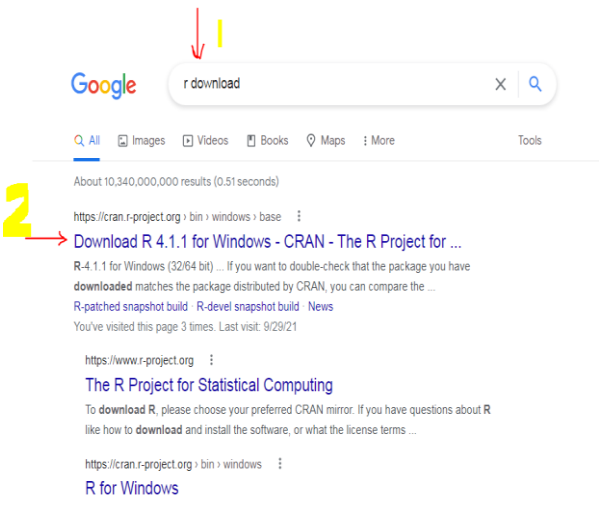
LAB 1

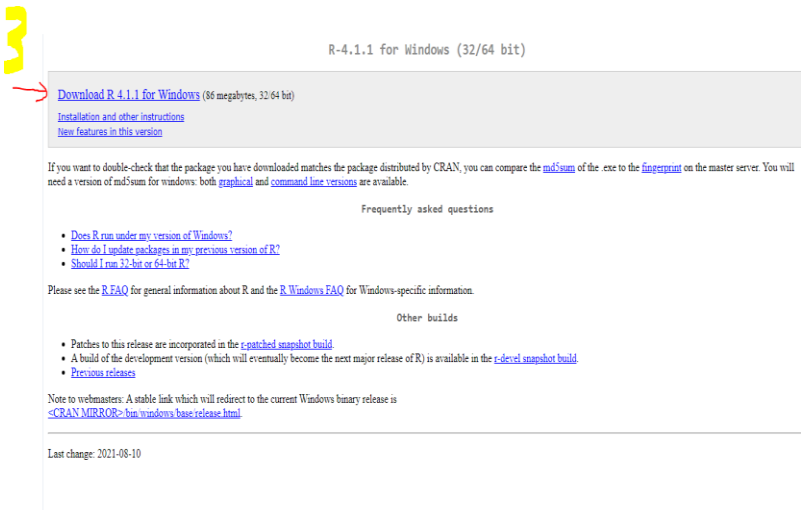
Objectives

- 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.

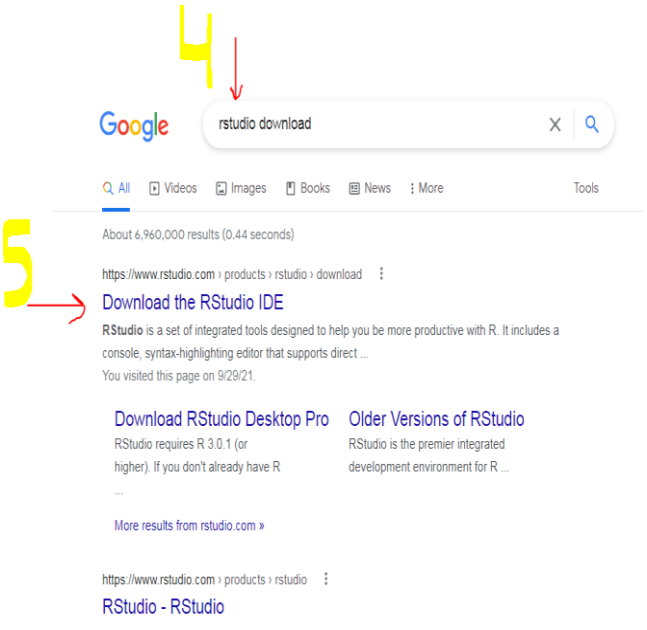
1) How to install R and Rstudio

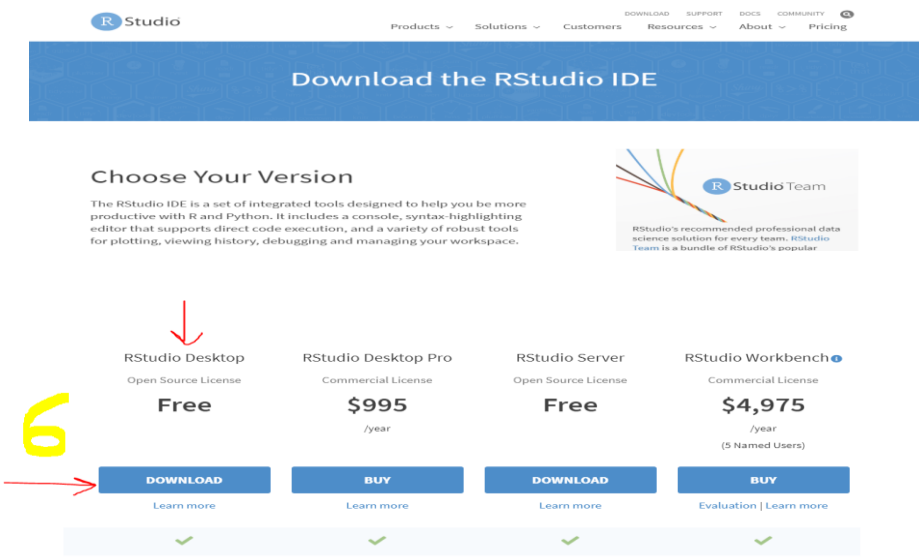
- simply follow the following steps to download R (https://cran.r-project.org/bin/windows/base/)





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





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

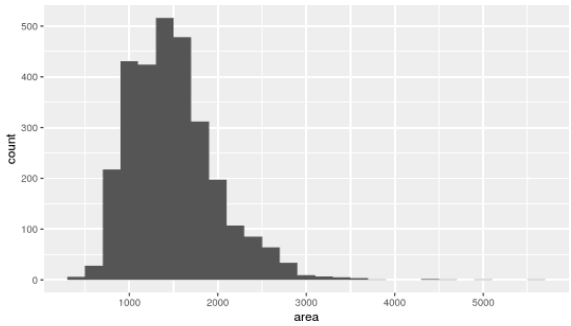
- We consider real estate data from the city of Ames, Iowa. The details of every real estate transaction in Ames is recorded by the City Assessor's office. Our particular focus for this lab will be all residential home sales in Ames between 2006 and 2010. This collection represents our population of 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 form 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 `samp1`. 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?