features are there in the data? What are the dimensions of the data? What are the column headers? Use the summary() and str() functions to explore...

What does the distribution of sale price look like?

Is the sale price (the variable we're interested in prediting) normally distributed? Find its mean, standard deviation, and plot a histogram of the distribution using ggplot2.

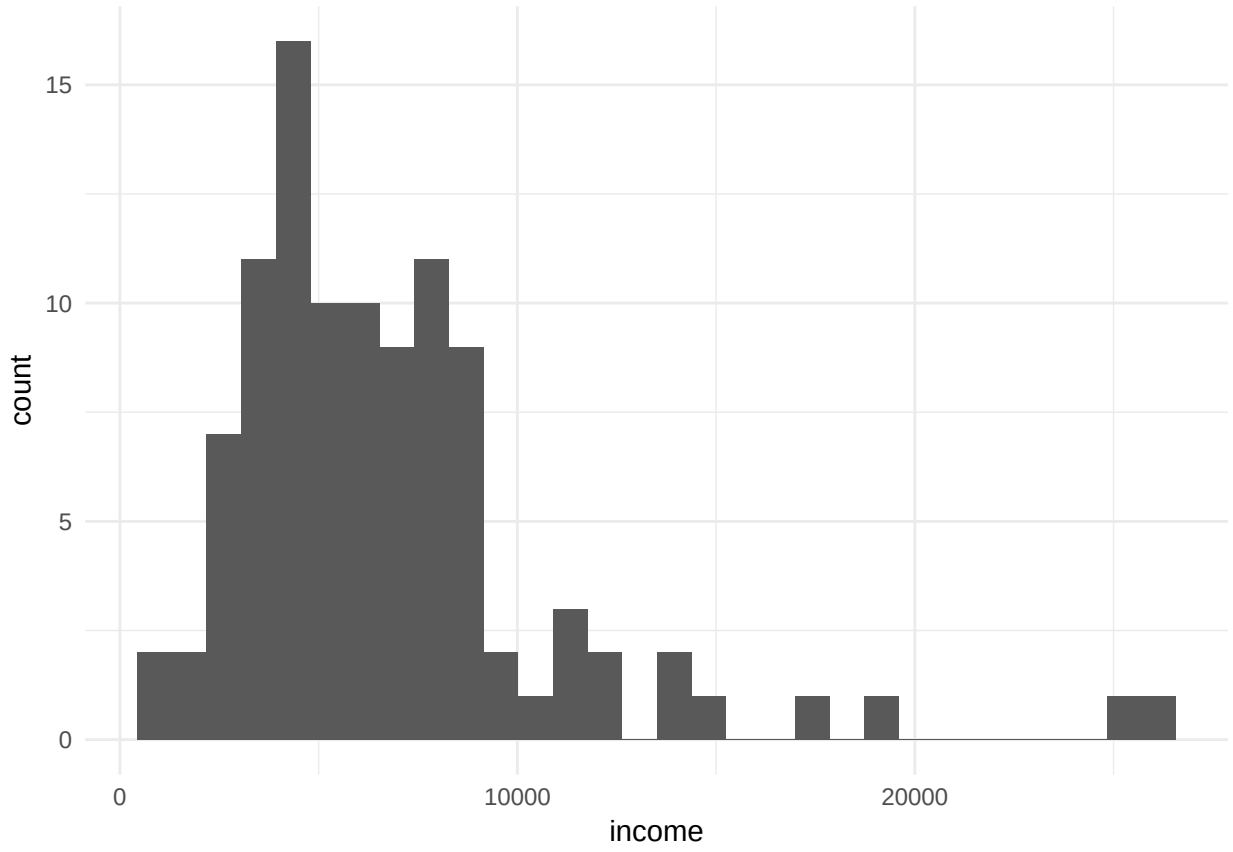
```
myprestige %>% ggplot(aes(x = income)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

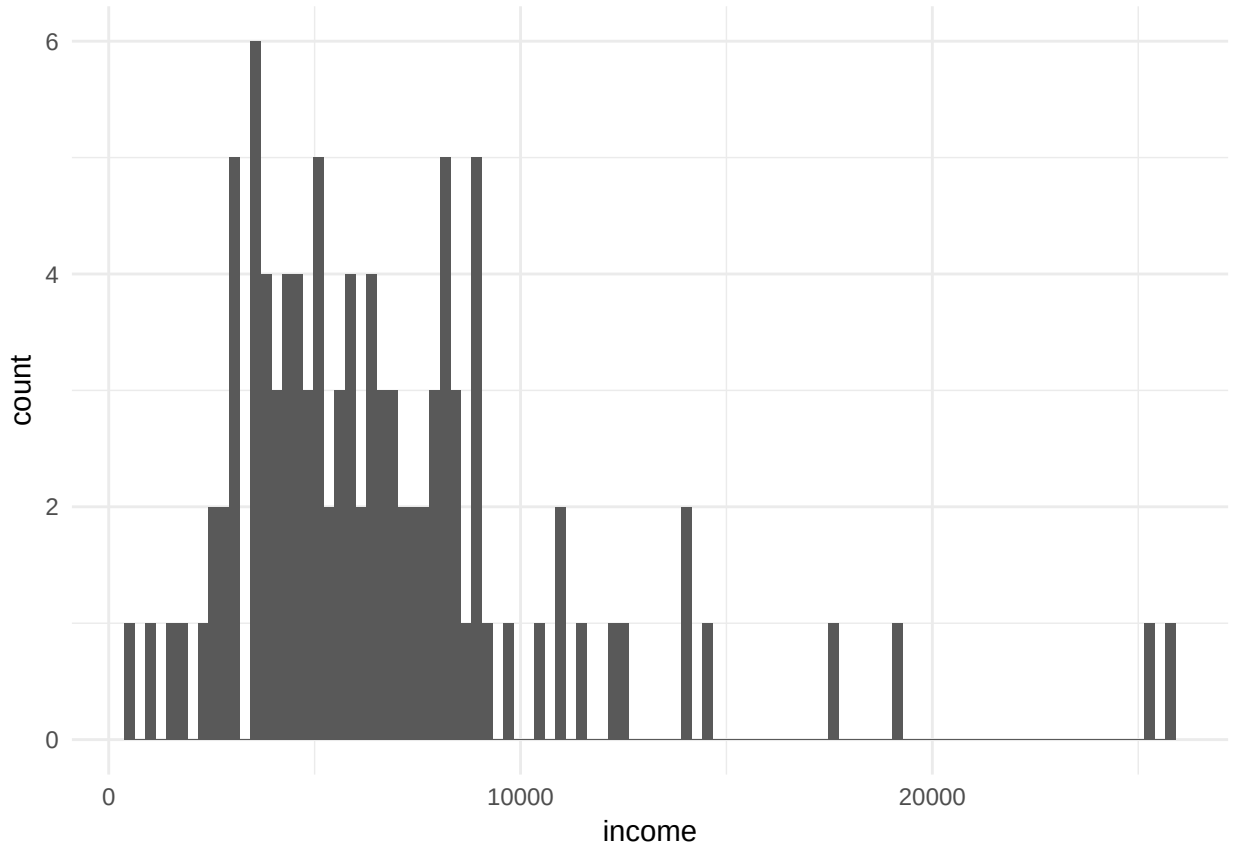



