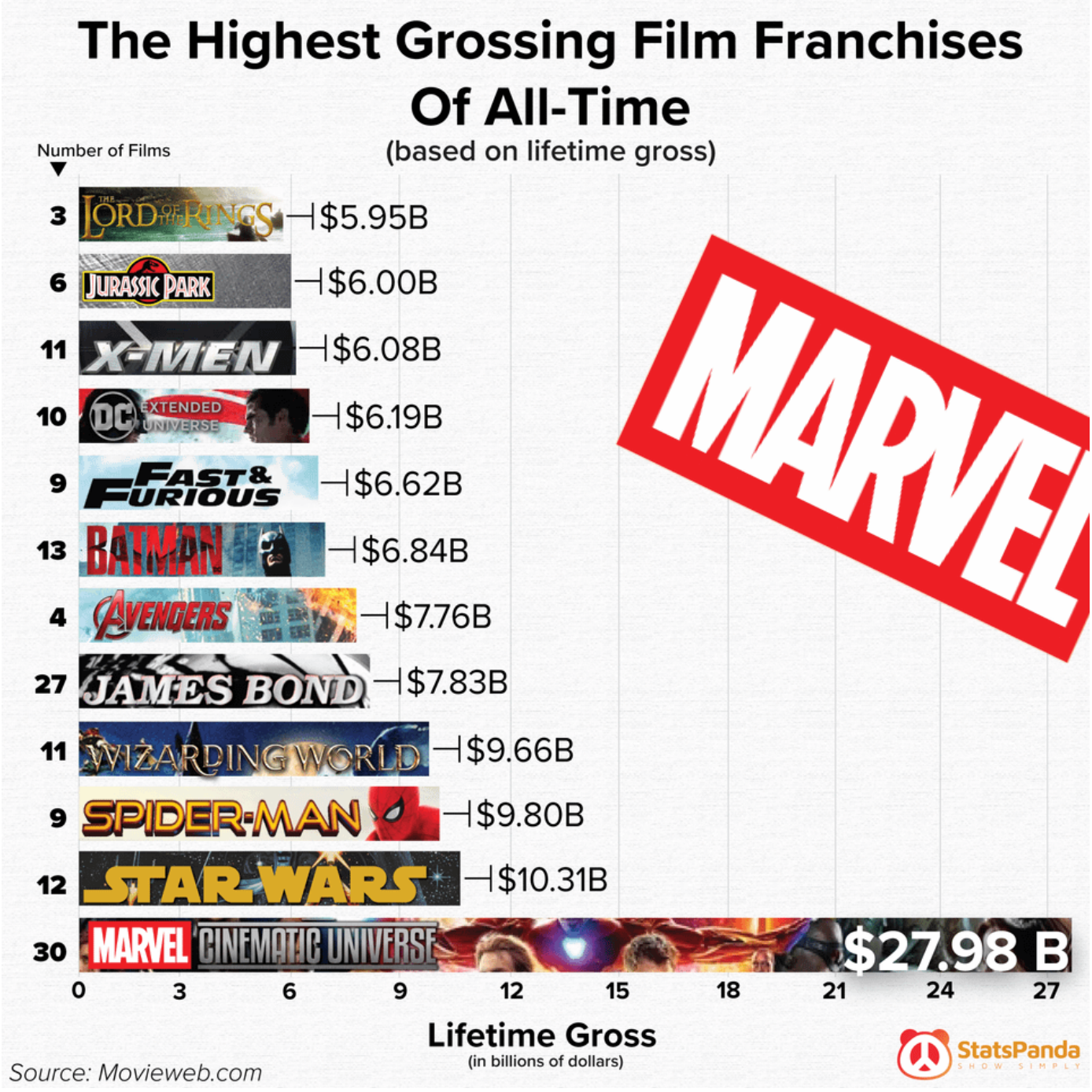
MXB262 Portfolio [n11198885]

2023-03-14

Week 3 Portfolio Questions–

Visualisation Title and Link:
https://www.reddit.com/r/dataisbeautiful/comments/11nsuwd/oc_the_highest_grossing_film_franchises_of_alltime/ (https://www.reddit.com/r/dataisbeautiful/comments/11nsuwd/oc_the_highest_grossing_film_franchises_of_alltime/)



The Highest Grossing Film Franchises Of All Time

1.(Marks: 2) Use the principles of story-telling practiced in the Week 3 practical to create a story board.

A. Describing the context

Who is the audience or audiences?:

- Business manager, investors, researcheres, general public(whos interested in the film)

What is the action the visualisation is aiming for? Consider each audience here:

- The action that the visualisation is aiming for may vary depending on the audience.

```
- Business manager : The visualisation aims to provide a quick
overview of the top companies in the world by market capitalization.

- Investors : The visualisation aims to provide information on the
market capitalization of the top companies in the world.

- Researchers : The visualisation aims to provide a clear and
concise overview of the top companies in the world by
market capitalization.

- General Public : The visualisation aims to provide a snapshot
of the biggest companies in the world by market capitalization.
```

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

- All information is given at once in a static image.
- The designer has chosen to order the companies by their market capitalization, from the highest to the lowest. This allows the audience to quickly see which companies have the highest value in the market.
- The use of a bar chart also makes it easy to compare the market capitalization values of each company.

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

- The data has been used in the visualisation to convey the action by presenting the market capitalization values of the top 10 companies in the world. The visualisation uses a bar chart to show the relative sizes of the market capitalizations of each company. The length of each bar represents the market capitalization value, and the bars are arranged in descending order to show the ranking of the companies. The use of colour to highlight the top three companies further emphasizes their position at the top of the ranking.

B. Genre

Which of the seven genres listed above best dsecribes the data visualisation?

- Annotated chart [Visual representation of the data, along with text annotations that provide additional information or context about the data].

C. Author-driven vs Reader-driven

Where on the spectrum from author- to reader-driven is this visualisation?

- Author-driven. The author has created the graph with a specific purpose and has chosen the data to be presented. The audience is able to interpret and understand the information being presented, but they do not have any control over the data or how it is presented. The visualisation is intended to convey specific information to the audience, and the audience is not able to interact with the data or modify the visualisation to suit their needs.

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)

Paper : Raoufi et al 2019

a. What is the main argument of the paper?

- The main argument of the paper is that visual communication methods and tools can be effectively used to assess sustainability performance in industry, and that collaboration between academia and industry is necessary for the development and implementation of such tools. The authors also emphasize the importance of collaboration between academia and industry in developing and implementing visual communication tools for sustainability assessment.

b. According to this paper, why is effective visual communication important (or not)?

- Better comprehension: Stakeholders, including managers, staff, and consumers, can gain from efficient visual communication by understanding complicated sustainability performance data and information.

- Engagement: By making sustainability performance evaluation more approachable and exciting, visual communication techniques can boost stakeholder participation.
- Improved decision-making: By emphasising important sustainability performance measures and trends, clear and succinct visual communication can support better decision-making.
- Better communication: By simplifying the flow of data and information about sustainability performance, visual communication tools can increase communication between various stakeholders, particularly those between academia and industry.
- Increased transparency: By making data easier to obtain and comprehend for stakeholders, visual communication techniques can aid in promoting transparency and accountability in the assessment of sustainability performance.

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?

- Clarity, relevance, accuracy, consistency, accessibility, interactivity, co-creation would be the key elements, and I agree with that.

d. What pitfalls are identified in the paper that can be avoided if we use effective visual communication?

- Data Misunderstanding: Without effective visual communication, stakeholders can misinterpret or misunderstand sustainability performance data, leading to inaccurate conclusions and decisions.
- Limited accessibility: Traditional methods of assessing sustainability performance may be inaccessible to certain stakeholders. B. For people with low reading and math skills. Effective visual communication makes sustainability performance data available to a wider range of stakeholders.
- Lack of transparency: Traditional methods of assessing sustainability performance lack sufficient transparency and accountability and can lead to skepticism and mistrust from stakeholders. Effective visual communication can help solve this problem by providing clear, transparent and accessible sustainability performance data and information.

Week 4 Portfolio Questions–

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(covdata)
```

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

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

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

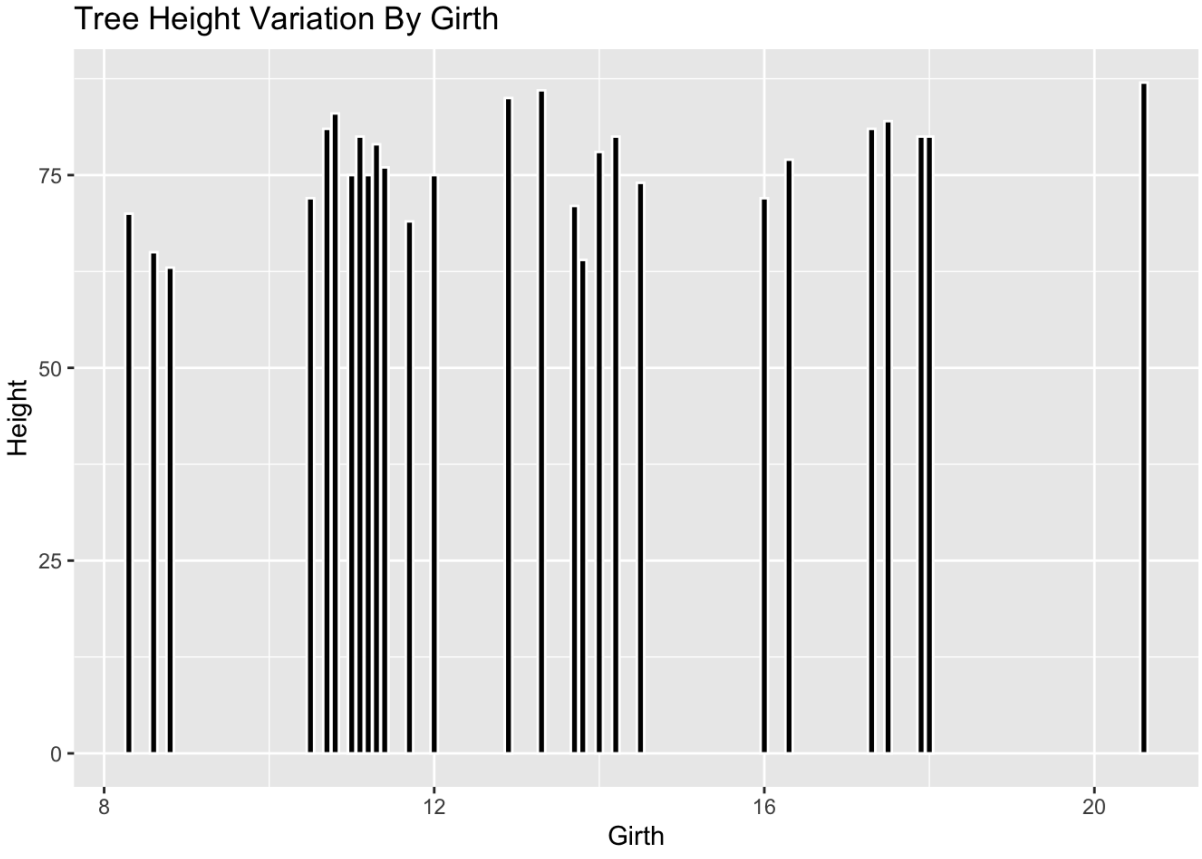
```
library(magrittr)
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ lubridate 1.9.2     ✔ tibble    3.2.1
## ✔ purrr     1.0.1     ✔ tidyr     1.3.0
## ✔ readr     2.1.4
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ tidyr::extract()   masks magrittr::extract()
## ✖ dplyr::filter()    masks stats::filter()
## ✖ dplyr::lag()       masks stats::lag()
## ✖ purrr::set_names() masks magrittr::set_names()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
me errors
```

```
library(knitr)
```

```
#Code that getting dataset
data(trees)
#Code to make graph to show the variation of tree heights
ggplot(trees, aes(x = Girth, y = Height)) +
  geom_bar(stat = "identity", fill = "Black", position = "dodge", color = "white") +
  labs(x = "Girth", y = "Height", title = "Tree Height Variation By Girth")
```



```
# I chose this trees dataset to show the variation of tree heights with plotting this dataset i
nto a bar graph.
# I decided to use the colours by their own values, in order to make them distinguishable, and
make labels to show the difference of heights by their Girth.
# It can be more effective when make this data's values more specifically, for instance there a
re some duplicate values in every columns and etc. And also in this graph which I made, when we
make those bar more bigger, its way more easy to see the differences betweeen their heights.
```

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.

