1. Data Visualisation

Introduction

There are many possible graphical techniques that can be used to explore and communicate patterns and structures in data. In this session a number of graphical methods will be introduced - some of which will be familiar to you and some which you will perhaps not have met before. Some graphics are available through standard applications like EXCEL whereas others require more specialised software like **DataDesk**, **Tableau** or **R**. Some indeed require to be drawn manually on paper or through the use of a drawing package such as Adobe Illustrator or Photoshop. The emphasis on this section is the creation of **static** graphics to communicate the results of some analysis. In the second part of the module we will explore **dynamic** and **interactive** graphics which are used to discover interesting patterns and structure in data. We will begin this session by describing two key steps in determining the appropriate graphical representation of your data - whether the data is **discrete** or **continuous** and the **dimensionality** of the data.

Discrete and continuous data formats

Graphics can be divided into those that are more appropriate for **discrete** data and those that are suitable for presentation of **continuous** data. The first step towards designing an appropriate graphic for your data is to determine if the data to be visualised are discrete or continuous. Discrete data has a limited number of distinct possible values or categories. For example, *day of week*, *county*, *educational Institute* and *eye colour* are discrete as these variables have distinct values or **categories**. Continuous data on the other hand has a large number of possible values which are usually **measurable**. For example, the grant awarded to an education training board for adult literacy, the longitude of a town or the number of domestic customers of a bank are three examples of measurable data.

Number of dimensions

The next step in the selection of an appropriate graphic is to determine the dimensions of the data set. We will use the terms **one-dimensional** (1D), **two-dimensional** (2D) and **multi-dimensional** data (MD) to describe the dimensions of a data set. It should also be noted that the terms 1D, 2D and MD data are often described as **univariate**, **bivariate** and **multivariate** data in statistical text books. One-dimensional data contains just one variable in a data set while 2D data contains two variables. MD data contains more than two variables.

For example, consider a data set of 100 data values with each data value the number of hours per week spent by an individual surfing the internet. There is only one variable in this data set (the number of hours surfing) so this data set can be considered 1D. If in addition to the number of hours we also know the *age* of each of the 100 respondents then the data set is 2D as it contains two variables - *number of hours* and *age*. If the sex of each respondent is added to the number of hours and their age then the data set is MD as it contains three variables which are *number of hours*, *age*, and *sex*.

According to Anthony Unwin, author of a number of textbooks on graphical data analysis, ‘to visualise a data set can require many different graphics analogous to a photographer taking several shots of the same image’ [1]. In the following sections of this Section we will demonstrate a number of ways to visualise 1D, 2D and MD data using the application **Tableau** and the R graphics libraries **ggplot2** and **vcd** (visualising categorical data). The intent is to illustrate to the analyst a number of visualisation options for some of the more common data formats.

It is hoped that on completion of this section of the course you will have been introduced to some new graphical forms, data sources and software applications for visualising data.

1.1 Visualisation of 1D Data

In this section a number of graphics useful for visualising one dimensional data will be illustrated. These graphical formats can also be used to visualise higher dimensional data. The following sections will demonstrate a number of methods suitable for one-dimensional discrete data followed by one-dimensional continuous data using a number of contemporary data sets.

1D Discrete Data

There are fewer graphical techniques available for the visualisation of discrete data when compared with continuous data. In this section we will outline three of the most common visualisations that can be used for discrete data which are Bar, Pareto and Time based charts.

i) Bar Chart

Bar charts visualise the total number of counts for the different categories of a discrete variable. For example, the plot below from the worksheet **Apprentice** in Tableau is a bar chart where each bar is the total count of apprentice registrations in Ireland in 2016 for each type of apprentice.