```
myprestige %>% ggplot(aes(x = income)) + geom_histogram() + theme_minimal()
```

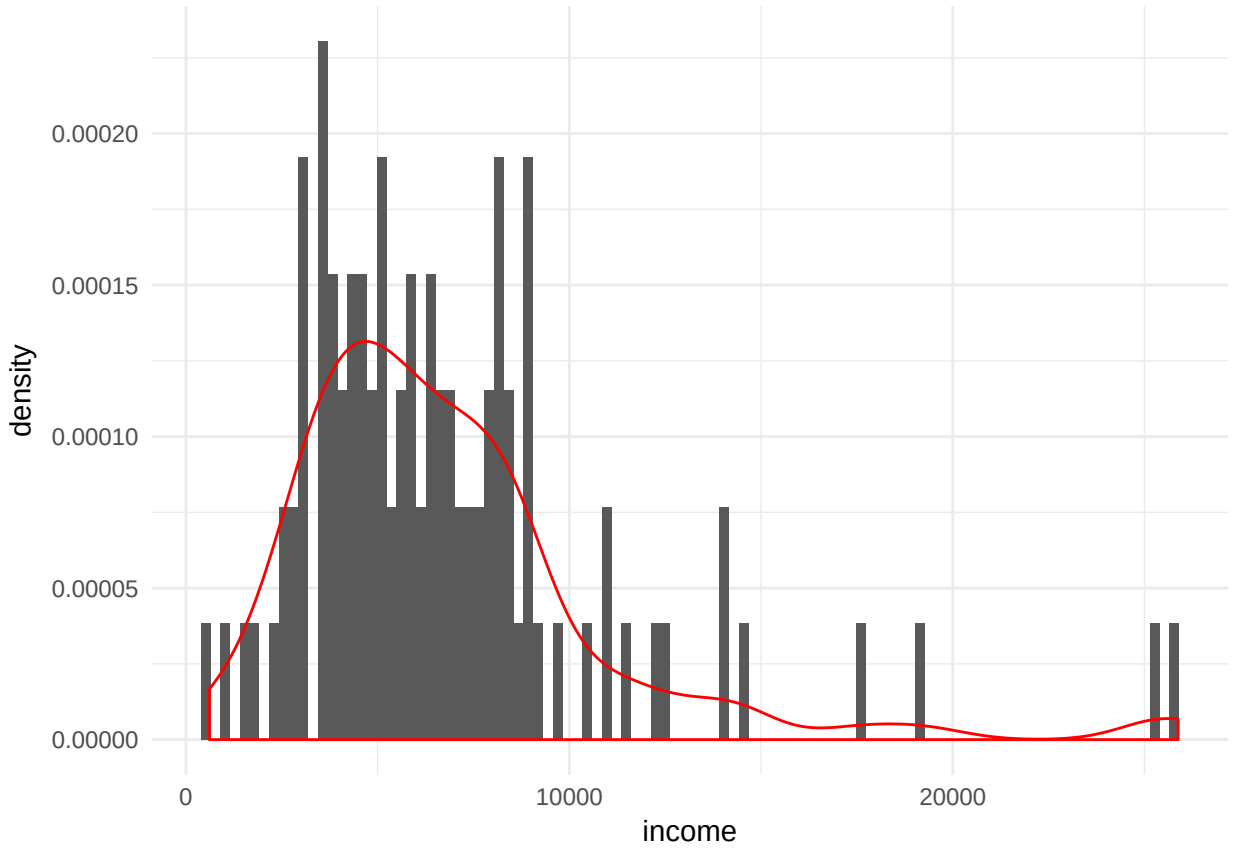
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



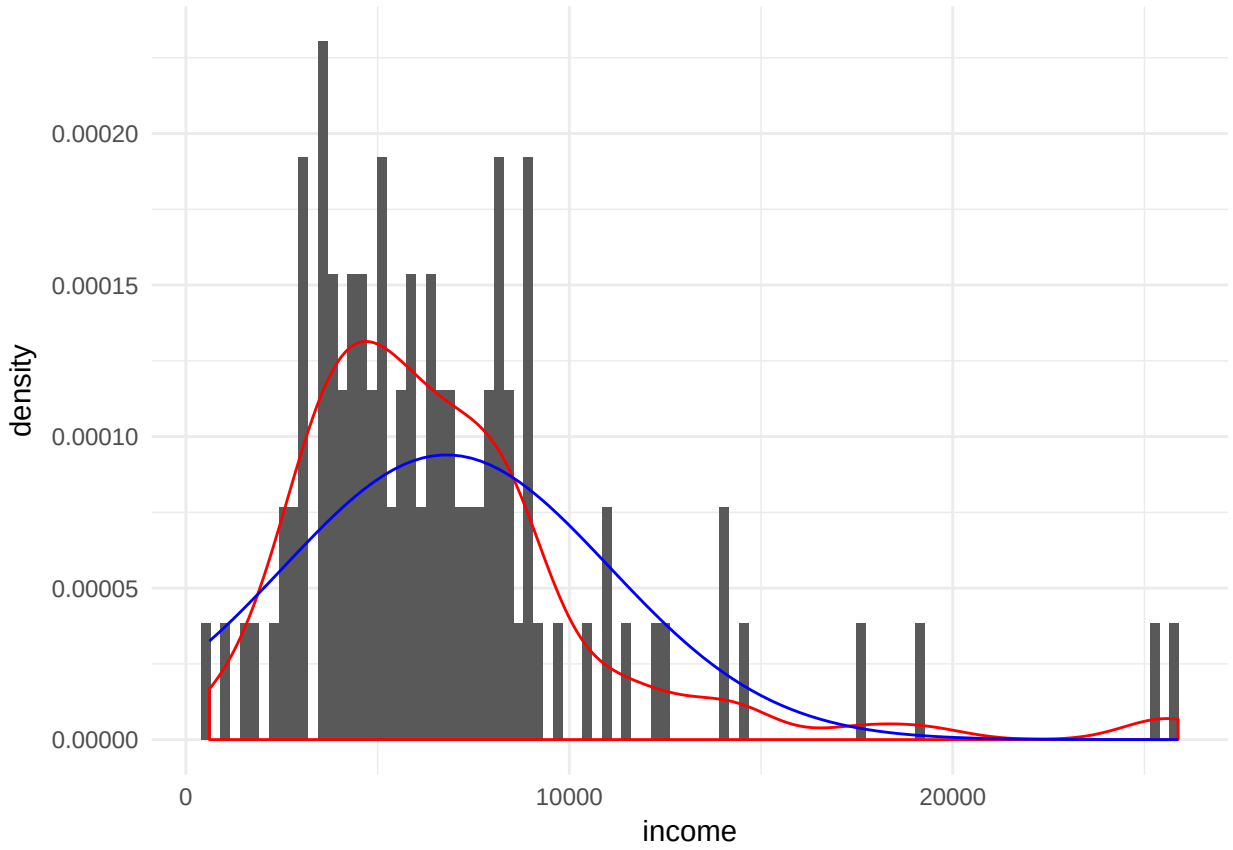

```
myprestige %>% ggplot(aes(x = income)) + geom_histogram(bins = 100) + theme_minimal()
```




```
myprestige %>% ggplot(aes(x = income)) + geom_histogram(bins = 100, aes(y =..density..)) + theme_minimal
```

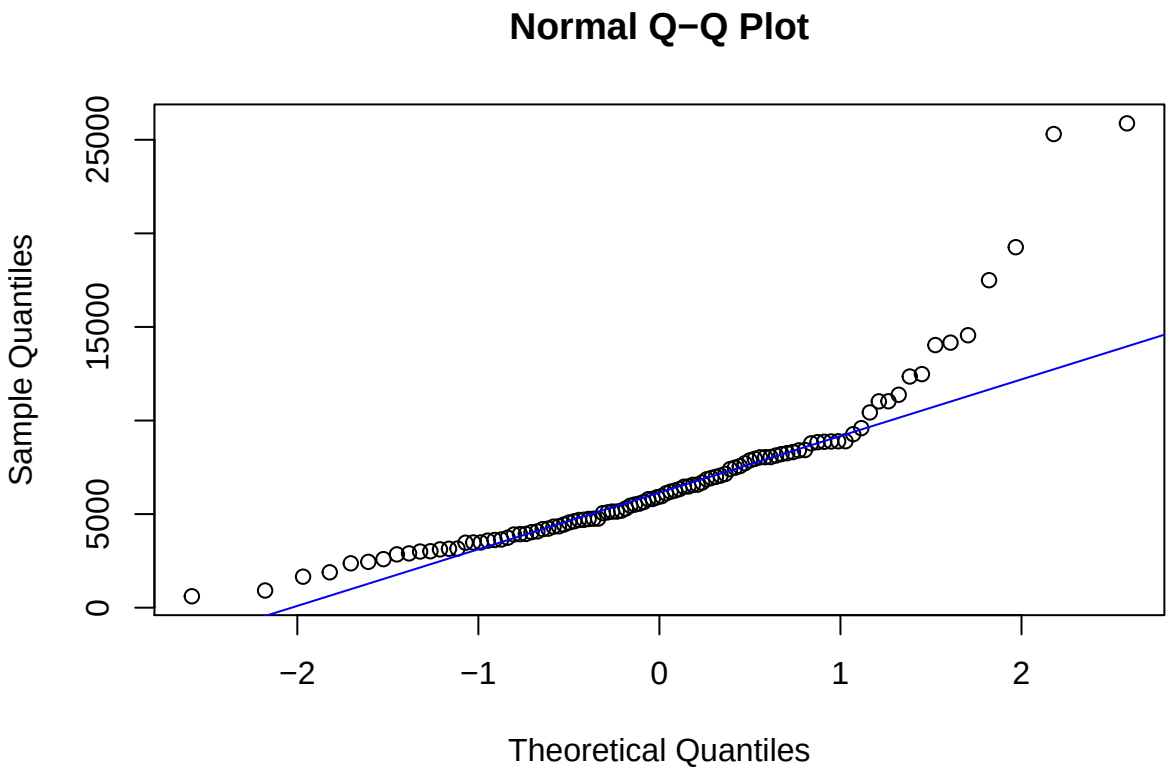



```
myprestige %>% ggplot(aes(x = income)) + geom_histogram(bins = 100, aes(y =..density..)) + theme_minimal
```



Plot a quantile-quantile plot (QQ plot) to "assess" normality. This plot compared the data we have (Sample Quantiles) with a theoretical sample coming from a normal distribution. Each point (x, y) corresponds to one of the quantiles of the second distribution (x-coordinate, theoretical) plotted against the same quantile of the first distribution (y-coordinate, our data). Thus the line is a parametric curve with the parameter which is the number of the interval for the quantile.

