n10852565_Portfolio

2024-03-19

Load libraries

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(babynames)
library(viridis)
```

```
## Loading required package: viridisLite
```

```
library(DT)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
##
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(ggrepel)
library(patchwork)
library(ggquiver)
library(gapminder)
library(emojifont)
library(palmerpenguins)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(ggalluvial)
library(ggiraphExtra)
library(dplyr)
library(ggthemes)
```

```
##
## Attaching package: 'ggthemes'
##
## The following object is masked from 'package:ggiraphExtra':
##
##     theme_clean
```

```
library(devtools)
```

```
## Loading required package: usethis
```

```
library(alluvial)
library(ggalluvial)
library(RColorBrewer)
library(igraph)
```

```
##
## Attaching package: 'igraph'
##
## The following objects are masked from 'package:lubridate':
##
##     %--%, union
##
## The following objects are masked from 'package:dplyr':
##
##     as_data_frame, groups, union
##
## The following objects are masked from 'package:purrr':
##
##     compose, simplify
##
## The following object is masked from 'package:tidyr':
##
##     crossing
##
## The following object is masked from 'package:tibble':
##
##     as_data_frame
##
## The following objects are masked from 'package:stats':
##
##     decompose, spectrum
##
## The following object is masked from 'package:base':
##
##     union
```

```
library(raster)
```

```
## Loading required package: sp
##
## Attaching package: 'raster'
##
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(sp)
library(tinytex)
```

Week 3

Question 1

(*Marks: 2*) Use the principles of story-telling practiced in the Week 3 practical to create a story board. Embed the image into your submission and also report where you got the visualisation from (including the URL) so we can click through and see its environment. Please choose from these categories: - one figure from a paper that one of your lecturers (not Kate!), or any academic you know from QUT or elsewhere has published. Most academics will have a Google Scholar website, so search them up and browse their papers - either fivethirtyeight, The Guardian., or another online news source. - either #tidytuesday on Twitter/X, r/dataisbeautiful on Reddit

The Big Map of Who Lived When?

A. Describing the context Who is the audience or audiences?:

General public and/or people interested in figures in history.

What is the action the visualisation is aiming for? Consider each audience here if you have multiple:

Learn which figures in history lived at the same time as others. Learn some relationships these figures potentially had with each other, such as friends, relatives, rivals, collaborators, etc. Understand some key figures in a particularly age or movement in history, such as the Age of Discovery, Enlightenment, Piracy, the Baroque Period, and the Lost Generation.

When can the communication happen, and what tools have been used to suggest an order:

All information is provided at once. The author forms a chronological order of events along the horizontal axis from around the 10th century to the present day. This chronological order makes it easier as a reader to read the visualisation from left to right. While there is not a discernible order in the vertical axis, the general shape of the visualisation suggests a top-left to bottom-right reading order. The positioning of the legend and title on the bottom-left side of the visualisation also supports this left-to-right ordering. Colour is used to differentiate between the figures' expertise, such as artists, entertainers, thinkers, etc.

How has the data been used to convey the action?:

Plots signifying the birth and death years of each figure. Lines and square blocks connecting particular figures and annotating them based on their relationship with each other. Text labelling in a smaller font size next to each figure's name to highlight what their exact expertise is on top of the coloured legend denoting their general expertise, e.g. Friedrich Nietzsche was a thinker who was a philosopher, vs. Isaac Newton was a

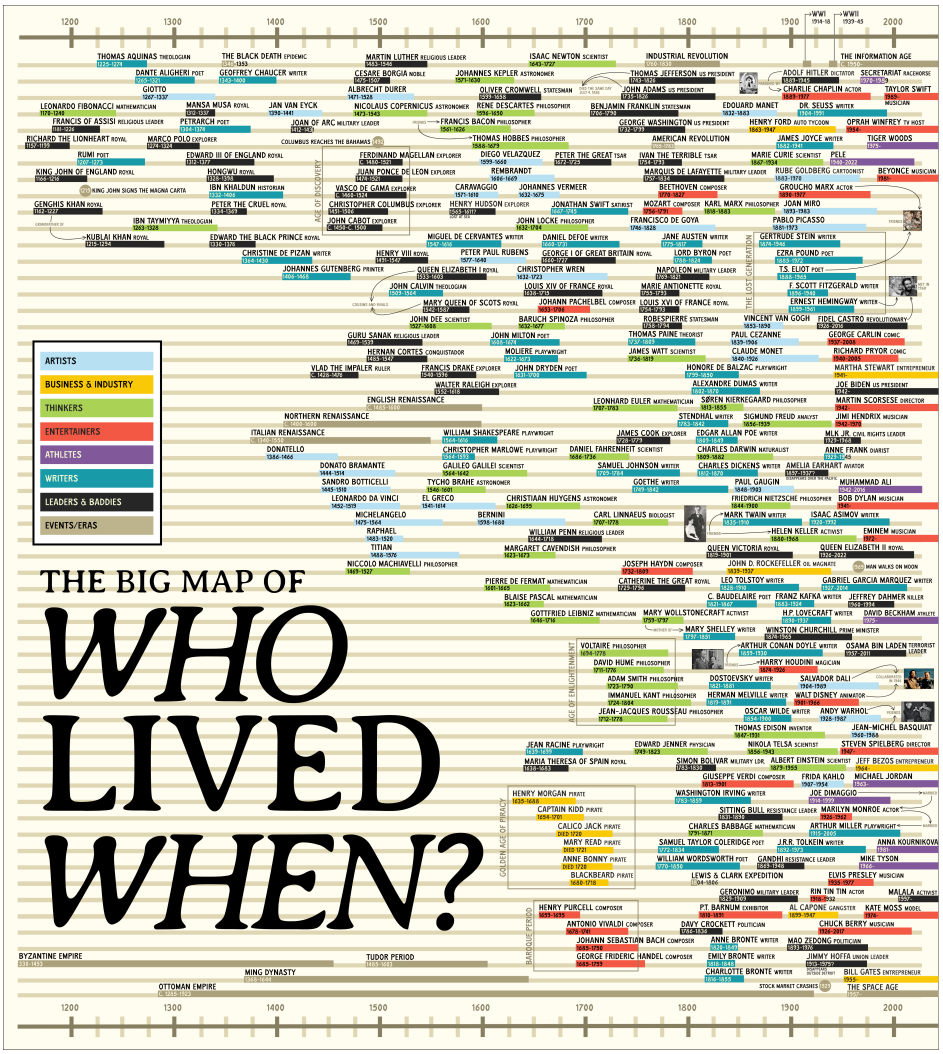


Figure 1: (OC) Who Lived When? The overlapping lives of historical figures, from 1200 to present. Created by u/profound_whatever from the r/dataisbeautiful subreddit: https://www.reddit.com/r/dataisbeautiful/comments/1ay3vuu/oc_who_lived_when_the_overlapping_lives_of/

thinker who was a scientist. These lines, square blocks, and text labelling annotate the timeline visualisation. Both the top and bottom of the horizontal axis display chronological years from left to right from ~mid-10th century to the present day.

B. Genre Which of the seven genres listed above best describes the data visualisation?

Flow chart, some annotation.

C. Author-driven vs. Reader-driven Where on the spectrum from author- to reader-driven is this visualisation?

This visualisation is generally reader-driven. While there is a chronological order in the horizontal, and it is generally easiest to start at the top-left of the visualisation, the reader can then jump to any point of the visualisation.

Question 2

(*Marks: 3*) Choose one of the papers in the readings from Week 3. In your own words (i.e. without using direct quotes from the paper), and using only information from the paper, answer the following questions: (maximum 300 words) a. What is the main argument of the paper? b. According to this paper, why is effective visual communication important (or not)? c. What are the key elements, considerations, or factors to be considered for effective visual communication addressed in the paper? Do you disagree with any? d. What pitfalls are identified in the paper that can be avoided if we use effective visual communication?

Reading: Improving Visual Communication of Science Through the Incorporation of Graphic Design Theories and Practices Into Science Communication

a. Implementing theory and practice from graphic design will aid the effectiveness of current scientific visual communicators. The effective visual communication of science is largely hindered by the fact that visualisations are generally used as a supplement rather than a natural, fundamental aspect of communication, and that these communicators fail to identify specific audiences and recontextualising these visualisations specifically to them.

b. In the current age, visual communication is displacing written communication similar to how the latter displaced oral communication. Visual communication improves an audience's understanding of scientific explanations, increasing their interest in the subject. The more effective the visual communication, the higher chance the audience may change their perspectives and actions.

c. Consider from the beginning the audience in which you are presenting a visualisation to, as they may perceive it differently to other audiences, particularly between specialists and non-specialists. Catering these visualisation to a specific audience increases the effectiveness of visual communication. Scientists need to learn how to understand and use visualisations effectively, learning from the field of design. I agree with these considerations, as if a particular scientific visualisation specifically appeals to me, then I would use the knowledge gained from it to either fuel some action or just to purely gain interest in the topic.

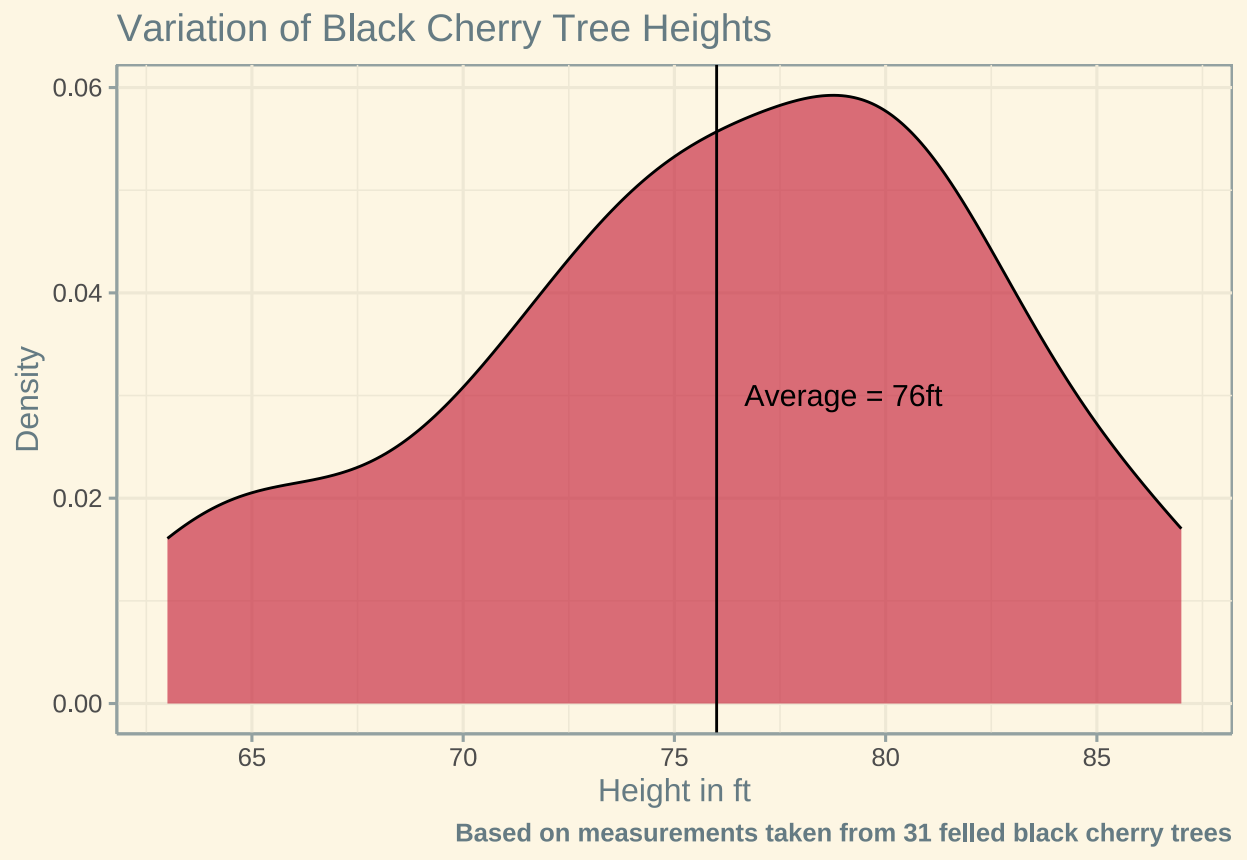
d. The choice and function of a chosen visualisation may not be entirely appropriate to the audience. Visualisations that are not well-incorporated with the written media diminish their impact on the audience.

Week 4

Question 3

(*Marks: 6*) Explore the datasets already loaded into your R workspace by typing `data()`, and use one of these to design a visualisation from the types covered in week 4 that shows variation of tree heights. Justify your choice of dataset, plot type, and variables. Give a short justification for your aesthetic choices and how they make the figure a more effective communication tool.

```
library(ggplot2)
library(ggthemes)

