6. Exploratory Data Analysis

Introduction

In earlier sections of the module a wide variety of graphics appropriate for ID, 2D and MD (or univariate, bivariate or multivariate) data were illustrated together with some guidelines on optimum ways of presenting such graphics. We were mainly concerned with **static** graphics used primarily to communicate, if possible to a wide audience, the results of data analysis. The gold standard adopted was in the words of Edward Tufte *to design graphics that* ***communicate*** *the main characteristics of the data in the least possible time (with the lowest amount of ink) to the viewer*. In this section we turn our attention to the use of **dynamic interactive graphics** as important data visualisation and exploration tools.

Dynamic graphics are in a sense 'work in progress' illustrations and are valuable aids in exploring data sets. The objective of exploratory analysis is to discover interesting patterns and structures in data sets that perhaps would remain hidden using more conventional approaches to data analysis. Conventional data analysis (and research programmes in general) are based on the scientific principles of formulating a hypothesis, collecting appropriate data and finally using some test statistic to decide on the validity of the hypothesis. On the other hand exploratory data analysis starts out with no preconceived hypothesis. This approach to data analysis has been made possible in recent years through the development of powerful interactive software.

This approach to data analysis was proposed by **John Tukey** in the 1960s and 1970s when he developed several simple, new and effective graphical displays. Included among these new innovations were the box plots we examined in earlier sessions and stem and leaf plots. Tukey suggested that we examine our data as a detective would examine the scene of a crime - not with a hypothesis (*I'll bet the butler did it*), but with an **open mind** and as few assumptions as possible.

According to Tukey by letting the data speak to us we hope to learn the *truths hidden beneath the random fluctuations, errors and general confusion seen in real data*. With his widely acclaimed text *Exploratory Data Analysis* published in 1972, Tukey had succeeded in creating a new found interest in statistical graphics and data visualisation.

In this section we introduce a software application called DataDesk. DataDesk is a very powerful yet comparatively inexpensive software application used to explore data. It was originally developed on the Apple Macintosh platform by Apple research fellow, **Paul Velleman**, but in recent years has become available on the Windows platform. The principle feature of DataDesk is the ability to interact at **speed** with multiple linked views of a dataset, so that, for example, selecting a subset of cases in one view highlights them in all other views. Our first task is to introduce the visual tools of DataDesk using a comparatively large data file containing a subset of variables from the Irish road accident data from 1996 to 2015.

6.1 Interactive Tools

As discussed one of the most powerful features of Data Desk the ability to interact with and explore data visually. A number of tools in DataDesk facilitate this interaction as shown in the **Tools** palette below. This is a floating palette and can be displayed on screen by selecting **Tools** from the **Modify** menu. In this section we will briefly describe each tool.

Slicing and Brushing

Brushing and slicing tools using the **brush** and **knife** can reveal joint patterns and relationships among many variables. For example, assume we have just two open plots - a bar chart of accidents by type and a bar chart of accidents by weekday. By selecting bar 1 (pedestrian accidents) on the primcoltype type barchart using the knife icon the distribution of the 1,200 or so pedestrian accidents by each weekday is shown on the weekday bar chart. From Figure 6.1 it seems that bar 6 (Friday) has the highest number of pedestrian accidents.