```
qqnorm(myprestige$income)
qqline(myprestige$income, col = "blue")
```

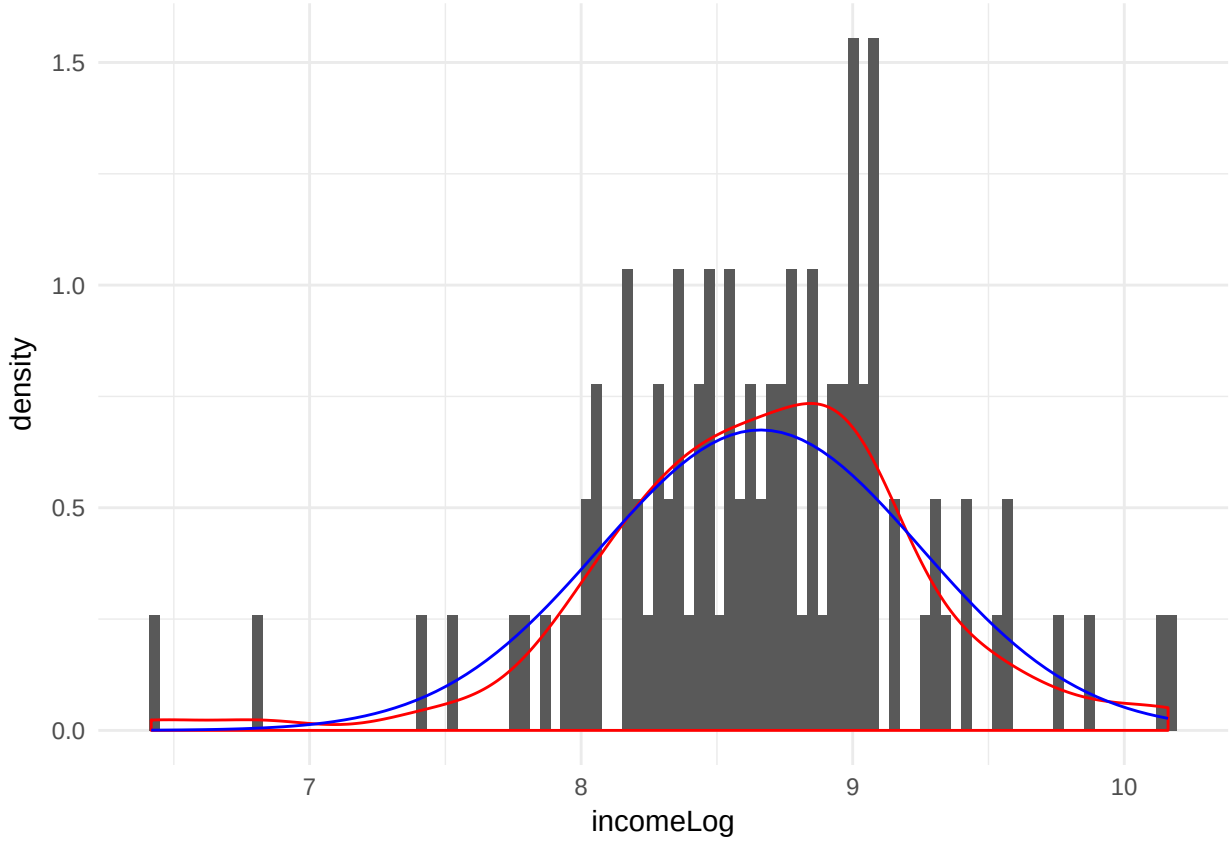


A standard way of transforming the data to be better approximated by a normal distribution is by using the log-transform?

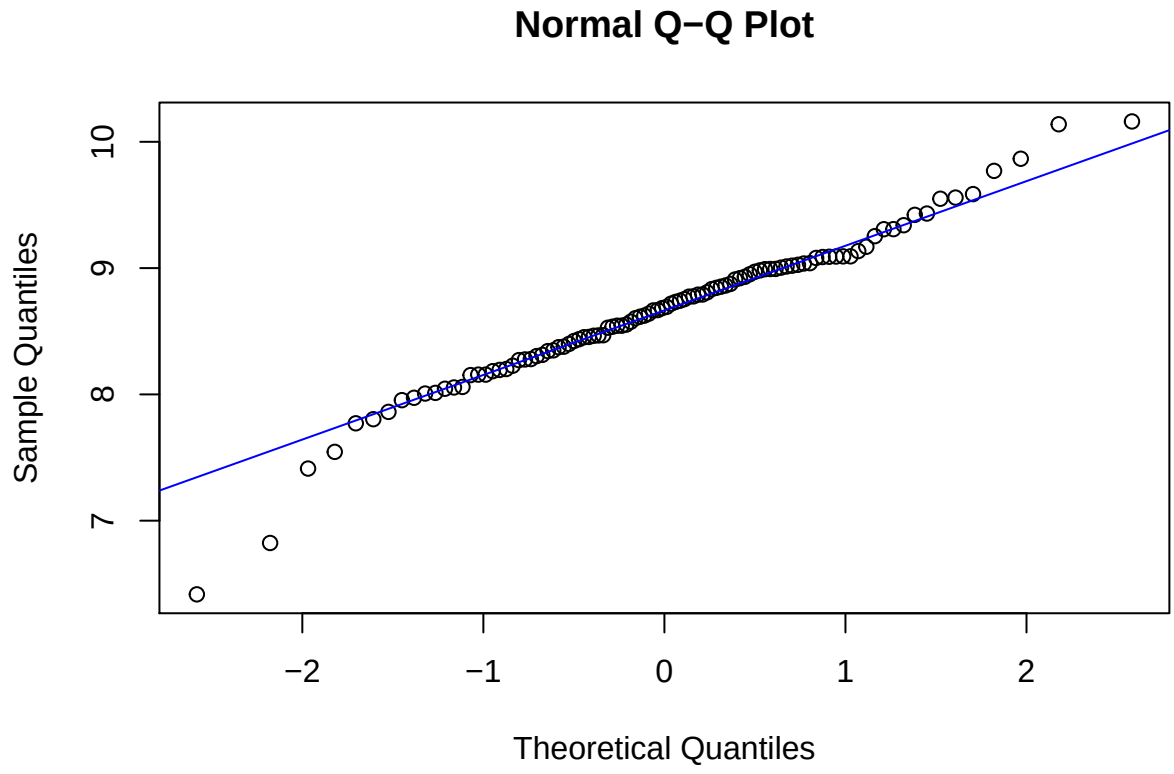
Carry out this transformation and use a histogram and QQ plot to see whether it works...

```
myprestige <- myprestige %>%
  mutate(incomeLog =  log(income + 1)) %>%
  mutate(income =  NULL)

# plot
myprestige %>% ggplot(aes(x = incomeLog)) + geom_histogram(bins = 100, aes(y =..density..)) + theme_mini
```




```
qqnorm(myprestige$incomeLog)
qqline(myprestige$incomeLog, col = "blue")
```



Missing data

What happens if we only use complete data? How much data is missing?

Topics used here (but not explored): Subsetting data frames The apply family

```
dim(myprestige)
```

```
## [1] 102   7
```

```
dim(myprestige[complete.cases(myprestige), ])
```

```
## [1] 98  7
```

```
colSums(sapply(myprestige, is.na)) [colSums(sapply(myprestige, is.na)) > 0]
```

```
## type
##    4
```

We need to combine the datasets for imputation, so that we don't have NAs in the test data as well!

```
# this is not applicable to myprestige but need to show rbind here
myprestige2 <- rbind(myprestige, myprestige)
```

How do we impute the missing data?

```
table(myprestige2$type, useNA = "always")
```

```
##
##   bc prof   wc <NA>
##   88   62   46    8
```

Read the metadata file and see that many of the NAs should be recoded as None since these features are lacking in the house.

(for the demo dataset we'll just add a factor of other)

```
myprestige2 <- myprestige2 %>% mutate(type = fct_explicit_na(type, na_level = "Ot"))
```

For the GarageYrBlt set to zero.

```
# no demo here
myprestige2 <-  myprestige2 %>% replace_na(list(income = 0)) # if there were a zero income
```

Lot frontage - set as median for the neighborhood.

```
# as a hint - use group_by() and mutate()
# will also need ifelse() function
myprestige2 %>% group_by(type) %>% summarise(incomeM = median(incomeLog))
```

```
## # A tibble: 4 x 2
##   type  incomeM
##   <fct>   <dbl>
## 1 bc       8.56
## 2 prof     9.09
## 3 wc       8.46
## 4 Ot       7.51
```

Now split data again

```
# no demo here
```

Basic exploratory data analysis of training data

How does the sale price depend on living area: X1stFlrSF, X2ndFlrSF, TotalBsmtSF? Create a variable TotalSqFt which is a combination of these 3. Does it better predict the house price?