```
#These are the codes to subset the data
data("apple_mobility")
Brisbane_mobility <- apple_mobility %>% filter(subregion_and_city =="Brisbane") %>% select(subr
egion_and_city, transportation_type, score) %>% group_by(subregion_and_city, transportation_typ
e) %>% summarise(total_score = sum(score))
```

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

```
Toronto_mobility <- apple_mobility %>% filter(subregion_and_city =="Toronto")%>% select(subregi
on_and_city, transportation_type, score) %>% group_by(subregion_and_city, transportation_type)
%>% summarise(total_score = sum(score))
```

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

```
Bangkok_mobility <- apple_mobility %>% filter(subregion_and_city =="Bangkok")%>% select(subregi
on_and_city, transportation_type, score) %>% group_by(subregion_and_city, transportation_type)
%>% summarise(total_score = sum(score))
```

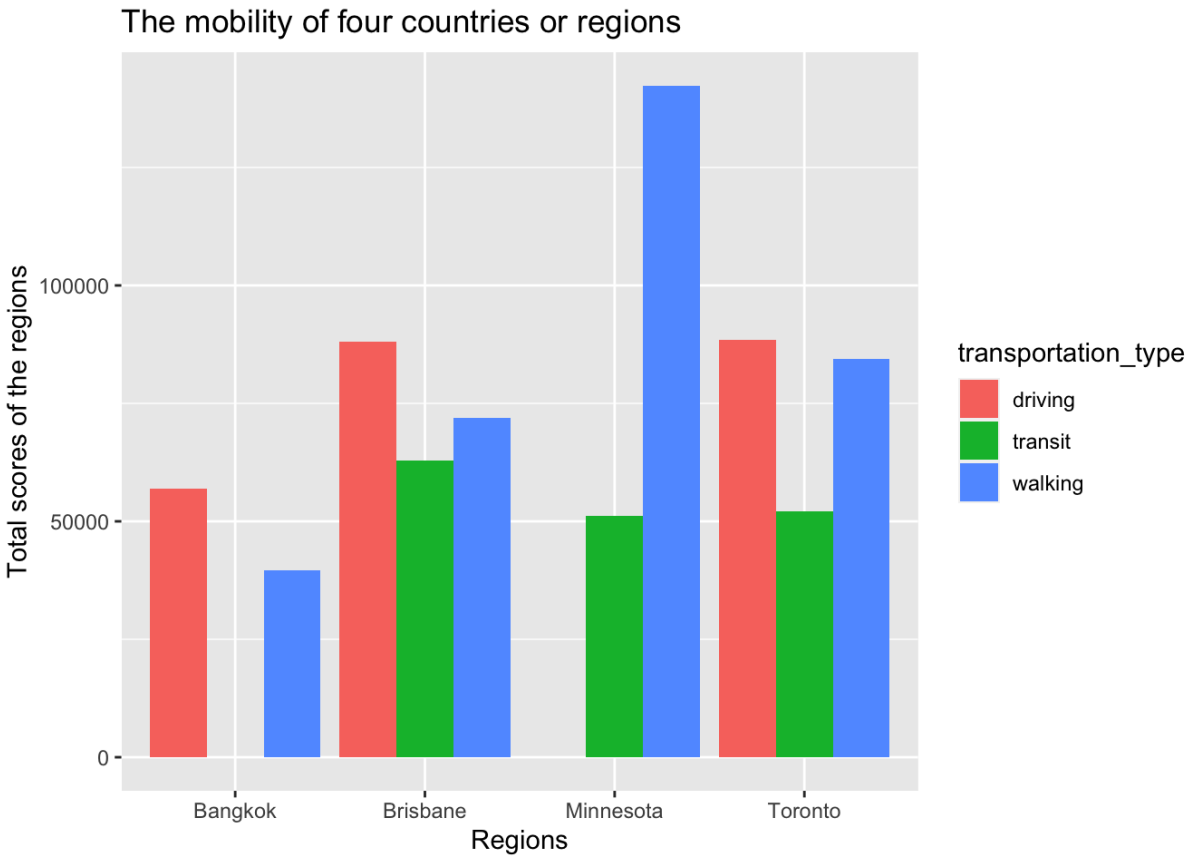
```
## `summarise()` has grouped output by 'subregion_and_city'. You can override
## using the `.groups` argument.
```

```
Minnesota_mobility <- apple_mobility %>% filter(subregion_and_city =="Minnesota") %>% select(su
bregion_and_city, transportation_type, score) %>% group_by(subregion_and_city, transportation_t
ype) %>% summarise(total_score = sum(score))
```

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

```
mobility_data <- rbind(Brisbane_mobility, Toronto_mobility, Bangkok_mobility, Minnesota_mobilit
y)

# Remove scientific notation in r
options(scipen=999)

#Code to plot the bar graph to compare those 4 countries/regions mobility
ggplot(mobility_data, aes(fill=transportation_type, y=total_score, x=subregion_and_city)) +
    geom_bar(position="dodge", stat="identity") +
    ggtitle("The mobility of four countries or regions")+
    labs( y = "Total scores of the regions", x = "Regions")
```

```
## Warning: Removed 2 rows containing missing values (`geom_bar()`).
```

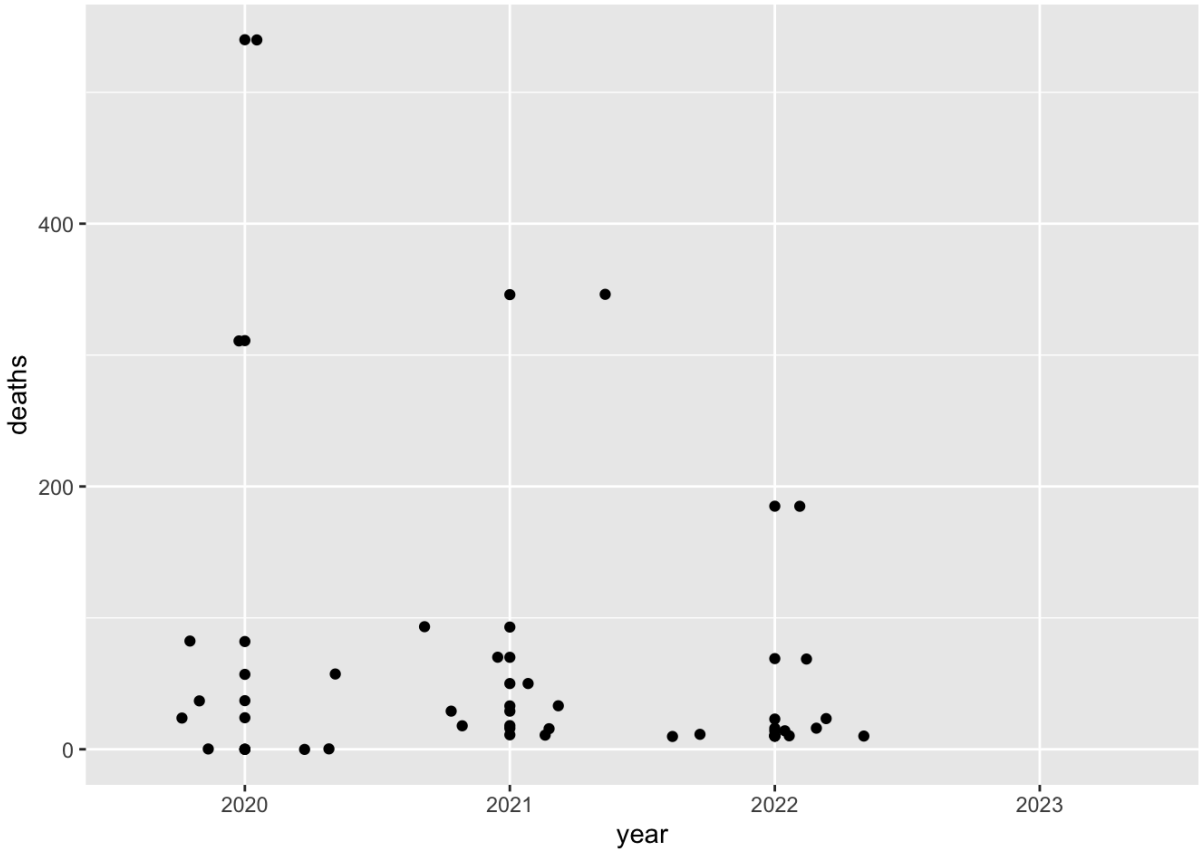



```
#The plot type chosen in this case is a grouped bar chart, which is appropriate for comparing t
he total mobility scores for different transportation types across multiple regions. The groupe
d bar chart enables the comparison of different transportation types for each region and facili
tates the comparison of mobility scores across regions.

#The x-axis label "Regions" and y-axis label "Total scores of the regions" are clear and concis
e, providing context to the reader about what the graph represents. The title of the graph "The
mobility of four countries or regions" clearly communicates the purpose of the graph.
```

5.(Mark 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.

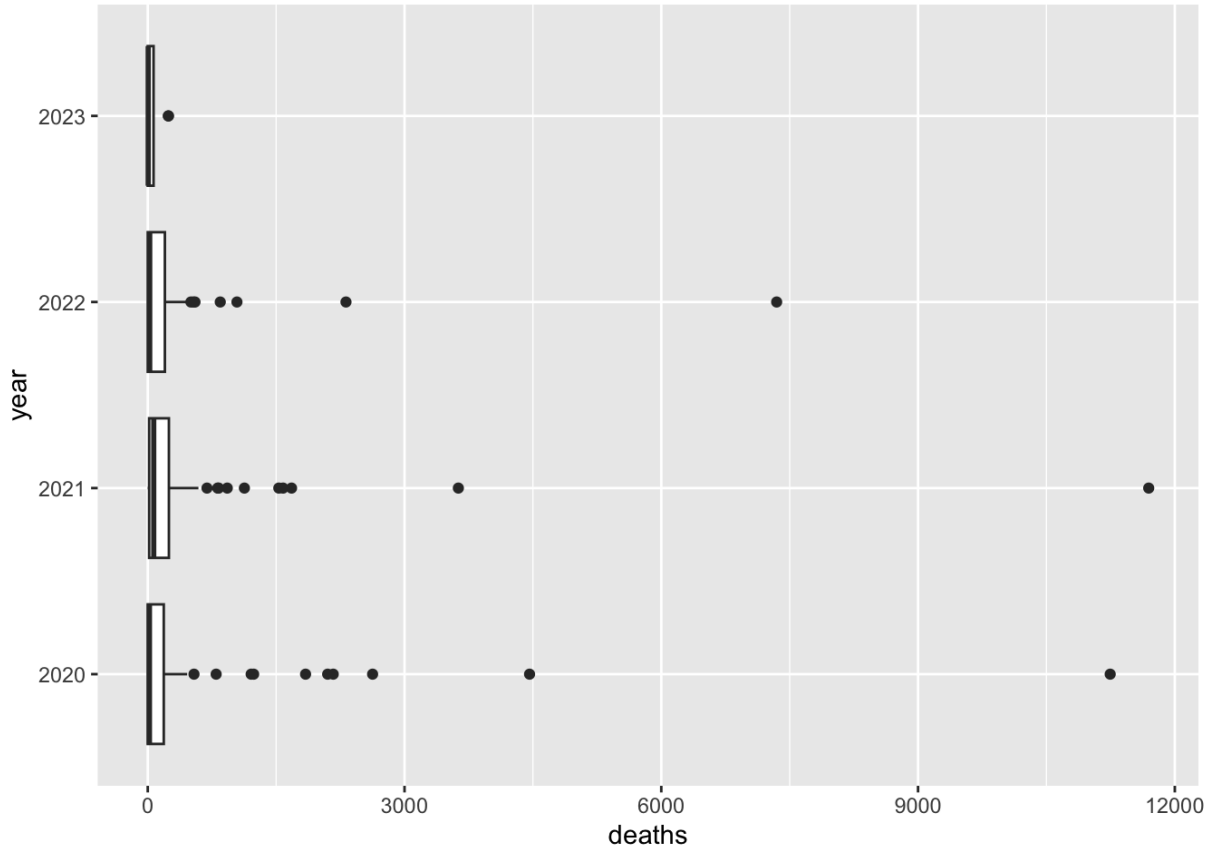
```
library(dplyr)

nchs_wss <- covdata::nchs_wss

nchs_wss_newyorkb <- nchs_wss %>% filter(state == "New York")

nchs_wss_year <- nchs_wss_newyorkb %>% filter(race_ethnicity == "Non-Hispanic Asian") %>% selec
t(year,race_ethnicity, deaths)

nchs_wss_newyork <-nchs_wss_newyorkb[-(1:7),]
nchs_wss_dens <-nchs_wss_year[-1,]

ggplot(nchs_wss_dens, aes(x = year, y = deaths)) +
  geom_point() +
  geom_jitter()
```

