Using ggplot2 and the rest of the tidyverse to present data

Introduction

Ggplot2 is a powerful package available for R that is used to create graphs, scatterplots and any other plots / graphical representations that you can think of (boxplots etc.). Ggplot2 comes as part of a larger grouping of packages, collectively called the 'tidyverse'. Some of these packages have already featured in our stat133 class, notably dplyr and readr. You can install the tidyverse with `install.packages("tidyverse")` . The tidyverse features six core packages when loaded with `library(tidyverse)` that are ready to use for data analysis and representation. The core packages are as follows:

- ggplot2: a system for creating graphics
- dplyr: a data manipulation package
- readr: a fast way to read data from rectangular formats (.csv etc.)
- tidyr: a set of functions to help you tidy up data
- purrr: an intuitive way to make functional programs
- tibble: a reworking of R's standard data frames for easier use

The tidyverse also comes with a lot of auxiliary packages that are not loaded automatically with `library(tidyverse)` . These packages are used for a range of more specialty applications, but must be individually loaded by name. A list of all the packages contained within the tidyverse can be found here: https://www.tidyverse.org/packages/

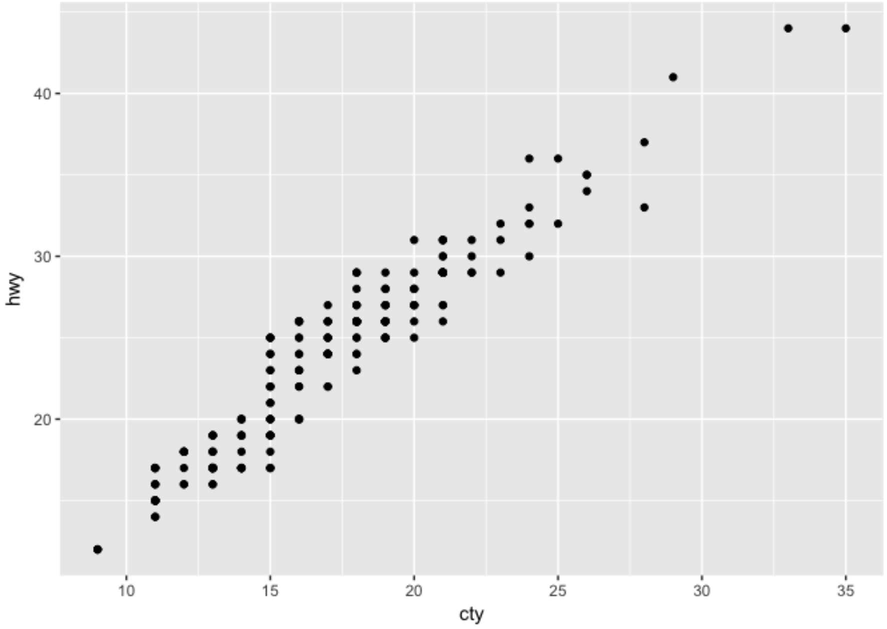
Ggplot2

Ggplot2 is the focus of this post:it is extremely powerful and can be used to make many different graphical representations of data. Full documentation for ggplot2 can be found here: http://ggplot2.tidyverse.org/reference/ . Here is an example of how to make a simple scatterplot showing the relationship between highway mpg (hwy) and city mpg (cty) from data `mpg` , which is information included in ggplot2 about fuel economy on cars (details here: http://ggplot2.tidyverse.org/reference/mpg.html):

```
library(tidyverse)
```