```
# no demo here, just ggplot pipes
# then make dummy variable
myprestige2$nonsense <- myprestige2$women + myprestige2$education
```

Identify and remove outliers with a high total square foot, but low price

Useful reference for dplyr

```
myprestige2 %>% arrange(desc(education))
```

```
##     education women prestige census type                      job
## 1       15.97 19.59     84.6   2711 prof       university.teachers
## 2       15.97 19.59     84.6   2711 prof       university.teachers
## 3       15.96 10.56     87.2   3111 prof                physicians
## 4       15.96 10.56     87.2   3111 prof                physicians
## 5       15.94  4.32     66.7   3115 prof             veterinarians
## 6       15.94  4.32     66.7   3115 prof             veterinarians
## 7       15.77  5.13     82.3   2343 prof                   lawyers
## 8       15.77  5.13     82.3   2343 prof                   lawyers
## 9       15.64  5.13     77.6   2113 prof                physicists
## 10      15.64  5.13     77.6   2113 prof                physicists
## 11      15.44  2.69     78.1   2141 prof                architects
## 12      15.44  2.69     78.1   2141 prof                architects
## 13      15.22 34.89     58.3   2391 prof    vocational.counsellors
## 14      15.22 34.89     58.3   2391 prof    vocational.counsellors
## 15      15.21 24.71     69.3   3151 prof               pharmacists
## 16      15.21 24.71     69.3   3151 prof               pharmacists
## 17      15.09 25.65     72.6   2133 prof                biologists
## 18      15.09 25.65     72.6   2133 prof                biologists
## 19      15.08 46.80     66.1   2733 prof secondary.school.teachers
## 20      15.08 46.80     66.1   2733 prof secondary.school.teachers
## 21      14.71  6.91     68.4   3117 prof  osteopaths.chiropractors
## 22      14.71  6.91     68.4   3117 prof  osteopaths.chiropractors
## 23      14.64  0.94     68.8   2153 prof          mining.engineers
## 24      14.64  0.94     68.8   2153 prof          mining.engineers
## 25      14.62 11.68     73.5   2111 prof                  chemists
## 26      14.62 11.68     73.5   2111 prof                  chemists
## 27      14.52  1.03     73.1   2143 prof           civil.engineers
## 28      14.52  1.03     73.1   2143 prof           civil.engineers
## 29      14.50  4.14     72.8   2511 prof                 ministers
## 30      14.50  4.14     72.8   2511 prof                 ministers
## 31      14.44 57.31     62.2   2311 prof                economists
## 32      14.44 57.31     62.2   2311 prof                economists
## 33      14.36 48.28     74.9   2315 prof             psychologists
## 34      14.36 48.28     74.9   2315 prof             psychologists
```

```
## 35      14.21 54.77     55.1   2331 prof            social.workers
## 36      14.21 54.77     55.1   2331 prof            social.workers
## 37      14.15 77.10     58.1   2351 prof                librarians
## 38      14.15 77.10     58.1   2351 prof                librarians
## 39      13.83 15.33     53.8   2183 prof       computer.programers
## 40      13.83 15.33     53.8   2183 prof       computer.programers
## 41      13.62 83.78     59.6   2731 prof   primary.school.teachers
## 42      13.62 82.66     72.1   3137 prof          physio.therapsts
## 43      13.62 83.78     59.6   2731 prof   primary.school.teachers
## 44      13.62 82.66     72.1   3137 prof          physio.therapsts
## 45      13.11 11.16     68.8   1113 prof        gov.administrators
## 46      13.11 11.16     68.8   1113 prof        gov.administrators
## 47      12.79 76.04     67.5   3156   wc       medical.technicians
## 48      12.79 76.04     67.5   3156   wc       medical.technicians
## 49      12.77 15.70     63.4   1171 prof               accountants
## 50      12.77 15.70     63.4   1171 prof               accountants
## 51      12.71 11.15     57.6   3337   wc       radio.tv.announcers
## 52      12.71 11.15     57.6   3337   wc       radio.tv.announcers
## 53      12.46 96.12     64.7   3131 prof                   nurses
## 54      12.46 96.12     64.7   3131 prof                   nurses
## 55      12.39  1.91     62.0   2161 prof                surveyors
## 56      12.39  1.91     62.0   2161 prof                surveyors
## 57      12.30  7.83     60.0   2163 prof               draughtsmen
## 58      12.30  7.83     60.0   2163 prof               draughtsmen
## 59      12.27  0.58     66.1   9111 prof                   pilots
## 60      12.27  0.58     66.1   9111 prof                   pilots
## 61      12.26  4.02     69.1   1130 prof          general.managers
## 62      12.26  4.02     69.1   1130 prof          general.managers
## 63      12.09 83.19     32.7   4161   wc               file.clerks
## 64      12.09 83.19     32.7   4161   wc               file.clerks
## 65      11.60 13.09     47.3   5171   wc          insurance.agents
## 66      11.60 13.09     47.3   5171   wc          insurance.agents
## 67      11.59 97.51     46.0   4111   wc               secretaries
## 68      11.59 97.51     46.0   4111   wc               secretaries
## 69      11.49 95.97     41.9   4113   wc                   typists
## 70      11.49 95.97     41.9   4113   wc                   typists
## 71      11.44  8.13     54.1   3373   Ot                  athletes
## 72      11.44  8.13     54.1   3373   Ot                  athletes
## 73      11.43 39.17     35.7   4193   wc             travel.clerks
## 74      11.43 39.17     35.7   4193   wc             travel.clerks
## 75      11.42  9.11     56.8   1175 prof       purchasing.officers
## 76      11.42  9.11     56.8   1175 prof       purchasing.officers
## 77      11.36 75.92     47.7   4143   wc        computer.operators
## 78      11.36 75.92     47.7   4143   wc        computer.operators
## 79      11.32 68.24     49.4   4131   wc               bookkeepers
## 80      11.32 68.24     49.4   4131   wc               bookkeepers
## 81      11.20 47.06     29.4   4191   wc                collectors
## 82      11.20 47.06     29.4   4191   wc                collectors
## 83      11.13 56.10     51.1   4192   wc           claim.adjustors
## 84      11.13  3.16     40.2   5133   wc     commercial.travellers
## 85      11.13 56.10     51.1   4192   wc           claim.adjustors
## 86      11.13  3.16     40.2   5133   wc     commercial.travellers
## 87      11.09 21.03     57.2   3314 prof        commercial.artists
## 88      11.09 24.44     47.1   5172   wc      real.estate.salesmen
```

```
## 89      11.09 21.03     57.2   3314 prof        commercial.artists
## 90      11.09 24.44     47.1   5172   wc      real.estate.salesmen
## 91      11.04 92.86     38.7   4171   wc              receptionsts
## 92      11.04 92.86     38.7   4171   wc              receptionsts
## 93      11.03 23.88     51.1   5191   wc                    buyers
## 94      11.03 23.88     51.1   5191   wc                    buyers
## 95      11.00 63.23     35.6   4197   wc             office.clerks
## 96      11.00 63.23     35.6   4197   wc             office.clerks
## 97      10.93  1.65     51.6   6112   bc                 policemen
## 98      10.93  1.65     51.6   6112   bc                 policemen
## 99      10.64 91.76     42.3   4133   wc          tellers.cashiers
## 100     10.64 91.76     42.3   4133   wc          tellers.cashiers
## 101     10.57  6.01     54.9   6141   bc         funeral.directors
## 102     10.57  6.01     54.9   6141   bc         funeral.directors
## 103     10.51 96.14     38.1   4175   wc       telephone.operators
## 104     10.51 96.14     38.1   4175   wc       telephone.operators
## 105     10.29  2.92     37.2   8537   bc        radio.tv.repairmen
## 106     10.29  2.92     37.2   8537   bc        radio.tv.repairmen
## 107     10.10  0.78     50.3   8582   bc        aircraft.repairmen
## 108     10.10  0.78     50.3   8582   bc        aircraft.repairmen
## 109     10.09  1.50     42.5   8311   bc           tool.die.makers
## 110     10.09  1.50     42.5   8311   bc           tool.die.makers
## 111     10.07 52.27     37.2   4173   wc             postal.clerks
## 112     10.07 52.27     37.2   4173   wc             postal.clerks
## 113     10.05 67.82     26.5   5137   wc              sales.clerks
## 114     10.05 67.82     26.5   5137   wc              sales.clerks
## 115     10.00 13.58     42.2   9511   bc               typesetters
## 116     10.00 13.58     42.2   9511   bc               typesetters
## 117      9.93  3.69     23.3   5145   bc service.station.attendant
## 118      9.93  0.99     50.2   8733   bc              electricians
## 119      9.93  3.69     23.3   5145   bc service.station.attendant
## 120      9.93  0.99     50.2   8733   bc              electricians
## 121      9.84 17.04     41.5   5130   wc         sales.supervisors
## 122      9.84 17.04     41.5   5130   wc         sales.supervisors
## 123      9.62  7.00     14.8   5143   Ot                 newsboys
## 124      9.62  7.00     14.8   5143   Ot                 newsboys
## 125      9.47  0.00     43.5   6111   bc             firefighters
## 126      9.47  0.00     43.5   6111   bc             firefighters
## 127      9.46 96.53     25.9   6147   Ot              babysitters
## 128      9.46 96.53     25.9   6147   Ot              babysitters
## 129      9.45 76.14     34.9   3135   bc             nursing.aides
## 130      9.45 76.14     34.9   3135   bc             nursing.aides
## 131      9.22  7.62     36.1   4172   wc             mail.carriers
## 132      9.22  7.62     36.1   4172   wc             mail.carriers
## 133      9.17 11.37     30.9   4153   wc           shipping.clerks
## 134      9.17 11.37     30.9   4153   wc           shipping.clerks
## 135      9.05  1.34     40.9   8731   bc        electrical.linemen
## 136      9.05  1.34     40.9   8731   bc        electrical.linemen
## 137      8.88  0.00     35.3   7711   bc      rotary.well.drillers
## 138      8.88  0.00     35.3   7711   bc      rotary.well.drillers
## 139      8.81  4.28     44.2   8313   bc                machinists
## 140      8.81  4.28     44.2   8313   bc                machinists
## 141      8.78  5.78     43.7   8515   bc          aircraft.workers
## 142      8.78  5.78     43.7   8515   bc          aircraft.workers
```

