# CT1100: Computer Systems

# Topic 3: The tibble and an introduction to ggplot2

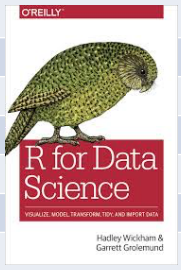
Prof. Jim Duggan,
School of Engineering & Informatics
National University of Ireland Galway.



1

# Topics to be Covered (R)

| Topic | Description |
|---|---|
| 1 | Introduction to R and R Studio Cloud |
| 2 | A program in R |
| 3 | The tibble – a way of storing information |
| 4 | Data Visualisation I |
| 5 | Data Transformation I |
| 6 | Running a Script in R |
| 7 | Data Visualisation II |
| 8 | Data Transformation II |
| 9 | Exploring Data |
| 10 | Communicating Results |




https://r4ds.had.co.nz



2

# Data Frames/Tibbles

- The most common way of storing data in R
- A two-dimensional structure, with rows (observations) and columns (variables)

```
> observations
# A tibble: 219,000 x 12
   station  year month   day  hour date                rain  temp  rhum
   <chr>   <dbl> <dbl> <int> <int> <dttm>              <dbl> <dbl> <dbl>
 1 ATHENRY  2017     1     1     0 2017-01-01 00:00:00   0     5.2    89
 2 ATHENRY  2017     1     1     1 2017-01-01 01:00:00   0     4.7    89
 3 ATHENRY  2017     1     1     2 2017-01-01 02:00:00   0     4.2    90
 4 ATHENRY  2017     1     1     3 2017-01-01 03:00:00   0.1   3.5    87
 5 ATHENRY  2017     1     1     4 2017-01-01 04:00:00   0.1   3.2    89
 6 ATHENRY  2017     1     1     5 2017-01-01 05:00:00   0     2.1    91
 7 ATHENRY  2017     1     1     6 2017-01-01 06:00:00   0     2      89
 8 ATHENRY  2017     1     1     7 2017-01-01 07:00:00   0     1.7    89
 9 ATHENRY  2017     1     1     8 2017-01-01 08:00:00   0     1      91
10 ATHENRY  2017     1     1     9 2017-01-01 09:00:00   0     1.1    91
# … with 218,990 more rows, and 3 more variables: msl <dbl>, wdsp <dbl>,
#   wddir <dbl>
```



# tibble abbreviations (variable types)

| Abbreviation | Data Type |
|---|---|
| int | integers |
| dbl | doubles (real numbers) |
| chr | character vectors (strings |
| dttm | date-times |
| lgl | logical |
| fctr | factor (categorical variables with fixed possible values) |
| date | dates |

# Challenge 3.1

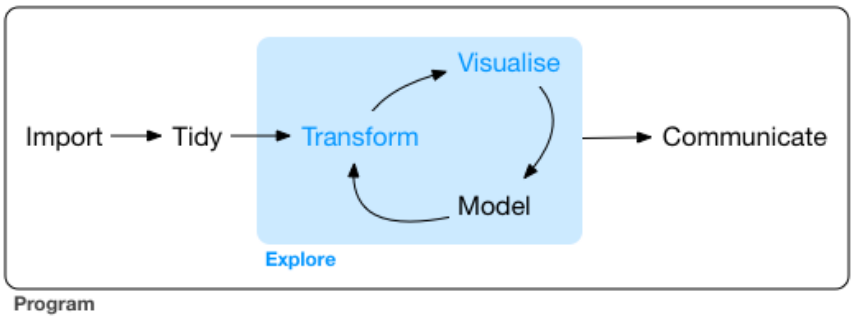
- Explore some of the following tibbles in R
  - mpg (ggplot2)
  - flights (nycflights13)
  - observations (aimsir17)
  - eirgrid17 (aimsir17)
  - stations (aimsir17)



5

# Data Exploration

"Data exploration is the art of looking at your data, rapidly generating hypotheses, quickly testing them, then repeating again and again and again." (Wickham and Grolemund 2017).






6

# Data Visualisation with **ggplot2**

"The simple graph has brought more information to the data analyst's mind that any other device." – John Tukey

```
> dt <- ggplot2::mpg
>
> dt
# A tibble: 234 × 11
   manufacturer      model displ  year   cyl      trans   drv   cty   hwy    fl    class
          <chr>      <chr> <dbl> <int> <int>      <chr> <chr> <int> <int> <chr>    <chr>
1          audi         a4   1.8  1999     4   auto(l5)     f    18    29     p compact
2          audi         a4   1.8  1999     4 manual(m5)     f    21    29     p compact
3          audi         a4   2.0  2008     4 manual(m6)     f    20    31     p compact
4          audi         a4   2.0  2008     4   auto(av)     f    21    30     p compact
5          audi         a4   2.8  1999     6   auto(l5)     f    16    26     p compact
6          audi         a4   2.8  1999     6 manual(m5)     f    18    26     p compact
7          audi         a4   3.1  2008     6   auto(av)     f    18    27     p compact
8          audi a4 quattro   1.8  1999     4 manual(m5)     4    18    26     p compact
9          audi a4 quattro   1.8  1999     4   auto(l5)     4    16    25     p compact
10         audi a4 quattro   2.0  2008     4 manual(m6)     4    20    28     p compact
# ... with 224 more rows
```



