HDDA Tutorial: Introduction and R: Solutions

Department of Econometrics and Business Statistics, Monash University

Tutorial 1

Motivation

1. Describe the role that exploratory data analysis plays in the overall statistical analysis of a dataset. Give examples to illustrate your answer.

Preliminary data analysis is very important. It helps identify data problems (such as outliers or other anomalous observations). Moreover, it helps you get a "feel" for your data and also helps to identify potential relationships. For example, boxplots can be used to help identify if spending is different according to gender. A scatterplot or histogram and descriptive statistics can help identify outliers either visually or due to strange ranges and large differences between the mean and median

2. Think of a problem from business or another discipline where the data are multivariate. How may the data be explained by a smaller number of unobserved variables.

One example could be a survey recording customer satisfaction for a hotel. Several survey questions could be asked for instance one may be about the cleanliness of the hotel, another may be about the friendliness of the staff, another about the facilities and another about the breakfast. These four questions may however be able to be reduced to a single variable about the general quality of the hotel.

Measurement

Think of an example of a non-metric variable and metric variable. For each variable answer the following:

1. Is the variable measured on a nominal, ordinal or ratio scale?

Mode of transport (bike, car or train) is an example of a non-metric variable (measured on a nominal scale) while stock return is an example of a metric variable (measured on a ratio scale). 2. What would be a good summary or plot to use to get an idea about this data? *The proportion individuals who use each mode of transport while the mean, mode and median all give some idea of the central tendency of stock returns*

Introduction to R

1. Open R Studio on your workstation and load the pacakge ggplot2. Once you have familiarised yourself with R, install R, R studio and the add-on package ggplot2 on to your own laptop.
2. Load the dataset Beer.rds which can be downloaded from Moodle.

```
Beer<-readRDS('Beer.rds')
```

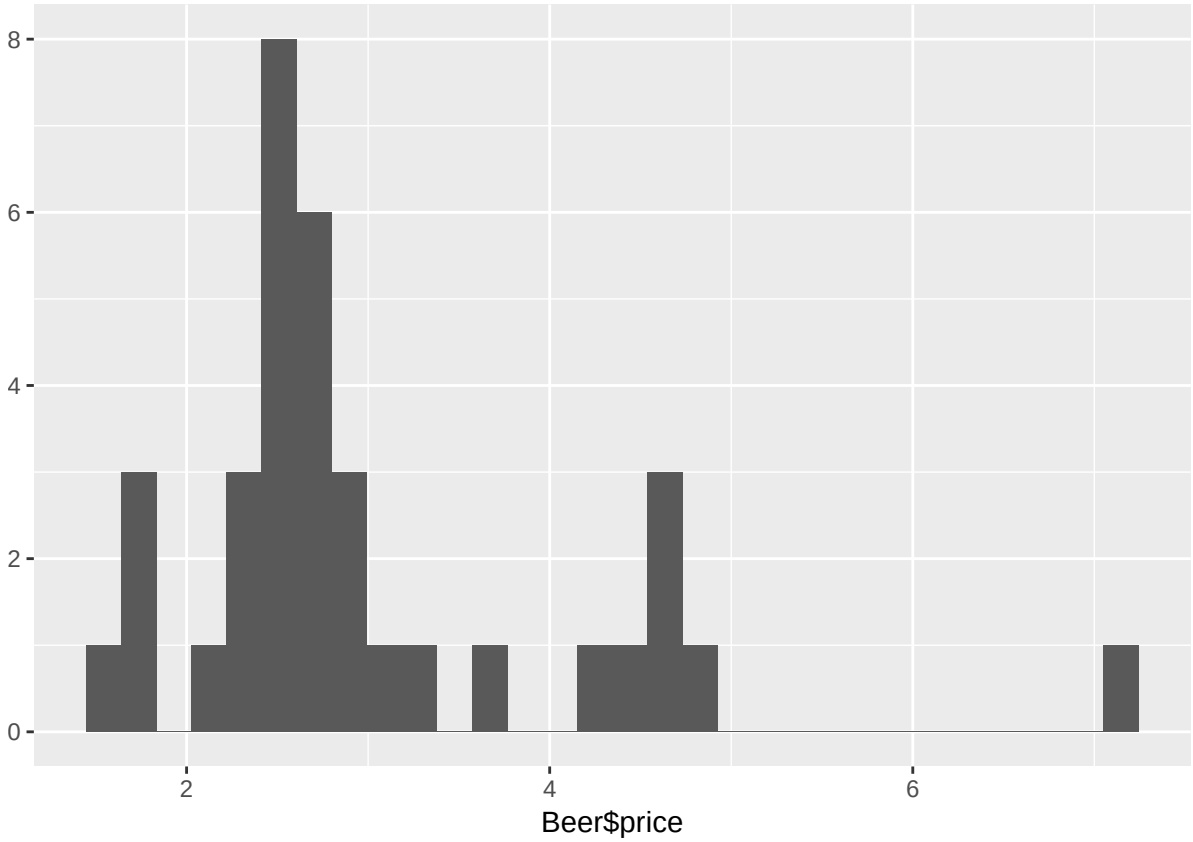
3. Produce a histogram of the price variable. Use the function `qplot` for this.

```
library(ggplot2)
```

```
## Registered S3 methods overwritten by 'ggplot2':
##   method         from
##   [.quosures     rlang
##   c.quosures     rlang
##   print.quosures rlang
```

```
qplot(Beer$price)
```

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



4. Do you identify any outliers in the data?

There is a clear outlier with a price above $7

5. Produce a cross tab of beer rating against origin. Use the function `table` for this.

```
library(ggplot2)
table(Beer$rating,Beer$origin)
```

```
##
##            USA Canada France Holland Mexico Germany Japan
##   VeryGood   7      2      1       1      0       0     0
##   Good      11      0      0       0      1       1     1
##   Fair       9      0      0       0      0       1     0
```