```
## 143      8.76 74.54     50.8   8534   bc        electronic.workers
## 144      8.76 74.54     50.8   8534   bc        electronic.workers
## 145      8.60 27.75     21.5   7182   bc              farm.workers
## 146      8.60 27.75     21.5   7182   bc              farm.workers
## 147      8.55 70.87     35.2   9517   bc               bookbinders
## 148      8.55 70.87     35.2   9517   bc               bookbinders
## 149      8.50 15.51     20.2   6123   bc                bartenders
## 150      8.50 15.51     20.2   6123   bc                bartenders
## 151      8.49  0.00     48.9   9131   bc           train.engineers
## 152      8.49  0.00     48.9   9131   bc           train.engineers
## 153      8.43 13.62     35.9   8513   bc              auto.workers
## 154      8.43 13.62     35.9   8513   bc              auto.workers
## 155      8.40  2.30     35.9   8333   bc       sheet.metal.workers
## 156      8.40  2.30     35.9   8333   bc       sheet.metal.workers
## 157      8.37  0.00     26.1   9313   bc              longshoremen
## 158      8.37  0.00     26.1   9313   bc              longshoremen
## 159      8.33  0.61     42.9   8791   bc                  plumbers
## 160      8.33  0.61     42.9   8791   bc                  plumbers
## 161      8.24  0.65     51.1   8780   bc      construction.foremen
## 162      8.24  0.65     51.1   8780   bc      construction.foremen
## 163      8.10  0.81     38.1   8581   bc            auto.repairmen
## 164      8.10  0.81     38.1   8581   bc            auto.repairmen
## 165      7.93  3.59     25.1   9173   bc              taxi.drivers
## 166      7.93  3.59     25.1   9173   bc              taxi.drivers
## 167      7.92  5.17     41.8   8335   bc                   welders
## 168      7.92  5.17     41.8   8335   bc                   welders
## 169      7.81  2.46     29.9   8785   bc            house.painters
## 170      7.81  2.46     29.9   8785   bc            house.painters
## 171      7.74 52.00     29.7   6121   bc                     cooks
## 172      7.74 52.00     29.7   6121   bc                     cooks
## 173      7.64 17.26     25.2   8215   bc            slaughterers.1
## 174      7.64 17.26     34.8   8215   bc            slaughterers.2
## 175      7.64 17.26     25.2   8215   bc            slaughterers.1
## 176      7.64 17.26     34.8   8215   bc            slaughterers.2
## 177      7.58 30.08     20.1   6193   bc        elevator.operators
## 178      7.58  9.47     35.9   9171   bc                bus.drivers
## 179      7.58 30.08     20.1   6193   bc        elevator.operators
## 180      7.58  9.47     35.9   9171   bc                bus.drivers
## 181      7.54 33.30     38.9   8213   bc                    bakers
## 182      7.54 33.30     38.9   8213   bc                    bakers
## 183      7.52  1.09     26.5   8798   bc    construction.labourers
## 184      7.52  1.09     26.5   8798   bc    construction.labourers
## 185      7.42 72.24     23.2   8221   bc                   canners
## 186      7.42 72.24     23.2   8221   bc                   canners
## 187      7.33 69.31     20.8   6162   bc                 launderers
## 188      7.33 69.31     20.8   6162   bc                 launderers
## 189      7.11 33.57     17.3   6191   bc                   janitors
## 190      7.11 33.57     17.3   6191   bc                   janitors
## 191      6.92  0.56     38.9   8781   bc                 carpenters
## 192      6.92  0.56     38.9   8781   bc                 carpenters
## 193      6.84  3.60     44.1   7112   Ot                    farmers
## 194      6.84  3.60     44.1   7112   Ot                    farmers
## 195      6.74 39.48     28.8   8278   bc         textile.labourers
## 196      6.74 39.48     28.8   8278   bc         textile.labourers
```

```
## 197      6.69 31.36     33.3   8267   bc           textile.weavers
## 198      6.69 31.36     33.3   8267   bc           textile.weavers
## 199      6.67  0.00     27.3   8715   bc        railway.sectionmen
## 200      6.67  0.00     27.3   8715   bc        railway.sectionmen
## 201      6.60  0.52     36.2   8782   bc                   masons
## 202      6.60  0.52     36.2   8782   bc                   masons
## 203      6.38 90.67     28.2   8563   bc     sewing.mach.operators
## 204      6.38 90.67     28.2   8563   bc     sewing.mach.operators
##     incomeLog nonsense
## 1    9.431963    35.56
## 2    9.431963    35.56
## 3   10.138915    26.52
## 4   10.138915    26.52
## 5    9.585965    20.26
## 6    9.585965    20.26
## 7    9.865993    20.90
## 8    9.865993    20.90
## 9    9.308465    20.77
## 10   9.308465    20.77
## 11   9.558459    18.13
## 12   9.558459    18.13
## 13   9.168893    50.11
## 14   9.168893    50.11
## 15   9.252729    39.92
## 16   9.252729    39.92
## 17   9.019059    40.74
## 18   9.019059    40.74
## 19   8.991562    61.88
## 20   8.991562    61.88
## 21   9.769899    21.62
## 22   9.769899    21.62
## 23   9.307830    15.58
## 24   9.307830    15.58
## 25   9.036463    26.30
## 26   9.036463    26.30
## 27   9.339437    15.55
## 28   9.339437    15.55
## 29   8.452548    18.64
## 30   8.452548    18.64
## 31   8.993427    71.75
## 32   8.993427    71.75
## 33   8.910046    62.64
## 34   8.910046    62.64
## 35   8.754161    68.98
## 36   8.754161    68.98
## 37   8.718173    91.25
## 38   8.718173    91.25
## 39   9.039077    29.16
## 40   9.039077    29.16
## 41   8.639234    97.40
## 42   8.535622    96.28
## 43   8.639234    97.40
## 44   8.535622    96.28
## 45   9.421573    24.27
```

```
## 46   9.421573    24.27
## 47   8.552753    88.83
## 48   8.552753    88.83
## 49   9.134754    28.47
## 50   9.134754    28.47
## 51   8.931023    23.86
## 52   8.931023    23.86
## 53   8.437067   108.58
## 54   8.437067   108.58
## 55   8.683216    14.30
## 56   8.683216    14.30
## 57   8.862200    20.13
## 58   8.862200    20.13
## 59   9.549167    12.85
## 60   9.549167    12.85
## 61  10.161226    16.28
## 62  10.161226    16.28
## 63   8.012018    95.28
## 64   8.012018    95.28
## 65   9.003562    24.69
## 66   9.003562    24.69
## 67   8.303257   109.10
## 68   8.303257   109.10
## 69   8.054840   107.46
## 70   8.054840   107.46
## 71   9.012743    19.57
## 72   9.012743    19.57
## 73   8.741935    50.60
## 74   8.741935    50.60
## 75   9.089979    20.53
## 76   9.089979    20.53
## 77   8.373554    87.28
## 78   8.373554    87.28
## 79   8.377701    79.56
## 80   8.377701    79.56
## 81   8.464214    58.26
## 82   8.464214    58.26
## 83   8.527737    67.23
## 84   9.080346    14.29
## 85   8.527737    67.23
## 86   9.080346    14.29
## 87   8.731982    32.12
## 88   8.852665    35.53
## 89   8.731982    32.12
## 90   8.852665    35.53
## 91   7.973155   103.90
## 92   7.973155   103.90
## 93   8.981807    34.91
## 94   8.981807    34.91
## 95   8.312871    74.23
## 96   8.312871    74.23
## 97   9.092907    12.58
## 98   9.092907    12.58
## 99   7.803435   102.40
```

```
## 100  7.803435   102.40
## 101  8.970813    16.58
## 102  8.970813    16.58
## 103  8.058960   106.65
## 104  8.058960   106.65
## 105  8.603371    13.21
## 106  8.603371    13.21
## 107  8.951181    10.88
## 108  8.951181    10.88
## 109  8.992682    11.59
## 110  8.992682    11.59
## 111  8.226841    62.34
## 112  8.226841    62.34
## 113  7.861342    77.87
## 114  7.861342    77.87
## 115  8.773849    23.58
## 116  8.773849    23.58
## 117  7.771067    13.62
## 118  8.874588    10.92
## 119  7.771067    13.62
## 120  8.874588    10.92
## 121  8.920389    26.88
## 122  8.920389    26.88
## 123  6.823286    16.62
## 124  6.823286    16.62
## 125  9.093357     9.47
## 126  9.093357     9.47
## 127  6.416732   105.99
## 128  6.416732   105.99
## 129  8.156510    85.59
## 130  8.156510    85.59
## 131  8.614683    16.84
## 132  8.614683    16.84
## 133  8.468423    20.54
## 134  8.468423    20.54
## 135  9.026057    10.39
## 136  9.026057    10.39
## 137  8.833608     8.88
## 138  8.833608     8.88
## 139  8.807921    13.09
## 140  8.807921    13.09
## 141  8.790878    14.56
## 142  8.790878    14.56
## 143  8.279697    83.30
## 144  8.279697    83.30
## 145  7.412764    36.35
## 146  7.412764    36.35
## 147  8.193677    79.42
## 148  8.193677    79.42
## 149  8.276649    24.01
## 150  8.276649    24.01
## 151  9.087721     8.49
## 152  9.087721     8.49
## 153  8.667680    22.05
```

