CT1100: Computer Systems

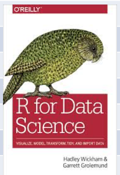
Topic 7: Statistical Transformations with ggplot2

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

Overview

- ggplot2 recap
- New data set: *ggplot2::diamonds*
- Geometric Objects
- Statistical Transformations
  - Bar Charts (Counts)
  - Box Plot (Distributions)

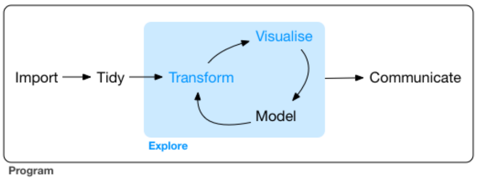
Topic	Description
1	Introduction to R and R Studio Cloud
2	A program in R
3	The tibble – a way of storing information
4	Data Visualisation I
5	Data Transformation I
6	Running a Script in R
7	Data Visualisation II
8	Data Transformation II
9	Exploring Data
10	Communicating Results

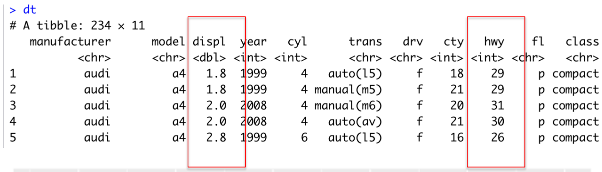


https://r4ds.had.co.nz

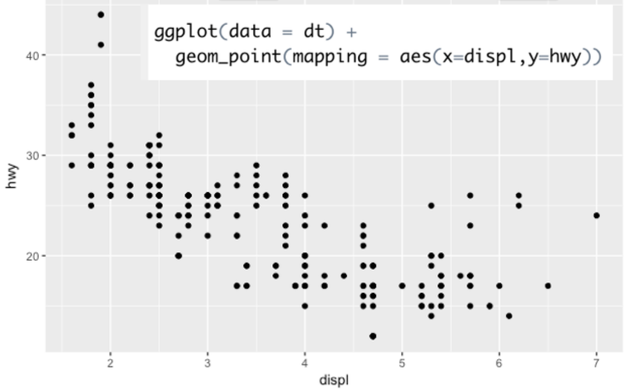
Recap - Data Exploration

“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).





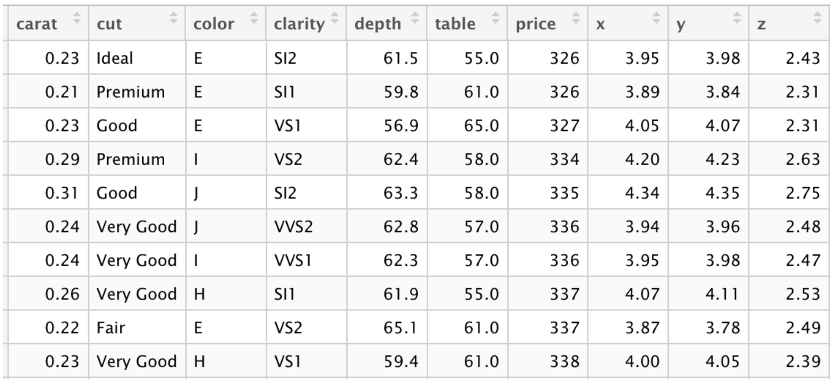
```
> dt
# A tibble: 234 x 11
   manufacturer      model displ  year   cyl      trans   drv   cty   hwy   fl   class
          <chr>      <chr> <dbl> <int> <int>      <chr> <chr> <int> <int> <chr>  <chr>
1          audi         a4   1.8  1999     4   auto(l5)     f    18    29     p compact
2          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
```



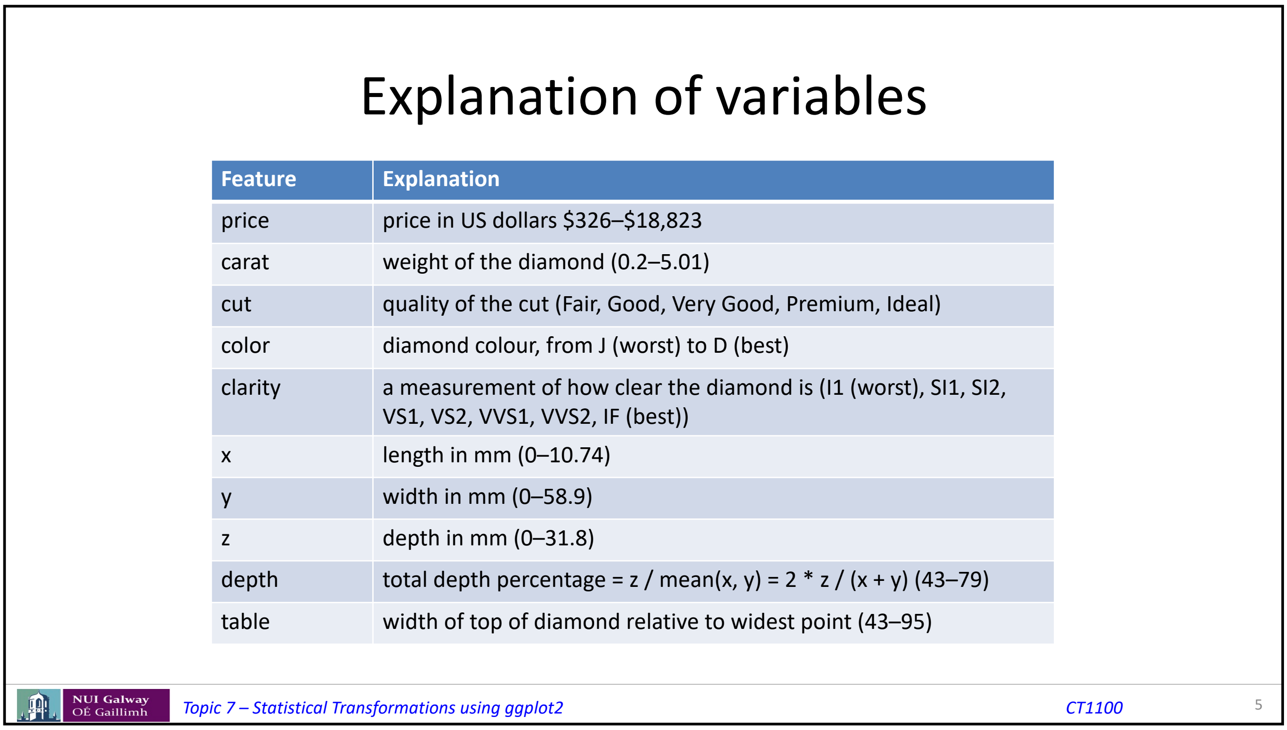
```
ggplot(data = dt) +
  geom_point(mapping = aes(x=displ,y=hwy))
```

diamonds data set (ggplot2)

A dataset containing the prices and other attributes of almost 54,000 diamonds.

