<u>Data Visualisation – Programming Exercise Assignment →</u>

We have been provided with the Marketing Management Analytics (MMA) of a Portuguese banking institution. The data is timestamped from 2012 and was published on the UC Irvine Machine Learning Repository (Moro et al 2012). The data has been pre cleaned from its semi-colon separated values, to enable an easy load into R-Studio. The dataset contains 21 variables/columns, and they are:

1. age: (numeric)
2. job: type of job (Categorical: 'admin.' ,'blue-collar' ,'entrepreneur' ,'housemaid' ,'management', 'retired', 'self-employed', 'services', 'student', 'technician' ,'unemployed', 'unknown')
3. marital: marital status (categorical: 'divorced', 'married', 'single', 'unknown')
4. k: The education type variable (categorical: 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown')
5. default: has credit in default? (Categorical: 'no', 'yes', 'unknown')
6. housing: has housing loan? (Categorical: 'no', 'yes', 'unknown')
7. loan: has personal loan? (Categorical: 'no', 'yes', 'unknown')
8. y: Outcome Variable for taking long term deposit (Categorical: 'yes', 'no')
9. contact: contact communication type (categorical: 'cellular', 'telephone')
10. month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
11. day of week: last contact day of the week (categorical: 'mon', 'tue', 'wed', 'thu', 'fri')
12. campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
13. pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
14. previous: number of contacts performed before this campaign and for this client (numeric)
15. poutcome: outcome of the previous marketing campaign (categorical: 'failure', 'nonexistent', 'success') Social and economic context attributes
16. emp.var.rate: employment variation rate - quarterly indicator (numeric)
17. cons.price.idx: consumer price index (CPI/inflation) - monthly indicator (numeric)
18. cons.conf.idx: consumer confidence index - monthly indicator (numeric)
19. euribor3m: 3-month measure on interest rate - daily indicator (numeric)
20. nr.employed: number of employees/employment rate - quarterly indicator (numeric)

To start exploring the data, I'll be using the R-Studio package 'Pysch'. This package enables me to get a quick understanding of my dataset. The describe() command of 'Psych', is very similar to the basic summary() command. The statistical output is shown in Figure 1, and it shows us information like, number of observations, mean value, standard deviation, minimum/maximum, and more. From the 4100 observations, the average age was 40.12. String Values have been highlighted by an Asterix (e.g. job *, and poutcome*). String values have been ordered alphabetically and assigned a number starting from 1. For example, for the variable 'y', the string values are 'no' or 'yes', and being sorted alphabetically, means that 'no' values are 1 and 'yes' values are 2. This is why the minimum and maximum are respectively 1 and 2, in Figure 1.

```
> describe(MMD)
               vars    n    mean     sd  median trimmed    mad     min     max   range  skew kurtosis   se
age               1 4100   40.12  10.32   38.00   39.44  10.38   18.00   88.00   70.00  0.71     0.44 0.16
job*              2 4100    4.83   3.61    4.00    4.60   4.45    1.00   12.00   11.00  0.41    -1.42 0.06
marital*          3 4100    2.18   0.61    2.00    2.22   0.00    1.00    4.00    3.00 -0.03    -0.29 0.01
k*                4 4100    4.78   2.15    4.00    4.93   2.97    1.00    8.00    7.00 -0.28    -1.21 0.03
default*          5 4100    1.20   0.40    1.00    1.12   0.00    1.00    3.00    2.00  1.55     0.43 0.01
housing*          6 4100    2.08   0.98    3.00    2.10   0.00    1.00    3.00    2.00 -0.16    -1.95 0.02
loan*             7 4100    1.35   0.74    1.00    1.19   0.00    1.00    3.00    2.00  1.72     1.03 0.01
contact*          8 4100    1.36   0.48    1.00    1.32   0.00    1.00    2.00    1.00  0.60    -1.64 0.01
month*            9 4100    5.29   2.30    5.00    5.37   2.97    1.00   10.00    9.00 -0.31    -1.02 0.04
day_of_week*     10 4100    3.01   1.39    3.00    3.01   1.48    1.00    5.00    4.00  0.00    -1.26 0.02
duration         11 4100  256.75 254.40  181.00  210.53 136.40    0.00 3643.00 3643.00  3.30    20.85 3.97
campaign         12 4100    2.54   2.57    2.00    1.99   1.48    1.00   35.00   34.00  4.01    25.30 0.04
pdays            13 4100  960.24 192.35  999.00  999.00   0.00    0.00  999.00  999.00 -4.76    20.66 3.00
previous         14 4100    0.19   0.54    0.00    0.06   0.00    0.00    6.00    6.00  4.02    22.03 0.01
poutcome*        15 4100    1.92   0.37    2.00    1.99   0.00    1.00    3.00    2.00 -0.84     3.55 0.01
emp.var.rate     16 4100    0.09   1.56    1.10    0.27   0.44   -3.40    1.40    4.80 -0.73    -1.04 0.02
cons.price.idx   17 4100   93.58   0.58   93.75   93.58   0.56   92.20   94.77    2.57 -0.22    -0.82 0.01
cons.conf.idx    18 4100  -40.50   4.59  -41.80  -40.58   6.52  -50.80  -26.90   23.90  0.28    -0.32 0.07
euribor3m        19 4100    3.62   1.73    4.86    3.81   0.16    0.64    5.04    4.41 -0.71    -1.40 0.03
nr.employed      20 4100 5166.47  73.66 5191.00 5178.54  55.00 4963.60 5228.10  264.50 -1.08     0.06 1.15
y*               21 4100    1.11   0.31    1.00    1.01   0.00    1.00    2.00    1.00  2.49     4.21 0.00
```

