Housing data analysis - Women Who Code workshop

Darya Vanichkina

13/03/2018

```
load.libraries <- c('tidyverse', 'forcats', 'corrplot', 'caret', 'Metrics', 'randomForest', 'xgboost',
# note car only for Darya to use the demo dataset
install.lib <- load.libraries[!load.libraries %in% installed.packages()]
for(libs in install.lib) install.packages(libs, dependences = TRUE)
sapply(load.libraries, library, character = TRUE)
knitr::opts_chunk$set(echo = TRUE)
```

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:

```
# getting help
?c
x <- c(1,2,3)
class(x)
y <- c(x, "2")
class(y)
```

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

Task 1

Load the data in from csv.

```
# read.csv("file")
myprestige <- Prestige
myprestige$job <- row.names(myprestige)
?Prestige
myprestige
head(myprestige)
```

Task 2

1. What features are there in the data?
2. What are the dimensions of the data?
3. What are the column headers?

Use the summary() and str() functions to explore…

```
summary(myprestige)
str(myprestige)
table(myprestige$type)
```

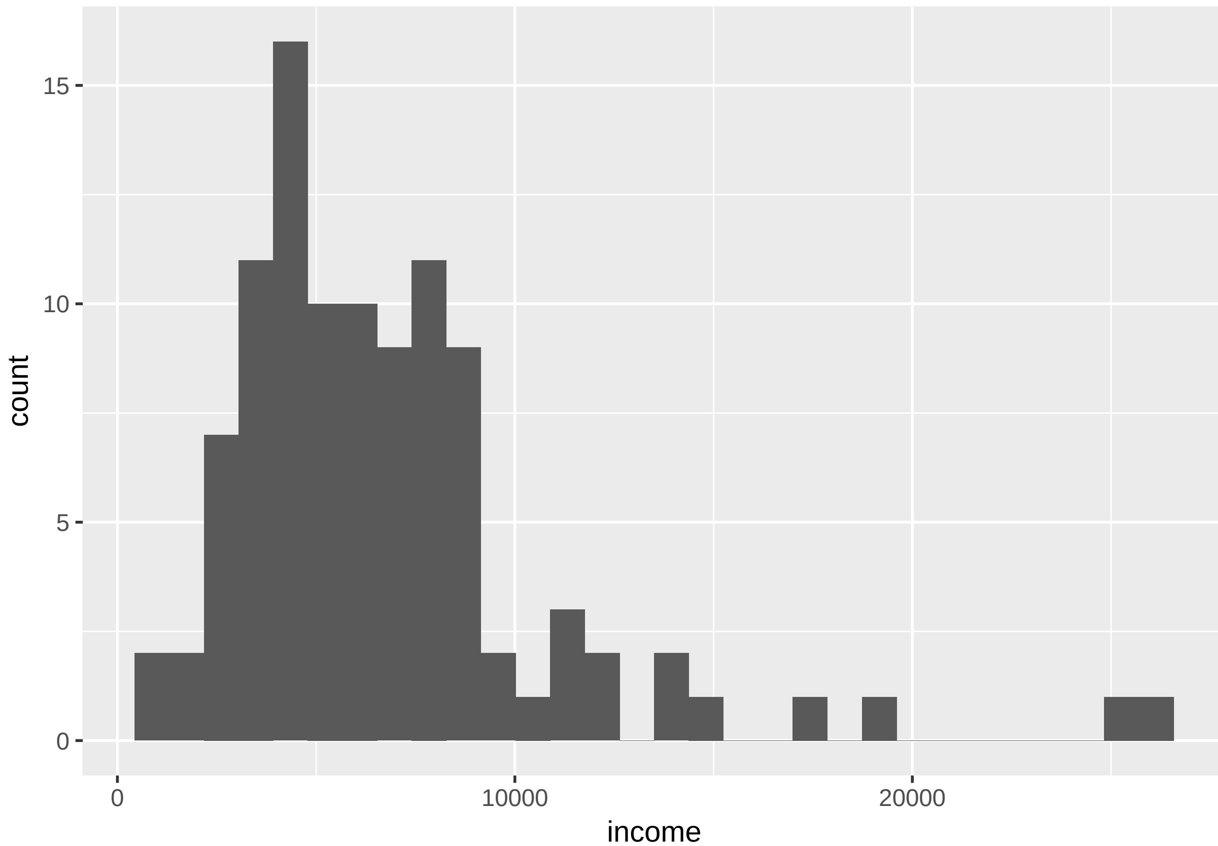
What does the distribution of sale price look like?