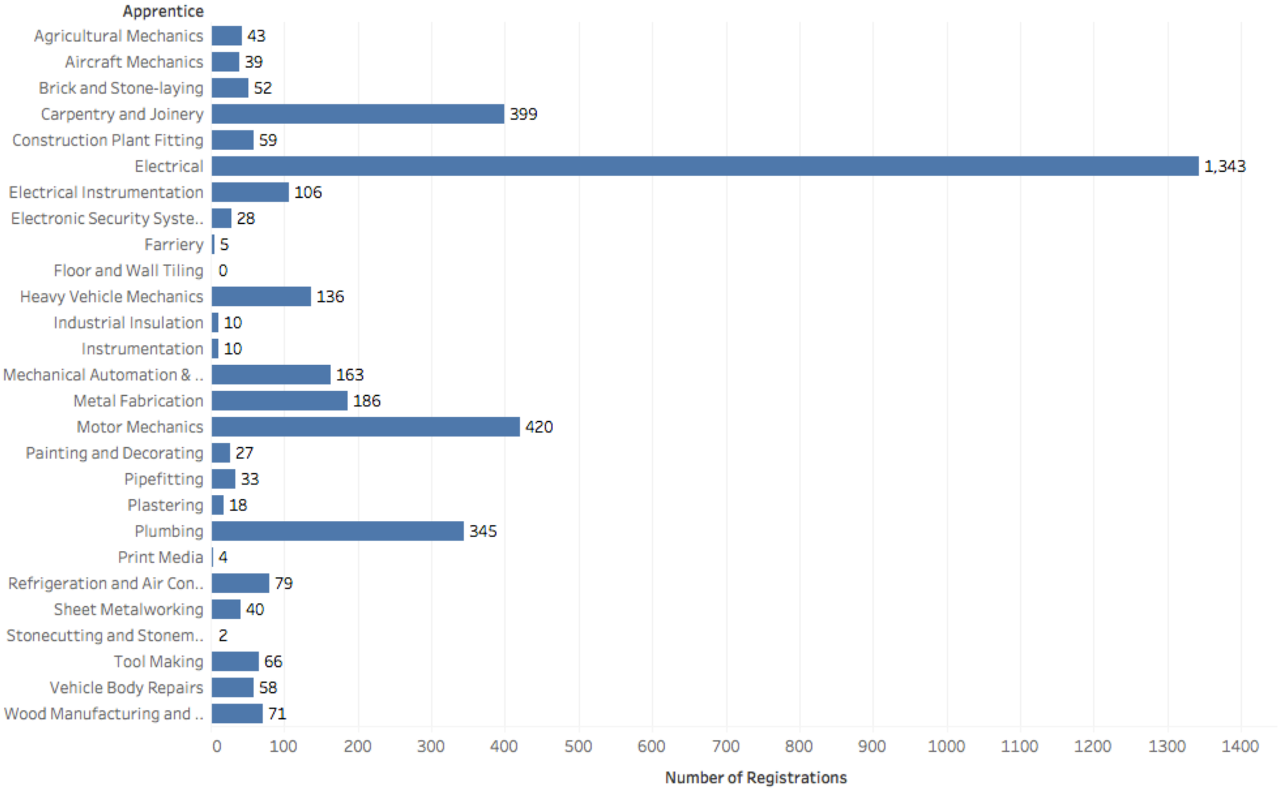


Figure 1.1: Bar chart of apprentice registrations, 2016

ii) Pareto Chart

Pareto charts are the same as a bar charts except that the bar counts are ordered (i.e. ranked) from highest to lowest in descending (or ascending) order. Pareto charts can be useful for highlighting the largest categories in a discrete variable. They are extensively used in continuous improvement strategies in business. For example, if there are many causes as to why a process is not working correctly ranking the causes by their frequency of occurrence can allow for prioritisation of remedial action. It is generally the case that when examining process problems it is found that they are caused by a small number of issues. This is sometimes referred to as the **80/20 rule** whereby roughly 80 per cent of process malfunctions can be attributed to just 20 per cent of all causes.

Example

Figure 1.2 is an example of a Pareto chart of the apprentice data shown in Figure 1.1 which can be obtained from the bar chart by clicking on the chart icon located on the top of the Tableau plot. This allows the apprentice categories to be sorted from highest to lowest or vice versa. The dominance of electrical apprentices is clearly evident from the plot.

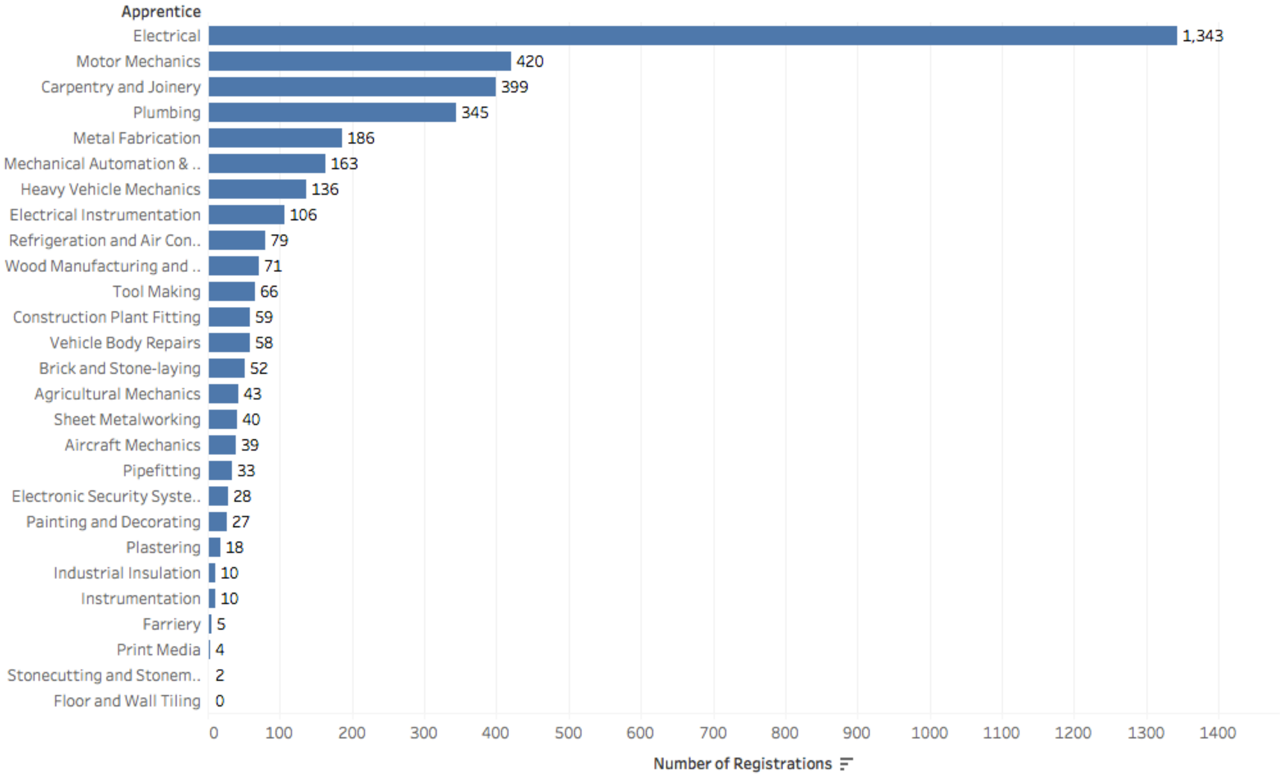


Figure 1.2: Pareto plot of apprentice registrations, 2016

iii) Time Plot

When data is plotted over time we can regard time as discrete or continuous. In this section we will regard time as a discrete variable. If the event at each point in time is a **count** the plot is one dimensional and can be plotted as a time series or line plot. In this case the x-axis measures time and the y-axis the corresponding count. An example of a time plot is provided in Figure 1.3 which plots the pedestrian foot flow in Capel Street in Dublin City Centre in 2015. To generate this plot select the worksheet **Capel Street** in Tableau and place the variable **Month** in the Columns shelf and **Total** in the Rows shelf.