```
## Warning: Removed 14 rows containing missing values (`geom_point()`).
## Removed 14 rows containing missing values (`geom_point()`).
```




```
ggplot(nchs_wss_newyork, aes(x = deaths, y = year)) + geom_boxplot()
```

```
## Warning: Removed 77 rows containing non-finite values (`stat_boxplot()`).
```



```
#The scatter plot with jittered points is used to visualize the relationship between year and d
eaths for the "Non-Hispanic Asian" population in New York. The boxplot compares the distributio
n of deaths across different years for the overall New York population. Jittering is applied to
avoid overlapping points, and a horizontal orientation is used in the boxplot to enhance readab
ility and comparison. These choices aid in effectively communicating the trends and variations
in deaths over time and across different groups in New York.
```

Week 6 Portfolio Questioins–

```
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(ggiraphExtra)
library(GGally)
```

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

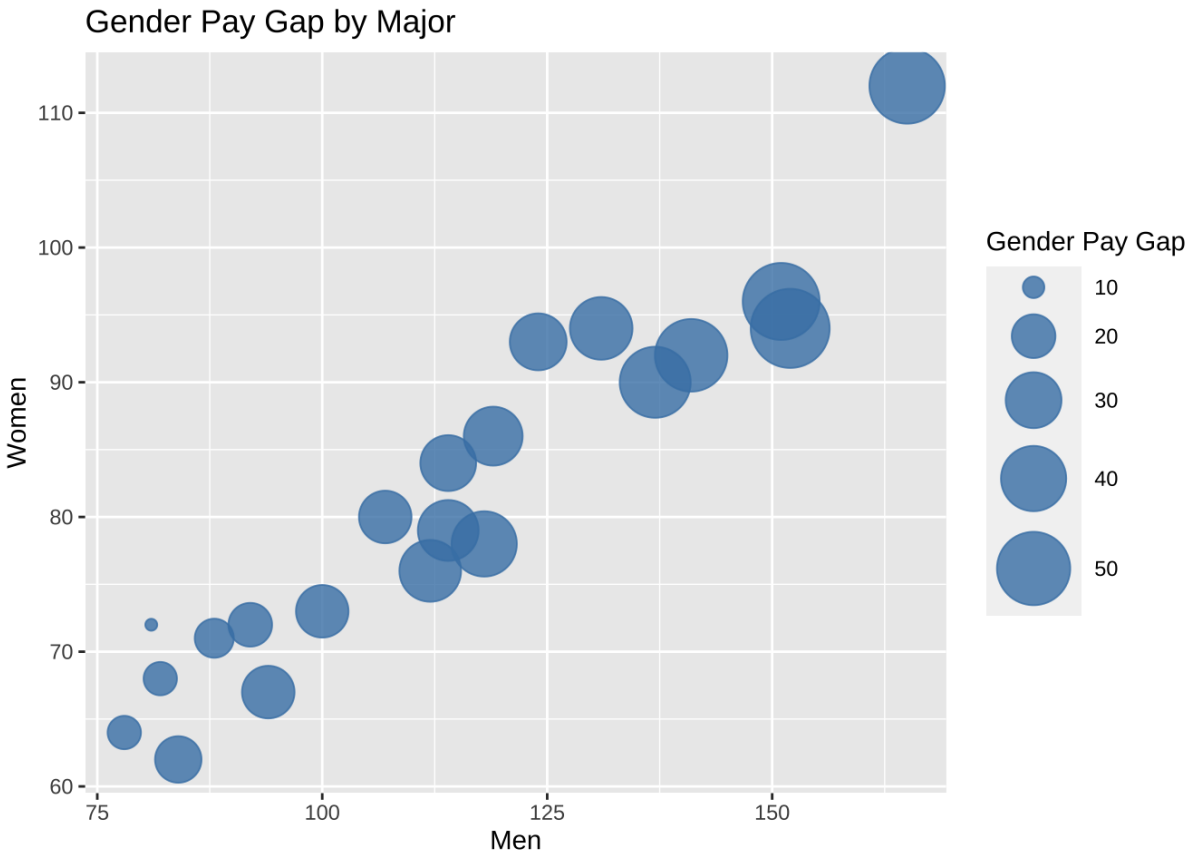
```
library(ggalluvial)
library(devtools)
```

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

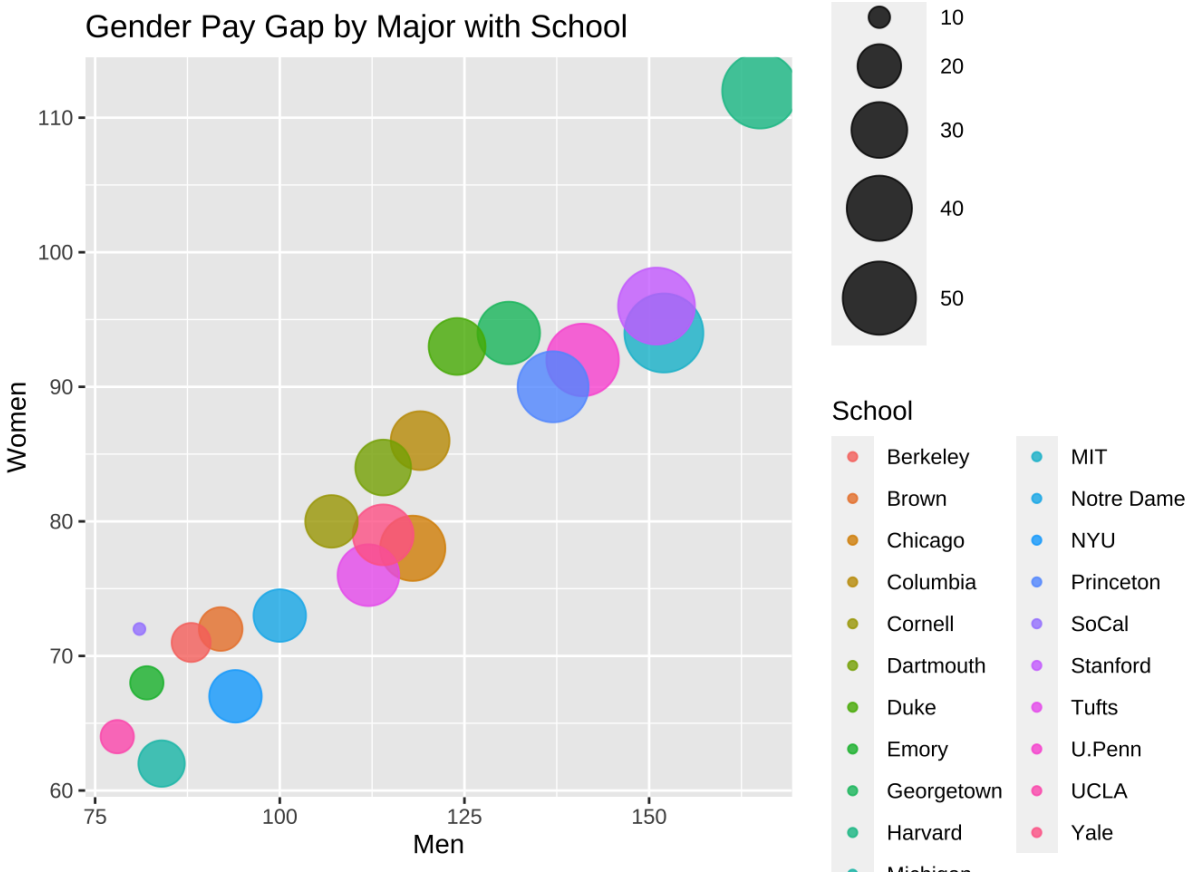
```
gender_pay_gap <- read.csv(
  "https://raw.githubusercontent.com/plotly/datasets/master/school_earnings.csv")
```

6.(Marks 5) Create two bubble plots (a.k.a multi-dimensional scatterplot) using the dataset gender_pay_gap. First, make the bubble plot you think is most useful.Second, add at least one more variable/preattentive attribute to your plot. what do you think is the first observation a reader makes from the second plot? why have you chosen to match the variable with the preattentive attributes you have chosen in the first plot? what audience might you use your first plot for vs your second?

```
# First Plot
ggplot(gender_pay_gap, aes(x = Men, y = Women, size = Gap)) +
  geom_point(color = "steelblue", alpha = 0.8) +
  scale_size_continuous(range = c(2, 15), name = "Gender Pay Gap") +
  ggtitle("Gender Pay Gap by Major")
```




```
# Second Plot
ggplot(gender_pay_gap, aes(x = Men, y = Women, size = Gap, color = School)) +
  geom_point(alpha = 0.8) +
  scale_size_continuous(range = c(2, 15), name = "Gender Pay Gap") +
  scale_color_discrete(name = "School") +
  ggtitle("Gender Pay Gap by Major with School")
```



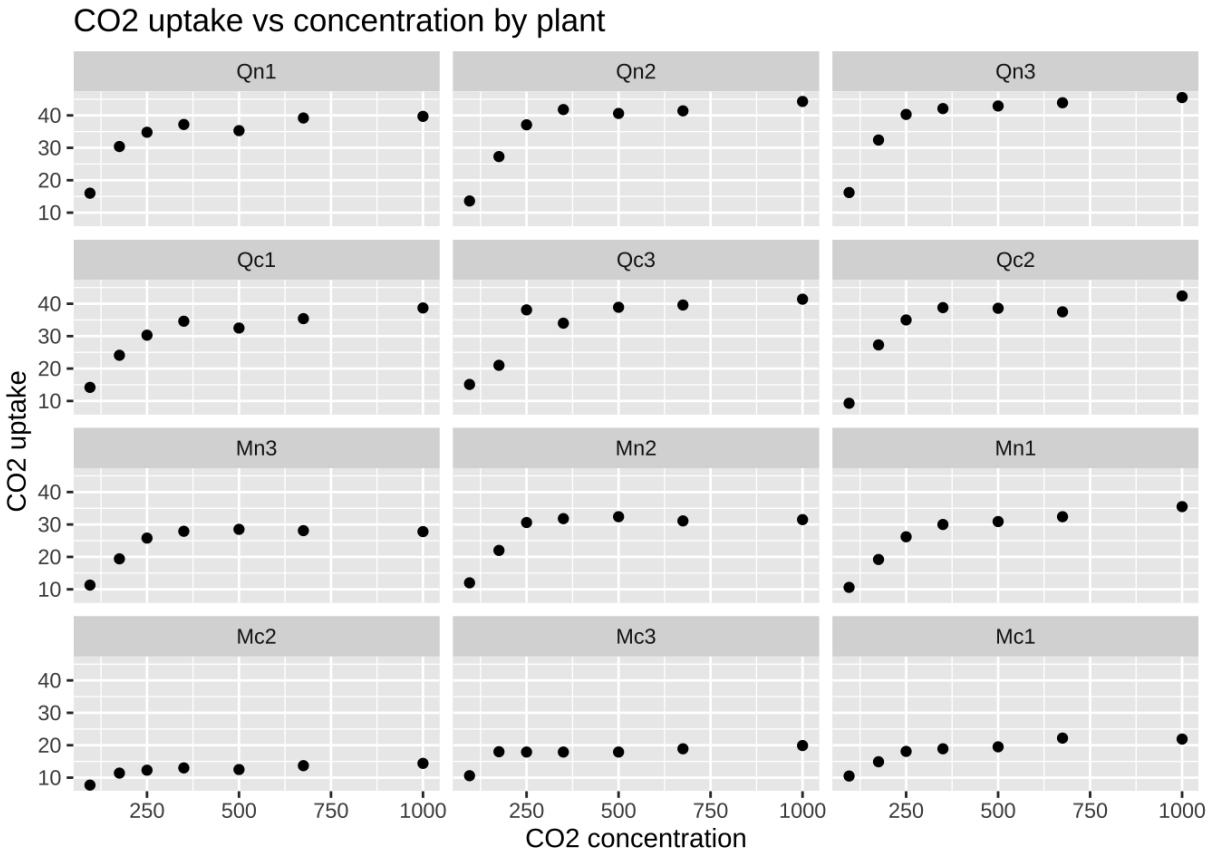
#A) From the second plot, the first observation a reader is likely to make is the relationship between the gender pay gap (represented by the size of the points) and the earnings of men and women in different majors. By incorporating the additional variable of "School" as a color aesthetic, the reader can also compare the gender pay gap across majors from different schools.

#B) In the first plot, the variable "Gender Pay Gap" is matched with the preattentive attribute of point size. The choice to match the variable with the size of the points is based on the principle of visual encoding and the goal of effectively communicating the magnitude of the gender pay gap. Using size as an attribute allows for a clear and intuitive representation of the pay gap, where larger points indicate a larger pay gap and smaller points indicate a smaller pay gap. This pairing helps the audience quickly grasp and compare the differences in pay gaps across different majors.

#C) The first plot, visualizing the gender pay gap by major, is suitable for a general audience interested in comparing pay disparities across fields of study. The second plot, incorporating the variable of "School," is more appropriate for analyzing pay gaps within each major across different institutions, appealing to researchers or individuals studying gender pay disparities in higher education.

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

