Notes Chapter 1 - Introduction to Data

```
## Warning: package 'ggplot2' was built under R version 3.2.1
```

```
## Please visit openintro.org for free statistics materials
##
## Attaching package: 'openintro'
##
## The following object is masked from 'package:datasets':
##
##     cars
```

Analyses are often a four step process:

1. Identify a question or problem.
2. Collect relevant data on the topic.
3. Analyze the data.
4. Form a conclusion.

The subject of statistics tries to make steps 2-4 objective, rigorous, and efficient.

1.1 Case Study

Summary Statistic: A single number summarizing a large amount of data.

1.2 Data Basics

1.2.1 Observations, Variables, and Data Matrices

A data matrix is a common way to display data. In it a row corresponds to a case (or unit of observation) and each column represents a variable. An example for a data matrix with 21 variables and 50 cases is the email50 dataset as loaded here:

```
email50 <- read.table("data/email50.txt", header = TRUE, sep = "\t")
head(email50)
```

```
##   spam to_multiple from cc sent_email               time image attach
## 1    0           0    1  0          1 2012-01-04 05:19:16     0      0
## 2    0           0    1  0          0 2012-02-16 12:10:06     0      0
## 3    1           0    1  4          0 2012-01-04 07:36:23     0      2
## 4    0           0    1  0          0 2012-01-04 09:49:52     0      0
## 5    0           0    1  0          0 2012-01-27 01:34:45     0      0
## 6    0           0    1  0          0 2012-01-17 09:31:57     0      0
##   dollar winner inherit viagra password num_char line_breaks format
## 1      0     no       0      0        0   21.705         551      1
## 2      0     no       0      0        0    7.011         183      1
## 3      0     no       0      0        0    0.631          28      0
## 4      0     no       0      0        0    2.454          61      0
```

```
## 5      9     no       0      0        1   41.623        1088      1
## 6      0     no       0      0        0    0.057           5      0
##   re_subj exclaim_subj urgent_subj exclaim_mess number
## 1       1            0           0            8  small
## 2       0            0           0            1    big
## 3       0            0           0            2   none
## 4       0            0           0            1  small
## 5       0            0           0           43  small
## 6       0            0           0            0  small
```

1.2.2 Types of Variables

Example data set county:

```
county <- read.table("data/county.txt", header = T, sep = "\t")
county_w_sb <- read.table("data/county_w_sm_ban.txt", header = T, sep = "\t")
county_w_sb_slim <- county_w_sb[,c(1,2,4,5,51,33,26,27,31,32,54)]
head(county_w_sb_slim)
```

```
##               name          state pop2010 pop2000 fed_spending poverty
## 1 Abbeville County South Carolina   25417   26167       169972    20.7
## 2    Acadia Parish      Louisiana   61773   58861       459879    20.1
## 3  Accomack County       Virginia   33164   38305       449275    15.6
## 4       Ada County          Idaho  392365  300904      3122360    10.2
## 5     Adair County       Missouri   25607   24977       203872    26.0
## 6     Adair County       Missouri   25607   24977       203872    26.0
##   home_ownership housing_multi_unit per_capita_income
## 1           77.4               7.7             16653
## 2           69.8               7.1             18116
## 3           74.1               5.6             22766
## 4           69.6              18.0             27915
## 5           61.1              24.3             17098
## 6           61.1              24.3             17098
##   median_household_income smoking_ban
## 1                   33143        none
## 2                   37261     partial
## 3                   41372        none
## 4                   55835     partial
## 5                   31176        none
## 6                   31176        none
```

```
head(county)
```

```
##             name   state pop2000 pop2010 fed_spend poverty homeownership
## 1 Autauga County Alabama   43671   54571  6.068095    10.6          77.5
## 2 Baldwin County Alabama  140415  182265  6.139862    12.2          76.7
## 3 Barbour County Alabama   29038   27457  8.752158    25.0          68.0
## 4    Bibb County Alabama   20826   22915  7.122016    12.6          82.9
## 5  Blount County Alabama   51024   57322  5.130910    13.4          82.0
## 6 Bullock County Alabama   11714   10914  9.973062    25.3          76.9
##   multiunit income med_income
## 1       7.2  24568      53255
```