ggplot(trees, aes(Height)) +
  geom_density(fill = "#c7233c", alpha=0.65) +
  scale_x_continuous(breaks=seq(65, 85, by = 5)) +
  labs(title = "Variation of Black Cherry Tree Heights",
       caption = "Based on measurements taken from 31 felled black cherry trees",
       x = "Height in ft",
       y = "Density") +
  theme_solarized() +
  theme(plot.caption = element_text(face = "bold")) +
  geom_vline(xintercept = mean(trees$Height)) +
  annotate('text', label="Average = 76ft", x=79, y=0.03, size = 4)
```



The visualisation above makes use of the `trees` dataset, which 'provides measurements of the diameter, height, and volume of timber in 31 felled black cherry trees'. Using the Height variable of this dataset, we

can design a visualisation showing a variation of tree heights `trees$Height`. As we are only interested in one variable for this visualisation (tree heights), and this variable is continuous, a density plot is appropriate. A histogram could also be used with an appropriate value of bins/binwidth, however a density plot was more effective in communicating the variation of tree heights overall in the dataset using a line rather than bins as the encoding object. Aesthetically, the plot was given a simple title and caption which instantly aids the reader in decoding the general message of the visualisation. A solid fill colour was used rather than a palette of colours as we are not interested in highlighting a second, categorical variable. Transparency was used such that the grid can still be seen which makes it easier to see the scale on both axes along the density plot and avoiding communication gap decoding issues. The scale on the x-axis was modified to include more ticks for ease of reading as well as avoiding inconsistency problems which may lead to distortion decoding issues. Both axes were given titles so the reader can instantly decode their meanings instead of inferring that one denotes height and the other denotes density. A vertical line was included to display the average tree height which is a valuable measurement takeaway for a visualisation regarding a variation of tree heights. Overall, these aesthetic choices aid…

Question 4

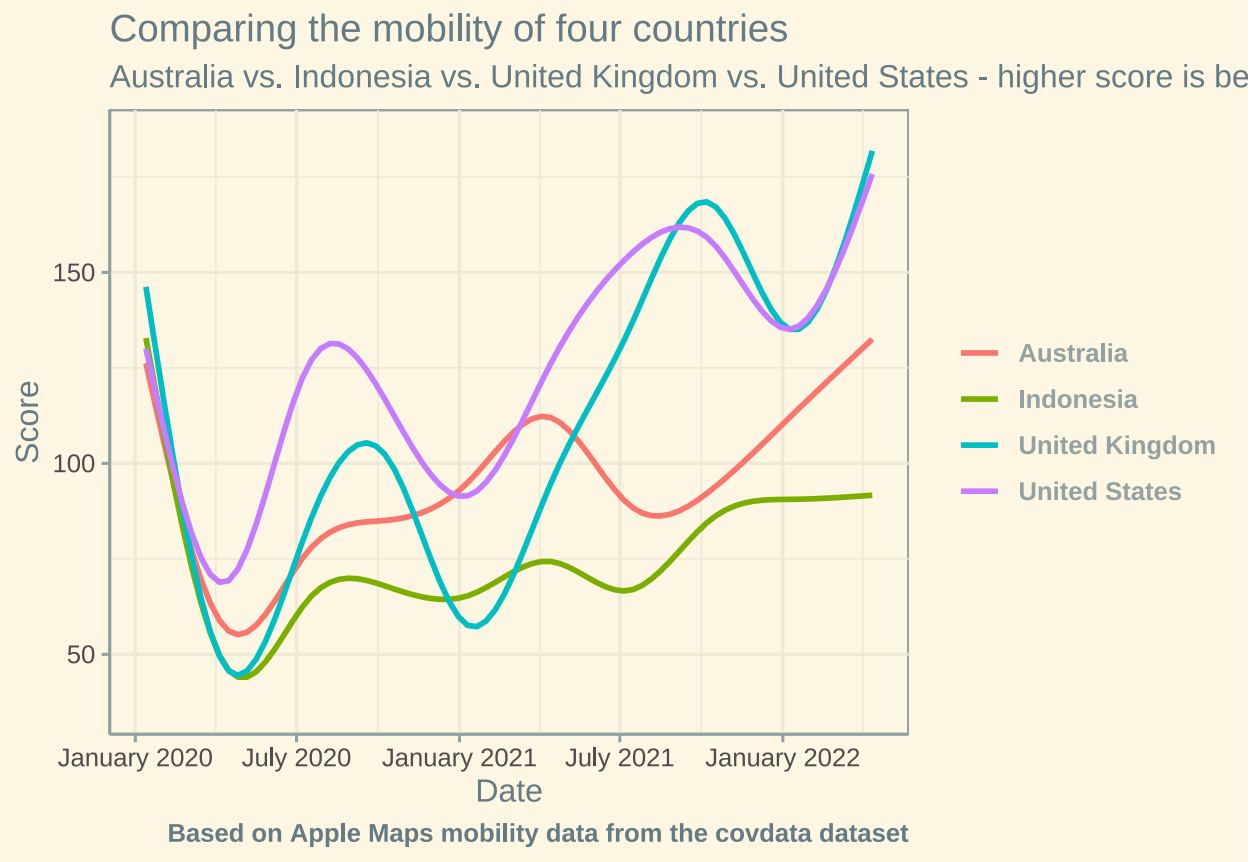
(**Marks 6**) Using the `covdata` dataset from the practical in Week 4, compare the mobility of four countries or regions. Include the code you write to subset the data (make sure it's being shown in your knitted markdown file by setting `echo=TRUE` when setting up the markdown code chunk). Justify your choice of plot type and variables. Give a short justification for your aesthetic choices and how they make the figure a more effective communication tool.

```
library(covdata)
```

```
##
## Attaching package: 'covdata'

## The following object is masked from 'package:datasets':
##
##     uspop
```

```
library(ggplot2)
library(ggthemes)

cbPalette <- c("#E69F00", "#56B4E9", "#009E73", "#CC79A7") # from http://www.cookbook-r.com/Graphs/Colo

mobility_subset <- apple_mobility %>%
  dplyr::filter(country == c("Australia", "United States", "Indonesia", "United Kingdom"))

ggplot(mobility_subset, aes(date, score, color=country)) +
  geom_line(stat="smooth", size=1) +
  scale_x_date(date_labels = "%B %Y") +
  scale_fill_manual(values=cbPalette) +
  labs(title = "Comparing the mobility of four countries",
       subtitle = "Australia vs. Indonesia vs. United Kingdom vs. United States - higher score is bette
       caption = "Based on Apple Maps mobility data from the covdata dataset",
       x = "Date",
       y = "Score") +
  theme_solarized() +
  theme(plot.caption = element_text(face = "bold")) +
  theme(legend.title = element_blank()) +
  theme(legend.text = element_text(face="bold"))
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



The visualisation above compares the mobility scores between four countries - Australia, Indonesia, the U.K., and the U.S. This was done through plotting the mobility scores per country through time, with higher scores denoting better mobility. A time-series line plot was used as it was most appropriate when comparing a continuous variable (score) against time for multiple categorical variables (country). This line plot was smoothed as we are mainly interested in an 'average' mobility score for the purposes of this visualisation. An area plot may also be used, however a line plot is more appropriate as using the former would require data being displayed in multiple plots rather than a single one like above which helps in comparison. A stacked area plot would be entirely inappropriate as well as we are not comparing the score variable as a 'whole' but rather as individual statistics per country variable. Aesthetically, a title, subtitle, and caption were included to aid the reader in instantly decoding the general aim (comparison) of the visualisation. The colours chosen for the country variables were chosen as they were the most colour-blind friendly as per R Cookbook. The line thickness was adjusted such that it was easier to discern within the plot grid. The axes labels were clearly defined, and the ticks specifically for the date axes were modified to display the month and year only, as potentially including specific days as well as using a numerical format may be redundant and negatively impact clutter decoding issues. A legend was also included to aid decoding and prevent a communication gap. This includes the colour associated with each country. Note that a 'country' label was removed from the legend for redundancy purposes. Overall, these aesthetic choices makes the figure a more effective communication tool, as well as aids the audience's ability to decode the message of the visualisation comparing the mobility of four countries and how they have changed since January 2020.

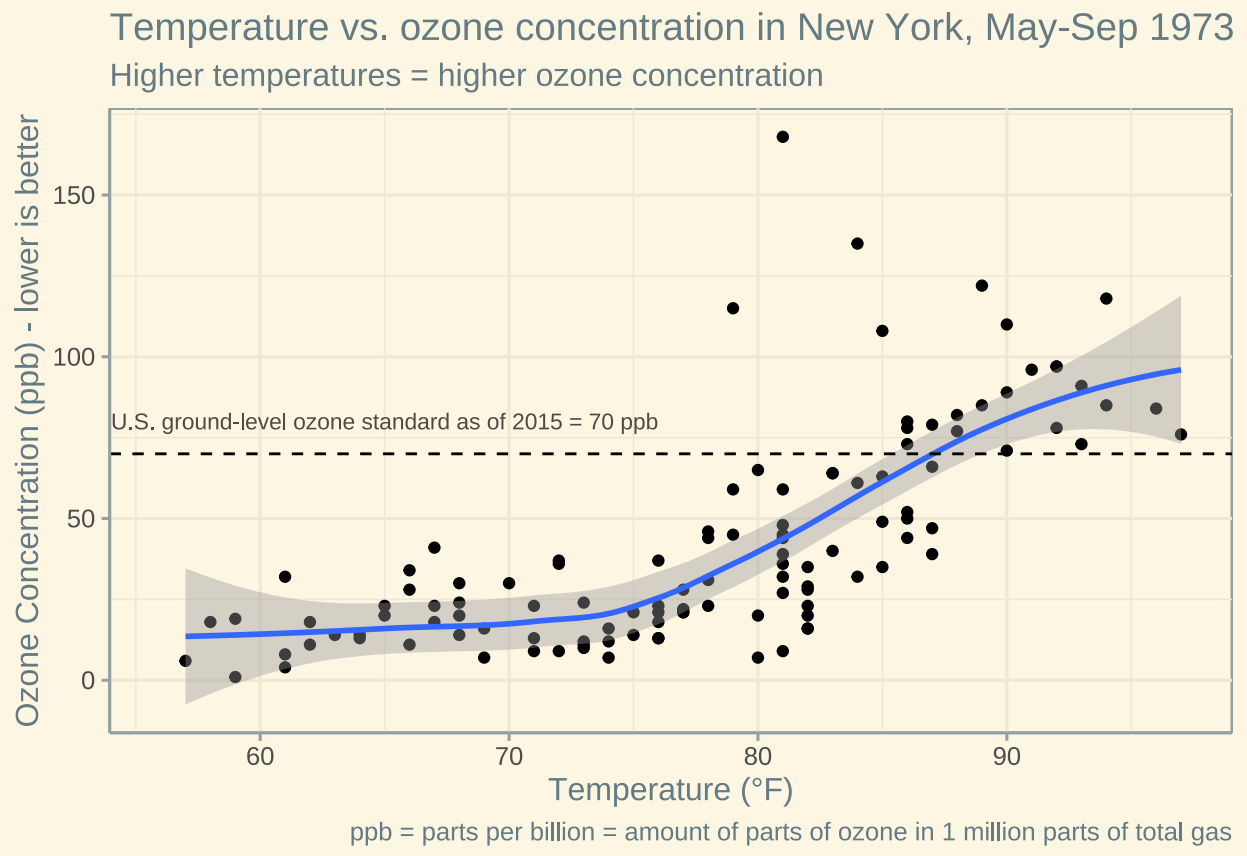
Question 5

(**Marks 10**) Explore any of the pre-loaded datasets you like. Choose one that we haven't explored yet in the unit materials. Produce two plots from the two variable plot types we explored in week 4, justify your

choice of plot type and variables, and explain what your audience might discover from that plot. Give a short justification for your aesthetic choices and how they make the figure a more effective communication tool.

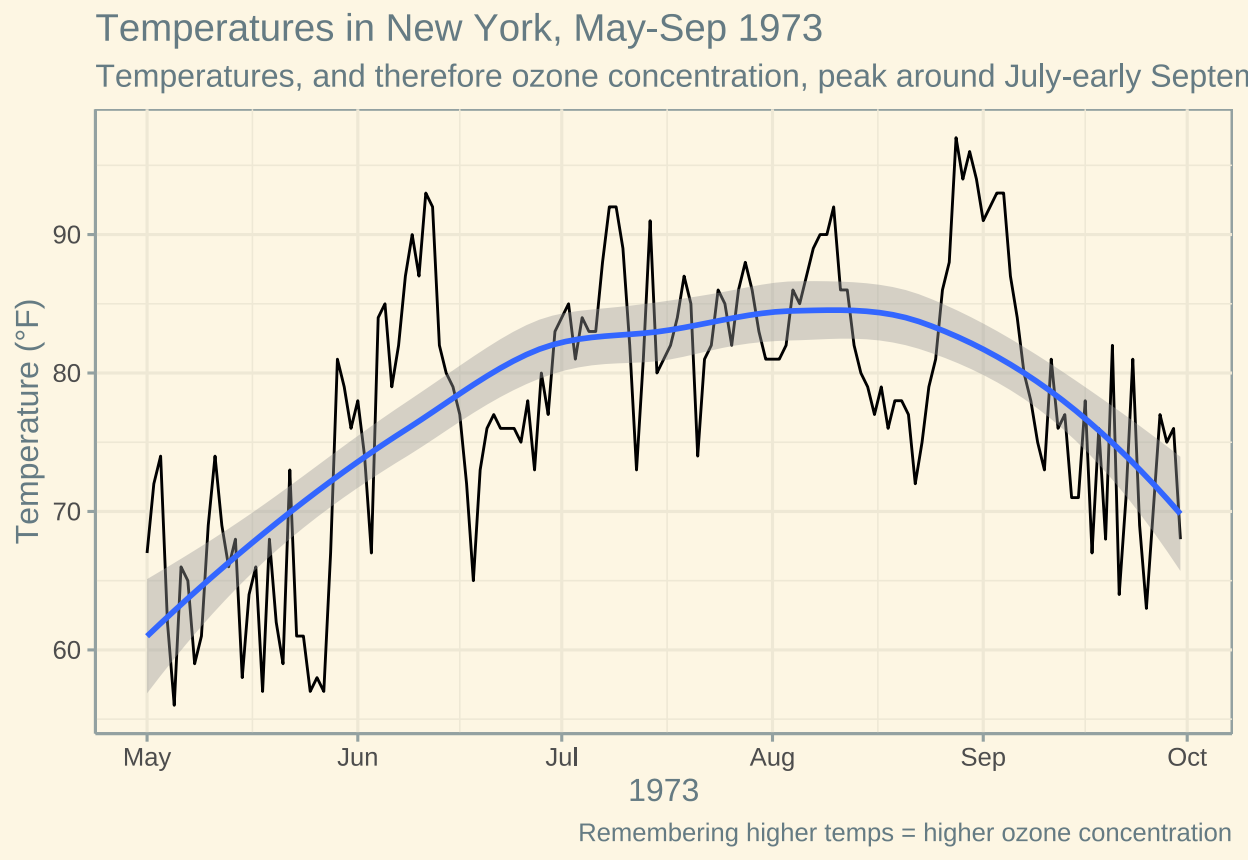
```
library(ggplot2)
library(ggthemes)

# ?airquality
# head(airquality)

# Scatterplot
ggplot(airquality, aes(x=Temp, y=Ozone)) +
  geom_point() +
  geom_smooth() +
  labs(title = "Temperature vs. ozone concentration in New York, May-Sep 1973",
       subtitle = "Higher temperatures = higher ozone concentration",
       caption = "ppb = parts per billion = amount of parts of ozone in 1 million parts of total gas",
       x = "Temperature (°F)",
       y = "Ozone Concentration (ppb) - lower is better") +
  theme_solarized() +
  geom_hline(yintercept = 70, linetype = "dashed") +
  annotate('text', label="U.S. ground-level ozone standard as of 2015 = 70 ppb", x=65, y=80, size = 3, a
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



```r
# Create date field in airquality df by combining Month and Day field and new Year field
Year <- 1973 # From documentation (?airquality)
airquality$Date <- as.Date(with(airquality,paste(Year,Month,Day,sep="-")), "%Y-%m-%d")

# Time-series line plot
ggplot(airquality, aes(x=Date, y=Temp)) +
  geom_line() +
  geom_smooth() +
  labs(title = "Temperatures in New York, May-Sep 1973",
       subtitle = "Temperatures, and therefore ozone concentration, peak around July-early September!",
       caption = "Remembering higher temps = higher ozone concentration",
       x = "1973",
       y = "Temperature (°F)") +
  theme_solarized()
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



These visualisations make use of a dataset measuring daily air quality in New York from May to September 1973. These measurements include 6 potential variables of which 4 were used - Temperature (measured in °F), Ozone (measured in ppb), and a combination of Month (1-12) and Day (1-31). The aim of these visualisations collectively is for the audience to learn that higher ozone concentrations take place around the middle of the year, specifically July-early September. This is done by first establishing a positive correlation and relationship between temperature and ozone concentration using a scatterplot, then again using temperature in a time-series line plot to visualise the change in temperature throughout May-September, knowing that higher temperatures correlate to higher ozone concentrations. As we are interested in two variables for both

visualisations, a scatterplot and time-series line plots are most appropriate, where the former is best suited to establish relationships, correlations, and trends, and the latter is best suited in measuring continuous variables through time. Aesthetically, both plots have the same key aspects. Both plots include a simple title, a supportive subtitle highlighting a key takeaway, a supportive caption enhancing understanding of the plots, and clearly-defined axes and axes titles, overall reducing potential encoding and decoding issues. Both plots make use of smoothing which averages key variables and makes it easier to discern correlations between two variables, which in this context is highly beneficial for our message. Overall, the plot type, variables, and aesthetic choices collectively make the figure a more effective communication tool and supports the audience in learning that higher ozone concentrations take place around the middle of the year.