7

# Fuel Economy Data Set (ggplot2::mpg)

This dataset contains a subset of the fuel economy data that the EPA makes available on http://fueleconomy.gov. It contains only models which had a new release every year between 1999 and 2008 - this was used as a proxy for the popularity of the car.

| | | | |
|---|---|---|---|
| **manufacturer** | manufacturer | **drv** | f = front-wheel drive, r = rear wheel drive, 4 = 4wd |
| **model** | model name | **cty** | city miles per gallon |
| **displ** | engine displacement, in litres | **hwy** | highway miles per gallon |
| **year** | year of manufacture | **fl** | fuel type |
| **cyl** | number of cylinders | **class** | "type" of car |
| **trans** | type of transmission | | |



8

# First Steps

- Generate a first graph to help answer the following question:
  - *Do cars with big engines use more fuel than cars with small engines*
- What might the relationship between engine size and fuel efficiency look like?

```
> dt <- ggplot2::mpg
>
> dt
# A tibble: 234 × 11
   manufacturer      model displ  year   cyl      trans   drv   cty   hwy    fl   class
          <chr>      <chr> <dbl> <int> <int>      <chr> <chr> <int> <int> <chr>   <chr>
1          audi         a4   1.8  1999     4   auto(l5)     f    18    29     p compact
2          audi         a4   1.8  1999     4 manual(m5)     f    21    29     p compact
3          audi         a4   2.0  2008     4 manual(m6)     f    20    31     p compact
4          audi         a4   2.0  2008     4   auto(av)     f    21    30     p compact
5          audi         a4   2.8  1999     6   auto(l5)     f    16    26     p compact
6          audi         a4   2.8  1999     6 manual(m5)     f    18    26     p compact
7          audi         a4   3.1  2008     6   auto(av)     f    18    27     p compact
8          audi a4 quattro   1.8  1999     4 manual(m5)     4    18    26     p compact
9          audi a4 quattro   1.8  1999     4   auto(l5)     4    16    25     p compact
10         audi a4 quattro   2.0  2008     4 manual(m6)     4    20    28     p compact
# ... with 224 more rows
```



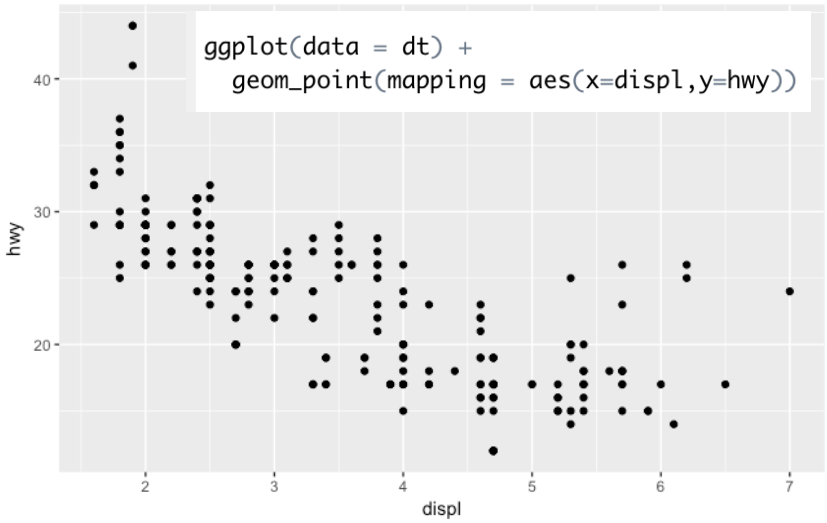
# Selecting data

```
> dt
# A tibble: 234 × 11
   manufacturer      model displ  year   cyl      trans   drv   cty   hwy    fl   class
          <chr>      <chr> <dbl> <int> <int>      <chr> <chr> <int> <int> <chr>   <chr>
1          audi         a4   1.8  1999     4   auto(l5)     f    18    29     p compact
2          audi         a4   1.8  1999     4 manual(m5)     f    21    29     p compact
3          audi         a4   2.0  2008     4 manual(m6)     f    20    31     p compact
4          audi         a4   2.0  2008     4   auto(av)     f    21    30     p compact
5          audi         a4   2.8  1999     6   auto(l5)     f    16    26     p compact
```

- Among the variables are:
  - **displ**, a car's engine size in litres
  - **hwy**, a car's fuel efficiency on the highway in miles per gallon

