Data Summaries and Exploration

Rei Sanchez-Arias, Ph.D.

Exploratory Data Analysis Example

House Prices Dataset

Data

For each neighborhood, a number of variables are given, such as the crime rate, the student/teacher ratio, and the median value of a housing unit in the neighborhood.

The file `BostonHousing.csv` contains information collected by the US Bureau of the Census concerning housing in the area of Boston, Massachusetts. The dataset includes information on 506 census housing tracts in the Boston area. The goal is to *predict the median house price* in new tracts based on information such as crime rate, pollution, and number of rooms. The response is the median house price (`MEDV`).

Variables	Description
CRIM	Crime rate
ZN	Percentage of residential land zoned for lots over 25,000 ft2
INDUS	Percentage of land occupied by non-retail business
CHAS	Does tract bound Charles River (= 1 if tract bounds river, = 0 otherwise)
NOX	Nitric oxide concentration (parts per 10 million)
RM	Average number of rooms per dwelling
AGE	Percentage of owner-occupied units built prior to 1940
DIS	Weighted distances to five Boston employment centers
RAD	Index of accessibility to radial highways
TAX	Full-value property tax rate per $10,000
PTRATIO	Pupil-to-teacher ratio by town
LSTAT	Percentage of lower status of the population
MEDV	Median value of owner-occupied homes in $1000s
CAT_MEDV	Is median value of owner-occupied homes in tract above $30,000 (CAT.MEDV = 1) or not (CAT_MEDV = 0)

Read dataset

```
library(tidyverse)
housing <- read_csv("https://raw.githubusercontent.com/reisanar/datasets/master/BostonHousing
# Print the first 6 observations
head(housing)
```

```
## # A tibble: 6 x 14
##      CRIM    ZN INDUS  CHAS   NOX    RM   AGE   DIS   RAD   TAX PTRATIO LSTAT  MEDV CAT_MED
##     <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>   <dbl> <dbl> <dbl>    <dbl
## 1 0.00632    18  2.31     0 0.538  6.58  65.2  4.09     1   296    15.3  4.98  24           0
## 2 0.0273      0  7.07     0 0.469  6.42  78.9  4.97     2   242    17.8  9.14  21.6         0
## 3 0.0273      0  7.07     0 0.469  7.18  61.1  4.97     2   242    17.8  4.03  34.7         1
## 4 0.0324      0  2.18     0 0.458  7.00  45.8  6.06     3   222    18.7  2.94  33.4         1
## 5 0.0690      0  2.18     0 0.458  7.15  54.2  6.06     3   222    18.7  5.33  36.2         1
## 6 0.0298      0  2.18     0 0.458  6.43  58.7  6.06     3   222    18.7  5.21  28.7         0
```

Data Summaries

summary()

Numerical summaries and graphs of the data are very helpful for data reduction. The information that they convey can assist in combining categories of a categorical variable, in choosing variables to remove, in assessing the level of information overlap between variables, and more.

```
# summary statistics of 5 features
summary(housing[ , c("CRIM", "ZN", "RM", "MEDV", "CHAS")])
##       CRIM                ZN               RM             MEDV           CHAS        
##  Min.   : 0.00632   Min.   :  0.00   Min.   :3.561   Min.   : 5.00   Min.   :0.00000  
##  1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.:5.886   1st Qu.:17.02   1st Qu.:0.00000  
##  Median : 0.25651   Median :  0.00   Median :6.208   Median :21.20   Median :0.00000  
##  Mean   : 3.61352   Mean   : 11.36   Mean   :6.285   Mean   :22.53   Mean   :0.06917  
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:6.623   3rd Qu.:25.00   3rd Qu.:0.00000  
##  Max.   :88.97620   Max.   :100.00   Max.   :8.780   Max.   :50.00   Max.   :1.00000
```

Other summaries

Summaries R Code

```
## # A tibble: 14 x 5
##    var_name var_mean  var_sd var_median var_miss
##    <chr>       <dbl>   <dbl>      <dbl>    <dbl>
##  1 CRIM       3.61     8.60       0.257        0
##  2 ZN        11.4     23.3        0            0
##  3 INDUS     11.1      6.86       9.69         0
##  4 CHAS       0.0692   0.254      0            0
##  5 NOX        0.555    0.116      0.538        0
##  6 RM         6.28     0.703      6.21         0
##  7 AGE       68.6     28.1       77.5          0
##  8 DIS        3.80     2.11       3.21         0
##  9 RAD        9.55     8.71       5            0
## 10 TAX      408.     169.       330            0
## 11 PTRATIO   18.5      2.16      19.0          0
## 12 LSTAT     12.7      7.14      11.4          0
## 13 MEDV      22.5      9.20      21.2          0
## 14 CAT_MEDV   0.166    0.372      0            0
```