Figure 1

The summary() command does a similar job to that of describe(). The string values need to be converted from character to factor format, to enable a count. From Figure 2 below, which shows the summary() command output, we can see that 451 people had subscribed to a term deposit as shown by output y.

```
> summary(MMD)
      age                 job           marital                        k       
 Min.   :18.00   admin.     :1005   divorced: 442   university.degree   :1259  
 1st Qu.:32.00   blue-collar: 883   married :2500   high.school         : 914  
 Median :38.00   technician : 688   single  :1147   basic.9y            : 572  
 Mean   :40.12   services   : 392   unknown :  11   professional.course: 533  
 3rd Qu.:47.00   management : 323                   basic.4y            : 428  
 Max.   :88.00   retired    : 165                   basic.6y            : 226  
                 (Other)    : 644                   (Other)             : 168  
    default          housing           loan             contact          month     
 no     :3300   no     :1832   no     :3334   cellular :2639   may    :1373  
 unknown: 799   unknown: 104   unknown: 104   telephone:1461   jul    : 707  
 yes    :   1   yes    :2164   yes    : 662                    aug    : 633  
                                                               jun    : 528  
                                                               nov    : 443  
                                                               apr    : 214  
                                                               (Other): 202  
 day_of_week    duration          campaign           pdays          previous     
 fri:762     Min.   :   0.0   Min.   : 1.000   999    :3940   Min.   :0.0000  
 mon:851     1st Qu.: 103.0   1st Qu.: 1.000   3      :  52   1st Qu.:0.0000  
 thu:856     Median : 181.0   Median : 2.000   6      :  42   Median :0.0000  
 tue:839     Mean   : 256.8   Mean   : 2.539   4      :  14   Mean   :0.1907  
 wed:792     3rd Qu.: 317.0   3rd Qu.: 3.000   7      :  10   3rd Qu.:0.0000  
             Max.   :3643.0   Max.   :35.000   10     :   8   Max.   :6.0000  
                                               (Other):  34                   
        poutcome     emp.var.rate      cons.price.idx  cons.conf.idx  
 failure    : 452   Min.   :-3.40000   Min.   :92.20   Min.   :-50.8  
 nonexistent:3506   1st Qu.:-1.80000   1st Qu.:93.08   1st Qu.:-42.7  
 success    : 142   Median : 1.10000   Median :93.75   Median :-41.8  
                    Mean   : 0.08517   Mean   :93.58   Mean   :-40.5  
                    3rd Qu.: 1.40000   3rd Qu.:93.99   3rd Qu.:-36.4  
                    Max.   : 1.40000   Max.   :94.77   Max.   :-26.9  

   euribor3m      nr.employed      y       
 Min.   :0.635   Min.   :4964   no :3649  
 1st Qu.:1.334   1st Qu.:5099   yes: 451  
 Median :4.857   Median :5191             
 Mean   :3.621   Mean   :5166             
 3rd Qu.:4.961   3rd Qu.:5228             
 Max.   :5.045   Max.   :5228             
```

Figure 2

When exploring new datasets, data is usually “Dirty” meaning that it needs to be cleaned, which is “often a long and difficult task” when manipulating large datasets (Dasu & Johnson 2003). As stated by Press (2016), “Data scientists spend 80% of their time preparing and manging data”. However, it’s a vital process when wanting to gather meaningful insight, especially when ‘unknown’ values create unnecessary noise.

Anscombe's Quartet provides a good reason for data cleansing. Created by English Statistician, Francis Anscombe, the purpose was to express the need to visualise data before deep analysing. The four datasets created all have similar simple descriptive statistics (Gupta 2022). Figure 3 shows that across the 4 datasets, the summary statistics like the mean, standard deviation (SD) and correlation coefficient (r).

Anscombe's Data											
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

Figure 3

However, when we come to visualising the data, we can see from Figure 4 that the scatter plots distribution is noticeable different across all 4 graphs, which the simple linear regressions hasn’t considered. We can see from the bottom two graphs, that outliers exist within those datasets, which have ultimately led to a misleading summary.