Task 3

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

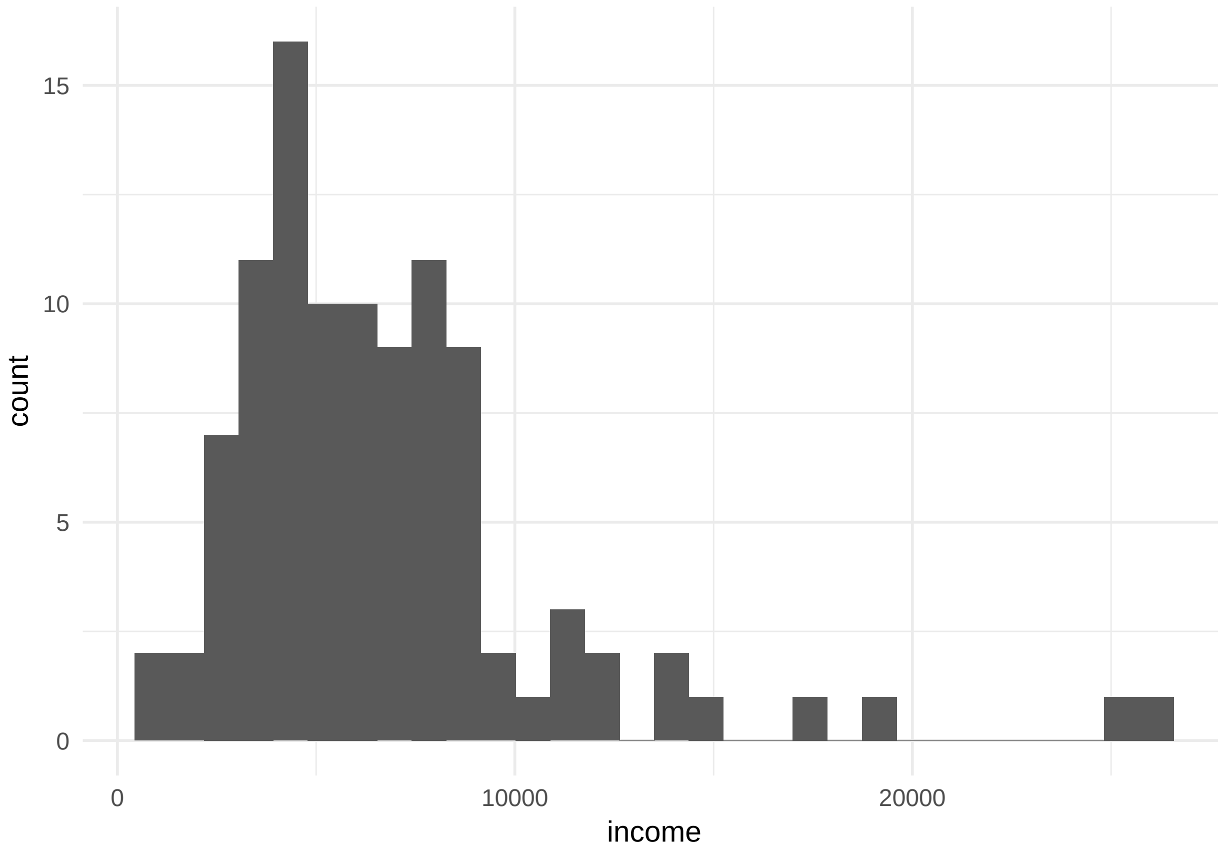
```
myprestige %>% ggplot(aes(x = income)) + geom_histogram()
```

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

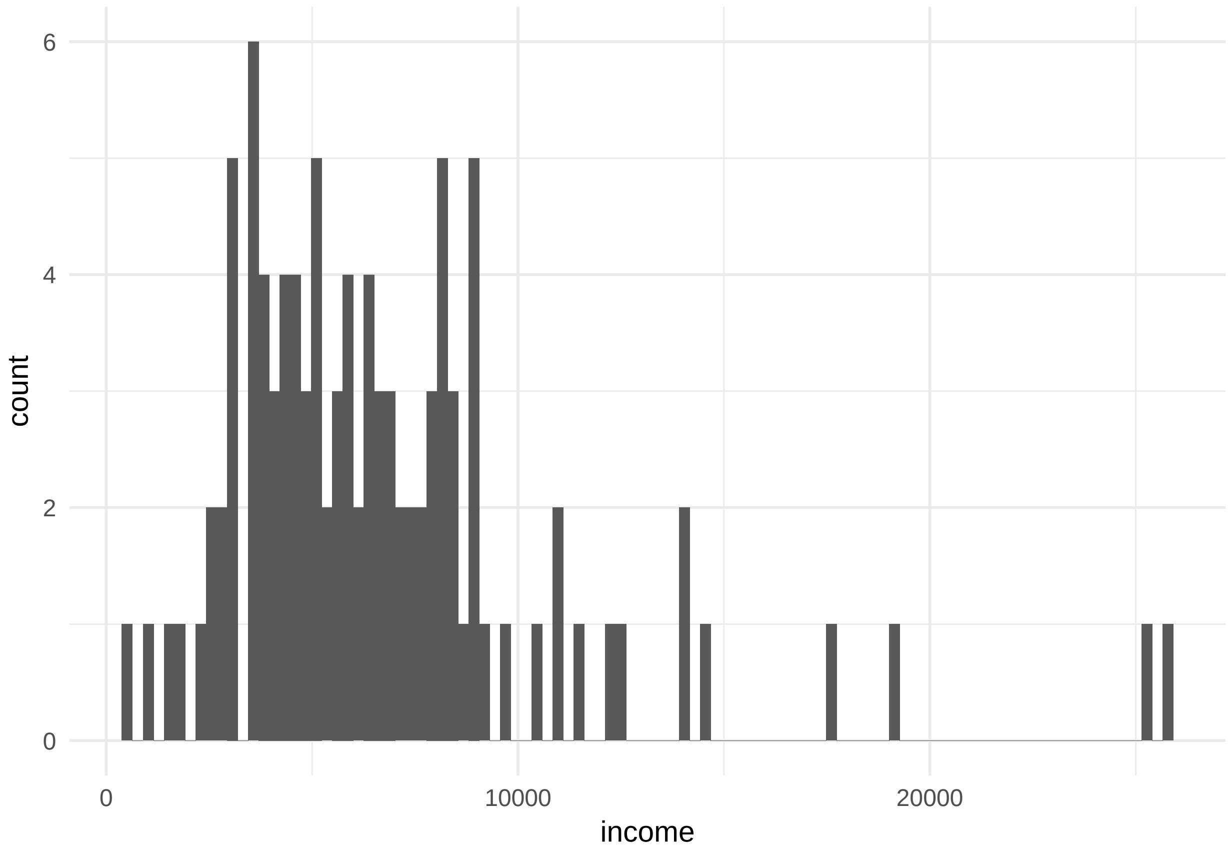



```r
myprestige %>% ggplot(aes(x = income)) + geom_histogram() + theme_minimal()
```