Week 6

Question 6

(**Marks 5**) Create two bubble plots (a.k.a multi-dimensional scatterplot) using the dataset downloaded by

```
gender_pay_gap <- read.csv(
  "https://raw.githubusercontent.com/plotly/datasets/master/school_earnings.csv"
)

head(gender_pay_gap)
```

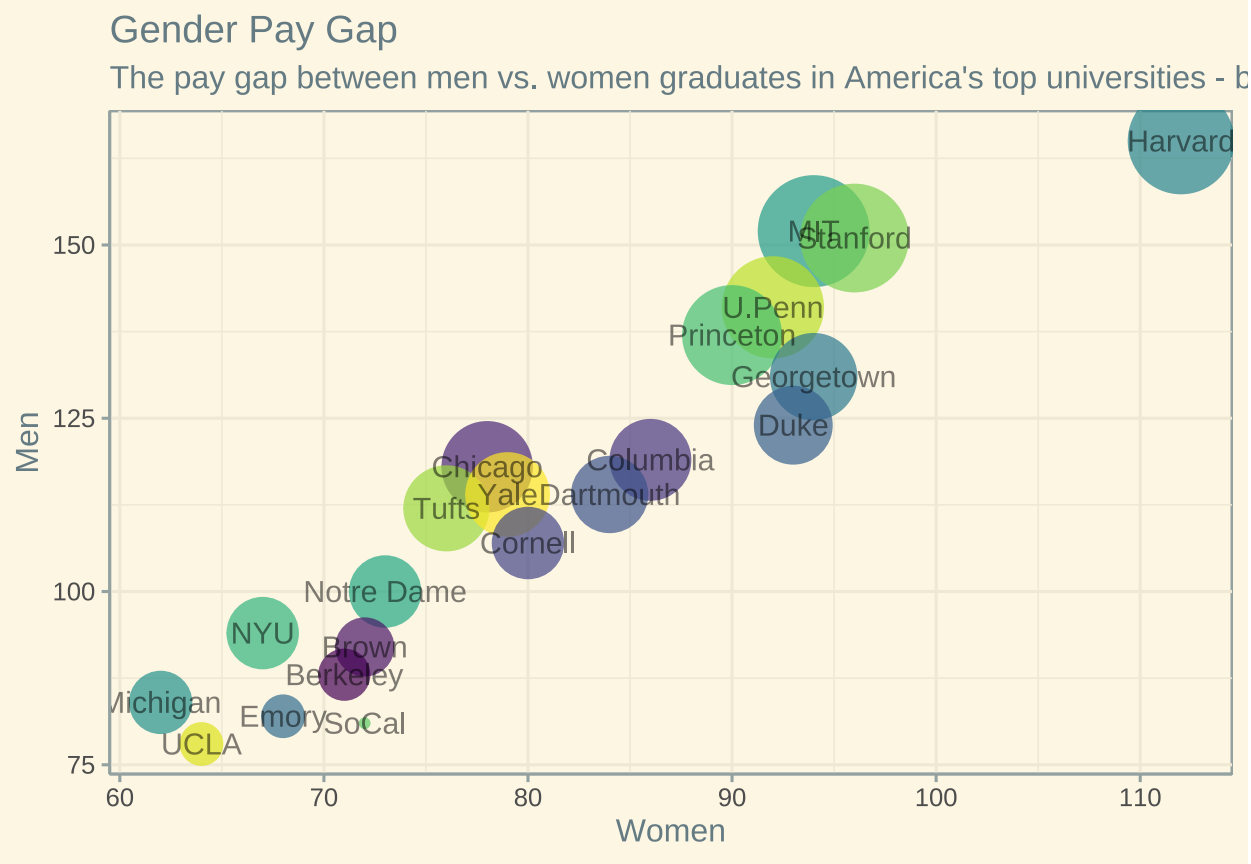
```
##      School Women Men Gap
## 1       MIT    94 152  58
## 2  Stanford    96 151  55
## 3   Harvard   112 165  53
## 4    U.Penn    92 141  49
## 5 Princeton    90 137  47
## 6   Chicago    78 118  40
```

```
dplyr::glimpse(gender_pay_gap)
```

```
## Rows: 21
## Columns: 4
## $ School <chr> "MIT", "Stanford", "Harvard", "U.Penn", "Princeton", "Chicago",~
## $ Women  <int> 94, 96, 112, 92, 90, 78, 94, 76, 79, 86, 93, 84, 67, 73, 80, 62~
## $ Men    <int> 152, 151, 165, 141, 137, 118, 131, 112, 114, 119, 124, 114, 94,~
## $ Gap    <int> 58, 55, 53, 49, 47, 40, 37, 36, 35, 33, 31, 30, 27, 27, 27, 22,~
```

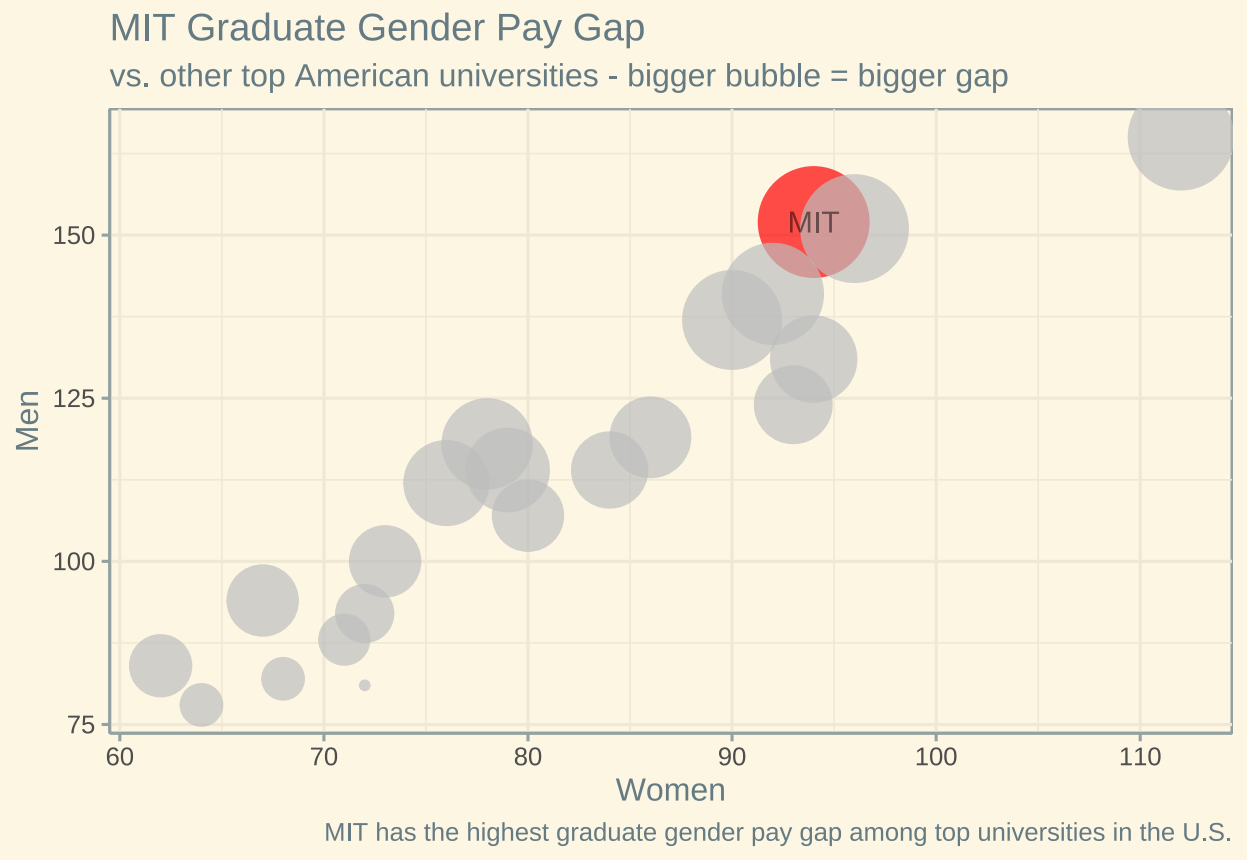
First, make the bubble plot you think is most useful.

```
ggplot(gender_pay_gap, aes(x=Women, y=Men, size=Gap, color=School)) +
  geom_point(alpha=0.7) +
  geom_text(aes(label = School), size = 4, color="black", alpha=0.5) +
  scale_size(range = c(1.4, 19), name="Gender Pay Gap") +
  scale_color_viridis(discrete=TRUE, guide=FALSE) +
  labs(title = "Gender Pay Gap",
       subtitle = "The pay gap between men vs. women graduates in America's top universities - bigger bu
  theme_solarized() +
  theme(legend.position = "none")
```



Second, add at least one more variable/preattentive attribute to your plot.

```
gender_pay_gap_highlight <- gender_pay_gap %>%
  mutate(Highlight = ifelse(School=="MIT", "MIT", "Other"))

ggplot(gender_pay_gap_highlight, aes(x=Women, y=Men, size=Gap, color=Highlight)) +
  geom_point(alpha=0.7) +
  geom_text(data = subset(gender_pay_gap_highlight, School == "MIT"), aes(label = Highlight), size = 4,
  scale_size(range = c(1.4, 19), name="Gender Pay Gap") +
  scale_color_manual(values=(c("red", "grey")), guide=FALSE) +
  labs(title = "MIT Graduate Gender Pay Gap",
       subtitle = "vs. other top American universities - bigger bubble = bigger gap",
       caption = "MIT has the highest graduate gender pay gap among top universities in the U.S.") +
  theme_solarized() +
  theme(legend.position = "none")
```



1. What do you think is the first observation a reader makes from the second plot?

The initial observation that a reader may make from the second plot is the MIT bubble plot and its comparative size to the other bubbles within the overall plot. This is the due to the red hue preattentive attribute that makes it stand out among the other, grey university bubbles. Text is also used to label the MIT bubble while other bubbles are not labelled. The audience can initially see that MIT is at the larger end of the bubble sizes, and further observations of the title and subtitle hammers down the message that MIT has the largest gender pay gap among top universities in the U.s..

2. Why have you chosen to match the variable with the preattentive attributes you have chosen in the first plot?

Size and colour together emphasises comparison of the School variable very effectively. Using bubble size to assign the values of each school's gender pay gap emphasises comparison which is the key take-away of the visualisation, and using a categorical colourmap for the schools is also very helpful in differentiating each school's bubble from each other as compared to using a static colour.

3. What audience might you use your first plot for vs your second?

The first plot is a very general comparison between all top universities in the U.S.. A suitable audience therefore would be individuals who are planning to study in these universities to compare their prospective future earnings against the opposite sex. The second plot is much more specific as it compares one particular school (MIT) to all other schools. Therefore, a suitable audience for the second plot are MIT students comparing their prospective future earnings against the opposite sex.

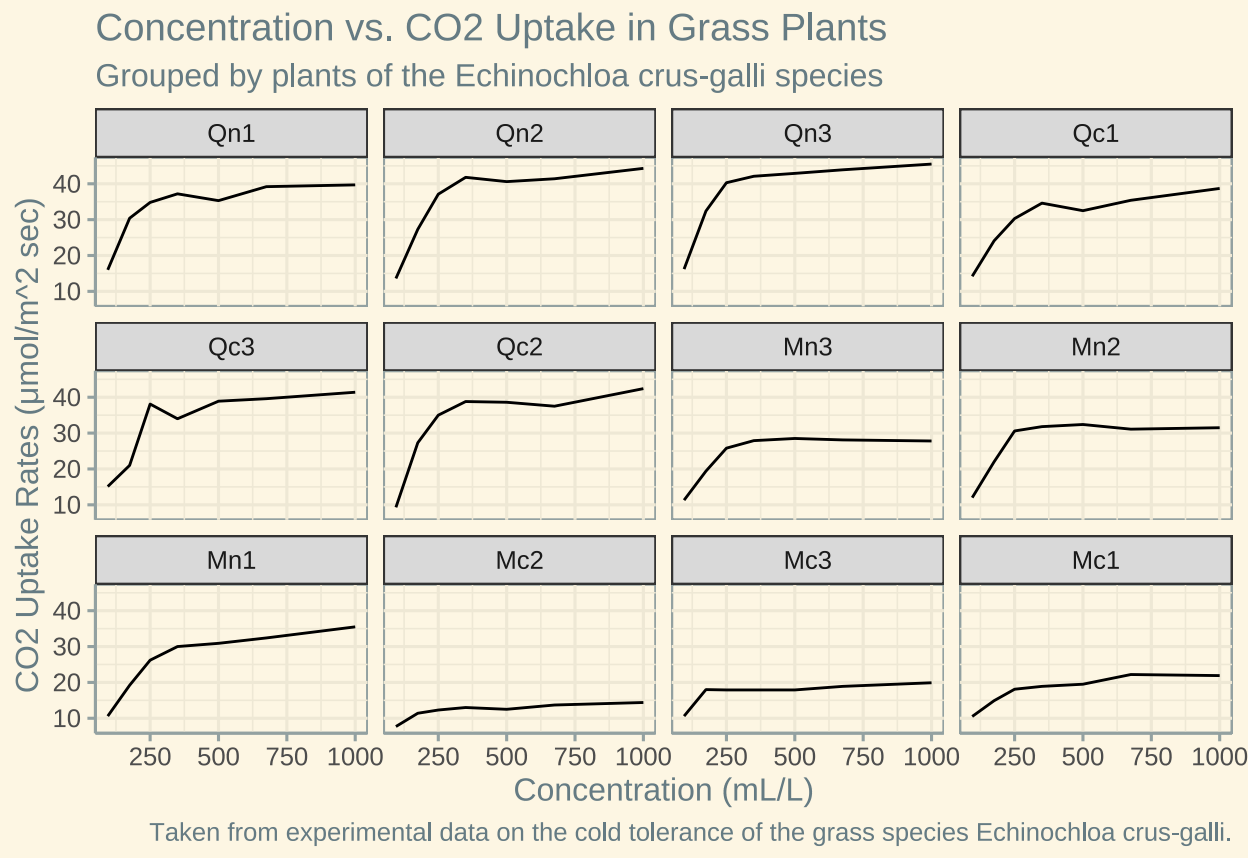
Question 7

(**Marks 3**) Using faceting to explore the relationship between C02 uptake, concentration, and plant using the `CO2` dataset.

```
head(CO2)
```

```
## Grouped Data: uptake ~ conc | Plant
##   Plant   Type  Treatment conc uptake
## 1   Qn1 Quebec nonchilled   95   16.0
## 2   Qn1 Quebec nonchilled  175   30.4
## 3   Qn1 Quebec nonchilled  250   34.8
## 4   Qn1 Quebec nonchilled  350   37.2
## 5   Qn1 Quebec nonchilled  500   35.3
## 6   Qn1 Quebec nonchilled  675   39.2
```

```
ggplot(CO2, aes(x=conc, y=uptake, group=Plant)) +
  geom_line() +
  labs(title = "Concentration vs. CO2 Uptake in Grass Plants",
       subtitle = "Grouped by plants of the Echinochloa crus-galli species",
       x = "Concentration (mL/L)",
       y = "CO2 Uptake Rates ( mol/m^2 sec)",
       caption = "Taken from experimental data on the cold tolerance of the grass species Echinochloa c
  theme_solarized() +
  facet_wrap(~Plant)
```



1. What do you think is the first observation a reader makes from the plot?

The first observation a reader makes is the title and subtitle of the visualisation. This describes the general message of the visualisation - the comparison between two variables grouped/faceted by a categorical third variable. From this, it can be discerned from an initial overview of the visualisation that concentration and uptake generally have a direct relationship among all grass plants of the Echinochloa crus-galli species included in the experimental dataset.