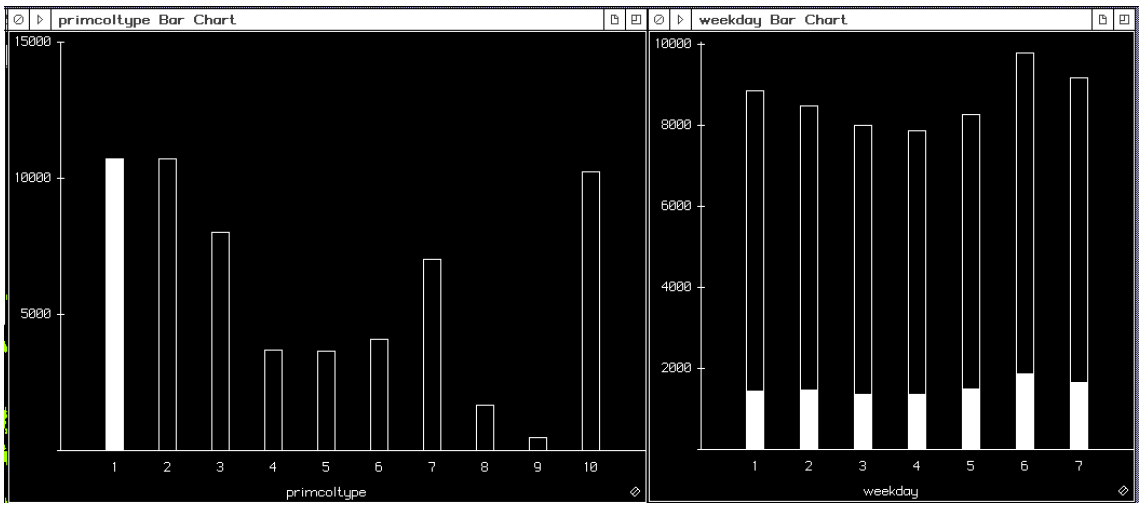
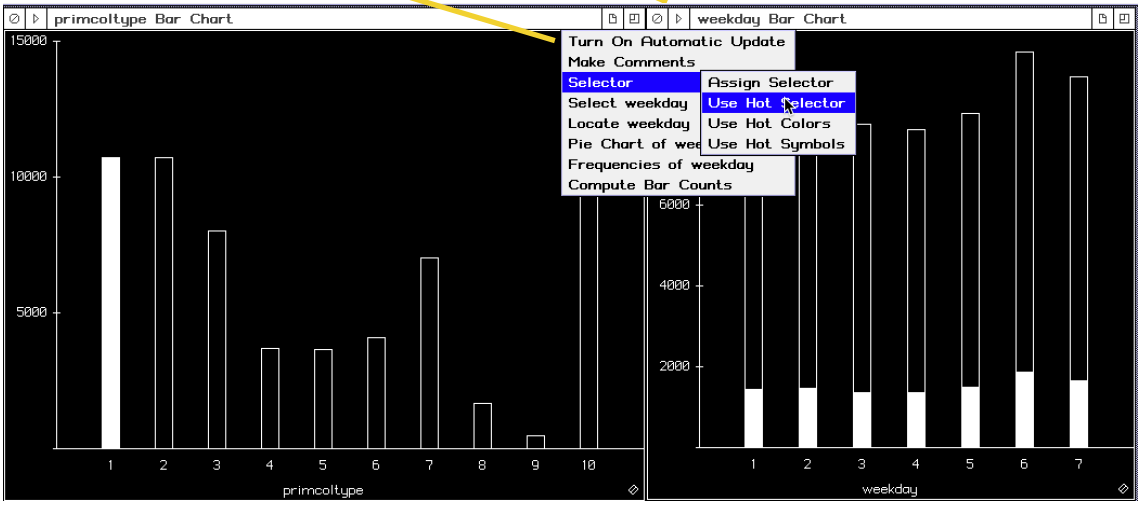


Figure 6.1: Selecting accident type using the knife

The Hot Selector

One problem with the above plot is that is can be hard to see the pattern of pedestrian accidents. This is because each bar includes non-pedestrian accidents with no highlighting. If non-pedestrian accidents comprise a large number of accidents (as is the case here) it is hard to compare the highlighted white areas. Visualising the white areas **alone** (i.e. just the pedestrian accidents) allows for easier comparisons between type and weekday. To display the white areas alone the following tasks are required:

- Select the **hyperview** window of the **weekday** bar chart.
- Choose **Selector** and the sub menu **Use Hot Selector.**
- Select **Turn on Automatic Update**



Now selecting any bar in the primcoltype chart (using the knife icon) shows those accidents **alone** in the weekday chart. This is shown in Figure 6.2 for pedestrian (top) and single vehicle (bottom) (i.e. bar values 1 and 2 of primcoltype, respectively).

