Marimekko chart

Data visualisation-End sem assignment

Dr.Ashish Anand **V.Saikiran(194161016)**

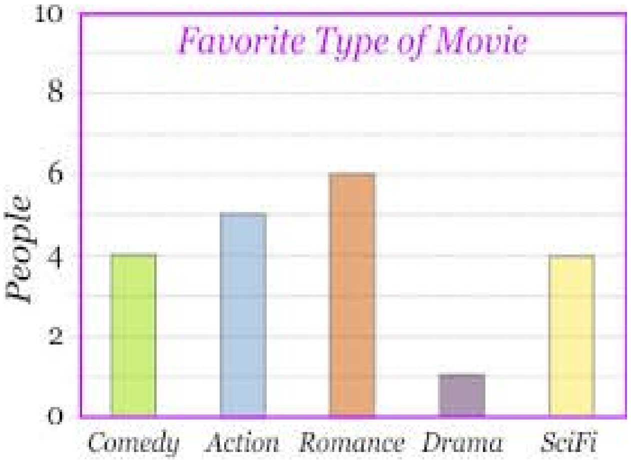
Introduction:

In statistical analysis we deal with different types of data. Types of data include:

- **Categorical Data (Nominal, Ordinal)**
- **Numerical Data (Discrete, Continuous, Interval, Ratio)**.

Categorical Data:

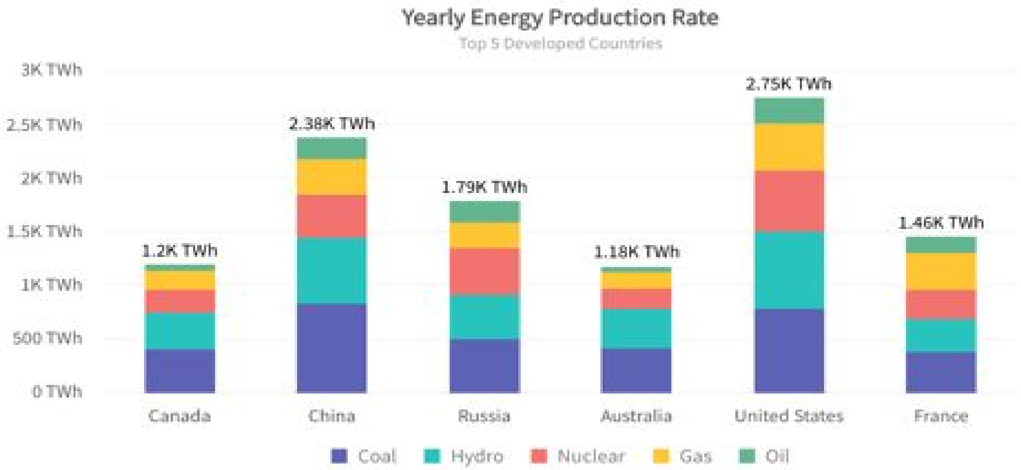
- As the **Marimekko chart** represents categorical data we mainly focus on it.
- In statistics, a **categorical variable** is a variable that can take on one of a limited, and fixed number of possible values.
- Examples include **race, sex, age group,mobile brands, countries** etc.
- Categorical data is usually represented using **bar charts, pie charts** etc.
- **Bar charts** are useful when there is one categorical variable and another variable representing numerical quantity such as count, frequency etc.



Graph Credits: www.mathsisfun.com

- The above bar graph represents **favorite genre of movies** among randomly selected people.
- What if the two category variables.

Example Category1 - Country Category2 - type of Energy



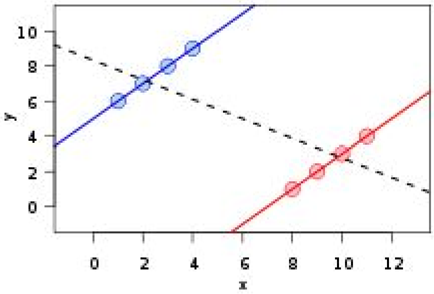
Graph Credits:exceljet.net

- In the above chart we can see the **country wise and Energy wise** number of terawatt hours(in 1000s) of production. The graph is representing two dimensions of the data **country** and **type of Energy**.
- But the catch here is the bars are just showing us the vary in heights keeping the width of the bars constant and same across different companies.
- This graph does not represent the **country's share** in the **whole world wide production of energy.**
- In such cases **Marimekko chart(or Mosaic chart)** is very much useful.