```
## 154  8.667680    22.05
## 155  8.789660    10.70
## 156  8.789660    10.70
## 157  8.466742     8.37
## 158  8.466742     8.37
## 159  8.843471     8.94
## 160  8.843471     8.94
## 161  9.091669     8.89
## 162  9.091669     8.89
## 163  8.664923     8.91
## 164  8.664923     8.91
## 165  8.348775    11.52
## 166  8.348775    11.52
## 167  8.776167    13.09
## 168  8.776167    13.09
## 169  8.422883    10.27
## 170  8.422883    10.27
## 171  8.044626    59.74
## 172  8.044626    59.74
## 173  8.543835    24.90
## 174  8.543835    24.90
## 175  8.543835    24.90
## 176  8.543835    24.90
## 177  8.183956    37.66
## 178  8.623893    17.05
## 179  8.183956    37.66
## 180  8.623893    17.05
## 181  8.342840    40.84
## 182  8.342840    40.84
## 183  8.271548     8.61
## 184  8.271548     8.61
## 185  7.544861    79.66
## 186  7.544861    79.66
## 187  8.006701    76.64
## 188  8.006701    76.64
## 189  8.152774    40.68
## 190  8.152774    40.68
## 191  8.575462     7.48
## 192  8.575462     7.48
## 193  8.200837    10.44
## 194  8.200837    10.44
## 195  8.156510    46.22
## 196  8.156510    46.22
## 197  8.399310    38.05
## 198  8.399310    38.05
## 199  8.454679     6.67
## 200  8.454679     6.67
## 201  8.692826     7.12
## 202  8.692826     7.12
## 203  7.954372    97.05
## 204  7.954372    97.05
```

```
myprestige2 %>% arrange(desc(education)) %>% select(education, incomeLog, women)
```

```
##     education incomeLog women
```

```
## 1       15.97  9.431963 19.59
## 2       15.97  9.431963 19.59
## 3       15.96 10.138915 10.56
## 4       15.96 10.138915 10.56
## 5       15.94  9.585965  4.32
## 6       15.94  9.585965  4.32
## 7       15.77  9.865993  5.13
## 8       15.77  9.865993  5.13
## 9       15.64  9.308465  5.13
## 10      15.64  9.308465  5.13
## 11      15.44  9.558459  2.69
## 12      15.44  9.558459  2.69
## 13      15.22  9.168893 34.89
## 14      15.22  9.168893 34.89
## 15      15.21  9.252729 24.71
## 16      15.21  9.252729 24.71
## 17      15.09  9.019059 25.65
## 18      15.09  9.019059 25.65
## 19      15.08  8.991562 46.80
## 20      15.08  8.991562 46.80
## 21      14.71  9.769899  6.91
## 22      14.71  9.769899  6.91
## 23      14.64  9.307830  0.94
## 24      14.64  9.307830  0.94
## 25      14.62  9.036463 11.68
## 26      14.62  9.036463 11.68
## 27      14.52  9.339437  1.03
## 28      14.52  9.339437  1.03
## 29      14.50  8.452548  4.14
## 30      14.50  8.452548  4.14
## 31      14.44  8.993427 57.31
## 32      14.44  8.993427 57.31
## 33      14.36  8.910046 48.28
## 34      14.36  8.910046 48.28
## 35      14.21  8.754161 54.77
## 36      14.21  8.754161 54.77
## 37      14.15  8.718173 77.10
## 38      14.15  8.718173 77.10
## 39      13.83  9.039077 15.33
## 40      13.83  9.039077 15.33
## 41      13.62  8.639234 83.78
## 42      13.62  8.535622 82.66
## 43      13.62  8.639234 83.78
## 44      13.62  8.535622 82.66
## 45      13.11  9.421573 11.16
## 46      13.11  9.421573 11.16
## 47      12.79  8.552753 76.04
## 48      12.79  8.552753 76.04
## 49      12.77  9.134754 15.70
## 50      12.77  9.134754 15.70
## 51      12.71  8.931023 11.15
## 52      12.71  8.931023 11.15
## 53      12.46  8.437067 96.12
## 54      12.46  8.437067 96.12
```

```
## 55      12.39  8.683216  1.91
## 56      12.39  8.683216  1.91
## 57      12.30  8.862200  7.83
## 58      12.30  8.862200  7.83
## 59      12.27  9.549167  0.58
## 60      12.27  9.549167  0.58
## 61      12.26 10.161226  4.02
## 62      12.26 10.161226  4.02
## 63      12.09  8.012018 83.19
## 64      12.09  8.012018 83.19
## 65      11.60  9.003562 13.09
## 66      11.60  9.003562 13.09
## 67      11.59  8.303257 97.51
## 68      11.59  8.303257 97.51
## 69      11.49  8.054840 95.97
## 70      11.49  8.054840 95.97
## 71      11.44  9.012743  8.13
## 72      11.44  9.012743  8.13
## 73      11.43  8.741935 39.17
## 74      11.43  8.741935 39.17
## 75      11.42  9.089979  9.11
## 76      11.42  9.089979  9.11
## 77      11.36  8.373554 75.92
## 78      11.36  8.373554 75.92
## 79      11.32  8.377701 68.24
## 80      11.32  8.377701 68.24
## 81      11.20  8.464214 47.06
## 82      11.20  8.464214 47.06
## 83      11.13  8.527737 56.10
## 84      11.13  9.080346  3.16
## 85      11.13  8.527737 56.10
## 86      11.13  9.080346  3.16
## 87      11.09  8.731982 21.03
## 88      11.09  8.852665 24.44
## 89      11.09  8.731982 21.03
## 90      11.09  8.852665 24.44
## 91      11.04  7.973155 92.86
## 92      11.04  7.973155 92.86
## 93      11.03  8.981807 23.88
## 94      11.03  8.981807 23.88
## 95      11.00  8.312871 63.23
## 96      11.00  8.312871 63.23
## 97      10.93  9.092907  1.65
## 98      10.93  9.092907  1.65
## 99      10.64  7.803435 91.76
## 100     10.64  7.803435 91.76
## 101     10.57  8.970813  6.01
## 102     10.57  8.970813  6.01
## 103     10.51  8.058960 96.14
## 104     10.51  8.058960 96.14
## 105     10.29  8.603371  2.92
## 106     10.29  8.603371  2.92
## 107     10.10  8.951181  0.78
## 108     10.10  8.951181  0.78
```