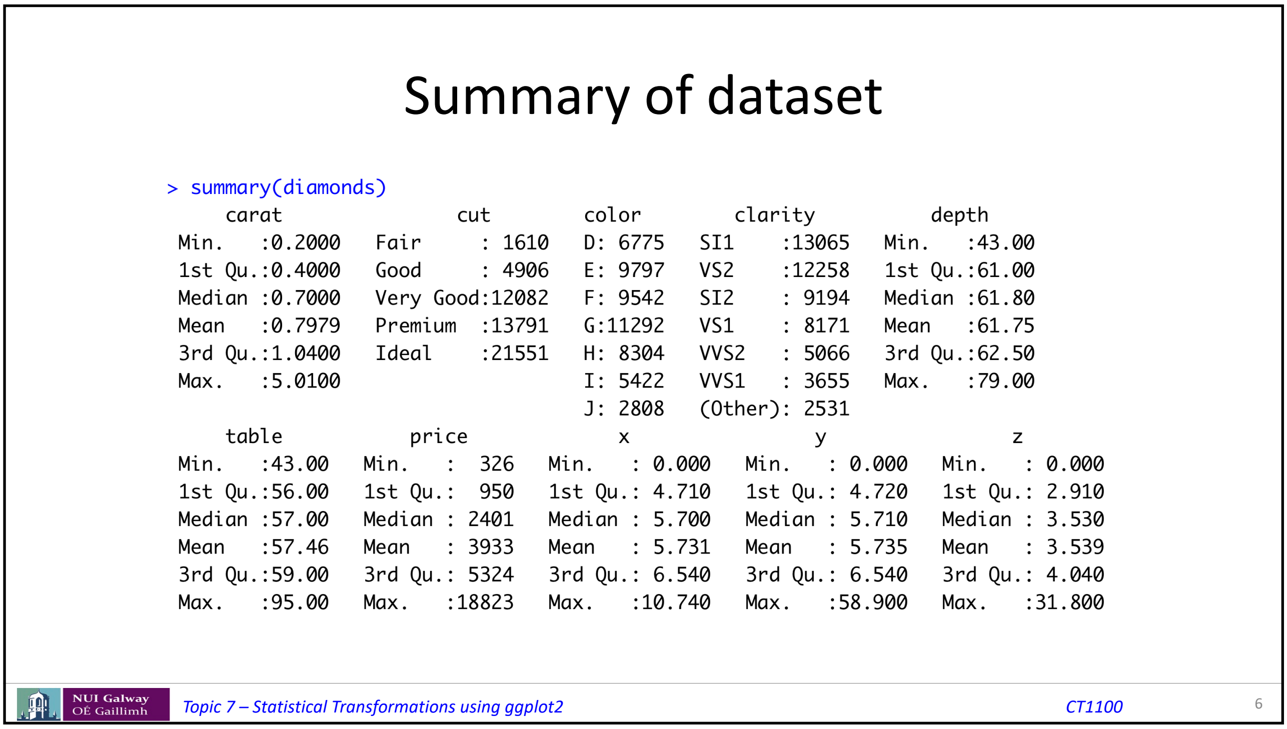


carat	cut	color	clarity	depth	table	price	x	y	z
0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75
0.24	Very Good	J	VVS2	62.8	57.0	336	3.94	3.96	2.48
0.24	Very Good	I	VVS1	62.3	57.0	336	3.95	3.98	2.47
0.26	Very Good	H	SI1	61.9	55.0	337	4.07	4.11	2.53
0.22	Fair	E	VS2	65.1	61.0	337	3.87	3.78	2.49
0.23	Very Good	H	VS1	59.4	61.0	338	4.00	4.05	2.39