```
data(CO2)

ggplot(CO2, aes(x=conc, y=uptake)) +
  geom_point() +
  facet_wrap(~Plant, ncol=3) +
  labs(title = "CO2 uptake vs concentration by plant",
       x = "CO2 concentration", y = "CO2 uptake")
```



#A) For several plant types, the reader may examine the relationship between CO2 uptake and CO2 concentration and how it varies for each species. The reader may also pick up on the various CO 2 concentration and uptake ranges for each plant. It is also simple to examine the correlations between CO2 uptake and concentration among plants because the reader can see the various aspect s that each variety of plant represents.

#B) This information can be effectively visualised by faceting to show the connections between CO2 uptake, concentration, and Plant type. Without having to build many plots, it enables the r eader to compare the correlations between CO2 uptake and concentration for each variety of plan t fast. Interpretation might be aided by the facet labels, which offer further details about th e Plant type being represented. Faceting can also make patterns and trends visible that might b e hidden in a single plot, particularly when working with several variables.

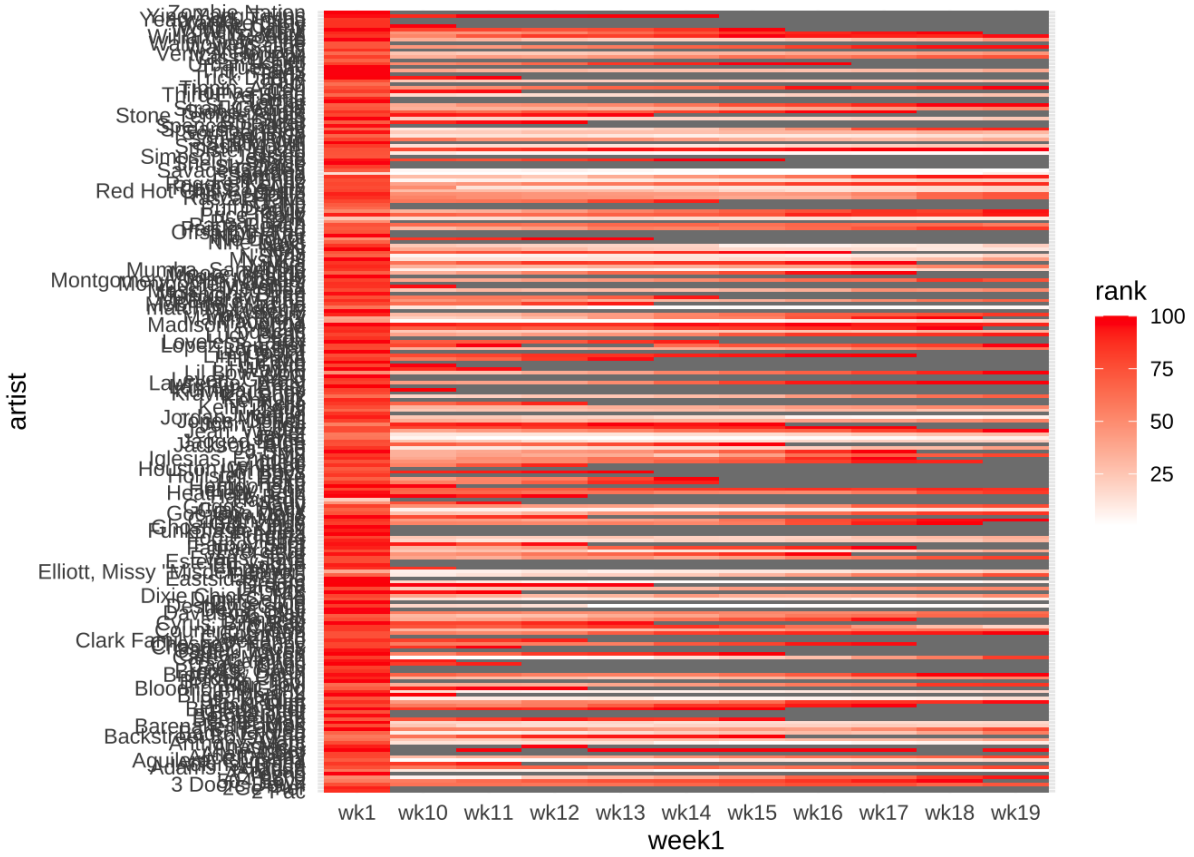
#C) A scatterplot matrix is yet another visualisation choice for investigating the link between CO2 uptake, concentration, and plant in the CO2 dataset. Each variable in a scatterplot matrix is plotted against every other variable in the dataset in a grid of scatterplots. This could be helpful for immediately spotting patterns or correlations in the pairwise interactions between the variables. Compared to faceting, it might also offer a more thorough picture of the data be cause all variables can be seen at once. For comparing relationships among various levels of a categorical variable, as in the case of plant types in the CO2 dataset, it might be less usefu l.

8.(Marks 7) You have a few choices for this question.

```
data("billboard")

# Pivot the data to long format
billboard_long <- billboard %>%
  pivot_longer(
    cols = starts_with("wk1"),
    names_to = "week1",
    values_to = "rank"
  ) %>%
  mutate(week = as.integer(stringr::str_replace(week1, "week", "")))
```

```
## Warning: There was 1 warning in `mutate()`.
## ℹ In argument: `week = as.integer(stringr::str_replace(week1, "week", ""))`.
## Caused by warning:
## ! NAs introduced by coercion
```

```
# Create the heatmap
ggplot(billboard_long, aes(x = week1, y = artist, fill = rank)) +
  geom_tile() +
  scale_fill_gradient(low = "white", high = "red") +
  theme_minimal()
```



#A) The first observation a reader might make from the plot is that there are many songs and artists represented, with varying levels of popularity (ranked by colour). The plot shows how the rankings change over time (weeks) for each artist, with some artists maintaining high rankings consistently and others fluctuating more.

#B) Because it makes it simple to compare the rank of various songs over time, the heatmap plot type was chosen as an effective way to visualise this data for comparisons. It is simple to recognise the most well-known songs and track changes in popularity over time thanks to the use of colour, in this case a gradient from white to red. A grid arrangement also makes it simple to spot patterns in the data, such as when several songs by the same artist are in demand at the same time.

#C) A stacked bar chart is a different type of visualisation that may be used to compare the ranking of songs over weeks and artists in the Billboard dataset.In this chart, the weeks could be represented on the x-axis, and the y-axis could represent the total number of songs in the top 100. Each bar could be segmented by the artists, and the colour of each segment could represent the rank of the song. This would allow the reader to compare the overall ranking of songs over time, as well as identify the top artists and their respective ranks.

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.

```
## Console

library(tidytuesdayR)
tuesdata <- tidytuesdayR::tt_load(2021, week = 15)
```

```
## --- Compiling #TidyTuesday Information for 2021-04-06 ----
```

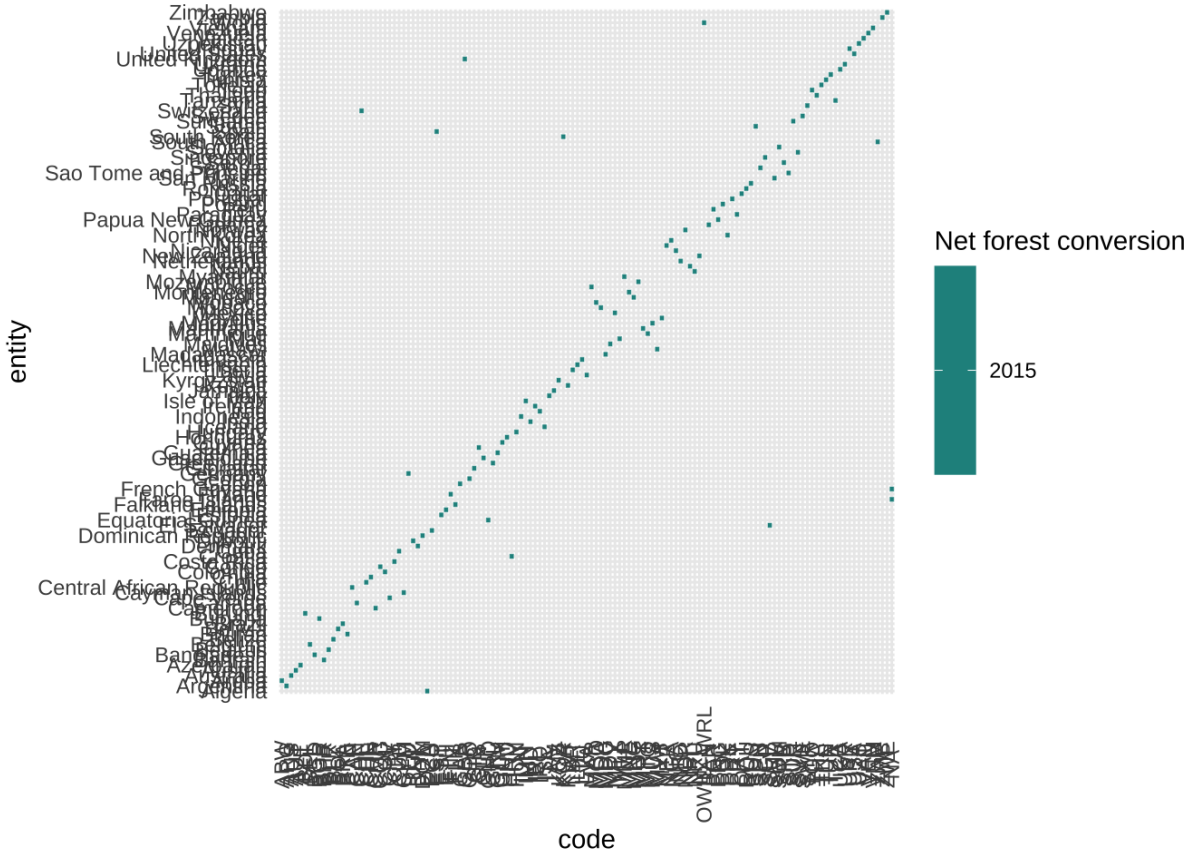
```
## --- There are 5 files available ---
```

```
## --- Starting Download ---
```

```
##
##  Downloading file 1 of 5: `forest.csv`
##  Downloading file 2 of 5: `forest_area.csv`
##  Downloading file 3 of 5: `brazil_loss.csv`
##  Downloading file 4 of 5: `soybean_use.csv`
##  Downloading file 5 of 5: `vegetable_oil.csv`
```

```
## --- Download complete ---
```

```
# Create a pivot table
forest_wide <- tuesdata$forest %>% pivot_wider(names_from = year, values_from = net_forest_conv
ersion)

# Create heatmap
ggplot(forest_wide, aes(x = code, y = entity)) +
  geom_tile(aes(fill = 2015), color = "white") +
  scale_fill_viridis(name = "Net forest conversion") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5))
```



```
#A) The first thing a reader notices about the plot is the locations (entities) along the y-axis, together with their appropriate codes, and the years along the x-axis. The hue of the tiles allows them to immediately determine the amount of nett forest conversion (positive or negative) in each country or region for the year 2015.

