3. Visualisation of Multidimensional (MD) Data

3.1 Introduction

Multidimensional data (MD) describes data sets where there are more than two variables. Visualisation of static MD data can be more difficult than 1D and 2D data as the printed page is limited to 2 dimensions or variables. However, data with three or more variables. can be visualised by extending the 2D graphic designs seen in earlier sections and by representing additional variables using **patterns**, **colours**, **shadings**, **sizes** and **labels**. These variables are referred to by Jacques Bertin in his text the *Semiology of Graphics* [7] as **differential variables**. As with 1D and 2D data the appropriate graphical representation will depend on whether the variables are discrete or continuous. In this section we will examine MD extensions to 2D scatterplots, trellis plots and mosaic plots. We will also illustrate the use of polygon maps to visualise data by county, electoral division (ED) and small areas (SA).

3.2 Scatterplot Maps

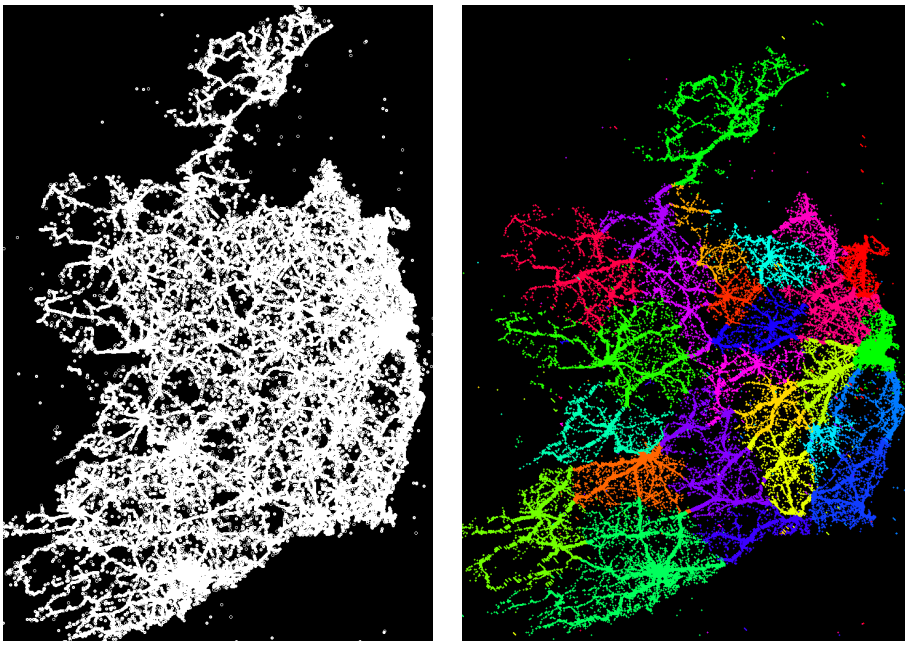


Figure 3.1: Plot of 2D road traffic accident data (left) and 3D data (right) with county encoded as colour.

If geocoded data is plotted a useful spatial scatterplot can be obtained.
We saw this in the previous section where the location of influenza cases was plotted. For example Figure 3.1 is a plot of the easting and northing of road accidents in Ireland

between 2004 and 2013. This map is in effect a 2D scatterplot with longitude and latitude representing two continuous variables which are plotted as an x,y pair. By encoding the county where the accident occurred using colour we now have a 3D representation. This is a 3D plot with the third dimension represented as colour of county. In this example longitude and latitude are continuous and county is discrete.

3.2 Bubble Plots

Another example of introducing additional dimensions to a 2D plot is to create a **Bubble Plot.** This is also a scatterplot with the size of the point or circle value linked to the numerical value of a third variable. The larger the numerical value of the third variable the larger the area of the circle and vice versa.

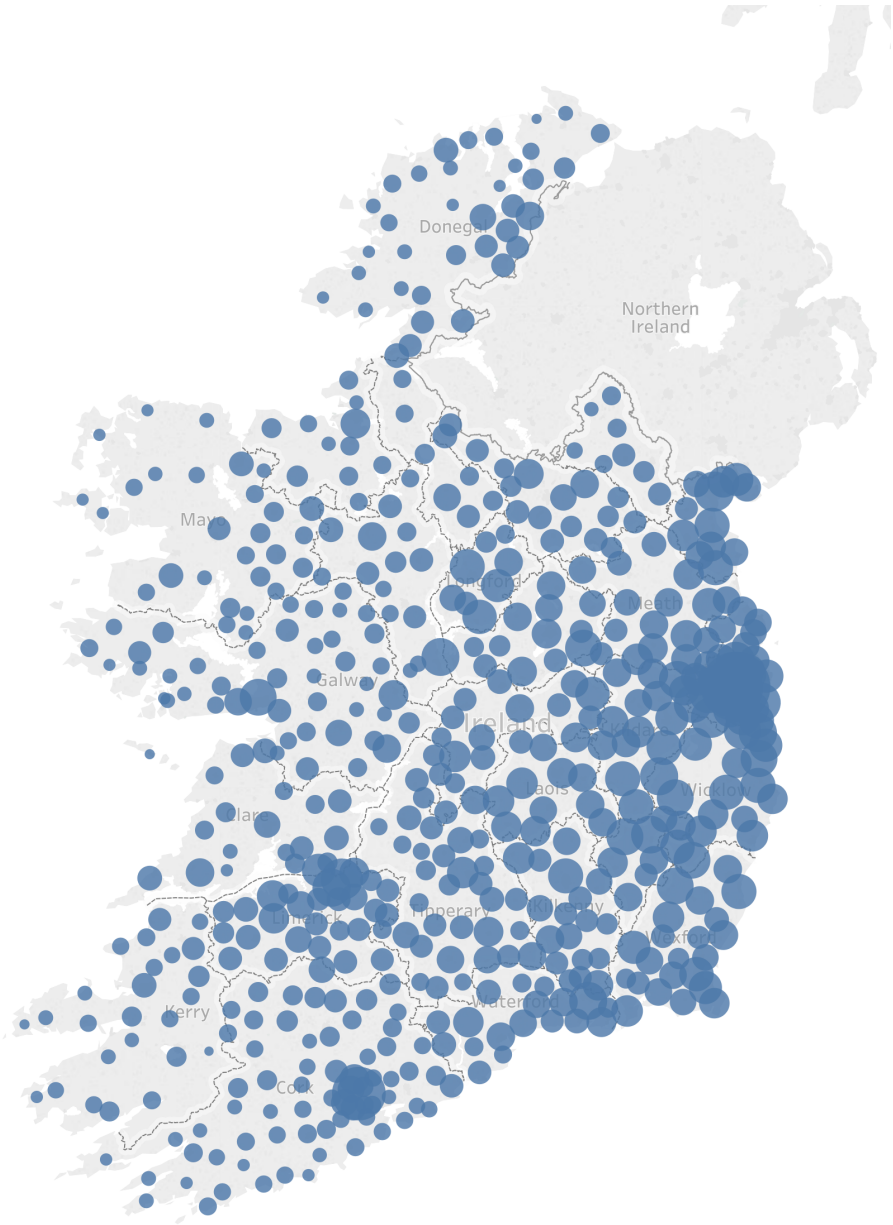


Figure 3.2: Bubble plot of Burglary rate per 100,000 population by garda station, 2003-2015

For example,the bubble plot in Figure 3.2 plots the number of *burglary and related crimes* reported to over 500 Garda Stations in Ireland. The data file is in the worksheet **Burgulary3D** in the Excel file *Data(2016).xlsx.* This plot is a 3D bubble plot with all three variables continuous - *longitude* and *latitude* of Garda Station and size of circle mark based on the number of crimes. To create a bubble plot in Tableau select the worksheet **Burglary by Garda Station**. Place *Longitude* on the column shelf, *Latitude* on the row shelf and drag *number of reported accidents* onto the size icon in the Marks menu. The plot can be extended to 4D by placing *Converting to English Garda Station* onto the label icon in the **Marks** section.

3.3 MD Trellis Plots

Many 2D plots can be extended to represent higher dimensions. For example a 2D dot plot of the location of students entering IADT in 2016 is shown in Figure 3.3.

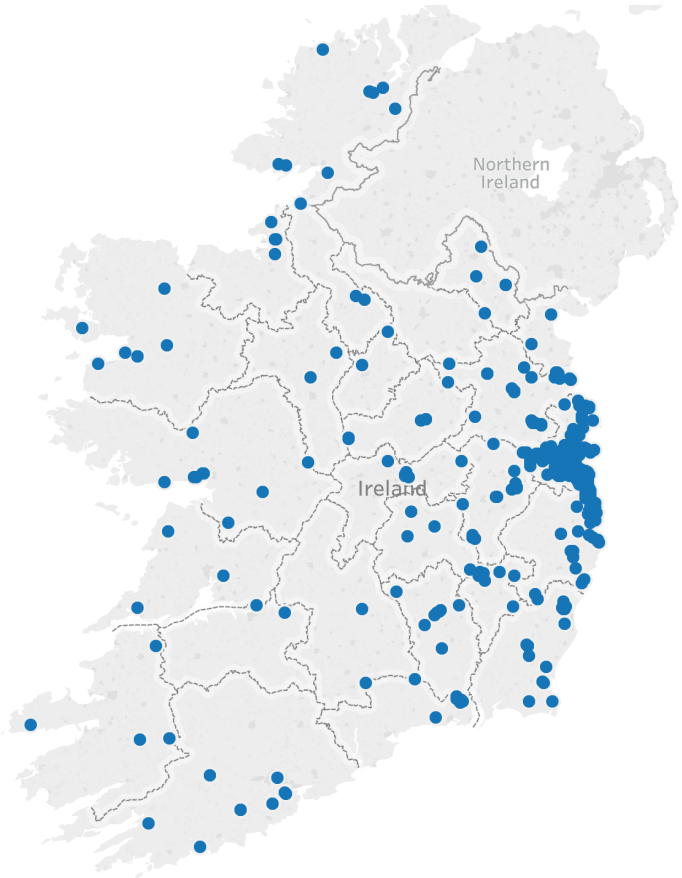


Figure 3.3: 2D plot of students locations entering IADT in 2016

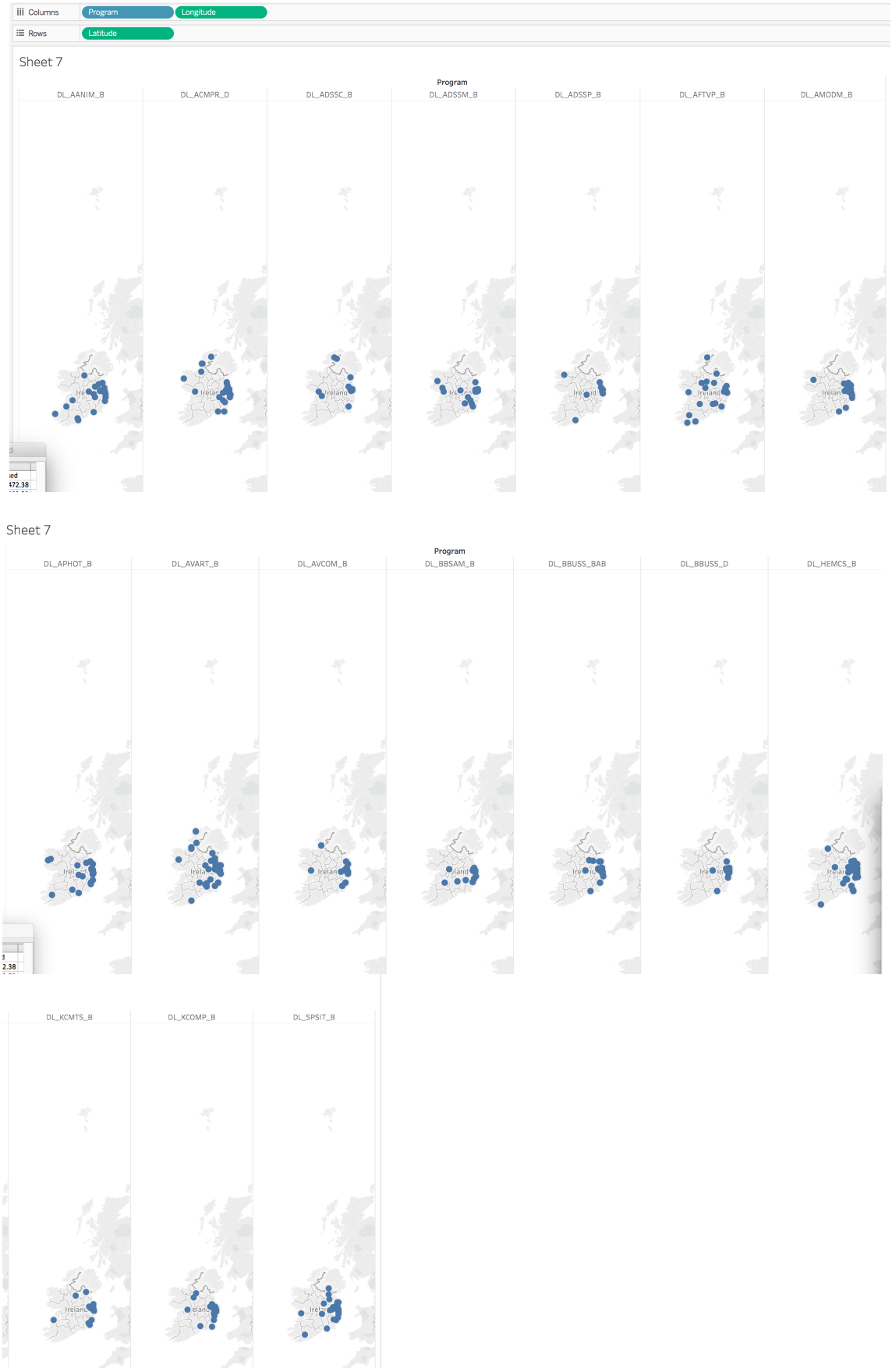


Figure 3.4: 3D trellis plot of student location by course

Using the dataset **Studentloc3D** provided in *Data(2016).xlsx* we can condition on course to create a 3D Trellis plot with each course having its own panel as shown in Figure 3.4. From this plot we can see that some courses attract a wide geographical spread while other courses have a strong Dublin flavour in their intake.

