2. Visualisation of 2D Data

This section will outline graphical visualisation for 2D (or bivariate) data according to the following data configurations:

- **both variables are continuous (e.g. visualising relationships)**
- **one variable is discrete and one continuous (e.g visualising subgroups)**
- **both variables are discrete (e.g. visualising tabular data)**

Both variables continuous (visualising relationships)

Scatterplot

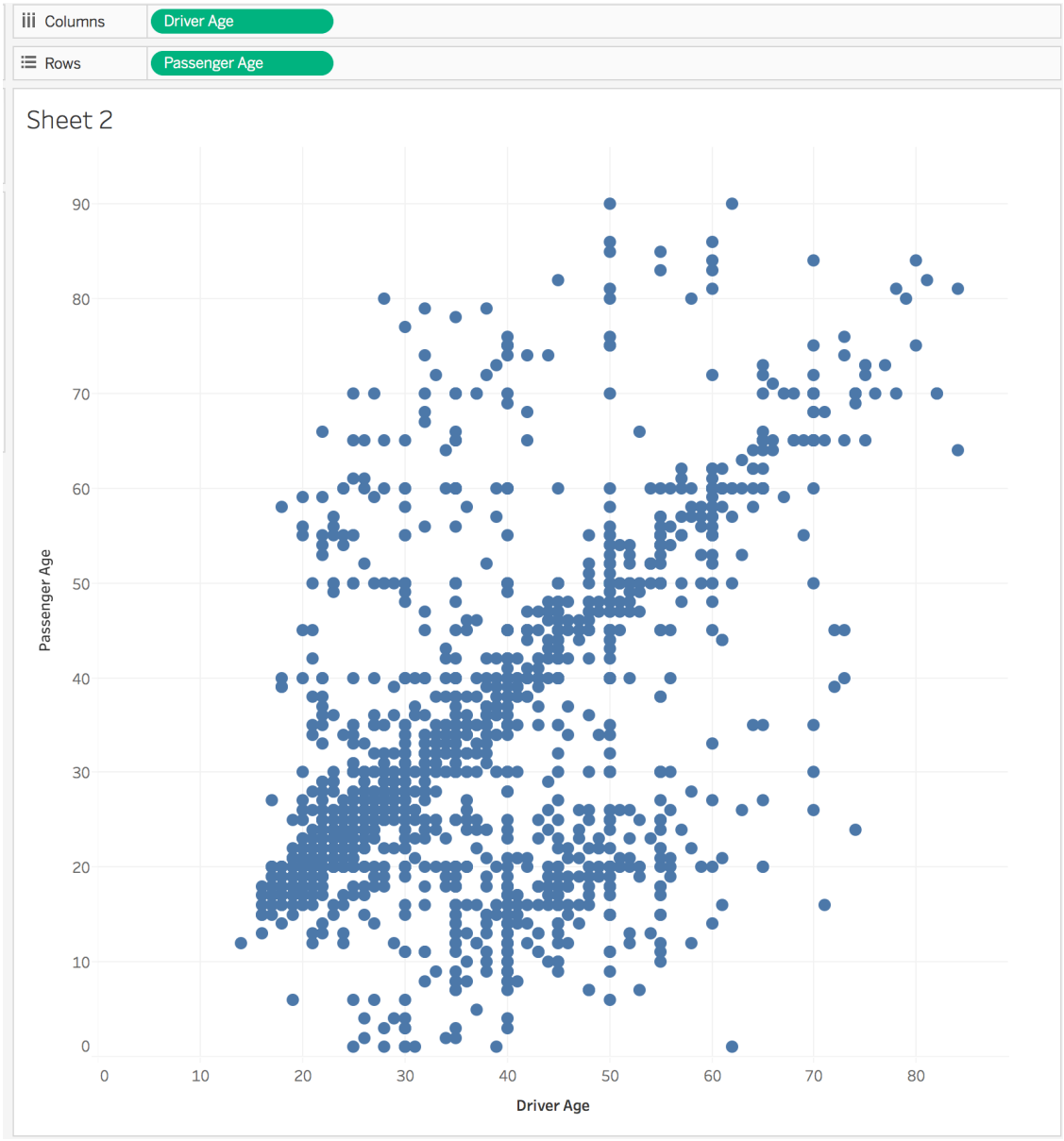
The most common graphic to visualise the relationship between two continuous variables is a scatterplot. This is a 2-dimensional graph that is used to investigate the **relationship** between two continuous variables. To draw a scatterplot one of the variables is plotted on the x-axis while the other variable is plotted on the y-axis. Each point on a scatterplot represents the x and y value for each data pair. The points on the plot are not joined by lines. Scatterplots are one of the most extensively used statistical graphics and are usually the first preliminary step taken when investigating the relationship between two variables. The co-ordinate system used to plot the data is known as **Cartesian** after the French philosopher **Renes Descartes** (1596-1650).

Example

Figure 2.1 is a random sample of the ages of 2,000 drivers and their front seat passengers involved in cars involved in serious injury accidents between 1995 and 2016. The full data set is provided in the worksheet **Driver and Front Seat Age** in the excel file *Data(2018).xls.*

To create the Tableau graph in Figure 2.1 select the worksheet **Driver & Passenger Age** in Tableau. Drag the **Driver Age** and **Passenger Age** variables to the Column and Row shelf, respectively and remember to deselect **Aggregate Measures** from the Analysis menu!

Figure 2.1 shows a strong linear trend through the diagonal together with evidence of two clusters in the top left and bottom right sections of the graph.



Example 2

Figure 2.2 plots the Adult Guidance grant (€) versus the Adult Literacy grant (€) awarded to 17 Education and Training Boards (ETBs) in 2016. There is a general positive relationship between the two grants i.e. the larger the guidance grant the larger the corresponding literacy grant. Note the data point on the top right hand side which is a departure from the trend with City of Dublin ETB having a noticeably higher literacy grant.