- Let us look at this chart using a **use case**.To demonstrate the use case we have to understand the significance of **simpson’s paradox** in Statistics.

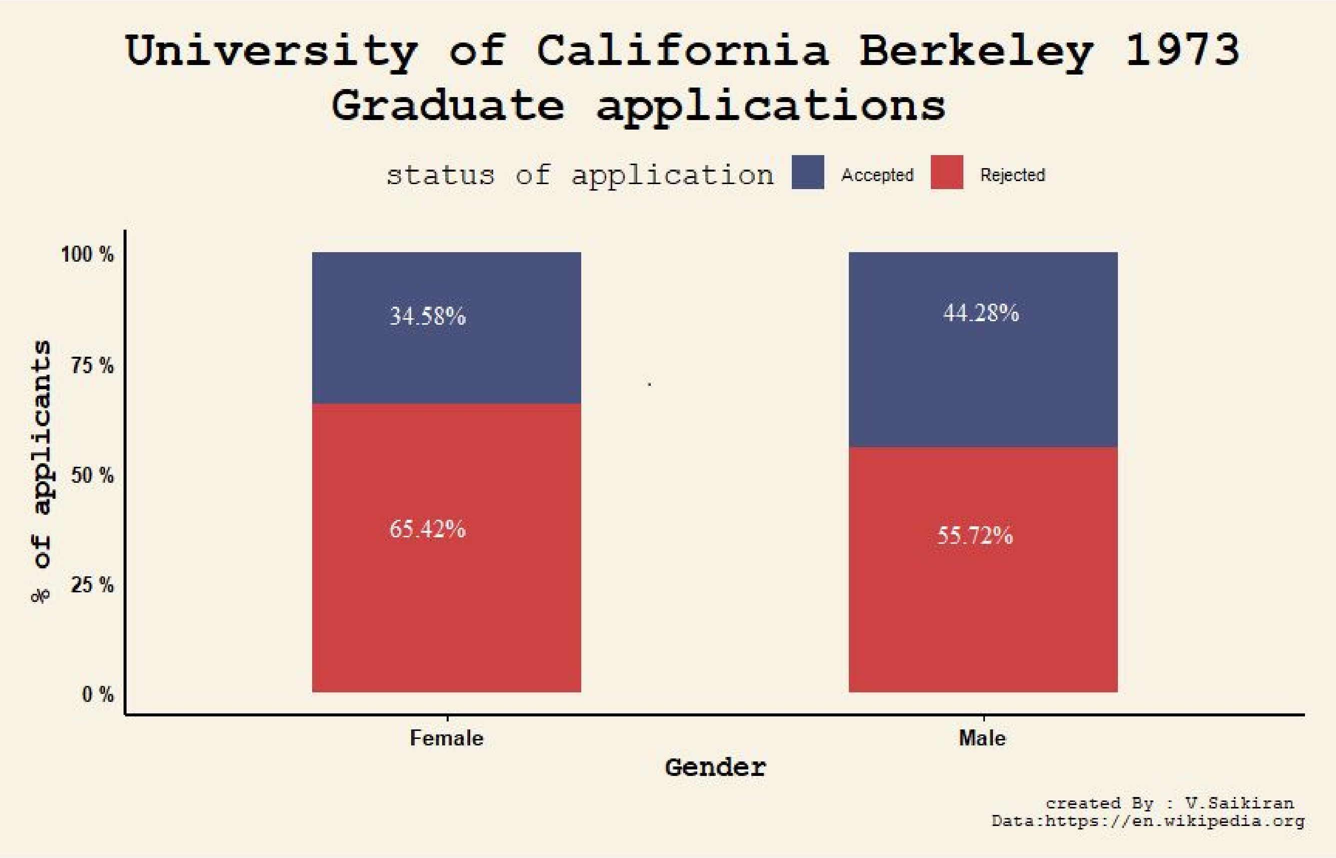
Simpson's paradox:

Simpson’s paradox, is a phenomenon in **probability** and **statistics**, which says that aggregated data sometimes behave in an **opposite(or reverse) trend** to the individual grouped data
Example picture from wikipedia:

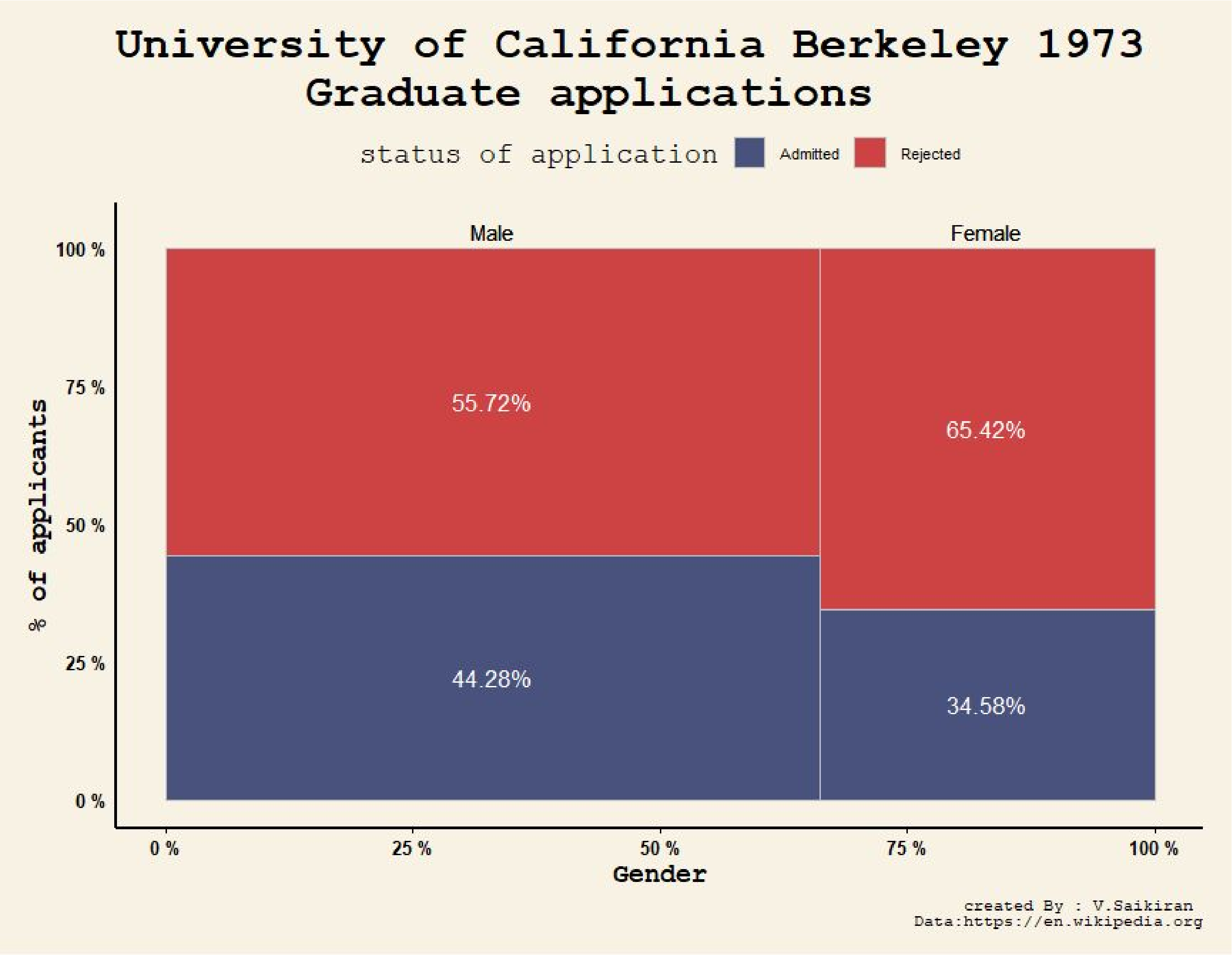


- We have 2 variables x, y which have **positive(or increasing trend)** relationships(we can see from colored lines).
- When we fit a **regression line** for the data, the line comes out to be a **negative(or decreasing trend)** kind of a relationship(dotted line) between x, y.
- This kind of phenomenon is called **simpson’s paradox**,which says that never trust the **data blindly**.Investigate thoroughly how actually the data is behaving.
- A famous use case demonstrating Simpson's paradox is a study of gender bias among **graduate school admissions to University of California, Berkeley.**
- There was an allegation against the graduate school admissions at **University of California, Berkeley in fall 1973.**

- The admission figures showed that men were more likely to get admitted when compared to women.
- However when the concerned authorities investigated department wise they found that 6 out of 85 departments were significantly biased against men, whereas 4 were significantly biased against women.
  Let us look at a plot demonstrating this.