Correlation

Next, we summarize relationships between two or more variables. For *numerical variables*, we can compute a complete **matrix of correlations** between each pair of variables, using the R function `cor()`.

```
# (sub) matrix with correlation coefficients for some variables
cor(housing[ , c("CRIM", "ZN", "RM", "MEDV", "CHAS")])
```

```
##              CRIM          ZN          RM        MEDV        CHAS
## CRIM  1.00000000 -0.20046922 -0.21924670 -0.3883046 -0.05589158
## ZN   -0.20046922  1.00000000  0.31199059  0.3604453 -0.04269672
## RM   -0.21924670  0.31199059  1.00000000  0.6953599  0.09125123
## MEDV -0.38830461  0.36044534  0.69535995  1.0000000  0.17526018
## CHAS -0.05589158 -0.04269672  0.09125123  0.1752602  1.00000000
```

Some notes

We see that most correlations are low and that many are negative. Pairs that have a very strong (positive or negative) correlation contain a lot of *overlap* in information and are good candidates for **data reduction** by removing one of the variables.

Another useful approach is **aggregation** by one or more variables. Below is the number of neighborhoods that bound the Charles River vs. those that do not. It appears that the majority of neighborhoods (471 of 506) do not bound the river.

```
# contingency table
table(housing$CHAS)
```

```
##
##   0   1
## 471  35
```

Working with categorical variables

Reducing the Number of Categories

When a categorical variable has many categories, and this variable is destined to be a predictor, many data mining methods will require converting it into many **dummy variables**. In particular, a variable with m categories will be transformed into either m or $m - 1$ dummy variables (depending on the method). This means that even if we have very few original categorical variables, they can greatly *inflate the dimension* of the dataset.

One way to handle this is to **reduce the number of categories** by combining close or similar categories. Combining categories requires incorporating *expert knowledge and common sense*.

> Let us compute the proportion of observations for which median value of owner-occupied homes in tract is above $30000 (`CAT.MEDV`), per percentage of residential land zoned for lots over 25,000 ft2 (`ZN`)