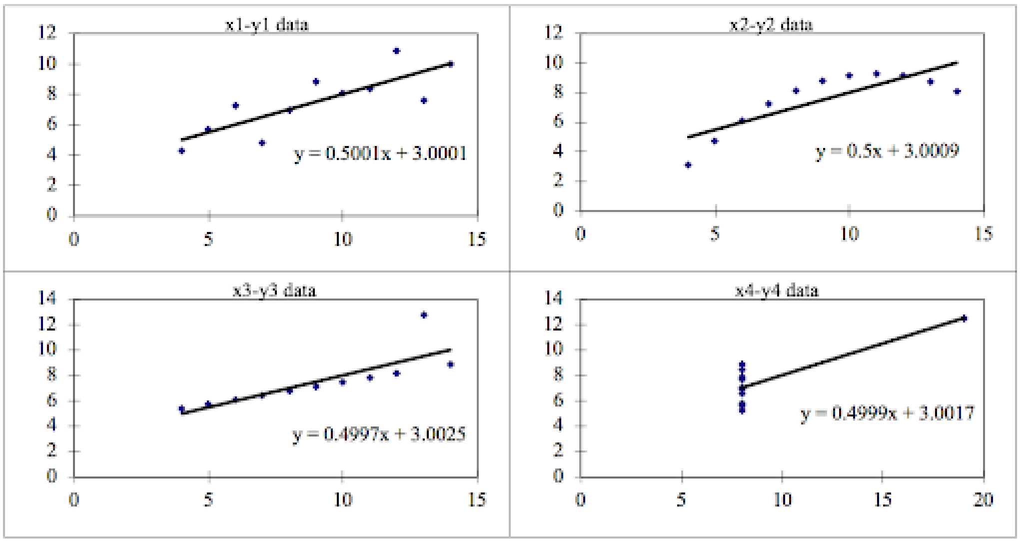


Figure 4

A good example skewness in our banking dataset is the columns 'pdays', which refers to the number of days passed since the client was last contacted. The number '999' was used to represent if a client was not previously contacted. Using R-studio, I've plotted a bar chart to represent the count of 'pdays' and this is displayed in Figure 5.

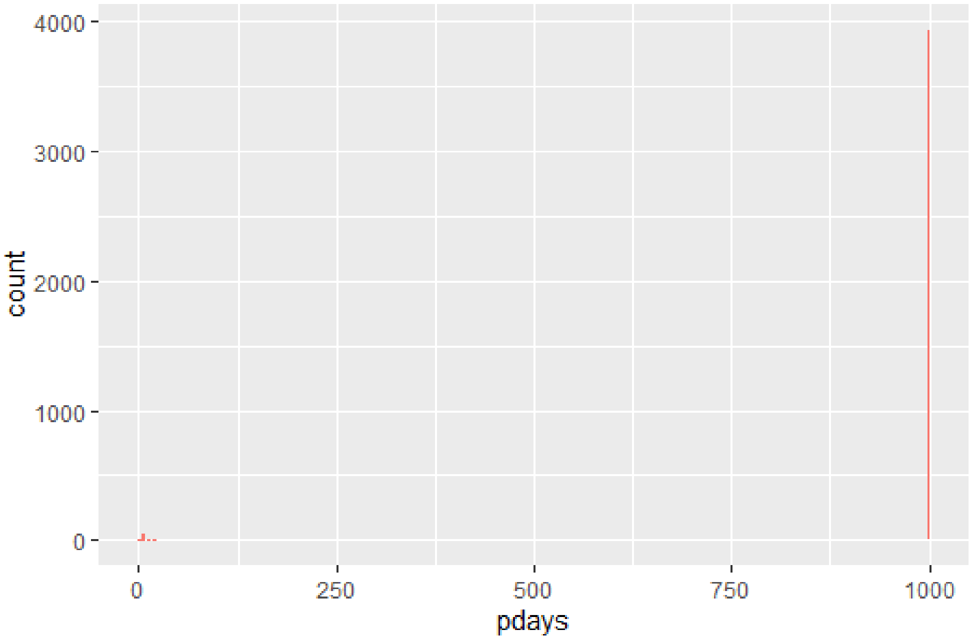


Figure 5

As Figure 5 shows, we have around 3900 entries of '999' representing those who haven't been contacted before. With those who have been contacted before towards the left of the bar chart. I've constructed the bar chart again but by filtering out the '999' input. Figure 6 shows the distribution now, and only 160 observations weren't assigned the value '999'. Now computing a summary_() command on the 'pdays' column, we get a mean value of 5.862, which is significantly different to our original mean values of 20.44.

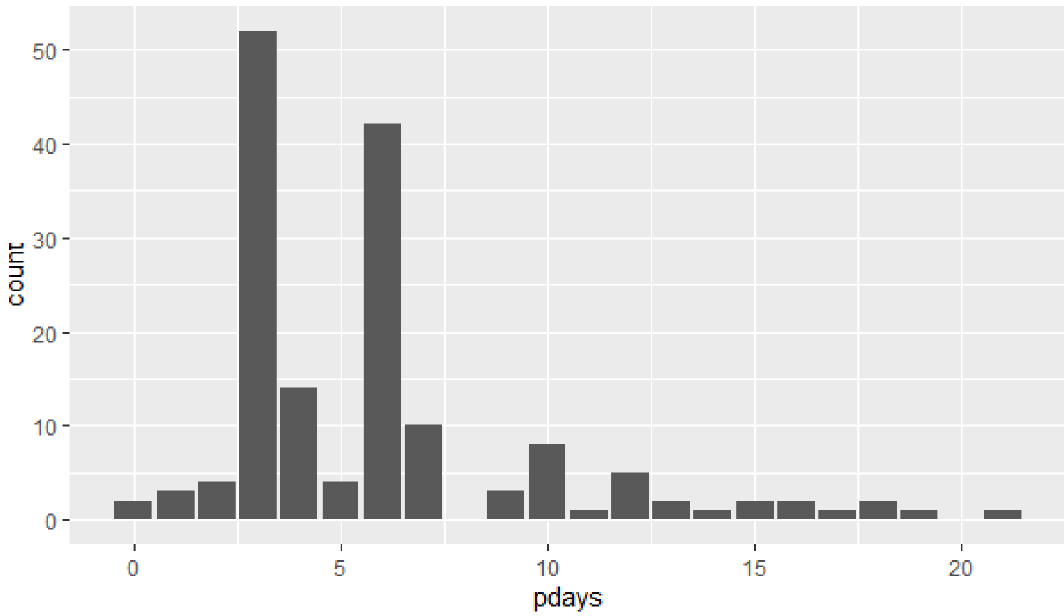


Figure 6

We also have skewness in the variable ‘Age’ as shown below in **Figure 7**. The skewness is to the right of the graph and there are minimal entries those over the age of 65. These should be considered as outliers. Omitting these values, give you the output in Figure 8 with an improved distribution. Omitting outliers is important, especially when it comes to identifying correlation between variables which I’ll be discussing later on.