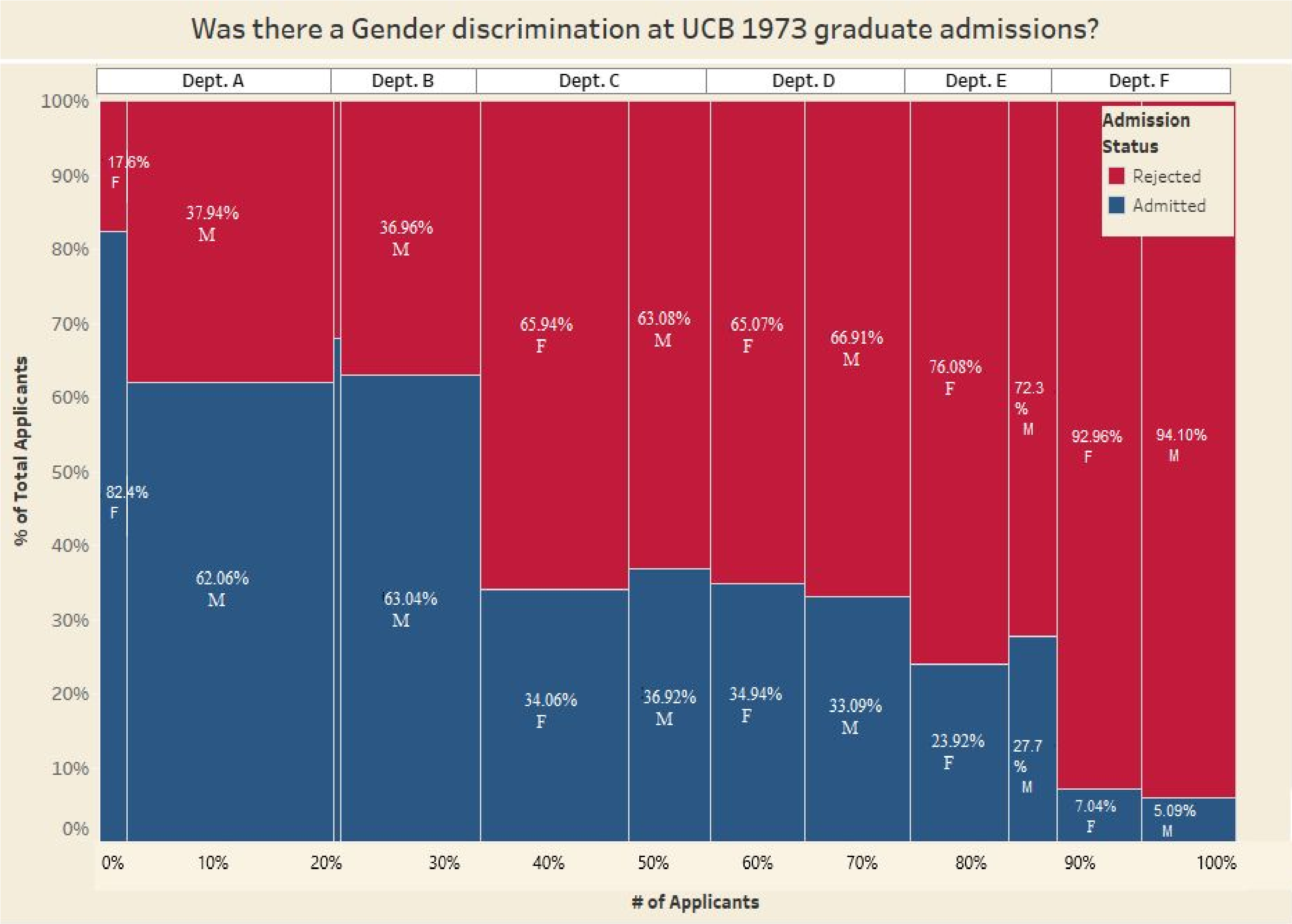


- **Stacked bar plot** version of the **whole data** of Graduate applications at UCB 1973 tells us that **44.28%** of **male applicants** and **34.58%** of **female applicants** were **admitted**.
- But the drawback of this plot is it is not showing the **percentage share of male and female applicants in the whole data.** This we can show by varying the width of the bars using **marimekko chart**.