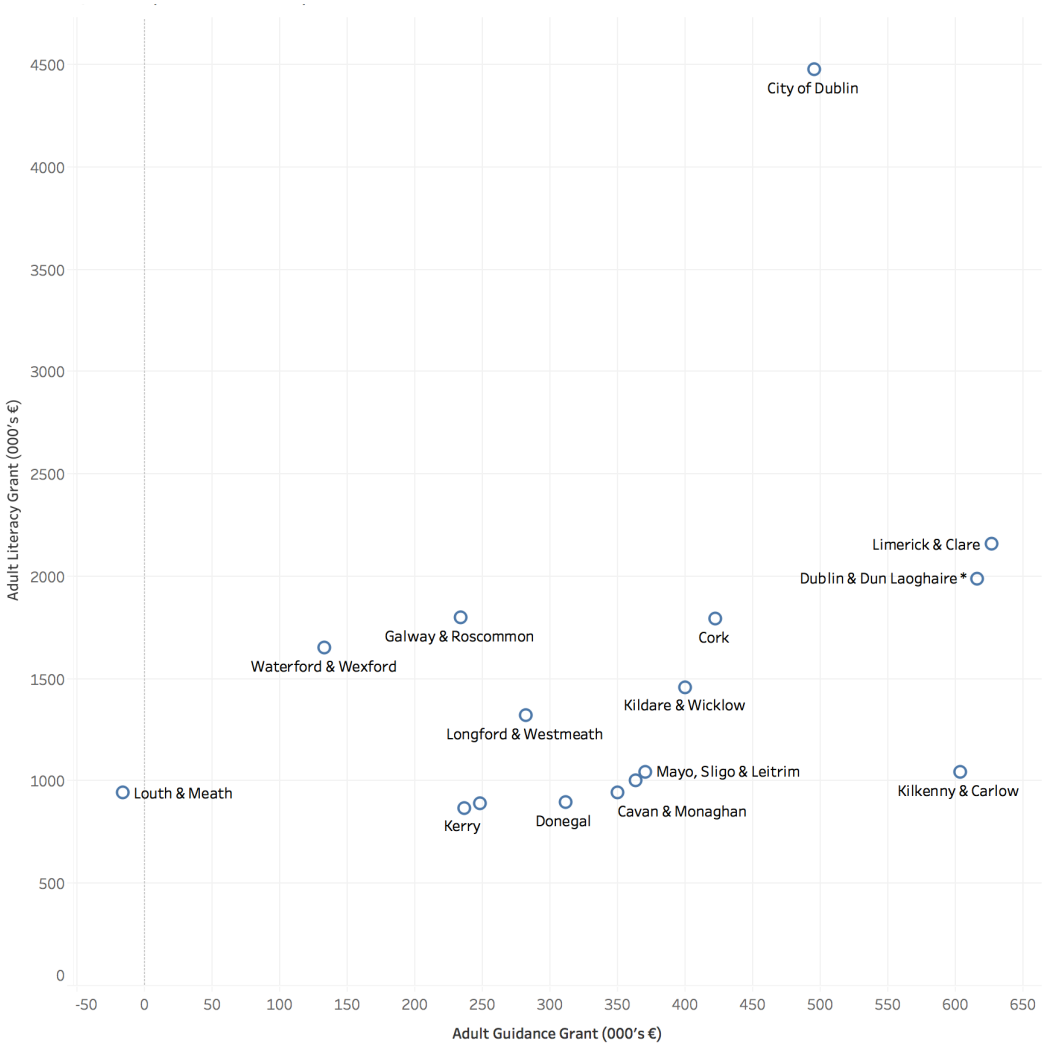


Figure 2.2 Scatterplot of Adult literacy (€'000) grant versus Guidance grant,2016

Example 3: Spatial Data

If a data set contains locational data it can be possible to geocode the data and create a continuous longitude and latitude variable for each record using on-line software like **batch geocode** [3]. However, in some instances locations might be provided at a higher level of spatial aggregation in which case the same longitude and latitude is given to many records. For example, Figure 2.3 illustrates the location at the suburb or town level of patients diagnosed with ecoli in the greater Dublin area.

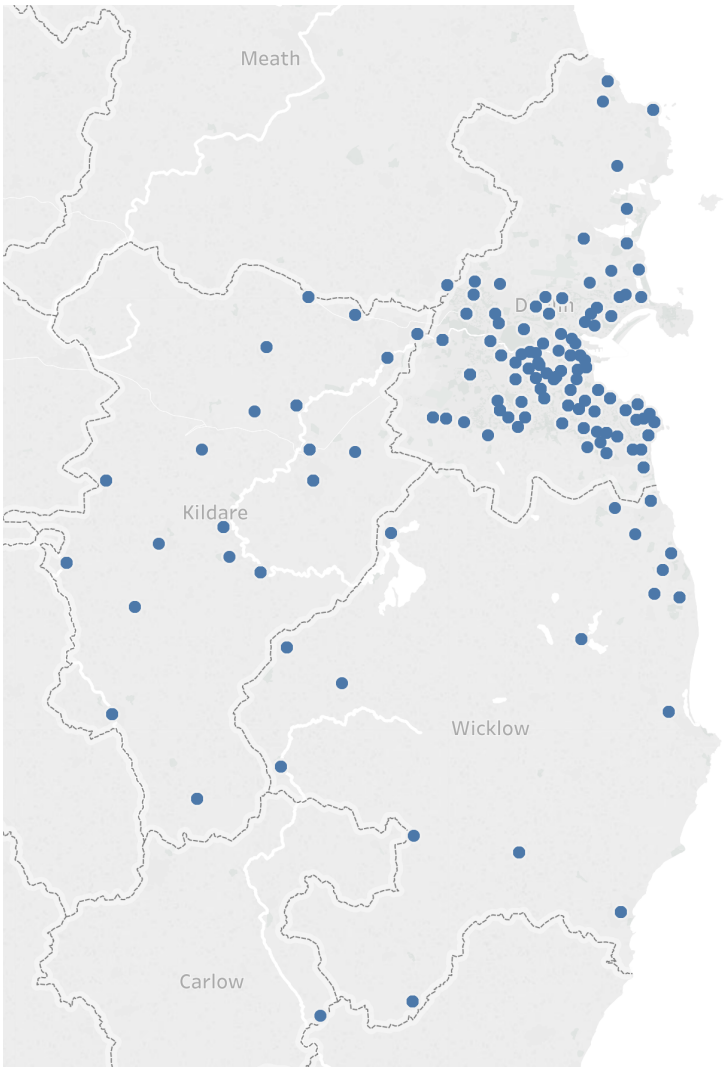


Figure 2.3: Plot of infection cases by town/suburb in greater Dublin area, 2015-16

Each point in Figure 2.3 represents a town or suburb where at least one infection has occurred. However, we cannot visualise the number of infections at each location. One partial solution in the absence of more granular data is to create a 2D jitter plot as shown in Figure 2.4.

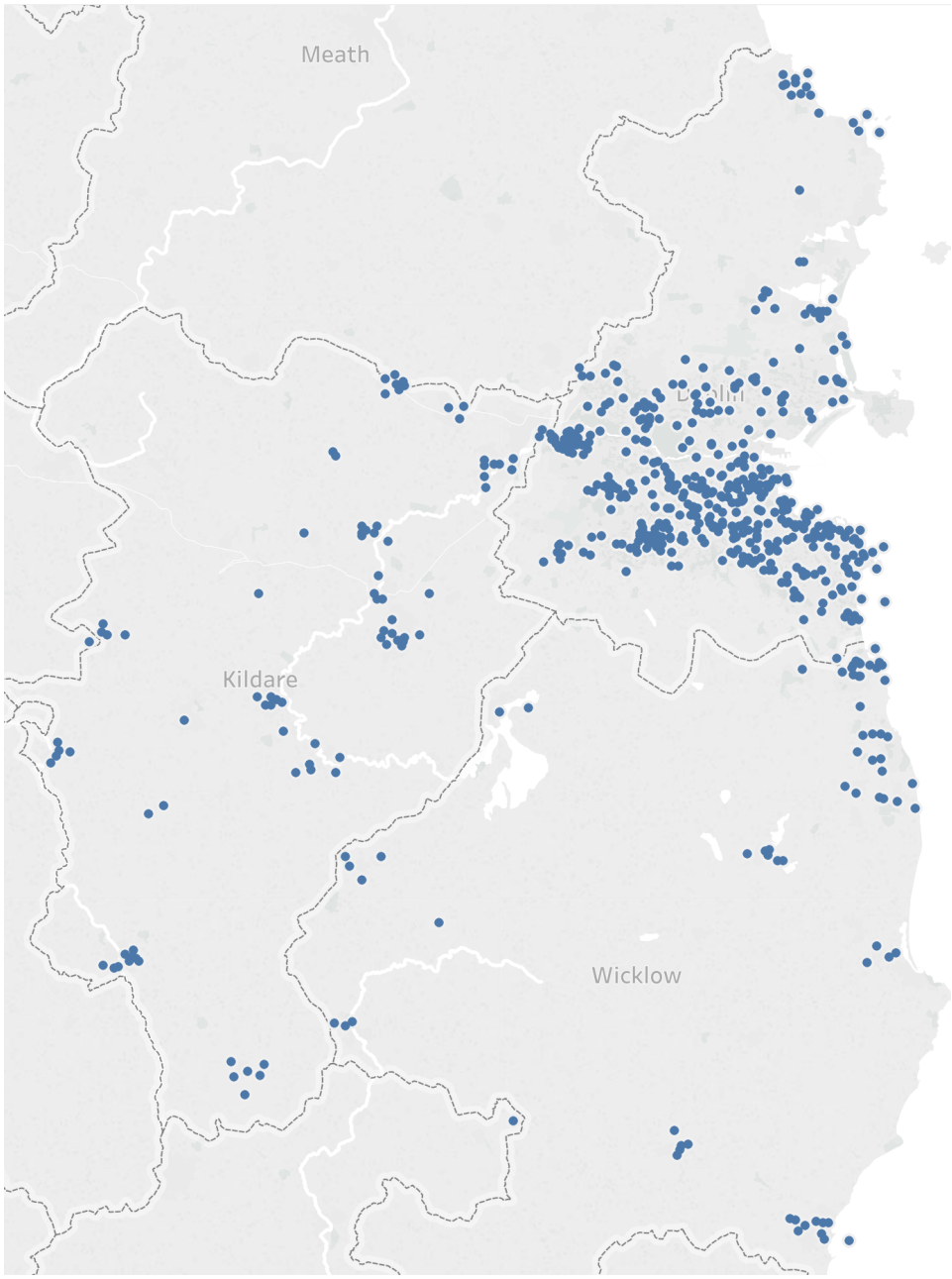


Figure 2.4: Jitter plot of infection cases by suburb, 2015-16

Figure 2.4 is an improved visualisation as all the individual cases of infection are now plotted for each area. It was created with the help of a statistical visualisation software application **Datadesk** by adding **noise** to the longitude and latitude of each record. This meant that no two records have the exact longitude and latitude. The technical details involve adding a standard normal variable with standard deviation 0.005 to the original longitude data and a standard normal variable with standard deviation 0.015 to the original latitude data. This 2D jitter plot could also be created in R using the geom_jitter() function.