#B) We can readily compare the nett forest conversion numbers across various nations and years thanks to the heatmap plot type that was selected as the best approach to visualise this data. Readers can quickly compare the values of a variable across many categories or time periods by using heatmaps, which employ colour to depict a variable's magnitude.

#C) A line plot is another visualisation choice for comparing nett forest conversion across nations and years. The year may be displayed on the x-axis, the nett forest conversion may be represented on the y-axis, and each country may be represented by a distinct coloured line. This would make it possible to analyse each country's historical trends and variations in nett forest conversion in more depth.
```

Week 7 Portfolio Questioins–

```
library(igraph)
```

```
##
## Attaching package: 'igraph'
```

```
## The following objects are masked from 'package:lubridate':
##
##     %--%, 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:dplyr':
##
##     as_data_frame, groups, union
```

```
## The following objects are masked from 'package:stats':
##
##     decompose, spectrum
```

```
## The following object is masked from 'package:base':
##
##     union
```

```
library(network)
```

```
##
## 'network' 1.18.1 (2023-01-24), part of the Statnet Project
## * 'news(package="network")' for changes since last version
## * 'citation("network")' for citation information
## * 'https://statnet.org' for help, support, and other information
```

```
##
## Attaching package: 'network'
```

```
## The following objects are masked from 'package:igraph':
##
##     %c%, %s%, add.edges, add.vertices, delete.edges, delete.vertices,
##     get.edge.attribute, get.edges, get.vertex.attribute, is.bipartite,
##     is.directed, list.edge.attributes, list.vertex.attributes,
##     set.edge.attribute, set.vertex.attribute
```

```
library(sna)
```

```
## Loading required package: statnet.common
```

```
##
## Attaching package: 'statnet.common'
```

```
## The following objects are masked from 'package:base':
##
##     attr, order
```

```
## sna: Tools for Social Network Analysis
## Version 2.7-1 created on 2023-01-24.
## copyright (c) 2005, Carter T. Butts, University of California-Irvine
##  For citation information, type citation("sna").
##  Type help(package="sna") to get started.
```

```
##
## Attaching package: 'sna'
```

```
## The following objects are masked from 'package:igraph':
##
##     betweenness, bonpow, closeness, components, degree, dyad.census,
##     evcent, hierarchy, is.connected, neighborhood, triad.census
```

```
library(ndtv)
```

```
## Loading required package: networkDynamic
```

```
##
## 'networkDynamic' 0.11.3 (2023-02-15), part of the Statnet Project
## * 'news(package="networkDynamic")' for changes since last version
## * 'citation("networkDynamic")' for citation information
## * 'https://statnet.org' for help, support, and other information
```

```
## Loading required package: animation
```

```
##
## 'ndtv' 0.13.3 (2022-11-20), part of the Statnet Project
## * 'news(package="ndtv")' for changes since last version
## * 'citation("ndtv")' for citation information
## * 'https://statnet.org' for help, support, and other information
```

```
library(EpiContactTrace)
library(RColorBrewer)
library(viridis)
library(circlize)
```

```
## ========================================
## circlize version 0.4.15
## 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:sna':
##
##     degree
```

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

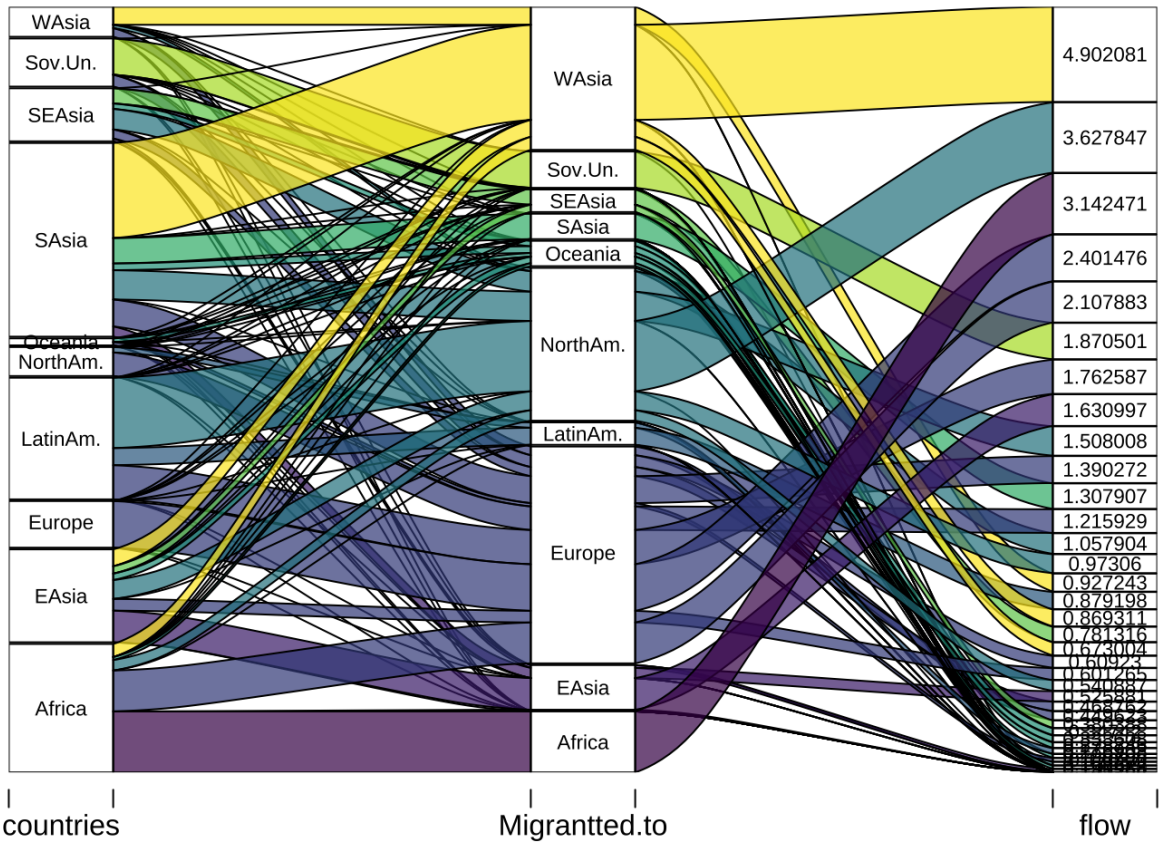
```
library(alluvial)
```

10.(Marks 10) A) 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.

```
network <- read.table("https://raw.githubusercontent.com/holtzy/data_to_viz/master/Example_data
set/13_AdjacencyDirectedWeighted.csv", header=TRUE)

colnames(network) <- c("Africa", "EAsia", "Europe", "LatinAm.",   "NorthAm.",   "Oceania", "SAs
ia", "SEAsia", "Sov.Un.", "WAsia")
rownames(network) <- colnames(network)

data_long <- gather(rownames_to_column(network), key = 'key', value = 'value', -rowname)

flows <- data_long %>%
  group_by(rowname, key) %>%
  summarize(flow = sum(value))
```

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

```
colnames(flows) <- c("countries", "Migrantted to", "flow")

mycolor <- viridis(nrow(network))

alluvial(
  flows[, c("countries", "Migrantted to", "flow")],
  freq = flows$flow,
  col = mycolor,
  alpha = 0.7,
  border = "#000000",
  cex = 0.7,
  gap.width = 0.02
)
```



#B) Alluvial plot visualises the migration network in terms of flows between countries, providing a detailed representation of migration patterns and volume. Chord diagram would present the same data with a focus on the connections between regions, emphasizing the relationship and intensity of migration between them.

#C) The use of color, width, and transparency in the alluvial plot enhances the audience's understanding by visually conveying the magnitude and diversity of migration flows, facilitating comparisons between countries and regions, and aiding in the interpretation of the overall migration network structure

#D) An alternative visualisation that could be more effective than the alluvial plot is a geographic map. The map would display the countries as distinct regions, and the migration flows would be represented by arrows or lines connecting the countries. This visual representation would provide a clear spatial context for the migration patterns, enabling viewers to easily identify the origin and destination countries. Additionally, the map could incorporate color coding or thickness of the lines to indicate the flow volume, further enhancing the understanding of the migration network.

#E) An alternative visualisation that can expand the message is a combination of a stacked bar chart and a world map. The stacked bar chart represents migration flows from each country to different regions, while the world map provides a geographic context. This combination allows for a comparison of migration patterns across countries and regions, but it may be less effective in illustrating precise flow patterns compared to the alluvial plot.

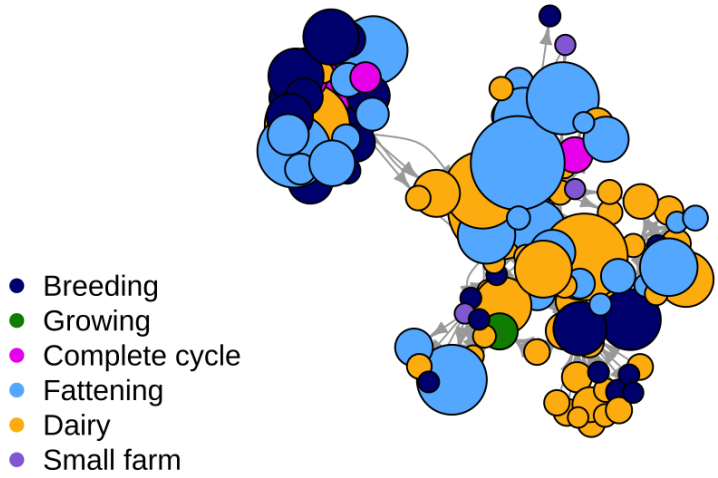
11.(Marks 8) A) Consider the cattle network from this week's workshop. Create a network visualisation where vertex size is used to represent the size of the farm in some way. Your plot will be assessed for accuracy and effectiveness, so design your visualisation and use pre-attentive attributes carefully.

```
attr1 <- read.csv("data/Farms/attr_farms.csv", stringsAsFactors = F)
edges <- read.csv("data/Farms/Edgelist_farms.csv", stringsAsFactors = F)
```

```
net_edges <- graph_from_data_frame(edges,directed=T)
net_edges
```

```
## IGRAPH 4726b5c 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 4726b5c (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
```

```
V(net_edges)$type <- as.character(attr1$type[match(V(net_edges)$name, attr1$farm.id)])
V(net_edges)$farm.size <- attr1$size[match(V(net_edges)$name, attr1$farm.id)]

type_colors <- c(
  "Breeding" = "navy",
  "Growing" = "green4",
  "Complete cycle" = "magenta2",
  "Fattening" = "steelblue1",
  "Dairy" = "darkgoldenrod1",
  "Small farm" = "mediumpurple"
)

size_range <- range(V(net_edges)$farm.size)
min_size <- 10  # Minimum vertex size
max_size <- 50  # Maximum vertex size

normalized_sizes <- scale(V(net_edges)$farm.size, center = min(size_range), scale = diff(size_r
ange))
vertex_sizes <- min_size + normalized_sizes * (max_size - min_size)

plot(
  net_edges,
  layout = layout.fruchterman.reingold,
  vertex.label = NA,
  vertex.color = type_colors[V(net_edges)$type],
  vertex.size = vertex_sizes,
  edge.arrow.size = 0.5
)

legend(
  "bottomleft",
  legend = names(type_colors),
  col = type_colors,
  pch = 19,
  cex = 1,
  bty = "n"
)
```


Breeding
Growing
Complete cycle
Fattening
Dairy
Small farm

```
#B1) The size of each farm vertex indicates the size of the farm as a whole. Smaller farms' ver
tex sizes will be smaller while larger farms' vertex sizes will be larger. We can determine the
relative sizes of farms within the network by comparing the vertex sizes.
#B2) Based on the strength of the linkages between connected farms, the network visualisation c
reated using the Fruchterman-Reingold arrangement positions connected farms closer to one anoth
er. If large farms have strong links with many other farms, they may display clustering or cent
ral positions within the network.
#B3) Large farms may have higher degrees of connectedness, which indicates that they interact o
r link with other farms in the network more frequently. They might act as key nodes in the netw
ork, linking several smaller farms.
#B4) It's important to note that the code differentiates across farms according to their charac
teristics (such as Breeding, Growing, Complete cycle, etc.). A specific specialisation inside t
he network may be indicated by large farms displaying a concentration within a certain farm typ
e.