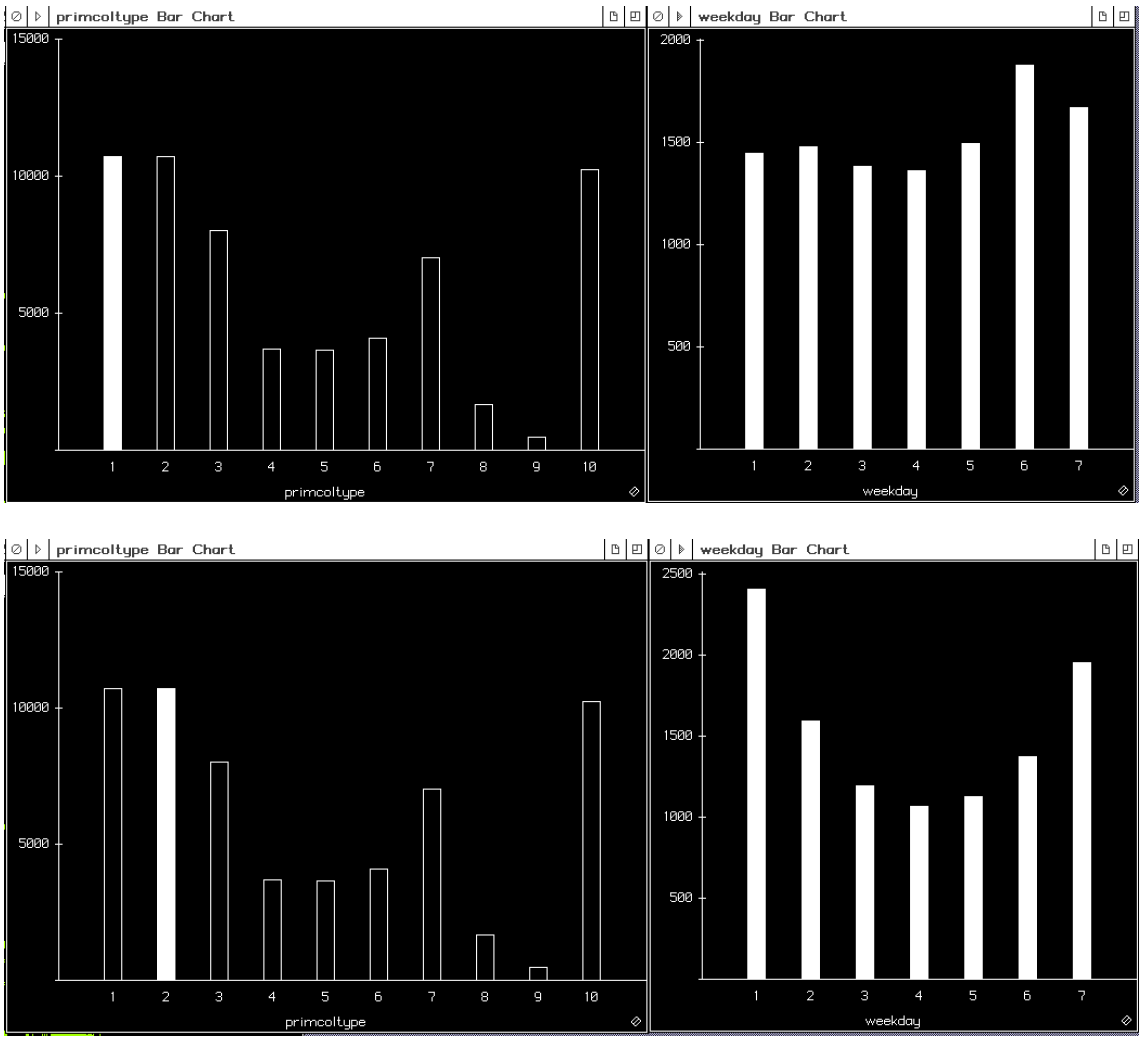
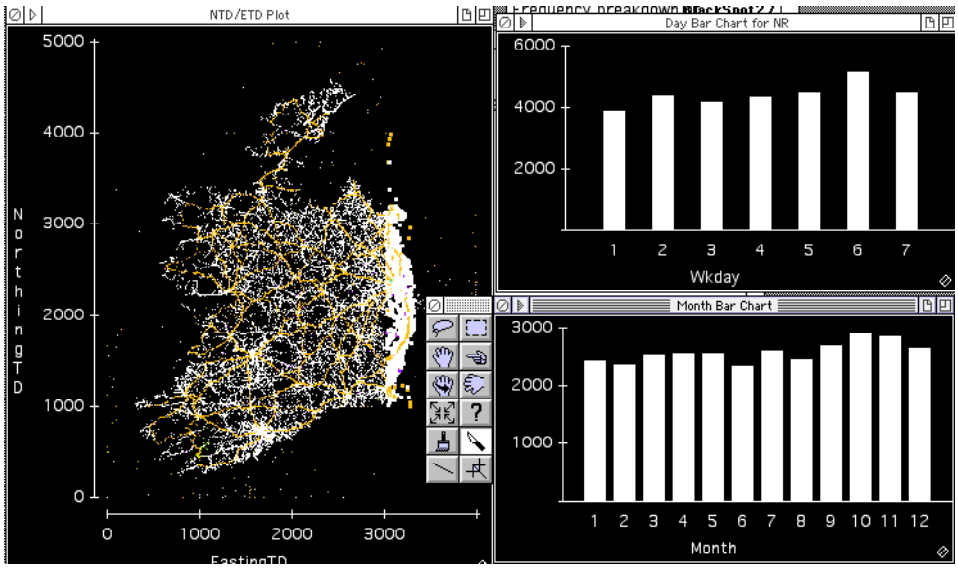


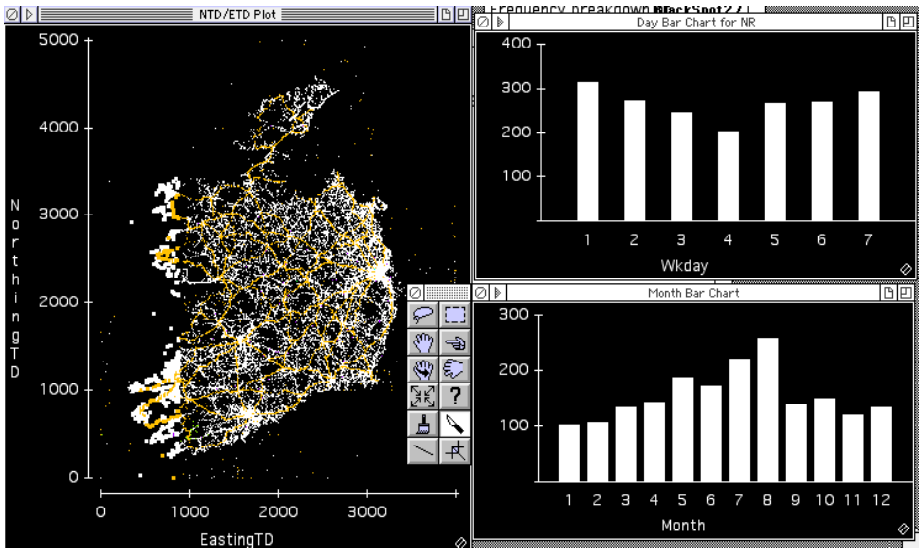
Figure 6.2: Selecting accident types 1 and 2 using the knife with the hot selector activated