Another example of a 3D trellis plot is visualising the distribution of grants awarded for seven schemes for 16 education training boards as shown in Figure 3.5. The data is provided in the worksheet **Grant3D** in *Data(2016).xlsx*.

+

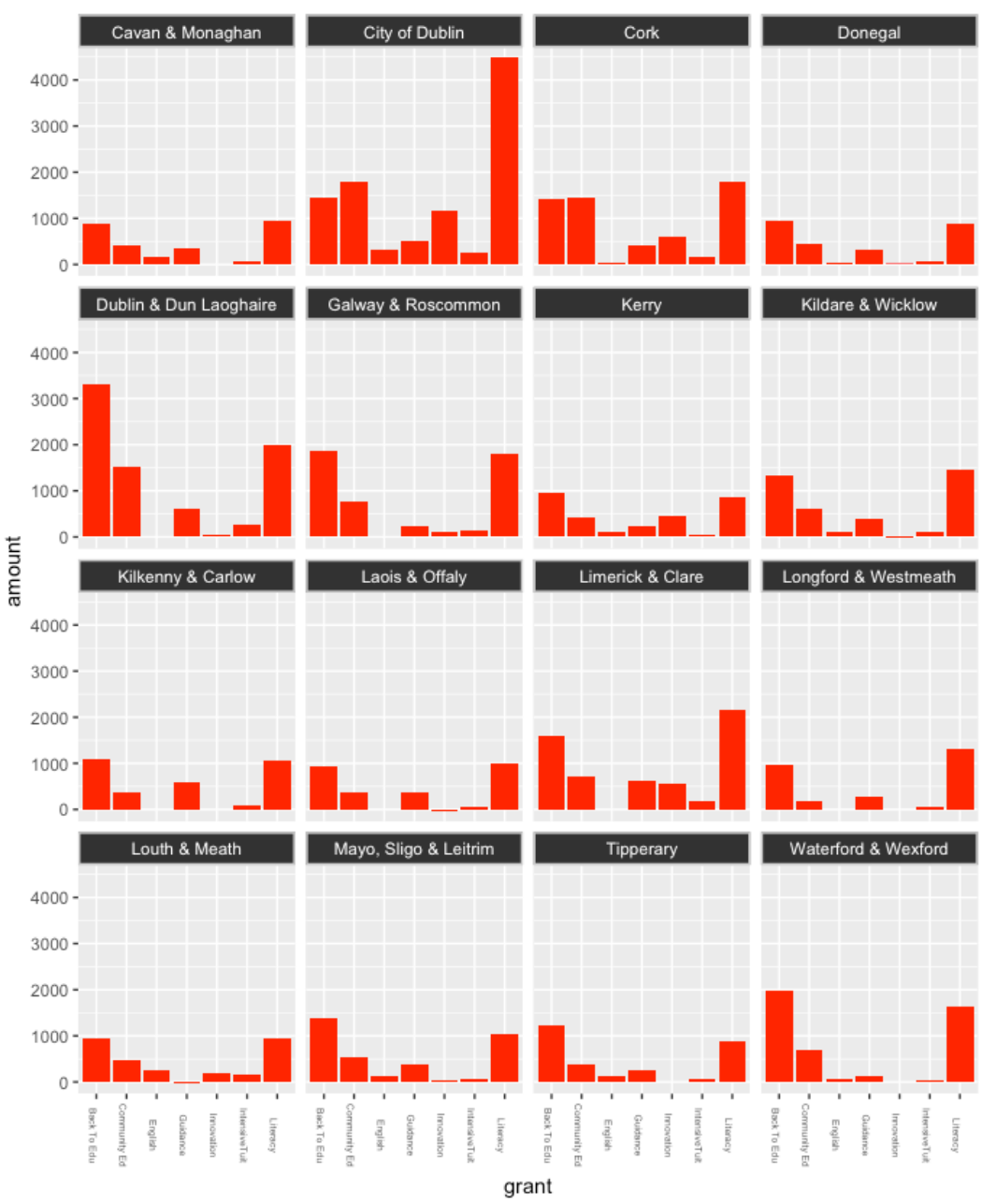


Figure 3.5: 3D trellis plot of grant award by educational training board

Trellis plots can be extended to four or more dimensions. For example, the data file **insurance4D** in *Data(2016).xlsx* contains four variables which are the premium charged to five age cohorts classified by gender and license status. Selecting the worksheet Trellis4D in Tableau we can create a 4D trellis plot with *premium* a continuous variable and *gender*, *license status* and *age cohort* discrete variables as shown in Figure 3.6. Moving the pills between the row and column shelf allows the analyst to easily obtain a number of different views of this data set.

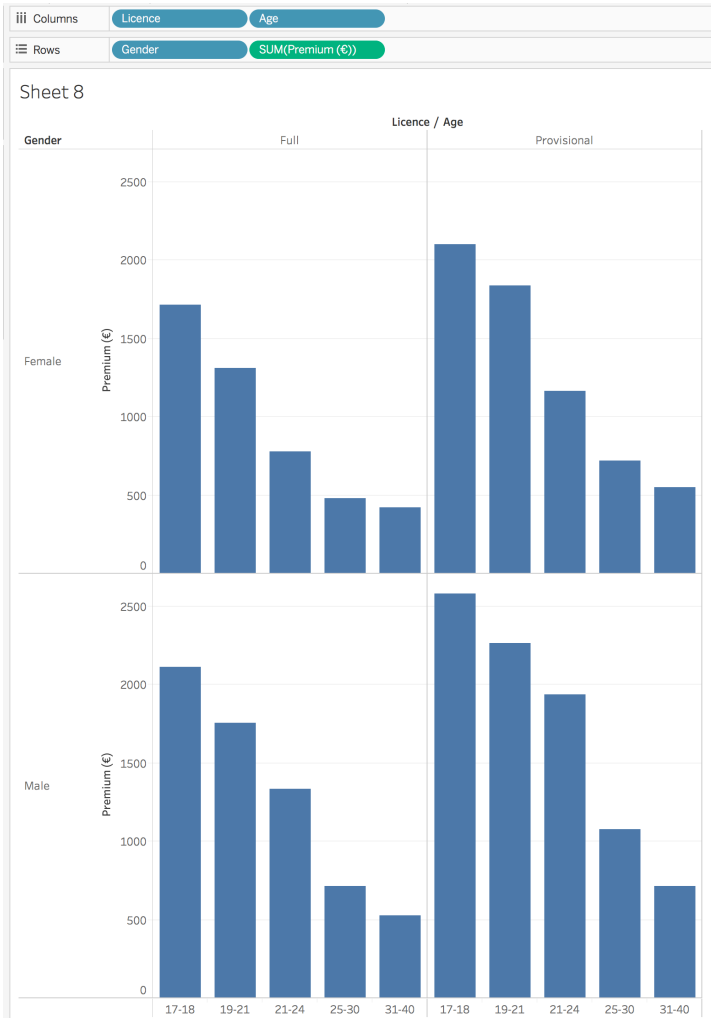
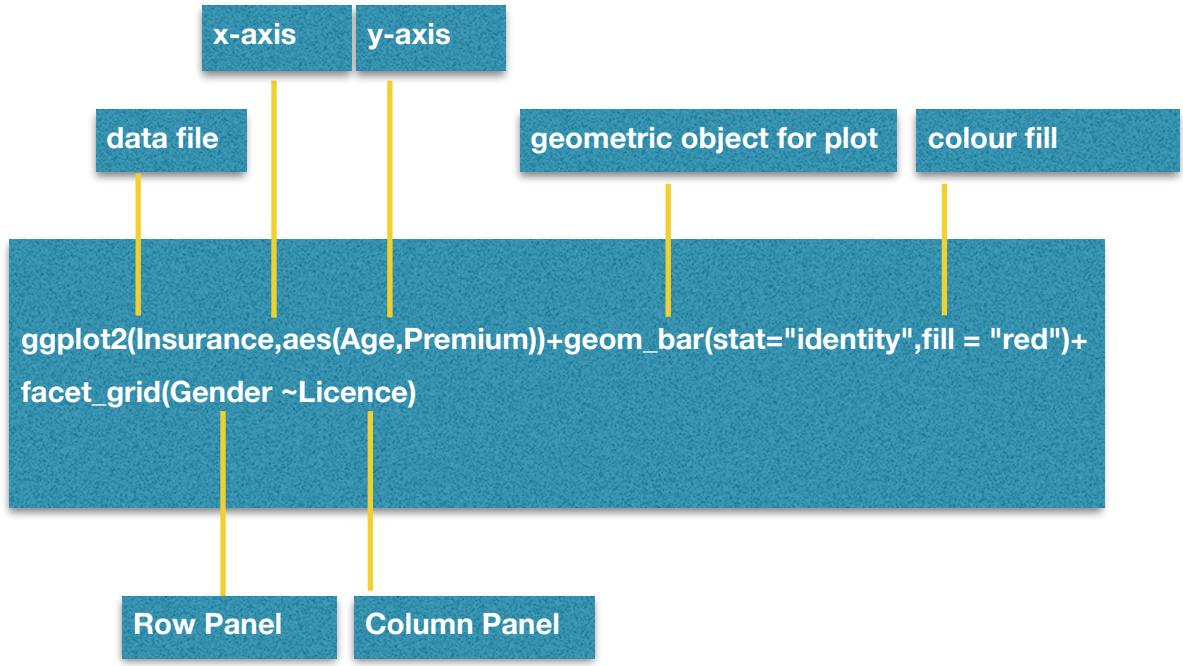


Figure 3.6: 4D trellis plot of premium classified by age, license and gender

The ggplot2 code to generate the Tableau plot in Figure 3.6 is provided below:



Inputting the above code into RStudio the trellis plot in Figure 3.7 is obtained.

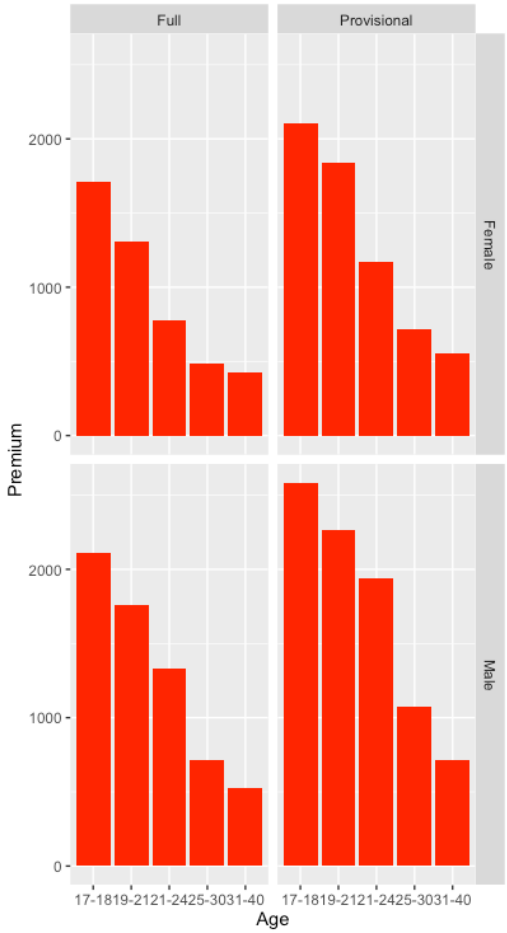


Figure 3.7: 4D trellis plot using ggplot2 of premium classified by age, license and gender

MD Mosaic Plots

The Mosaic plots outlined in Chapter 2 on bivariate data can be extended to higher dimensions but there is a limit to the number of dimensions that can be displayed clearly. The following table sourced from *Private Motor Statistics* published by the Central Bank of Ireland provides the number of Third Party Fire and Theft policies in the Irish market by age, sex and license status and can be found in the excel sheet **InsuranceMosaic** in *Data(2018).* The data set therefore contains three discrete variables and is suitable for a 3D mosaic plot representation.