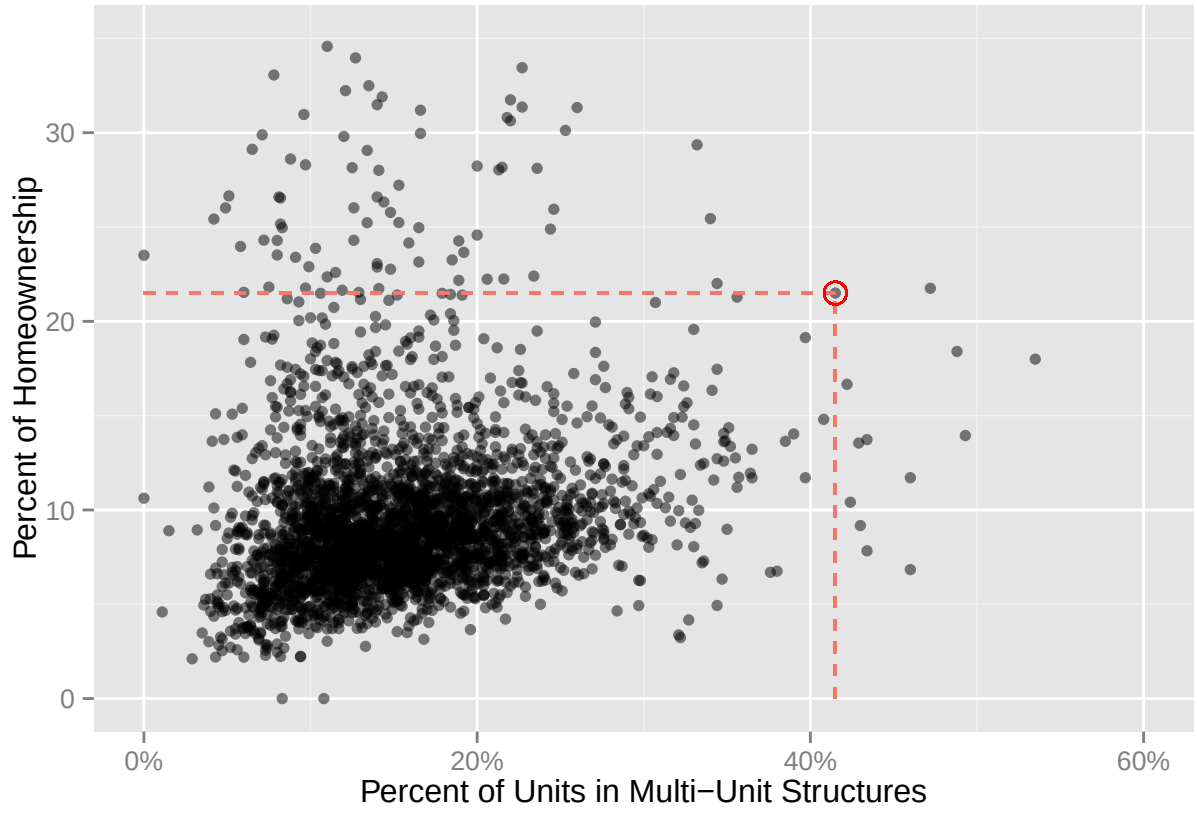
```
## 2      22.6  26469      50147
## 3      11.1  15875      33219
## 4       6.6  19918      41770
## 5       3.7  21070      45549
## 6       9.9  20289      31602
```

Variables can be:

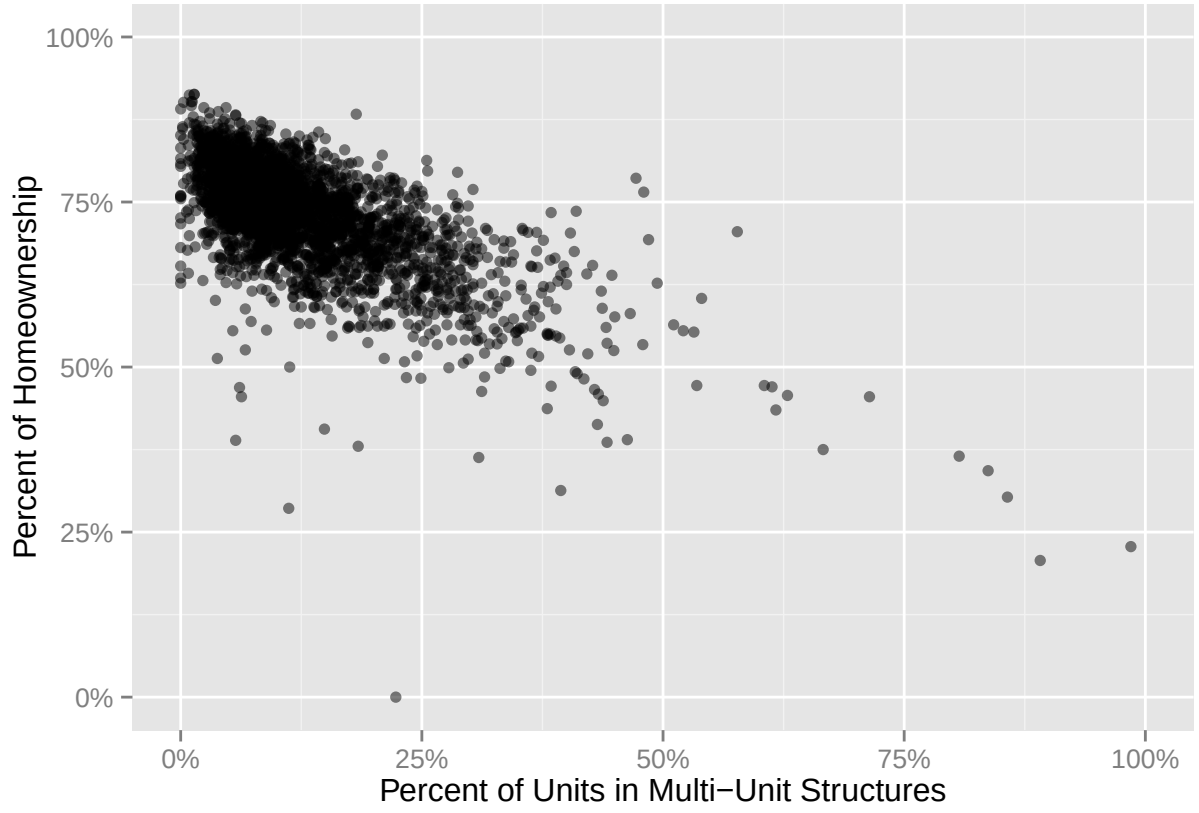
- Numerical: Variable is a number and it makes sense to perform calculations based on it. E.g. fed_spending or pop2010 in the county dataset. As a counter example, zip-codes are not numerical variables because calculations with zip- codes do not make sense. Numerical variables can be split into:
  - Continuous: Variable can take values in a certain range. E.g. fed_spending.
  - Discrete: Variable can take only values of certain steps. E.g. pop2010, a population of 25.34 does not make sense.
- Categorical: Categorical variables take values which categorize observations. E.g. the state or smoking_ban column in the country dataset. Categorical variables can be:
  - Ordinal: These variables have a (natural) ordering.
  - Regular: These variables have no ordering.

1.2.3 Relationship Between Variables

```
## Warning: Removed 32 rows containing missing values (geom_point).
```



Variables can be associated (also called dependent), which means they have some kind of relationship with each other, or independent. Furthermore the association (or dependency) can be positive (see plot above) or negative (see plot below).



1.3 Overview of Data Collection Principles

After identifying the problem upon which research is supposed to be conducted, it has to be considered how data are collected so that they are reliable and useful for research.

1.3.1 Populations and Samples

Example research questions:

1. What is the average mercury content in swordfish in the Atlantic Ocean?
2. Over the last 5 years, what is the average time to complete a degree for Duke undergraduate students?
3. Does a new drug reduce the number of deaths in patients with severe heart disease?

Research question often refer to a target population. These populations are in many cases so large that it is too expensive to collect data for every case. Instead a sample is taken, representing an (often small) subset of the population.

1.3.2 Anecdotal Evidence

Anecdotal Evidence is evidence that is based on too few cases (often low single digit number) and which are often not representative of the population. We are more likely to remember unusual or personal cases, which is why these often cloud our judgement. Instead we should collect sufficient, representative data and use it to answer a research question in a sound way.

Example answers to the questions above based on anecdotal evidence:

1. A man on the news got mercury poisoning from eating swordfish, so the average mercury concentration in swordfish must be dangerously high.
2. I met two students who took more than 7 years to graduate from Duke, so it must take longer to graduate at Duke than at many other colleges.
3. My friend's dad had a heart attack and died after they gave him a new heart disease drug, so the drug must not work.

