R – Udacity

**Exploratory Data Analysis with Chris Saden and Facebook Data Scientists**

In this course, you'll be learning explore data using R, a powerful and useful language for statistical analysis. You'll learn to explore data quantitatively and visually, you'll hear about example analyses done at Facebook, and you'll build proficiency with R.

Chris Saden will be your instructor, along with a team of data scientists at Facebook. Chris is a Full Stack Software Engineer at Udacity. He brings his experience of teaching high school mathematics to online learning.

**[Facebook processes more than 500 terabyte of data a day (2012)](http://news.cnet.com/8301-1023_3-57498531-93/facebook-processes-more-than-500-tb-of-data-daily/" \t "_blank)**

One of Facebook's tools, **[Presto](http://prestodb.io/" \t "_blank)** (mainly used for adhoc analysis), processes over 1 petabyte of data per day.

Learn more about **[salmonella](http://www.cdc.gov/salmonella/" \t "_blank)**.

**[Link to Chris's search](http://www.google.com/trends/explore" \l "q=chicken%2C%20music%2C%20movies&cmpt=q" \t "_blank)** on Google Trends.

Response to Chris’s graph:

interest in chicken has been consistent; interest in movies has risen at slow and steady pace; interest in music is declining

Thanks for your thoughts! There's so much to say here. See what Chris noticed in the solution video.

[**VIEW INTRO**](https://classroom.udacity.com/nanodegrees/nd002/parts/9f3e2782-9409-4c7a-b3d7-6197edfee54e/modules/67c46ff5-5a70-402e-881b-558d374953c6/lessons/6261e7b3-3e2c-465d-b066-7ade56e8f93a/concepts/8245785490923)

**[Google Trends](http://www.google.com/trends/" \t "_blank)**

**[Google Flu Trends:](http://gking.harvard.edu/publications/parable-google-flu%C2%A0traps-big-data-analysis" \t "_blank)** Limitations of Predicting the Flu with Data

**[Google Trends](http://www.google.com/trends/" \t "_blank)**   
  
Share your search by starting a discussion in the forum (see the resources tab) and upvote your favorite ones! Students enrolled in the Nanodegree program can also share their findings in the Nanodegree program's Slack channel.

Carnatic vocal music :got no trend by regions

Changed to Carnatic music: then got good summary

# Thanks for completing that!

That's alright. You may not have a sense of adventure now. We hope this course sparks that feeling for you.

Chris mentions the forum here, which are the discussions forums or discussions. The forums are linked in the resources tab. Students enrolled in the Nanodegree program can also share their results in the Slack channel for the program.

**[What is the Netflix Prize?](http://en.wikipedia.org/wiki/Netflix_Prize" \t "_blank)**   
  
**[Netflix Prize Dataset Visualization](http://flowingdata.com/2007/12/11/netflix-prize-dataset-visualization/" \t "_blank)**

Why do you want to learn EDA?

Gain insights about my blog posts, how they are doing, my readers, etc. I find SEO and Analytics are interlinked, so, this learning to analyze my personal data would be greatly beneficial.

Thank you for your response! We'll post a word cloud of many responses at a later date.

**[Exporatory Data Anaylsis](http://en.wikipedia.org/wiki/Exploratory_data_analysis" \t "_blank)**

At 1:00 Chris discusses assessing and validating assumptions. Here are some more assumptions we might consider. For one, we should expect that mpg is in a certain range, so we could validate that assumption. Or we might wonder how precisely hp is measured or reported here. (Is it always an integer?)

Nathan Yau's Flowing Data: **[The Growth of Televisions](http://flowingdata.com/2009/09/23/tv-size-over-the-past-8-years/" \t "_blank)**

Quiz Qs:

Median TV size in 2006:

Largest TV size increases happened from \_\_ to \_\_\_ :

What else stands out inNathan’s work?What connections are there to his

**[The R Programming Language](http://en.wikipedia.org/wiki/R_(programming_language)" \t "_blank)**

**[Foodbourne Chicago finds Dodgy Restaurants with Tweets](http://blog.revolutionanalytics.com/2013/08/foodborne-chicago.html" \t "_blank)**   
  
**[Sensitivity and Specificity and ROC Curves for Binary Classifiers](http://blog.yhathq.com/posts/roc-curves.html" \t "_blank)** (more advanced reading; focus on understanding TPR and FPR).   
  