By activating the **hot selector** and **turn on automatic update** options the linked plots will automatically update when the user selects different bars on the primcoltype chart. The difference in weekday patterns is striking with **single vehicle accident** (code 2) having a strong U shape compared with the Pedestrian profile. The hot selector is thus a useful visualisation tool allowing analysts to explore their data and detect patterns that may not be identified using more conventional static displays. While plotly, Tableau and other interactive graphics can create the above plots DataDesk has the advantage of having more advanced dynamic and interactive capability.

The hot selector together with the knife icon can be used to dynamically visualise relationships between **any number** of plots open on the desktop. For example, the screenshot below contains three plots - a scatterplot map of accidents and two bar charts of accidents by weekday and month. In the scatterplot the knife has sliced over the edge of the **east cost**. Because the bar charts have their *hot selector* and *turn on automative update* activated the distribution of east cost accidents can now be visualised by weekday and month dynamically. The plot suggests that accidents along the east cost are fairly constant by weekday and month.



If the selector is moved across to the **west coast** the charts update automatically once the **hot selector** and **Turn on Automatic Update** remain activated. The distribution of accidents by weekday and month is markedly different to the east coast as shown in the screenshot below. Accidents by weekday in the west are lowest during midweek and highest at the weekends while accidents by month are highest during the summer months and lowest during the winter months.



This example illustrates the power of interactive visualisation using 4 data variables - longitude, latitude, month and weekday.

Brush

The brush tool is used primarily to explore scatterplots and when placed on the display a rectangle is formed that can be used to **brush** through scatterplots. As you brush through the scatterplot, points are temporarily highlighted, as are the corresponding points in all open linked displays. In addition, if the hot selector option is activated all displays will **dynamically update** as the brush travels over the display. Brushing scatterplots was developed by the data visualisation pioneer **William Cleveland** who also invented the **trellis plot** discussed in Sections 2 and 3.

For example Figure 6.3 shows three plots - a scatterplot map of accidents in Ireland and two bar charts of **accident type** (PCTD) and **light condition** (LightD) at the time of the accident. In the screenshot Donegal has been selected by the brush. Both bar charts have the **hot selector** and **Turn on Automatic Update** activated. From the top bar chart values H-C, OSV and Pdn are the collision types in Donegal reporting the highest accident frequencies. These abbreviations refer to *Head-On*, *Single Vehicle* and *Pedestrian* collisions, respectively. The bottom bar chart of light conditions reports the highest frequency for D&V (daylight and good visibility).

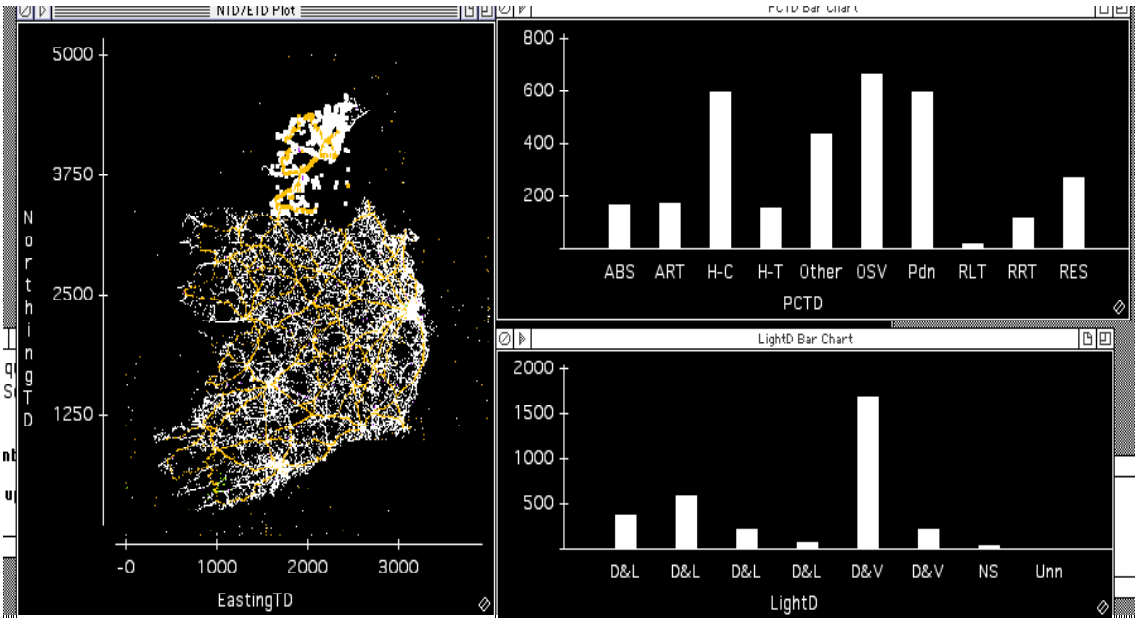


Figure 6.3: Brushing **Donegal** with linked bar charts of collision type and daylight

Moving the brush to the Dublin area a change in the pattern of collision type and light condition frequencies is evident as shown in Figure 6.4. The accident category reporting the highest collision type frequencies are Pdn and Other (Other accidents generally refer to pedal cyclists). As with Donegal most accidents in the Dublin area occur with code D&V (daylight and good visibility) but very few with code D&L (dark with no lighting) as might be expected in a predominately urban area with street lighting.