1.3.3 Sampling from a Population

Unless impossible, samples should always be choosen **randomly** from a population. When cases are selected manually, there is a chance of introducing **bias** into the sample. A basic way of sampling randomly is called **simple random sample**, which is similar to a raffle lottery. Here each case has the same chance of being included and there is no connection between the cases in the sample.

Another type of bias is the **non-response bias**, which is very common in surveys. It happens when a certain part of the population is more likely to respond to the survey. Especially when response rates are low, one must be cautious.

A third common bias is the one of having a **convenience sample**. Here cases that are easily accessible are more likely to be included in the sample. This could happen when a survey is taken only is a small part of the area for which it is supposed to be.

1.3.4 Explanatory and Response Variables

An **explanatory variable** is one that might affect the **response variable**. In datasets with many variables there are often many of these relationships. Other terms are **independent variable** and **dependent variable** for the explanatory (does not depend on the response variable) and the response variable respectively.

E.g. in the county dataset one might say that the poverty rate is the explanatory variable for federal spending.

Caution: Just because two variables are associated there is not necessarily a **causal relationship** between them. **Association does not imply causation.**

Also even if there is association between two variables, the direction might not be clear. One example for this is the homeownership rate and the percentage of multi-unit structures in the county dataset.

1.3.5 Introducing Observational Studies and Experiments

Observational Studies and Experiments are the two primary types of data collection.

In **observational studies** data is collected in a way that does not interfere with how the data arises. E.g. medical records or already existing company records. Generally observational studies can only provide evidence of a naturally occuring association, while they can not show causal connection.

In order to prove causality an **experiment** needs to be conducted. Here explanatory and response variables are defined upfront. Afterwards the sample is split into groups. If this is done randomly we speak of a

randomized experiment. Each group is then assigned a different input for the explanatory variables. E.g. when researching a medical treatment, one group could get the treatment while the other group gets a placebo.

1.4 Observational Studies and Sampling Strategies

1.4.1 Observational Studies

Data in observational studies are collected by monitoring what occurs. This allows the researcher to show association, but it is dangerous, and therefore not recommended, to use them to infer causality. For this an experiment is usually required.

Confounding variables, which are correlated to both the explanatory and the response variable, play a great role in observational studies. E.g. sunscreen usage and skin cancer are associated variables, however one does not cause the other. A confounding variable in this case is sun exposure.

Confounding variables are a reason why observational studies can not prove causality well. In order to do so one must exhaust the search for confounding variables, which is very hard to do.

Confounding variables are also called **lurking variables**, **confounder**, or **confounding factor**.

Observational studies come in two forms:

- **Prospective**: Individuals/cases are identified and data is collected as events unfold. E.g. observing a group of similar individuals over time to see if certain aspects of their behaviour influence cancer risk.
- **Retrospective**: Data is collected after events have taken place. E.g. medical records.

1.4.2 Three Sampling Methods

Simple random sampling is the most intuitive form of random sampling. All cases are randomly drawn from the total population. This is comparable to a raffle lottery. In general simple random sampling is characterized by each case having the same probability of being included in the sample and by there being no relationship between the cases selected and not selected to be included in the sample.

In **stratified sampling** the population is divided into groups of cases, called _strata___, which are similar to each other with respect to the outcome of interest. Then from each stratum a number of cases are choosen by simple random sampling. This technique is very useful if the cases within each stratum are very similar to each other. However the statistical methods described here need to be extended to be used on data collected using stratified sampling.

In **cluster sampling** the population is divided into groups of cases called **clusters**. Afterwards a fixed number of cases are sampled from some clusters via simple random sampling. It is not required to sample from all clusters. Cluster sampling is most helpful when there is a lot of case-by-case variability withing the clusters, but the clusters themselves are very similar to each other. It is often a more economical sampling method than the other two. However the methods in this book need to be extended before the can be used on data collected via cluster sampling.

1.5 Experiments

Studies in which researchers assign treatments to cases are called **experiments**. When this assignment is done in a randomized way, the study is a **randomized experiment**, which are fundamentally important when trying to show a causal connection between variables.

1.5.1 Principles of Experiment Design

Controlling: After assigning treatments to cases, the researcher must do his best to control any other factors that might influence the outcome of the observed variable. It is essential to track all these other factors in variables.