**[What's a tweet?](http://en.wikipedia.org/wiki/Twitter" \t "_blank)**

Read about **[Github](http://en.wikipedia.org/wiki/GitHub" \t "_blank)** on Wikipedia.   
  
**[Try Github](https://github.com/" \t "_blank)** for yourself and start contributing to open-source projects.

Quiz:

R package used by Corey – textcat

How many tweets did the system flag? 10 to 20

What did you find most interesting about the article? Great combination of research and application, commendable balancing of constraints

To install ggplot2 and RColorBrewer, you can run the following lines of code in RStudio.

**install**.packages('ggplot2', dependencies = TRUE) **install**.packages('RColorBrewer', dependencies = TRUE) # **use** **library**() **to** **load** packages once they're installed! library(ggplot2) library(RColorBrewer)

If you are using a Mac, skip the next video which shows how to install R and RStudio on a PC. The second installation video is for you!

If you are using Linux, visit **[cran.rstudio.com](http://cran.rstudio.com/" \t "_blank)** to learn how to install the programming language R. Then, go to **[rstudio.com](http://rstudio.com/" \t "_blank)** to install RStudio.

Download and install the R programming language **FIRST** at **[http://cran.rstudio.com](http://cran.rstudio.com/" \t "_blank)**.   
  
After you install R, you can download and install RStudio from **[http://www.rstudio.com](http://www.rstudio.com/" \t "_blank)**.   
  
**There have been a few reports of the latest version of Rstudio failing to initialize. If you encounter these issue you can try using an older stable version:**

**[https://s3.amazonaws.com/rstudio-dailybuilds/RStudio-0.98.944.dmg](https://s3.amazonaws.com/rstudio-dailybuilds/RStudio-0.98.944.dmg" \t "_blank)**

If you have trouble installing RStudio, please google your problem first to get help. After that, we recommend posting a discussion. Be as specific as possible with your problem and include the current version of your operating system.

**We do not recommend installing R and RStudio using Homebrew or MacPorts**.

Installed R and RStudio? Yes

Great! It's time to learn some basics of R.   
  
There is SO MUCH to learn about R, and we won't have time to cover it all.   
  
We'll show you many of the basics when it comes to loading, subsetting, and visualizing data so you can perform EDA!

All four panels are configurable. The layout discussed is the default layout for RStudio.   
You can change the default layout by going to Tools -> Options…

This quiz assumes the default layout.

**[A Beginner's Guide to R: Introduction](http://www.computerworld.com/article/2497143/business-intelligence-beginner-s-guide-to-r-introduction.html?null" \t "_blank)** by Sharon Machlis

## [Quick Guide to R Layout](http://dss.princeton.edu/training/RStudio101.pdf" \t "_blank) by Oscar Torres-Reyna.

Quiz clarification: "Clear the workspace" means removing all variables from the current working memory.

We recommend anyone new to R and RStudio to try **Swirl** (statistics with interactive R learning). Swirl is a software package for the R statistical programming language. Its purpose is to teach statistics and R commands interactively.

Type the following commands in the **Console**, pressing Enter or Return after each line:

install.packages("swirl") **library**(swirl) **swirl**()

Note that the > symbol at the beginning of the line is R's prompt for you type something into the console. We include it here so you know that the above commands are to be typed into the console and not elsewhere. The part you type begins after >. 

### Download the [EDA Course Materials zip file](https://d17h27t6h515a5.cloudfront.net/topher/2016/November/581a4977_eda-course-materials/eda-course-materials.zip" \t "_blank) to get all of the files used in this course.

### OR

### Download the file, [demystifying.R](https://d17h27t6h515a5.cloudfront.net/topher/2016/October/57fb4aa7_demystifying/demystifying.r" \t "_blank), that Chris mentions in this video.