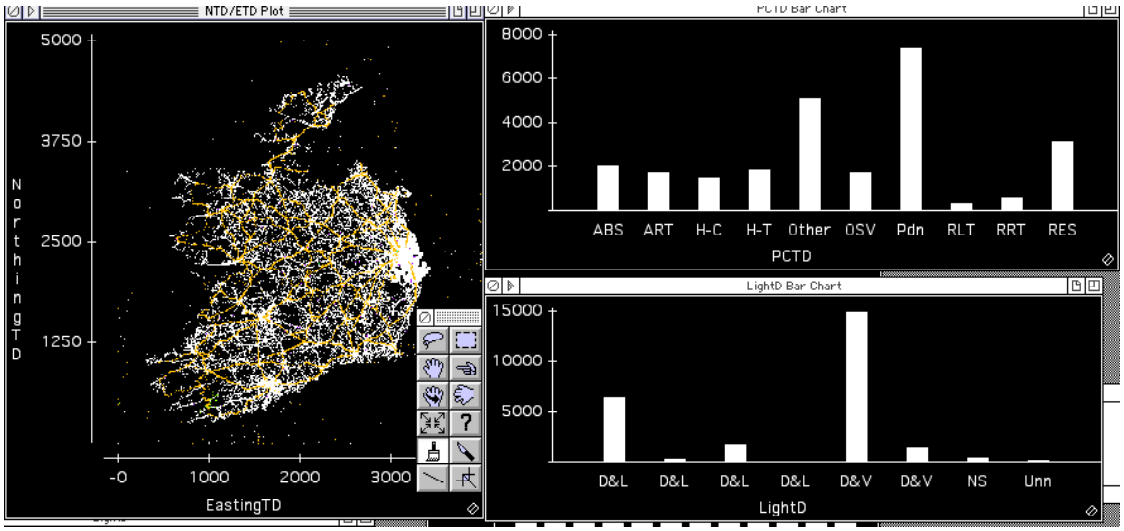


Figure 6.4 Brushing **Dublin** with linked bar charts of collision type and daylight

Lasso, Rectangle and Pointer

To select points with the **Lasso**, drag the icon around the points. When the button is released the shape drawn is automatically closed, and all enclosed points are selected. This is similar to the lasso icon in Tableau.

To select points with the **Rectangle**, hold down the mouse button and drag out a rectangle on the plot. When you release the button, all enclosed points are selected.

The **Pointer** operates on all plots. When the mouse is pressed, it selects the data point or part of the plot it is pointing to. It selects individual points in scatterplots but selects entire bars in histograms and entire wedges in pie charts.

Identifier

The **Identifier** tool plots a cross hair cursor that looks like a bomb sight. Place it over a plotted point and press the mouse to display the points case number. To display the value of a variable corresponding to the selected point ensure the variable window is open.

Grabber

The **Grabber** icon repositions the contents of a plot within its window providing a natural way to adjust displays.

Resize

This tool resizes the plot by zooming in or out. Place the mouse curser inside the plot window. If the mouse is near the centre of the plot the cursor icon changes to reflect zooming in. Move the mouse away from the centre of the plot and the curser changes to reflect zooming out.

Refocus

Refocus allows the user to highlight a selected region of a display which then expands to fill the screen. For example, selecting the Dublin area using the refocus icon in Figure 6.5 (left plot with selected area coloured) the Dublin area now expands to fit in the whole window (right hand plot).