Time Series Plots

If data have been collected over time and the time has been recorded then it is possible to plot data as a line of time series plot. In this example we are assuming that the time measurement and the quantity that is measured at each time stamp is **continuous**. Time series plots can reveal surprising trends and patterns which may otherwise remain hidden in a data set.

Example

Figure 2.5 contains 100 measurements of the paste height (in mm) of PCB boards measured before they passed into a wave solder of a computer manufacturing facility. The data are collected in time order while the data set is provided in the worksheet **Paste Height** in *Data(2018).xls*

.

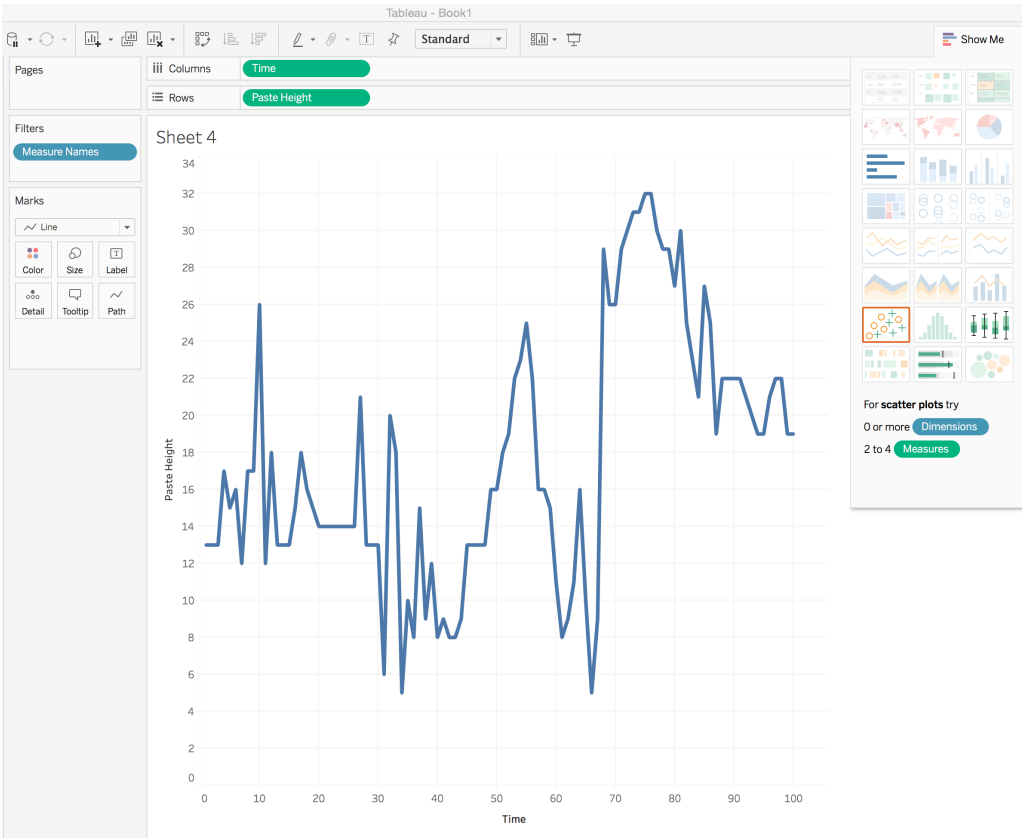


Figure 2.5: Paste height measurements (mm) from an industrial process

Plotting the data in Table it is clear that the process has shifted upwards from about sample point 70 as shown in Figure 2.5. Plotting this data as a one dimensional histogram or box plot would hide this shift. This example is a reminder that if the time order of data is has been recorded it can be a particularly useful variable.

To create the line plot in Tableau select the worksheet **Paste Height**. Place **Time** in the column shelf and **Paste Height** in the row shelf. Deselect Aggregate Measures from the Analysis menu and using the **Marks** panel change the symbol to **Line**. Note that if a data set has a time variable expressed as a **date** then Tableau can split the date into **years, quarters, months and days** as seen in Section 1. Statistical techniques for modelling line plots include **Time Series Analysis** and **Statistical Process Control (SPC)**.

ii) One discrete and one continuous variable (visualising subgroups)

This data configuration can be used to compare subgroups where the discrete variable is the **grouping variable**. For example, if we have a 2D data set of grants amounts (continuous) awarded to a number of educational training boards (ETB) (discrete) we can use the methods of this section to visualise the distribution of grants for each ETB by using the grouping variable ETB. Most of the methods outlined in this section can be considered extensions to the 1D techniques introduced previously in Section 1.

Multiway box plots

Multiway box plots are extensions to the 1D box-plot but in this case we have a box plot for each value of the discrete (grouping) variable. For example, if we want to visualise the distribution of motor insurance quotations for policyholders aged 20, 30 and 50 years of age then we can consider **quotation** (in €) as the continuous variable and **age** as the discrete grouping variable (as it has just three fixed values). We normally refer to the different values of the discrete variable as **levels**. In this example the discrete variable has three levels - those aged 20, 30 and 50 years of age as shown in Table 2.1.

Age of Consumer		
20	**30**	**50**
2,543	644	579
3,285	800	508
2,840	536	738
2,609	538	536
2,440	691	459
3,191	614	691
2,636	565	560
	664	404
	459	579
	668	666

Table 2.1: Insurance quotations (€) classified by age of policyholder

We can visualise this data set using a **multiway box plot** a **multiway violin plot** or a **multiway dot plot** where instead of just one geometric object as we saw in the 1D case we now have three geometric objects corresponding to the distribution of quotes for each of the three discrete age cohorts. A multiway box plot for this data is shown in Figure 2.6.

To create multiway box plots software applications usually require data to be entered in two columns. One column contains the continuous variable (quotes) while the other column the discrete variable (age of consumer). In this example the data would be entered in two columns containing 27 rows as shown in Table 2.2.

Quote (€)	Age
2,543	20
3,285	20
-	-
644	30
800	30
-	-
666	50

Table 2.2: Structure of data required for multiway box plots

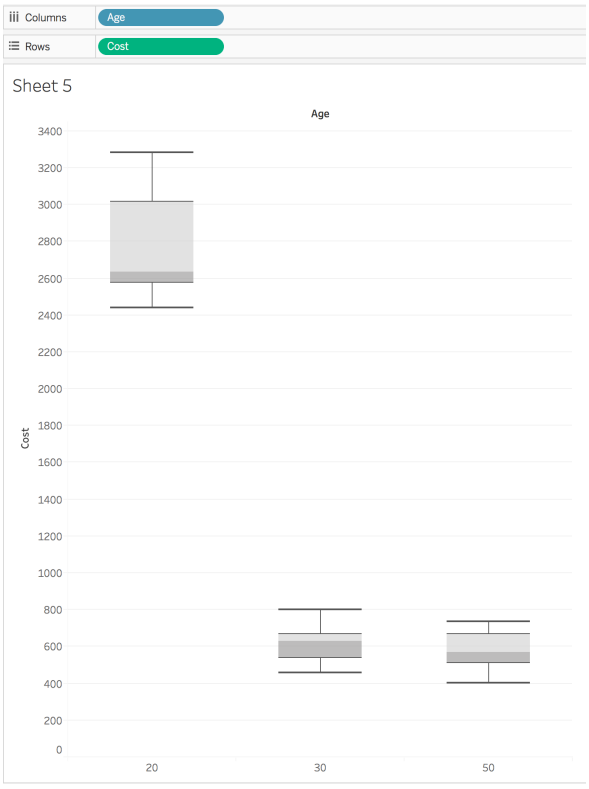


Figure 2.6: Multiway box-plots of motor insurance quotes (€) by age

To create Figure 2.6 select the worksheet **Insurance Quote** in Tableau. Place **Age** in the columns shelf and **Cost** in the rows shelf. Ensure that Age is placed in the Dimensions panel (as it is discrete) and that its format is set to **string** by selecting *string* from the **Change Data Type** drop down menu. Finally, deselect **Aggregate Measures** from the **Analysis** menu.