Swirl commands:

| You can exit swirl and return to the R prompt (>) at any time by pressing the Esc key. If you are already at the prompt,

| type bye() to exit and save your progress. When you exit properly, you'll see a short message letting you know you've done

| so.

| When you are at the R prompt (>):

| -- Typing skip() allows you to skip the current question.

| -- Typing play() lets you experiment with R on your own; swirl will ignore what you do...

| -- UNTIL you type nxt() which will regain swirl's attention.

| -- Typing bye() causes swirl to exit. Your progress will be saved.

| -- Typing main() returns you to swirl's main menu.

| -- Typing info() displays these options again.

Links to learn R

**[Quick R](http://www.statmethods.net/" \t "_blank)**   
**[R Cookbook](http://www.cookbook-r.com/" \t "_blank)**   
**[R-Bloggers](http://www.r-bloggers.com/" \t "_blank)**   
**[StackOverflow About R](http://stackoverflow.com/tags/r/info" \t "_blank)**   
**[StackOverflow R FAQ](http://stackoverflow.com/questions/tagged/r-faq%20" \t "_blank)**   
**[Google's R Style Guide](https://google.github.io/styleguide/Rguide.xml" \t "_blank)**

Read this **[fantastic tutorial](http://flowingdata.com/2015/02/18/loading-data-and-basic-formatting-in-r/" \t "_blank)** on loading data into R by Nathan Yau.

**[Download the stateData.csv](https://s3.amazonaws.com/udacity-hosted-downloads/ud651/stateData.csv" \t "_blank)**   
  
**[How to Subset Data](http://www.statmethods.net/management/subset.html" \t "_blank)**

Follow along with Chris in this video using the RMD file, **[What\_is\_a\_RMD\_file](https://s3.amazonaws.com/udacity-hosted-downloads/ud651/What_is_a_RMD_file.Rmd" \t "_blank)**.   
  
The second file, **[demystifyingR2.Rmd](https://d17h27t6h515a5.cloudfront.net/topher/2016/October/57fb4c3b_demystifyingr2-v3/demystifyingr2-v3.rmd" \t "_blank)**, is the file that Chris mentions at the end of the video. Work through this file and use it to answer the next quiz!   
  