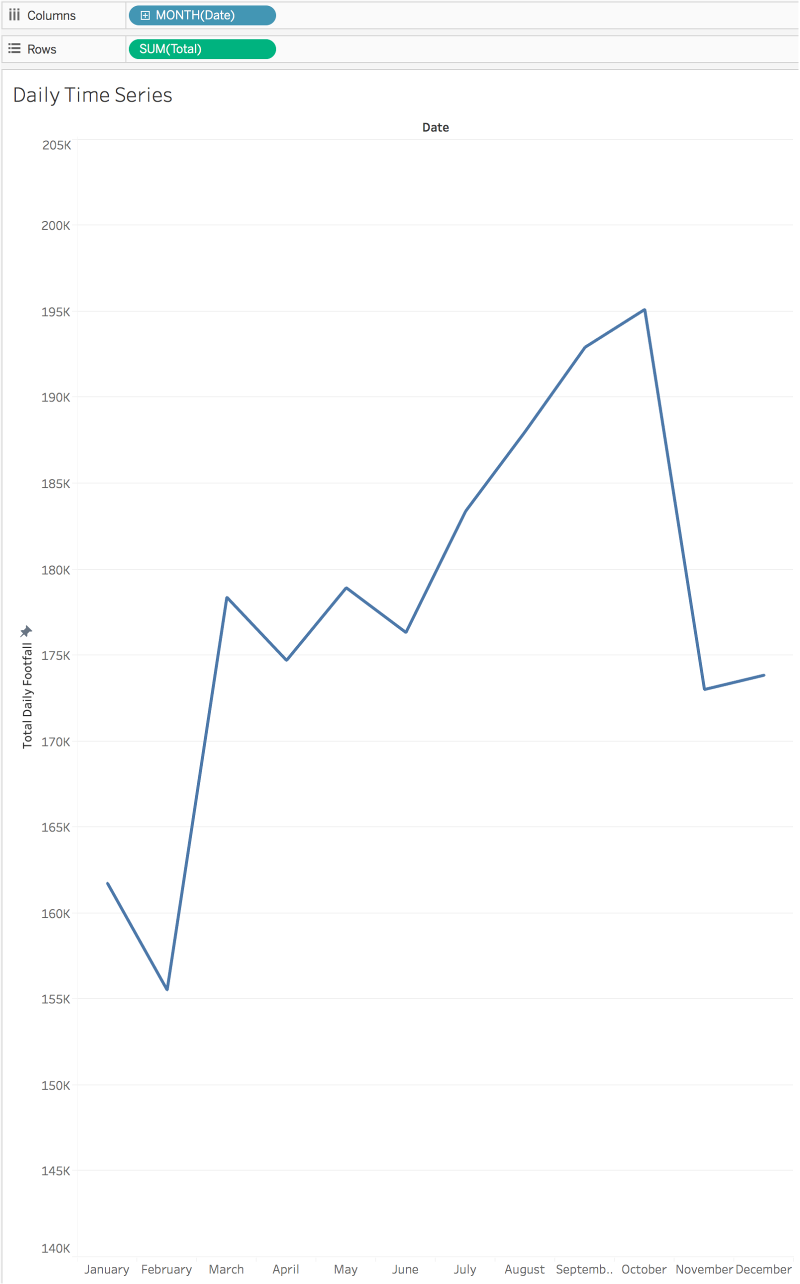


Figure 1.3: Time series plot of pedestrian flows by month in Capel Street, 2016

If time is recorded in a **date** format i.e. **day/month/year** Tableau will recognise the format and allow the Month variable to be collapsed into days. This is obtained by selecting the **plus sign** in the **Month** icon. This plots the foot flow by **days** as shown in Figure 1.4.

