Assignment 2: Data Visualization

Due beginning of class: Monday September 25

1. Visual fractions can be used to provide a sense of the size of that fraction (provided it is not too small). In the document `VisualFractions.pdf`, you will find an introduction to some graphical primitives that will allow you to draw some visual fractions using circles (the `VisualFractions.Rmd` contains the code in an Rmarkdown file which you might find helpful).

   (a) (10 marks) Read that document and complete the definition of the function `visualFraction`$(\cdots)$.

   (b) Show the results of your program on $\frac{3}{100}$ and also on $\frac{37}{1000}$. In each case, show the results **both when** `random` $=$ `FALSE` **and when** `random` $=$ `TRUE`.

   (c) (2 marks) Explain why the case `random` $=$ `TRUE` might be of interest.

2. In class, a time scale was used to indicate the average time it would take to first win Lotto 649, purchasing 1 ticket per weekly draw. Assume that a winning ticket is one which matches the 6 numbers drawn from 1 to 49.

   (a) (1 mark) Suppose p is the probability of winning the grand prize. Write down the value for p for Lotto 649.

   (b) (1 mark) Write down the probability of winning (**for the first time**) on the nth draw (i.e. losing on the first $n-1$ draws).

   (c) (1 mark) Determine the expected number of draws you must play (1 ticket each draw) before winning for the first time.

   (d) (1 mark) Show how the average time to win Lotto 649 when playing 1 ticket per weekly 649 draw turns into the long wait given for the Homo sapiens example (as described in the slides)

3. Colour blindness: Consider the following oppositional colour palette:

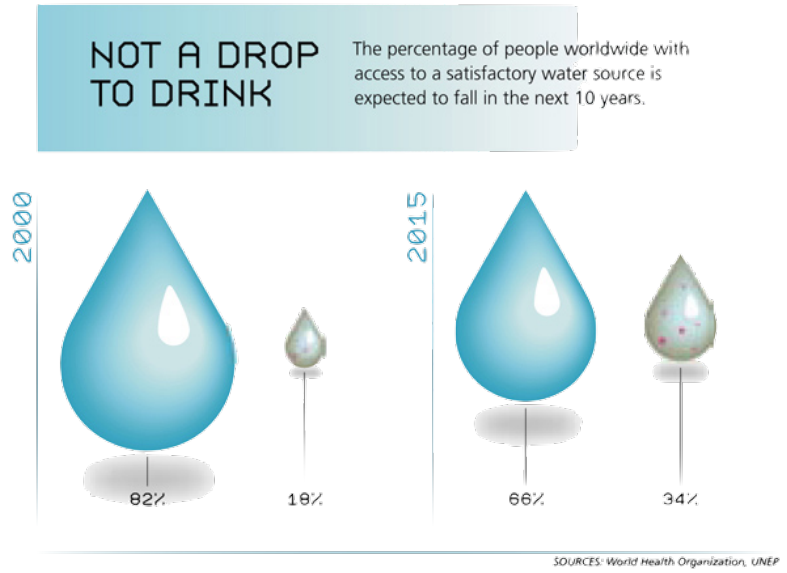
Figure 1: Oppositional colour palettes

   Each of these two palettes uses opposing colours, either yellow to blue or green to red, to provide a palette that changes continuously from one extreme to the other. The image itself is available on the course website where this assignment was located. Up load this image to the website http://www.color-blindness.com/coblis-color-blindness-simulator/ and explore how well each of the two palettes compare for the various sorts of colour blindness.

   a. (3 marks) How well do each of these palettes work for those unable to see colour at all? Why?
   b. (3 marks) Which oppositional colour pair seems best over all – yellow-blue or green-red? Explain your answer.
   c. (4 marks) Given what we discussed in class about the photo-receptors in the human retina, what characteristic of the photo-receptors might explain your choice in part (b)?

4. Visual representations.

   (a) The following diagram was produced by the World Health Organizationa few years back.