- From this marimekko chart we can see that around 65% of the applications are from the male students and 35% are from female students.
- Among them **55.72% of male applications** were **rejected** and **65.42%** of **female applications** were **rejected**. **Yes,** we can clearly say that there is **gender discrimination** in the process of admissions.
- But we have seen that, according to **Simpson's paradox**, **aggregated data** may behave in the **reverse direction** of the actual trend of the individual grouped data.
- The data from the **six largest departments** of University of California Berkeley has been collected and a marimekko chart has been plotted.

This Marimekko chart is telling us a different story.



- Out of **6** departments, **4** departments had been in **more** favor to **female** applicants compared to male applicants.
- Department A has admitted **82.4%** of **female applicants** whereas only **62.06%** of **male applicants** were admitted.
- Similarly **Department B, D, F** were also in favor of **female applicants**.

- But the **confounding factor** we can see from the data is that in the case of **Dept. A, B** there were very **less no. of applications**(bar width tells us) from **female candidates** where the rate of acceptance is **high**.
- Where as in **Dept. F** the no. of **female applicants** are more which has **very low acceptance rates.**
- Examining **aggregate data** on graduate admissions was clearly showing us a **misleading pattern** of bias against **female applicants**.
- There was a research paper by Bickel et al. which concluded that **female applicants** tended to apply to most **competitive departments** which had **low rates of admission**, whereas **men** tended to apply to **less-competitive departments** with **high rates of admission** among the qualified applicants.
- This use case is the best example to demonstrate **Simpson's paradox** with the help of **Marimekko charts.**

Lets answer some of the questions:

What does a Marimekko chart display?

- Generally Marimekko charts are used to display **categorical data** over a pair of variables.
- The major difference it has compared to **stacked bar charts** is that in this **both the axes** are **variable** with **percentage scale** which determines both the **width and height** of each category.
- That is compared to the standard absolute value **stacked bar chart**, each of the **primary bars** will now have the **same length** but **different widths**.
- With the help of **Marimekko charts** we can detect the **relationships** between categories with the help of **two axes.**We can treat this as a two way 100% stacked bar graph.
- In this chart column widths are scaled in such a way that the total width of the categories matches the desired chart width.

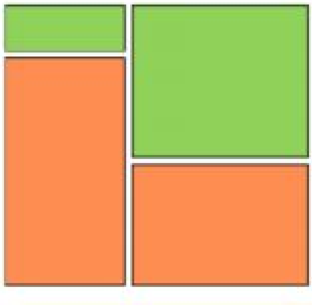
What are the advantages and disadvantages of the Marimekko chart?

Advantages:

1. Using **Marimekko charts** one can capture two dimensions in one chart.That is the **width and height** of the **bars carry information**.
2. With a Marimekko chart, you can quickly **spot large and small segments** with the help of the **width** of **vertical bars.**

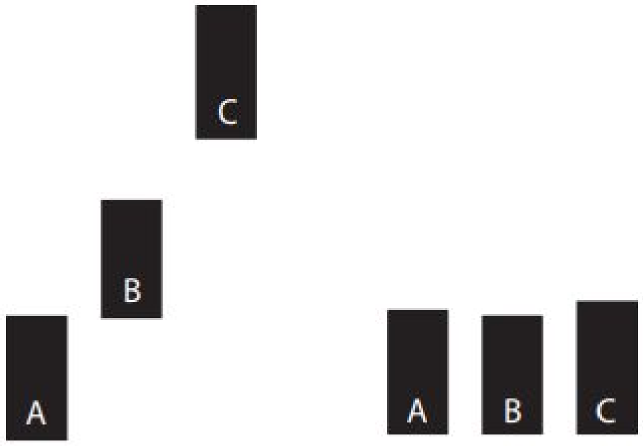
Disadvantages:

1. **Marimekko charts** may be **hard to read**, especially when there are **many segments and sub categories**.
2. We can say that **one segment** has a **larger width** than **another segment** but we can't distinguish how large it is(**numerically**).Even the **stacked bar charts** have the same disadvantage.
3. Comparisons of **rectangle sizes** are complicated since the rectangles can vary in size in the sense of **aspect ratios.**Looking at the below image, can we say which rectangle(left orange vs right green) has high value?

   We cannot say them because they vary in heights and widths.But the overall area of those rectangles are the same .i.e., they both have the same value.