This data format is sometimes referred to as a **tidy** or **long** format. The original format where each level of the discrete variable has its own column i.e. 20, 30 and 50 is known as a **wide** format. It is best to convert your data to long format prior to analysis as this can provide a greater degree of flexibility to the analyst. The R graphics package **ggplot2** (and other statistical software) requires data to be in long format. The R library application **tidyr** can convert data from wide to long format readily and the procedure is outlined in Appendix 1. Tableau can also convert from wide to long format using the **pivot** command by selecting the age columns from the **Data Source** worksheet, right clicking on one of the column headers and selecting **pivot** from the drop down menu.

Facet/Small Multiple/Trellis Plot

We can also visualise data with one discrete and one continuous configuration using what is known as a **Facet** plot. Facet plots involve the construction of a plot enclosed in a panel for each value of the discrete grouping variable. Facet plots are also known as **Small Multiples** after Edward Tufte [4] or **Trellis plots** after **William Cleveland** [5,6]. To create a trellis plot using the previous example on insurance quotations we can use the code:

```
ggplot2(QuoteLong,aes("",Cost))+ geom_boxplot() +
facet_wrap(~Age)+xlab("Age of Consumer") +
ggtitle("Distribution of Quotations by Age")
```

This code contains a new component called **facet_wrap(~Age)**. This means create a plot for each age and place the panels in columnar format as shown in Figure 2.6. Facet plots can be extended to higher dimensions and can also be used for representing a wide range of data configurations which we will see later. The principal requirement for creating a Facet plot is that **at least** one variable must be a discrete (grouping) variable. The slight visual difference between Figure 2.6 and 2.7 is that ggplot2 places a boundary line around each boxplot.

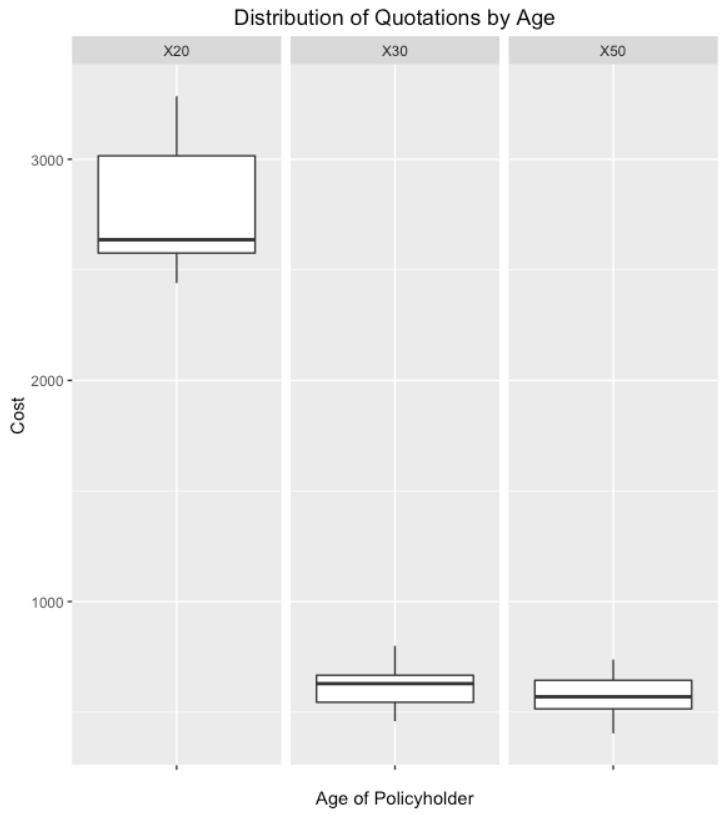


Figure 2.7: Trellis plot of motor insurance quotes (€) by age

iii) Both variables discrete (visualising tables)

In this section we will examine a number of graphical methods that can be used to visualise data when both variables are discrete. Data in this format is commonly represented as **tables** with the intersection of each row and column containing a **count**. We will outline a number of techniques for visualising this data configuration including bar, trellis, heatmaps, highlight tables and the lesser known mosaic charts.

Stratified bar charts/Trellis charts

When both variables are discrete we can extend the bar chart in the previous section on 1D data and plot a Trellis or stratified (side by side) bar chart. For example, the number of registered students classified by gender (discrete) and course (discrete) in iadt between 2011 and 2015 is provided in the worksheet **2D gender by course** in *Data(2018).xls*. This data is in provided in both wide and long formats. Either format can be handled by Tableau. For the wide format Tableau puts the Male and Female columns into a combined variable called **Measure Names**. The number of registrations are placed into another variable pill called **Measure Values** as shown in Figure 2.8.

However, for further analysis it is generally better to convert the data to long format with two columns representing the variables **gender** and **course** with a third column a count of the number of students in each gender and course combination.

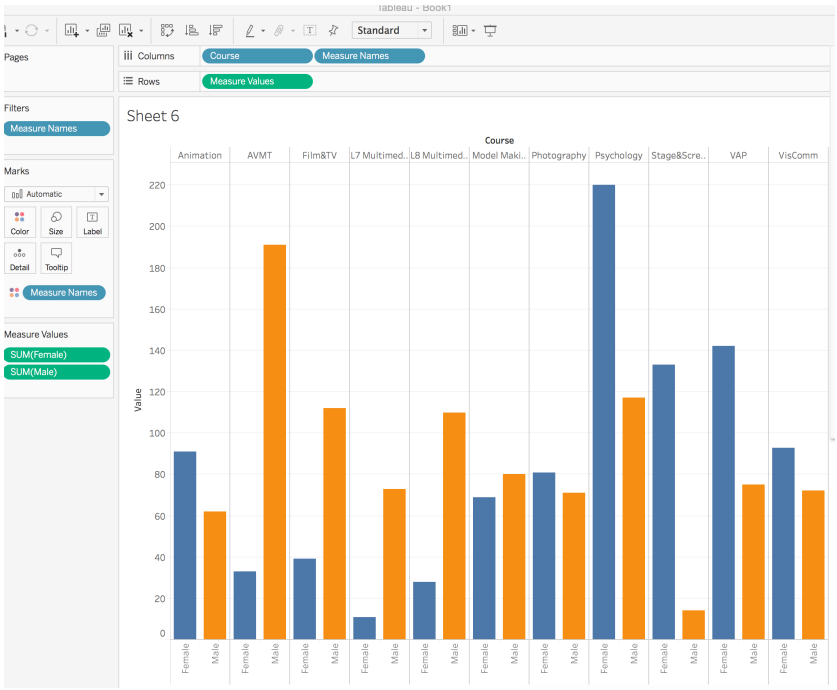


Figure 2.8: Trellis plot of gender distribution by programme, Faculty of Film, Art & Creative Technology, 2009-13

In this example the y-axis is a **count** which is not a variable - hence this plot can be considered 2D (course, gender) rather than 3D (course, gender, count). Note that the above chart is strictly a trellis chart as each course is delineated by a line representing an individual panel. Without this vertical line is becomes the more traditional stratified bar chart.

.

Another example of Trellis plots is shown in Figure 2.9 which is based on road traffic accidents in Ireland by weekday and accident type between 2005 and 2013. The data are shown in the Excel worksheet **RoadAccident** in *Data(2018).xls* and provides two discrete variables **weekday** and **type**. This data set is 2D with both variables discrete. A trellis plot using Tableau can be generated by selecting the worksheet **Road Accident by Type and Day**. Place **Weekday** on the Columns shelf and **Type** and **Number** on the Rows shelf - though the analyst can experiment by interchanging the row and column variables as required.