There is a difference between using the hash or pound symbol# inside and outside of a {r} block of code.

```{r} # The hash or pound symbol inside the r block creates # a comment. These three lines of are not code and cannot be # executed. x <- [1:10] mean(x) ```

If you use the # symbol outside of an {r} block of code, you can create headers of text.   
  
**[Markdown Tutorial](https://www.youtube.com/watch?v=6A5EpqqDOdk" \t "_blank)**: You can also use Markdown to format your posts in the discussions.   
  
You'll need to install and load the package knitr in order to use the **KNIT HTML** button . Run the following commands in the RStudio console to install and load knitr.

**install**.packages('knitr', dependencies = **T**) **library**(knitr)

Download the **[Reddit Survey Responses](https://s3.amazonaws.com/udacity-hosted-downloads/ud651/reddit.csv" \t "_blank)**   
  
Learn more about **[data types](http://www.statmethods.net/input/datatypes.html" \t "_blank)**.

**[Who is Reddit?](http://www.redditblog.com/2011/09/who-in-world-is-reddit-results-are-in.html" \t "_blank)**

Download the **[Reddit Survey Responses](https://s3.amazonaws.com/udacity-hosted-downloads/ud651/reddit.csv" \t "_blank)**   
  
To run the qplot function you must have the ggplot2 library installed and loaded. You can do that by running these two lines of code.

**install**.packages('ggplot2', dependencies = **T**) **library**(ggplot2)

For those on Mac OS X 10.9 (or more recent OS X updates), the above line of code may display an error in the Console and fail to load ggplot2. This is due to a recent software update deleting the X11 symlink. You will need to *reinstall* XQuartz from [**http://xquartz.macosforge.org/**](http://xquartz.macosforge.org/) and then run the following lines of code:

**install**.packages('devtools', dependencies = **T**) **library**(devtools) install\_version("colorspace","1.2-4")

You will then be able to load ggplot2 in RStudio. See the following stackoverflow links for reference [**here**](http://stackoverflow.com/questions/28984243/ggplot2-sourcing-error-x11-library-is-missing) and [**here**](http://stackoverflow.com/questions/28952128/loading-ggplot2-opens-up-x11).

At 0:27 of the solution video, the Instructor misspeaks. Cheese is not binary for the reddit data set. It is a categorical variable with 11 levels. We wouldn't rank Brie over Monterey Jack cheese (although this Instructor would). Cheese would be binary if the variable represented whether respondents liked cheese.   
-Chris

**[Learn how to set and order factor levels](http://stats.idre.ucla.edu/r/modules/factor-variables/" \t "_blank)**.

There a mistake in the vector assigned to the parameter levels.   
The last age.range group in both solutions should be '65 or Above'.

Have questions? Head to the **[forums](https://forums.udacity.com/?forum_path=c/nd002-data-analysis-with-r" \t "_blank)** for discussion with the Udacity Community.

We recommend anyone new to R and RStudio to try **Swirl** (statistics with interactive R learning). Swirl is a software package for the R statistical programming language. Its purpose is to teach statistics and R commands interactively.

Type the following commands in the **Console**, pressing Enter or Return after each line:

install.packages("swirl") **library**(swirl) **swirl**()

Note that the > symbol at the beginning of the line is R's prompt for you type something into the console. We include it here so you know that the above commands are to be typed into the console and not elsewhere. The part you type begins after >. 

### Download the [EDA Course Materials zip file](https://d17h27t6h515a5.cloudfront.net/topher/2016/November/581a4977_eda-course-materials/eda-course-materials.zip" \t "_blank) to get all of the files used in this course.

### OR

### Download the file, [demystifying.R](https://d17h27t6h515a5.cloudfront.net/topher/2016/October/57fb4aa7_demystifying/demystifying.r" \t "_blank), that Chris mentions in this video.

**[Data Wrangling with MongoDB: Data Manipulation and Retrieval](https://www.udacity.com/course/ud032" \t "_blank)**

Read this **[fantastic tutorial](http://flowingdata.com/2015/02/18/loading-data-and-basic-formatting-in-r/" \t "_blank)** on loading data into R by Nathan Yau.

**[Tidy Data](http://vita.had.co.nz/papers/tidy-data.pdf" \t "_blank)** by Hadley Wickham

**[Tidy Data Presentation](http://courses.had.co.nz.s3-website-us-east-1.amazonaws.com/12-rice-bdsi/slides/07-tidy-data.pdf" \t "_blank)** by Hadley Wickham (June 2012)

Getting Help

**[Quick R](http://www.statmethods.net/" \t "_blank)**   
**[R Cookbook](http://www.cookbook-r.com/" \t "_blank)**   
**[R-Bloggers](http://www.r-bloggers.com/" \t "_blank)**   
**[StackOverflow About R](http://stackoverflow.com/tags/r/info" \t "_blank)**   
**[StackOverflow R FAQ](http://stackoverflow.com/questions/tagged/r-faq%20" \t "_blank)**   
**[Google's R Style Guide](https://google.github.io/styleguide/Rguide.xml" \t "_blank)**

Read this **[fantastic tutorial](http://flowingdata.com/2015/02/18/loading-data-and-basic-formatting-in-r/" \t "_blank)** on loading data into R by Nathan Yau.

**[Download the stateData.csv](https://s3.amazonaws.com/udacity-hosted-downloads/ud651/stateData.csv" \t "_blank)**   
  
**[How to Subset Data](http://www.statmethods.net/management/subset.html" \t "_blank)**

Follow along with Chris in this video using the RMD file, **[What\_is\_a\_RMD\_file](https://s3.amazonaws.com/udacity-hosted-downloads/ud651/What_is_a_RMD_file.Rmd" \t "_blank)**.   
  
The second file, **[demystifyingR2.Rmd](https://d17h27t6h515a5.cloudfront.net/topher/2016/October/57fb4c3b_demystifyingr2-v3/demystifyingr2-v3.rmd" \t "_blank)**, is the file that Chris mentions at the end of the video. Work through this file and use it to answer the next quiz!   
  
There is a difference between using the hash or pound symbol# inside and outside of a {r} block of code.