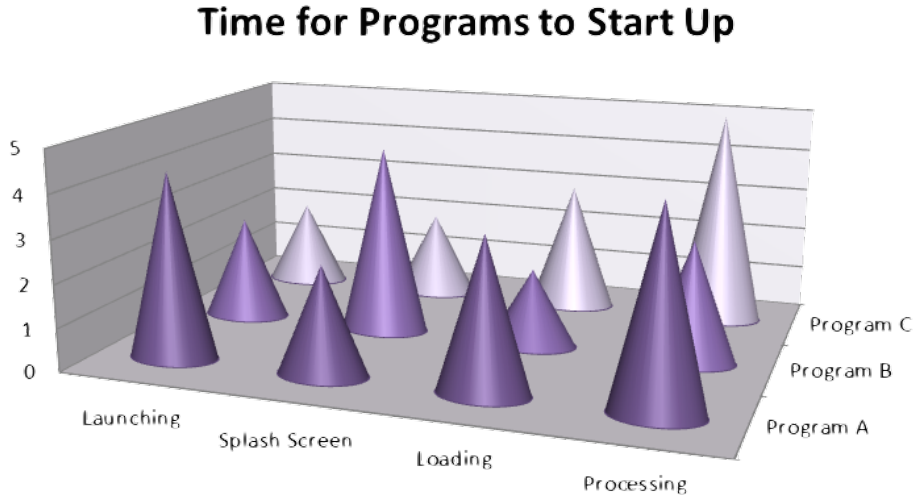
i. (4 marks) Suppose the percentages were not actually displayed. What visual features of the display are available to the reader to decode the numbers presented? Which of these correspond to the percentages?

ii. (4 marks) From Cleveland and McGill's ordering of elementary tasks identify which of these tasks are used where in this diagram. Which of these elementary tasks are most likely to be used in decoding the numbers by the reader? Comment on the likely accuracy of this decoding.

iii. (4 marks) Create a table representing the same numerical information. Give the pros and cons of the table compared to the original diagram. (See 'Tables.Rmd' in the folder for this assignment)

(b) The following diagram was produced with considerable enthusiasm from a site called premiermicrosoft.wordpress.com https://premiermicrosoft.wordpress.com/2012/02/12/how-to-make-a-graphchart-in-microsoft-word/index.htm (not to be confused with Microsoft's https://premier.microsoft.com) where it is called a "cone chart".

To quote the author: "Now, jazz up some effects for your graph. You can customize your current graph, by clicking Design in Chart Tools." …sigh.



i. (5 marks) Write out all of the values that appear in the diagram which are categorical and say how they are encoded in the diagram.

ii. (3 marks) Consider how the values of 'Time for Programs to Start Up' encoded. From Cleveland and McGill's ordering of elementary tasks identify which of these tasks are used in the encoding of

these values. Which elementary tasks are most likely to be used in decoding the values by the reader? Comment on the likely accuracy of this decoding.

iii. (2 marks) Critically assess the contribution of the scales appearing at the left and back of the plot.

5. Tables are an important way to display symbolic numbers. In the document `Tables.pdf` (and perhaps more importantly its source file `Tables.Rmd`) you will find some examples of manipulating tables using the `knitr R` package. Please consult those files (downloading them and opening them from RStudio) and familiarize yourself with the material found there. It will be very helpful to you in undertaking the analysis in this questions.

   Here you are going to work on some Statistics Canada data to produce an interesting table. The data are on aboriginal populations taken from a Statistics Canada website http://www.statcan.gc.ca/tables-tableaux/sum-som/l01/cst01/demo60a-eng.htm.

   This data is available in a "Comma Separated Values" or "csv" file named "aboriginal.csv". This should appear in the same location as this file of questions. Download both the ".Rmd" file of questions and the"csv" file of data and place them in the same directory on your machine.

   In `RStudio` check what the current working directory is using the `getwd()` command in `R`. For example in my case the current working directory is

```
getwd()
```

```
## [1] "/Users/rwoldford/Documents/Admin/courses/Data Visualization/Assignments/Fall17/Assignment 2"
```

   As it turns out, this is where I have saved the data file. Had a different string have been returned for the directory, I would have needed to set the working directory to the appropriate place using `setwd()` as follows:

```
setwd("/Users/rwoldford/Documents/Admin/courses/Data\ Visualization/Assignments/Fall17/Assignment\ 2")
```

   Now that the directory is set to the data location, we can read the csv file into a data frame and show the results.

```
data <- read.csv("aboriginal.csv")
# The first few columns look like
kable( data[, 1:4])
```

	CA	NL	PE	NS
Aboriginal.identity.population	3.7539895	4.6852840	1.2890727	2.6769204
North.American.Indian	2.2343213	1.5511076	0.9127827	1.6875394
Métis	1.2476541	1.2924232	0.2868746	0.8504136
Inuit	0.1615824	0.9418509	0.0223539	0.0359876
Non.aboriginal	96.2459945	95.3157148	98.7109273	97.3236333

```
data3 <- data
other <- NULL
for (i in 1:ncol(data3)) {
  other <- append(other, data3[,i][1] - sum(data3[,i][c(2:4)]))
}
data3[1,] <- other
rownames(data3)[1] <- "Other.aborginal"

kable(t(data3), digits = 2)
```