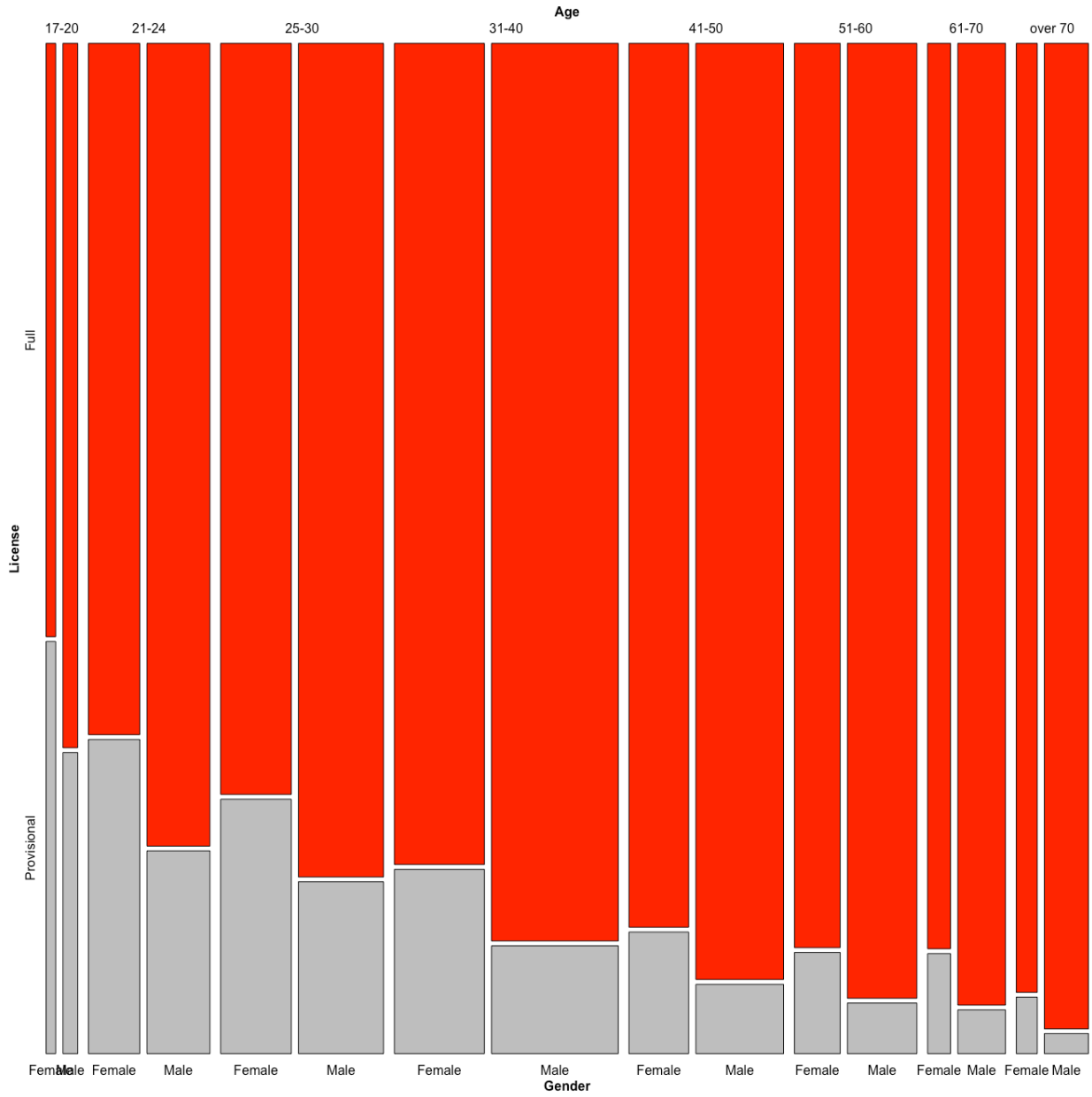
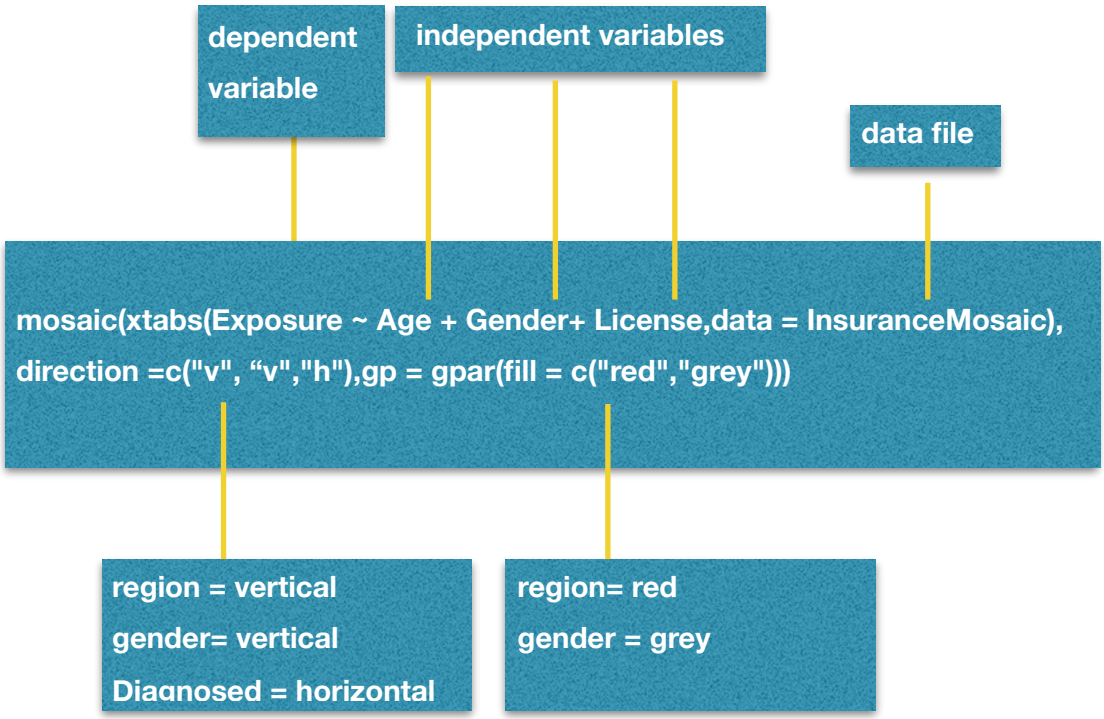


Figure 3.8:Mosaic plot of Third Party Fire &Theft Policies cancer in Ireland classified by age, gender and license.

The R code to create Figure 3.8 using the library vcd is shown below:



There are a number of ways of organising the display and it is easy to change the variables around using R. In Figure 3.8 **Gender** and **Age** are placed on the x-axis while **License** is placed on the y-axis. We can see from this plot that the proportion of males is higher than females for all cohorts. Also the proportion of female provisional licenses is larger than males for all ages while the overall number of policies is largest for the age group 31-40. However placing **Gender** on the y-axis in place of License we can now get an alternative view showing that as age increases from the 18-20 cohort the proportion of males increases while for most age cohorts the male/female breakdown for provisional licenses is broadly the same.

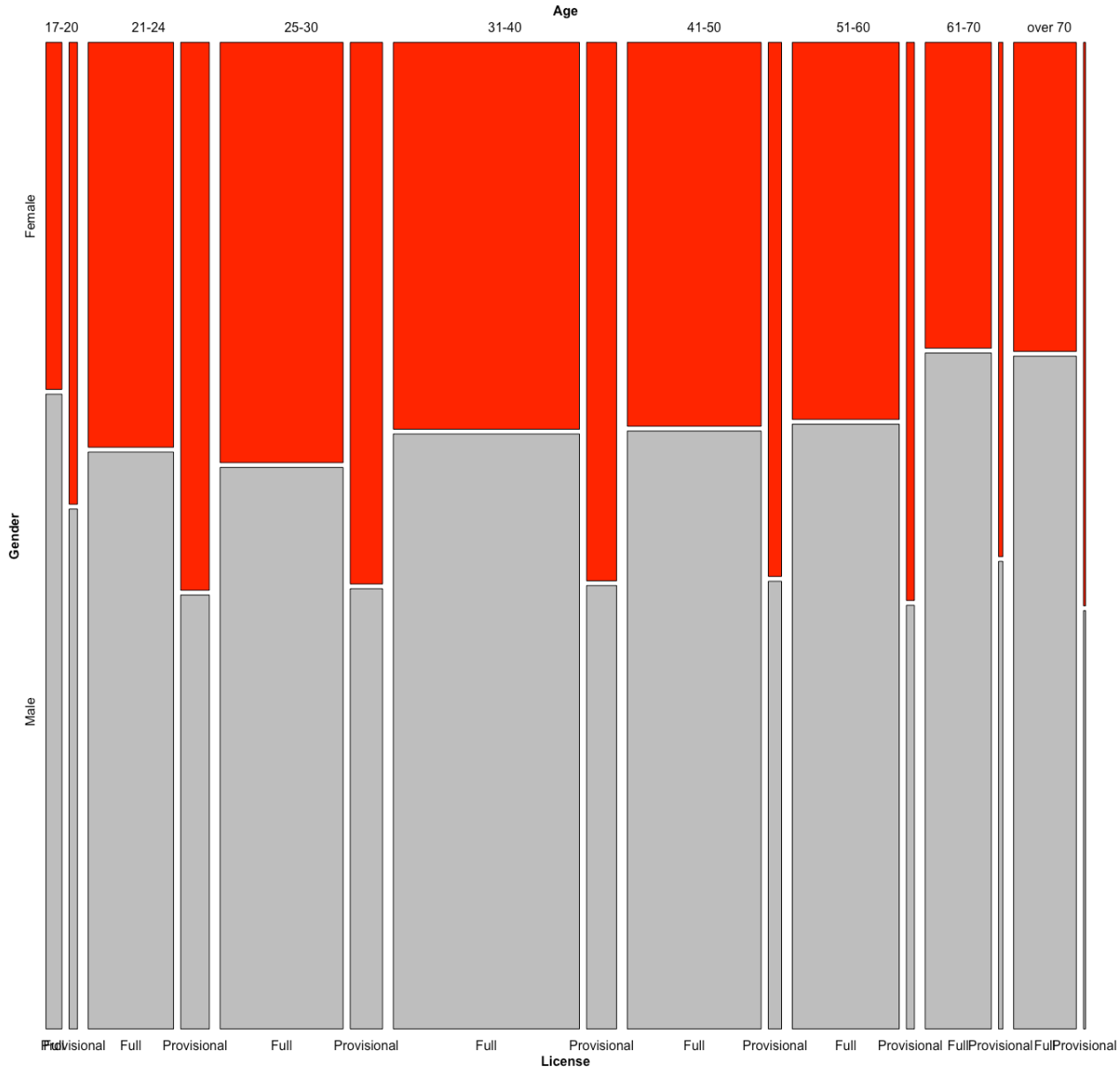


Figure 3.9:Mosaic plot of Third Party Fire &Theft Policies cancer in Ireland classified by age, gender and license with gender on the y-axis.

Finally, the HyperCard plot shown in Figure 4.0 includes a useful scale for both x and y- axis. The plot also contains a side bar containing the proportion of policies in each age cohort. The sidebar can be a useful reference point to allow comparison against the main plot to assess if there is over or underrepresentation of that age cohort in the gender/ license categories. For example, the 17-20 and 21-24 cohorts seem to be overrepresented in provisional licenses with approximately 10 per cent of males compares with just 2 or 3 percent of all policies.

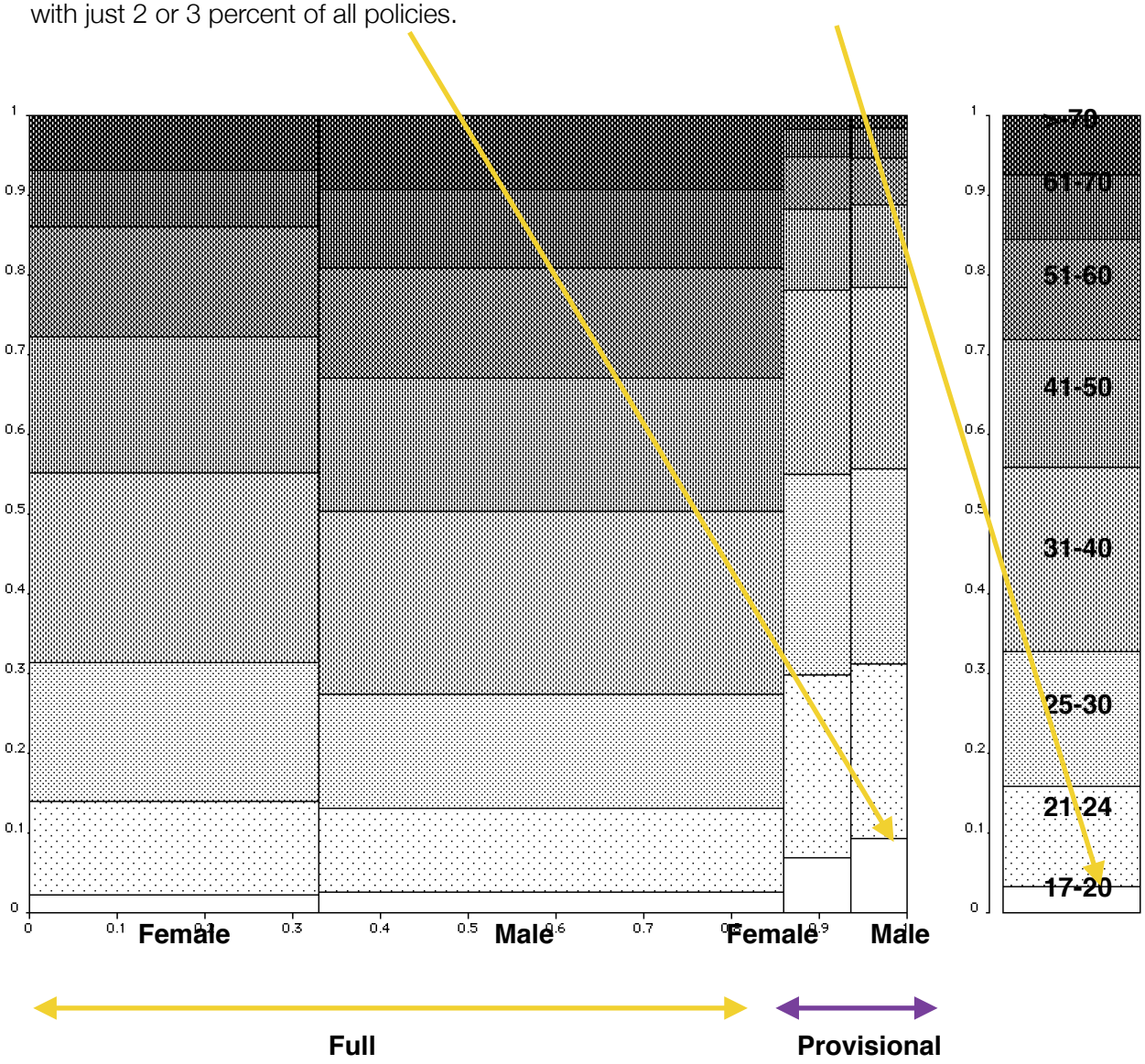


Figure 4.0 :Mosaic plot of Third Party Fire &Theft Policies cancer in Ireland classified by age, gender and license with age on the y-axis. with scale and sidebar

From the last three plots we can see that different orientations provide additional insights which is true to Unwin's analogy with photography mentioned earlier. Of course Mosaics can take some time to explain to the viewer which can have significantly impact on their effectiveness. As the artist **David Hockney** said in response to a question about what it is about his work that goes straight through to understanding and feeling for a large number of people:

I'm interested in ways of looking and trying to think of it in simple ways. If you can communicate that, of course, people will respond. Everybody does look, its just a question of how **hard** *they're willing to look, isn't it* (my emphasis).

The analogy with artists and their work runs through graphics and visualisation. Mosaics are similar in appearance to the work of **Mondrian** and software applications developed by Unwin and his colleagues are named after some of the great artists such as **Mondrian** and **Manet**.

Mosaics were invented in Hartigan and Kleiner (2) but long before then plots similar to Mosaics known as divided bar charts were used by the great French cartographer/ Engineer **Minard** who used them to communicate the transport of goods on French railways in the 1800's as shown in Figure 4.1.