```{r} # The hash or pound symbol inside the r block creates # a comment. These three lines of are not code and cannot be # executed. x <- [1:10] mean(x) ```

If you use the # symbol outside of an {r} block of code, you can create headers of text.   
  
**[Markdown Tutorial](https://www.youtube.com/watch?v=6A5EpqqDOdk" \t "_blank)**: You can also use Markdown to format your posts in the discussions.   
  
You'll need to install and load the package knitr in order to use the **KNIT HTML** button . Run the following commands in the RStudio console to install and load knitr.

**install**.packages('knitr', dependencies = **T**) **library**(knitr)

Download the **[Reddit Survey Responses](https://s3.amazonaws.com/udacity-hosted-downloads/ud651/reddit.csv" \t "_blank)**   
  
To run the qplot function you must have the ggplot2 library installed and loaded. You can do that by running these two lines of code.

**install**.packages('ggplot2', dependencies = **T**) **library**(ggplot2)

For those on Mac OS X 10.9 (or more recent OS X updates), the above line of code may display an error in the Console and fail to load ggplot2. This is due to a recent software update deleting the X11 symlink. You will need to *reinstall* XQuartz from [**http://xquartz.macosforge.org/**](http://xquartz.macosforge.org/) and then run the following lines of code:

**install**.packages('devtools', dependencies = **T**) **library**(devtools) install\_version("colorspace","1.2-4")

You will then be able to load ggplot2 in RStudio. See the following stackoverflow links for reference [**here**](http://stackoverflow.com/questions/28984243/ggplot2-sourcing-error-x11-library-is-missing) and [**here**](http://stackoverflow.com/questions/28952128/loading-ggplot2-opens-up-x11).

At 0:27 of the solution video, the Instructor misspeaks. Cheese is not binary for the reddit data set. It is a categorical variable with 11 levels. We wouldn't rank Brie over Monterey Jack cheese (although this Instructor would). Cheese would be binary if the variable represented whether respondents liked cheese.   
-Chris

**[Learn how to set and order factor levels](http://stats.idre.ucla.edu/r/modules/factor-variables/" \t "_blank)**.

There a mistake in the vector assigned to the parameter levels.   
The last age.range group in both solutions should be '65 or Above'.

install.packages('ggplot2') library(ggplot2) tShirts <- factor(c('medium', 'small', 'large', 'medium', 'large', 'large'), levels = c('medium','small','large')) tShirts qplot(x = tShirts) tShirts <- ordered(tShirts, levels = c('small', 'medium', 'large')) tShirts qplot(x = tShirts)

**[Data Wrangling with MongoDB: Data Manipulation and Retrieval](https://www.udacity.com/course/ud032" \t "_blank)**

Read this **[fantastic tutorial](http://flowingdata.com/2015/02/18/loading-data-and-basic-formatting-in-r/" \t "_blank)** on loading data into R by Nathan Yau.

**[Tidy Data](http://vita.had.co.nz/papers/tidy-data.pdf" \t "_blank)** by Hadley Wickham

**[Tidy Data Presentation](http://courses.had.co.nz.s3-website-us-east-1.amazonaws.com/12-rice-bdsi/slides/07-tidy-data.pdf" \t "_blank)** by Hadley Wickham (June 2012)

**Take notes and code along with Chris by working in this** **[Lesson 3 RMD file](https://s3.amazonaws.com/udacity-hosted-downloads/ud651/lesson3_student.rmd" \t "_blank)**!

If you are comfortable with basic statistics (mean, median, boxplots, and histograms) and the qplot syntax in R, then consider skipping to the exercises or jumping to [**Lesson 4**](https://classroom.udacity.com/courses/ud651/lessons/755298985/concepts/8644089190923) or [**Lesson 5**](https://classroom.udacity.com/courses/ud651/lessons/701610057/concepts/8737286360923).

**[Download the Pseudo Facebook Data Set](https://s3.amazonaws.com/udacity-hosted-downloads/ud651/pseudo_facebook.tsv" \t "_blank)**!   
  
**Take notes and code along with Chris by working from this** **[Lesson 3 RMD file](https://s3.amazonaws.com/udacity-hosted-downloads/ud651/lesson3_student.rmd" \t "_blank)**!   
  