# Creating a ggplot

```
ggplot(data = dt) +
  geom_point(mapping = aes(x=displ,y=hwy))
```

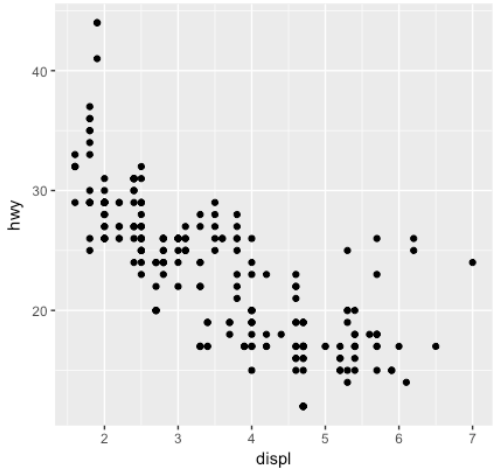





11

# Interpreting the plot

- The plot shows a negative relationship between engine size (displ) and fuel efficiency (hwy)
- Cars with big engines use more fuel






12

# Challenge 3.2

- For a given weather station, and over one month of the year, explore the relationship between atmospheric pressure and wind speed. What do you think it might look like?

```
> observations
# A tibble: 219,000 x 12
   station  year month   day  hour date                rain  temp  rhum   msl  wdsp wddir
   <chr>   <dbl> <dbl> <int> <int> <dttm>              <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
 1 ATHENRY  2017     1     1     0 2017-01-01 00:00:00     0   5.2    89 1022.     8   320
 2 ATHENRY  2017     1     1     1 2017-01-01 01:00:00     0   4.7    89 1022      9   320
 3 ATHENRY  2017     1     1     2 2017-01-01 02:00:00     0   4.2    90 1022.     8   320
 4 ATHENRY  2017     1     1     3 2017-01-01 03:00:00   0.1   3.5    87 1022.     9   330
```

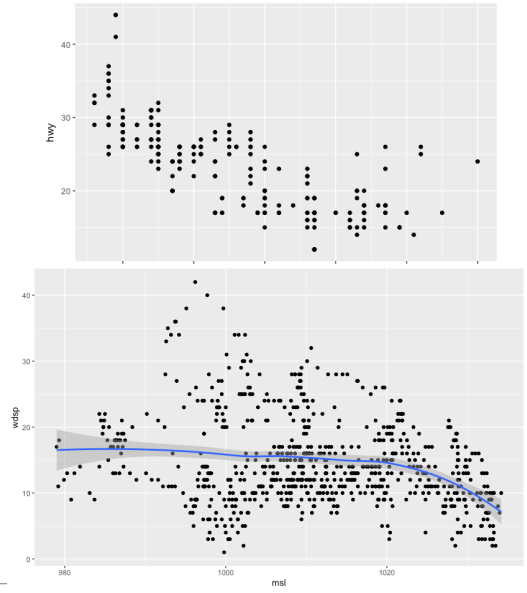


13

# Lecture 3 - Summary

| Topic | Description |
|---|---|
| 1 | Introduction to R and R Studio Cloud |
| 2 | A program in R |
| 3 | The tibble – a way of storing information |
| 4 | Data Visualisation I |
| 5 | Data Transformation I |
| 6 | Running a Script in R |
| 7 | Data Visualisation II |
| 8 | Data Transformation II |
| 9 | Exploring Data |
| 10 | Communicating Results |


```
> observations
# A tibble: 219,000 x 12
   station  year month   day  hour date                rain  temp  rhum
   <chr>   <dbl> <dbl> <int> <int> <dttm>              <dbl> <dbl> <dbl>
 1 ATHENRY  2017     1     1     0 2017-01-01 00:00:00     0   5.2    89
 2 ATHENRY  2017     1     1     1 2017-01-01 01:00:00     0   4.7    89
 3 ATHENRY  2017     1     1     2 2017-01-01 02:00:00     0   4.2    90
 4 ATHENRY  2017     1     1     3 2017-01-01 03:00:00   0.1   3.5    87
 5 ATHENRY  2017     1     1     4 2017-01-01 04:00:00   0.1   3.2    89
 6 ATHENRY  2017     1     1     5 2017-01-01 05:00:00     0   2.1    91
 7 ATHENRY  2017     1     1     6 2017-01-01 06:00:00     0   2      89
 8 ATHENRY  2017     1     1     7 2017-01-01 07:00:00     0   1.7    89
 9 ATHENRY  2017     1     1     8 2017-01-01 08:00:00     0   1      91
10 ATHENRY  2017     1     1     9 2017-01-01 09:00:00     0   1.1    91
# … with 218,990 more rows, and 3 more variables: msl <dbl>, wdsp <dbl>,
#   wddir <dbl>
```






14