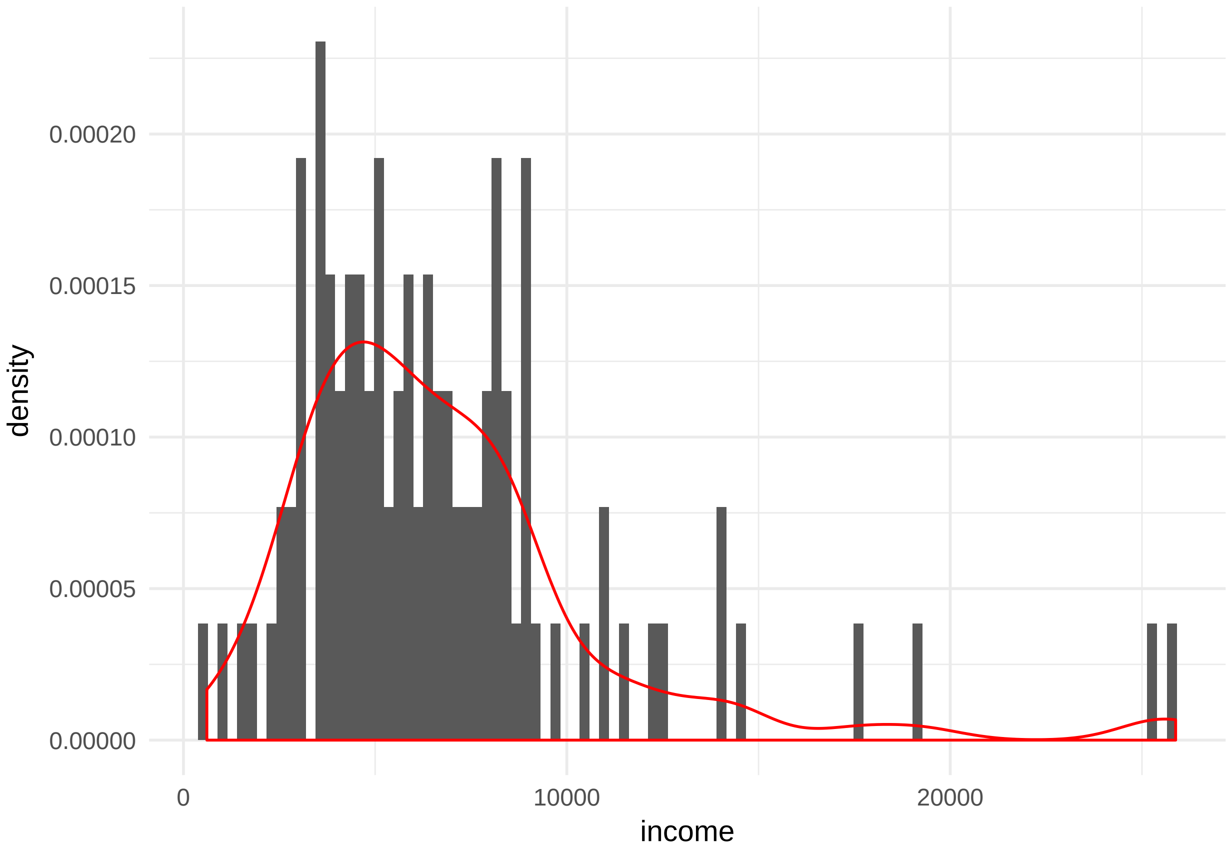
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



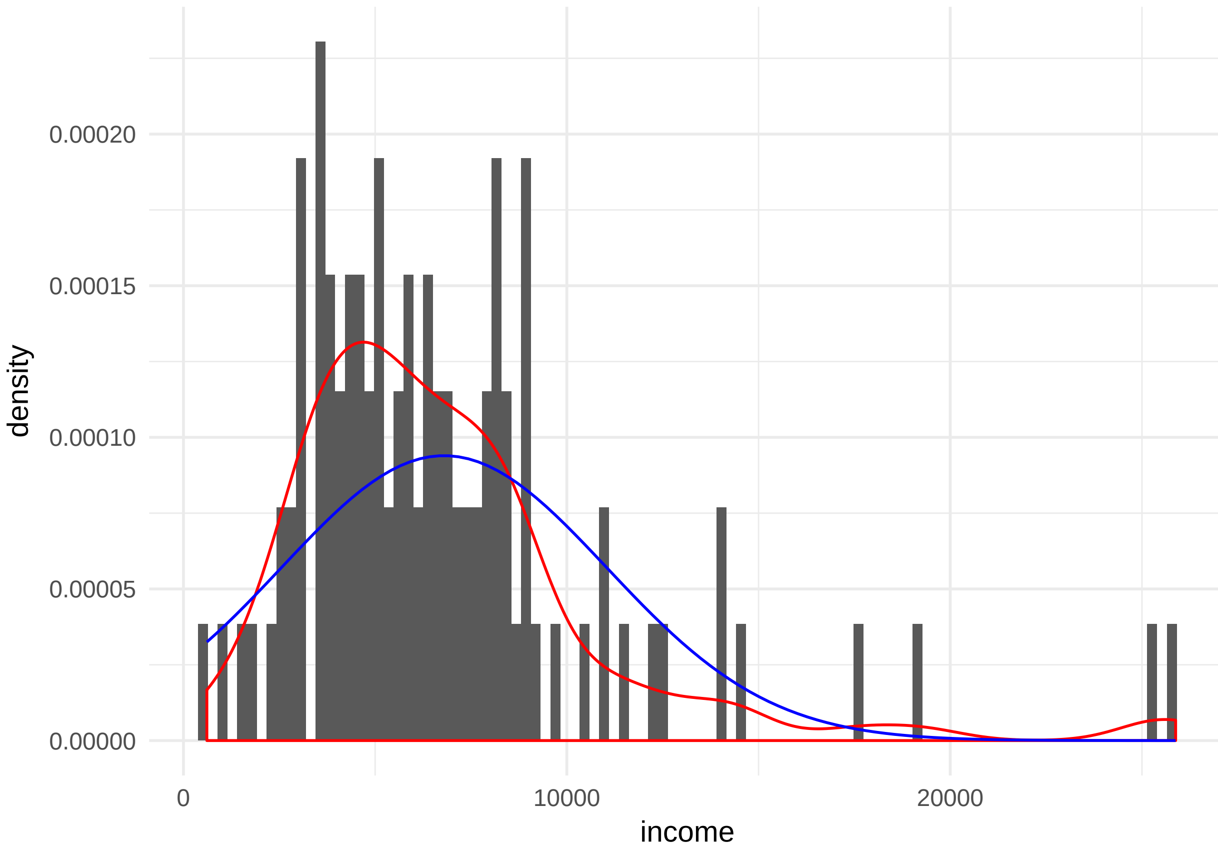

```
myprestige %>% ggplot(aes(x = income)) + geom_histogram(bins = 100) + theme_minimal()
```




```
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
```