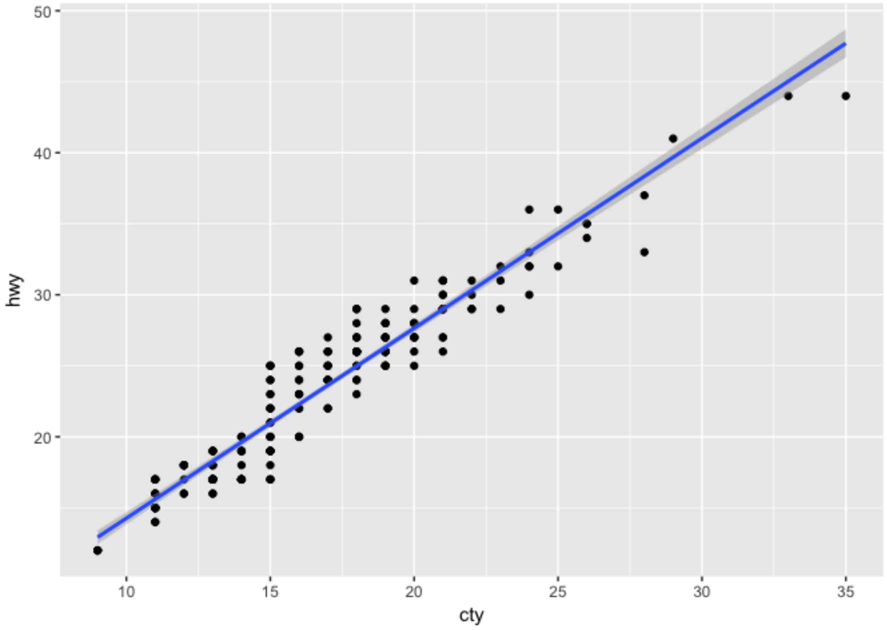
```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr

## Warning: package 'tidyr' was built under R version 3.4.2

## Warning: package 'purrr' was built under R version 3.4.2

## Conflicts with tidy packages ----------------------------------------------

## filter(): dplyr, stats
## lag():    dplyr, stats
```

```
ggplot(mpg, aes(x= cty, y = hwy )) + geom_point()
```



We can then use ggplot2 to map more information about this data on to a graph. In this graph we will use `geom_smooth` to add a line of best fit to our data. The designation of `method = 'lm'` tells R that we are going to do a linear regression. Other types of methods can also be used in the geom_smooth functions.

```
ggplot(mpg, aes(x= cty, y = hwy)) + geom_point() + geom_smooth(method = 'lm')
```



More information on how to best manipulate ggplot2 can be found here: http://r-statistics.co/Complete-Ggplot2-Tutorial-Part1-With-R-Code.html . This is a very thorough tutorial of all the powers of ggplot2.

The Core Tidyverse

readr

Readr is used instead of R's own base functions to import tabular data. It tends to be around 10x faster than the equivalent base functions and also is more consistent. One of the biggest perks of readr is no longer having to specify `stringsAsFactors = FALSE` . A full tutorial can be found here: https://blog.rstudio.com/2015/04/09/readr-0-1-0/

dplyr

Dplyr is the ultimate in data manipulation. It uses a few key verbs in order to allow easy manipulation of data, particularly when compared to R's base functions. An example of this filtering out only the cars with both city and highway mpg >= 20 is shown below.

```
# base R
mpg_high_cty_and_hwy <- mpg[mpg$cty>=20 &
                                mpg$hwy>=20, ]

# dplyr
mpg_high_cty_and_hwy <- filter(mpg, cty>=20, hwy>=20)
```

A tutorial for the rest of dplyr can be found here: https://rpubs.com/justmarkham/dplyr-tutorial

purrr

Purrr is made to enhance R's functional programming. It is designed to provide a more consistent set of tools than R's base functions. Tutorial here: https://jennybc.github.io/purrr-tutorial/

tidyr

Tidyr provides a set of tools to help you tidy up data into a consistent form. Data goes into columns, and every column is a variable. More information can be found here: http://data.library.virginia.edu/a-tidyr-tutorial/

tibble

Tibble provides data frames that are lazier than R's data frames. This forces you to confront problems in data earlier, and helps to prevent confusion in the long run. More info on how to use tibble can be found here: http://tibble.tidyverse.org/

Conclusion

As you can see, the tidyverse provides a comprehensive set of packages for streamlining data preparation and visualisation. It is an essential set of tools in any useR's arsenal.