#C) The message conveyed in this visualisation is a representation of a cattle network where fa
rm sizes are visually encoded using vertex sizes. The purpose of the visualisation is to provid
e insight into the distribution and relationships of farms in the network based on their size.
The audience for this visualisation includes researchers, policy makers, and stakeholders invol
ved in animal husbandry and agricultural planning. You may be interested in understanding the s
tructure of cattle networks, identifying farm sizes, and possibly examining patterns and relati
onships between farm sizes and other factors in the network.

#D) The use of vertex size in the visualization allows for easy comparison of farm sizes, conve
ying the relative magnitudes effectively. The color coding of farm types adds an additional lay
er of information, facilitating the understanding of farm diversity and aiding in identifying p
atterns or clusters within the network.

#E) An alternative visualization that can communicate the same message is a bubble chart. In th
is chart, each farm would be represented by a bubble, where the x-axis represents the farm size
and the y-axis represents some measure of importance or influence within the network. The size
of each bubble could correspond to the farm's size, similar to the vertex sizes in the network
visualization. Additionally, the color of the bubble could represent the farm type.
This bubble chart would provide a different perspective on the data compared to the network vis
ualisation. It would allow for a more direct comparison of farm sizes and their corresponding i
mportance within the network. By aligning the bubbles along the x-axis according to farm size,
the audience can easily identify the range and distribution of farm sizes. The varying sizes of
the bubbles would further emphasize the relative magnitudes of the farm sizes. But on the other
hand, the bubble chart may be less effective in depicting the network structure and relationshi
ps between farms. It does not explicitly show the connections or interactions between farms, un
like the network visualisation. The absence of these connections may limit the audience's abili
ty to identify clusters, central hubs, or patterns of farm relationships within the network. Th
e bubble chart would be more focused on the individual attributes of each farm, rather than the
network as a whole.
```

Week 8 Portfolio Questioins–

```
library(cowplot)
```

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

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

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

```
library(ggplot2)
library(ggrepel)
library(ggspatial)
library(rnaturalearth)
library(rnaturalearthdata)
```

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

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

```
library(sf)
```

```
## Linking to GEOS 3.10.2, GDAL 3.4.2, PROJ 8.2.1; sf_use_s2() is TRUE
```

```
library(maps)
```

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

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

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

```
library(maptools)
```

```
## Loading required package: sp
```

```
## Checking rgeos availability: TRUE
## Please note that 'maptools' will be retired during 2023,
## plan transition at your earliest convenience;
## some functionality will be moved to 'sp'.
```

```
library(rgeos)
```

```
## rgeos version: 0.6-3, (SVN revision 696)
##  GEOS runtime version: 3.10.2-CAPI-1.16.0
##  Please note that rgeos will be retired during October 2023,
## plan transition to sf or terra functions using GEOS at your earliest convenience.
## See https://r-spatial.org/r/2023/05/15/evolution4.html for details.
##  GEOS using OverlayNG
##  Linking to sp version: 1.6-0
##  Polygon checking: TRUE
```

```
##
## Attaching package: 'rgeos'
```

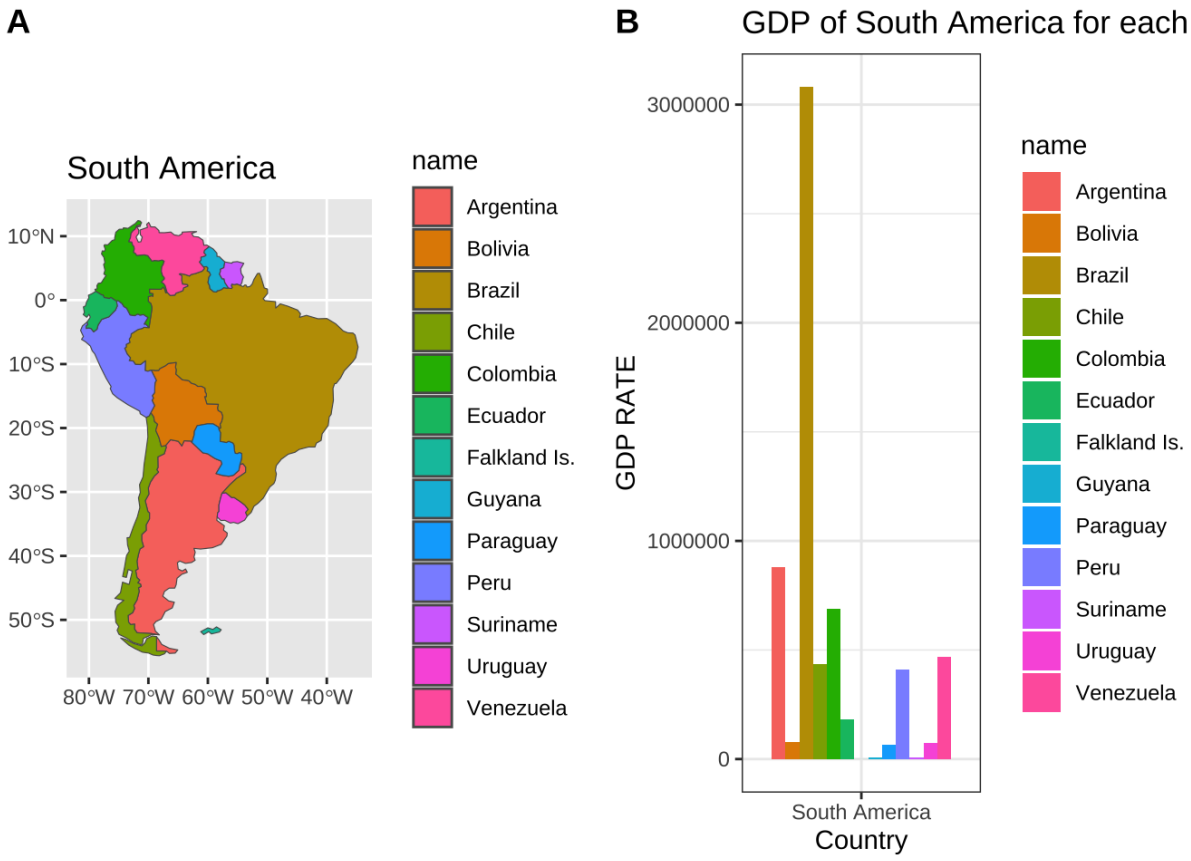
```
## The following object is masked from 'package:dplyr':
##
##     symdiff
```

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.

```
continent <- "South America"
continent_map <- ne_countries(continent = continent, returnclass = "sf")

class(continent_map)
```

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

```
world_plot <- ggplot(data = continent_map) +
    geom_sf(aes(fill = name))+
    ggtitle("South America")


world_gdp_plot <- ggplot(continent_map, aes(x = continent, y = gdp_md_est)) +
  geom_bar(stat = "identity", position = "dodge", aes(fill = name)) +
  labs(x = "Country", y = "GDP RATE") +
  ggtitle(paste("GDP of", continent,"for each countries")) +
  theme_bw()

plot_grid(world_plot, world_gdp_plot,
          labels = c("A", "B"),
          ncol = 2, nrow = 1)
```




```
##Map Visualization:
# Purpose: The map visualization displays the countries in South America, with polygons filled
based on the "gdp_md_est" attribute. It provides a spatial overview of GDP distribution across
the continent.
# Target Audience: Researchers, policymakers, or individuals interested in understanding the GD
P variations in South American countries.
# Intended Message: The map highlights the relative economic strength of different countries in
South America based on their GDP estimates.

##Bar Chart Visualization:
# Purpose: The bar chart represents the GDP rates of South American countries, comparing them s
ide by side. Each bar corresponds to a country, and the height of the bar represents the GDP va
lue.
# Target Audience: Economists, analysts, or individuals interested in comparing the GDP rates o
f South American countries.
# Intended Message: The bar chart provides a clear visual comparison of the GDP rates among the
countries in South America, allowing viewers to identify countries with higher or lower GDP val
ues.

# While these visualizations focus on the same geographical region (South America) and use the
"gdp_md_est" attribute, they serve different purposes. The map visualization provides a spatial
overview, while the bar chart facilitates direct comparisons. Both visualizations offer insight
s into GDP, but they present the information in distinct ways. Consider using them separately b
ased on the specific information needs of the target audience.
```

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.

```
whiskey_loc <- read.csv("whiskey_data/tas_locations.csv")
whiskey_sale <- read.csv("whiskey_data/whiskey_sales_tasmania.csv")

counties <- st_as_sf(map("county", plot = FALSE, fill = TRUE))
counties <- subset(counties, grepl("Tasmania", counties$ID))
counties$area <- as.numeric(st_area(counties))
head(counties)
```

```
## Warning in `[<-.data.frame`(`*tmp*`, is_list, value = list(`2` =
## "<s_GEOMET>")): replacement element 1 has 1 row to replace 0 rows
```

0 rows

```
world <- ne_countries(scale = "medium", returnclass = "sf")
class(world)
```

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

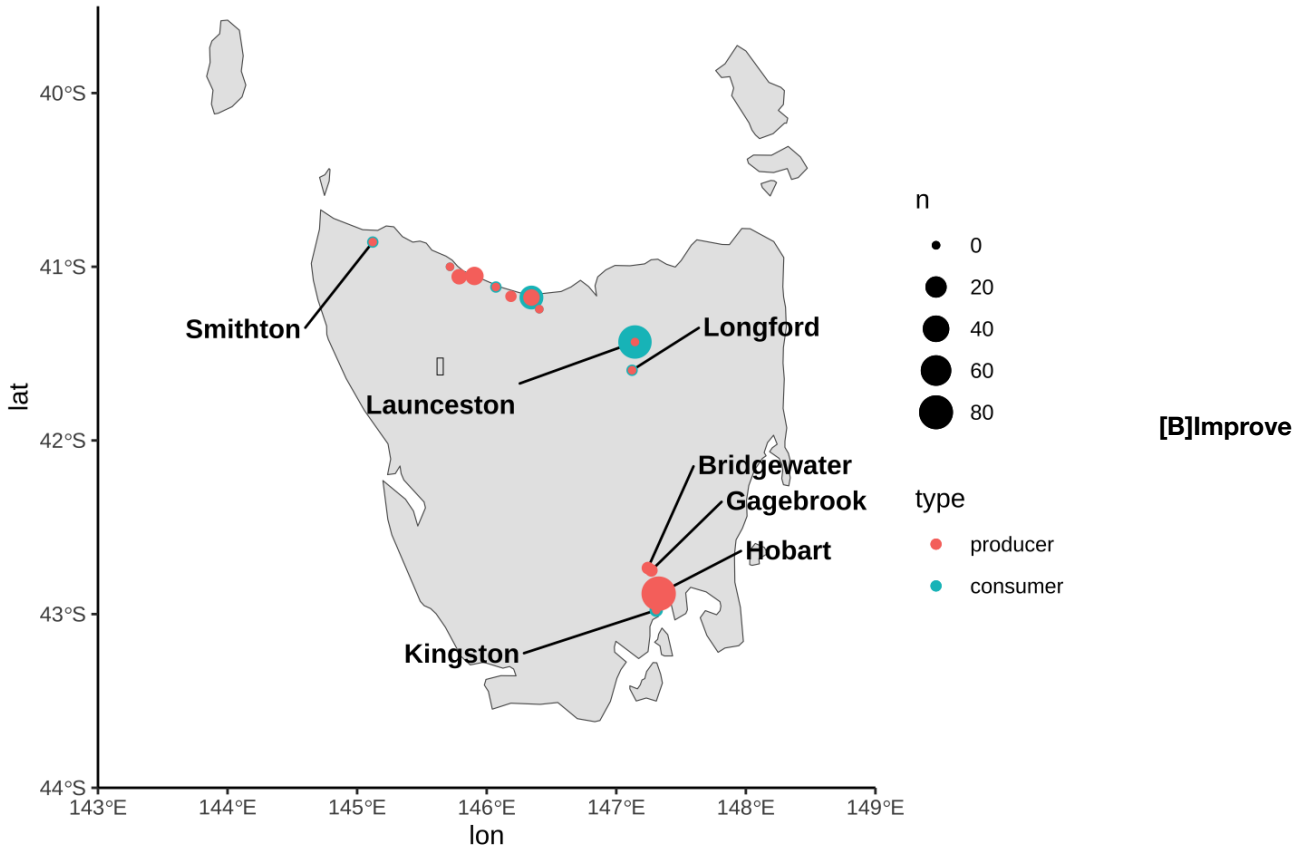
```
whiskey_latlng <- data_frame(state = rep("Tasmania",7), city = c("Hobart", "
Launceston", "Gagebrook", "Bridgewater", "Smithton", "Longford", "Kingston"), lat = c(-42.8821,
-41.4332, -42.7483, -42.7347, -40.8578, -41.5958, -42.9769), lon = c(147.3272, 147.1441, 147.27
06, 147.2442, 145.1206, 147.1218, 147.3083))
```