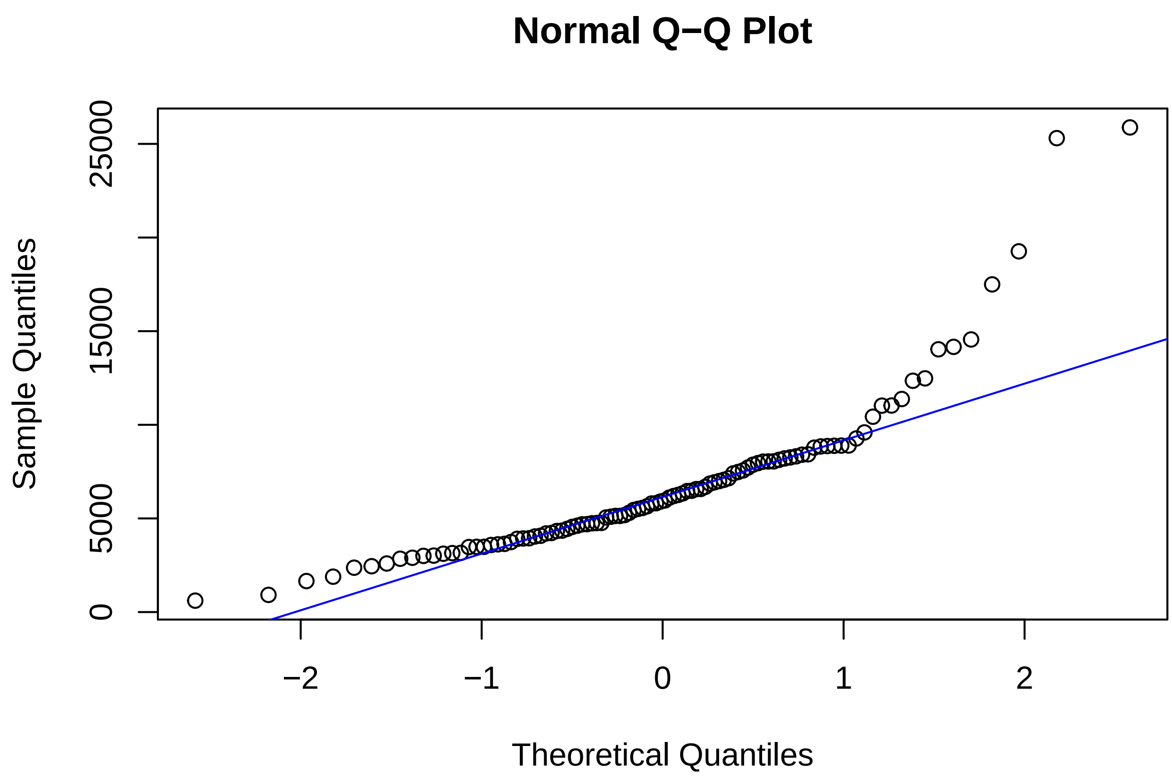


Task 4

1.Plot a quantile-quantile plot (QQ plot) to "assess" normality.

`Note: This plot compares the data we have (Sample Quantiles) with a theoretical sample coming from a no`