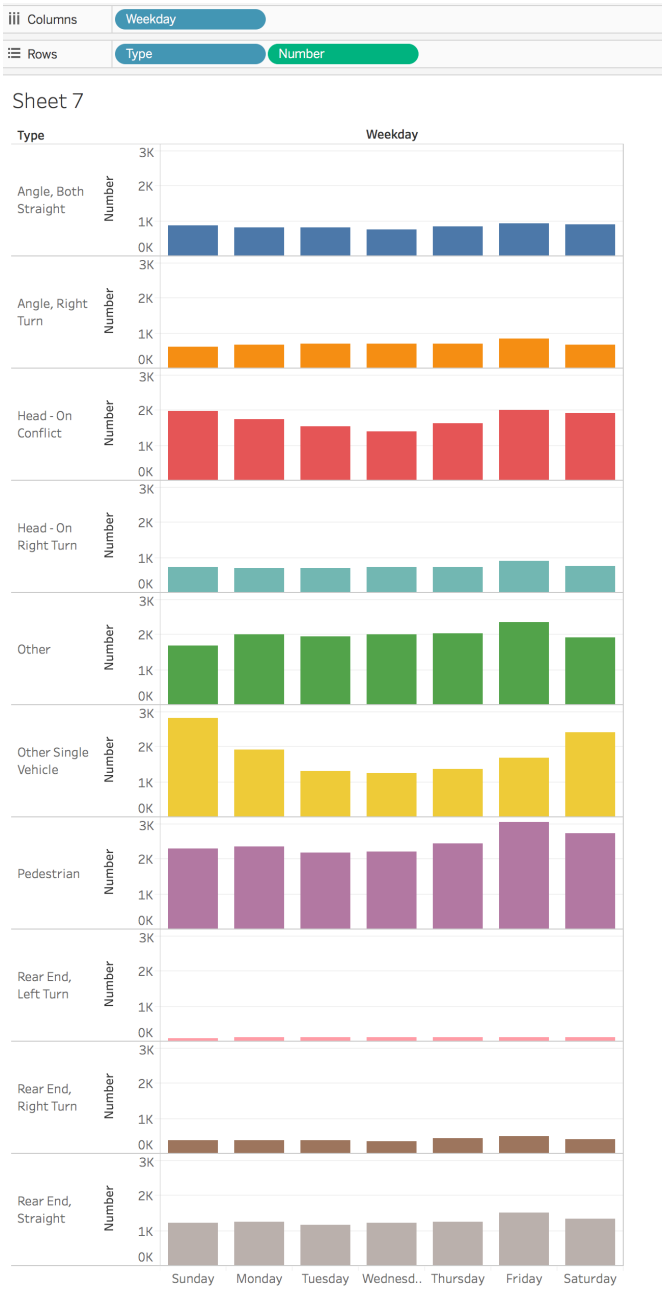


Figure 2.9: Trellis plot using Tableau of road traffic accidents by accident type and weekday, 2009-13

As mentioned earlier the R graphics package ggplot2 can also create Trellis plots as shown in Figure 2.10 and provides some additional flexibility by allowing the analyst to specify the number of rows and columns in the Trellis plot by using **nrow** and **ncol** arguments as shown on page 32.

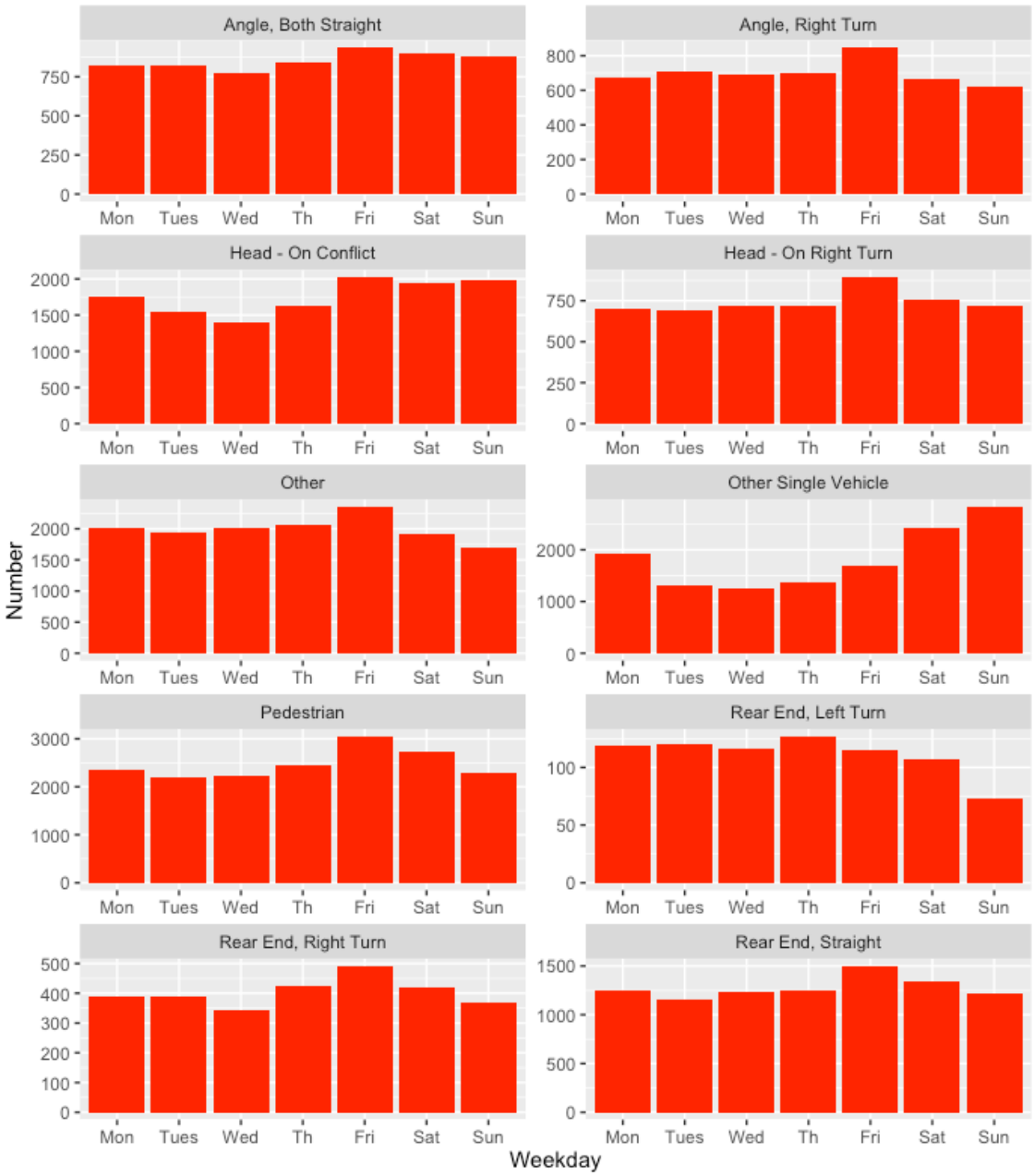
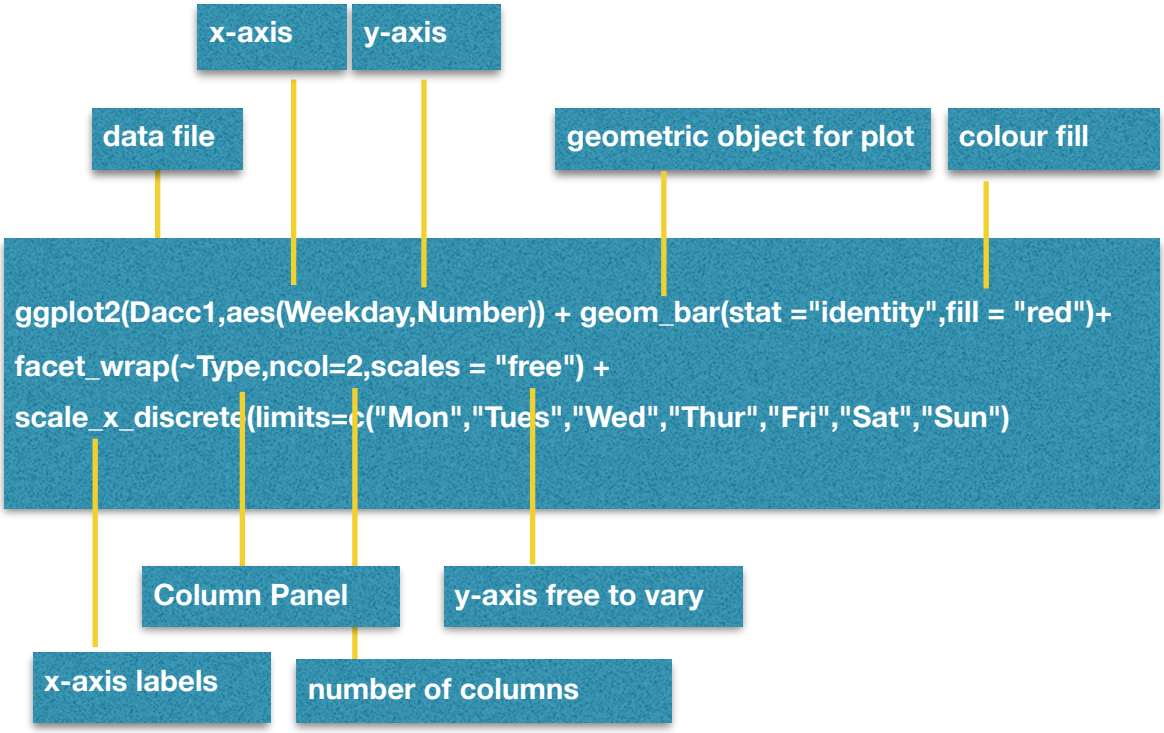