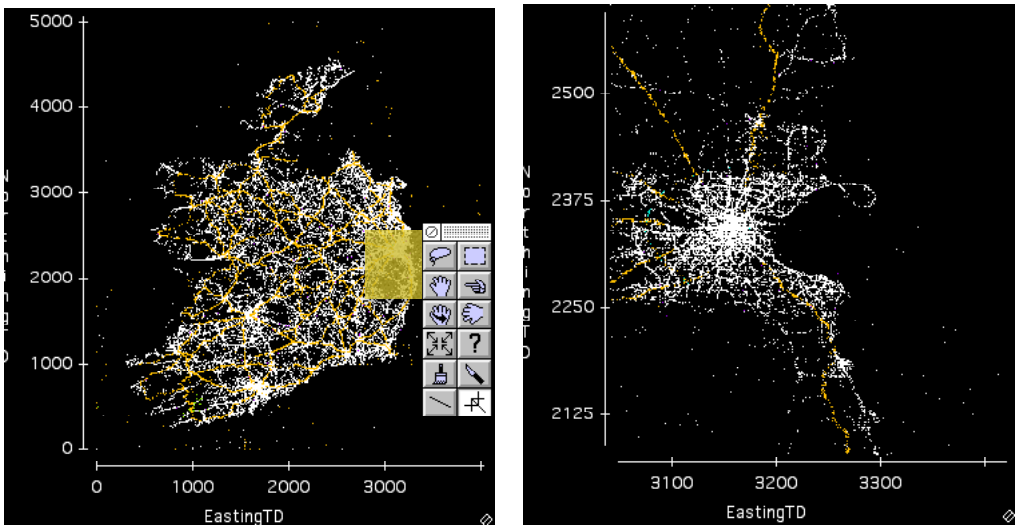


Figure 6.5: Using the refocus tool in the Dublin area

Rotate

The ability to rotate data is a powerful tool for understanding the relationship between three or more variables. Rotating plots often detect patterns that are obscured in conventional two-dimensional representations. To compute a rotating plot select the y (up-down axis), x (left right axis) and z (in-out axis) variables and select **Rotating Plot** from the **Plot** menu or use the rotate icon in the palette menu.

The first computer application developed for rotating data was known as the **PRIM** system developed by Fisherkeller, Friedman and Tukey in 1972 at the **Stanford Linear Accelerator Centre**. It ran on an IBM system and required a few million dollars worth of computer and display hardware, (the display unit was $400,000 alone) and cost several hundred dollars an **hour** to use - so it remained a prototype system! A video image of John Tukey in front of the PRIM is in Figure 6.6.

Figure 6.6: John Tukey with PRIM computer

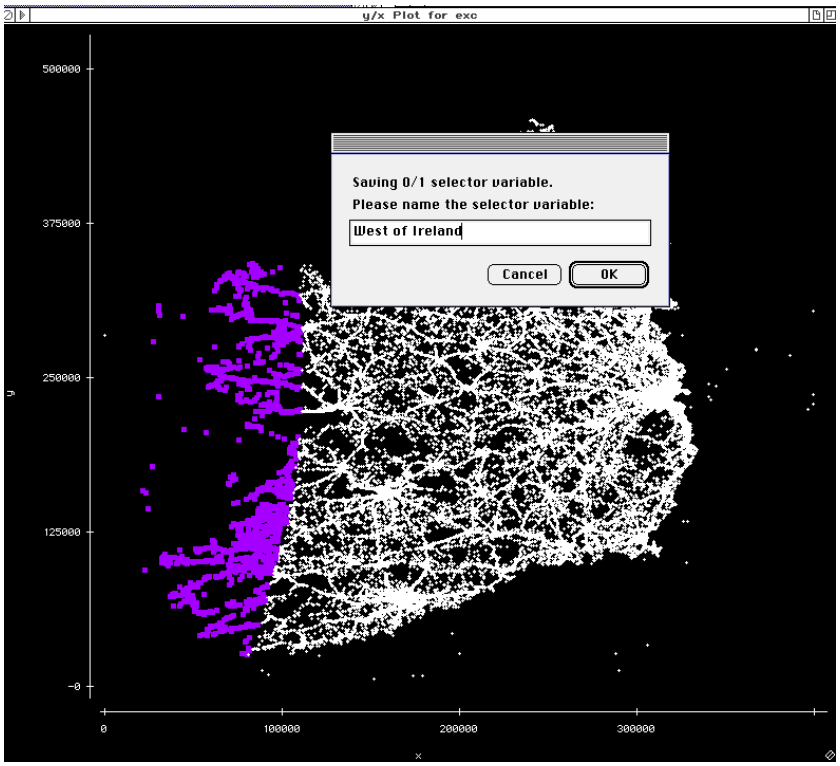
PRIM stood for **P**rojection, **R**otation, **I**solation and **M**asking the elementary operations that were found to be a basis for using plot rotations in data analysis. The plot operations of Data Desk include each of these PRIM operations and enhances them beyond what was possible in the prototype systems.

The principal benefit of rotating plots is that it allows the analyst to readily see patterns or clusters in a data set that would otherwise remain hidden using conventional data analysis procedures. These patterns can then be isolated for further analysis.

6.2 Creating interactive filters by recording screen selections

There are many situations in data analysis when we would like to create customised interactive filters. For example, we may want to compare accidents in the west of Ireland with accidents along the east coast. The visual tools of DataDesk can be used to allow direct selection of cases from the viz. Tableau also allows for this direct interaction using the lasso and other tools but not as seamlessly as DataDesk. For example, to create a filter comprising west of Ireland accidents the following approach can be adopted:

- Select the region of the plot you are interested in using the knife, lasso etc.
- Choose **Selector** from the **Manip** menu
- Select the menu-item **Record**