```
## Warning: `data_frame()` was deprecated in tibble 1.1.0.
## ℹ Please use `tibble()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
whiskey_latlng <- st_as_sf(whiskey_latlng, coords = c("lon", "lat"), remove = FALSE, crs = 432
6, agr = "constant")

Producer <- c(0, 77, 29, 0, 0, 0, 0, 0, 1, 1, 1, 0, 3, 0)
Consumer <- c(83, 0, 8, 12, 1, 2, 2, 6, 0, 0, 0, 0, 0, 0)

whiskey_consum <- cbind(whiskey_loc, Consumer)
whiskey_sizecl <- cbind(whiskey_loc, Producer, Consumer)

ggplot(data = world) +
  geom_sf()+
  geom_sf(data = counties) +
  geom_text_repel(data = whiskey_latlng, aes(x= lon, y= lat, label= city),
                  fontface= "bold", nudge_x = c(1, -1.5, 2, 2, -1), nudge_y = c(0.25,
                  -0.25, 0.5, 0.5, -0.5)) +
  geom_point(data = whiskey_sizecl, aes(x = lon, y = lat, size = Producer, color = "#E41A1C"))+
  geom_point(data = whiskey_sizecl, aes(x = lon, y = lat, size = Consumer, color = "#377EB8"))+
  labs( color = "type",  size = "n")+
  scale_colour_discrete(labels = c("producer", "consumer"))+
  coord_sf(xlim = c(143, 149), ylim = c(-44, -39.5), expand = FALSE) +
  xlab("lon") + ylab("lat") +
  theme_bw() + theme(panel.border = element_blank(), panel.grid.major = element_blank(),
  panel.grid.minor = element_blank(), axis.line = element_line(colour = "black"))
```

```
## Warning in x + params$x: longer object length is not a multiple of shorter
## object length
```

```
## Warning in y + params$y: longer object length is not a multiple of shorter
## object length
```

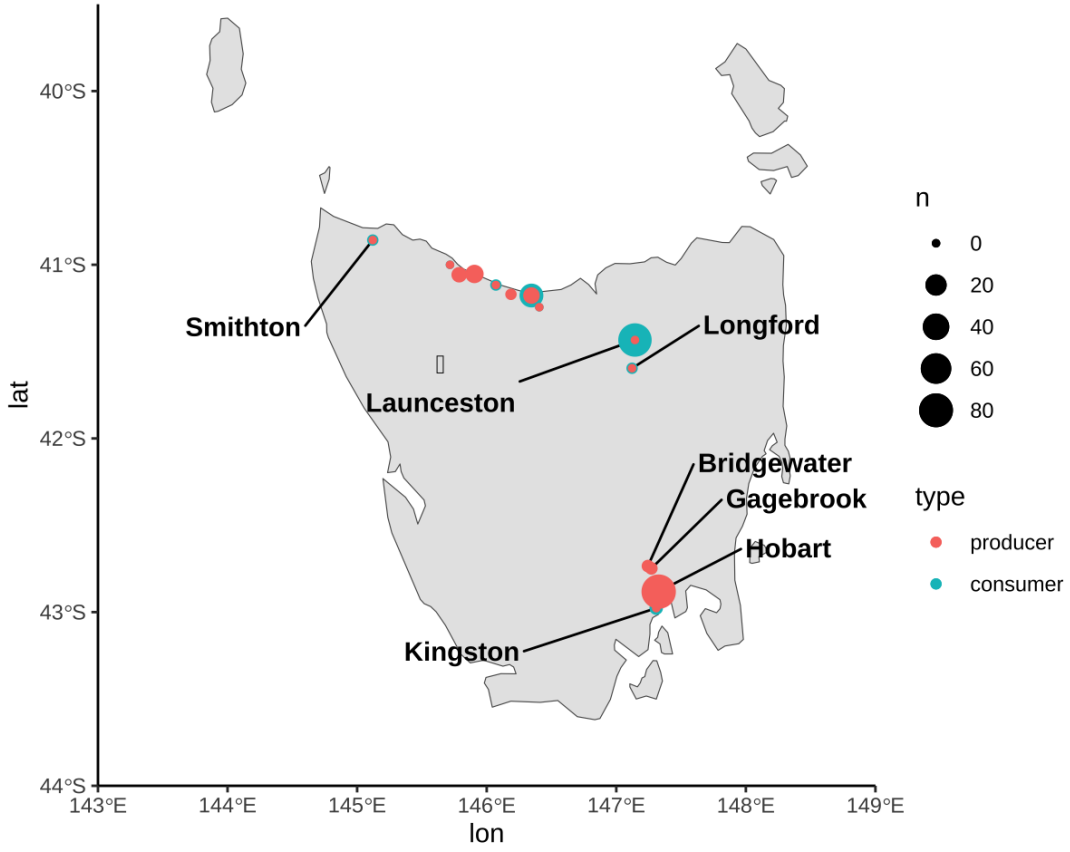


[B]Improve upon this visualisation to be more effective at communicating a message. Identify the audience and the message, and justify your visualisation choices within that context. Include both visualisations (the one from part (a) and part (b))

```
#First plot
ggplot(data = world) +
  geom_sf()+
  geom_sf(data = counties) +
  geom_text_repel(data = whiskey_latlng, aes(x= lon, y= lat, label= city),
                  fontface= "bold", nudge_x = c(1, -1.5, 2, 2, -1), nudge_y = c(0.25,
                  -0.25, 0.5, 0.5, -0.5)) +
  geom_point(data = whiskey_sizecl, aes(x = lon, y = lat, size = Producer, color = "#E41A1C"))+
  geom_point(data = whiskey_sizecl, aes(x = lon, y = lat, size = Consumer, color = "#377EB8"))+
  labs( color = "type",  size = "n")+
  scale_colour_discrete(labels = c("producer", "consumer"))+
  coord_sf(xlim = c(143, 149), ylim = c(-44, -39.5), expand = FALSE) +
  xlab("lon") + ylab("lat") +
  theme_bw() + theme(panel.border = element_blank(), panel.grid.major = element_blank(),
  panel.grid.minor = element_blank(), axis.line = element_line(colour = "black"))
```

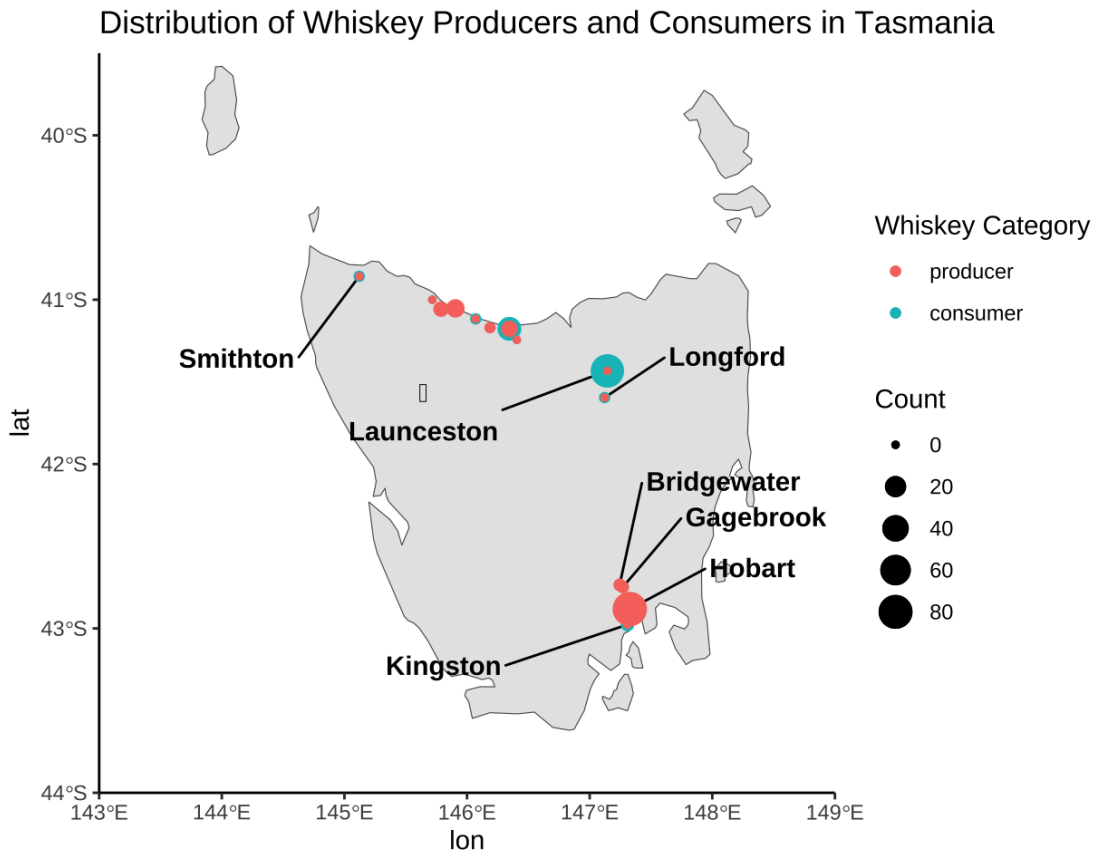
```
## Warning in x + params$x: longer object length is not a multiple of shorter
## object length
```

```
## Warning in y + params$y: longer object length is not a multiple of shorter
## object length
```




```
#Second plot
ggplot(data = world) +
  geom_sf()+
  geom_sf(data = counties) +
  geom_text_repel(data = whiskey_latlng, aes(x= lon, y= lat, label= city),
                  fontface= "bold", nudge_x = c(1, -1.5, 2, 2, -1), nudge_y = c(0.25,
                  -0.25, 0.5, 0.5, -0.5)) +
  geom_point(data = whiskey_sizecl, aes(x = lon, y = lat, size = Producer, color = "#E41A1C"))+
  geom_point(data = whiskey_sizecl, aes(x = lon, y = lat, size = Consumer, color = "#377EB8"))+
  labs( color = "Whiskey Category",  size = "Count", title = "Distribution of Whiskey Producers
and Consumers in Tasmania")+
  scale_colour_discrete(labels = c("producer", "consumer"))+
  coord_sf(xlim = c(143, 149), ylim = c(-44, -39.5), expand = FALSE) +
  xlab("lon") + ylab("lat") +
  theme_bw() + theme(panel.border = element_blank(), panel.grid.major = element_blank(),
  panel.grid.minor = element_blank(), axis.line = element_line(colour = "black"))
```

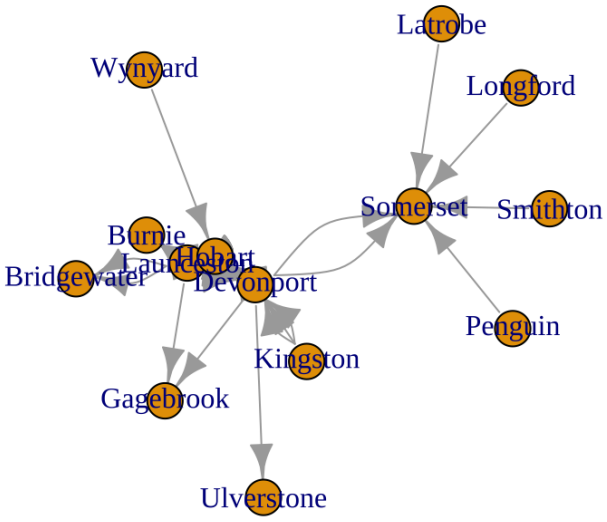
```
## Warning in x + params$x: longer object length is not a multiple of shorter
## object length

## Warning in x + params$x: longer object length is not a multiple of shorter
## object length
```




```
#Audience: The audience for the second visualization remains the same: individuals interested i
n understanding the distribution of whiskey producers and consumers in Tasmania.