2. Why/why not is faceting an effective way to visualise this data to see these relationships?

Faceting is an effective way to visualise the relationship within this data as it can visualise multi-dimensional comparisons in a much clearer way relative to overlaying multiple plots, for example. Describing the data would also be easier, as the author can refer to 'Plant Qn1' for example. Though, the audience would have to exert extra cognitive strain and working memory to flick between the panels. Sometimes, faceting could become overwhelming when visualising multiple panels or including another dimension (variable), though in this case it is still easily digestible and discerning the slope of the line across the panels (determining the direct relationship) is relatively trivial.

3. What is another visualisation option (you do not need to make this plot)?

Using parallel plots could be an alternative to the above visualisation. Two vertical lines would represent concentration and uptake, and colour could be used as a preattentive attribute to label the 9 different plants. Depending on how the lines appear in the visualisation, we can determine if there is a general correlation between concentration and uptake. Based on the visualisation above, we could possible infer that the lines in a similar parallel plot would run parallel, indicating a positive correlation.

Question 8

(**Marks 7**) You have a few choices for this question. Use either:

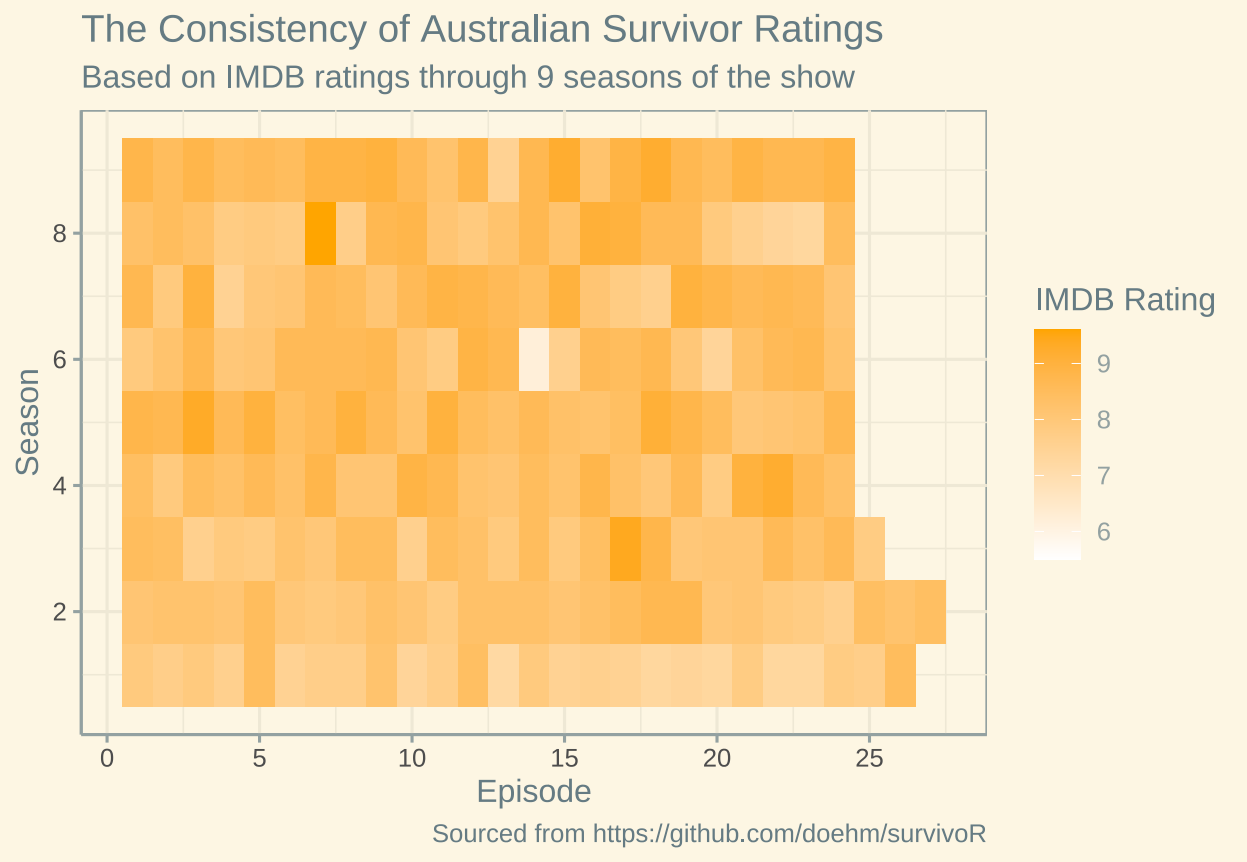
- `survivoR` dataset (from `devtools::install_github("doehm/survivoR")`),
- the `billboard` dataset (built in), or
- deforestation data, downloaded by running the following code once only. Then, explore the data downloaded by typing `tuesdata$` into the console in R studio and viewing each of the options that pop up.

Create one visualisation using any of those three datasets, of either:

- heat map
- radar plot
- alluvial

```
# devtools::install_github("doehm/survivoR")

survivor_data <- survivoR::episodes
episodes_filtered <- survivoR::episodes %>%
  filter(version=="AU")

ggplot(episodes_filtered, aes(x=episode, y=season)) +
  geom_tile(aes(fill = imdb_rating)) +
  scale_fill_continuous(low = "white", high = "orange", limits = c(5.5,9.6)) +
```

```
  scale_x_continuous(breaks = c(0,5,10,15,20,25)) +
  scale_y_continuous(breaks = c(2,4,6,8)) +
  labs(title = "The Consistency of Australian Survivor Ratings",
       subtitle = "Based on IMDB ratings through 9 seasons of the show",
       x = "Episode",
       y = "Season",
       fill = "IMDB Rating",
       caption = "Sourced from https://github.com/doehm/survivoR") +
  theme_solarized()
```



Questions:

1. what do you think is the first observation a reader makes from the plot?

The first observation a reader makes from this plot is the consistency of the heatmap throughout the seasons and episodes of the Australian version of Survivor. There is not much of a change in colour or discernible upwards or downwards trend in ratings as the show went on, thereby a conclusion can be made that the show has been very consistent throughout its runtime, which is the main message of the plot as explicitly communicated by the title.

2. why/why not is your chosen plot type an effective way to visualise this data to make this comparison?

In terms of communicating ratings consistency with this data, a heatmap is an effective way as it makes use of colour and hue as pre-attentive attributes. Specifically, the *lack* of hue between episodes and seasons is what makes this plot effective in communicating a certain *consistency* of the show throughout its runtime.

The use of radar plots or alluvials are ineffective since we are only making use of a small amount of variables which makes such plots too trivial and irrelevant.

3. what is another visualisation option (you do not need to make this plot)?

Alternatively, we can implement more variables and create a radar plot grouped and coloured by season to see its average ratings, episode lengths, viewers, number of episodes, etc. This would provide further analysis on various statistics relating to each season rather than just ratings.

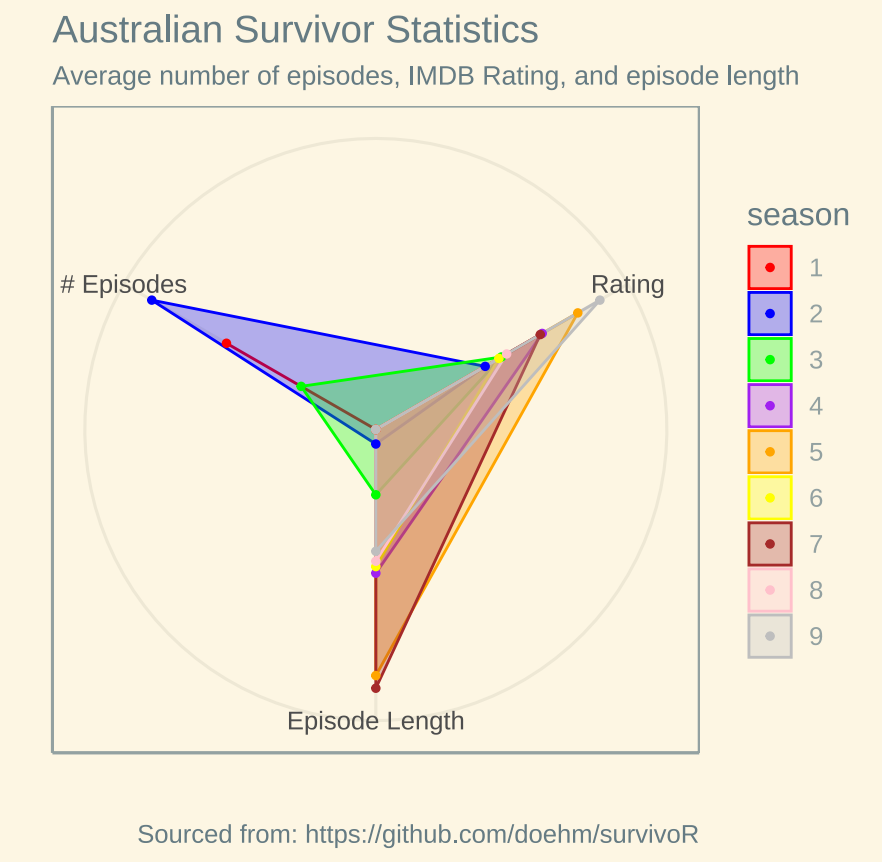
Question 9

(**Marks 7**) Using a different combination of dataset and plot type from the options in the previous question, create another visualisation. It is fine to either use the same dataset or the same visualisation type, but not both.

```
# devtools::install_github("doehm/survivoR")

episodes_average <- episodes_filtered %>%
  group_by(season) %>%
  summarise(avg_imdb_rating = mean(imdb_rating, na.rm = TRUE),
            avg_episode_length = mean(episode_length, na.rm = TRUE),
            num_episodes = n()) # Number of observations in current group https://www.rdocumentation.or

episodes_average$season <- as.factor(episodes_average$season)

head(episodes_average)
```

```
## # A tibble: 6 x 4
##   season avg_imdb_rating avg_episode_length num_episodes
##   <fct>            <dbl>              <dbl>        <int>
## 1 1                 7.70               60.1           26
## 2 2                 8.18               60.6           27
## 3 3                 8.25               62.6           25
## 4 4                 8.43               65.5           24
## 5 5                 8.58               69.4           24
## 6 6                 8.24               65.3           24
```

```
ggRadar(episodes_average, aes(x = c(avg_imdb_rating, avg_episode_length, num_episodes),
                              group = season),
        rescale = TRUE,
        size = 1) +
  scale_color_manual(values = c("red", "blue", "green", "purple", "orange", "yellow", "brown", "pink",
  scale_fill_manual(values = c("red", "blue", "green", "purple", "orange", "yellow", "brown", "pink", "
  theme_solarized() +
  scale_y_discrete(breaks = NULL) +
  labs(title = "Australian Survivor Statistics",
       subtitle = "Average number of episodes, IMDB Rating, and episode length",
       caption = "Sourced from: https://github.com/doehm/survivoR") +
  scale_x_discrete(labels = c("Rating", "Episode Length", "# Episodes")) +
  theme(plot.subtitle = element_text(size = 10))
```



Questions:

1. what do you think is the first observation a reader makes from the plot?

Australian Survivor has had varying relative averages of episode lengths and # of episodes throughout its runtime. The average IMDB rating seems to be relatively close to each other, though it can be discerned that season 9 had the highest rating. Season 7 had the highest episode length, and season 1 and 3 had shorter number of episodes compared to the other seasons.

2. why/why not is your chosen plot type an effective way to visualise this data to make this comparison?

When visually comparing multiple variables grouped by a categorical variable, a radar plot is effective, as the reader can view the shape and area of the polygons as pre-attentive attributes to compare with others. It is very effective for comparisons, is novel and interesting, and trivially visualises data in a way that is accessible to general audiences.

3. what is another visualisation option (you do not need to make this plot)?

Alluvials can be created to see the relationships between average # of episodes, episode lengths, and ratings.