Figure 2.10: Trellis plot using ggplot2 of road traffic accidents by accident type and weekday, 2009-13.

The visualisation produced in Figure 2.10 is generated using the ggplot2 code:



The data file is **Dacc1** with Weekday and Number plotted on the x and y axis, respectively. The **geom_bar** is a bar chart filled with the colour red while the **facet_wrap** assigns the trellis panels to accident type. The y-axis scale which is a count of each accident type by day is free to vary as the code specifies **scales = "free"**. The last line **scale_x_discrete** forces ggplot2 to place the days in the order Mon, Tues etc.

Heatmap

Heatmaps are a simple visualisation technique that can be used to give a general overview of information provided in a table. Heatmaps represent the data in a cell using symbols which are scaled according to the size of the data in each cell. For example, the plot below is a Tableau heat map of the distribution of cancer type by smoking status between 2009 and 2014 where the value codes for smoking status are **C** = Current, **N** = Never Smoked, **X** = Ex Smoker **Z** = Unknown.

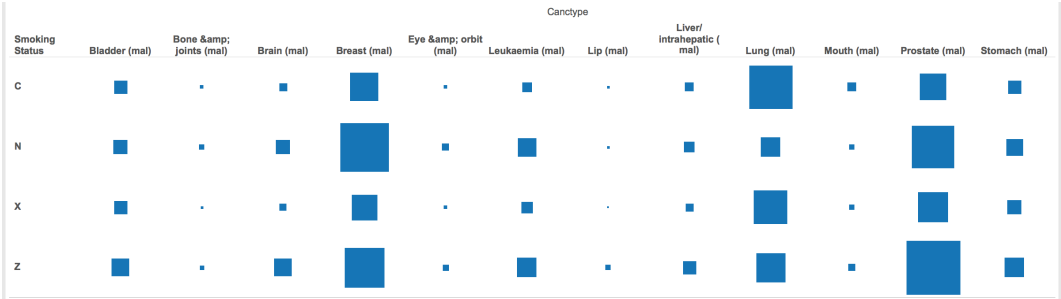
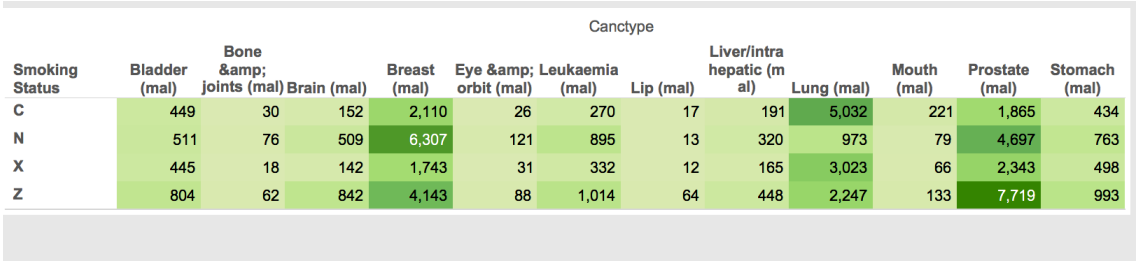


Figure 2.11: Heatmap of cancer diagnosis by smoking status

The plot is created by selecting the worksheet **Heatmap2D** in Tableau. Place **Canctype** in the column shelf and **Smoking Status** in the row shelf. Finally select the Heatmap icon on the **Show Me** graph palette in Tableau. The source data are provided in the worksheet **smokingcancer** in *Data(2018).xls*.

Highlight Tables

Highlight tables are another simple visualisation technique which encodes discrete data in a table such that the larger the data the deeper the colour saturation and vice versa. A highlight table for cancer by smoking status is illustrated in Figure 2.12.



	Canctype											
Smoking Status	Bladder (mal)	Bone & joints (mal)	Brain (mal)	Breast (mal)	Eye & orbit (mal)	Leukaemia (mal)	Lip (mal)	Liver/intra hepatic (mal)	Lung (mal)	Mouth (mal)	Prostate (mal)	Stomach (mal)
C	449	30	152	2,110	26	270	17	191	5,032	221	1,865	434
N	511	76	509	6,307	121	895	13	320	973	79	4,697	763
X	445	18	142	1,743	31	332	12	165	3,023	66	2,343	498
Z	804	62	842	4,143	88	1,014	64	448	2,247	133	7,719	993

Figure 2.12: Highlight table of cancer diagnosis by smoking status

The plot is created by selecting the worksheet **Heatmap2D** in Tableau. Place **Canctype** in the column shelf and **Smoking Status** in the row shelf. Finally select the highlight table icon on the **Show Me** graph palette in Tableau. The source data are provided in the worksheet **smokingcancer** in Data(2018).xls.

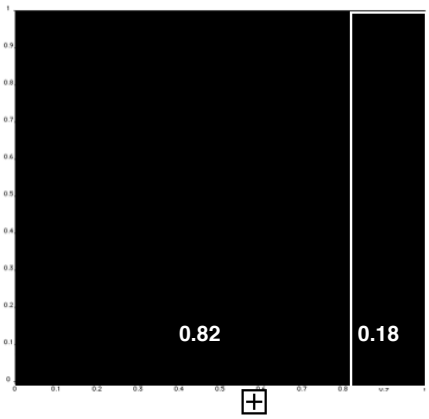
Mosaic Plots

Mosaic plots are a rarer graphic which visualises data configured as a table where each cell is a **count**. Mosaic plots represent the cell counts using '**tiles**' whose size is proportional to the cell count. Mosaic plots are generally used to provide a graphical assessment of the relationship between the rows and columns of a table. Despite their effectiveness as visualisation tools few software applications include mosaic charts.

	17-18	19-20	21-24	TOTAL
Male	124	300	1,273	1,697
Female	43	89	240	372
TOTAL	**167**	**389**	**1,513**	**2,069**

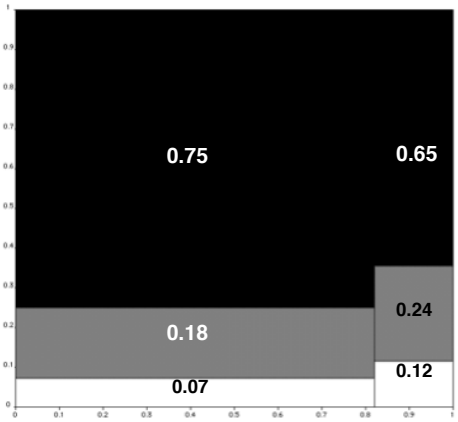
To calculate a mosaic plot for this data we first decide what variable to place on each axis. Selecting gender for the x-axis and age for the y-axis we implement the following three steps:

Step 1: Divide the x-axis axis proportionality according to **sex** as shown below.



There are 1,697 males and 372 females out of a total of 2,069 policyholders. Therefore the proportion of males is 1,697/2,069 = 0.82 and the proportion of females is 372/2,069 = 0.18. The x-axis is then divided 0.82 for males and 0.18 for females

Step 2 Divide the y-axis proportionality according to **age** for each sex as shown below.