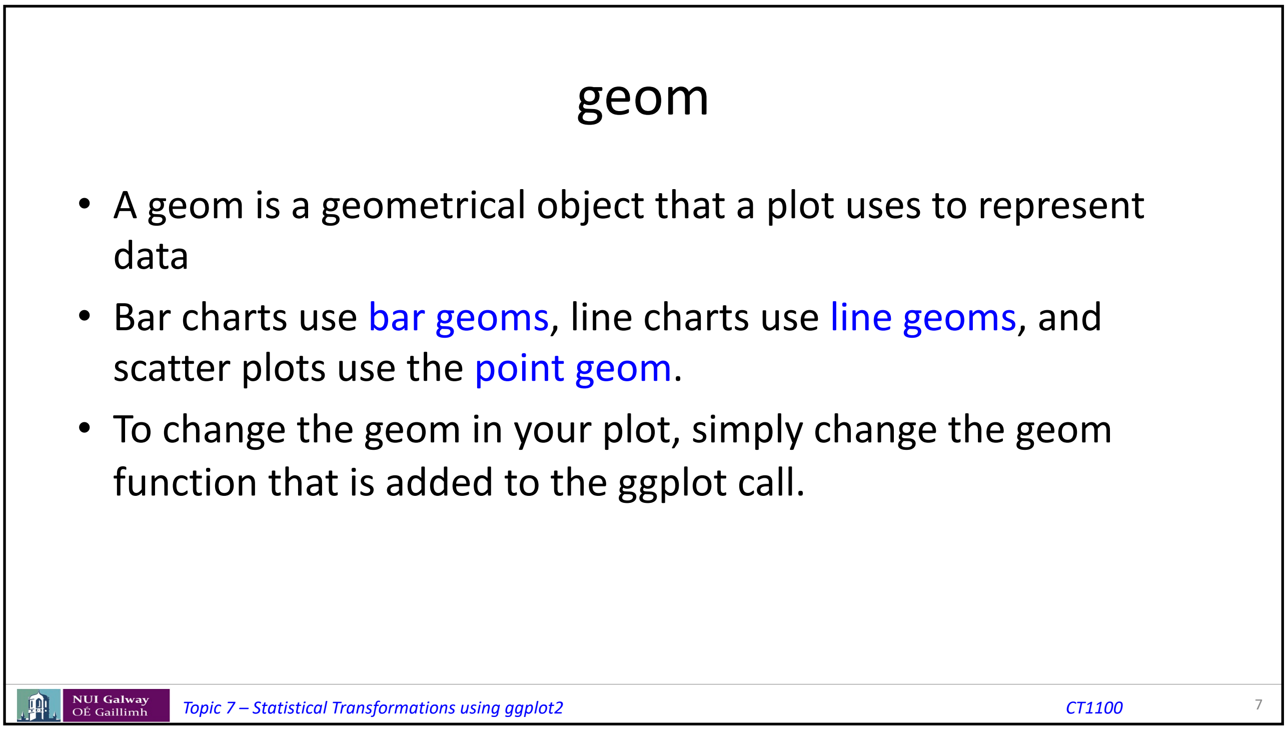
Explanation of variables

Feature	Explanation
price	price in US dollars $326–$18,823
carat	weight of the diamond (0.2–5.01)
cut	quality of the cut (Fair, Good, Very Good, Premium, Ideal)
color	diamond colour, from J (worst) to D (best)
clarity	a measurement of how clear the diamond is (I1 (worst), SI1, SI2, VS1, VS2, VVS1, VVS2, IF (best))
x	length in mm (0–10.74)
y	width in mm (0–58.9)
z	depth in mm (0–31.8)
depth	total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43–79)
table	width of top of diamond relative to widest point (43–95)



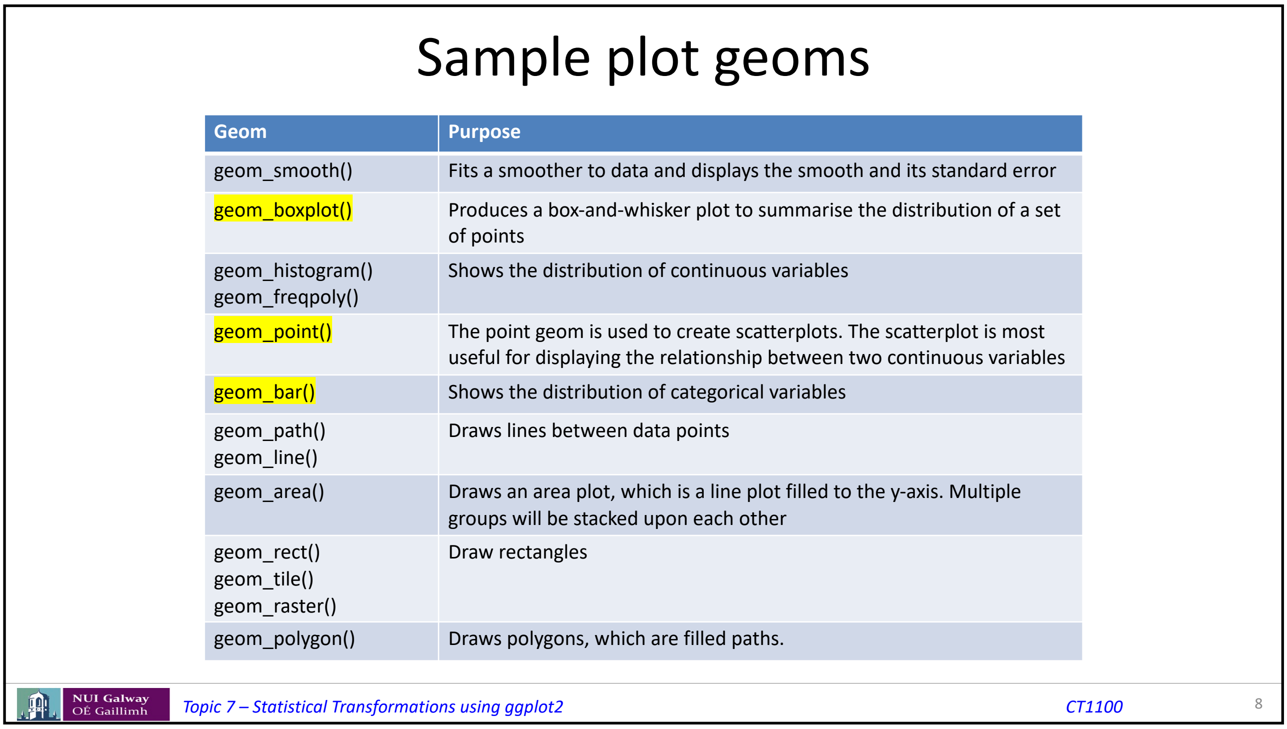
Summary of dataset

```
> summary(diamonds)
     carat                cut         color        clarity          depth      
 Min.   :0.2000   Fair     : 1610   D: 6775   SI1    :13065   Min.   :43.00  
 1st Qu.:0.4000   Good     : 4906   E: 9797   VS2    :12258   1st Qu.:61.00  
 Median :0.7000   Very Good:12082   F: 9542   SI2    : 9194   Median :61.80  
 Mean   :0.7979   Premium  :13791   G:11292   VS1    : 8171   Mean   :61.75  
 3rd Qu.:1.0400   Ideal    :21551   H: 8304   VVS2   : 5066   3rd Qu.:62.50  
 Max.   :5.0100                     I: 5422   VVS1   : 3655   Max.   :79.00  
                                    J: 2808   (Other): 2531                  
     table           price             x                y                z         
 Min.   :43.00   Min.   :  326   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000  
 1st Qu.:56.00   1st Qu.:  950   1st Qu.: 4.710   1st Qu.: 4.720   1st Qu.: 2.910  
 Median :57.00   Median : 2401   Median : 5.700   Median : 5.710   Median : 3.530  
 Mean   :57.46   Mean   : 3933   Mean   : 5.731   Mean   : 5.735   Mean   : 3.539  
 3rd Qu.:59.00   3rd Qu.: 5324   3rd Qu.: 6.540   3rd Qu.: 6.540   3rd Qu.: 4.040  
 Max.   :95.00   Max.   :18823   Max.   :10.740   Max.   :58.900   Max.   :31.800  
```



geom

- A geom is a geometrical object that a plot uses to represent data
- Bar charts use bar geoms, line charts use line geoms, and scatter plots use the point geom.
- To change the geom in your plot, simply change the geom function that is added to the ggplot call.