Week 7

Question 10

(**Marks 10**) 1. In the workshop, we visualised a migration network as a chord diagram. Use the same data to create a matching alluvial plot of this same migration network (see lecture for example). Your plot

will be assessed for accuracy and effectiveness, so design your visualisation and use pre-attentive attributes carefully.

```
# Load dataset from github
migration <- read.table("https://raw.githubusercontent.com/holtzy/data_to_viz/master/Example_dataset/13_
```

```
head(migration)
```

```
##                 Africa East.Asia   Europe Latin.America North.America  Oceania
## Africa        3.142471  0.000000 2.107883      0.000000      0.540887 0.155988
## East Asia     0.000000  1.630997 0.601265      0.000000      0.973060 0.333608
## Europe        0.000000  0.000000 2.401476      0.000000      0.000000 0.000000
## Latin America 0.000000  0.000000 1.762587      0.879198      3.627847 0.000000
## North America 0.000000  0.000000 1.215929      0.276908      0.000000 0.000000
## Oceania       0.000000  0.000000 0.170370      0.000000      0.000000 0.190706
##               South.Asia South.East.Asia Soviet.Union West.Asia
## Africa                 0        0.000000            0  0.673004
## East Asia              0        0.380388            0  0.869311
## Europe                 0        0.000000            0  0.000000
## Latin America          0        0.000000            0  0.000000
## North America          0        0.000000            0  0.000000
## Oceania                0        0.000000            0  0.000000
```

```
# short names
colnames(migration) <- c("Africa", "EAsia", "Europe", "LatinAm.",   "NorthAm.",   "Oceania", "SAsia", "S
rownames(migration) <- colnames(migration)
```

```
head(migration)
```

```
##            Africa    EAsia   Europe LatinAm. NorthAm.  Oceania SAsia   SEAsia
## Africa   3.142471 0.000000 2.107883 0.000000 0.540887 0.155988     0 0.000000
## EAsia    0.000000 1.630997 0.601265 0.000000 0.973060 0.333608     0 0.380388
## Europe   0.000000 0.000000 2.401476 0.000000 0.000000 0.000000     0 0.000000
## LatinAm. 0.000000 0.000000 1.762587 0.879198 3.627847 0.000000     0 0.000000
## NorthAm. 0.000000 0.000000 1.215929 0.276908 0.000000 0.000000     0 0.000000
## Oceania  0.000000 0.000000 0.170370 0.000000 0.000000 0.190706     0 0.000000
##          Sov.Un.    WAsia
## Africa         0 0.673004
## EAsia          0 0.869311
## Europe         0 0.000000
## LatinAm.       0 0.000000
## NorthAm.       0 0.000000
## Oceania        0 0.000000
```

```
# The data is in the form of an adjacency matrix, but we need it in a three column matrix instead: firs
migration_long <- gather(rownames_to_column(migration), key='key', value='value', -rowname)
```

```
head(migration_long)
```

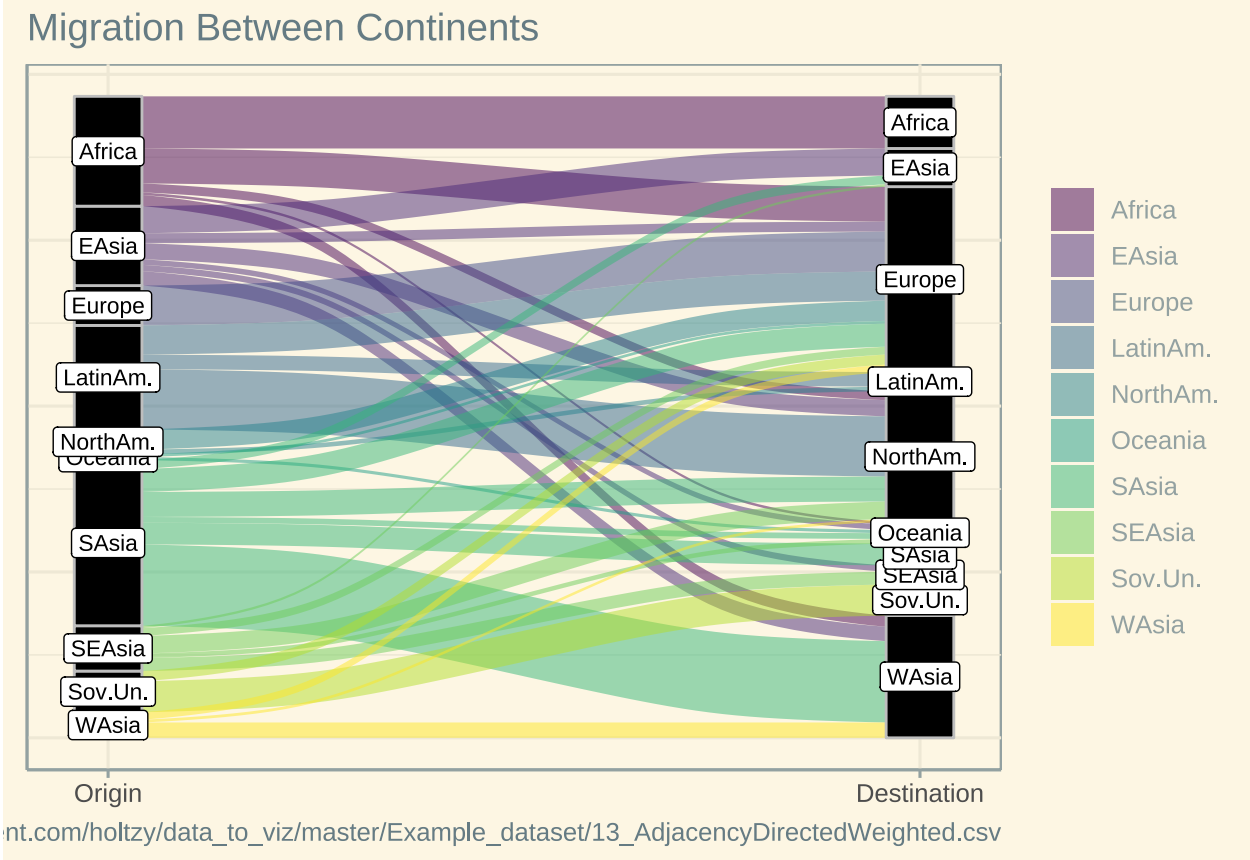
```
##    rowname    key    value
## 1   Africa Africa 3.142471
## 2    EAsia Africa 0.000000
```

```
## 3   Europe Africa 0.000000
## 4 LatinAm. Africa 0.000000
## 5 NorthAm. Africa 0.000000
## 6  Oceania Africa 0.000000
```

```
migration_nonzero <- migration_long[apply(migration_long!=0, 1, all),] # https://scales.arabpsychology.

head(migration_nonzero)
```

```
##    rowname    key    value
## 1   Africa Africa 3.142471
## 12   EAsia  EAsia 1.630997
## 17   SAsia  EAsia 0.525881
## 18  SEAsia  EAsia 0.145264
## 21  Africa Europe 2.107883
## 22   EAsia Europe 0.601265
```

```
# colour palette
migration_color <- viridis(nrow(migration)) # 10 colours

ggplot(migration_nonzero, aes(y=value, axis1=rowname, axis2=key)) +
  geom_alluvium(aes(fill=rowname),width=1/12) +
  geom_stratum(width = 1/12, fill = "black", color="grey", stratum_padding = 2) +
  geom_label(stat="stratum", aes(label=after_stat(stratum)), size = 3) +
  scale_x_discrete(limits=c("Origin", "Destination"), expand=c(.05, .05)) +
  scale_fill_manual(values = migration_color) +
  labs(title = "Migration Between Continents",
       caption = "Source: https://raw.githubusercontent.com/holtzy/data_to_viz/master/Example_dataset/13
       fill = "Continent",
       y = NULL) +
  theme_solarized() +
  theme(axis.text.y=element_blank(),
        axis.ticks.y=element_blank(),
        legend.title=element_blank())
```



2. In 1-2 sentences, propose any differences in the who/what/when context between these two visualisations.

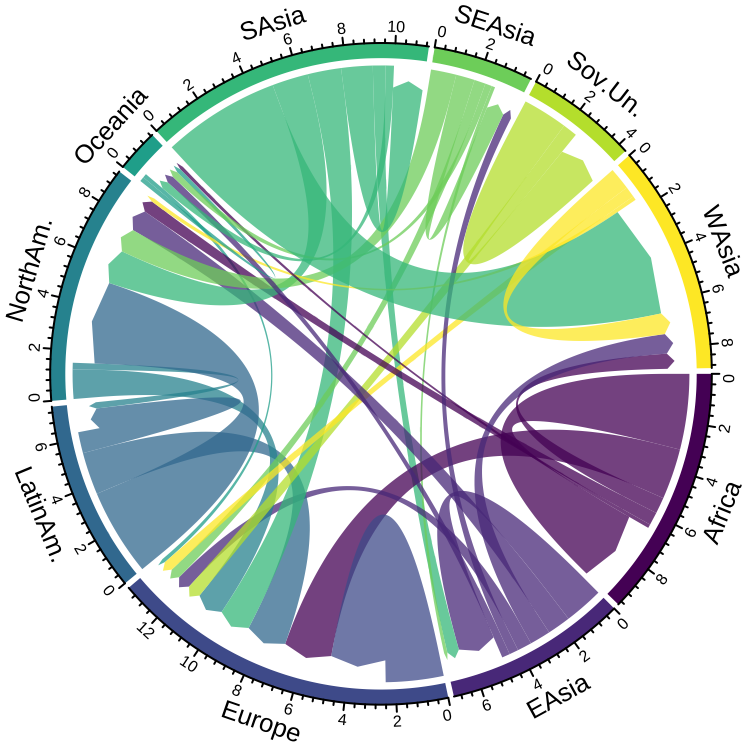
The original chord diagram:

```
library(circlize)
```

```
## ========================================
## circlize version 0.4.16
## CRAN page: https://cran.r-project.org/package=circlize
## Github page: https://github.com/jokergoo/circlize
## Documentation: https://jokergoo.github.io/circlize_book/book/
##
## If you use it in published research, please cite:
## Gu, Z. circlize implements and enhances circular visualization
##   in R. Bioinformatics 2014.
##
## This message can be suppressed by:
##   suppressPackageStartupMessages(library(circlize))
## ========================================

##
## Attaching package: 'circlize'
```

```
## The following object is masked from 'package:igraph':
##
##     degree
```

```
# Original chord diagram
chordDiagram(
  x = migration_long,
  grid.col = migration_color,
  transparency = 0.25,
  directional = 1,
  direction.type = c("arrows", "diffHeight"),
  diffHeight  = -0.04,
  #  annotationTrack = "grid",
  #  annotationTrackHeight = c(0.05, 0.1),
  link.arr.type = "big.arrow",
  link.sort = TRUE)
```



In comparison to the chord diagram, the alluvial visualisation is arguably easier to understand for a general population, whereas the former would require some extent of visualisation knowledge to dissect, especially without useful annotations such as a title. Therefore, the alluvial diagram would likely be directed towards average individuals from these continents who may be wanting to migrate to another continent, while the chord diagram is aimed towards data scientists working with governments from countries within these continents to analyse modern immigration patterns and taking action (e.g. creating/improving embassies and other diplomatic and consular efforts between countries).

3. How have your aesthetic choices improved an audiences' understanding of the visualisation, and helped to communicate within this context?

Encoded annotations such as a title as well as displaying a legend will improve an audiences' understanding of the visualisation much better than simply displaying the visualisation as is. This is supported by the labels of the continents in the strata within the alluvial. As the visualisation is reader-driven, while the labels themselves is sufficient in understanding *where people first originated* before migrating when reading the visualisation left-to-right, the legend further helps the audience understand this when reading right-to-left as well (for each continent, *where have people migrated from?*). Colour is used to encode the continents to better discern them from each other as opposed to a monochrome colourmap which does not facilitate *comparison* as is the main action of this visualisation. Size is also a preattentive attribute, as the bigger the size of each 'line' in the alluvial encodes a larger amount of people migrating from one specific continent to another.

4. Describe (but do not implement) a different visualisation that can communicate the same message with the same variables in the data. How might this other visualisation be more effective than this alluvial visualisation?

A network visualisation can communicate the same message with the same variables. The thickness of edge lines within a network denotes the weight of the relationship vertices (continents) much like in the alluvial. Arrows in the network would further encode that the flow between continent vertices are *directed* rather than *undirected*. Label data from the strata in the alluvial can be reused to label the continent vertices in the network. Colour can be reused to denote the continent vertices and/or edges between continents. Networks may be more effective in having a clear organisation of the visualisation as compared to the alluvial. In the alluvial, some strata are tiny and hard to discern between larger strata which may negatively impact decoding where people are migration to/from (e.g. Oceania). In a network, the vertices may be clearer, though size can again be reused to encode amounts of people migrating from a continent. A network can use strategies such as minimising edge crossing, having uniform edge lengths, preventing overlaps, and symmetry to create a visually pleasing visualisation while still having effective encoding and having the same message as the alluvial.

5. Describe (but do not implement) a different visualisation that can communicate the same message, which includes additional information/data that can expand the message. You do not need to find this data, but please describe it. How might this other visualisation be more or less effective than this alluvial visualisation?

Using further data about the countries within these continents, especially countries that amount to a higher proportion of migrating people within said continent, we can again use a network to separate the data into communities. A community in this instance are countries from the same continent. The size of the country labels can further denote proportions of migration within each continent community. This can have the same message as the alluvial diagram (migration between continents) but display it with a community of countries rather than one continent variable. This further aids the 'who' context of the visualisation as it caters better towards the average individual from these specific countries (rather than continents) in seeing where their fellow countrymen have migrated to. The advantages of a network opposed to an alluvial mentioned above also apply here.

Question 11

(**Marks 8**) 1. Consider the cattle network from this week's workshop. Create a network visualisation where vertex size is used to represent the size of the farm in some way. Your plot will be assessed for accuracy and effectiveness, so design your visualisation and use pre-attentive attributes carefully.

```
# Load data
attr1 <- read.csv("data_w7/Farms/attr_farms.csv", stringsAsFactors = F) # load attr.farms.Rdata
edges <- read.csv("data_w7/Farms/Edgelist_farms.csv", stringsAsFactors = F) # load the edgelist_farms.R
```

```r
# Inspect data
head(attr1)
```

```
##   X farm.id      type size    long      lat state
## 1 1       1 Fattening  354 6194009 567599.8     0
## 2 2       3     Dairy  203 6181087 559118.7     0
## 3 3       2 Fattening  250 6196090 562336.4     0
## 4 4       5     Dairy  994 6207398 474646.4     0
## 5 5      15     Dairy  544 6221094 413707.2     0
## 6 6      13     Dairy   87 6205884 421178.5     0
```

```r
head(edges)
```

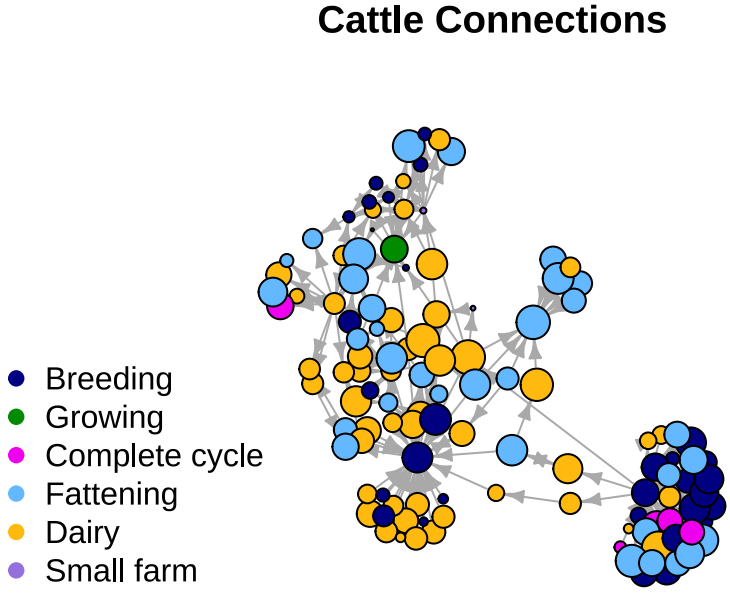
```
##   Origin Dest       Date breeding.cows steers heifers calves batch.size
## 1      1   25 2008-12-10             0     43       0      0         43
## 2      1   25 2008-12-10             0     41       0      0         41
## 3      3   88 2009-07-30             0      0       0     24         24
## 4      2   25 2008-12-10             0     42       0      0         42
## 5      5   13 2008-12-20             0      0      12      0         12
## 6      5    7 2008-12-20             0      0       8      0          8
```

```r
# Create `igraph` object
net_edges <- graph.data.frame(edges, directed=T)
net_edges
```

```
## IGRAPH 3d17bc4 DN-- 120 356 --
## + attr: name (v/c), Date (e/c), breeding.cows (e/n), steers (e/n),
## | heifers (e/n), calves (e/n), batch.size (e/n)
## + edges from 3d17bc4 (vertex names):
##  [1] 1 ->25  1 ->25  3 ->88  2 ->25  5 ->13  5 ->7   15->18  15->18  15->18
## [10] 15->19  15->19  15->7   15->7   15->88  15->88  15->106 15->26  13->4
## [19] 13->14  13->14  13->17  13->20  13->8   13->10  13->7   13->88  13->93
## [28] 23->22  23->21  18->19  18->7   18->7   18->88  18->88  18->106 18->106
## [37] 9 ->13  9 ->10  9 ->10  9 ->10  9 ->10  9 ->10  9 ->10  9 ->10  9 ->10
## [46] 9 ->10  9 ->10  9 ->10  4 ->29  4 ->29  4 ->29  11->13  11->13  35->116
## [55] 12->13  12->13  12->10  12->10  19->106 14->11  14->8   17->13  17->13
## + ... omitted several edges
```

```r
# Add production to each vertex
V(net_edges)$type <- as.character(attr1$type[match(V(net_edges)$name,attr1$farm.id)])

# Add herd size
V(net_edges)$farm.size <- attr1$size[match(V(net_edges)$name,attr1$farm.id)]

# Custom colours by creating an attribute for vertex colour
V(net_edges)$color <- V(net_edges)$type

V(net_edges)$color <- gsub("Fattening","steelblue1",V(net_edges)$color)
V(net_edges)$color <- gsub("Dairy","darkgoldenrod1",V(net_edges)$color)

V(net_edges)$color <- gsub("Small farm","mediumpurple",V(net_edges)$color)
```

```r
V(net_edges)$color <- gsub(pattern="Breeding",replacement="navy",x=V(net_edges)$color)
V(net_edges)$color <- gsub(pattern="Complete cycle",replacement="magenta2",x=V(net_edges)$color)
V(net_edges)$color <- gsub(pattern="Growing",replacement="green4",x=V(net_edges)$color)

# Plot network
plot.igraph(simplify(net_edges),
            layout=layout.fruchterman.reingold,
            vertex.label=NA,
            vertex.color=V(net_edges)$color,
            vertex.size=log(V(net_edges)$farm.size)*2, # https://www.digitalocean.com/community/tutoria
            edge.arrow.size=.5)

# Make a legend
legend("bottomleft",
       legend=c("Breeding","Growing","Complete cycle","Fattening","Dairy","Small farm"),
       col=c("navy","green4","magenta2","steelblue1","darkgoldenrod1","mediumpurple"),
       pch=19,cex=1,bty="n")

# Add title and subtitle
title(main = "Cattle Connections",
      sub = "The flow of different types of cattle between different categories of farms. Vertex size ba
```