```
## 109     10.09  8.992682  1.50
## 110     10.09  8.992682  1.50
## 111     10.07  8.226841 52.27
## 112     10.07  8.226841 52.27
## 113     10.05  7.861342 67.82
## 114     10.05  7.861342 67.82
## 115     10.00  8.773849 13.58
## 116     10.00  8.773849 13.58
## 117      9.93  7.771067  3.69
## 118      9.93  8.874588  0.99
## 119      9.93  7.771067  3.69
## 120      9.93  8.874588  0.99
## 121      9.84  8.920389 17.04
## 122      9.84  8.920389 17.04
## 123      9.62  6.823286  7.00
## 124      9.62  6.823286  7.00
## 125      9.47  9.093357  0.00
## 126      9.47  9.093357  0.00
## 127      9.46  6.416732 96.53
## 128      9.46  6.416732 96.53
## 129      9.45  8.156510 76.14
## 130      9.45  8.156510 76.14
## 131      9.22  8.614683  7.62
## 132      9.22  8.614683  7.62
## 133      9.17  8.468423 11.37
## 134      9.17  8.468423 11.37
## 135      9.05  9.026057  1.34
## 136      9.05  9.026057  1.34
## 137      8.88  8.833608  0.00
## 138      8.88  8.833608  0.00
## 139      8.81  8.807921  4.28
## 140      8.81  8.807921  4.28
## 141      8.78  8.790878  5.78
## 142      8.78  8.790878  5.78
## 143      8.76  8.279697 74.54
## 144      8.76  8.279697 74.54
## 145      8.60  7.412764 27.75
## 146      8.60  7.412764 27.75
## 147      8.55  8.193677 70.87
## 148      8.55  8.193677 70.87
## 149      8.50  8.276649 15.51
## 150      8.50  8.276649 15.51
## 151      8.49  9.087721  0.00
## 152      8.49  9.087721  0.00
## 153      8.43  8.667680 13.62
## 154      8.43  8.667680 13.62
## 155      8.40  8.789660  2.30
## 156      8.40  8.789660  2.30
## 157      8.37  8.466742  0.00
## 158      8.37  8.466742  0.00
## 159      8.33  8.843471  0.61
## 160      8.33  8.843471  0.61
## 161      8.24  9.091669  0.65
## 162      8.24  9.091669  0.65
```

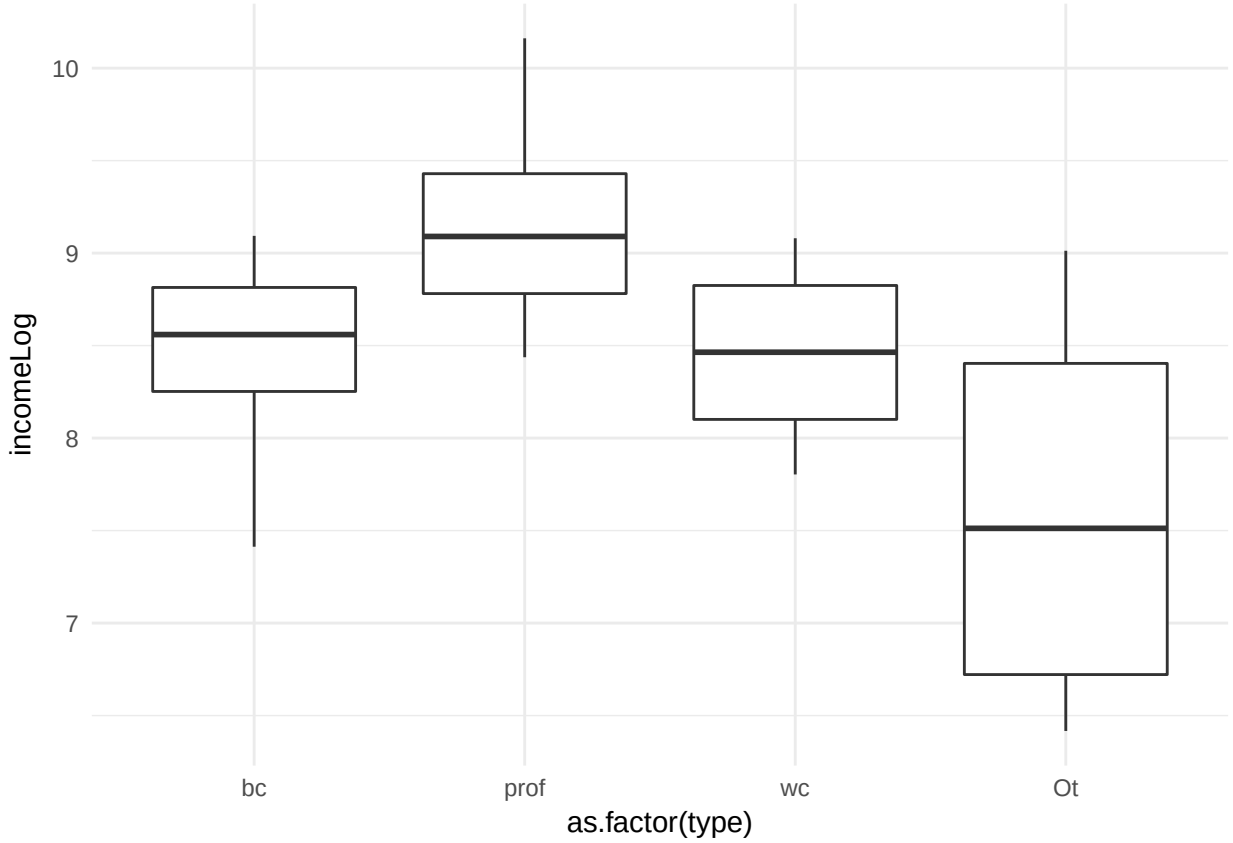
```
## 163      8.10  8.664923  0.81
## 164      8.10  8.664923  0.81
## 165      7.93  8.348775  3.59
## 166      7.93  8.348775  3.59
## 167      7.92  8.776167  5.17
## 168      7.92  8.776167  5.17
## 169      7.81  8.422883  2.46
## 170      7.81  8.422883  2.46
## 171      7.74  8.044626 52.00
## 172      7.74  8.044626 52.00
## 173      7.64  8.543835 17.26
## 174      7.64  8.543835 17.26
## 175      7.64  8.543835 17.26
## 176      7.64  8.543835 17.26
## 177      7.58  8.183956 30.08
## 178      7.58  8.623893  9.47
## 179      7.58  8.183956 30.08
## 180      7.58  8.623893  9.47
## 181      7.54  8.342840 33.30
## 182      7.54  8.342840 33.30
## 183      7.52  8.271548  1.09
## 184      7.52  8.271548  1.09
## 185      7.42  7.544861 72.24
## 186      7.42  7.544861 72.24
## 187      7.33  8.006701 69.31
## 188      7.33  8.006701 69.31
## 189      7.11  8.152774 33.57
## 190      7.11  8.152774 33.57
## 191      6.92  8.575462  0.56
## 192      6.92  8.575462  0.56
## 193      6.84  8.200837  3.60
## 194      6.84  8.200837  3.60
## 195      6.74  8.156510 39.48
## 196      6.74  8.156510 39.48
## 197      6.69  8.399310 31.36
## 198      6.69  8.399310 31.36
## 199      6.67  8.454679  0.00
## 200      6.67  8.454679  0.00
## 201      6.60  8.692826  0.52
## 202      6.60  8.692826  0.52
## 203      6.38  7.954372 90.67
## 204      6.38  7.954372 90.67
```

```
myprestige2 %>% arrange(desc(education)) %>% filter(education >= 15.96)
```

```
##   education women prestige census type                job incomeLog
## 1     15.97 19.59     84.6   2711 prof university.teachers  9.431963
## 2     15.97 19.59     84.6   2711 prof university.teachers  9.431963
## 3     15.96 10.56     87.2   3111 prof          physicians 10.138915
## 4     15.96 10.56     87.2   3111 prof          physicians 10.138915
##   nonsense
## 1    35.56
## 2    35.56
## 3    26.52
## 4    26.52
```

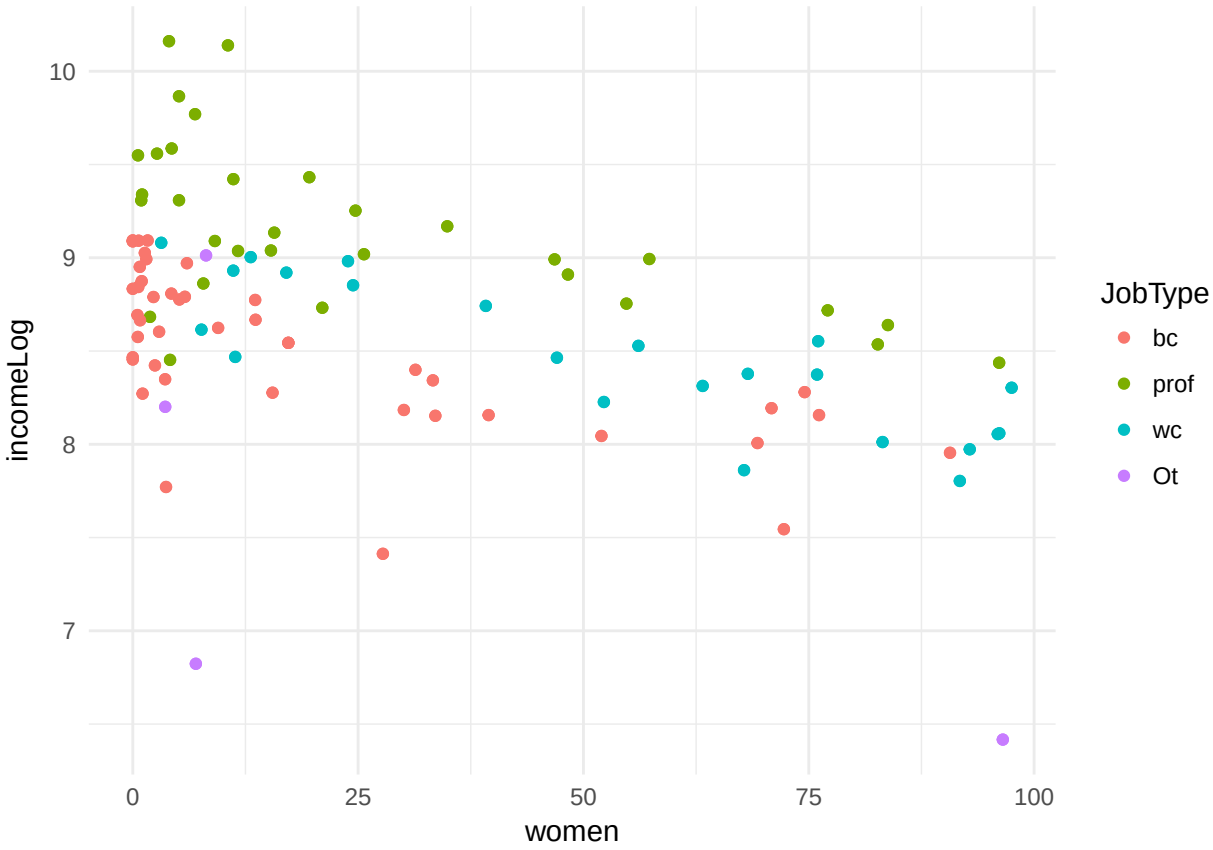
Does having more bedrooms increase sale price?