Sample plot geoms

Geom	Purpose
geom_smooth()	Fits a smoother to data and displays the smooth and its standard error
geom_boxplot()	Produces a box-and-whisker plot to summarise the distribution of a set of points

geom_histogram()
geom_freqpoly()
	Shows the distribution of continuous variables

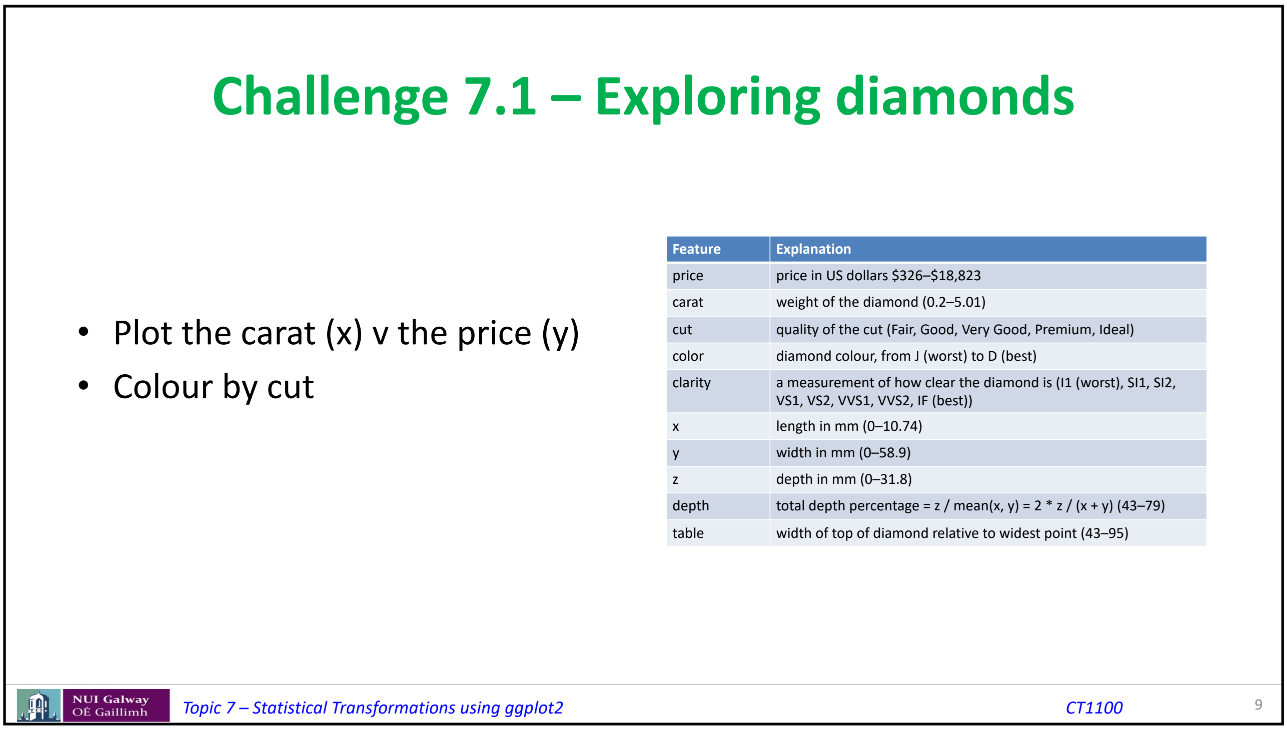
geom_point()	The point geom is used to create scatterplots. The scatterplot is most useful for displaying the relationship between two continuous variables
geom_bar()	Shows the distribution of categorical variables

geom_path()
geom_line()
	Draws lines between data points

geom_area()	Draws an area plot, which is a line plot filled to the y-axis. Multiple groups will be stacked upon each other

geom_rect()
geom_tile()
geom_raster()
	Draw rectangles

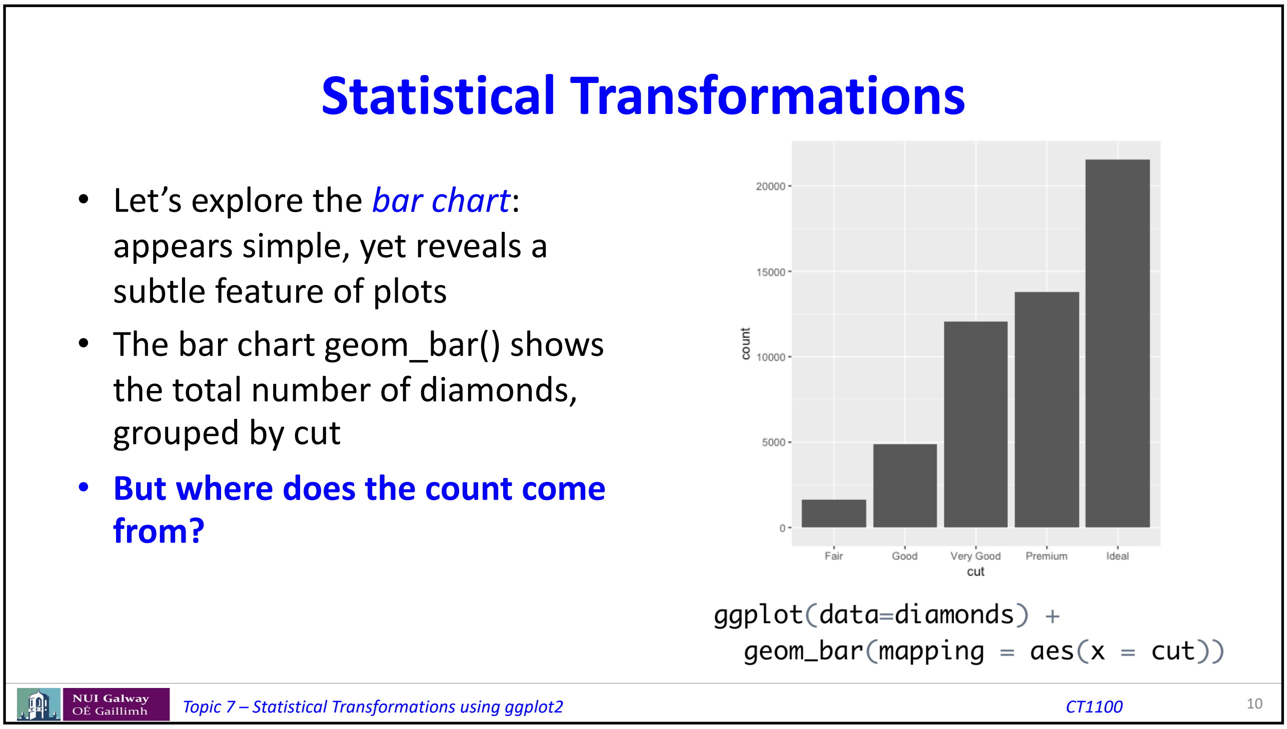
geom_polygon()	Draws polygons, which are filled paths.