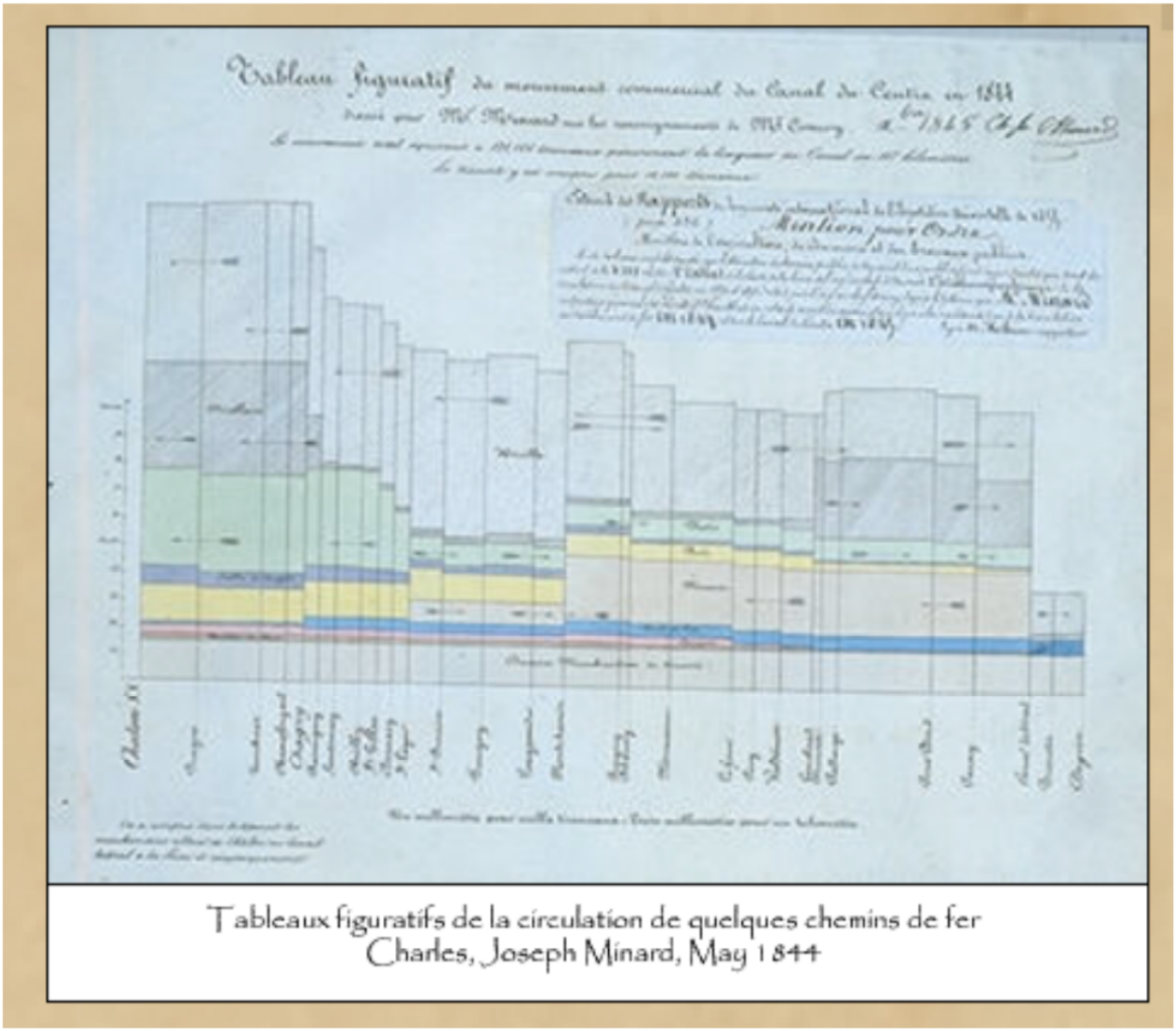


Figure 4.1: Divided bar chart by Minard showing transport of passengers and goods

4. Visualising geospatial data using Tableau

In this section we will map CSO census data using what are known as **shapefiles** which are an open data format used in geographic information systems (GIS) software like Arcview to represent spatial data. Shape files are usually contained in a folder with five files that must be present. These files have extensions .cpg, .dbf, .prj,.shx and .shp. The .**shp** extension is the main shape file and is the file selected in Tableau.

Shape files are developed and regulated by the company ESRI (Environmental Systems Research Institute) based in California. Tableau can now directly read shape files opening up the possibility of visualising geospatial data at the county, electoral division (ED) and what are known as small areas as well as a host of other geospatial boundaries including garda stations, garda districts, garda divisions and church boundaries to name a few. For example, if data is available at the county level effective visualisations of the kind shown in Figure 3.9 can be obtained using the appropriate shape file. The plot is the number of males who report the highest level of education at Technical/Vocational level in the 2016 census. Counties reporting high rates e.g. Cork and parts of Dublin show deeper colours and vice versa.

In this section we will use detailed data from the latest CSO data from the 2016 census released in July 2017. We will restrict the examples in this section to the education theme which is categorised by CSO as Theme 10 and map data at the county, electoral division and small area level.

4.1 Mapping County Data using Shape files

We will start by recreating the plot in Figure 4.1 using the following procedure:

- Open Tableau and select **connect to Spatial file**.
- Select the folder **Admin_Counties_Generalised_20m_OSi_National_Boundaries** This is the county shape file containing five separate files.
- Select the file **Admin_Counties_Generalised_20m_OSi_National_Boundaries.shp**

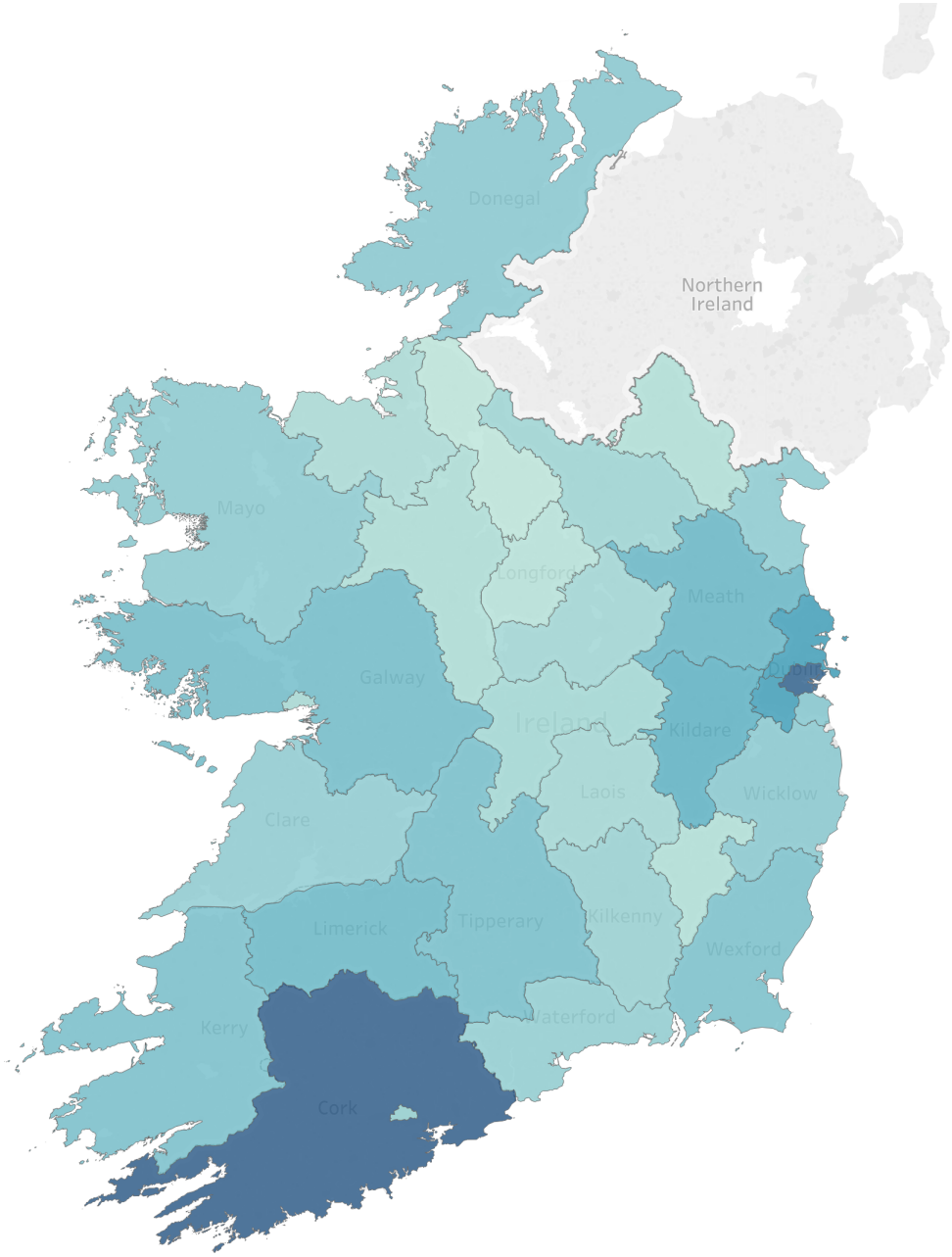


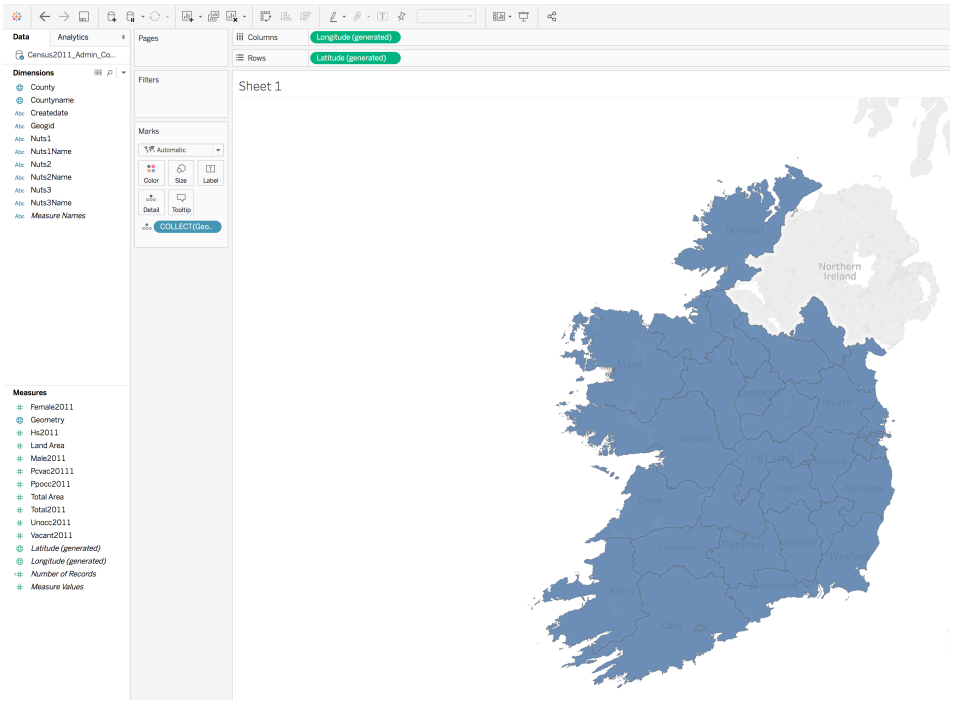
Figure 4.1: Highest level of education reported as Technical/Vocational by county

The screen below now appears showing the raw data with the variables in the file at the top of each column in bold.

Sort fields | Data source order

Contae	Gaeilge	Province	ID	Centroid X	Centroid Y	Global ID	Guid	Shape__Are	Shape__Len	Geometry
Baile Átha Cliath	null	Leinster	0	712153	735780	ec7df314-e9ad-4b31...	EC7DF314-E9AD-4B3...	0.017327	1.2584	MULTIPOLYGON
Gaillimh	null	Connacht	1	530363	726364	fa241879-9706-4005...	FA241879-9706-400...	0.006836	0.7459	MULTIPOLYGON
Corcaigh	null	Munster	2	567611	571949	eb9428b8-b443-407...	EB9428B8-B443-407...	0.005171	0.4665	POLYGON
Laois	null	Leinster	3	644825	693938	76a974d4-60f2-4167...	76A974D4-60F2-416...	0.230083	3.2426	POLYGON
Ros Comáin	null	Connacht	4	582317	778883	2d3544e6-a828-4bf7...	2D3544E6-A828-4BF...	0.346826	5.1602	POLYGON
Port Láirge	null	Munster	5	625792	604797	8256b4ab-2ebb-46e...	8256B4AB-2EBB-46E...	0.244034	5.1616	MULTIPOLYGON
Baile Átha Cliath	null	Leinster	6	704391	726785	6c7d5f02-1d93-42bf-...	6C7D5F02-1D93-42B...	0.030097	1.1223	POLYGON
Lú	null	Leinster	7	699245	798377	0f2662ee-a38d-4001...	0F2662EE-A38D-400...	0.112963	3.0435	MULTIPOLYGON
Muineachán	null	Ulster	8	669141	824829	78ae611d-f9da-4ec9...	78AE611D-F9DA-4EC...	0.178008	3.6176	POLYGON
An Mhí	null	Leinster	9	688919	762584	7d3d6ef0-2287-488a...	7D3D6EF0-2287-488...	0.318150	4.7763	POLYGON
Baile Átha Cliath	null	Leinster	10	715859	750573	da8a143e-499d-440...	DA8A143E-499D-440...	0.062004	2.8677	MULTIPOLYGON
Liatroim	null	Connacht	11	599868	819517	76cf8f82-cd46-41f7-...	76CF8F82-CD46-41F...	0.218370	3.7898	MULTIPOLYGON
An Iarmhí	null	Leinster	12	635445	753975	c26ac177-3d04-4f56...	C26AC177-3D04-4F5...	0.249160	3.5139	POLYGON

Now select sheet1 from the bottom tab. Double click on the file **Geometry** in the measures panel. The map below now appears.