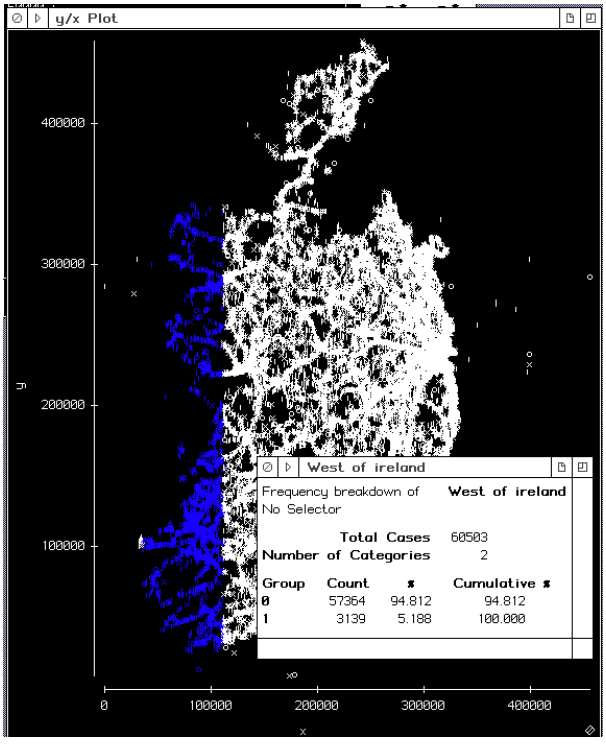
Give the selector a name **e.g. West of Ireland** and a new variable with this name will be visible on the desktop.

To analysis road accidents restricted to the **West of Ireland** apply the following steps:

- Select the **West of Ireland** variable
- Select **Selector** from the **Special** menu
- Select the menu-item **Assign**

All plots and graphs computed with the selector variable **West of Ireland** highlighted will now be computed for west of Ireland accidents only.

A frequency breakdown of the **West of Ireland** variable contains value codes 1 and 0. The value 1 corresponds to accidents in the West of Ireland and the value 0 to accidents everywhere else in Ireland. Selecting the number 1 in the frequency table will visually highlight **West of Ireland** accidents while selecting 0 with highlight all other accidents. This is a useful check that the selector variable is computed correctly as shown below.

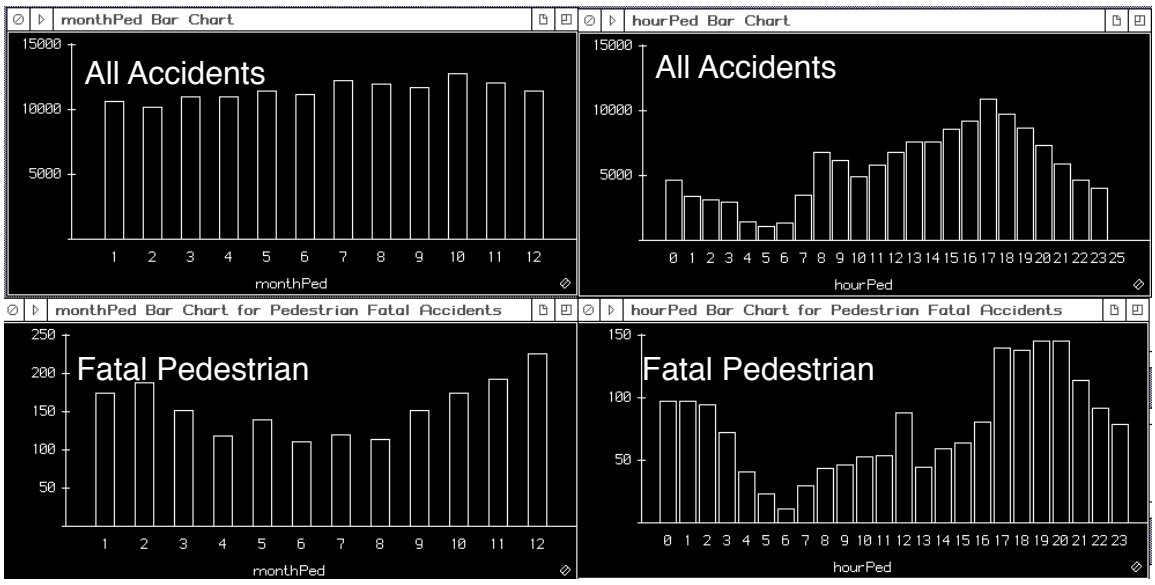


Exercises

Exercises 1-3 are based on reported road traffic accidents in Ireland between 1990 and 2015. The data files are contained in folders provided the DataDesk file **Road Traffic Accidents Accidents.dsk** located the DataDesk folder