Randomization: By assigning cases to the treatment and control group in a randomized way, the researcher not only avoids bias, he also accounts for variables that can not be controlled by making sure they affect both groups equally.

Replication: The more cases are included in a study, the more accurately the effect between explanatory and response variable can be estimated. In addition to replicating cases within a study (by simply having many of them), a whole study can be replicated by another researcher to varify findings.

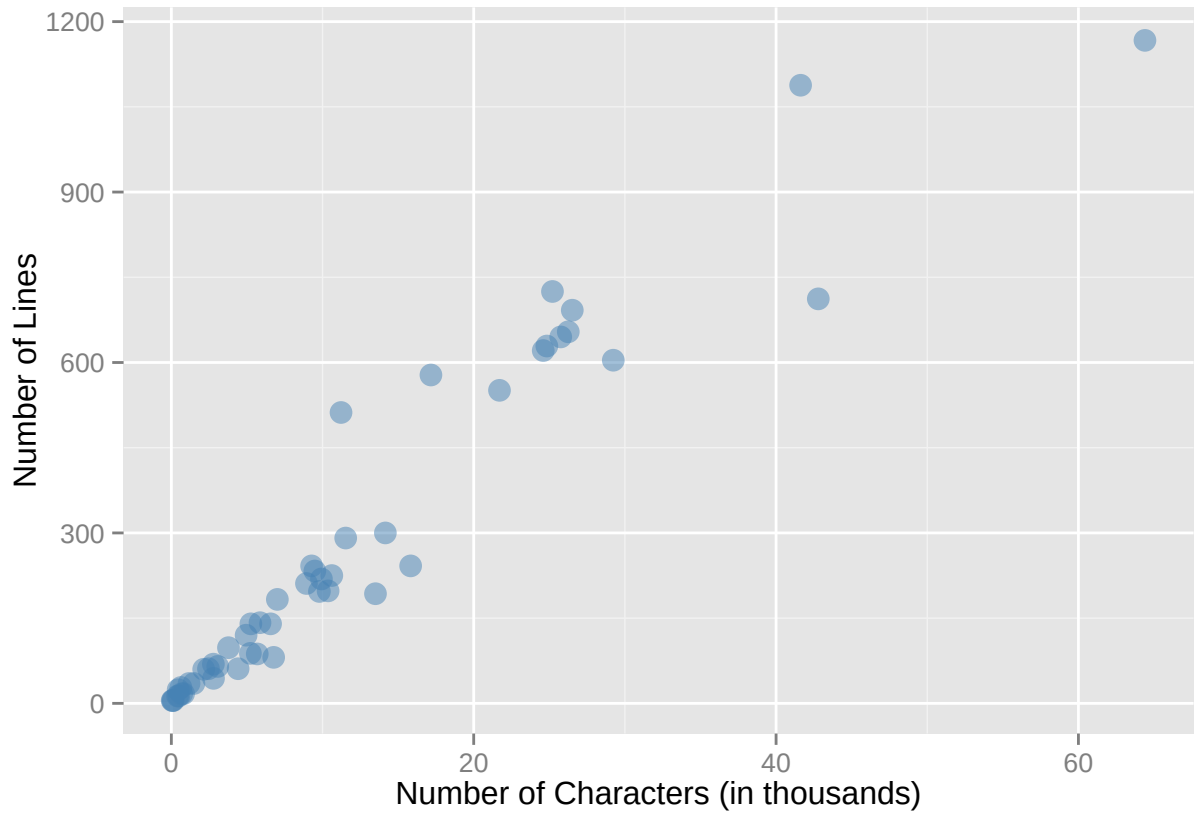
Blocking: When the researcher suspects that uncontrollable variables affect the outcome of the study, he may first group cases into **blocks**. Afterwards he randomizes for each block seperately. The methods in this book need to first be extended to analyze data from experiments that use blocking.

1.5.2 Reducing Bias in Human Experiments

While the researcher must always do his best to aviod bias in a study it often arises unintentionally. This is especially common in studies involving humans in any way. One way to avoid unintentional bias is to design a **blind** study, in which the patient is unaware about which treatment he receives. To aid this a **placebo** can be used. In some cases slight but real improvements happen for cases that receive the placebo (e.g. because of emotional effects). This is called the **placebo effect**. Because bias can also be introduced by the person administrating the study he should also be unaware of which case belongs to which group. These studies are then **double-blind**.