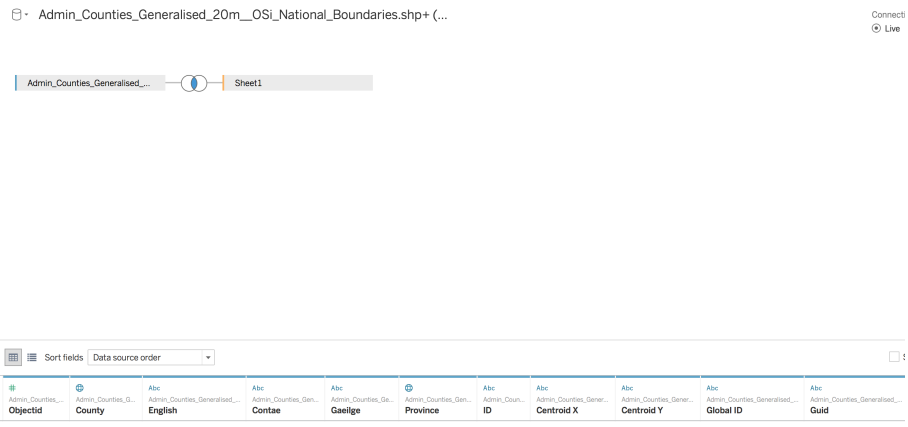


Joining shape files to CSO census files

We now need to link this visualisation with the CSO census data file. In this example we will use an extract of Census data 2016 that contains variables relating to the education theme (Theme 10). The data extract is contained in the file *County2016.xls*. located in the **Theme 10 CSO Census Extract** in the **Mapping Files 2018** folder.

Click on the **Data Source** panel tab on the bottom of the Tableau screen and then select **Add** from the top left of the resulting window. Select **Excel file** from **Add a Connection** panel and then *County2016.xls*

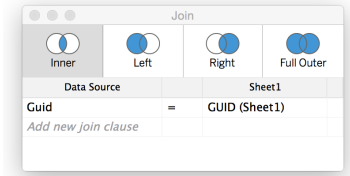
The following screen should now appear:

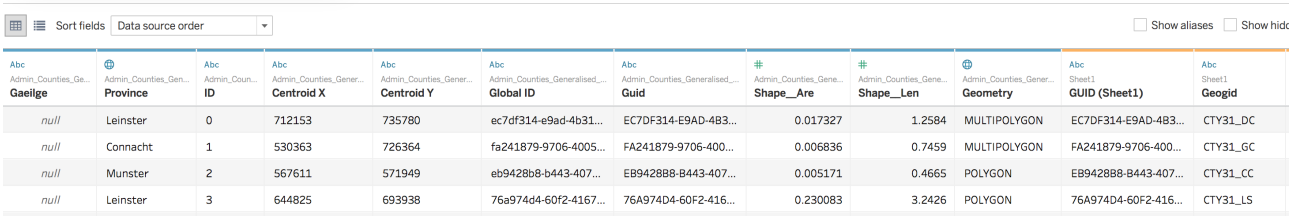


To link the two data sets we must use a variable that is common to both files. **Guid** is one such variable that is provided in both the shape and *County2016.xls* files. To link both data sources apply the following procedure:

Select **Inner** and then for Data Source select **Guid**.
Select **Sheet 1** (which is the *County2016.xls* file) and the name **Guid1** (Tableau changes the variable name automatically from Guid to Guid1).

Note: Tableau can sometimes identify and automatically select common variables from both files and join the files. However, the analyst can override the selections by clicking on variable name and selecting a different variable.





Now click the **sheet 1** tab in tableau. The dimensions and measures pane now have the the *County2016.xls* variables in a folder called sheet 1 in both the dimensions and measures section.

Drag the variable **T10 4 Tvm** (which is the file that records the number of males with vocational/technical education) onto the **Colour** icon in the **Marks** section to obtain the map of numbers of males with highest level of education at vocational/technical encoded by colour as shown in Figure 4.2.

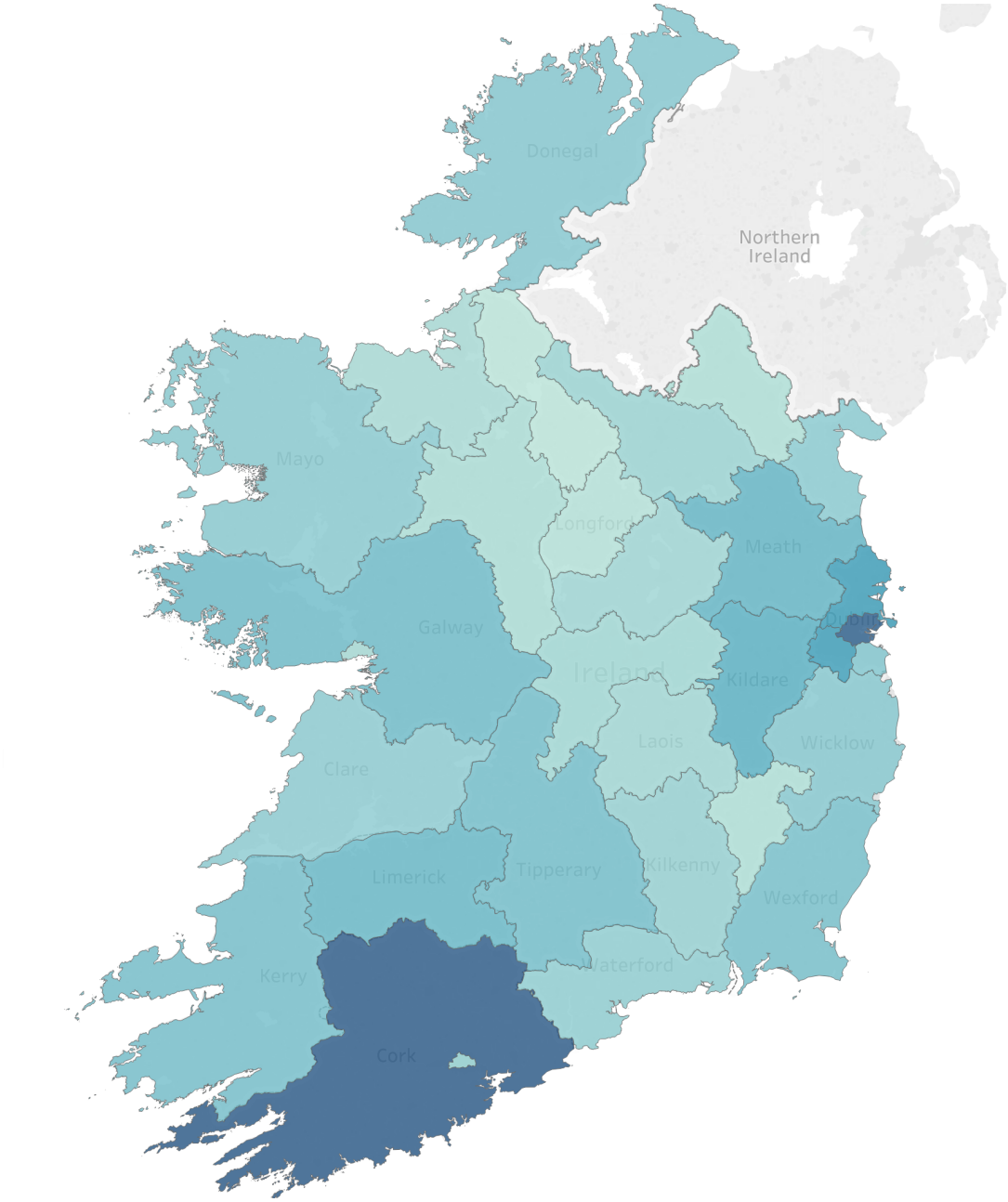
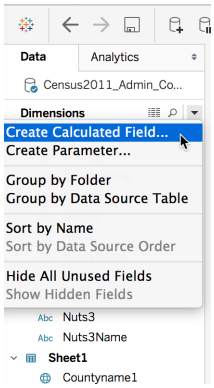


Figure 4.2: Highest level of education reported as Technical/Vocational by county

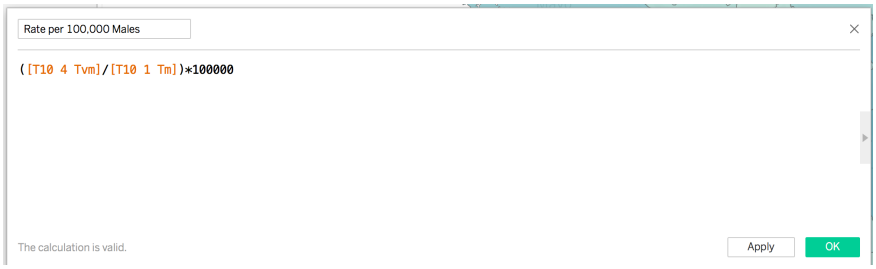
Adjusting data using calculated fields

The map in Figure 3.4 shows, as might be expected, deeper hues in Dublin City and Cork County, as these have the highest population densities. To adjust the data by the male population of each county we can divide **T10 4 Tvm** by the male population of each

County which is contained in the variable **T10_1_Tm** to get a rate per county population by creating a **calculated field**. Calculated fields can be computed selecting the icon beside the dimensions panel and selecting **Create Calculated Field** as shown below:



A blank dialog box appears. To calculate the rate, say per 100,000 county population, divide the two variables and multiply by 100,000 to obtain a technical/vocational rate per 100,000 population.



Finally, **enter** the name of this new calculated variable as say Rate per 100,000 Males at the top of the dialog box and select **apply** from the bottom of the tab.
This new variable now appears in the **Measures** section.

Finally, drag **Rate per 100,000** onto the Color icon in the Marks section to obtain the population adjusted map as shown in Figure 4.3. Note that the population adjusted data reveals Dublin (and Dunlaoghaire Rathdown) as having one of the lowest rates while the border region reports high rates per 100,000 male county population.

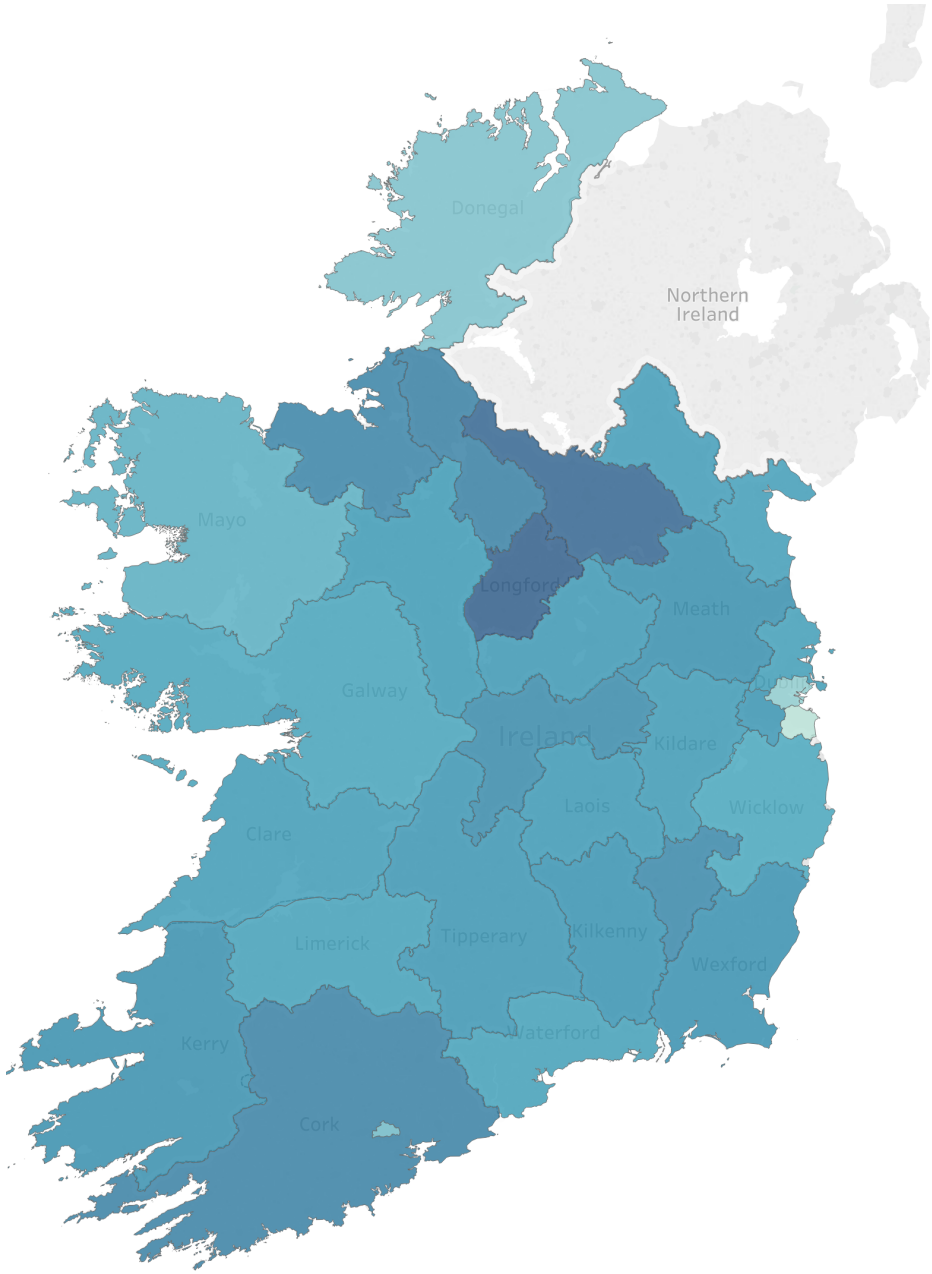


Figure 4.3: Male rate of technical/vocational education per male county population, 2016

4.2 Mapping Electoral Division Data using Shape files

Electoral Divisions (EDs) are the smallest legally defined administrative areas in the State. There are 3,409 ED's in the shape file (a small number of small EDs have been merged for confidentiality reasons). It is possible to visualise data using electoral division (ED)

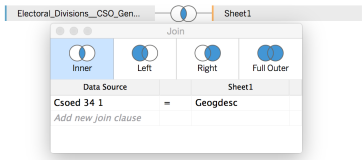
boundaries using the same approach as outlined in the previous section on mapping county data. To map data at the ED level we use the shape file:

Electoral Divisions__CSO_Generalised_20M.shp

The Excel file recording the variables collected under Theme 10 (education) for each electoral division is provided in the file:

ED2016.xlsx.

located in the **Mapping Files 2016** folder. To join the two files we need to identify a variable that is common to both files which is **cso 34 1** from the shape file and **Geogdesc** from the CSO file. The result of joining the two files is shown below