#Message: The second visualization aims to communicate the distribution of whiskey producers an
d consumers in Tasmania through a more comprehensive presentation, including a title and a lege
nd.

#Improvements:
#Title: The second visualization already includes a title, which provides a clear message and c
ontext.
#Legend and Axis Labels: The second visualization has already addressed the legend and axis lab
el improvements discussed for the first visualization.
```

[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).

```
edges1 <- whiskey_sale[, c("producer", "consumer")]

# Create a graph object from the edge list
g <- graph_from_data_frame(edges1, directed = TRUE)

# Set visual attributes for the network plot
visual <- list(
  layout = layout_with_fr(g),  # Choose a layout algorithm (e.g., Fruchterman-Reingold)
  vertex.size = 10,            # Set the size of the vertices
  vertex.label = V(g)$name,    # Use the vertex names as labels
  vertex.color = "lightblue",  # Set the vertex color
  edge.color = "gray",         # Set the edge color
  vertex.title = "connection between producers and consumers "
)

# Plot the network
plot(g)
```




```
# Part b (Distribution of Whiskey Producers and Consumers in Tasmania):
#Audience: This visualization is intended for a general audience interested in understanding th
e spatial distribution of whiskey producers and consumers in Tasmania.

# Part C (Network Plot of Whiskey Producers and Consumers):
#Audience: This visualization is aimed at an audience interested in exploring the relationships
between whiskey producers and consumers in Tasmania.
```

Week 8 Portfolio Questioins–

```
library(cartogram)
library(maptools)
library(sf)
library(rgeos)
library(ggplot2)
library(raster)
```

```
##
## Attaching package: 'raster'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(sp)
library(rgdal)
```

```
## Please note that rgdal will be retired during 2023,
## plan transition to sf/stars/terra functions using GDAL and PROJ
## at your earliest convenience.
## See https://r-spatial.org/r/2022/04/12/evolution.html and https://github.com/r-spatial/evolu
tion
## rgdal: version: 1.6-6, (SVN revision 1201)
## Geospatial Data Abstraction Library extensions to R successfully loaded
## Loaded GDAL runtime: GDAL 3.4.2, released 2022/03/08
## Path to GDAL shared files: /Library/Frameworks/R.framework/Versions/4.2/Resources/library/rg
dal/gdal
## GDAL binary built with GEOS: FALSE
## Loaded PROJ runtime: Rel. 8.2.1, January 1st, 2022, [PJ_VERSION: 821]
## Path to PROJ shared files: /Library/Frameworks/R.framework/Versions/4.2/Resources/library/rg
dal/proj
## PROJ CDN enabled: FALSE
## Linking to sp version:1.6-0
## To mute warnings of possible GDAL/OSR exportToProj4() degradation,
## use options("rgdal_show_exportToProj4_warnings"="none") before loading sp or rgdal.
```

```
library(viridis)
```

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 10, and justify how your choices (including the numeric breaks) have made your visualisations more effective.

```
Lythrum <- raster("invasive_plant_rasters_2019/sumrast_allassumptions.avg_Lythrum salicaria.ti
f")
Frangula <- raster("invasive_plant_rasters_2019/sumrast_allassumptions.avg_Frangula alnus.tif")

summary(Lythrum)
```

```
## Warning in .local(object, ...): summary is an estimate based on a sample of 100000 cells (1
6.01% of all cells)
```

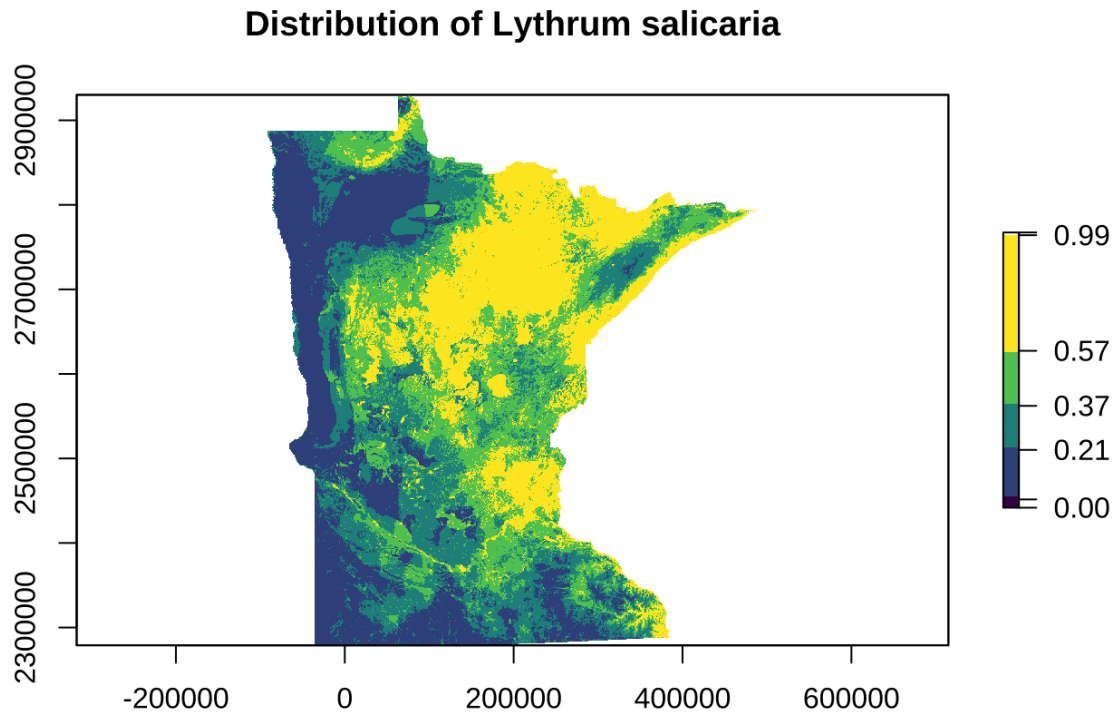
```
##         sumrast_allassumptions.avg_Lythrum.salicaria
## Min.                                      0.03686781
## 1st Qu.                                   0.21478477
## Median                                    0.37880574
## 3rd Qu.                                   0.57473505
## Max.                                      0.99149007
## NA's                                 262555.00000000
```

```
summary(Frangula)
```

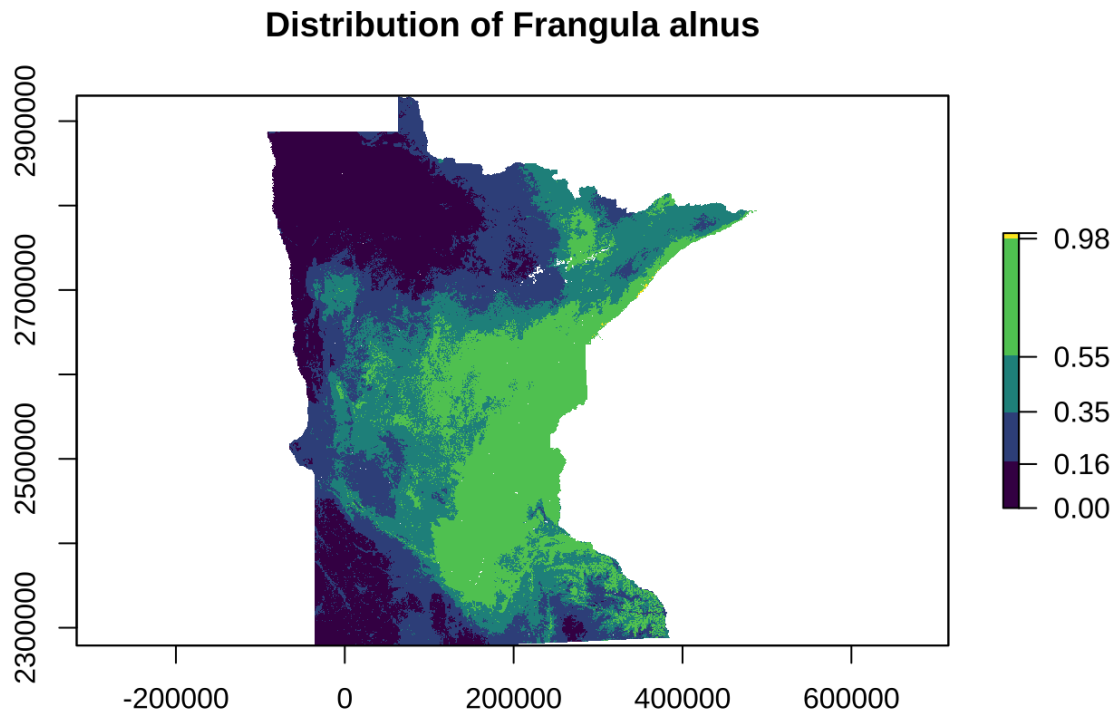
```
## Warning in .local(object, ...): summary is an estimate based on a sample of 100000 cells (1
6.01% of all cells)
```

```
##         sumrast_allassumptions.avg_Frangula.alnus
## Min.                                   0.03362482
## 1st Qu.                                0.16304430
## Median                                 0.35875545
## 3rd Qu.                                0.55030139
## Max.                                   0.98712343
## NA's                              263080.00000000
```

```
breaks_Lythrum <- c(0,0.03,0.21,0.37,0.57,0.99,1)
breaks_Frangula <- c(0,0.16,0.35,0.55,0.98,1)

collll <- viridis(5)

plot(Lythrum, col = collll, breaks = breaks_Lythrum, legend = TRUE,
     main = "Distribution of Lythrum salicaria")
```




```
plot(Frangula, col = collll, breaks = breaks_Frangula, legend = TRUE,
     main = "Distribution of Frangula alnus")
```




```
#Numeric breaks : I used their Min., 1st Qu, Median, 3rd Qu, Max values which I got from summar
y() and it helped me approaching to defining numeric breaks for the color scheme. It allows you
to capture the range of values and highlight different intervals or levels of presence effectiv
ely.

#Colour: Using viridis(5) I can have a balanced distribution of colors to represent distinct le
vels or categories of the invasive plant species data. The five colors in the palette will allo
w viewers to easily differentiate between different levels of presence or intensity in the dist
ribution of the invasive plant species.
```

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 setence argument that you'll make to the Department to convince them of your recommendation, referring to your maps and your figures from this question.

```
#As a data scientist working with the Minnesota Department of Natural Resources, I have analyze
d the distribution of two invasive plant species, Lythrum salicaria (Lythrum) and Frangula alnu
s (Frangula), using both spatial maps and non-spatial visualizations. Based on the maps, it is
evident that Lythrum has a larger distribution compared to Frangula, covering a wider area acro
ss Minnesota. Additionally, the non-spatial visualization further supports this observation, sh
owing a higher overall presence of Lythrum compared to Frangula. Therefore, considering the ext
ent and intensity of the invasive species, I recommend focusing on controlling Lythrum as the p
rimary target for management efforts to mitigate its significant impact on Minnesota's ecosyste
ms.
```