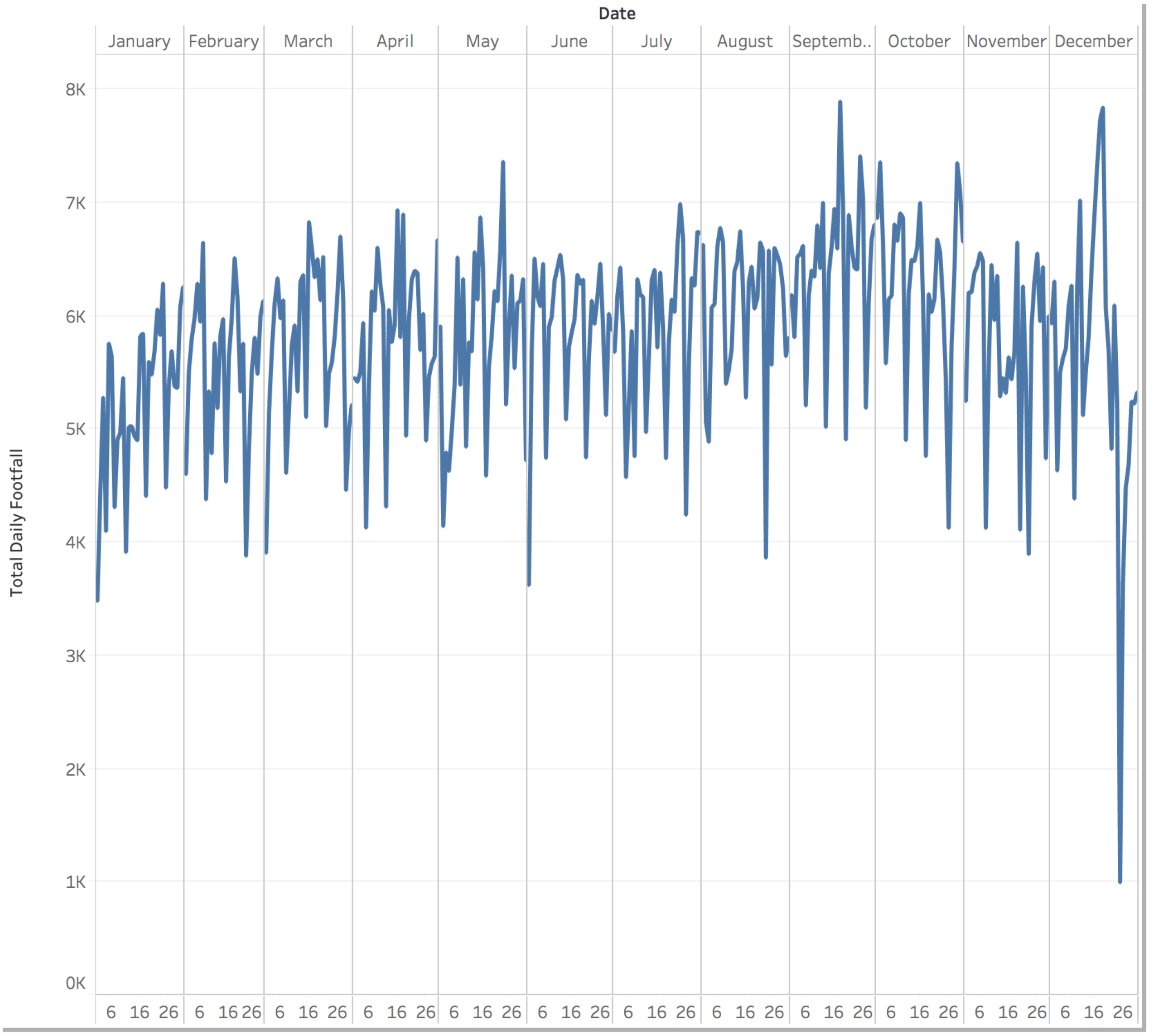


Figure 1.4: Time series plot of pedestrian flow in Capel Street by day of month

ID Continuous Data

i) Histogram

Histograms are one of the most common graphics used for representing 1D continuous data. A histogram is a plot of the number of observations in a particular interval. Histograms allow us to observe the general shape (or distribution) of the data and identify if any extreme points exist in the data set which may otherwise remain hidden amongst a possibly large amount of raw numbers. These extreme points (also known as outliers) may be the result of an error in recording a value, equipment malfunction, etc. Whatever the reason, extreme or unusual data points should be investigated and a decision taken whether to include them or not in the calculation of summary statistics. It is best to illustrate the construction of a histogram by way of an example.

Example

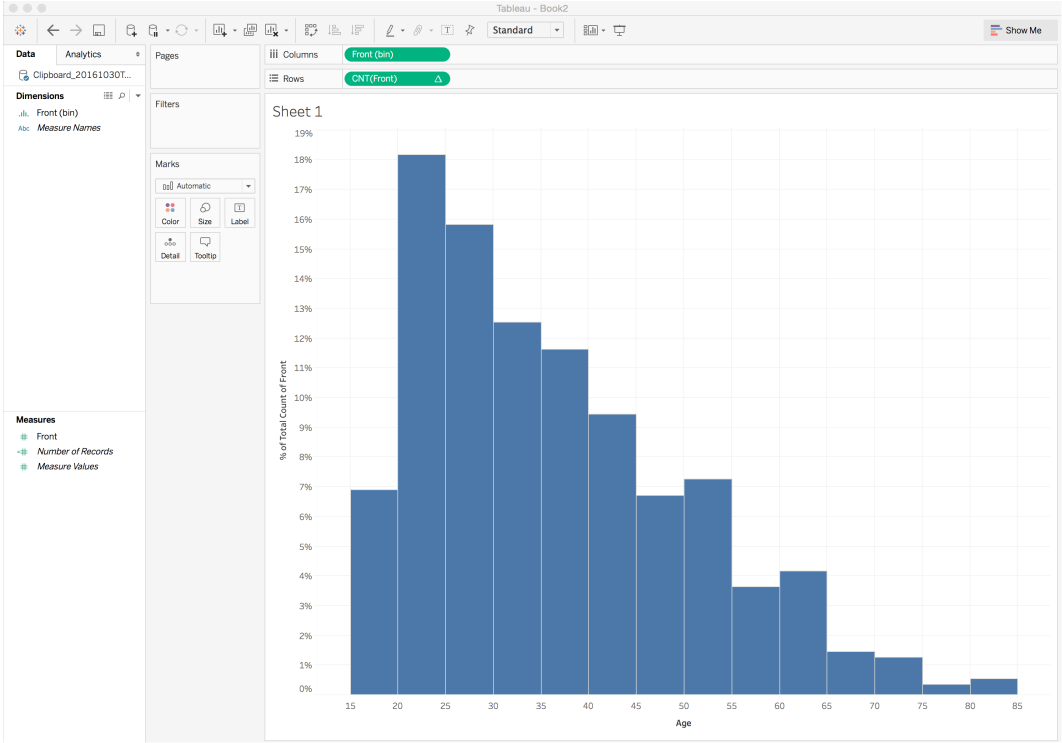
The age of 50 front seat passengers involved in road traffic accidents in Ireland between 2008 and 2013 is shown in Table 1.1. The complete file comprising 550 ages is provided in the Excel worksheet **Passenger Age** in *Data(2018).xlsx*

26	26	40	40	27	52	35	35	40	33
16	16	64	35	61	28	23	23	50	35
34	34	48	45	55	22	28	28	50	28
23	23	21	36	19	18	25	25	28	24
30	30	28	41	35	33	24	24	35	26

Table 1.1: Age of front seat passengers involved in road traffic accidents

It is hard to draw any conclusions about this data set in its current form. We need to rearrange this data so that a clearer picture can be obtained. We can achieve this by use of a histogram. To create a histogram we need to divide the data into intervals of some convenient length which is called the **bin size** and record the number of points in each bin. We can then plot the number of points falling in each bin as a histogram.

To create a histogram select the Tableau worksheet tab **Front Seat Passenger**. Select the variable *Front* and then **create Bins** from the menu obtained by clicking on the *Front* variable icon (or what is called the **pill** in Tableau). Enter 5 in the **Size of bins** panel. A new variable called **Front(bin)** is created which collects the data into cohorts each containing five year age grouping. This variable is placed on the **Dimensions** panel which is Tableau's location for discrete data.



Columns: Front (bin)

Rows: CNT(Front)

Figure 1.5: Histogram of age of front seat passengers

The Columns and Rows slots above the chart are called the **shelf** in Tableau. Place the discrete **Front(bin)** variable in the Column shelf and the continuous count of ages in the variable **Front** onto the rows shelf. Change the pill statistic from **Sum(Front)** to **CNT(Front)** by selecting **count** from the menu item **Measure(Count)** in the pills drop down menu. Finally, ensure that **Front(bin)** is changed from discrete to continuous by selecting the menu item **Convert to Continuous** from the dropdown menu.

The bin size of the histogram can be easily changed as can most of the graphical components e.g. colour, labels and so on. It is also possible to create a filter of **Front(bin)** by selecting **Show Filter** from the dropdown menu. A **slider** then appears on the canvas which allows the analyst to select and view the impact of different bin sizes on the histogram.

Exercise

The worksheet **Cancer1D** in the Excel file *Data(2018).xlsx* contains the

month since diagnosis of approximately 49,000 cancer patients diagnosed between 2009 and 2013 in Ireland. Using Tableau plot a histogram for this data set using a bin size of 6.

ii) Box and Whisker Plots

Box plots (also known as Box and Whisker plots) are graphical aids invented by the statistician John Tukey who was one of the pioneers of exploratory data analysis and author of the landmark text **Exploratory Data Analysis** [2]. Box plots allow five important components of a continuous dataset to be readily visualised. The five quantities, sometimes referred to as **five number summaries**, are the **maximum** and **minimum** values, the **median** (50th percentile) and the **upper** (75th percentile) and **lower quartiles** (25th percentile) of the data set. An example of a box plot is shown in Figure 1.6. The box part of the display extends from the lower quartile to the upper quartile of the data set. This distance is known as the **interquartile range (IQR)** and is used as a measure of the variability of the data set. The whiskers extend from the upper and lower quartiles to the maximum and minimum values of the data set, respectively.