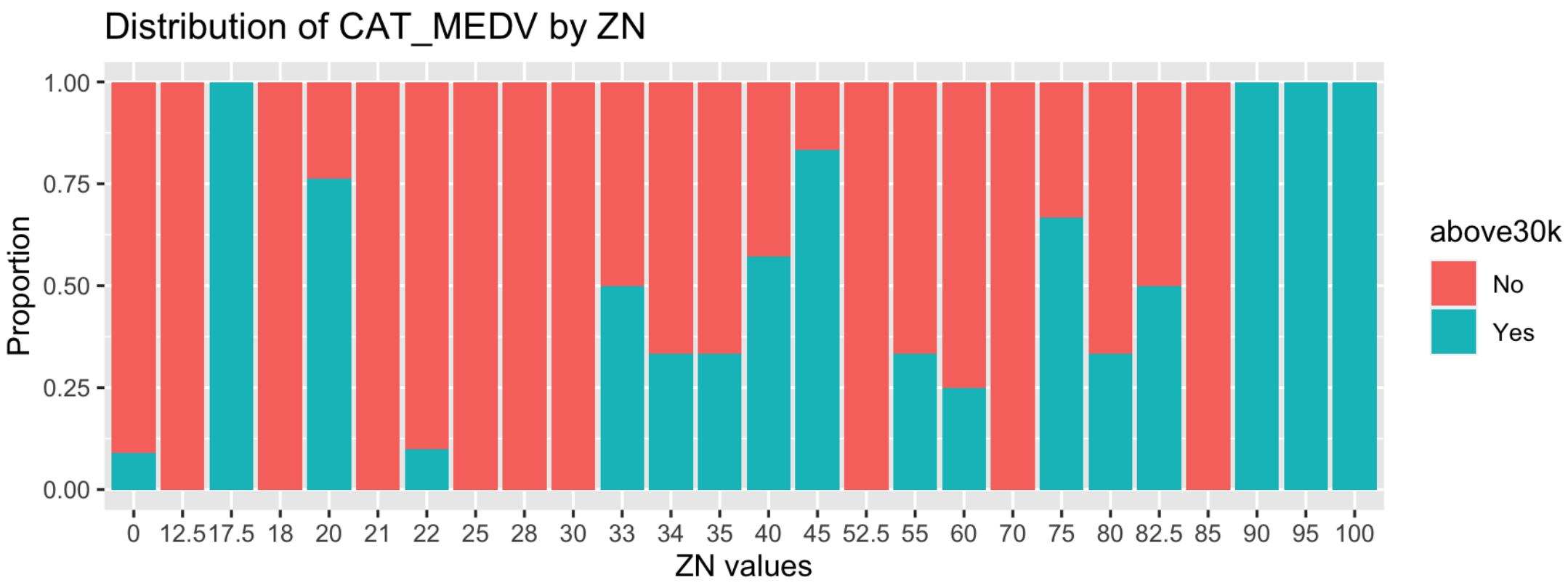
Reducing categories

Less categories R Code

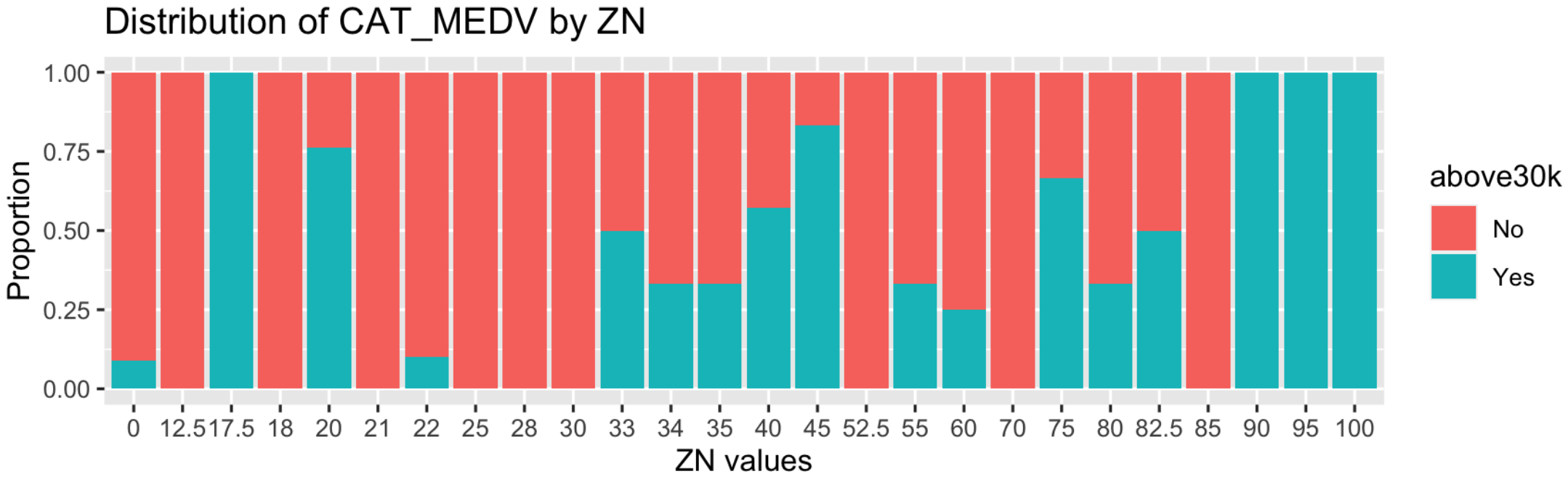
```
## # A tibble: 14 x 5
##    var_name var_mean  var_sd var_median var_miss
##    <chr>       <dbl>   <dbl>      <dbl>    <dbl>
##  1 CRIM       3.61     8.60       0.257        0
##  2 ZN        11.4     23.3        0            0
##  3 INDUS     11.1      6.86       9.69         0
##  4 CHAS       0.0692   0.254      0            0
##  5 NOX        0.555    0.116      0.538        0
##  6 RM         6.28     0.703      6.21         0
##  7 AGE       68.6     28.1       77.5          0
##  8 DIS        3.80     2.11       3.21         0
##  9 RAD        9.55     8.71       5            0
## 10 TAX      408.     169.       330            0
## 11 PTRATIO   18.5      2.16      19.0          0
## 12 LSTAT     12.7      7.14      11.4          0
## 13 MEDV      22.5      9.20      21.2          0
## 14 CAT_MEDV   0.166    0.372      0            0
```

Visualization

Visualization R Code



Visualization (cont.)



- We see that the distribution of outcome variable `CAT_MEDV` is broken down by `ZN` (treated here as a categorical variable).
- We can observe that the distribution of `CAT_MEDV` is identical for `ZN` = 17.5, 90, 95, and 100 (where all neighborhoods have `CAT_MEDV = 1`, that is `above30k = Yes`).
- These four categories can then be combined into a single category. Similarly, categories `ZN` = 12.5, 25, 28, 30, and 70 can be combined.

Categorical to Numerical

Sometimes the categories in a categorical variable represent *intervals*.

Common examples are age group or income bracket.

One approach: If the interval values are known (e.g., category `"2"` is the age interval 20-30), we can replace the categorical value (`"2"` in the example) with the mid-interval value (here "25").

The result will be a numerical variable which no longer requires multiple dummy variables.