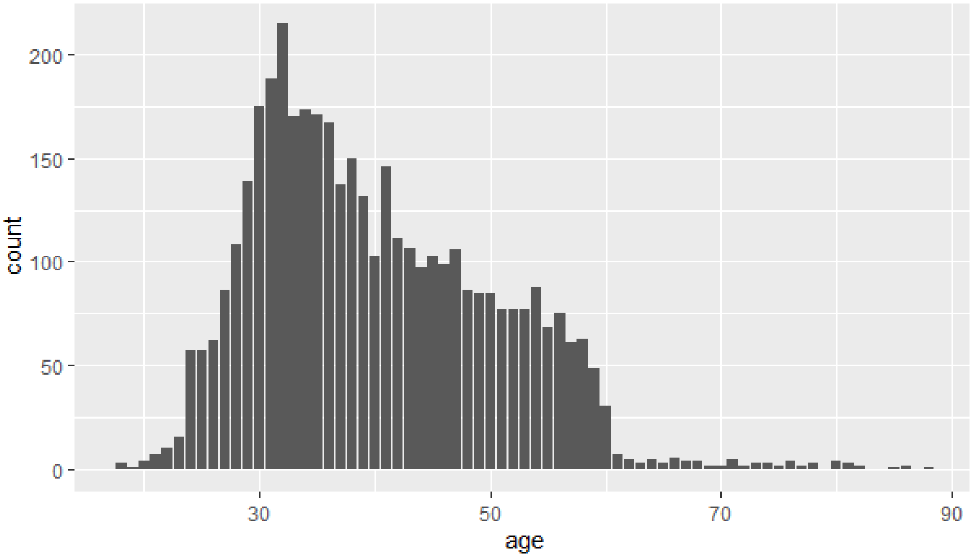


Figure 7

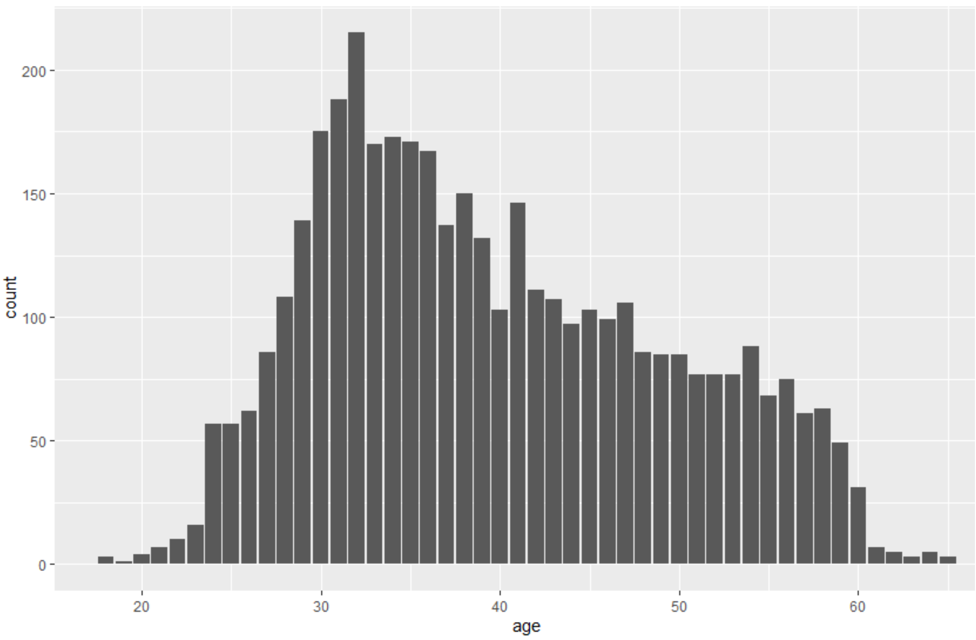


Figure 8

The use of visualisations has been useful in identifying outliers and skewed distribution, helping to avoid bias results. The mutate() command from the tidyverse package has been useful in altering values the dataset. Figure 9 shows the mutations that I’ve carried out. We can see the adjustments made previously to variable ‘pdays’ and ‘age’. I’ve altered values such as ‘unknown’ in the dataset to NA/blank, for variables like ‘marital’ and ‘housing’. Also, ‘duration’ had a very similar distribution to ‘age’ and to avoid skewness in data, I’ve limited the values between a range of 30 to 600 (duration of call in seconds. Although, as mentioned by Moro et al (2012) who released the marketing data, the duration is not known before a call and ‘duration’ should “only be included for benchmark purposes”. Lastly, we have added new column ‘y_numeric’ which replaces the string values of “yes” and “no”, with binary values of 1 and 0 for the outcome variable y. Left 'y' to be a factor variable which is useful for colouring my visualisations.

```
#Transforming our Data with Mutate Function
MMD <- MMD%>%
  mutate(pdays=ifelse(pdays==999,NA,pdays),
         age=ifelse(age>65,NA,age),
         y_numeric=ifelse(y=="no",0,1),
         marital=ifelse(marital=='unknown',NA,marital),
         default=ifelse(default=='unknown',NA,default),
         housing=ifelse(housing=='unknown',NA,housing),
         loan=ifelse(loan=='unknown',NA,loan),
         duration=ifelse(duration<30,NA,ifelse(duration>600,NA,duration)))
```