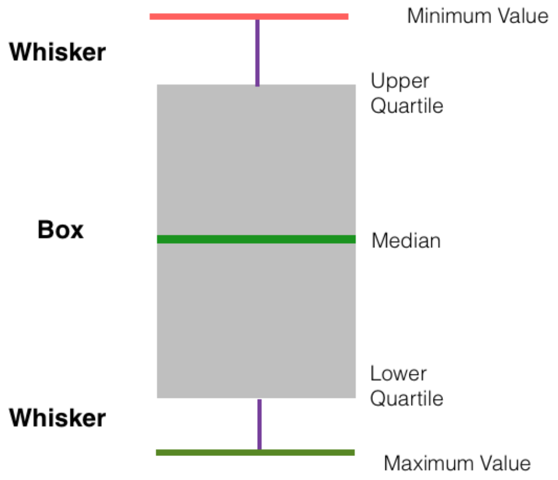


Figure 1.6: Box and whisker plot

Example

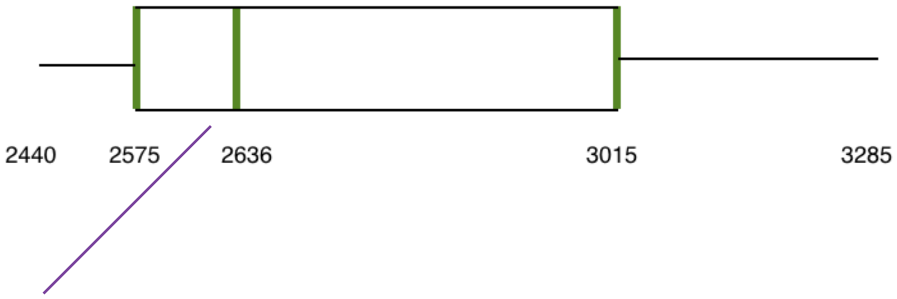
The following seven quotes (in €) for non-comprehensive insurance were obtained by a 20 year old with three years no claims bonus:

Quote(€)	2,543	3,285	2,840	2,609	2,440	3,191	2,636

To compute a box plot the data are first sorted or ranked from lowest to highest (or highest to lowest) as follows:

Rank	1	2	3	4	5	6	7
Data	2,440	2,543	2,609	2,636	2,840	3,191	3,285

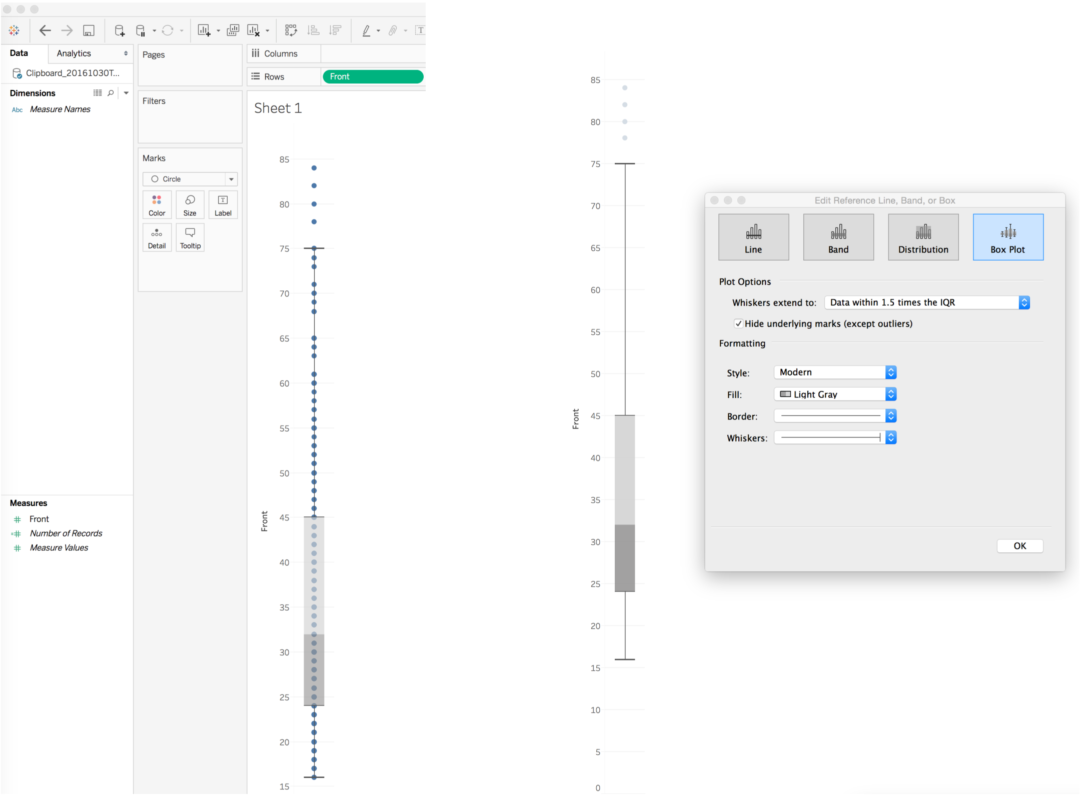
The maximum and minimum values are €3,285 and €2,440, respectively. The median is 2,636. The upper quartile is (€3,191+€2,840)/2 = €3,015 while the lower quartile is (€2,609+ €2,543)/2 = €2,576. The box plot for this data set illustrating the five figure summary is shown below.



If the **median** line in the box is not half-way between the upper and lower quartiles we say the data is not symmetric. We describe such a data set as **skewed**. In the above plot the region between the median and the upper quartile is substantially larger than the region between the median and the lower quartile. We describe such a data set as **positively skewed**. If the distance between the median and the lower quartile is larger than the distance between the median and the upper quartile then the data set can be described as **negatively skewed**. While box plots are very efficient at communicating the important characteristics of a data set they have do have a disadvantage in that they do not visualise the size of the data set i.e the number of data points.

