Data Analytics Lab 1: Examining data

Learning Objectives

- Revision of basic R commands
- Some interesting R libraries
- Examining the extent of missing data
- Basic plots and tables

1 Starting `R`

To start `R` do the following

- Click on box (9 dots) on bottom left hand corner of screen. Choose Rstudio. If you cannot see it use Search facility

To access datafiles

Data files are stored on BlackBoard. Download them from there.

To Log out

- Choose the symbol at the top right corner of the screen (looks like a wheel).
- Choose Log out

2 The Data

You have been given a description of the Ames data. The object of this lab is to explore the data. You should at this stage know all the basic R commands and how to use scripting.

2.1 Read in Data

The data set SmallAmes is stored on Tholos in the folder $ST4003/get/labs$. It is also on Blackboard. Import the dataset. The easiest way (I think)is to use the Import Dataset command within Rstudio (top left part of screen). Choose **From Text(readr...)** option, Check that it is imported as a data frame as opposed to a *tbl_df*. Use command **class**. To convert to a data frame use the command **as.data.frame**

2.2 Types of Attributes

The **R** command **sapply** is a handy command for determining types of data in **R**. To get a list for a data frame use the following command

```
sapply(<data frame name>,class)
```

See http://www.statmethods.net/input/datatypes.html for a good description of the different data types. The family of **apply** functions are well worth checking out.
See https://www.r-bloggers.com/using-apply-sapply-lapply-in-r/

2.3 Cleaning data

2.3.1 Check for Duplicates

One way to do this is to use command **duplicated**.

```
 chdup<-duplicated(SmallAmes)   or whatever the name data frame is
```

The variable **chdup** is a logical vector (TRUE or FALSE). Use the **table** command to obtain a frequency. The command **unique** saves the unique cases.

There is an interesting library called **janitor** which provides useful functions for cleaning data. There is a vignette on the CRAN website which describes the various functions. The functions include

- Clean dataframe names with **clean_names**
- **tabyl()** an alternative to **table**
- Crosstabulate two variables with **crosstab()**
- Format a crosstab table with **adorn_crosstab()** . (The name crosstab comes from SPSS)
- Useful for finding duplicate values for specific combinations of variables.

I have not used this extensively but I thought it looked very useful.

2.4 Missing data

The simplest starting point is to use the **summary** command

```
summary <dataset name>
```

Check the results carefully. Use the **as.factor** command to change a variable to a categorical variable. Use the **summary** command. What changed? And remember keep asking why why why?

The package **VIM** is also useful and it provides good summaries of the % of missing data for each variable together with pattern of missing data across variables. The command **aggr** provides all this information. I have read the data in a data frame called Ames. Use the command **aggr**

```
oaggr<- aggr(Ames)
summary(oaggr)
```

How do you suggest handling the missing data?

2.5 Near Zero variance variables

We will use the function **nearZeroVar** from the library **caret**. Use the following commands and have a look at output. Remember to load library **caret**. You should have a look at the documentation to see what else you can do.

```
x<-nearZeroVar(Ames,saveMetrics=TRUE)
x
```

- **freqratio** = ratio of highest frequency to the second highest frequency. If it is greater than 19 there may be a problem
- **percentUnique**: Percent unique values; < 10% a problem
- **ZeroVar**: Zero Variance True/False
- **nzv**: True or False according to definition of **freqratio** and **percentUnique**

You can change the defaults i.e. 10% and the 19 mentioned above.

2.6 Exploring data

The next step is examine data preferable using graphs and descriptive statistics. We talked about this in class. Now to help you get back to R produce an appropriate table or graph to illustrate the following:

- Distribution of Sales Prices
- Relationship of Sales Prices with Lot area
- Relationship of Sales price with Type of Dwelling (Bldg Type)
- Relationship of Sales price with Overall condition of House
- Relationship of Sales Price with Availability of Air Conditioning
- Now use the binary version of Sales Price called Salecat and repeat the above analyses.

3 Other useful packages in R

R Book Garrett Grolemund and Hadley Wickham have written a very good book called **R for Data Science**. You will find it online at http://r4ds.had.co.nz/

dplyr This is a very useful package for formatting and describing data.

forcats Good for dealing with factors