You could also use the following code to load the Pseudo Facebook data.   
read.delim('pseudo\_facebook.tsv')   
  
The read.delim() function defaults to the tab character as the separator between values and the period as the decimal character. Run ?read.csv or ?read.delim in the console for more details.

**Note**: The use of the scale\_x\_discrete() layer as shown in the video is deprecated as of ggplot2 version 2.0. You can use scale\_x\_continuous() instead to get the break points, or use ggplot() syntax as shown below.

**[How to read Histograms and use them in R](http://flowingdata.com/2014/02/27/how-to-read-histograms-and-use-them-in-r/" \t "_blank)** (note this post uses the base graphics package of R to create histograms; we'll be using the ggplot2 graphics package in this course)   
  
Run the following code in R to get other themes.

**install**.packages('ggthemes', dependencies = TRUE) **library**(ggthemes)

The ggthemes package was developed by Jeffery Arnold. Check out **[examples](https://github.com/jrnold/ggthemes" \l "examples" \t "_blank)** of the themes on the github page.   
  
Chris is using theme\_minimal()with the font size set to 24, which is why his output might look slightly different than yours. You can set the same theme in R by running the following code, or you can set the theme to a choice of your own.   
theme\_set(theme\_minimal(24))   
  
  
Instead of using the qplot() function, you can also use the ggplot() function to create the histogram:

ggplot(aes(x = dob\_day), data = pf) + geom\_histogram(binwidth = 1) + scale\_x\_continuous(breaks = 1:31)

When you register at a websites, some sites have default setting that use the 1st day of the month under the birth day field. Some also use a default month of January.

In the case of Facebook, Jan and the 1st are the first drop down items that appear on the site. It's important to think about such details and the data generating process when exploring data.

**Note**: The use of the scale\_x\_discrete() layer as shown in the video is deprecated as of ggplot2 version 2.0. You can use scale\_x\_continuous() instead to get the break points, or use ggplot() syntax as shown below.

**[How to read Histograms and use them in R](http://flowingdata.com/2014/02/27/how-to-read-histograms-and-use-them-in-r/" \t "_blank)** (note this post uses the base graphics package of R to create histograms; we'll be using the ggplot2 graphics package in this course)   
  
Run the following code in R to get other themes.

**install**.packages('ggthemes', dependencies = TRUE) **library**(ggthemes)

The ggthemes package was developed by Jeffery Arnold. Check out **[examples](https://github.com/jrnold/ggthemes" \l "examples" \t "_blank)** of the themes on the github page.   
  
Chris is using theme\_minimal()with the font size set to 24, which is why his output might look slightly different than yours. You can set the same theme in R by running the following code, or you can set the theme to a choice of your own.   
theme\_set(theme\_minimal(24))   
  
  
Instead of using the qplot() function, you can also use the ggplot() function to create the histogram:

ggplot(aes(x = dob\_day), data = pf) + geom\_histogram(binwidth = 1) + scale\_x\_continuous(breaks = 1:31)

When you register at a websites, some sites have default setting that use the 1st day of the month under the birth day field. Some also use a default month of January.

In the case of Facebook, Jan and the 1st are the first drop down items that appear on the site. It's important to think about such details and the data generating process when exploring data.

**[Bernstein, M. S., Bakshy, E., Burke, M., & Karrer, B. (2013). Quantifying the invisible audience in social networks. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2013), pp. 21-30](http://hci.stanford.edu/publications/2013/invisibleaudience/invisibleaudience.pdf" \t "_blank)**.

**Note**: The use of the scale\_x\_discrete() layer as shown in the video is deprecated as of ggplot2 version 2.0. You can use scale\_x\_continuous() instead, or use ggplot() syntax as shown below.

It's best to create the plot in RStudio on your computer and examine at full screen. **[Faceting](http://www.cookbook-r.com/Graphs/Facets_(ggplot2)/" \t "_blank)** in ggplot2

Equivalent ggplot syntax:

ggplot(data = pf, aes(x = dob\_day)) + geom\_histogram(binwidth = 1) + scale\_x\_continuous(breaks = 1:31) + facet\_wrap(~dob\_month)