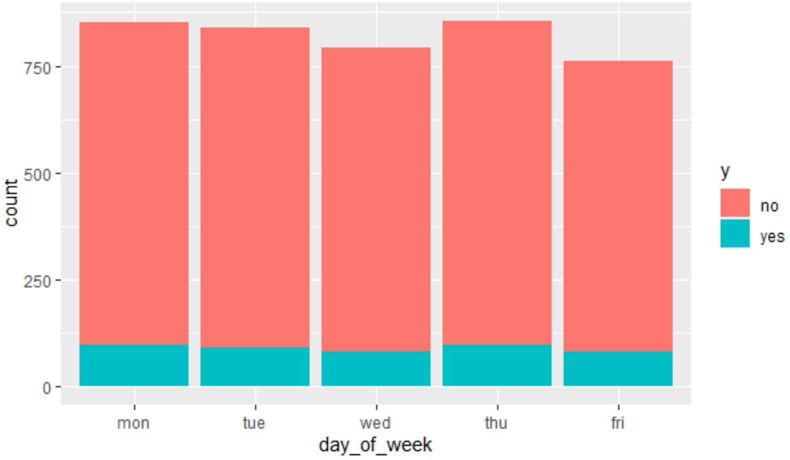
Figure 9

As a result of cleaning the data further, by mutating columns and assigning NA values, we get a summary table as shown in Figure 10. The main outcome is that out of the 4100 observations, only 11% resulted in a successful long-term deposit.

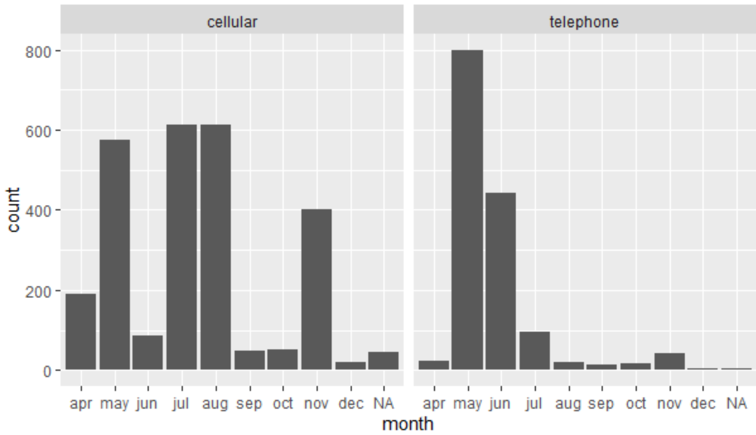
```
> summary(MMD)
      age                  job             marital                       k
 Min.   :18.00   admin.      :1005   divorced: 442   university.degree  :1259
 1st Qu.:32.00   blue-collar: 883   married :2500   high.school         : 914
 Median :38.00   technician : 688   single  :1147   basic.9y            : 572
 Mean   :39.66   services   : 392   NA's    :  11   professional.course: 533
 3rd Qu.:47.00   management : 323                   basic.4y            : 428
 Max.   :65.00   retired    : 165                   basic.6y            : 226
 NA's   :55      (Other)    : 644                   (Other)             : 168
 default      housing      loan          contact          month       day_of_week
 no  :3300   no  :1832   no  :3334   cellular :2639   may    :1373   mon:851
 yes :   1   yes :2164   yes : 662   telephone:1461   jul    : 707   tue:839
 NA's: 799   NA's: 104   NA's: 104                    aug    : 633   wed:792
                                                      jun    : 528   thu:856
                                                      nov    : 443   fri:762
                                                      (Other): 368
                                                      NA's   :  48
    duration        campaign          pdays          previous
 Min.   : 30.0   Min.   : 1.000   Min.   : 0.000   Min.   :0.0000
 1st Qu.:107.0   1st Qu.: 1.000   1st Qu.: 3.000   1st Qu.:0.0000
 Median :173.0   Median : 2.000   Median : 6.000   Median :0.0000
 Mean   :206.6   Mean   : 2.539   Mean   : 5.862   Mean   :0.1907
 3rd Qu.:274.0   3rd Qu.: 3.000   3rd Qu.: 6.000   3rd Qu.:0.0000
 Max.   :600.0   Max.   :35.000   Max.   :21.000   Max.   :6.0000
 NA's   :515                      NA's   :3940
        poutcome      emp.var.rate       cons.price.idx  cons.conf.idx
 failure    : 452   Min.   :-3.40000   Min.   :92.20   Min.   :-50.8
 nonexistent:3506   1st Qu.:-1.80000   1st Qu.:93.08   1st Qu.:-42.7
 success    : 142   Median : 1.10000   Median :93.75   Median :-41.8
                    Mean   : 0.08517   Mean   :93.58   Mean   :-40.5
                    3rd Qu.: 1.40000   3rd Qu.:93.99   3rd Qu.:-36.4
                    Max.   : 1.40000   Max.   :94.77   Max.   :-26.9

   euribor3m      nr.employed      y            index          y_numeric
 Min.   :0.635   Min.   :4964   no :3649   Min.   :   1   Min.   :0.00
 1st Qu.:1.334   1st Qu.:5099   yes: 451   1st Qu.:1026   1st Qu.:0.00
 Median :4.857   Median :5191              Median :2050   Median :0.00
 Mean   :3.621   Mean   :5166              Mean   :2050   Mean   :0.11
 3rd Qu.:4.961   3rd Qu.:5228              3rd Qu.:3075   3rd Qu.:0.00
 Max.   :5.045   Max.   :5228              Max.   :4100   Max.   :1.00
```