Challenge 7.1 – Exploring diamonds

- Plot the carat (x) v the price (y)
- Colour by cut

Feature	Explanation
price	price in US dollars \$326–\$18,823
carat	weight of the diamond (0.2–5.01)
cut	quality of the cut (Fair, Good, Very Good, Premium, Ideal)
color	diamond colour, from J (worst) to D (best)
clarity	a measurement of how clear the diamond is (I1 (worst), SI1, SI2, VS1, VS2, VVS1, VVS2, IF (best))
x	length in mm (0–10.74)
y	width in mm (0–58.9)
z	depth in mm (0–31.8)
depth	total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43–79)
table	width of top of diamond relative to widest point (43–95)



Statistical Transformations

- Let’s explore the *bar chart*: appears simple, yet reveals a subtle feature of plots
- The bar chart geom_bar() shows the total number of diamonds, grouped by cut
- **But where does the count come from?**

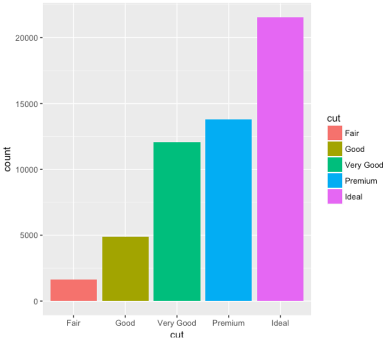
```
ggplot(data=diamonds) +
  geom_bar(mapping = aes(x = cut))
```

Explanation

- Many graphs, like scatterplots, plot the raw values of the dataset
- However, other graphs (e.g. bar charts) *calculate new values to plot*
  - **Bar charts, histograms and frequency polygons** bin your data and plot bin counts, the number of points that fall in each bin/category
  - **Smoothers** fit a model to your data and the plot predictions from the model
  - **Boxplots** compute a robust summary of the distribution and display a specially formatted box

fill aesthetic for bar charts

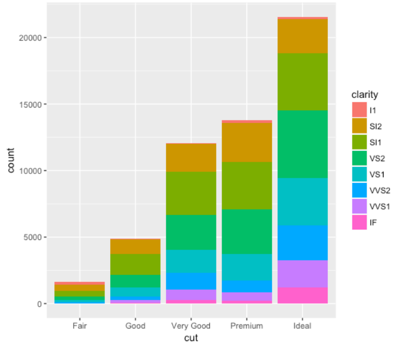
- Bar charts can be coloured using the fill aesthetic



```
ggplot(data=diamonds) +
  geom_bar(mapping=aes(x=cut,fill=cut))
```

- When a different variable is used, the graph has further detail

```
ggplot(data=diamonds) +
  geom_bar(mapping=aes(x=cut,fill=clarity))
```