Sort fields Data source order — Show aliases — Show hidden fields

Centroid Y	Guid	Csoed 3409	Osied 3441	Csoed 34 1	Shape__Are	Shape__Len	Geometry	Guid1	Geogid	Geogdesc	T10 1 Lt15M
616289.230043841	2AE196291BD413A3...	14095	247006	Aughwilliam	0.0020651	0.27565	POLYGON	2AE196291BD413A3...	ED3409_14095	Aughwilliam	8
727734.553337762	2AE196291CBE13A3...	05009	267020	Blackrock-Carysfort	0.0001761	0.05950	POLYGON	2AE196291CBE13A3...	ED3409_05009	Blackrock-Carysfort	26
637693.505219255	2AE1962923CC13A3...	07006	97028	Coolaghmore	0.0034492	0.31131	POLYGON	2AE1962923CC13A3...	ED3409_07006	Coolaghmore	16
929172.005661047	2AE1962920971 3A3...	33125	57129	Rathmullan	0.0025359	0.28671	MULTIPOLYGON	2AE1962920971 3A3...	ED3409_33125	Rathmullan	26
635898.045774622	2AE1962923921 3A3...	19093	77005	Ardagh	0.0020464	0.29314	POLYGON	2AE1962918191 3A3...	ED3409_11019	Ardagh	41
821863.816083927	2AE196291A6313A3...	34034	177001	Aghabog	0.0029907	0.30411	POLYGON	2AE196291A6313A3...	ED3409_34034	Aghabog	19
662018.848222025	2AE196291E7D13A3...	14058	247077	Kilnahue	0.0021114	0.28489	POLYGON	2AE196291E7D13A3...	ED3409_14058	Kilnahue	12
713163.275751818	2AE1962922B213A3...	06065	87033	Droichead Nua (New...	0.0025835	0.30306	POLYGON	2AE1962922B213A3...	ED3409_06065	Droichead Nua (New...	31
805352.722110978	2AE1962918BE13A3...	29032	157141	Sraheen	0.0035552	0.33138	POLYGON	2AE1962918BE13A3...	ED3409_29032	Sraheen	48
786561.108231558	2AE196291D8F13A3...	32030	27022	Castlerahan	0.0039112	0.41316	POLYGON	2AE196291D8F13A3...	ED3409_32030	Castlerahan	25
667565.937848404	2AE1962922001 3A3...	07090	97013	Baunmore	0.0023174	0.28733	POLYGON	2AE1962922001 3A3...	ED3409_07090	Baunmore	3
666953.160567704	2AE196291E9513A3...	14059	247084	Limerick	0.0028073	0.29648	POLYGON	2AE196291E9513A3...	ED3409_14059	Limerick	16
612860.747786429	2AE196291BA213A3...	24016	228016	Gracedieu	0.0009446	0.13086	POLYGON	2AE196291BA213A3...	ED3409_24016	Gracedieu	29
733252.957227345	2AE1962923191 3A3...	27062	67202	Sliabh an Aonaigh	0.0078171	0.67979	POLYGON	2AE1962923191 3A3...	ED3409_27062	Sliabh an Aonaigh	13
633713.921970016	2AE196291E6413A3...	14014	247027	Bree	0.0037958	0.33349	POLYGON	2AE196291E6413A3...	ED3409_14014	Bree	77
762918.640948974	2AE1962918601 3A3...	11076	167020	Clonmacduff	0.0018681	0.27687	POLYGON	2AE1962918601 3A3...	ED3409_11076	Clonmacduff	11
806259.755782044	2AE196291DA913A3...	32048	27024	Cavan Rural	0.0049886	0.50534	POLYGON	2AE196291DA913A3...	ED3409_32048	Cavan Rural	109
590892.885068379	2AE196291F6613A3...	18223	47037	Ballynamona	0.0053755	0.52752	POLYGON	2AE196291F6613A3...	ED3409_18223	Ballynamona	27
731650.358365839	2AE196291CFE13A3...	03006	267049	Clondalkin-Dunawley	0.0006404	0.14541	POLYGON	2AE196291CFE13A3...	ED3409_03006	Clondalkin-Dunawley	214
657921.911160026	2AE1962918741 3A3...	01025	17019	Garryhill	0.0031739	0.32762	POLYGON	2AE1962918741 3A3...	ED3409_01025	Garryhill	22
781029.973479602	2AE1962919041 3A3...	09027	137035	Granard Urban	0.0011069	0.16606	POLYGON	2AE1962919041 3A3...	ED3409_09027	Granard Urban	27

Following the same procedure as before by dragging the variable **T10 4 Tvm** onto the colour marker to obtain the map by electoral division as shown in Figure 4.4. As before in this illustration deeper hues represent higher number of males with highest education level designated as Technical/Vocational. We could also adjust this data by population of ED to obtain a population adjusted rate by creating a **calculated field** and entering the

formula (T10 4 Tvm/T10_1_TM)*100,000. This formula is the number of males in each ED with technical/vocational level of education divided by the total number of males in each ED multiplied by 100,000. We will leave the task of generating this visualisation to the reader!

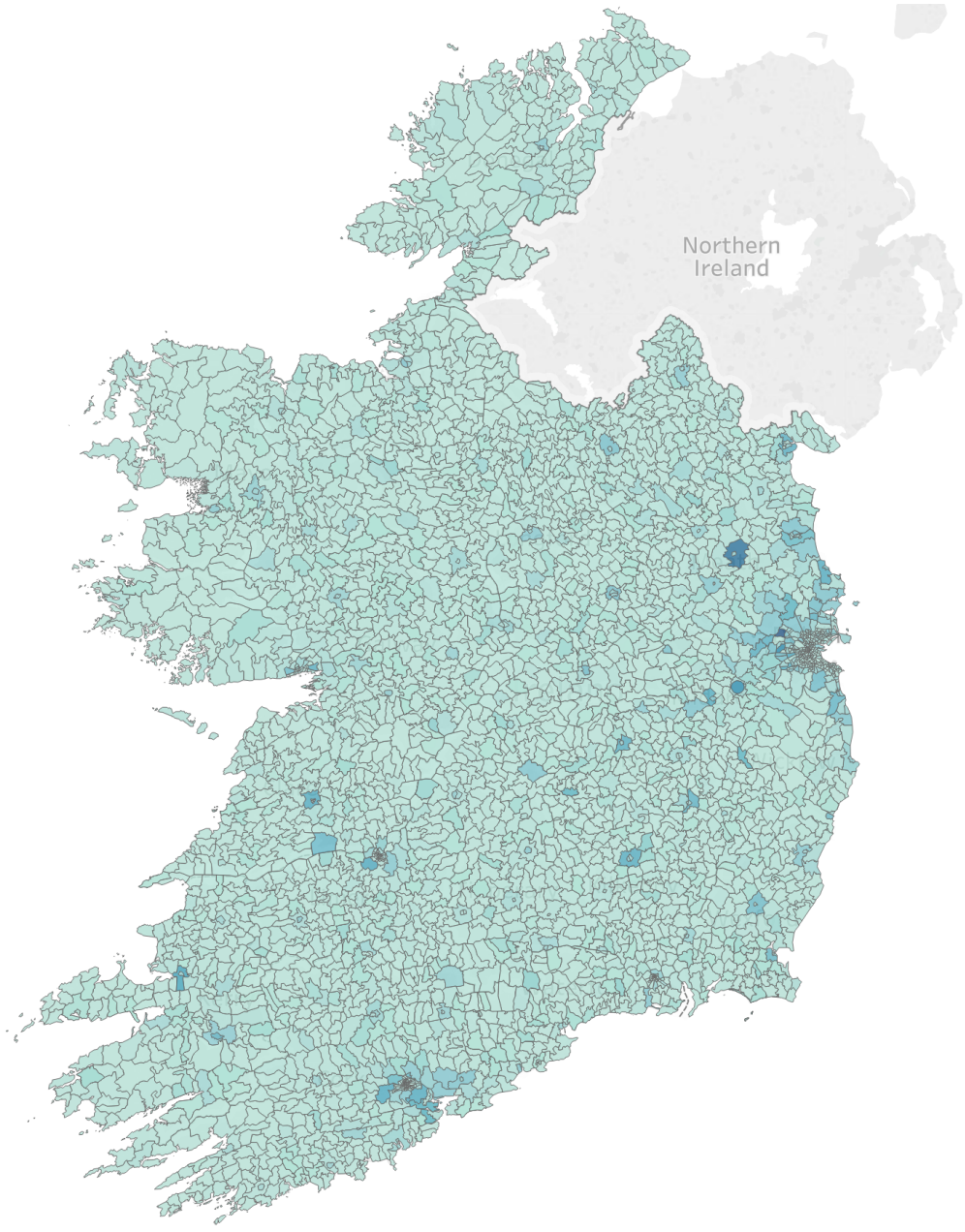


Figure 4.4: Number of males with technical/vocational education by electoral division, 2016

4.3 Mapping Small Area Data using Shape files

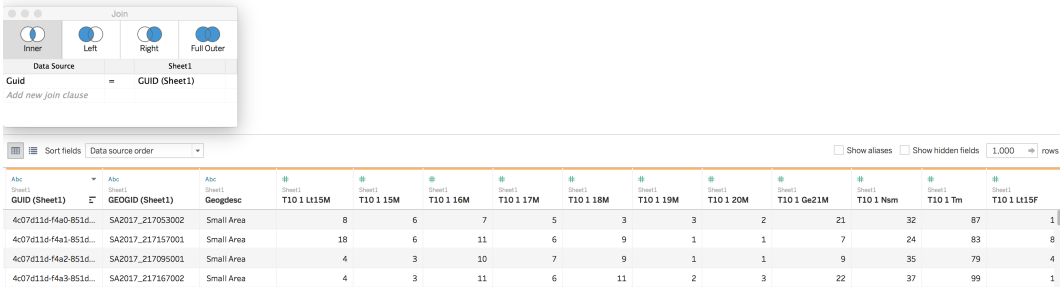
We can data to an even higher geographic resolution using Small Area boundary data. Small Area boundaries were created by the **National Centre of Geocomputation at NUI Maynooth** for Ordnance Survey Ireland and are considerably smaller and more homogeneous (or similar) than Electoral Divisions. There are approximately 18,488 Small Area units in comparison to 3,409 Electoral Divisions. A Small Area boundary is usually composed of approximately 80-120 households and have an average size of 3.5km^2. In comparison, an Electoral Division can has an average size of over 20km^2.

To generate a Small Area map we use the same approach is adopted as in the previous sections as follows:

Open the shape file **Small Areas__Generalised_20m_OSi_National Boundaries.shp**.

Add the Excel file recording the number of males with highest level of education vocational/technical for each small area is provided in the file:
SA2016.xlsx which is located in the Mapping Files 2016 folder.

To link the two files we can use variable name **Guid** which is an id for each small area and id contained in both files. Tableau may automatically select a different variable for joining but this can be overrided by the user. The result of joining the two files is shown in the illustration below:



Select the sheet 1 tab, double click on **geometry** and drag the data file **T10 4 Tvm** onto the colour marks to produce the map shown in Figure 4.5. As before deeper colour hues represent small areas with high number of males with highest education level designated as Technical/Vocational.

We could also adjust this data by male SA population as we did for the county data. Again we will leave the task of generating this visualisation to the reader!

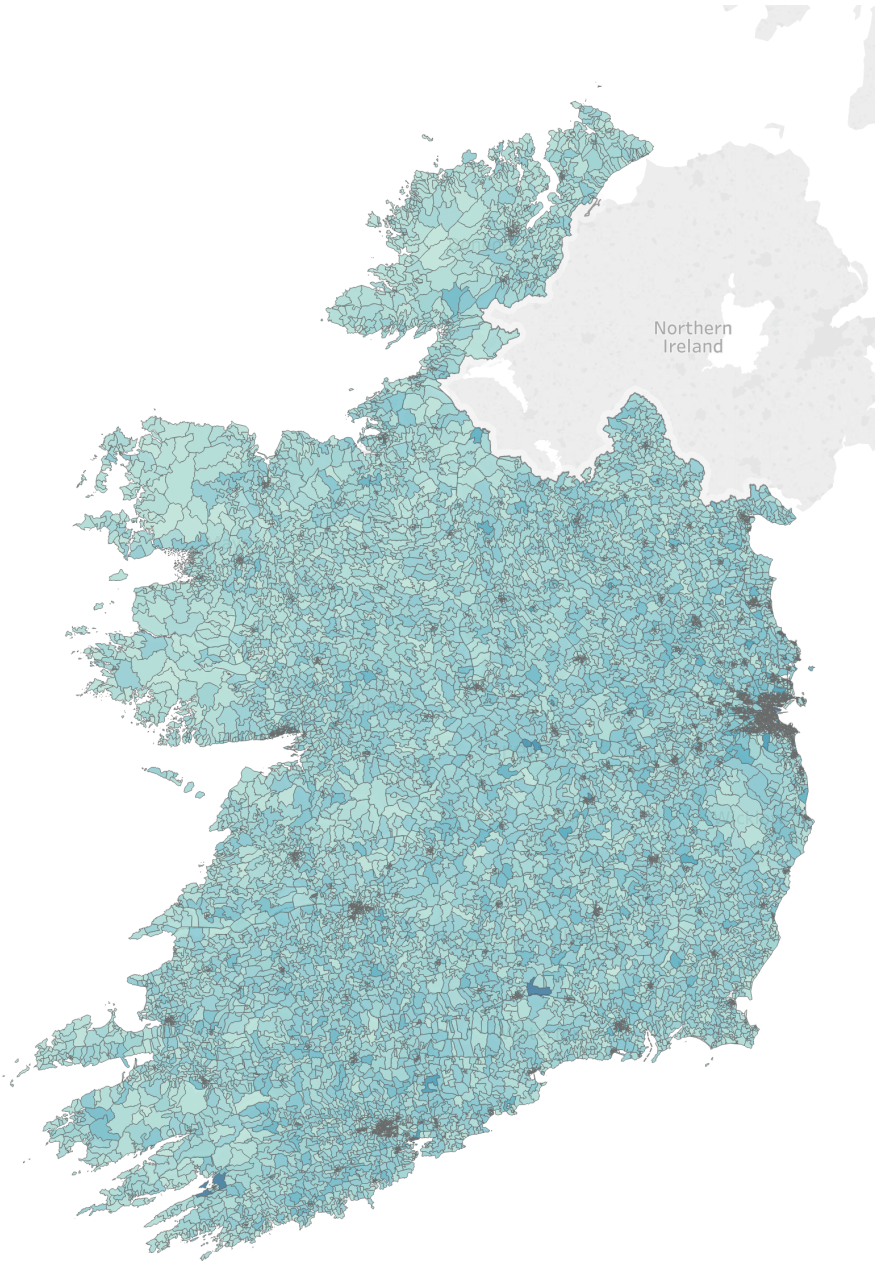


Figure 3.13: Number of males with technical/vocational education by small area, 2016

Using Tableau's zoom capability we can investigate the distribution of males with highest education level defined as technical/vocational in city areas as shown in Figure 4.6.