The number of males aged 17-18 is 124 while the total number of males is 1,697. Therefore the proportion of **males** aged 17-18 is 124/1,697 = 0.07. The calculation is computed for all six cells for both genders as shown in the table overleaf.

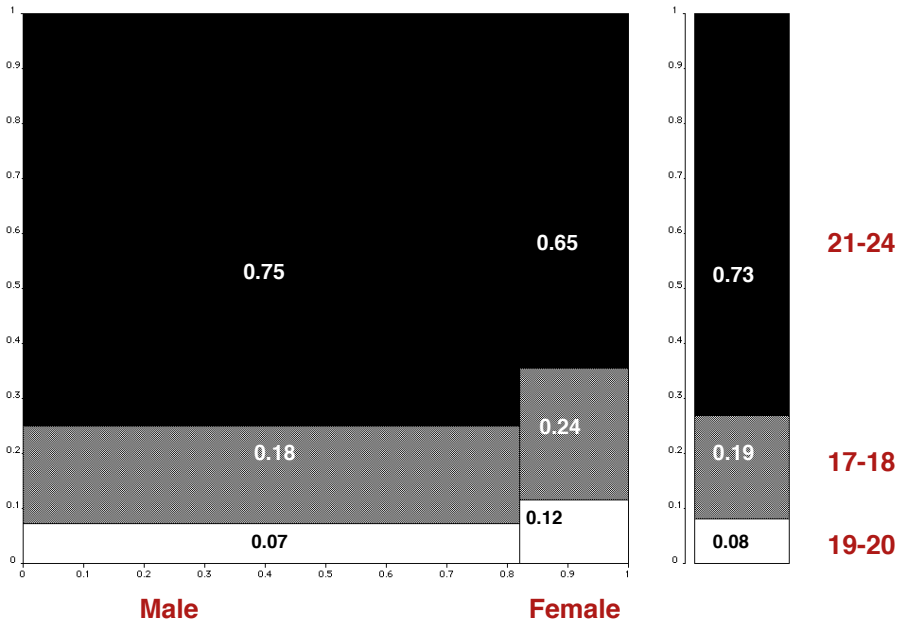
	17-18	19-20	21-24	Total
Male	0.07	0.18	0.75	1.00
Female	0.12	0.24	0.65	1.00

Step 3: Calculate the proportion of **all** policyholders according for each age i.e. the overall proportion of 17-18, 19-20 and 21-24 year olds.

There are 167 policyholders aged 17-18 therefore the proportion of **all** policyholders aged 17-18 is 167/2,069 = 0.08. The remaining entries are shown in the table below:

	17-18	19-20	21-24	Total
Proportion	0.08	0.19	0.73	1.00

These proportions are normally represented as a side bar adjacent to the plot as shown in the final plot below.



The width of the x-axis provides a graphical assessment of the proportions of male and female exposure (0.82 and 0.18) while the y-axis gives the proportion of male and female exposure within each gender. In addition, the side bar on the right of the plot gives the overall proportions of 17-18, 19-20 and 21-24 policyholders in the table.

Mosaic plot are a very effective visual representation of the principal characteristics of data that are organised in tables. In this example they allow the viewer to see, at a glance, how the distribution of policyholder age varies according to policyholder gender.

For example, from the plot it is clear that there is a higher proportion of female policyholders aged 17-18 compared with males aged 17-18. The overall female/male proportions are also visible from the x-axis while the overall distribution of ages can be visualised from the side bar. The side bar shows the proportion of policyholders we should expect in each age cohort assuming there is no relationship between age and gender. We can see from inspection of the plot that there is a slight overrepresentation of female 17-18 year olds and an under representation of female 21-24 year olds.

Mosaic plots can be extended to visualise three and higher dimensions which we will see in Section 3.Tableau does not provide mosaic plots but they can be generated using the R graphing library package **vcd** (visualising categorical data) as shown in Figure 2.13. The package vcd does not provide a side bar or scaling from 0 to 1 which are shown in Figure 2.13 which was programmed using a now unsupported Apple software application called **Hypercard**.

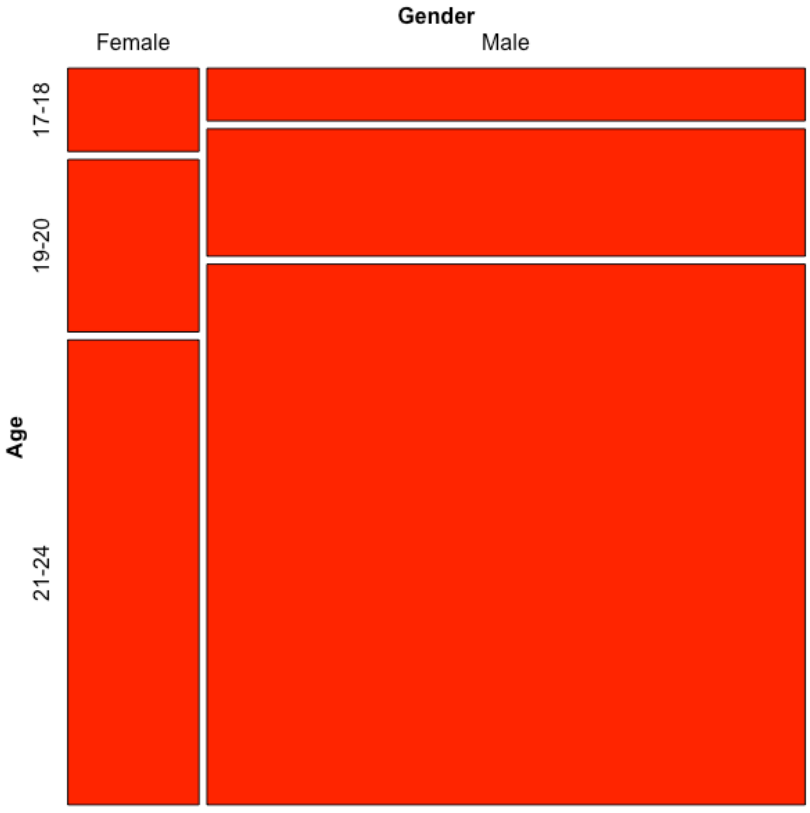


Figure 2.13: Mosaic plot using R library vcd

The R code for drawing this mosaic plot is shown below together with an explanation of each of the terms.

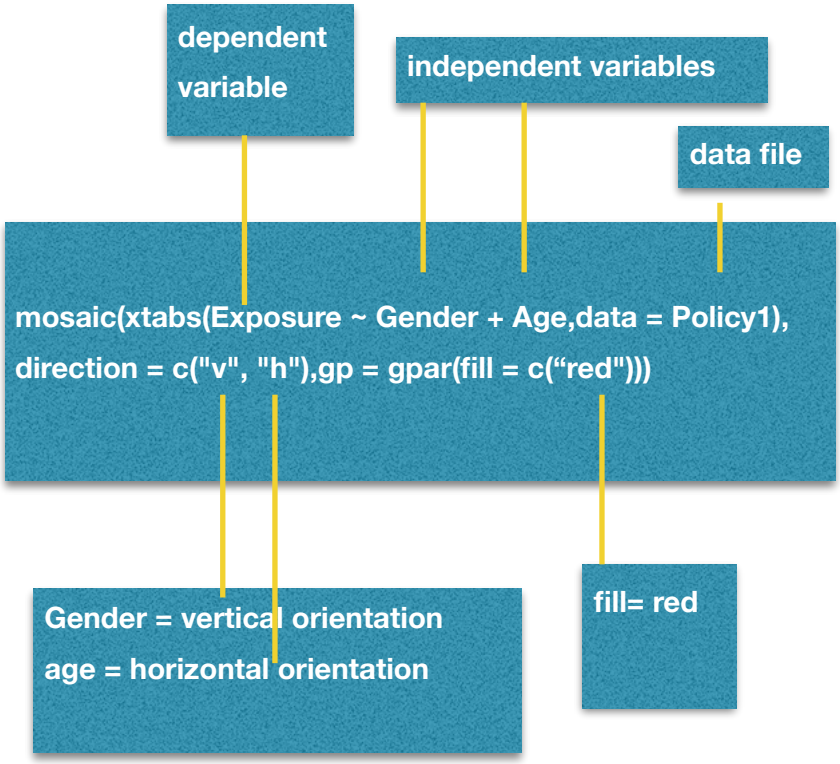


Figure 2.14: Code for creating mosaic plot using R package vcd

Exercises

. The data in the Excel worksheet **grant2D** in the *ExerciseData(2018).xlsx* records the grant (€ 000s) awarded to 16 education and training boards classified by category of award in 2016.

i) State giving a reason if this data set is 1D, 2D or MD?

ii) List the name of each variable in the data set and state if they are discrete or continuous.

iii) Using Tableau or R compute a trellis plot with where each panel represents an **etb** category for the following graphic types:

box plot
dot plot
violin plot
histogram.

iv) Using the results of iii) summarise the principle features of the distribution of grant awards.

v) Write a brief note on which of the above plots you believe is most effective at communicating the distribution of grant awards.