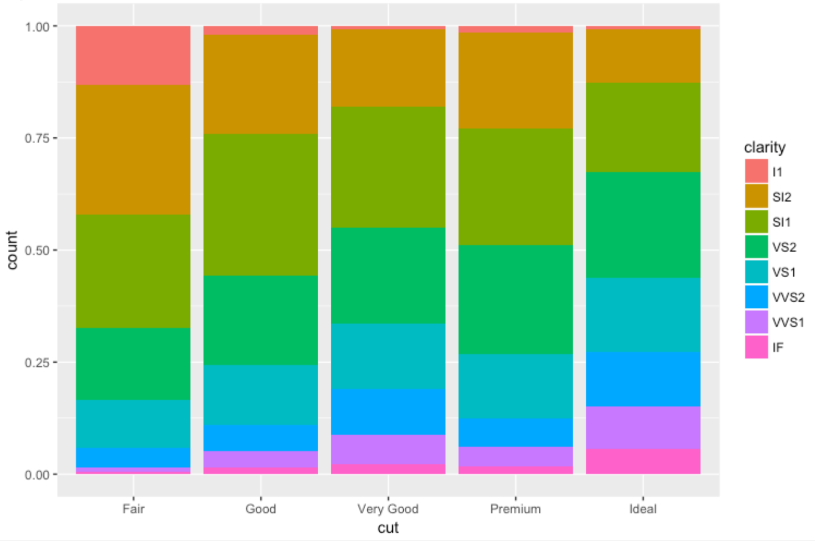


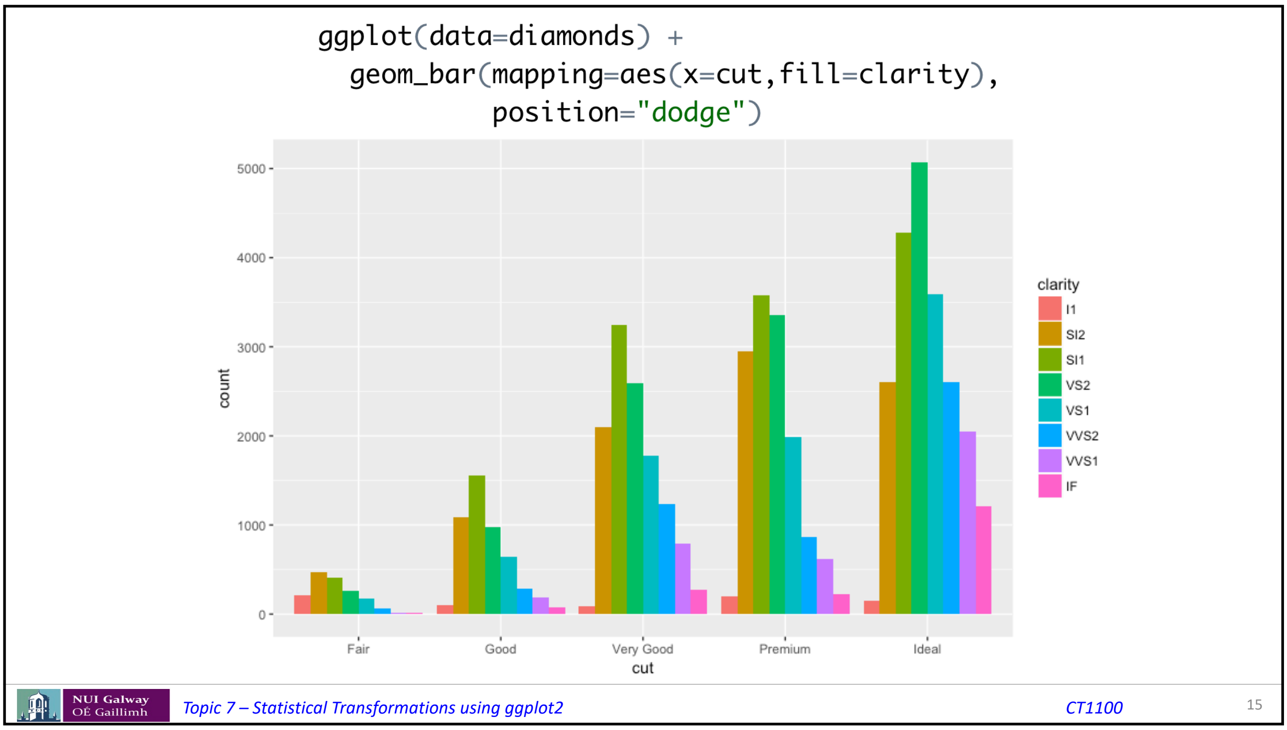
Stacking options

- Stacking is performed automatically by the position adjustment specified by the **position** argument
- Examples include “identity”, “fill” and “dodge”

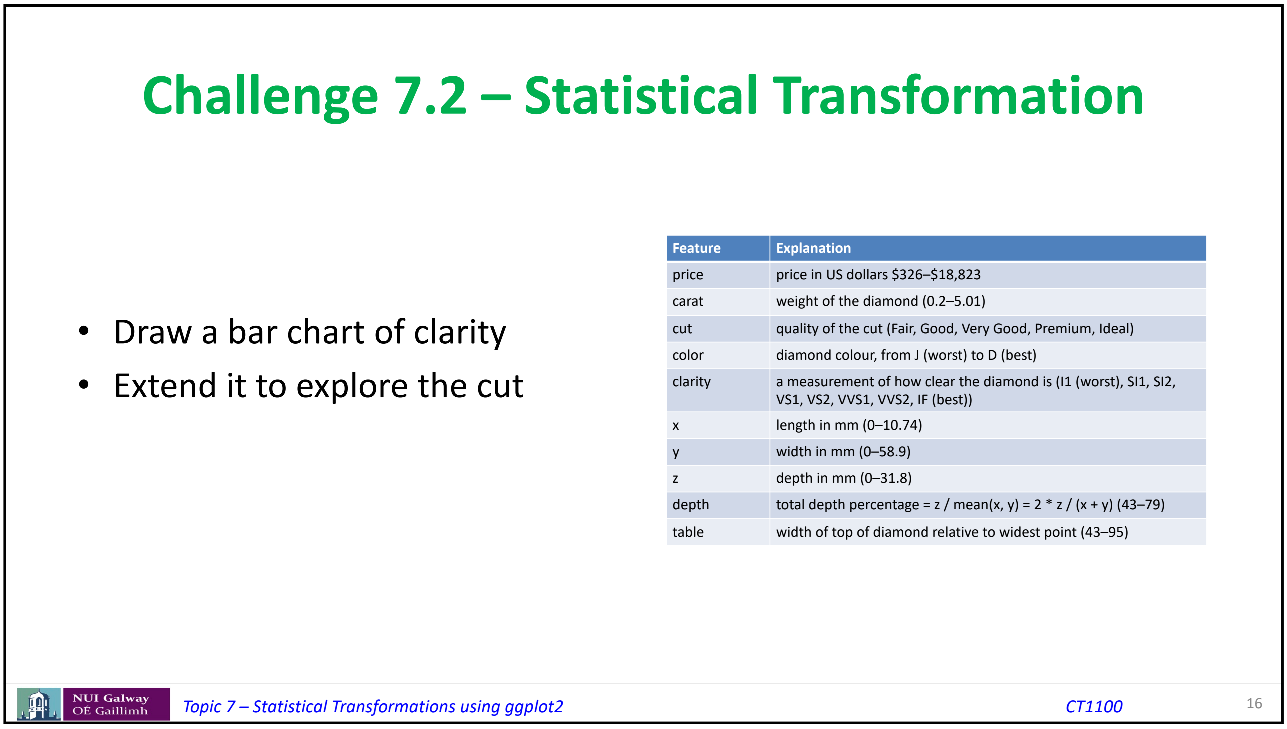
- “fill”
  - Works like stacking, but each stacked bar is the same height
  - Makes it easier to compare proportions
- “dodge”
  - Places objects directly beside one another
  - Makes it easier to compare individual values

```
ggplot(data=diamonds) +
  geom_bar(mapping=aes(x=cut,fill=clarity),
           position="fill")
```