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



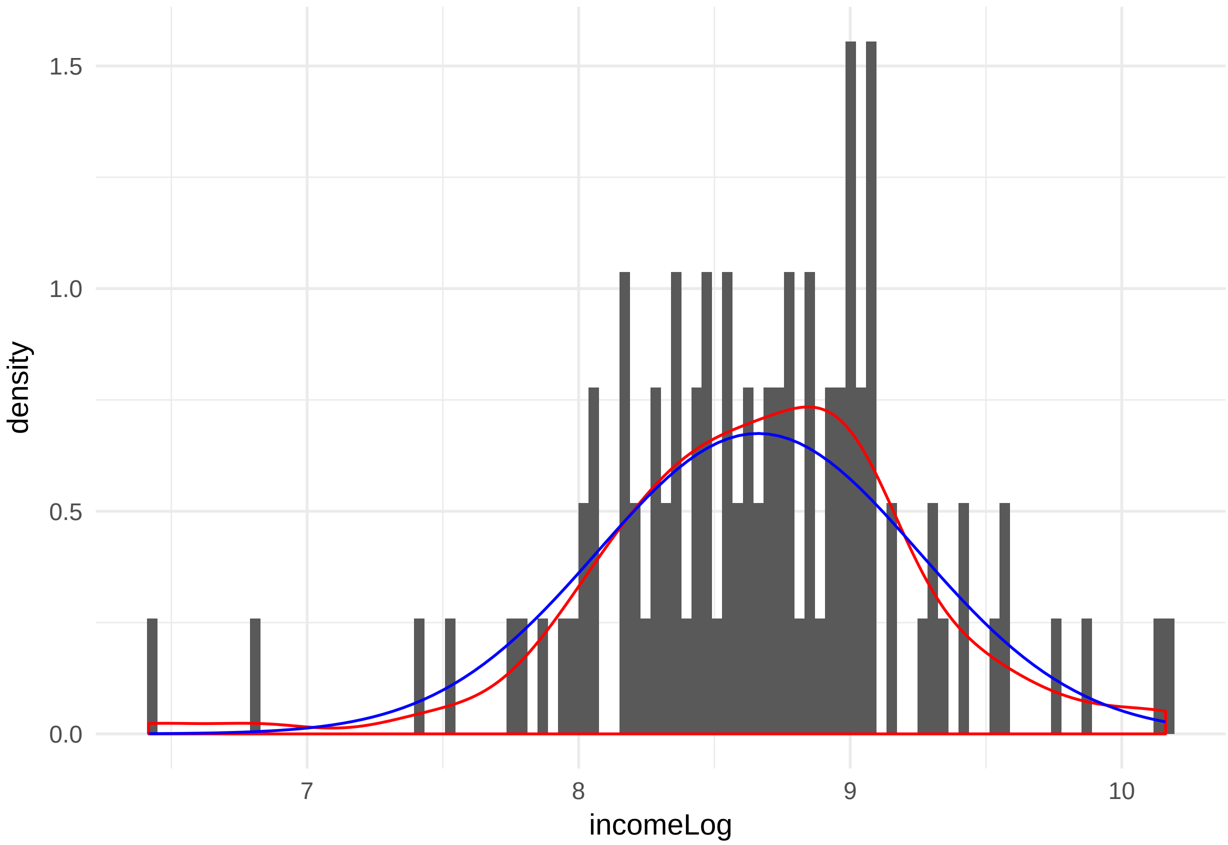
A standard way of transforming the data to be better approximated by a normal distribution is by using the log-transform?

Task 5

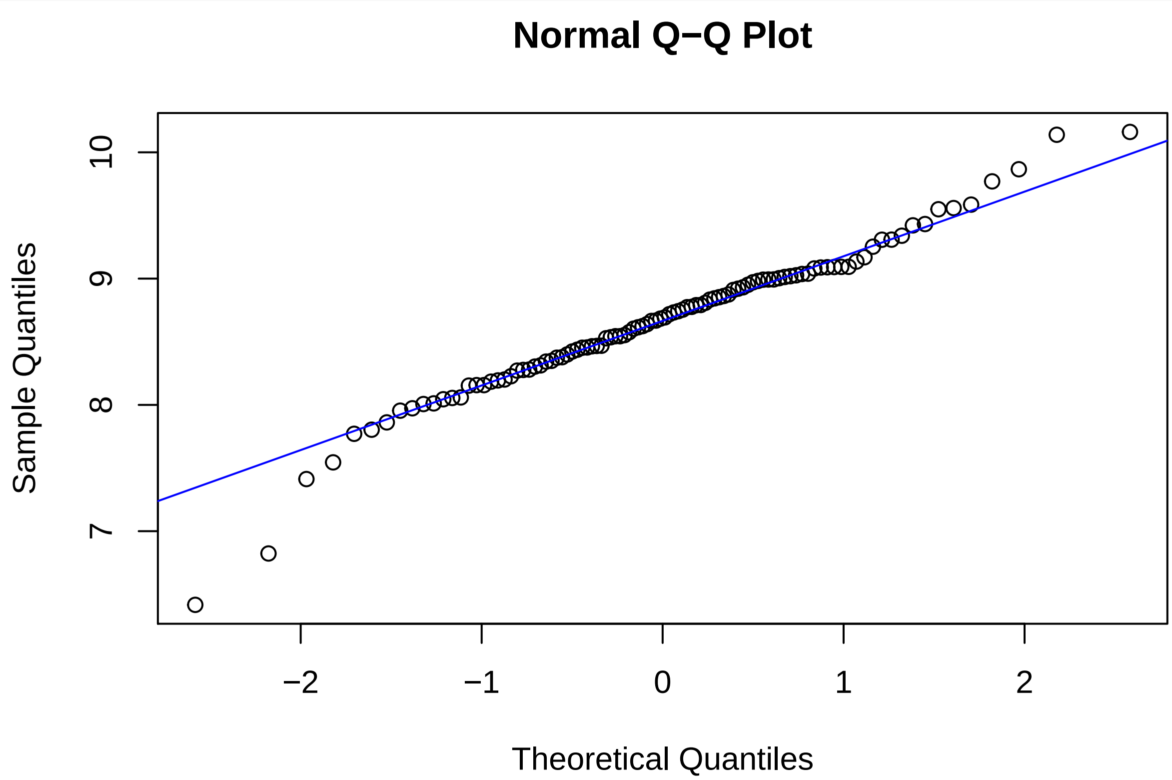
1. Carry out this transformation
2. Use a histogram and QQ plot to see whether it works…

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

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




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



Missing data

Task 6

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!

Task 7

Combine the testing and training data.

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

How do we impute the missing data?

Task 8

Explore the data using the table() function (variable by variable).

```
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.

Task 9

Recode the NA values that should be None using mutate() and fct_explicit_na().

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

Task 10

For the GarageYrBlt - set NA values using replace_na() to zero.

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

Task 11

For Lot frontage - set it to be the median for the neighborhood using group_by() and mutate().