w of different types of cattle between different categories of farms. Vertex size based on s

2. Do you notice any patterns about where large farms are located within the network?

The Fruchterman-Reingold force-directed layout, as is used in the network, is a layout in which vertices that

share more connections are closer to each other in the resulting graph. On the network, we can see that most of the relatively larger farms usually have more connections to other farms, thus forming clusters within the graph. We can see that most of the farms in the main cluster are of a consistently larger variety than the farms elsewhere. The other farms also have some large farms but not to the same consistency as the farms in the main cluster. Smaller farm vertices seem to be less connected to other farms and thus spatially far apart from each other. Dairy and fattening seem to generally have consistently larger farm sizes compared to other farm types. Therefore, farms that have more connections with other farms and thus form clusters within the network, as well as dairy and fattening farms in particular, usually are the relatively larger farms in the cattle movement data.

3. Describe the message being communicated with this visualisation, and who the audience is.

The message is explicitly stated in the subtitle down the bottom of the visualisation which is for the audience to examine and learn about cattle movement data between farms. A likely audience for this visualisation are cattle farmers who, with this visualisation, can better plan their farms based on the relationships between different types of farms and relative sizes of said farms presented in the dataset.

4. How have your aesthetic choices improved an audiences' understanding of the visualisation, and helped to communicate within this context?

Colour is used a preattentive attribute to denote the different type of farms in the visualisation, which is further encoded in the legend on the bottom left. Size is also used as a preattentive attribute to encode the size of the farms and presents an extra dimension/variable to measure relationships with within the visualisation. An appropriate, relevant title and subtitle is included to support the audiences' understanding of the visualisation. Arrows instead of lines are used to encode movement to-and-from farms rather than just a simple, arbitrary relationship between each farm vertex. Lastly, positioning of the vertices is encoded using the Fruchterman-Reingold force-directed layout where farms closer to each other have more connections to each other.

5. Suggest (but do not implement) a different visualisation that can communicate the same message – either with the same variables in the data, or suggest additional information that would be useful. How might this other visualisation be more or less effective than this network visualisation?

A chord diagram can be used to visualise the movement of the cattle. Different farm types can be used as the nodes of the diagram. Size/thickness of the chords can denote the proportions of specific farm types that have cattle move to another specific farm type. This can simplify the network visualisation by boiling it down to simply proportions of farm types that have cattle move to other farm types without showing every single farm, different farm sizes, etc. A less complicated visualisation may benefit cattle farmers who likely do not have the visualisation skills to decode the many complicated aspects of the network diagram.

Week 8

Question 12

(*Marks 6*) Create a map using the data contained in `rnaturalearth` with a map extent that is smaller than the entire world (e.g. only shows one continent, or is bounded in some other way). Colour polygons according to some aspect of the data. In the workshop you created the entire world coloured by the square root of the total population, so you cannot visualise population count for this question. Then, represent that same data using at least one other plot type that we learned earlier in the unit, very briefly (one sentence) identify a target audience and an intended message for each of your two visualisations. Note whether they serve the same purpose, or work best when presented together.

```r
# Load libraries
library(cowplot)
```

```
##
## Attaching package: 'cowplot'

## The following object is masked from 'package:ggthemes':
##
##     theme_map

## The following object is masked from 'package:patchwork':
##
##     align_plots

## The following object is masked from 'package:lubridate':
##
##     stamp
```

```r
library(ggspatial)
library(rnaturalearth)
library(rnaturalearthdata)
```

```
##
## Attaching package: 'rnaturalearthdata'

## The following object is masked from 'package:rnaturalearth':
##
##     countries110
```

```r
library(sf)
```

```
## Linking to GEOS 3.11.2, GDAL 3.8.2, PROJ 9.3.1; sf_use_s2() is TRUE
```

```r
library(maps)
```

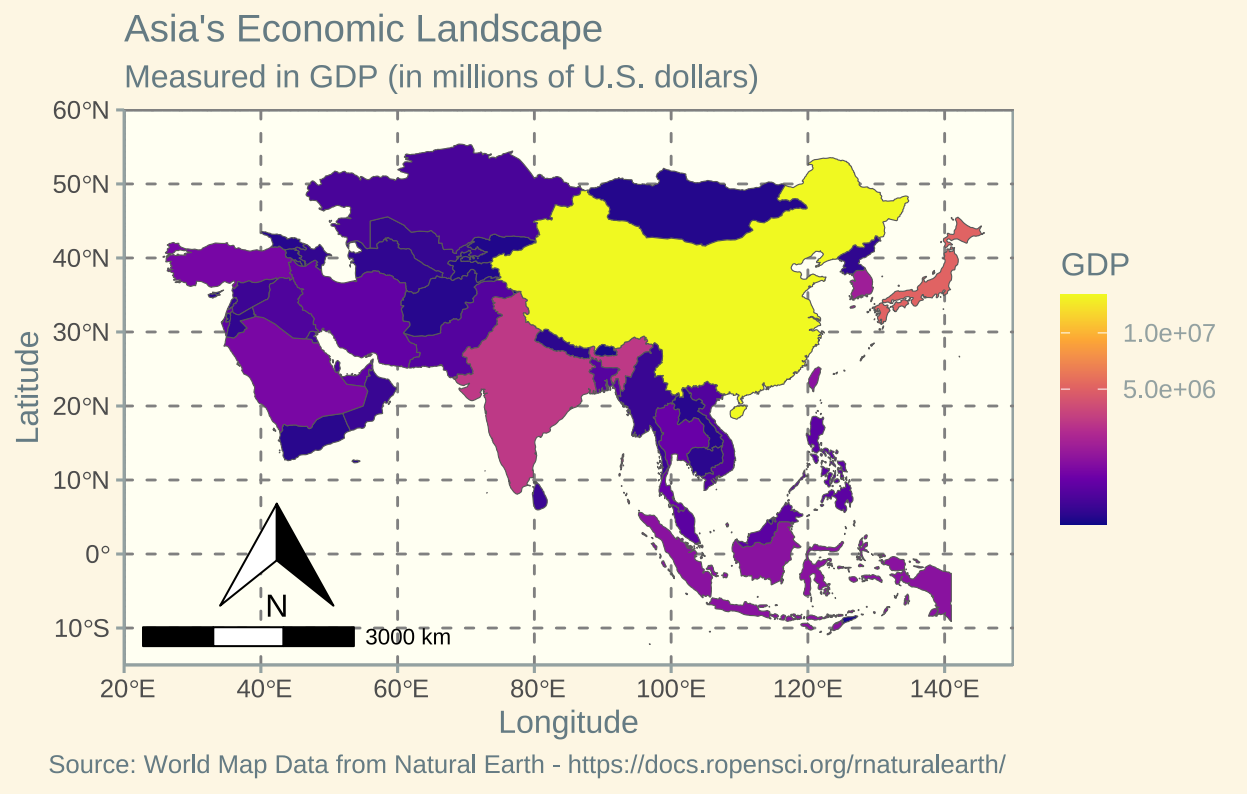
```
##
## Attaching package: 'maps'

## The following object is masked from 'package:viridis':
##
##     unemp

## The following object is masked from 'package:purrr':
##
##     map
```

```r
# Pull world data (Asia)
world_asia <- ne_countries(scale = "medium", returnclass = "sf", continent = "asia")
class(world_asia) # should be 'sf'
```

```
## [1] "sf"         "data.frame"
```

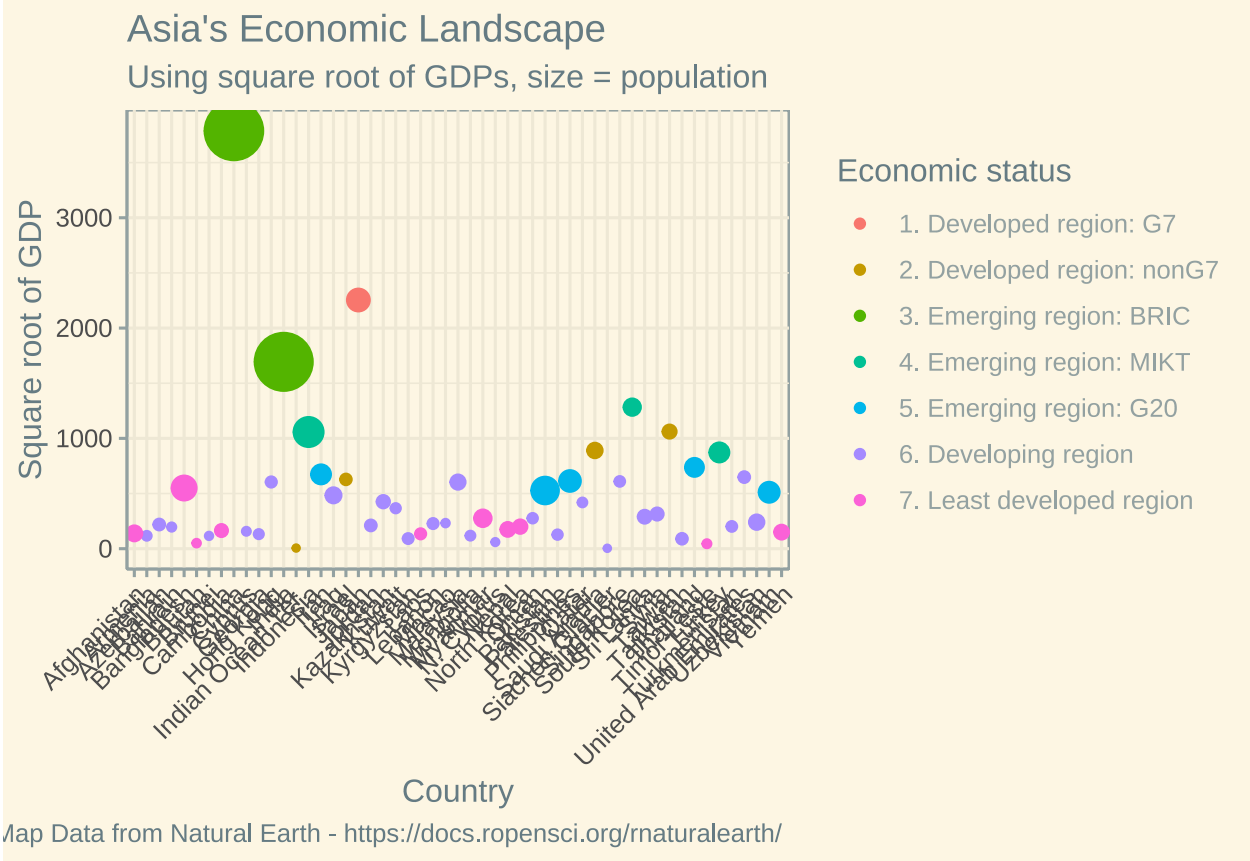
```r
# Pull world labels (for annotating)
asia_points <- st_centroid(world_asia)
asia_points <- cbind(world_asia, st_coordinates(st_centroid(world_asia$geometry)))

# Plot map of Asia's economic landscape (using gdp_md)
ggplot(data = world_asia) +
  geom_sf(aes(fill = gdp_md)) +
  scale_fill_viridis_c(option = "plasma", trans = "sqrt") +
  ggspatial::annotation_scale(location = "bl", width_hint = 0.25) +
  ggspatial::annotation_north_arrow(location = "bl", which_north = "true", pad_x = unit(0.5, "in"), pad
  coord_sf(xlim = c(20, 150), ylim = c(-15, 60), expand = FALSE) +
  ggtitle("Asia's Economic Landscape", subtitle = "Measured in GDP (in millions of U.S. dollars)") +
  xlab("Longitude") +
  ylab("Latitude") +
  theme_solarized() +
  theme(panel.grid.major = element_line(color = gray(.5), linetype = "dashed", size = 0.5),
        panel.background = element_rect(fill = "#fffff2")) +
  labs(fill = "GDP", caption = "Source: World Map Data from Natural Earth - https://docs.ropensci.org/r
```

```
## Scale on map varies by more than 10%, scale bar may be inaccurate
```




```r
# Pull all world data
world <- ne_countries(scale = "medium", returnclass = "sf")
```

```r
# Plot bubble plot alternative

ggplot(world_asia, aes(x=name, y=sqrt(gdp_md), size = pop_est, color = economy)) +
  geom_point() +
  scale_size(range = c(1,10)) +
  labs(title = "Asia's Economic Landscape",
       subtitle = "Using square root of GDPs, size = population",
       x = "Country",
       y = "Square root of GDP",
       color = "Economic status",
       caption = "Source: World Map Data from Natural Earth - https://docs.ropensci.org/rnaturalearth/
  theme_solarized() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  guides(size = FALSE)
```