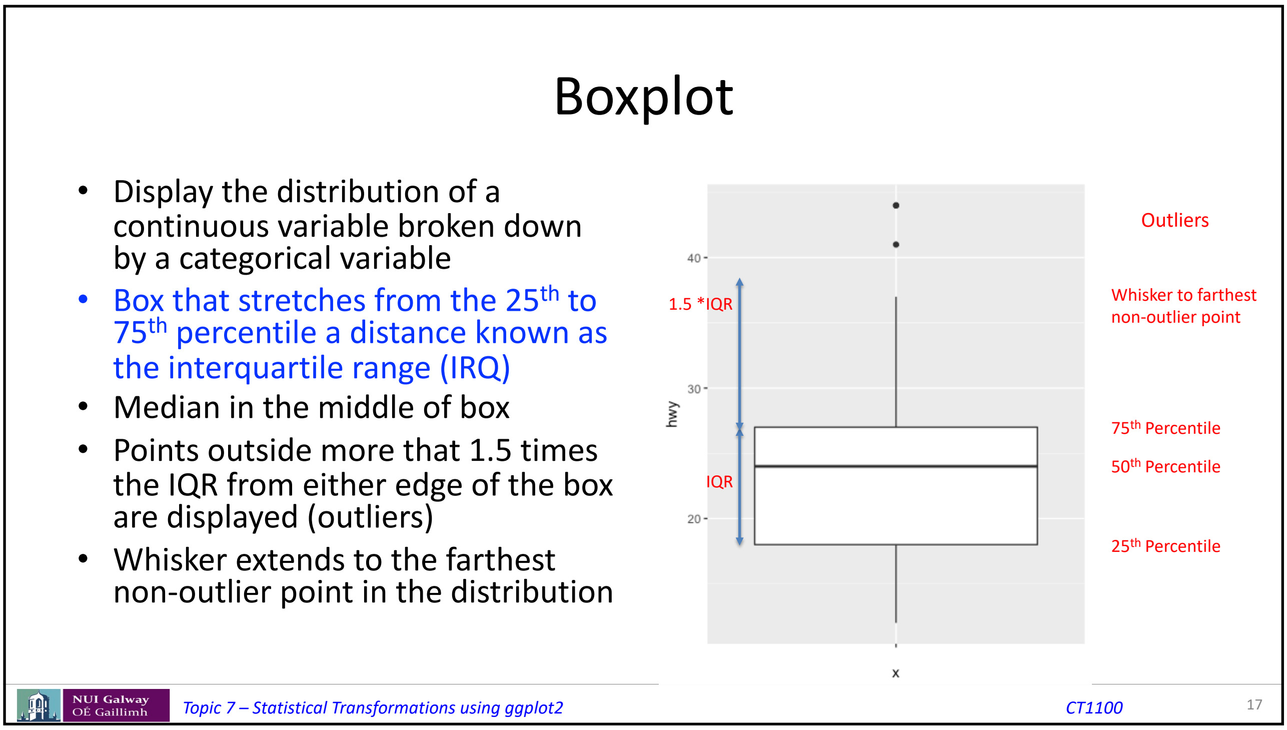
```
ggplot(data=diamonds) +
  geom_bar(mapping=aes(x=cut,fill=clarity),
           position="dodge")
```



Challenge 7.2 – Statistical Transformation

- Draw a bar chart of clarity
- Extend it to explore the cut

Feature	Explanation
price	price in US dollars $326–$18,823
carat	weight of the diamond (0.2–5.01)
cut	quality of the cut (Fair, Good, Very Good, Premium, Ideal)
color	diamond colour, from J (worst) to D (best)
clarity	a measurement of how clear the diamond is (I1 (worst), SI1, SI2, VS1, VS2, VVS1, VVS2, IF (best))
x	length in mm (0–10.74)
y	width in mm (0–58.9)
z	depth in mm (0–31.8)
depth	total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43–79)
table	width of top of diamond relative to widest point (43–95)



Boxplot

- Display the distribution of a continuous variable broken down by a categorical variable
- Box that stretches from the 25th to 75th percentile a distance known as the interquartile range (IRQ)
- Median in the middle of box
- Points outside more that 1.5 times the IQR from either edge of the box are displayed (outliers)
- Whisker extends to the farthest non-outlier point in the distribution