Example

A box and whisker plot of the age of 550 front seat passengers involved in road traffic accidents is shown in Figure 1.7 (left). The plot was created in Tableau by selecting the worksheet **Front Seat Passenger** and placing the variable **Front** on the **Rows** shelf and selecting the box icon in the **show me** graph panel on the right of the canvas. Note that the menu item **Aggregated Measures** from the **Analysis menu** should be deselected.



Whiskers extend to: Data within 1.5 times the IQR

✓ Hide underlying marks (except outliers)

Style: Modern

Fill: Light Gray

Figure 1.7: Box and whisker plots using Tableau

The box plot can be easily modified by clicking in the box area of the plot and selecting **Edit**. This gives a variety of options for altering the appearance of the plot. For example, it is straightforward to remove the actual data points by selecting **hide underlying points** to create the plot above right which now does not show the individual quote data.

iii) Violin Plot

Violin plots are a less well known graphic that can be used to visualise continuous data. These plots visualise the complete distribution of a variable whereas box plots show the median, quartiles, range and degree of skewness. The shape of a violin plot also gives an indication of the size of the data set. For example, a violin plot of passenger age in road traffic accidents which was used in previous examples is shown in Figure 1.8. This plot suggests that passengers aged under 30 are the dominant cohort in this data set.

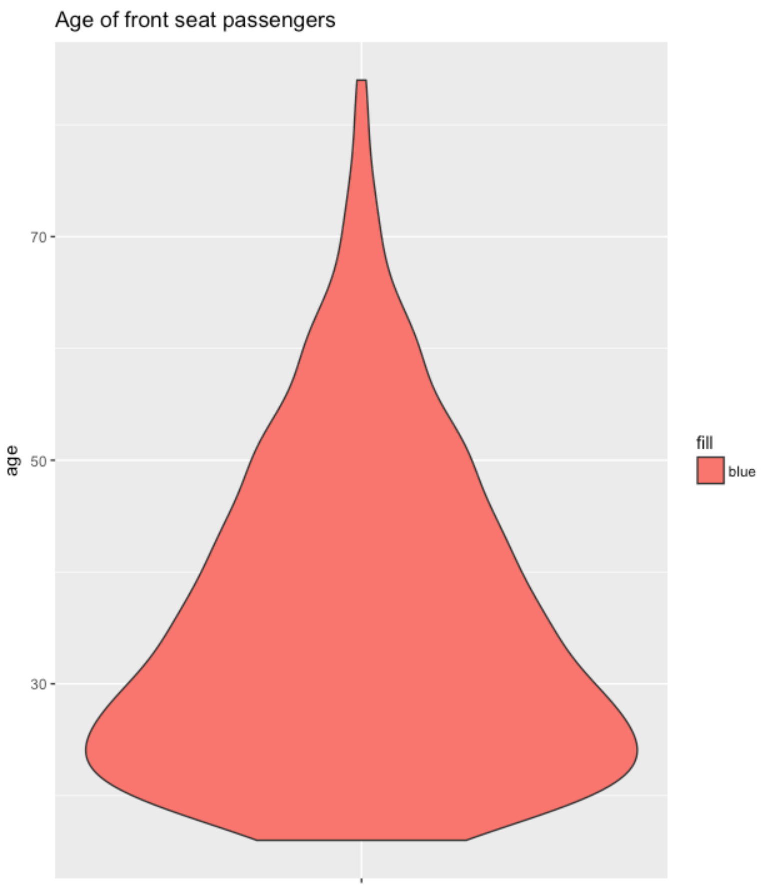
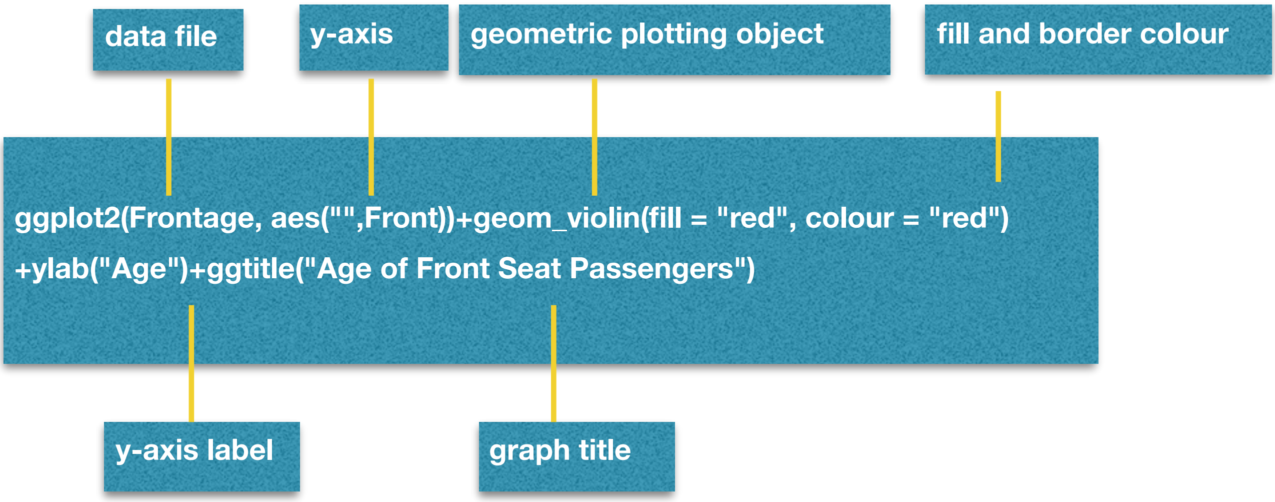


Figure1.8: Violin plot of the age of 550 front seat passengers involved in Irish road traffic accidents

