

LAB 1

Objectives

- How to install R and Rstudio
- In this lab we will explore the data using the `dplyr` package and visualize it using the `ggplot2` package for data visualization. The data can be found in the companion package for this lab (`statsr`).
- Insert a population data and summarize its statistics.

1) How to install R and Rstudio

- simply follow the following steps to download R (<https://cran.r-project.org/bin/windows/base/>)

2

Google search for "r download" results in the CRAN website for Windows.

Download R 4.1.1 for Windows - CRAN - The R Project for ...

R-4.1.1 for Windows (32/64 bit) ... If you want to double-check that the package you have downloaded matches the package distributed by CRAN, you can compare the ...

R-patched snapshot build · R-devel snapshot build · News

You've visited this page 3 times. Last visit: 9/29/21

https://www.r-project.org · The R Project for Statistical Computing

To download R, please choose your preferred CRAN mirror. If you have questions about R like how to download and install the software, or what the license terms ...

https://cran.r-project.org · bin · windows · R for Windows

3

R-4.1.1 for Windows (32/64 bit)

Download R 4.1.1 for Windows (86 megabytes, 32/64 bit)

Installation and other instructions

New features in this version

If you want to double-check that the package you have downloaded matches the package distributed by CRAN, you can compare the [md5sum](#) of the .exe to the [fingerprint](#) on the master server. You will need a version of md5sum for windows: both [graphical](#) and [command line versions](#) are available.

Frequently asked questions

- Does R run under my version of Windows?
- How do I update packages in my previous version of R?
- Should I run 32-bit or 64-bit R?

Please see the [R FAQ](#) for general information about R and the [R Windows FAQ](#) for Windows-specific information.

Other builds

- Patches to this release are incorporated in the [r-patched snapshot build](#)
- A build of the development version (which will eventually become the next major release of R) is available in the [r-devel snapshot build](#)
- [Previous releases](#)

Note to webmasters: A stable link which will redirect to the current Windows binary release is [<CRANMIRROR>-bin/windows/base/release.html](#).

Last change: 2021-08-10

- simply follow the following steps to download RStudio (<https://www.rstudio.com/products/rstudio/download/>)

4

Google search for "rstudio download" results in the RStudio website.

Download the RStudio IDE

RStudio is a set of integrated tools designed to help you be more productive with R. It includes a console, syntax-highlighting editor that supports direct ...

You visited this page on 9/29/21.

Download RStudio Desktop Pro · Older Versions of RStudio

RStudio requires R 3.0.1 (or higher). If you don't already have R ...

More results from rstudio.com »

https://www.rstudio.com · products · rstudio · RStudio - RStudio

5

R Studio

Products · Solutions · Customers · Resources · About · Pricing

Download the RStudio IDE

Choose Your Version

The RStudio IDE is a set of integrated tools designed to help you be more productive with R and Python. It includes a console, syntax-highlighting editor that supports direct code execution, and a variety of robust tools for plotting, viewing history, debugging and managing your workspace.

RStudio Team

RStudio's recommended professional data science solution for every team. RStudio Team is a bundle of RStudio's popular ...

RStudio Desktop	RStudio Desktop Pro	RStudio Server	RStudio Workbench
Open Source License	Commercial License	Open Source License	Commercial License
Free	\$995 /year	Free	\$4,975 /year (5 Named Users)
DOWNLOAD	BUY	DOWNLOAD	BUY
Learn more	Learn more	Learn more	Evaluation Learn more
✓	✓	✓	✓

6

STEP 1: the step's objective is to install the required packages

```
install.packages("dplyr", dependencies = TRUE)
install.packages("ggplot2", dependencies = TRUE)
install_github("StatsWithR/statsr", dependencies = TRUE)
```

STEP 2: the step's objective is to load the required packages

```
library(statsr)
library(dplyr)
library(ggplot2)
```