Outliers

Whisker to farthest non-outlier point

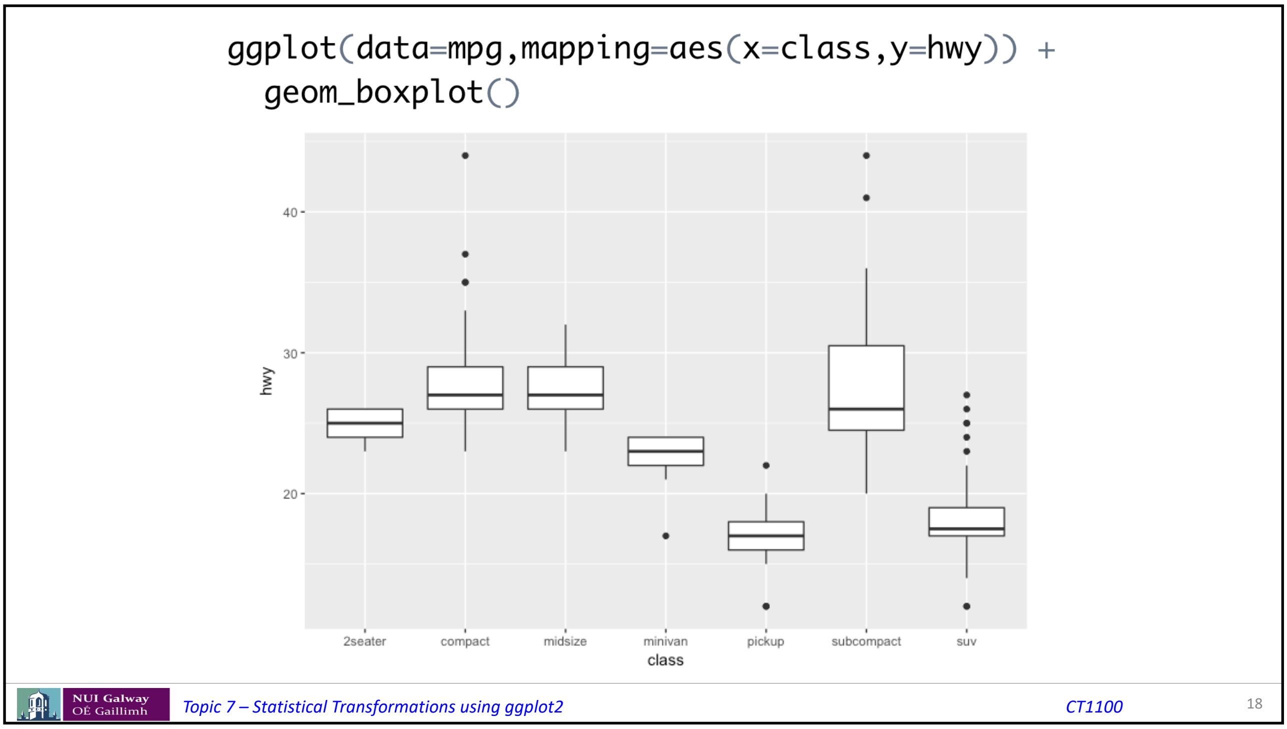
75th Percentile

50th Percentile

25th Percentile

Topic 7 – Statistical Transformations using ggplot2 CT1100 17

17



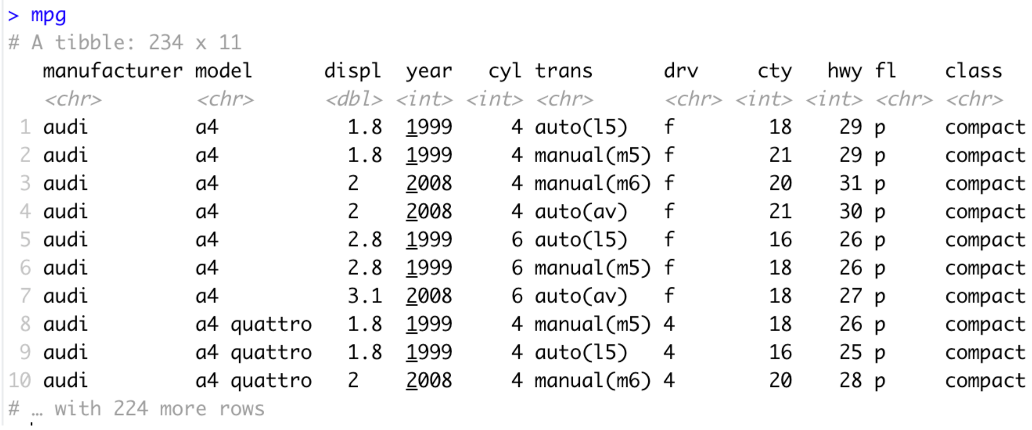
```
ggplot(data=mpg,mapping=aes(x=class,y=hwy)) +
  geom_boxplot()
```

Topic 7 – Statistical Transformations using ggplot2 CT1100 18

18

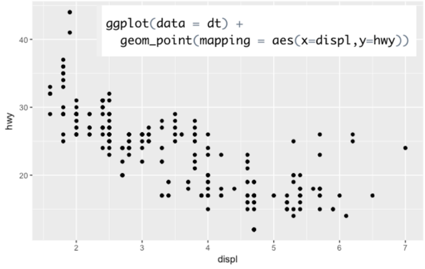
Challenge 7.3 – Statistical Transformation

- Draw a boxplot of hwy by manufacturer

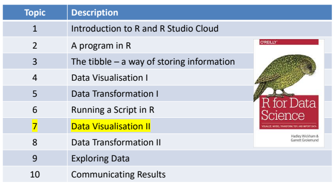


```
> mpg
# A tibble: 234 x 11
   manufacturer model      displ  year   cyl trans      drv     cty   hwy fl    class
   <chr>        <chr>      <dbl> <int> <int> <chr>      <chr> <int> <int> <chr> <chr>
 1 audi         a4           1.8  1999     4 auto(l5)   f        18    29 p     compact
 2 audi         a4           1.8  1999     4 manual(m5) f        21    29 p     compact
 3 audi         a4           2    2008     4 manual(m6) f        20    31 p     compact
 4 audi         a4           2    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    2008     4 manual(m6) 4        20    28 p     compact
# … with 224 more rows
```

Summary – ggplot2

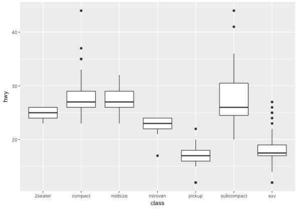


```
ggplot(data = dt) +
  geom_point(mapping = aes(x=displ,y=hwy))
```

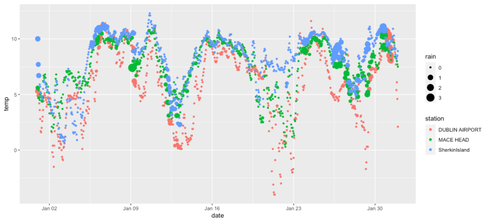


Topic	Description
1	Introduction to R and R Studio Cloud
2	A program in R
3	The tibble – a way of storing information
4	Data Visualisation I
5	Data Transformation I
6	Running a Script in R
7	Data Visualisation II
8	Data Transformation II
9	Exploring Data
10	Communicating Results

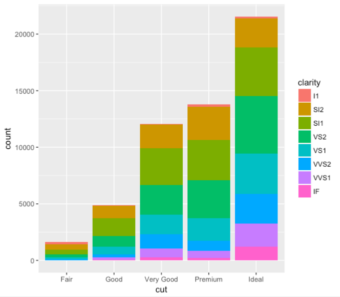
O'REILLY
R for Data Science
VISUALIZE, MODEL, TRANSFORM, TIDY, AND IMPORT DATA
Hadley Wickham &
Garrett Grolemund



hwy
2seater
compact
midsize
minivan
pickup
subcompact
suv
class



temp
Jan 02
Jan 09
Jan 16
Jan 23
Jan 30
date
rain
station
DUBLIN AIRPORT
MACE HEAD
SherkinIsland



count
Fair
Good
Very Good
Premium
Ideal
cut
clarity
I1
SI2
SI1
VS2
VS1
VVS2
VVS1
IF