Tableau does not calculate violin plots but we can use the R graphics library **ggplot2** to compute graphics plots using the following code:



data file

y-axis

geometric plotting object

fill and border colour

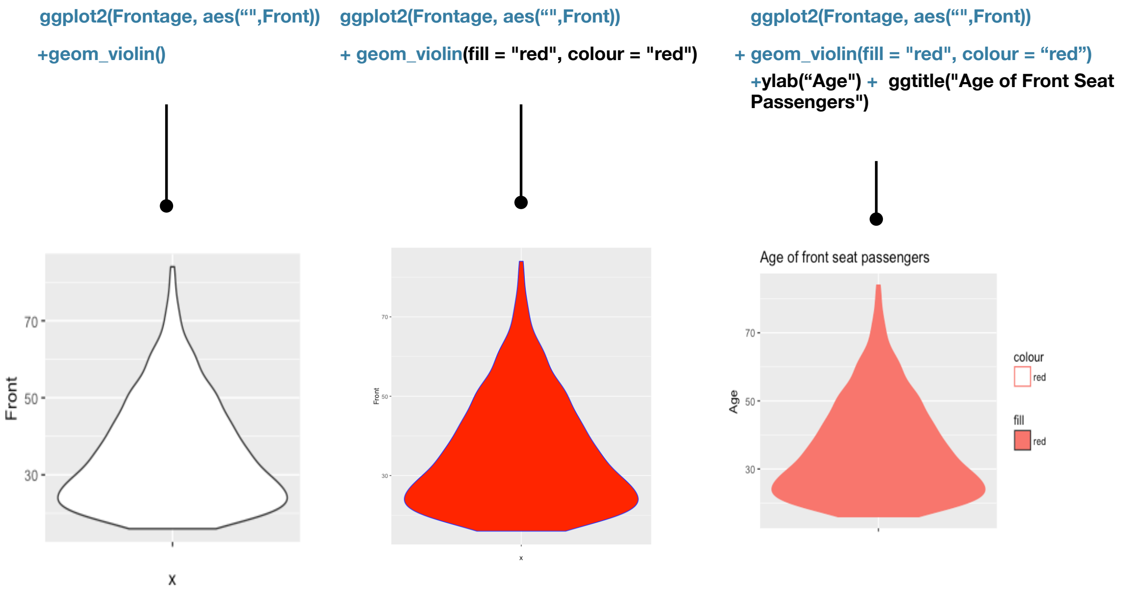
```
ggplot2(Frontage, aes("",Front))+geom_violin(fill = "red", colour = "red")
+ylab("Age")+ggtitle("Age of Front Seat Passengers")
```

y-axis label

graph title

The ggplot2 code draws the plot by successively adding graphical components. The first term **FrontAge** is the data file with **aes("",Front)** referring to the x-axis and y-axis variables. There is no x-axis variable hence the term **""**. The next component **geom_violin()** plots the graphic with no colour. Adding **geom_violin(fill = "red" colour = "blue")** included two more aesthetics - the colour of the fill of the plot and the colour of the border around the plot. The final two components added are **ylab("Age") +ggtitle("Age of front seat passengers")** which are the label for the y-axis and graph title, respectively.

There are many additional objects that can be added to this line of code as one of the advantages of ggplot2 is that it allows complete control of all the graphical elements in a plot. The three panels below show the cumulative effect of adding just three different graphical components (called layers) to the violin plot. Appendix 1 contains more details on available layers that can be added using ggplot2 while Appendix 4 provides an overview of some additional data visualisation packages available in R.



```
ggplot2(Frontage, aes("",Front))
+geom_violin()
```

```
ggplot2(Frontage, aes("",Front))
+ geom_violin(fill = "red", colour = "red")
```

```
ggplot2(Frontage, aes("",Front))
+ geom_violin(fill = "red", colour = "red")
+ylab("Age") +  ggtitle("Age of Front Seat
Passengers")
```

Figure 1.9: Successive graphical components being added to a violin plot

A listing of ggplot2 and vcd code to generate all the plots in this Chapter is provided in Appendix 2.

iv) Dot Plots

Dot plots are the most simple of ID continuous visualisations where each point is represented as a point as shown in Figure 1.10. This plot is based on the ages of 550 front seat passengers involved in road traffic accidents. To generate this plot in Tableau drag the variable **front** from the worksheet **Front Seat Passenger** to the bar chart icon and then deselect **Aggregated Measures** in the Analysis menu. To alter the shape and other characteristics of the dot plot change the **Marks** to Circle and then alter the Size, Colour etc as shown in Figure 1.10.

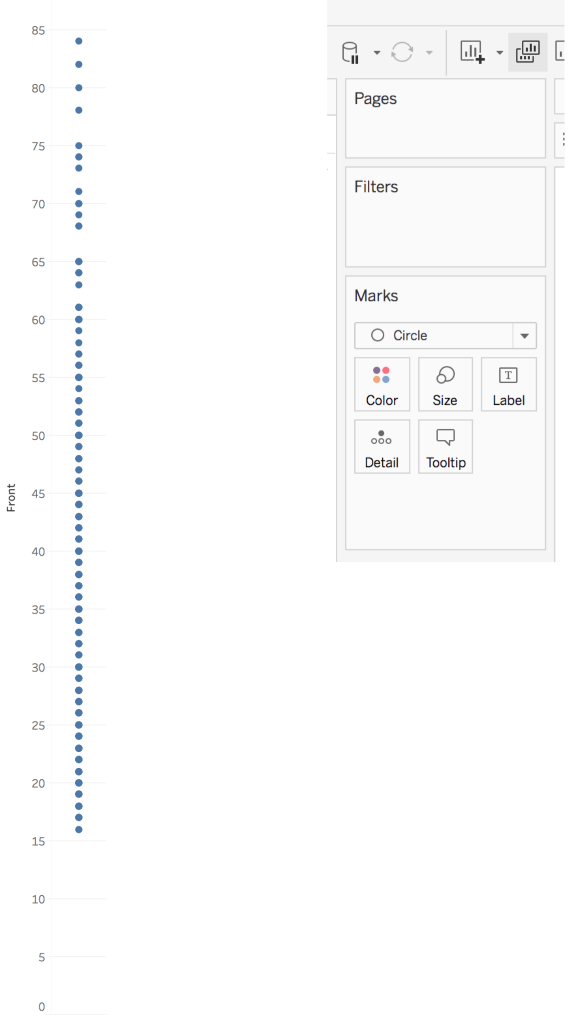


Figure 1.10: Dot plot using age of front seat passenger

v) Jitter Plots

Dot plots are not a particularly useful visualisation of large data sets as there can be considerable overplotting. For example, in the last example we might expect many passengers to have the same age given that our data set contains 550 ages. One useful remedy is to compute a jitter plot of the age of passengers.