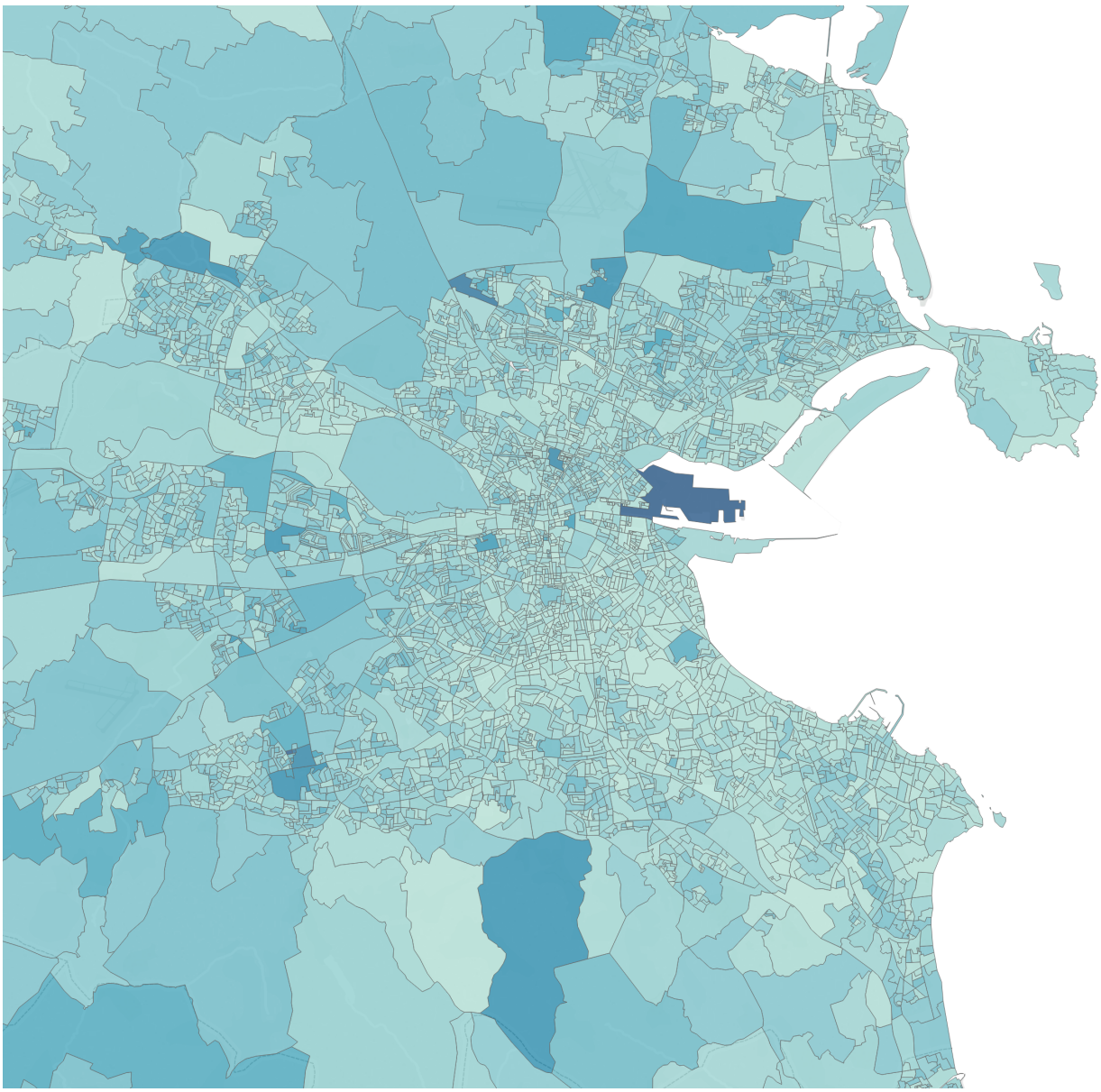


Figure 4.6: Number of males with technical/vocational education in greater Dublin area, 2016

Exercises

1. The data in the Excel worksheet **cancer3D** in *ExerciseData(2018).xlsx* records **cancer type**, **age cohor**t and **sex** for over 50,000 patients diagnosed between 2009—2013.

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.
ii) Create a Trellis chart for this data set using all available variables.
iv) Using i) summarise the principle features of this data set

2. The data in the **crime3D** in *ExerciseData(2018).xlsx* records the number of **Attempts or threats to murder, assaults, harassments and related offences** in 2015 by garda station.

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.
ii) Create an effective visualisation of this data set
v) Using i) summarise the principle features of this data set

3. Theme 8 in the CSO Census 2016 file (*CensusCounty2016.xls*) examines the principal status of census respondents classified by age and gender by county/city. One of the variables records male respondents who are looking for their first job. The variable is called **T8_1_LFFJM** in the excel file *CensusCounty2016.xls*.

i) Visualise the geographical distribution of this variable by **county** joining the shape file called *Admin_Counties_Generalised_20m__OSi_National_Boundaries* with the CSO Census file *CensusCounty2016.xls*.

ii) Generate a calculated field to visualise the distribution of males looking for their first job per 10,000 county population.

iii) Using results of i) and ii) write a short note on the geographic distribution this variable

4. Theme 10 in the CSO Census 2016 file (*CensusED2016.xls*) is based around Education by **electoral division**. One of the variables lists citizens aged 15 years and over who have highest level of education recorded as *advanced certificate/completed apprentice*. The variable is called **T10_4_ACCAT** and is located in the excel file *CensusED2016.xls* in the mapping folder.

i) Visualise the geographical distribution of this variable by **Electoral Division (ED)** by **joining** the shape file *Electoral_Divisions__CSO_Generalised_20M* with the CSO Census file *CensusED2016.xls*.

ii) Using a **calculated field** visualise the distribution of this variable per 1,000 ED population.

iii) Using results of i) and ii) write a short note on the geographic distribution this variable

5. Theme 13 in the CSO Census 2016 examines the occupations of citizens by **small area**. Using the variable **T13_1_STOT** records numbers defined as *Skilled Trade Occupations* and is located in Census*SA2016.xls* in the mapping folder.

i) Create a **small area** map that shows clearly the distribution of this variable in Dublin City.

ii) Using a **calculated field** visualise the distribution of this variable per 100 SA population.

iii) Write a short note on the geographic distribution of this variable in Dublin City highlighting any interesting patterns.

6. The data in the worksheet **ecoli4D** in *ExerciseData(2018).xlsx* records location, gender and diagnosis of three diseases *Cryptosporidiosis*, *Verotoxigenic Escherichia coli infection* and *Giardiasis* during 2015 and 2016 in the greater Dublin area. The location data in the form of longitude and latitude has been jittered by adding noise in the form of a standard normal probability distribution.

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 chart for this data containing a map in each panel and the remaining gender and disease variables placed in rows and columns.

iv) Summarise the principal features of the plot computed in iii).

7. The data in the worksheet **nursingHome3D** in *ExercisesData(2018).xls* contains details on the **Age**, **Length of Service** and **Occupation** for 500 nursing home employees.

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.

ii) Create a mosaic chart for this data that investigates **Length of Service** conditioned on **Age** and **Occupation**

8. The distribution of the age of 853 patients diagnosed with Influenza during 2015 and 2016 is provided in the Excel worksheet **influenza4D** in *ExercisesData(2018).xls*

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 2D box plots and 2D violin plots of:

a) age by patient type
b) age by disease type
c) age by gender

iv) Compute a trellis plot with each panel having a box plot. Put patient type in rows and sex in columns.

v) Using iii) and iv) summarise the principle features of this data set.

9. The data in the worksheet **ShareSegmentation** in *ExerciseData(2018).xlsx* records the percentage share of motor insurance cover by age cohort, license status and company for Third Party Fire & Theft cover in the Irish market.

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 chart for this data with each panel showing the percentage of cover on the y-axis and the age of the policyholder on the x-axis. The panel rows will be company and the panel columns license status.

iv) Summarise the principal features of the plot computed in iii).

9. The data in the worksheet **PremiumTime** in *ExerciseData(2018).xlsx* records the average premium charged to policyholders of motor insurance cover by age cohort, license status, and sex between 2011 and 2015 for Third Party Fire & Theft cover in the Irish market.

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 chart for this data set with each panel showing the premium charged to the policyholder on the y-axis and the age of the policyholder on the x-axis. The panel rows will be gender and the panel columns license status.

iv) Summarise the principal features of the plot computed in iii).

Appendix 1: Data Visualisation using ggplot2

The R package ggplot2 was developed by Hadley Wickham and deploys a novel approach to statistical plots based on the textbook **The Grammar of Graphics** by Leland Wickinson. ggplot2 assumes that a statistical graphic can be created from the same few components. These components are known as graphical layers. The following table outlines the main layers that can be used to create plots in ggplot2:

Layer	Description	Example Code
Data	Maps data to aesthetic attributes which define how the data should be perceived i.e. which variable is placed on the x-axis and which is placed on the y-axis.	*ggplot2(grantsolaslong,aes(grant,Amount))*
Geometric object	Defines what mark is displayed on the plot e.g. bar, point, box plot.	*geom_bar(stat="identity",fill = "red")*
Statistical transformations	creates a smooth line or a predicted regression line though the data.	*geom_smooth(method = 'lm')*
Co-ordinates	The co-ordinate system used for the plot e.g. cartesian, polar or log.	*scale_Amount_log10()*
Faceting	Creates subplots or conditional plots based on one or more discrete grouping variables.	*facet_wrap(~ETB,ncol = 4,scales = "fixed")*
Theme	Allows control over the plot appearance e.g. lines, axis titles, background etc	*Theme(axis.text.x=element_text(angle = -90,size =5))*

Table A.1: Worked example using ggplot2 code

Example

The following example showing the distribution of grants to 16 educational training boards shown on page 52 should make the basic operation behind ggplot2 clearer.
The code to create this plot involves five layers with each layer separated by +. The first layer is:

ggplot2(grantsolaslong,aes(grant,Amount))+

This layer sets up the data file which is called grantsolaslong and specifies the x and y-axis as *grant* and *Amount* respectively. Running this code generates the top left in Figure A.1 overleaf.

The second layer adds the geometry of the plot (called geom) as follows

geom_bar(stat="identity",fill = “red")

In this case the geometry is a bar chart . The text *stat = identity* means just use the raw counts while *fill = red* means fill the bars with the colour red. The resultant bar chart is shown top right in Figure A.1 and illustrates the amount awarded for each grant for all educational and training boards (etb).

These first two layers are the minimum required for a plot to be drawn.

The third layer is the facet (or trellis or small multiple) layer which allows a conditional plot of grant amounts for each etb to be created and is shown bottom left. This layer is coded as:

facet_wrap(~ETB,ncol = 4,scales = "fixed")+

This code *~ETB* allocates the variable ETB as the splitting or group variable. *ncol = 4* means that the final plot will have four columns. *scales = fixed* means that the scale for each plot is the same.

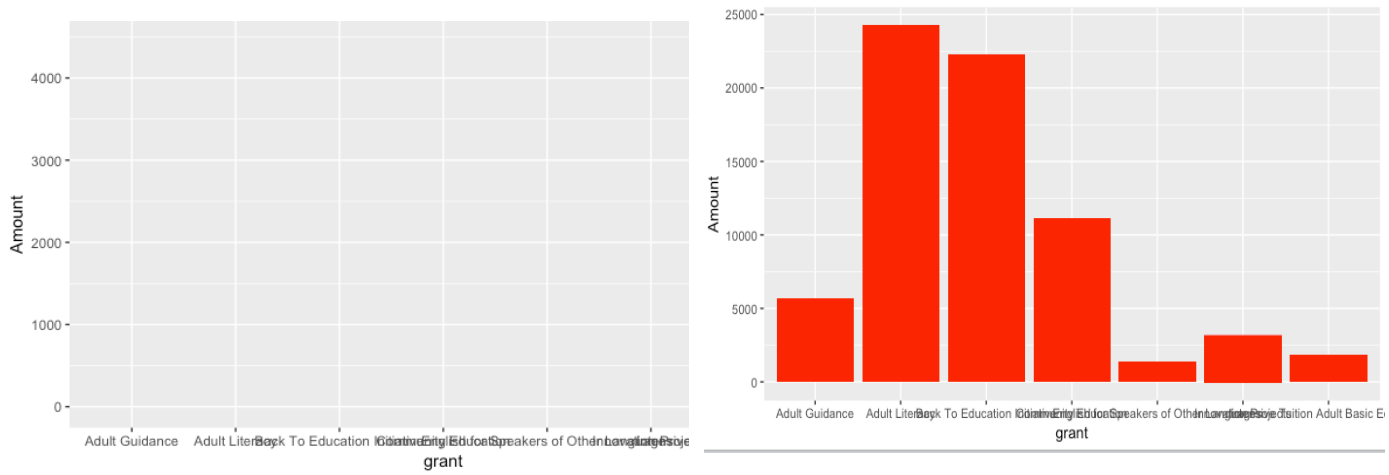
The fourth and fifth layers are known as theme layers and allow control over the plot elements.

theme(axis.text.x=element_text(angle = -90,size =5))+

This layer allows us to change the x axis text labels to a vertical orientation (coded as *angle = -90*) with font size of 5 (*size = 5*) This plot is shown in bottom right.

ggplot2(grantsolaslong,aes(grant,Amount))

+ geom_bar(stat="identity",fill = “red")



+facet_wrap(~ETB,ncol = 4,scales = “fixed")

+ theme(axis.text.x=element_text(angle = -90,size =5))