4. As the rectangles in the Mekko charts are at different heights with different baseline, it becomes difficult to compare them.



These figures explain the above disadvantage.In the **left figure** we cannot exactly say which is the tallest rectangle among A,B,C as they don't have a **common baseline** whereas in the **right image** we can say **C is the tallest**.

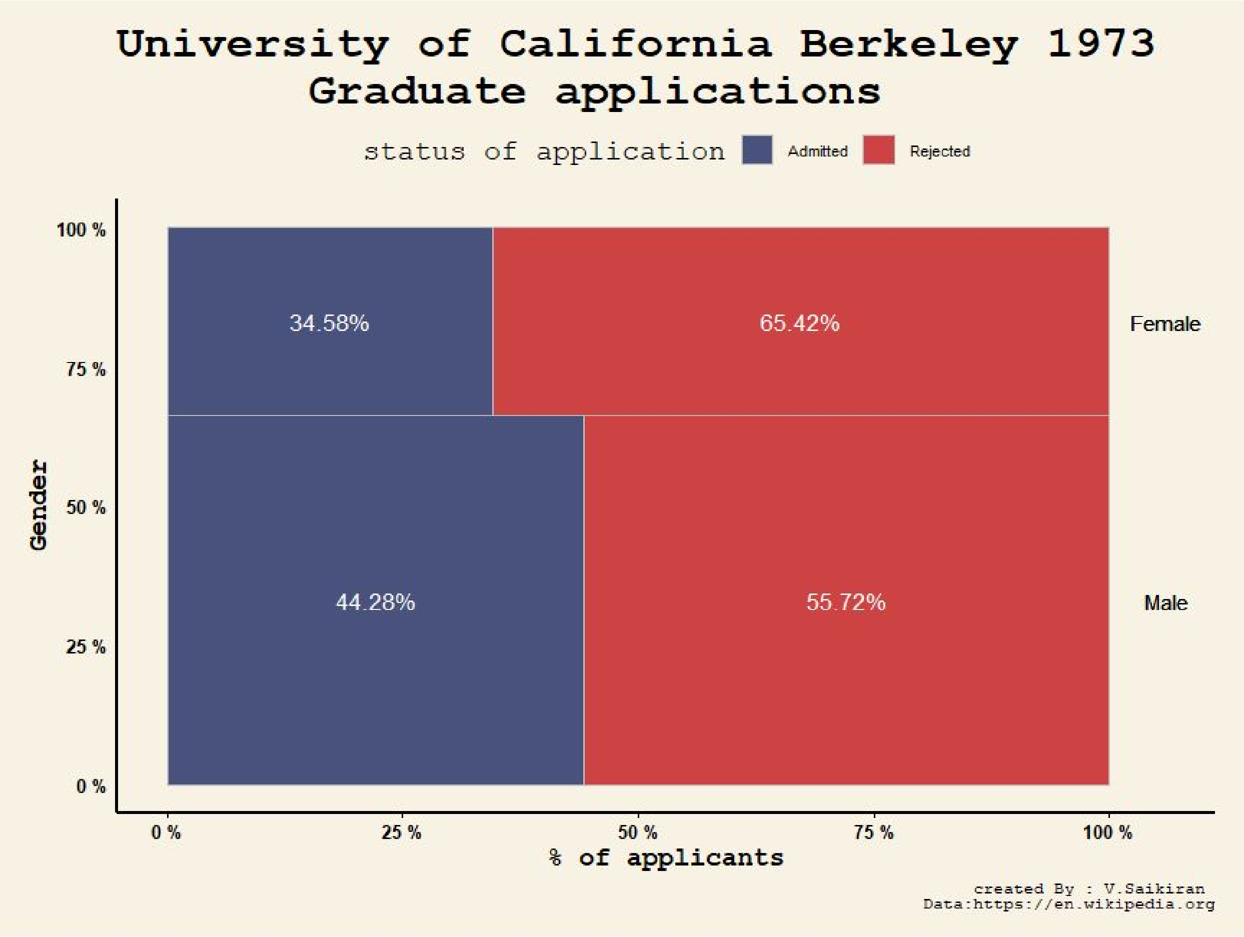
Note: We can overcome some of these **disadvantages** by representing **actual numbers or percentages** over every rectangle using which the **comparisons** become easy.

Is there a possibility to enhance the visual information by using additional attributes?