Figure 10

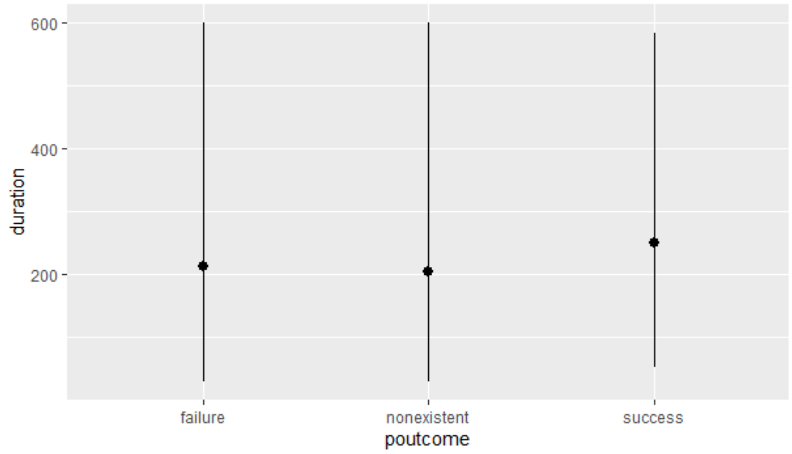
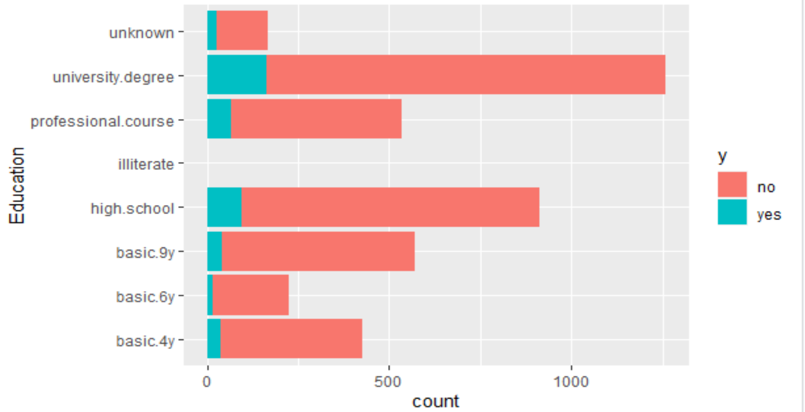
In the following table, I've used a mixture of visualisations to describe the dataset.

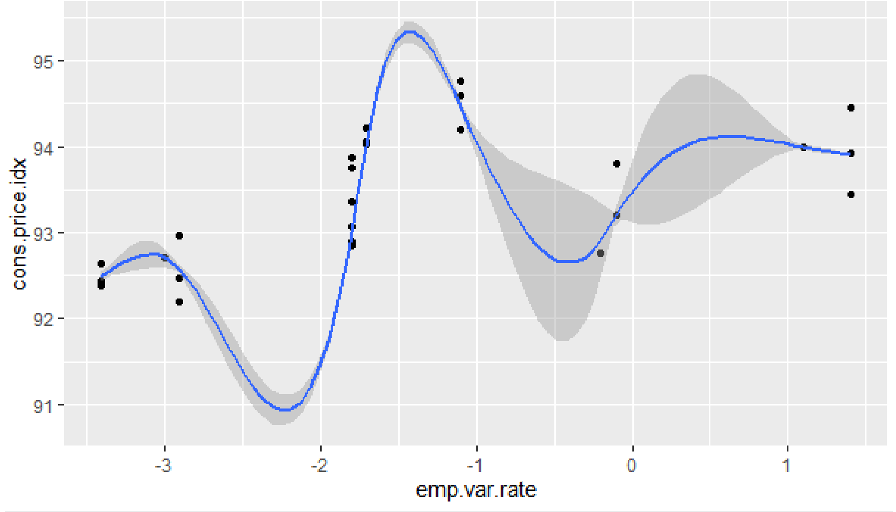
<table>
<tr><th>Variable and Explanation</th><th>Visualisation</th></tr>
<tr><td>Day_of_week & Y – The number of calls made during the week remained very constant, with minor dips on Wednesday and Thursday. The outcome of a customer opening a deposit was influenced by the day of the week, as shown by the small distribution (green coloured bars).</td><td>


</td></tr>
<tr><td>Month & contact – The volume of calls fluctuates between each month significantly, unlike day_of_week. This is likely due to season demand.

Method of contact by telephone was higher in the months of May-June, but contact by cellular dramatically increased. Cellular and telephone acting as substitutes.</td><td>


</td></tr>
</table>

Duration, poutcome & y – The graph shows the duration of calls, and splits it by whether they've been contacted previously, with a success and failure flag also. The star is the mean outcome, and it shows that previously 'successful' contacts resulted in longer phone calls and more success. However, a big issue here is the range. The minimum/maximum bars are very large, meaning that the statistical significance is quite small.	
K (education) and Y – The success of opening a deposit is higher for those with a 'university degree'. However, due to the imbalance of 'yes' to 'no' (11% to 89%). It makes it hard to draw causation from this correlation.	

Consumer Price index & employment Variance – A positive correlation between the two variables, showing cyclical employment. As consumer price rises, the employment variance changes from negative to positive.	