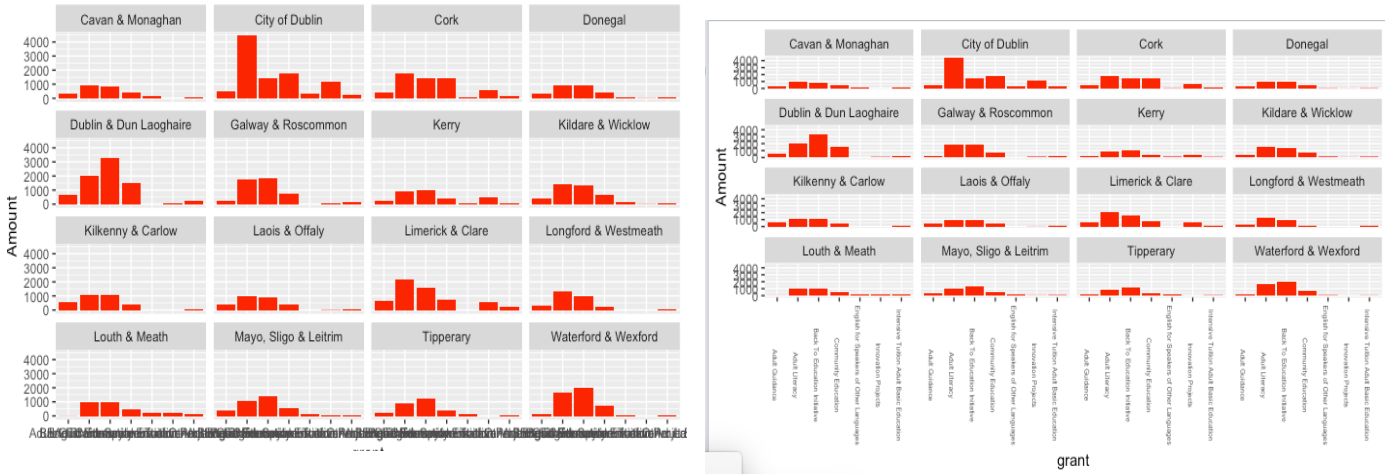


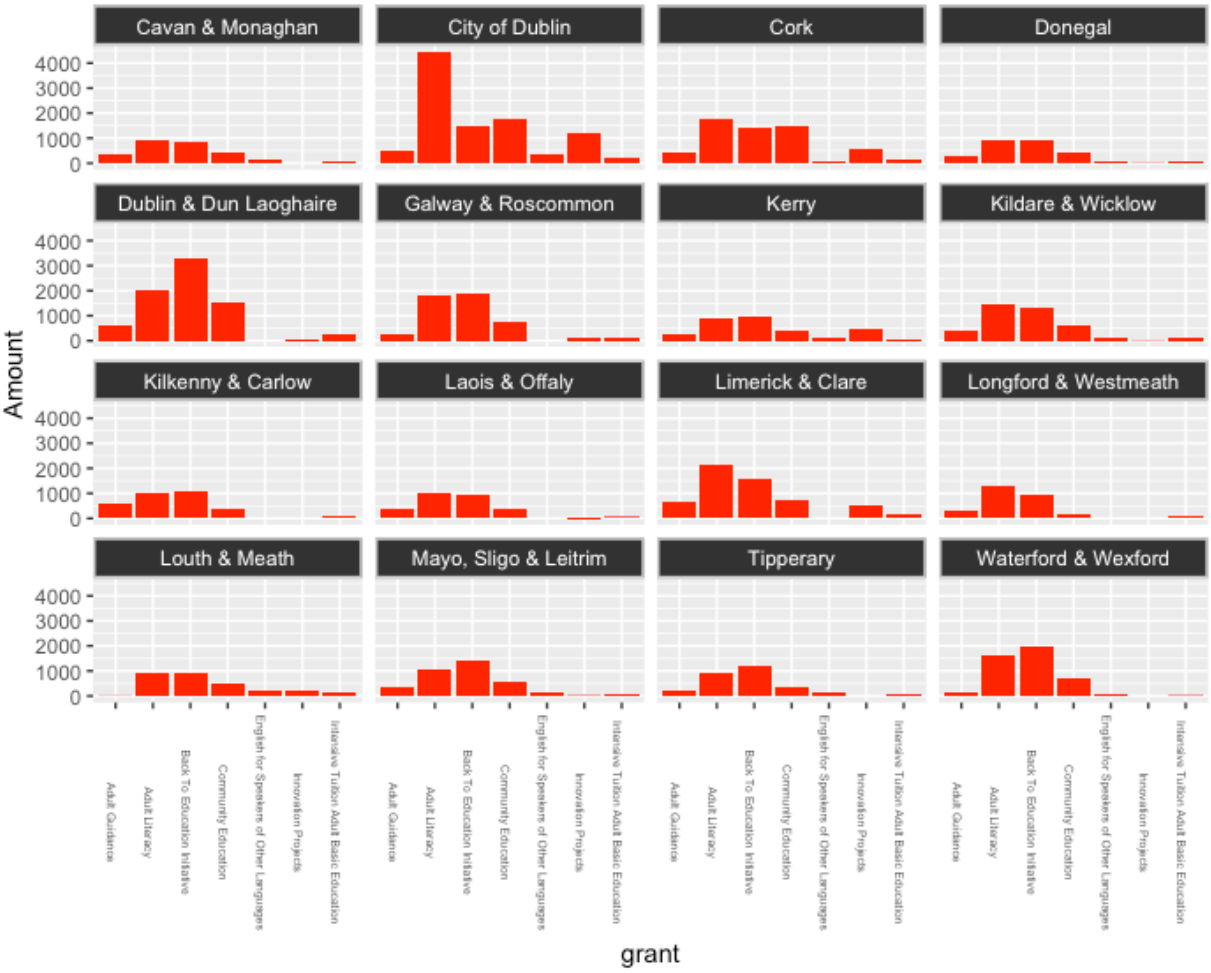
Figure A.1: Creating a plot by the addition of layers in ggplot2

The final layer sets the text located on the top of each panels to white text with the rectangle containing the text set to black. The code of this background is:

theme(strip.background = element_rect(fill = "grey20",colour = "grey80",size =1),strip.text = element_text(colour = "white"))

The complete plot with associated code is shown below:

```
ggplot(grantsolaslong,aes(grant,Amount))+
geom_bar(stat="identity",fill = "red")+
facet_wrap(~ETB,ncol = 4,scales = "fixed")+
theme(axis.text.x=element_text(angle = -90,size =5))+
theme(strip.background = element_rect(fill = "grey20",colour = "grey80",size =1),strip.text = element_text(colour = "white"))
```



Appendix 2: R Code to Generate Graphics

The R code used to generate most of the plots used is provided below. Please note that the data files need to be imported and the libraries ggplot2 and vcd need to be installed. The appendix also contains the code used to convert data from wide format to long format with requires the R library tidyr.

```
ggplot2(Appreg,aes(Category,Number)) + geom_bar(stat ="identity",fill = "Blue") +
coord_flip()
```
[Figure 1.1]

```
ggplot2(Appreg, aes(x = reorder(Category,Number), y = Number)
+geom_bar(stat="identity", fill ="red") + coord_flip()
```
[Figure 1.2]

```
CapelStreet$Date <- as.Date(CapelStreet$Date,format = "%d/%m/%y")
ggplot2(CapelStreet,aes(Date,PedestrianFlow)) + geom_line(size =1,colour = "blue")
+xlab("Year")+ylab("Number")
```
[Figure 1.4]

```
ggplot2(FrontAge, aes(Front))+geom_histogram(bins = 10, show.legend=NA) + xlab("Age")
+ylab("number")+ggtitle("Age of front seat passengers")
```
[Figure 1.5]

```
ggplot2(FrontAge, aes("",Front))+geom_boxplot() + xlab("")+ylab("number")+ggtitle("Age of
front seat passengers")
```
[Figure 1.7]

```
ggplot2(FrontAge, aes("",Front, fill = "blue"))+geom_violin() + xlab("")+ylab("number")
+ggtitle("Age of front seat passengers")
```
[Figure 1.8]

```
ggplot2(FrontAge, aes("",Front))+geom_point() + xlab("")+ylab("number")+ggtitle("Age of
front seat passengers")
```
[Figure 1.10]

```
ggplot2(FrontAge, aes("",Front,size = 3,colour = "grey"))+geom_jitter() + xlab("")
+ylab("number")+ggtitle("Age of front seat passengers")
```
[Figure 1.11]

```
ggplot2(DrPaAge, aes(DrAge,PasAge))+geom_point(size =1, colour = "black") +
xlab("Driver Age")+ylab("Passenger Age")
```
[Figure 2.1]

ggplot2(Quote,aes("",Quote)) + geom_boxplot(fill = "red")+ facet_wrap(~Age, scales= "fixed") **[Figure 2.6]**

ggplot2(Dacc1,aes(Weekday,Number)) + geom_bar(stat ="identity",fill = "Blue")+ facet_wrap(~Type,ncol=2,scales = "free")+ scale_x_discrete(limits= c("Mon","Tues","Wed","Th","Fri","Sat","Sun")) **[Figure 2.9]**

mosaic(xtabs(Exposure ~ Gender + Age,data = Policy1), direction = c("v", "h"),gp = gpar(fill = c("red"))) **[Figure 2.13]**

ggplot2(BurglaryRate,aes(Longitude,Latitude,size = Rate)) + geom_point()+ xlab("Longitude") + ylab("Latitude") + ggtitle("Crime Rate by Garda Station per 100,000 Population") **[Figure 3.2]**

ggplot2(StLoc,aes(Latitude,Longitude)) + geom_point(size = 2)+ xlab("") + ggtitle("Distribution of Students by Program") + facet_wrap(~Program,ncol = 4) **[Figure 3.4]**

ggplot2(grantsolaslong,aes(grant,Amount))+geom_bar(stat="identity",fill = "red") + facet_wrap(~ETB,ncol = 4,scales = "fixed") + theme(axis.text.x=element_text(angle = -90,size =5))+theme(strip.background = element_rect(fill = "grey20",colour = "grey80",size =1),strip.text = element_text(colour = "white")) **[Figure 3.5]**

ggplot2(Ins2,aes(Age,Premium))+geom_bar(stat="identity",fill = "red")+ facet_grid(Gender ~Licence) + geom_text(aes(label=Premium), vjust=-1.0) **[Figure 3.7]**

mosaic(xtabs(Number ~ Region + Gender+Diagnosed,data = Canchild), direction =c("v","v","h"),gp = gpar(fill = c("red","grey"))) **[Figure 3.8]**

Appendix 3: Converting data from wide to long format using R library Tidyr

The data in table A.2 (*Quotewide in Data(2018).xls*) in containing insurance quotations for 27 consumers aged 20, 30 and 50 is in wide format. This is so because each value of the grouping variable age has its own column. To convert this file to long format each of the age columns will be merged to form two new variables called age and quotation. This is called long format.

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 A.2: Quotation data in wide format

The data in table A.1 is stored in R as QuoteWide and to convert to long format we can use the R library **tidy** and write the code:

```
QuoteLong <-gather(QuoteWide,key = Age,value = Quotation, 1:3)
```

QuoteWide is the name of the existing data set, **key = Age** means that we will be merging the three age columns into one single column which will be called Age. This new variable will contain values 20, 30 or 50 and in statistical terminology is called the independent variable. The code **value = Quotation, 1:3** means that the quote for each value of age will be placed in a new column called Quotation while **1:3** means that the quotes are contained in columns 1 to 3.

In statistical terminology quote is called the dependent variable. The code **QuoteWide <-** means put the output of gather into a new variable called **QuoteLong**. Applying this code the variable QuoteLong is now generated with two columns - Age and Cost as shown below in Table A.2.

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

Table A.3: Quotation data in long format

The new long format QuoteLong can be exported in .csv format using the code

```
write.table(QuoteLong, file = "QuoteLong", sep = ",", row.names = FALSE);
```

The new file QuoteLong will then appear in the R directory.

Appendix 4: Other R based (Static) data visualisation packages

Package	Description
graphics	Sometimes referred to a the base graphics environment as it was the first R package for data visualisation.
lattice	Specialises in lattice of facet plots originally proposed by Bill Cleveland - one of the early pioneers of data visualisation.
vcd	Visualising categorical data. This package provides techniques for visualising categorical data including mosaic plots.
GGlally	Sometimes regarded as a helper package of ggplot22. GGally allows for sophisticated visualisation of scatterplot matrices, parallel co-ordinate plots and network graphs. Main function call is ggpairs(). The equivalent function in the graphics package is called pairs()
ggmap	This package is based on ggplot2 and allows for mapping of web based maps e.g. google maps.
map	Creates maps of countries and regions but is limited to a small number of countries.

References

[1] Unwin, U. (2015), Graphical Data Analysis with R, CRC Press.

[2] Tukey, J. (1977), Exploratory data analysis, Addison-Wesley.

[3] https://www.findlatitudeandlongitude.com/batch-geocode/#.WahCCzOZM7w Accessed August 2018

[4] Tufte, E.R. (1983), The Visual Display of Quantitative Information, Graphics Press.

[5] Cleveland, W.S. (1993), Visualising Data, Hobart Press.

[6] Cleveland, W.S. (1994), The Elements of Graphing Data, Hobart Press.

[7] Bertin, J. (1967), Semiologie Graphique, Paris: Editions Gauthier-Villars. English translation by W.J. Berg as Semiology of Graphics, Madison, WI: University of Wisconsin Press, 1983.

Further Reading

Few, S. (2004), Show me the numbers: designing tables and graphics to enlighten, Analytics Press.

Teutonico, Donato (2015), ggplot22 Essentials, PACKT publishing.

Tufte, E.R. (2007), Beautiful Evidence, Graphics Press.

Tufte, E.R. (1990), Envisioning Information, Graphics Press.

Tufte, E.R. (1997), Visual Explanations: Images, Evidence and Narrative, Graphics Press.

Unwin, U. et. al. (2007), Graphics of Large Data Sets, Visualizing a Million, Springer.

Yau, Nathan (2013), Data Points – Visualization that Means Something, Wiley.

Wickham, Hadley (2016), ggplot22 – Elegant Graphics for Data Analysis, Springer

Willis, Graham (2012), Visualizing Time, Springer