Question 1: Data provided in DataDesk folder: Pedestrian

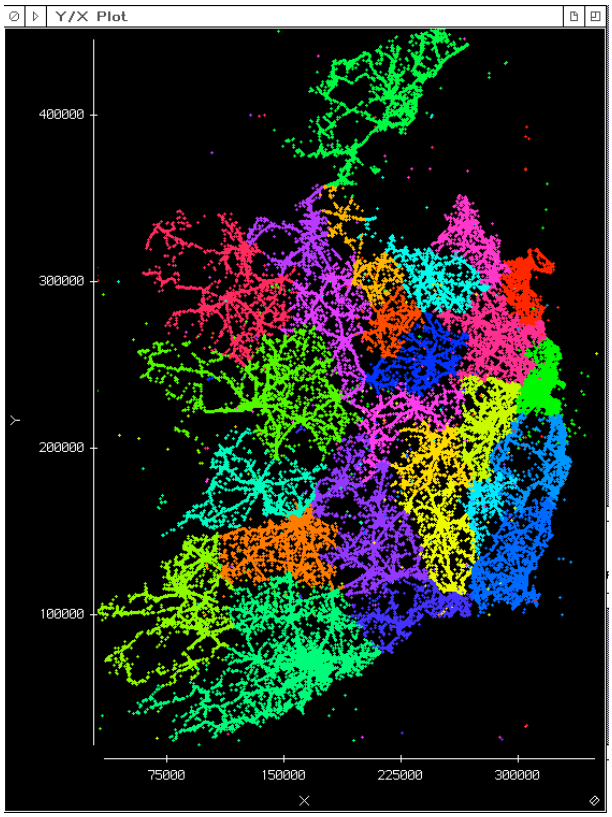
i) Contrast the bar chart of **all accidents in Ireland** by month and by hour with the bar chart of **fatal pedestrian accidents** by month and by hour as shown below. What are the main differences by month and by hour between all accidents and fatal pedestrian accidents?



ii) Using the knife tool slice the hourly bar chart for fatal pedestrian accidents and see if you can find any clues to the cause of the monthly U shape in accidents. Remember to activate **hot selector** and **turn on automatic update** on the month bar chart. The data files are provided in the folder **Pedestrian**.

Question 2: Data provided in DataDesk folder: Map

i) The scatterplot map below is the location of reported injury accidents in Ireland between 1999 and 2015 with county encoded by colour. Using the **knife** and **brush** tools examine any difference between east and west coast accidents by hour, weekday and month. The data files are provided in the folder **Map**.

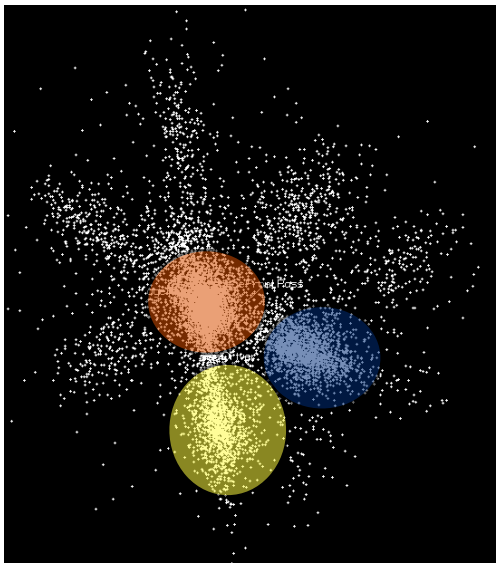


ii) Can you suggest any reason why a difference in patterns between the east and west coast is evident for the variables **hour**, **prcoltype** and **month**?

Question 3: Data provided in DataDesk folder: Rotating Plot

i) Create a **rotating plot** using the variables age of driver (ageDriver), front seat passenger age (ageFrontPass) and rear seat passenger age (ageRearPass).

ii) Using the tools of DataDesk experiment with the rotating plot until you get the star shaped image shown below.



iii) Using the selection tools of DataDesk e.g. lasso select and **record** the **three** coloured clusters shown above giving each cluster a name.

iv) Create an appropriate chart for each of the following variables in the folder:

ageDriver	**ageFrontPass**	**ageRearPass**
GenderDriver	**GenderFrontPass**	**GenderRearPass**
prcoltypR	**weekdayR**	**hourR**

v) Activate **hot selector** and **turn on automatic update** for each of the nine charts computed in iv)

vi) Select **1** on the **frequency breakdown table** for each of the three recorded clusters and observe any difference in patterns across each of the nine charts variables. See can you can identify a **road user profile** associated with each of the three clusters.

vii) Import the graphical displays consisting of 27 plots (9 charts by 3 clusters) into a word processing document

Question 4

The data in the Excel worksheet **RTA** in the file *ExercisesData(2018).xlsx* records19 variables associated with road accidents collected by An Garda Siochana.

i) Create a scatterplot map using the longitude and latitude variables.

ii) Use the lasso tool or otherwise create three selector variables by **recording** selections from the scatterplot accident map computed in i) corresponding roughly to accidents in i) West of Ireland ii) East Coast and iii) Cork City.

iii) Create a frequency breakdown of each of the three recorded variables in ii) and investigate differences in the patterns in plots of the following variables:

- time of day defined as **hour** in the worksheet **RTA**
- primary collision type defined as **primarycoll** in **RTA**
- gender of each of the three vehicle occupants driver, front seat passenger and rear seat passenger defined as **sexdriver**, **sexfrontpass** and **sexrearpass** in **RTA**.
- age of each of the three vehicle occupants driver, front seat passenger and rear seat passenger defined as **agedriver**, **agefrontpass** and **agerearpass** in **RTA**

Note: ensure **hot selector** and **turn on automatic update are activated** for each chart.

iv) Import the graphical displays into a word processing document.

v) Write a report on your observations of the different patterns of the three accident locations for the variables in iii).