In Figure 11, I've carried out some basic visualisation of the variables 'job', 'marital', 'default', 'housing', 'campaign', and 'previous'. Housing is the only variable that looks to have an even split of distribution compared to the other 5 variables. For example, for 'default', only 1 person defaulted, with a majority of nearly 3200 saying they haven't defaulted. This same individual who did default, didn't successfully open a long-term deposit. It would be statistically inaccurate to suppose that having defaulted before, you're guarantee to not deposit. The sample size is too small, especially with around 750 people answering 'unknown' to this question.

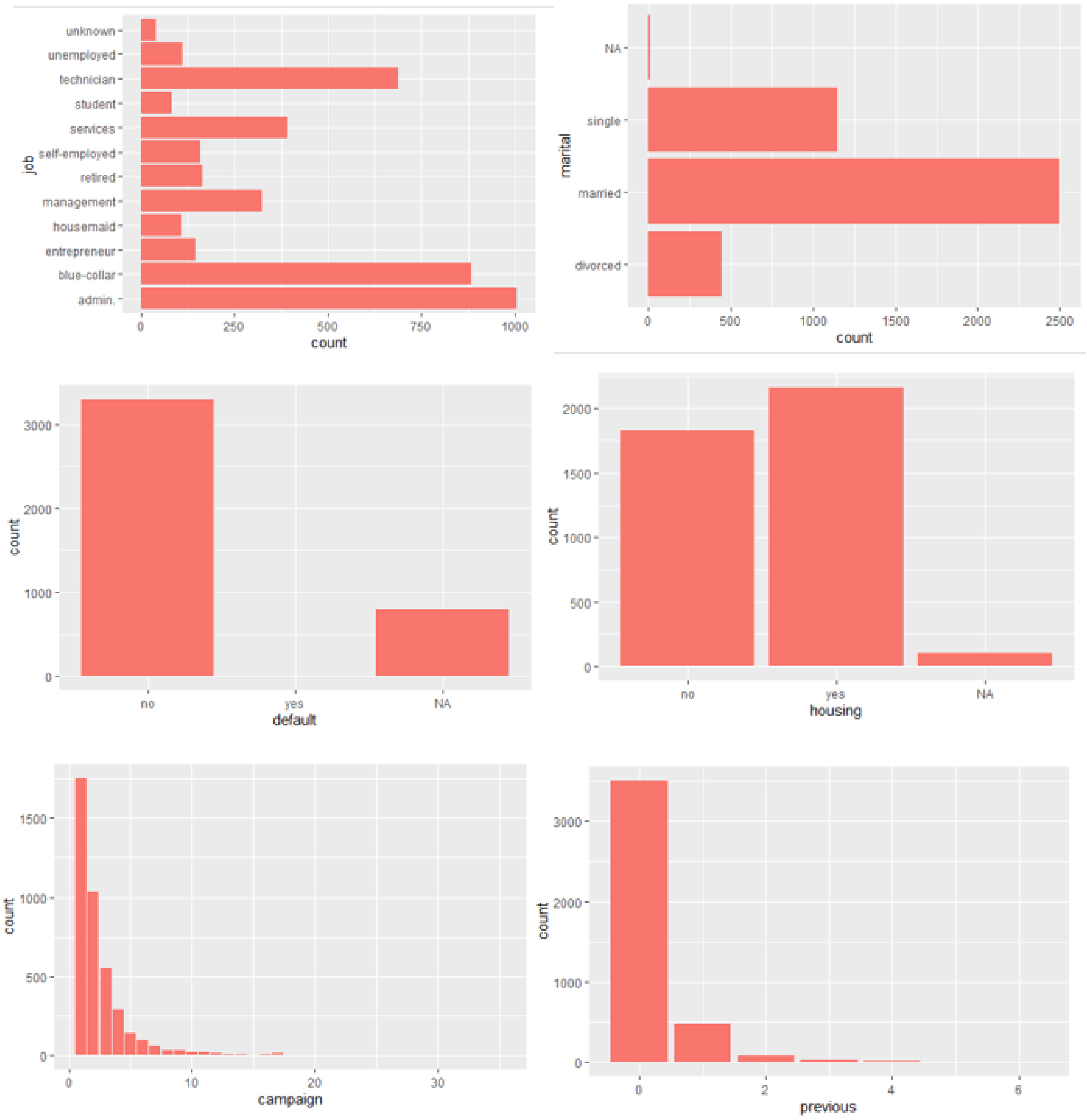


Figure 11

To help us understand the relationship between variables I computed a correlation matrix between variables in the dataset. I've only used numeric variables, as show in Figure 12, as it's difficult to say string values are better than each other according to a scale (e.g. can't calculate the numerical difference between 'self employed' and 'retired' for job factor). I've stored the numeric values into a different variable (MMD_numerical) in my R Script, and then used the cor() function, along with ggplot. I did have to transform the dataset from wide to long, with the help of the 'reshape2' library to melt the transformation.