	Other.aborginal	North.American.Indian	Métis	Inuit	Non.aboriginal
CA	0.11	2.23	1.25	0.16	96.25
NL	0.90	1.55	1.29	0.94	95.32
PE	0.07	0.91	0.29	0.02	98.71
NS	0.10	1.69	0.85	0.04	97.32
NB	0.11	1.72	0.59	0.03	97.55
PQ	0.06	0.88	0.38	0.15	98.54

	Other.aborginal	North.American.Indian	Métis	Inuit	Non.aboriginal
ON	0.07	1.32	0.61	0.02	97.98
MB	0.21	8.88	6.33	0.05	84.53
SK	0.23	9.58	5.04	0.02	85.12
AB	0.12	2.99	2.63	0.05	94.22
BC	0.15	3.18	1.46	0.02	95.19
YT	0.81	20.80	2.65	0.84	74.91
NT	0.62	30.78	8.72	10.13	49.73
NU	0.17	0.34	0.44	84.01	15.02

(a) (10 marks) Reformat this table to make whatever patterns it contains more easily apprehended. Show each step that you choose to follow by displaying the table that results from each step. Say why you chose to make that step by referencing the rules we had for reformatting tables. Write down a summary of whatever patterns you have uncovered.

(b) (4 marks) Note that the category 'Aboriginal.identity.population' includes the "Aboriginal groups (North American Indian, Métis and Inuit), multiple Aboriginal responses and Aboriginal responses not included elsewhere". Replace the data on 'Aboriginal.identity.population' by 'Other.aboriginal' that is the difference between 'Aboriginal.identity.population' and the North American Indian, Métis and Inuit groups. Again, give the table the best presentation and summarize whatever pattern exists.

(c) (4 marks) Whatever marginal (row or column) pattern you identified in the previous part, build the table of deviations from that pattern, display it, and comment on what you see.

6. **Graduate students** (bonus undergraduates): Suppose we have n-dimensional real and linearly independent vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$ and $\mathbf{y}$. The vector $\mathbf{y}$ is the sum of two n-dimensional real vectors μ and $\mathbf{r}$

$$\mathbf{y} = \mu + \mathbf{r}$$

where μ is restricted to be a linear combination of the vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$. That is

$$\mu = \theta_1 \times \mathbf{x}_1 + \theta_2 \times \mathbf{x}_2 + \cdots + \theta_p \times \mathbf{x}_p$$

for some unknown real constants $\theta_1, \theta_2, \ldots, \theta_p$, or equivalently

$$\mu = \mathbf{X}\theta$$

where $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_p]$ is an $n \times p$ matrix and $\theta = (\theta_1, \theta_2, \ldots, \theta_p)^T$ is a $p \times 1$ vector.

(a) (5 marks) For any $\mathbf{y}$, neither μ nor $\mathbf{r}$ are uniquely defined. Suppose we choose particular vectors $\widehat{\mu}$, and $\widehat{\mathbf{r}}$ (with $\mathbf{y} = \widehat{\mu} + \widehat{\mathbf{r}}$) to be such that they are orthogonal to one another (whatever values any θ_i take). That is, $\widehat{\mu}^T\widehat{\mathbf{r}} = 0$.

Prove that this additional constraint implies that

$$\widehat{\mu} = \mathbf{P}\mathbf{y}$$

where $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ and hence show that $\widehat{\mathbf{r}} = (\mathbf{I}_n - \mathbf{P})\mathbf{y}$.

(b) (2 marks) Show that $\mathbf{P}$ is an idempotent matrix, that is that $\mathbf{P}^2 = \mathbf{P}$.

(c) (2 marks) Show that if $\mathbf{P}$ is an idempotent matrix, then so must be $(\mathbf{I}_n - \mathbf{P})$.

(d) (2 marks) Show that $\widehat{\mathbf{r}}$ is in fact orthogonal to $\widehat{\mu}$.

(e) (5 marks) Show that $\widehat{\mu} = \mathbf{P}\mathbf{y}$ is the choice of μ which minimizes the squared length of $\mathbf{r}$. That is it minimizes $\mathbf{r}^T\mathbf{r} = (\mathbf{y} - \mu)^T(\mathbf{y} - \mu)$.