```
# 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
```

Task 12

Split back into training (trainHC) and test (testHC) sets (because kaggle training set had prices, test didn't).

```
# no demo here
```

Basic exploratory data analysis of training data

Task 13

1. How does the sale price depend on living area: X1stFlrSF, X2ndFlrSF, TotalBsmtSF? (use a scatterplot to visualise this)
2. Create a variable TotalSqFt which is a combination of these
3. Does it better predict the house price? (again, just using scatterplot at this point)

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

Task 14

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

Useful reference for dplyr

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

```
##   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
## 5     15.94  4.32     66.7   3115 prof       veterinarians  9.585965
## 6     15.94  4.32     66.7   3115 prof       veterinarians  9.585965
##   nonsense
## 1    35.56
## 2    35.56
## 3    26.52
## 4    26.52
```

```
## 5    20.26
## 6    20.26
```

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

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

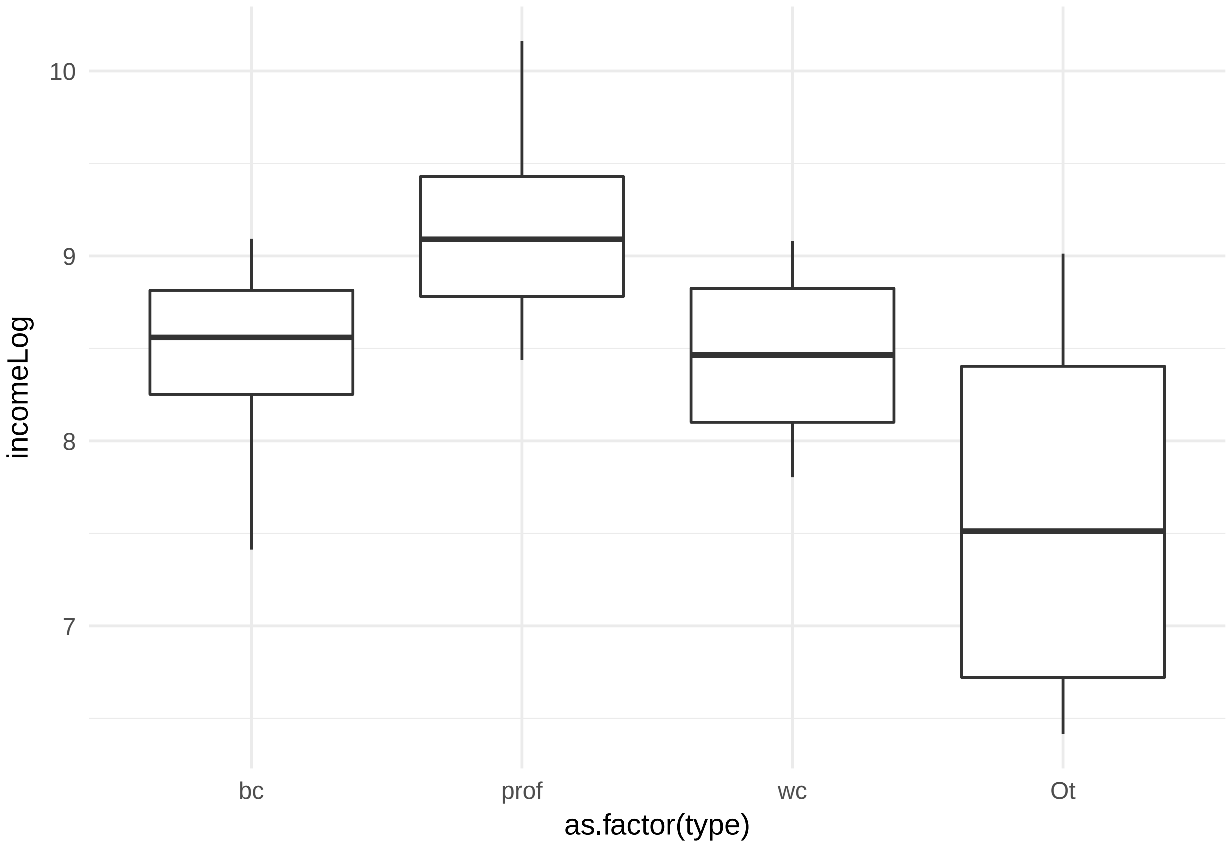
```
##   education women prestige census type                 job incomeLog
## 1     15.97 19.59     84.6   2711 prof university.teachers  9.431963
## 2     15.97 19.59     84.6   2711 prof university.teachers  9.431963
## 3     15.96 10.56     87.2   3111 prof          physicians 10.138915
## 4    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