STEP 3: the step's objective is to load Population data

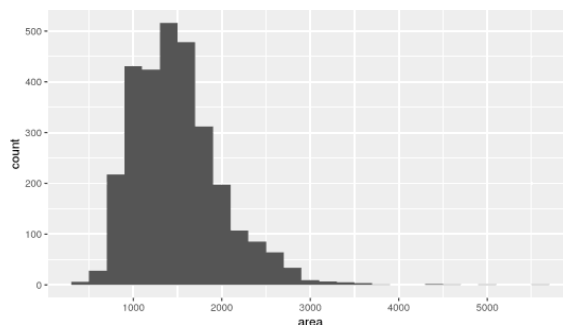
- We consider real estate data from the city of Ames, Iowa. The details of every real estate transaction in Ames is recorded by the City Assessor's office. Our particular focus for this lab will be all residential home sales in Ames between 2006 and 2010. This collection represents our population of interest. In this lab we would like to learn about these home sales by taking smaller samples from the full population. Let's load the data.

```
data(ames)
```

- We see that there are quite a few variables in the data set, enough to do a very in-depth analysis. For this lab, we'll restrict our attention to just two of the variables: the above ground living area of the house in square feet (`area`) and the sale price (`price`).
- We can explore the distribution of areas of homes in the population of home sales visually and with summary statistics. Let's first create a visualization, a histogram:

STEP 4: the step's objective is to summarize the statistics of population

```
ggplot(data = ames, aes(x = area)) +
  geom_histogram(binwidth = 200)
```



Let's also obtain some summary statistics. Note that we can do this using the `summarise` function. We can calculate as many statistics as we want using this function, and just string along the results. Some of the functions below should be self explanatory (like `mean`, `median`, `sd`, `IQR`, `min`, and `max`). A new function here is the `quantile` function which we can use to calculate values corresponding to specific percentile cutoffs in the distribution. For example `quantile(x, 0.25)` will yield the cutoff value for the 25th percentile (Q1) in the distribution of `x`. Finding these values are useful for describing the distribution, as we can use them for descriptions like *"the middle 50% of the homes have areas between such and such square feet"*.

```
ames %>%
  summarise(mu = mean(area), pop_med = median(area),
    sigma = sd(area), pop_iqr = IQR(area),
    pop_min = min(area), pop_max = max(area),
    pop_q1 = quantile(area, 0.25), # first quartile, 25th percentile
    pop_q3 = quantile(area, 0.75)) # third quartile, 75th percentile
```

```
# A tibble: 1 x 8
      mu pop_med sigma pop_iqr pop_min pop_max pop_q1 pop_q3
  <dbl>   <dbl> <dbl>   <dbl>   <int>   <int>   <dbl>   <dbl>
1 1500.    1442   506.    617.    334    5642    1126    1743.
```

Discussion

Which of the following is **false**?

1. The distribution of areas of houses in Ames is unimodal and right-skewed. **TRUE**
2. 50% of houses in Ames are smaller than 1,499.69 square feet. **FALSE**
3. The middle 50% of the houses range between approximately 1,126 square feet and 1,742.7 square feet. **TRUE**
4. The IQR is approximately 616.7 square feet. **TRUE**
5. The smallest house is 334 square feet and the largest is 5,642 square feet. **TRUE**

STEP 5: the step's objective is to take a random sample from the population

- In this lab we have access to the entire population, but this is rarely the case in real life. Gathering information on an entire population is often extremely costly or impossible. Because of this, we often take a sample of the population and use that to understand the properties of the population.
- If we were interested in estimating the mean living area in Ames based on a sample, we can use the following command to survey the population.

```
sampl <- ames %>%
  sample_n(size = 50)
```

- This command collects a simple random sample of `size 50` from the `ames` dataset, which is assigned to `sampl`. This is like going into the City Assessor's database and pulling up the files on 50 random home sales. Working with these 50 files would be considerably simpler than working with all 2930 home sales. n sale price of homes in Ames?