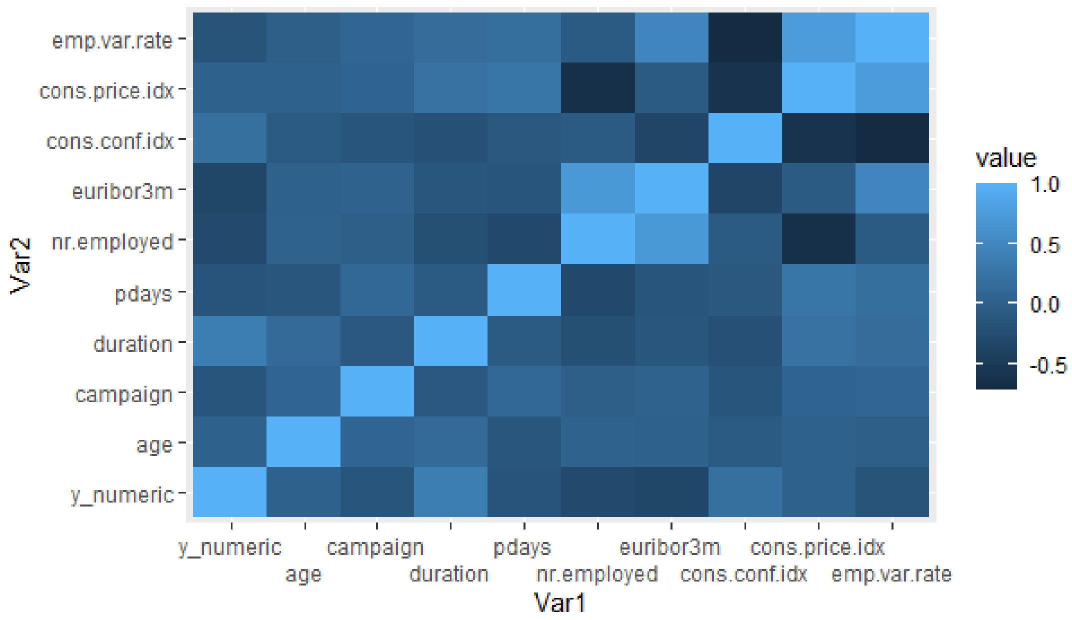


Figure 12

The key in Figure 12, explains the stronger the correlation, the lighter the blue colour is (e.g. value towards 1.0). We can see that matrix of matching variables have a perfect correlation of 1 which makes sense. However, we also have variables producing a strong correlation of more than 0.7. For example, 'emp.var.rate' and 'cons.price.idx', which I discussed earlier with the regression and scatter plot. A strong correlation exists between 'euribor3m' (interest rates) and 'nr.employed' (number of employees). This strong correlation makes sense because when interest rates rise, the cost to businesses rise, resulting in firms reducing the size of workforce. Consequently, employment falls.

The matrix shows weak correlation/no correlation between variables like age, and y (outcome). The weak correlation suggests that age doesn't influence the 'y' outcome of opening a long-term deposit. But caution should be placed on the dispersion shown in Figure 10, as only 11% of people opened a long-term deposit. A large difference to those who didn't.

Binomial regressions can be used to show relationships as well. From the matrix in Figure 12, consumer confidence had a weak correlation with influencing the deposit of long-term deposits. Using ggplot again, we can model this relationship by a regression to get the following output. The weak correlation is seen by the slight elevation of the blue regression line of Figure 13.

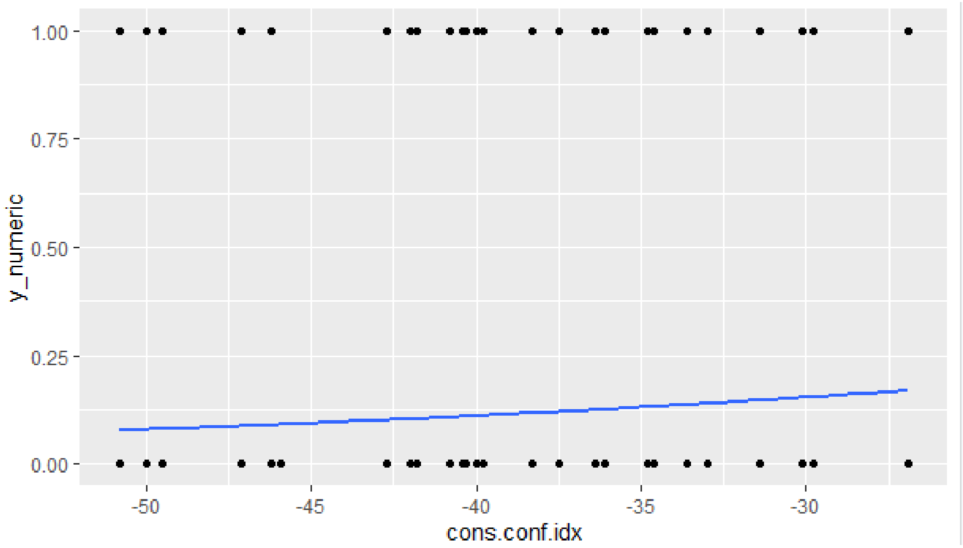


Figure 13

To conclude, the use of visualisation provides Data Scientists the opportunity to understand their dataset in more detail, whilst identifying trends and outliers. With the banking data set, I identified variables/columns with skewed distribution, and adjusting outliers that could lead to biasness. To gain a better understanding of what influences an individual to open a long-term deposit (y= "yes"), more data is needed. Only 11% of the 4100 respondents opened a long-term deposit, meaning the data was heavily outweighed by those who didn't.

References

Gupta, S (2022) Anscombe's Quartet: What Is It and Why Do We Care? Available at https://builtin.com/data-science/anscombes-quartet [Accessed: 25th September 2023]

Dasu, T and Johnson, T (2003) Exploratory data mining and data cleaning. John Wiley & Sons.

Press, G (2016) Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says. Forbes. Available at: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/?sh=3cb0355c6f63 [Accessed: 26th September 2023]