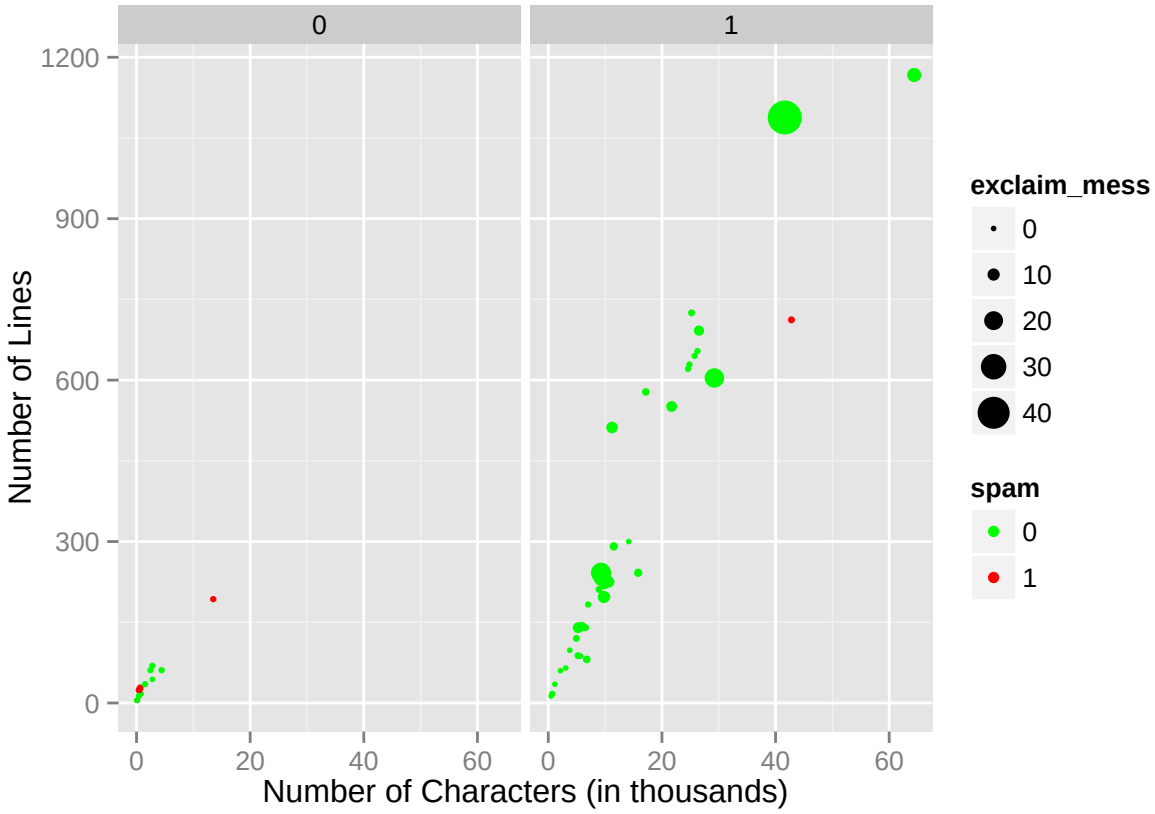
1.6 Examining Numerical Data

1.6.1 Scatterplots for Paired Data



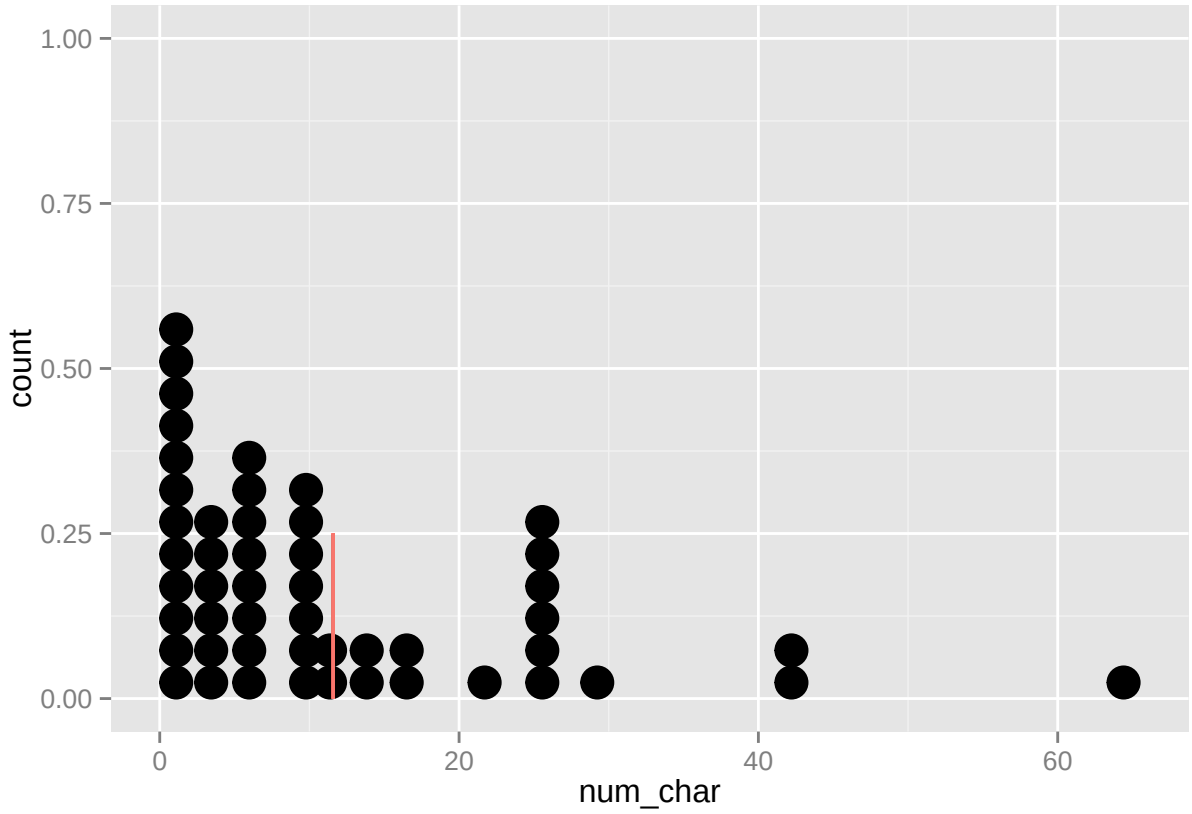
This **scatterplot** visualizes the relationship between two variables from the email50 dataset. Scatterplots are helpful when illustrating association between variables. However they can not show causality. Also they might miss relationships. E.g. in the email50 dataset, many of the more verbose mails are HTML, which increases the number of characters without increasing the amount of content.

Scatterplots can be enhanced in many ways. E.g. the plots below adds three more variables, by varying color, size, and facet by the spam, exclaim_mess, and format variables.



1.6.2 Dot Plots and the Mean

```
## stat_bindot: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



A **dot plot** is essentially a one variable scatterplot. It is useful to show the **mean** (average) of the data. The mean of the sample data is called **sample mean**, usually denoted by $\bar{x}$ (x_bar), and computed as follows:

$$\bar{x} = \frac{\sum(x_1 + x_2 + ... + x_n)}{n}$$

Here **n** is the number of observations (exlcuding NAs) and $\mathbf{x_i}$ are the values observed for each variable. The equivalent calculation in R:

```
sum(email50$num_char) / length(email50$num_char)
```

```
## [1] 11.59822
```

```
mean(email50$num_char)
```

```
## [1] 11.59822
```

The mean of the populations, the **population mean** is represented by the greek letter mu: μ. A subscript can be added to show which variable it refers to: $\mu_{\mathbf{x}}$.

It is often important to calculate the **weighted mean**, especially when using data that is already aggregated. Instead of dividing by the number of observations the weighted mean divides by the sum of the weights w_i that each variable has:

$$\bar{x} = \frac{\sum(x_1 * w_1 + x_2 * w_2 + ... + x_n * w_n)}{\sum(w_1 + w_2 + ... + w_n)}$$

E.g. for the county dataset one could compute the mean income per county as the normal mean. However for the average income per person in the US, a weighted mean needs to be used.

```
county <- read.table("data/county.txt", header = T, sep = "\t")
mean(county$income)
```

```
## [1] 22504.7
```

```
weighted.mean(county$income, county$pop2010)
```

```
## [1] 27348.43
```

```
sum(as.double(county$income) * as.double(county$pop2010)) / sum(county$pop2010)
```

```
## [1] 27348.43
```