Task 15

Use a geom_boxplot() to explore this

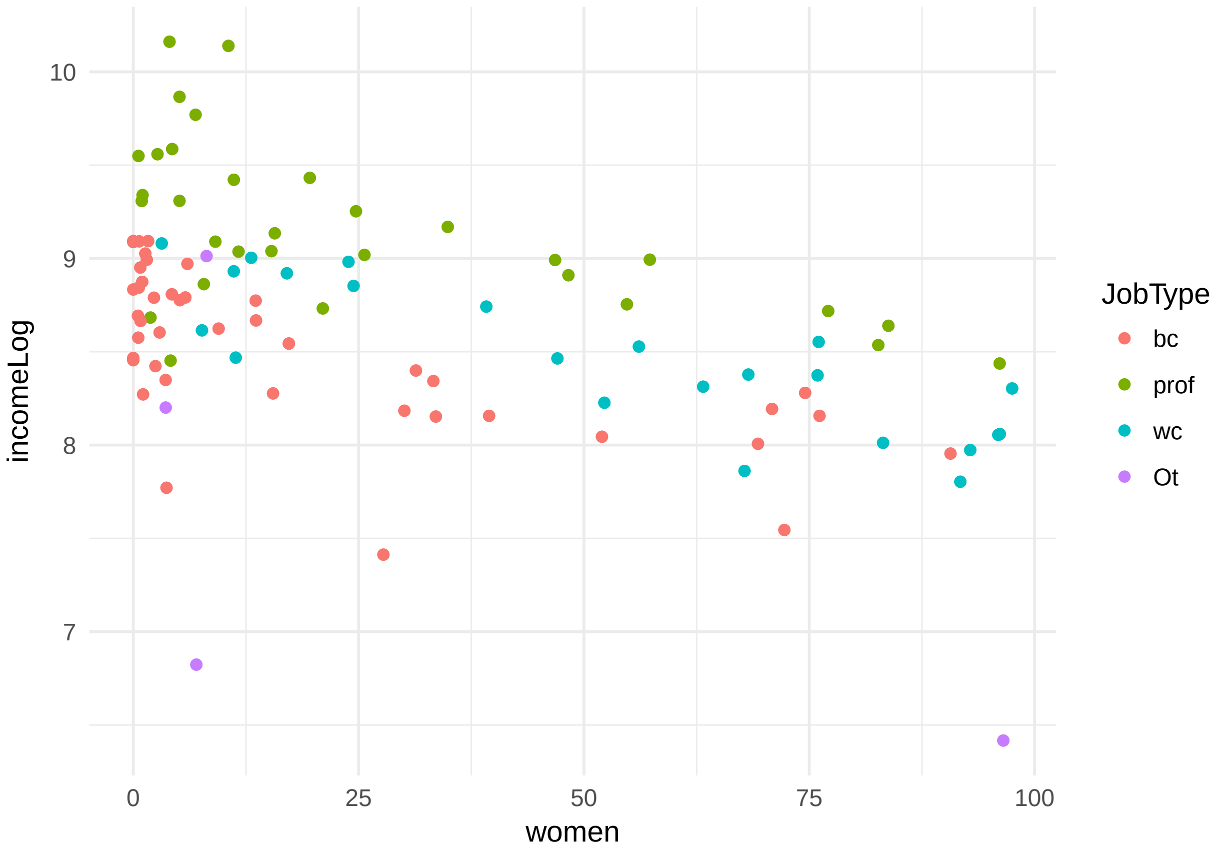
```
myprestige2 %>% ggplot(aes(x=as.factor(type), y = incomeLog)) +   geom_boxplot() + theme_minimal()
```



Task 16

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?

Task 17

1. Investigate this generally and then
2. . . . 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

Task 18

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
```

Task 19

Convert Bldgtype to numeric

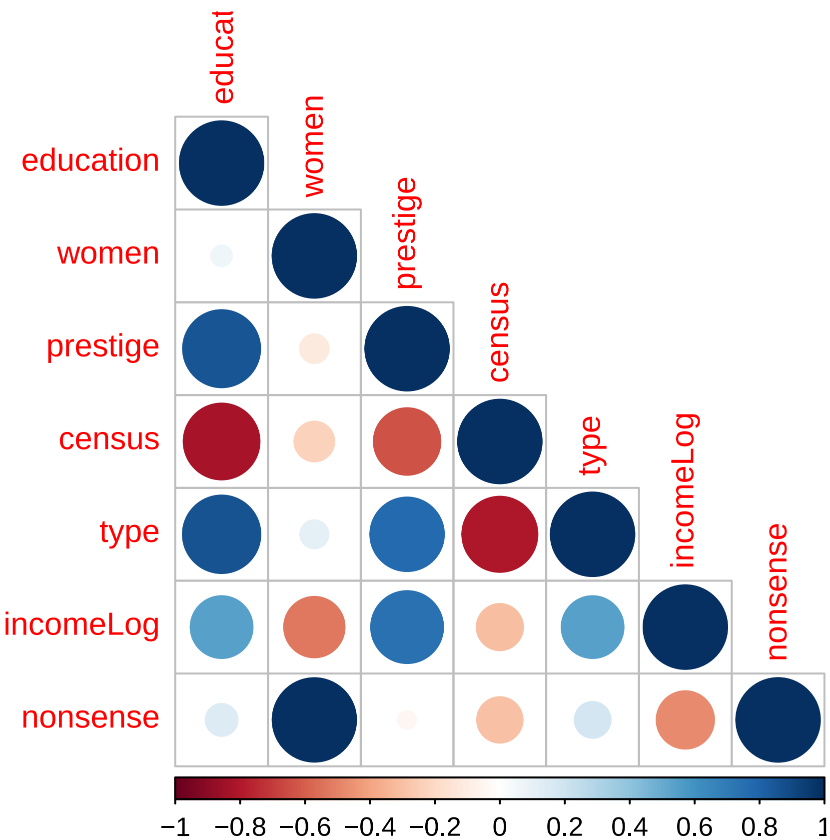
```
# no need for code
```

What variables are correlated with each other and with price?

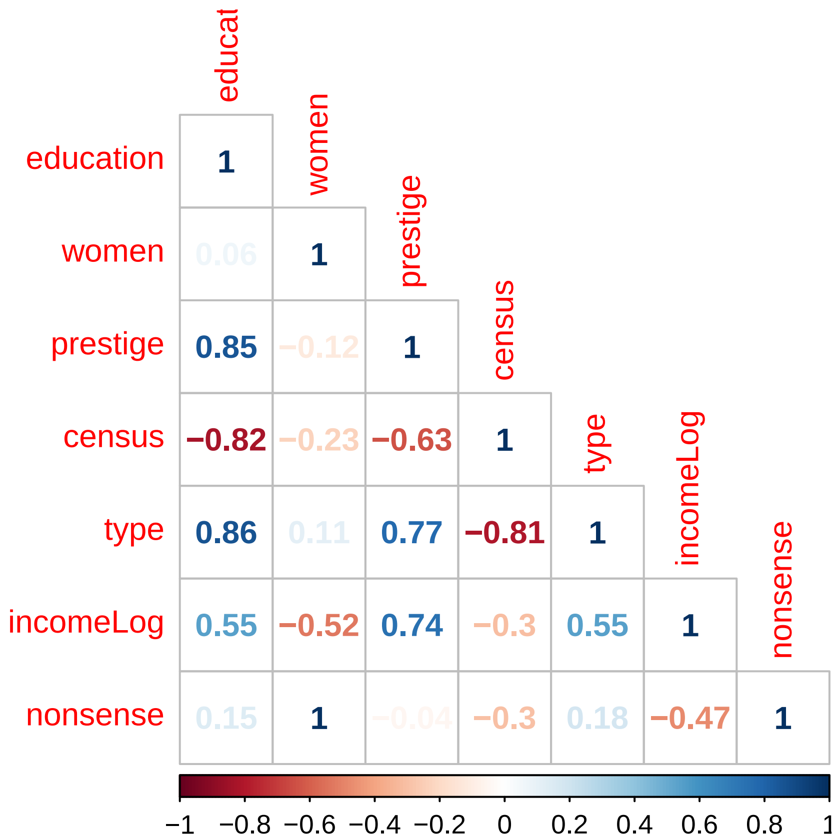
Task 20

1. Plot a correlation plot using corrplot() for all numeric variables and
2. ... 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
```