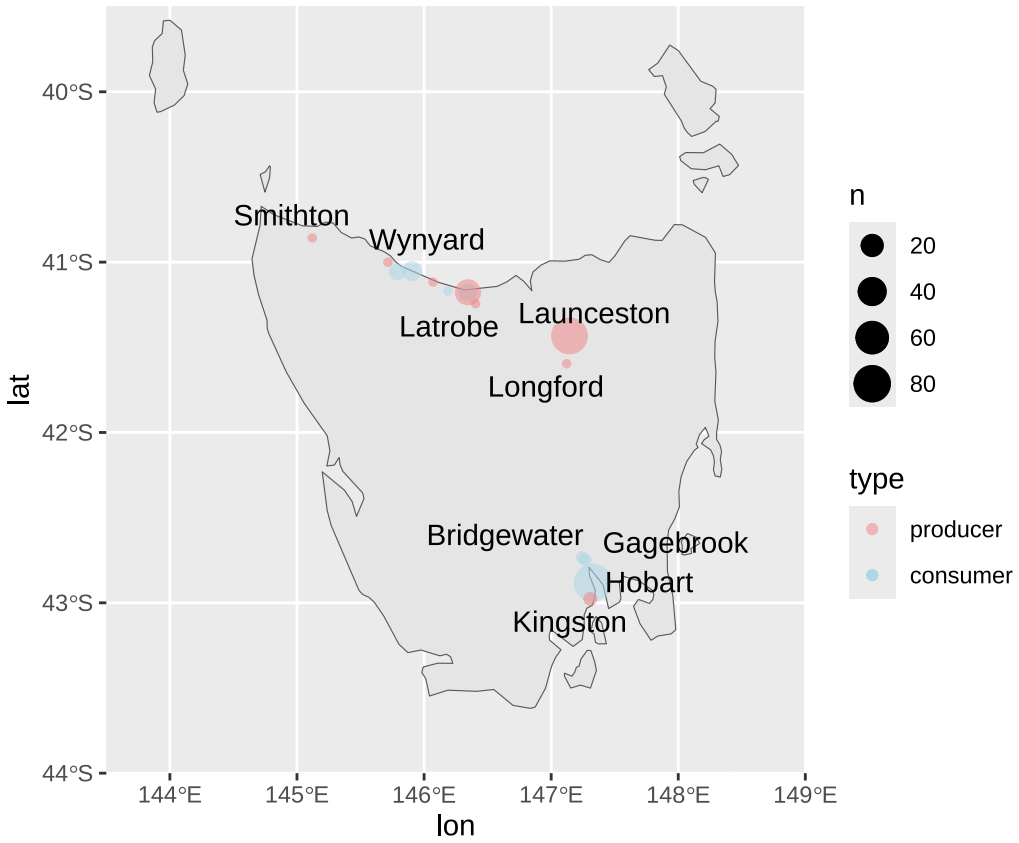


The cartogram map visualisation is aimed towards the Asian population in learning about their economic landscape and determining quick comparisons between their country and other countries. As scatter and bubble plots can encode an actual data value, the second visualisation is aimed more towards international economic analysts aiming to analyse the economic landscape of Asia. These two visualisations could work well in tandem together, where the cartogram can be presented first to determine quick comparisons between countries, then the bubble plot presented second to explicitly encode GDP values and expand upon the first visualisation.

Question 13

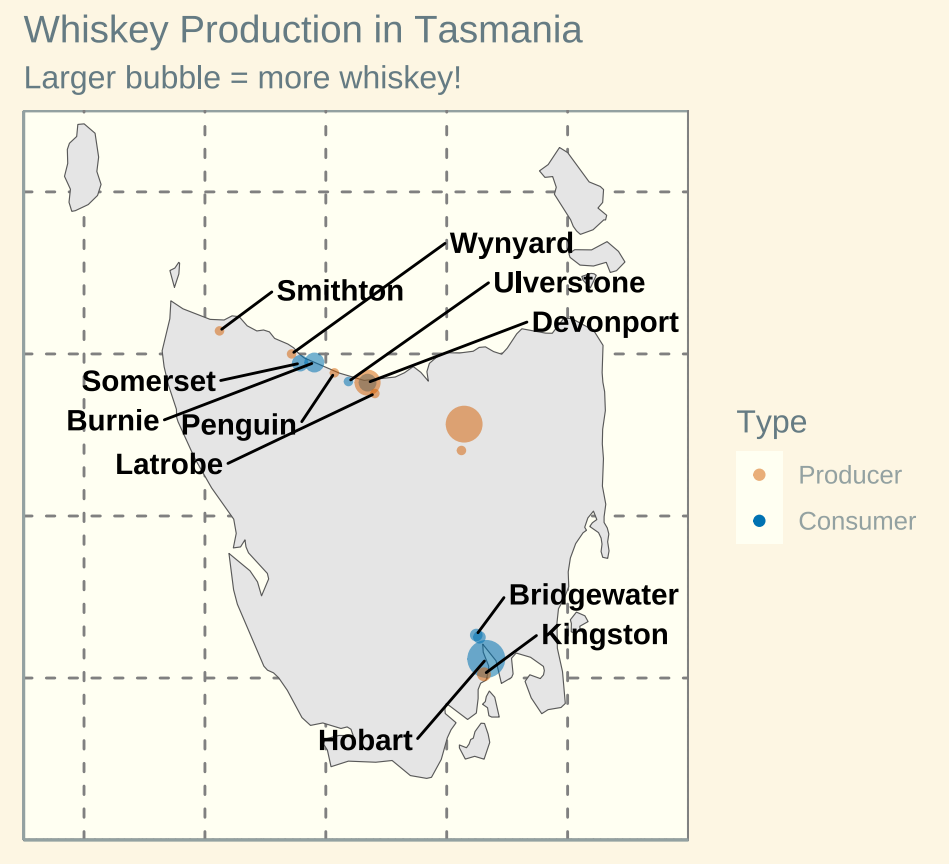
(*Marks 14*) A. Reproduce the following map using these locations and transport datasets Download these locations and transport datasets. In the dataset, one row represents the sale of one barrel of whiskey (this is made up data!). You'll need to draw from your past experiences in this unit wrangling data and changing elements of the ggplot outputs. Remember the textbook you have used in earlier workshops - this will be very valuable for manipulating the data. Also explore the help file for the count() function, and the help file for 'Efficiently bind multiple data frames by row and column' using the package dplyr.

```
# Read in CSV data
tas_locations <- read.csv("whiskey_data/tas_locations.csv")
whiskey_sales_tasmania <- read.csv("whiskey_data/whiskey_sales_tasmania.csv")

# Count how many times a place has produced
count_produced <- whiskey_sales_tasmania %>% count(producer, name="produced_count") %>% rename(place = p

# Count how many times a place has consumed
count_consumed <- whiskey_sales_tasmania %>% count(consumer, name="consumed_count") %>% rename(place = c

# Combine counts to single dataframe
combined_count <- bind_rows(count_produced, count_consumed) %>%
  group_by(place) %>%
  summarise(produced_count = sum(produced_count, na.rm = TRUE),
            consumed_count = sum(consumed_count, na.rm = TRUE)) %>%
  ungroup()

# Arrange both locations and counts by place (for cbind)
tas_locations <- tas_locations %>% arrange(place)
combined_count <- combined_count %>% arrange(place)

# Combine locations and count
tas_locations_count <- cbind(combined_count, tas_locations[, c("lat", "lon")])

# Remove zeroes (change to NA - won't display on plot)
tas_locations_count <- tas_locations_count %>%
  mutate(
    produced_count = ifelse(produced_count == 0, NA, produced_count),
    consumed_count = ifelse(consumed_count == 0, NA, consumed_count))

# Plot Tasmania map and produced and consumed
ggplot(data = world) +
  geom_sf() +
  geom_point(data = tas_locations_count, aes(x = lon, y = lat, size = consumed_count, color = "lightcora
  geom_point(data = tas_locations_count, aes(x = lon, y = lat, size = produced_count, color = "lightblu
  geom_text_repel(data = tas_locations, aes(x = lon, y = lat, label = place)) +
  scale_color_manual(name = "type", values = c("lightcoral", "lightblue"), labels = c("producer", "consu
  coord_sf(xlim = c(143.5, 149), ylim = c(-44, -39.5), expand = FALSE) +
  labs(size = "n", color = "type") +
  guides(alpha = FALSE)
```



B. Improve upon this visualisation to be more effective at communicating a message. Identify the audience and the message, and justify your visualisation choices within that context. Include both visualisations (the one from part (a) and part (b))

```
ggplot(data = world) +
  geom_sf() +
  geom_point(data = tas_locations_count, aes(x = lon, y = lat, size = consumed_count, color = "#D55E00"
  geom_point(data = tas_locations_count, aes(x = lon, y = lat, size = produced_count, color = "#0072B2"
  geom_text_repel(data = tas_locations, aes(x = lon, y = lat, label = place), fontface = "bold", nudge_
  scale_color_manual(name = "Type", values = c("#D55E00", "#0072B2"), labels = c("Producer", "Consumer"
  coord_sf(xlim = c(143.5, 149), ylim = c(-44, -39.5), expand = FALSE) +
  labs(title = "Whiskey Production in Tasmania",
       subtitle = "Larger bubble = more whiskey!",
       size = "Amount") +
  theme_solarized() +
  theme(panel.grid.major = element_line(color = gray(0.5), linetype = "dashed", size = 0.5),
        panel.background = element_rect(fill = "#fffff2"),
        axis.text.x = element_blank(),
        axis.text.y = element_blank(),
        axis.ticks = element_blank(),
        axis.title.x = element_blank(),
        axis.title.y = element_blank()) +
  guides(alpha = FALSE, size = FALSE)
```



This visualisation is for market geo-analysts working in the whiskey industry to analyse which cities in Tasmania produce and/or consume the most amount of whiskey. This new visualisation adds an active title and descriptive subtitle. The `n` legend has been removed as encoding explicit values in a map visualisation is largely ineffective, and darting between the bubble and the legend would not help increase effectiveness or understanding. The colours have been changed to pop out better on the map. The city labels have been reworked to stand out better but also added lines to specifically encode the location of each city on the map. Both axes labels, values, and ticks have been removed as latitude and longitude is insignificant in this visualisation. Visualising this within a map may help these geo-analysts better plan transport routes between their producers/consumers.

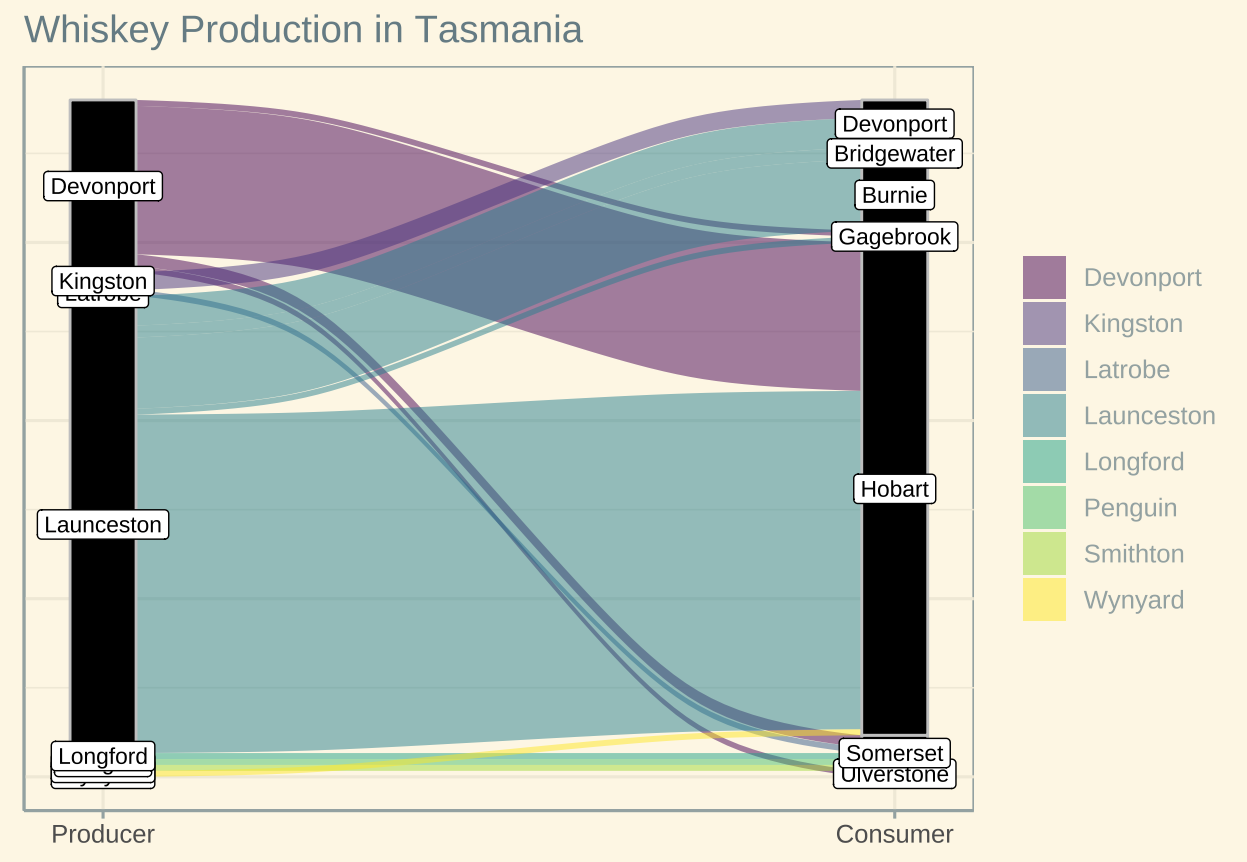
C. Visualise this data using another type of visualisation style we have examined in previous weeks. Explain how the audience is different between your maps in part (b) and (c).

```
whiskey_alluvial <- whiskey_sales_tasmania %>%
  group_by(producer, consumer) %>%
  summarise(count = n()) %>%
  ungroup()
```

```
## `summarise()` has grouped output by 'producer'. You can override using the
## `.groups` argument.
```

```
whiskey_color <- viridis(length(unique(whiskey_alluvial$producer)))

ggplot(whiskey_alluvial, aes(y=count, axis1=producer, axis2=consumer)) +
  geom_alluvium(aes(fill=producer), width=1/12) +
  geom_stratum(width=1/12, fill="black", color="grey", stratum_padding=2) +
```

```
geom_label(stat="stratum", aes(label=after_stat(stratum)), size = 3) +
scale_x_discrete(limits=c("Producer", "Consumer"), expand=c(.05,.05)) +
scale_fill_manual(values=whiskey_color) +
labs(title = "Whiskey Production in Tasmania",
     fill = "Producer",
     y = NULL) +
theme_solarized() +
theme(axis.text.y = element_blank(),
      axis.ticks.y = element_blank(),
      legend.title = element_blank())
```



This alluvial visualisation will cater towards market data and business analysts in general rather than specifically geo-analysts who may be using the data to plan transport routes between producers and consumers. Using this visualisation, these analysts can identify the trends in main producers and consumers and plan general business decisions (e.g. marketing or operations) around this knowledge.

Week 9

In the following questions, you will download data from this University of Minnesota websiteLinks to an external site. which shows the spatial distribution of various invasive plants in the US state of Minnesota. Download the .tiff files (invasive_plant_rasters_2019.zip). Unzip those files to your computer and move the folder into your working directory. It will contain 14 raster files and one text file which describes the data.