```
myprestige2 %>% ggplot(aes(x=as.factor(type), y = incomeLog)) +   geom_boxplot() + theme_minimal()
```



Visualise both number of bedrooms (as a factor) and TotalSqFt as a scatterplot to see if a trend is visible.

```
myprestige2 %>% ggplot(aes(x=women, y = incomeLog, colour = as.factor(type))) + geom_point() + theme_mi
```



Are newer or more recently renovated properties more expensive? Investigate this generally and then specifically for 2 - 4 bedroom properties.

```
# no code
```

Lets convert kitchen quality to numeric (we'll see why we need this later):

From the metadata we know it can be:

- Ex Excellent
- Gd Good
- TA Typical/Average
- Fa Fair
- Po Poor

Recode this to numeric values using mutate() and recode().

```
myprestige2 <- myprestige2 %>% mutate(type = dplyr::recode(type, `prof` = 4L, `wc` = 3L, `bc` = 2L, `Ot
summary(myprestige2$type)
```

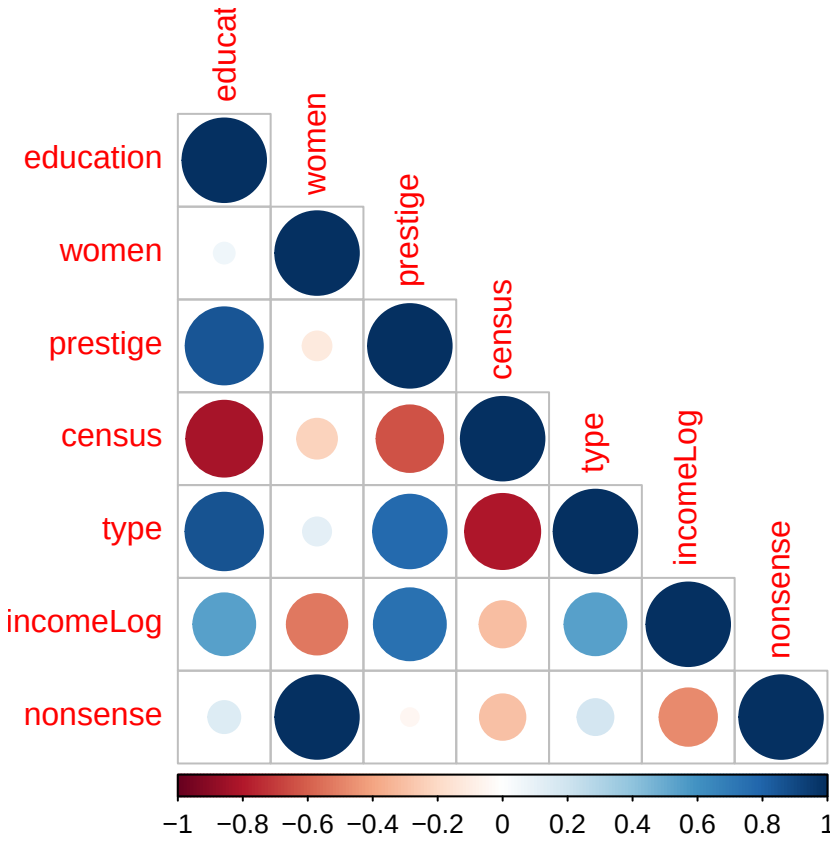
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   2.794   4.000   4.000
```

Convert Bldgtype to numeric

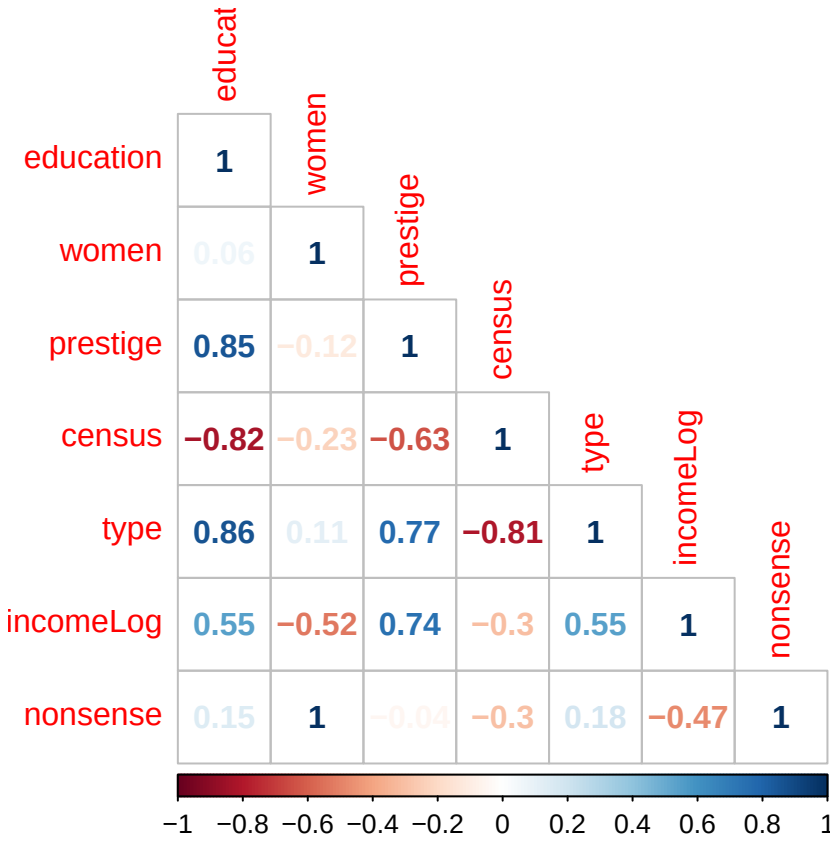
```
# no need for code
```

What variables are correlated with each other and with price? Plot a correlation plot using corrplot() for all numeric variables and those that show the top correlation with LogSalePrice.

```
myprestige2num <- myprestige2[ , sapply(myprestige2, is.numeric)]
corrplot(cor(myprestige2num, use="everything"), method="circle", type="lower",  sig.level = 0.01, insig
```




```
corrplot(cor(myprestige2num, use="everything"), method="number", type="lower",  sig.level = 0.01, insig
```



Use the createDataPartition() function to separate the training data into a training and testing subset. Allocate 50% of the data to each class. Run set.seed(12) before this.

```
set.seed(12)
partitionD <- createDataPartition(y = myprestige2num$incomeLog, p = 0.5, list=FALSE)
myprestige2train <- myprestige2num[partitionD,]
myprestige2test <- myprestige2num[-partitionD,]
```

Fit a linear model considering the "top 10"" correlated (top 9, ignore LogSalePrice for obvious reasons).

```
lm_myprestige1 <- lm(incomeLog ~ education, data=myprestige2train)
lm_myprestige2 <- lm(incomeLog ~ education + women + prestige, data=myprestige2train)
summary(lm_myprestige1)
```

```
##
## Call:
## lm(formula = incomeLog ~ education, data = myprestige2train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.08902 -0.23147  0.05233  0.33801  0.81662
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.31736    0.20527   35.65  < 2e-16 ***
## education    0.12562    0.01858    6.76 9.35e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.5083 on 100 degrees of freedom
## Multiple R-squared:  0.3137, Adjusted R-squared:  0.3068
## F-statistic:  45.7 on 1 and 100 DF,  p-value: 9.354e-10
```

```
summary(lm_myprestige2)
```

```
##
## Call:
## lm(formula = incomeLog ~ education + women + prestige, data = myprestige2train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.22509 -0.08212  0.06043  0.17939  0.44993
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.783801   0.139658  55.735  < 2e-16 ***
## education   -0.010393   0.023902  -0.435    0.665
## women       -0.007319   0.001107  -6.614 1.98e-09 ***
## prestige     0.025594   0.003948   6.483 3.65e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3314 on 98 degrees of freedom
## Multiple R-squared:  0.714,  Adjusted R-squared:  0.7052
## F-statistic: 81.55 on 3 and 98 DF,  p-value: < 2.2e-16
```

Use predict() to predict house prices using our top10 model on the "test" portion of the training dataset. Use rmse to assess the root mean square error (our metric of accuracy).

```
prediction_lm1 <- predict(lm_myprestige1, myprestige2test, type="response")
prediction_lm2 <- predict(lm_myprestige2, myprestige2test, type="response")

# rmse?
rmse(myprestige2test$incomeLog, prediction_lm1)
```

```
## [1] 0.4814839
```

```
rmse(myprestige2test$incomeLog, prediction_lm2)
```

```
## [1] 0.2732863
```

All other models - just work in the housing template/final housing template files.

Where to from here

- DataCamp
- R-Bloggers
- RStudio webinars
- Our data today: LOTS more info and analysis
- ISWR
- EOSL
- AnalyticsEdgeMIT
- Anything Hadley Wickham does***