Task 21

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,]
```

Task 22

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

```
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
```

```r
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
```

Task 23

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

```r
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
```

```r
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 - kaggle
- Introductory Statistics with R
- Elements of Statistical Learning
- AnalyticsEdgeMIT
- Anything Hadley Wickham does***

Task 24

1. Use randomForest() to train a random forest model on all of the variables.
2. Use predict() and rmse() to make the prediction and assess the accuracy respectively.
3. Was a linear (on 9 features) or random forest model more accurate?

```
# randFor <- randomForest(LogSalePrice ~ ., data=trainHCtrain)
# # Predict using the test set
# prediction_rf <- predict(randFor, trainHCtest)
# trainHCtest$randFor <- prediction_rf
# # rmse?
# rmse(trainHCtest$LogSalePrice, trainHCtest$randFor)
```

Task 25

1. Use xgboost to predict house prices from numeric features of training dataset.
2. Use xgb.plot.importance() to assess which variables are most important for predicting house prices.

```
# trainHCtrainNum <- as(as.matrix(trainHCtrain[ , sapply(trainHCtrain, is.numeric)]), "sparseMatrix")
# trainHCtestNum <-  as(as.matrix(trainHCtest[ , sapply(trainHCtest, is.numeric)]), "sparseMatrix")
#
# trainD <- xgb.DMatrix(data = trainHCtrainNum, label = trainHCtrainNum[,"LogSalePrice"])
#
# #Cross validate the model
# cv.sparse <- xgb.cv(data = trainD,
#                     nrounds = 600,
#                     min_child_weight = 0,
#                     max_depth = 10,
#                     eta = 0.02,
#                     subsample = .7,
#                     colsample_bytree = .7,
#                     booster = "gbtree",
#                     eval_metric = "rmse",
#                     verbose = TRUE,
#                     print_every_n = 50,
#                     nfold = 4,
#                     nthread = 2,
#                     objective="reg:linear")
#
# #Train the model
# #Choose the parameters for the model
# param <- list(colsample_bytree = .7,
#               subsample = .7,
#               booster = "gbtree",
#               max_depth = 10,
#               eta = 0.02,
```

```
#               eval_metric = "rmse",
#               objective="reg:linear")
#
#
# #Train the model using those parameters
# bstSparse <-
#   xgb.train(params = param,
#              data = trainD,
#              nrounds = 600,
#              watchlist = list(train = trainD),
#              verbose = TRUE,
#              print_every_n = 50,
#              nthread = 2)
#
# testD <- xgb.DMatrix(data = trainHCtestNum)
#
# prediction <- predict(bstSparse, testD) #Make the prediction based on the half of the training data s
#
# #Put testing prediction and test dataset all together
#
# prediction <- as.data.frame(as.matrix(prediction))
# colnames(prediction) <- "xgboost"
# trainHCtest$xgboost <- prediction$xgboost
#
#
# #Test with RMSE
# rmse(trainHCtest$LogSalePrice, trainHCtest$xgboost)
#
# # Feature importance
# importance_matrix <- xgb.importance(dimnames(trainD)[[2]], model = bstSparse)
# xgb.plot.importance(importance_matrix[1:10])
```

Task 26

1.Use the glmnet library to train a ridge regression model. 2. Is it more or less accurate than XGBoost?

```
# trainHCtrainNumMatrix <- as.matrix(trainHCtrain[ , sapply(trainHCtrain, is.numeric)])
# trainHCtestNumMatrix  <-  as.matrix(trainHCtest[ , sapply(trainHCtest, is.numeric)])
# # cross validation for glmnet
# glm.cv.ridge <- cv.glmnet(trainHCtrainNum[,c(1:38,40)], trainHCtrainNum[,"LogSalePrice"], alpha = 0)
# penalty.ridge <- glm.cv.ridge$lambda.min
# glm.ridge <- glmnet(x = trainHCtrainNum[,c(1:38,40)], y = trainHCtrainNum[,"LogSalePrice"], alpha = 0
# y_pred.ridge <- as.numeric(predict(glm.ridge, trainHCtestNum[,c(1:38,40)] ))
# rmse(trainHCtest$LogSalePrice, y_pred.ridge)
```