Question 14

(*Marks 8*) Create two maps showing two different invasive plants – one which has a larger distribution than the other. Create numeric breaks in the colours that make sense given the distribution of the data. Customise the visualisation based on the workshop in Week 9, and justify how your choices (including the numeric breaks) have made your visualisations more effective.

```
library(raster)

alliaria <- raster("invasive_plant_rasters_2019/sumrast_allassumptions.avg_Alliaria petiolata.tif")

summary(alliaria)
```

```
##         sumrast_allassumptions.avg_Alliaria.petiolata
## Min.                                       1.303218e-03
## 1st Qu.                                    2.338973e-02
## Median                                     1.054104e-01
## 3rd Qu.                                    3.546943e-01
## Max.                                       9.743232e-01
## NA's                                       2.625550e+05
```

```
alliaria
```

```
## class      : RasterLayer
## dimensions : 704, 887, 624448  (nrow, ncol, ncell)
## resolution : 655, 925  (x, y)
## extent     : -91806.12, 489178.9, 2278944, 2930144  (xmin, xmax, ymin, ymax)
## crs        : +proj=aea +lat_0=23 +lon_0=-96 +lat_1=29.5 +lat_2=45.5 +x_0=0 +y_0=0 +ellps=GRS80 +towgs
## source     : sumrast_allassumptions.avg_Alliaria petiolata.tif
## names      : sumrast_allassumptions.avg_Alliaria.petiolata
## values     : 0.00129215, 0.9783568  (min, max)
```

```
# setMinMax(alliaria)

alliaria@crs
```

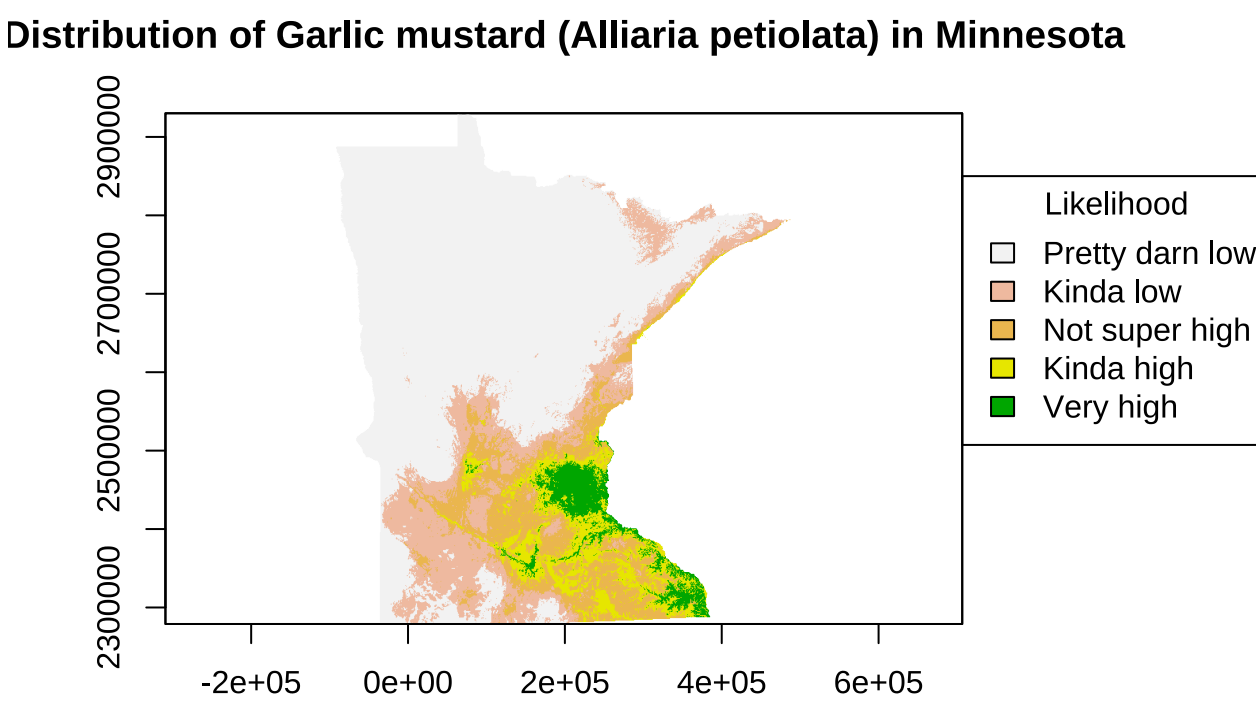
```
## Coordinate Reference System:
## Deprecated Proj.4 representation: NA
```

```
alliaria@extent
```

```
## class      : Extent
## xmin       : -91806.12
## xmax       : 489178.9
## ymin       : 2278944
## ymax       : 2930144
```

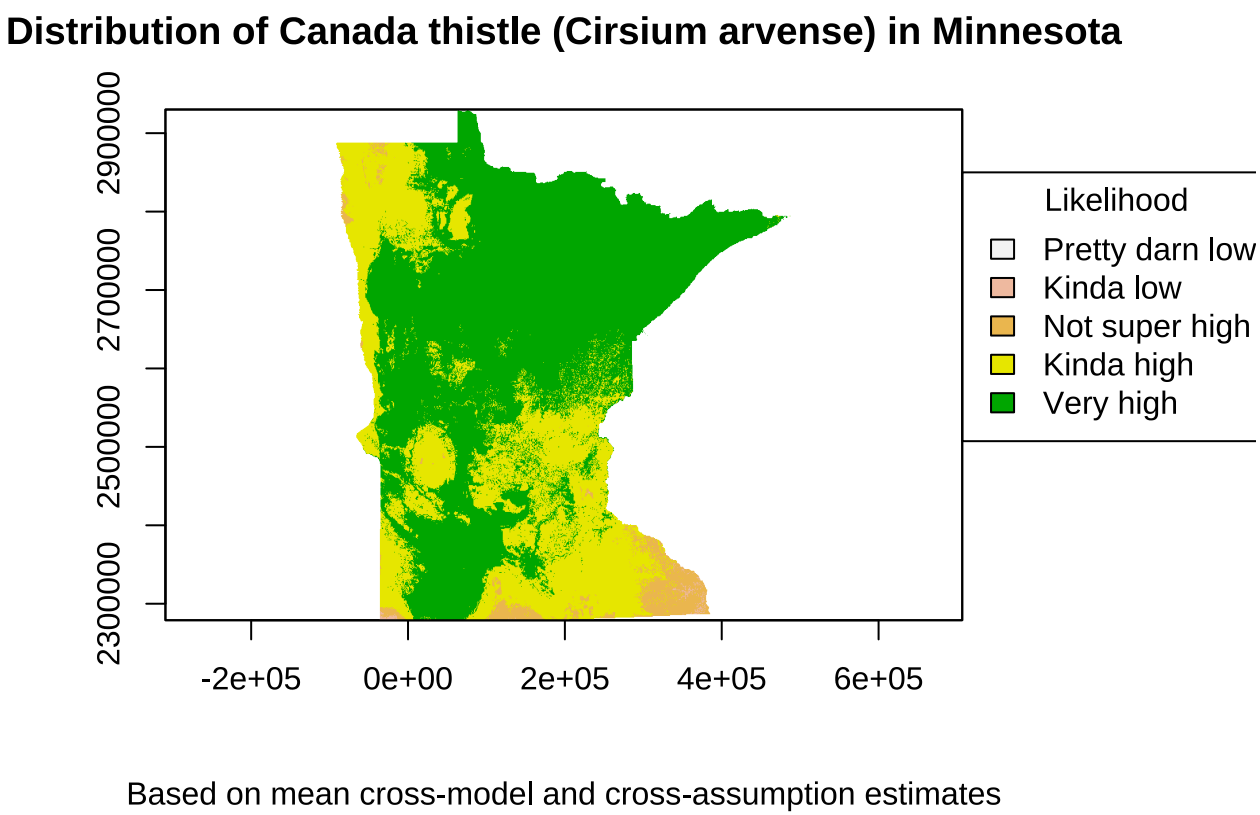
```
col <- rev(terrain.colors(5))
brk = c(0, 0.2, 0.4, 0.6, 0.8, 1)

# First, expand right side of the image by clipping the rectangle to make
```

```r
# room for the legend and then turn xpd off
par(xpd = FALSE, mar=c(5.1, 4.1, 4.1, 4.5))

# Second, plot with no legend
plot(alliaria,
     col=col,
     breaks=brk,
     main = "Distribution of Garlic mustard (Alliaria petiolata) in Minnesota",
     sub = "Based on mean cross-model and cross-assumption estimates",
     legend = FALSE)

# Third, turn xpd back on to force the legend to fit next to the plot.
par(xpd = TRUE)

# Fourth, add a legend outside of the plot
legend(par()$usr[2], 2850000,
       legend = c("Pretty darn low", "Kinda low", "Not super high", "Kinda high", "Very high"),
       fill = col,
       title = "Likelihood")
```




```r
cirsium <- raster("invasive_plant_rasters_2019/sumrast_allassumptions.avg_Cirsium arvense.tif")

# First, expand right side of the image by clipping the rectangle to make
# room for the legend and then turn xpd off
par(xpd = FALSE, mar=c(5.1, 4.1, 4.1, 4.5))
```

```r
# Second, plot with no legend
plot(cirsium,
     col=col,
     breaks=brk,
     main = "Distribution of Canada thistle (Cirsium arvense) in Minnesota",
     sub = "Based on mean cross-model and cross-assumption estimates",
     legend = FALSE)

# Third, turn xpd back on to force the legend to fit next to the plot.
par(xpd = TRUE)

# Fourth, add a legend outside of the plot
legend(par()$usr[2], 2850000,
       legend = c("Pretty darn low", "Kinda low", "Not super high", "Kinda high", "Very high"),
       fill = col,
       title = "Likelihood")
```



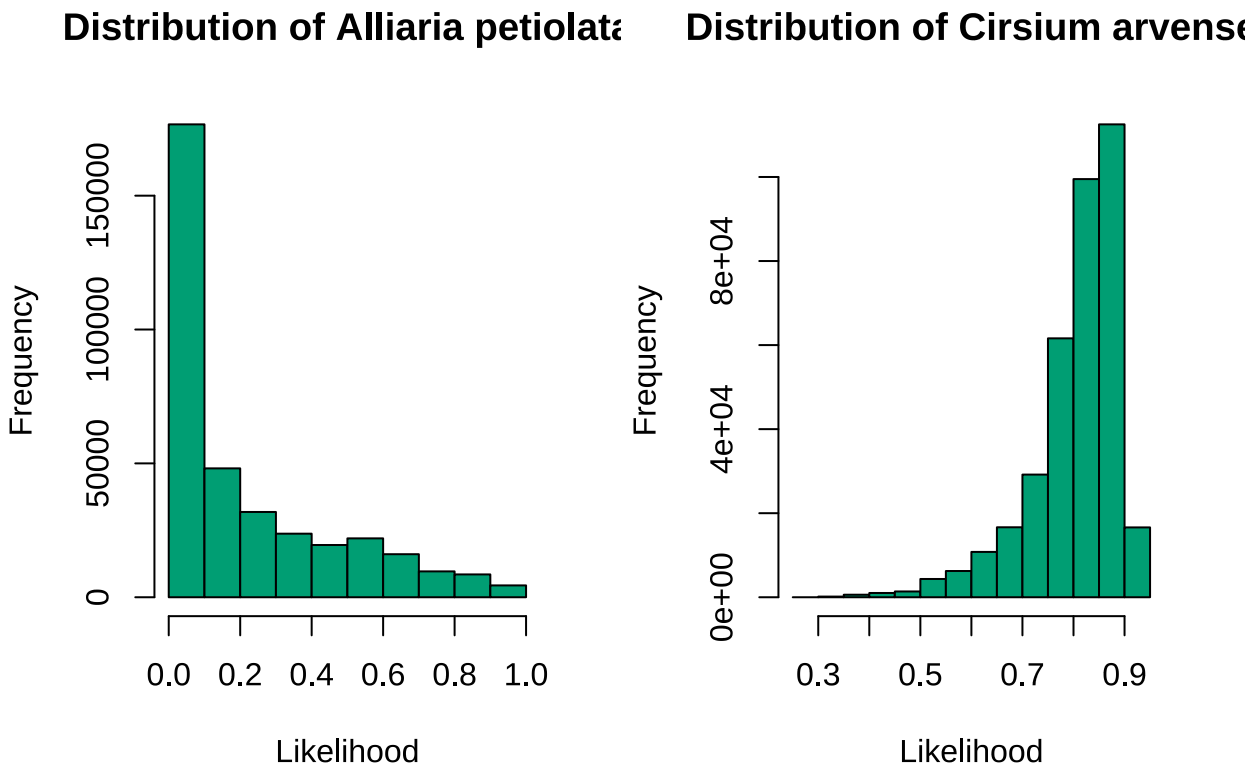
These visualisations include a suitable, descriptive title, as well as a relevant subtitle describing how these likelihood/potential distributions were calculated to explicitly communicate the message of the spatial visualisation in analysing the likelihood of a species of plant growing around the state of Minnesota. A suitable colour palette was used that is familiar to audiences in an environmental context. It is very self-explanatory for the audience that 'green' sections in the spatial visualisation indicates there is a high likelihood that this species grows there due to this preattentive attribute. Numeric breaks were included for the legend which better emphasises categorical likelihood groups (instead of fully continuous). These breaks were defined at 0%, 25%, 50%, 75%, and 100%. The legend texts were modified to better define categorical rather than numerical likelihoods. This effectively encodes a range of values which is effective

for environmental scientists as a target audience yet still understandable enough for the layman.

Question 15

(*Marks 5*) You are a data scientist working with the Minnesota Department of Natural Resources who are trying to decide which invasive plant to control first. They are seeking your recommendation, which must be supported by data. You are armed with your maps from the previous question. Create one other, non-spatial, simple visualisations which can help illustrate your argument for which of the two invasive plants is the biggest problem (this might have two panels – one for each of the two species). Write a 3 sentence argument that you'll make to the Department to convince them of your recommendation, referring to your maps and your figures from this question.

```
par(mfrow = c(1,2))

hist(alliaria,
     main = "Distribution of Alliaria petiolata",
     col = "#009E73",
     maxpixels = 22000000,
     breaks = 10,
     xlab = "Likelihood")

hist(cirsium,
     main = "Distribution of Cirsium arvense",
     col = "#009E73",
     maxpixels = 22000000,
     breaks = 10,
     xlab = "Likelihood")
```



As is evident by the map, the Cirsium arvense species is a relatively highly invasive plant species which has spread around Minnesota. They can highly likely be found in a large majority of the state, while the Alliaria petiola species in turn can only be found in a few concentrated areas as well as scattered scarcely in some counties. Turning to the histograms, it is evident that the Cirsium arvense has several orders of magnitude higher frequency in comparison to the Alliaria petiola. A majority of the bars in the Cirsium arvense histogram lie on the right side, indicating it is very likely to be found in more areas at a high frequency. Therefore, the Cirsium arvense species is the relatively more invasive plant, and must be controlled first.