2. The following table presents the cross-classification of 1,490 passengers and crew by gender and class who died on the Titanic on April 15, 1912.

Gender	First Class	Second Class	Third Class	Crew
Male	118	154	387	670
Female	4	13	89	3

i) State giving a reason if this data set is 1D, 2D or MD?

ii) List the name of each variable in the data set and state if they are discrete or continuous.

iii) Compute a **mosaic** plot for this data set using class as the x-axis

iv) Using i) summarise the principle features of this data set

3. The data set below is based on the average premium for comprehensive cover (€) paid for 24 policyholder segments for eight companies operating in the Irish market.

Segment	1	2	3	4	5	6	7	8
17-20,Female,Full	1,043	-	1,643	1,952	1,882	1,456	1,735	1,452
17-20,Female,Prov	1,124	400	1,705	1,706	2,467	2,055	2,220	1,877
17-20,Male,Full	1,716	705	2,388	2,152	2,675	2,204	2,687	1,842
17-20,Male,Prov	2,272	1,480	5,776	2,115	4,015	3,613	3,183	2,365
21-24,Female,Full	791	1,167	853	1,062	1,010	844	947	1,075
21-24,Female,Prov	1,165	1,320	1,209	1,322	1,725	1,281	1,301	1,468
21-24,Male,Full	1,507	1,049	1,454	1,582	1,749	1,598	1,874	1,352
21-24,Male,Prov.	2,095	1,586	1,333	1,914	2,581	2,005	2,598	1,849
25-30,Female,Full	598	526	484	573	600	545	607	603
25-30,Female,Prov	920	695	599	729	1,036	922	943	771
25-30,Male,Full	1,039	845	646	800	1,024	1,011	912	730
25-30,Male,Prov	1,476	1,153	717	918	1,792	1,536	1,634	986
31-50,Female,Full	425	477	425	443	555	455	523	551
31-50,Female,Prov	733	511	619	730	740	712	729	593
31-50,Male,Full	756	652	520	525	651	643	612	572
31-50,Male,Prov	1,228	843	637	680	889	971	828	638
51-70,Female,Full	381	475	425	426	569	463	543	575
51-70,Female,Prov	626	501	706	834	695	617	662	702
51-70,Male,Full	589	617	487	490	653	622	736	626
51-70,Male,Prov	1,021	766	739	829	794	971	695	672
over 70,Female,Full	292	514	475	475	496	397	448	537
over 70,Female,Prov	517	648	586	595	570	533	547	801
over 70,Male,Full	448	623	504	531	579	569	679	590
over 70,Male,Prov	976	830	628	750	665	734	883	816

i) State giving a reason if this data set 1D, 2D or MD?

ii) List the name of each variable in the data set and state if they are discrete or continuous.

iii) Compute a highlight chart, a heat map and a Trellis plot for this data set.

iv) Which graphic do you think is most suitable for this data set?

v) Assuming you are a company considering entering the Irish market. Based on the visualisations computed in iii) what segments would you consider entering?

4. The data set below represents the age at which full-time education stopped for Dublin City residents in 2016 classified gender of policyholder

Gender	18	19	20
Male	16,433	5,049	4,921
Female	18,138	6,319	5,749

i) State giving a reason if this data set is 1D, 2D or MD?

ii) List the name of each variable in the data set and state if they are discrete or continuous.

iii) Compute a **mosaic** plot for this data set using age on the x-axis

iv) Using i) summarise the principle features of this data set.

5. The data in the Excel worksheet **Apprentice2D** in the file *ExerciseData(2018).xlsx* records the number of apprentices classified by type of apprentice and educational training board (etb) in wide and long data formats.

i) State, giving a reason, if this data set is 1D, 2D or MD?

ii) List the name of each variable in the data set and state, giving a reason, if they are discrete or continuous variables.

iii) Using Tableau or R compute the following graphics:
multiway box plot
multiway dot plot plot
Let the x-axis represent apprentice type and the y-axis the number of apprentices
(**note**: etb is not required for this section)

iv) Using Tableau or R compute:

a) Trellis bar plot where each panel represents an apprentice type.
Let the x-axis represent etb and the y-axis the number of apprentices.

b) Using Tableau compute a **heatmap** and a **highlight table**.

6. The data in the table below records the number of H. Influenza cases by clinical diagnosis between 2004—2014. The raw data is provided in the Excel worksheet **Influenza2D** in the file *ExercisesData(2018).xlsx*.

	Year										
Clinical Diagnosis	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
Septicaemia	8	14	13	6	3	9	9	11	11	14	15
Pneumonia	5	0	3	6	3	8	5	12	12	4	12
Meningitis	3	9	3	2	2	2	1	3	2	2	7
Bacteraemia (w/o focus)	1	0	1	1	2	0	0	3	5	6	9
Other	1	2	1	0	0	0	0	3	4	7	7
Epiglottitis	1	3	3	1	1	0	2	0	0	3	1
Cellulitis	1	1	2	1	1	0	0	1	0	0	0
Meningitis & Septicaemia	1	0	1	0	1	1	1	1	1	0	0
Osteomyelitis	1	0	0	0	0	0	0	0	0	0	0
Septic arthritis	0	1	0	0	1	0	0	0	0	0	0
Not specified	16	4	11	14	8	23	10	10	6	5	10

i) State giving a reason if this data set is 1D, 2D or MD?

ii) List the name of each variable in the data set and state if they are discrete or continuous.

iii) Create a trellis plot for this data set with a panel for each Diagnosis

iv) Using i) summarise the principle features of this data set.

7. The data in the Excel worksheet **Cancer2D** in the file *ExerciseData(2018).xlsx* records the number of months since diagnosis classified by cancer type between 2009 and 2013.

i) State giving a reason if this data set 1D, 2D or MD?

ii) List the name of each variable in the data set and state if they are discrete or continuous.

iii) Compute a multiway box, violin and jitter plot for this data set.

iv) Using iiii) summarise the principle features of this data set

v) Which of the three graphics in computed in iii) do you think is most suitable for this data set?

8. The disease classification and gender of 641 patients are provided in the Excel worksheet **Infection2d** in *ExerciseData(2018).xlsx*.

i) State giving a reason if this data set is 1D, 2D or MD?
ii) List the name of each variable in the data set and state if they are discrete or continuous.
iii) Compute a **mosaic** plot for this data set using patient type on the x-axis
iv) Using i) summarise the principle features of this data set.

9. The average cost per claim (€ 000s)for comprehensively insured vehicles between between 1997 and 2015 is provided in the Excel worksheet **CostComp** in *ExerciseData(2018).xlsx*.

i) State giving a reason if this data set is 1D, 2D or MD?
ii) List the name of each variable in the data set and state if they are discrete or continuous.
iii) Compute an appropriate plot for this data set
iii) Using i) summarise the principle features of this data set.

10. The number of policies for the top four companies in the Irish market for comprehensive and third party fire & theft cover in 2015 is provided in the Excel worksheet **MarketShare** in *ExerciseData(2018).xlsx*.

i) State giving a reason if this data set is 1D, 2D or MD?
ii) List the name of each variable in the data set and state if they are discrete or continuous.
iii) Compute a **mosaic** plot for this data set using Company id on the x-axis
iv) Using i) summarise the principle features of this data set.