Jitter plots add some **noise** to the data which allows points with the same value to be offset slightly so that no two points have the same exact value. An example of a Jitter plot using the passenger age data is shown in Figure 1.11. The larger density of points in the younger ages is clear from this plot.

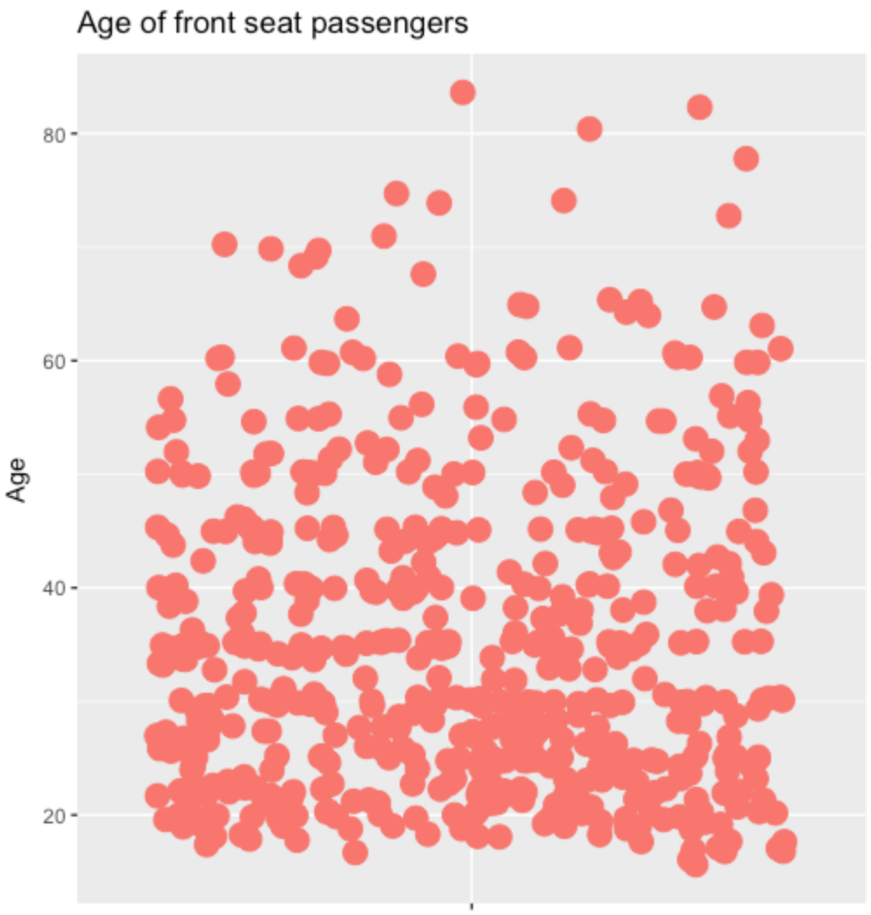


Figure 1.11 Jitter Plot of passenger age using ggplot2

Tableau does not directly provide for jitter plots. The R code for a jitter plot is generally the same as the violin plot except the geometric object **geom_jitter** replaces **geom_violin** as shown below.

```
ggplot2(FrontAge,aes("",Front)) + geom_jitter() + ylab("Age")
+ ggtitle("Age of Front Seat Passenger")
```

Exercises

1. The number of pedestrians killed in Ireland by month between 1997 and 2016 is provided in the table below:

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

ii) State giving a reason if the data set is discrete or continuous.

iii) Using **Tableau** create a bar and a Pareto chart for this data set. State giving a reason why which chart is the most effective in communicating the story of this data set.

iv) Write a short paragraph on your observations on pedestrian fatalities in Ireland based on the data contained in this table.

v) Spend a short time reflecting on your answer to part iv) and list any additional variables that you think may be useful for exploring further the topic of pedestrian fatal accidents in Ireland.

Month	Number Killed
January	175
February	188
March	153
April	119
May	140
June	112
July	120
August	114
September	152
October	175
November	193
December	227

2. The 194 grants awarded by Solas to educational training boards (ETB) for a range of services in 2016 are provided in the **Grant1D** tab in the Excel file *ExerciseData(2018).xls*. Using Tableau and/or R compute the following plots for this data file:

i) Histogram
ii) Box Plot
iii) Violin Plot
iv) Dot Plot
v) Jitter Plot

vi) Rank each of the above plots in order of their usefulness as a visualisation of the distribution of grants.
vii) Using the results of i) to v) summarise the principle features of the distribution of grants.

3. The age in years of 853 Dublin area patients diagnosed with Influenza during 2015 and 2016 is provided in the Excel worksheet **Influenza1D** in *ExerciseData(2018).xls*. Using Tableau and R compute the following plots for this data file:

i) Histogram
ii) Box Plot
iii) Violin Plot
iv) Dot Plot
v) Jitter Plot

vi) Rank each of the above plots in order of their usefulness as a visualisation of the distribution of age.
vii) Using the results of i) to v) summarise the principle features of the age distribution of influenza patients.

4. The Cancer diagnosis of just under 50,000 Irish patients are provided in the Excel worksheet **Cancer1D** in *ExerciseData(2018).xls.* Using Tableau compute the following plots for this data file:

i) Bar Chart
ii) Pareto Chart

5. Daily pedestrian flows from a high definition camera located on Henry Street in Dublin for each day of 2015 are recorded in the worksheet **Henry Street** in *ExerciseData(2018).xls.*

i) Using Tableau compute a time plot for this data set breaking down the date into months, quarters and days.

ii) Using i) summarise the principle features of the time plot computed in i).

6. The number of apprentice registrations in Ireland by year between 2009 and 2016 patients are provided in the file **Registrations** in *ExerciseData(2018).xls.*

i) Using Tableau compute a time plot for this data set

ii) Using i) summarise the principle feature(s) of the time plot computed in i).

7. The date of diagnoses of diabetes in the midland health region patients are provided in the file **Diabetes1D** in *ExerciseData(2018).xls*

i) Using Tableau compute a time plot for this data set breaking down the date into months, quarters and days.

ii) Using i) summarise the principle features of the time plot computed in i)

8. The number of comprehensively insured Irish vehicles between 2006 and 2015 is provided in the file **ExpComp** in *ExerciseData(2018).xls*

i) Using Tableau compute an appropriate plot for this data set

ii) Using i) summarise the principle features of the time plot computed in i)