1. A **Marimekko chart** with **x, y axis labelled**, **height** and **width** of the **bars** properly **distinguished**, **different colors** used to **distinguish** the **subcategories** inside every bar is **well enough** to get a lot of insights about the data.
2. Adding **more no. of attributes** makes the chart **clumsy** and becomes **difficult** to **read** and **understand**.

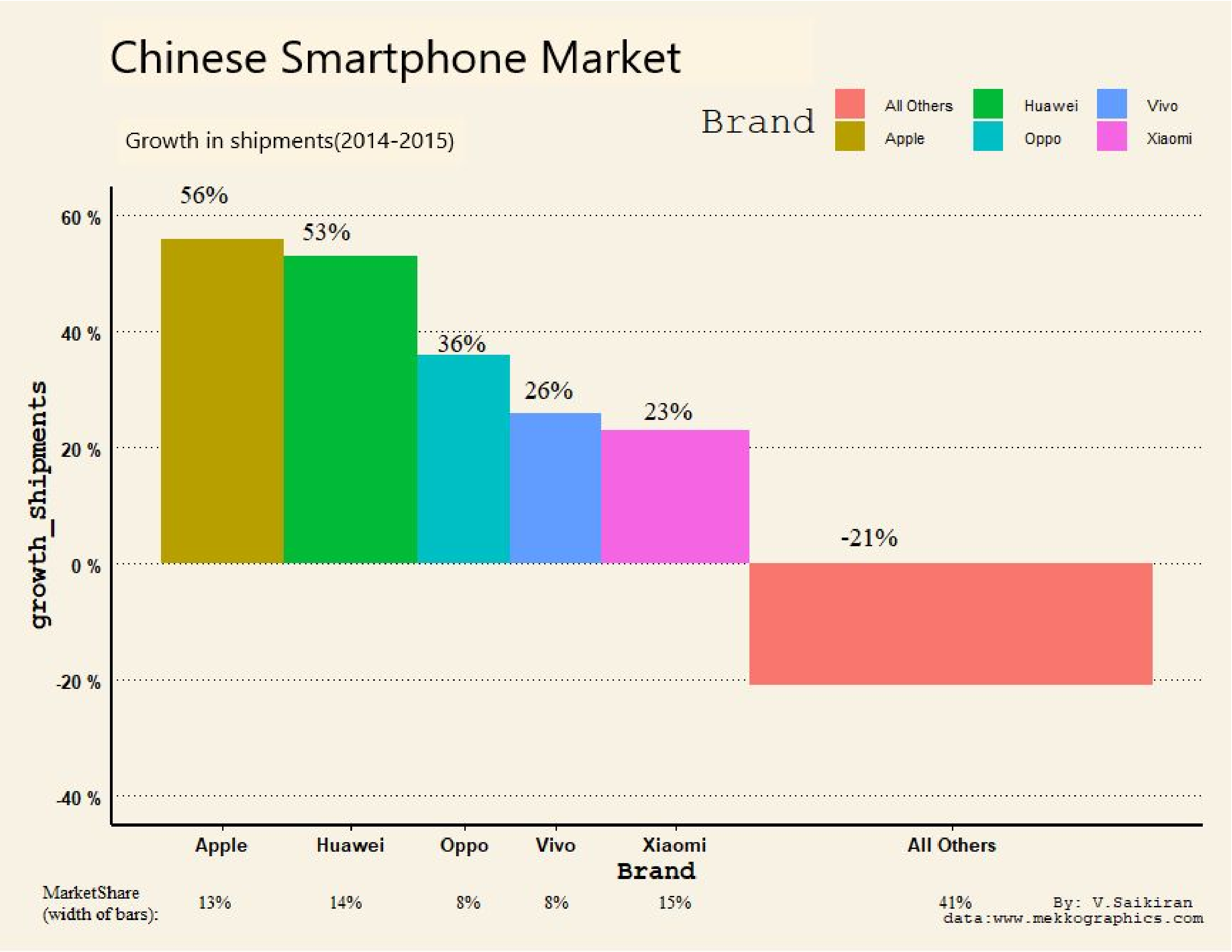
If there is any variant of this chart?

Yes, the Marimekko chart has got a variant which has horizontal columns instead of vertical columns. Below chart demonstrates this:



Gender is represented as the bars on y axis and % of applicants on the x axis. This is a horizontal version of the previous chart that we saw earlier.

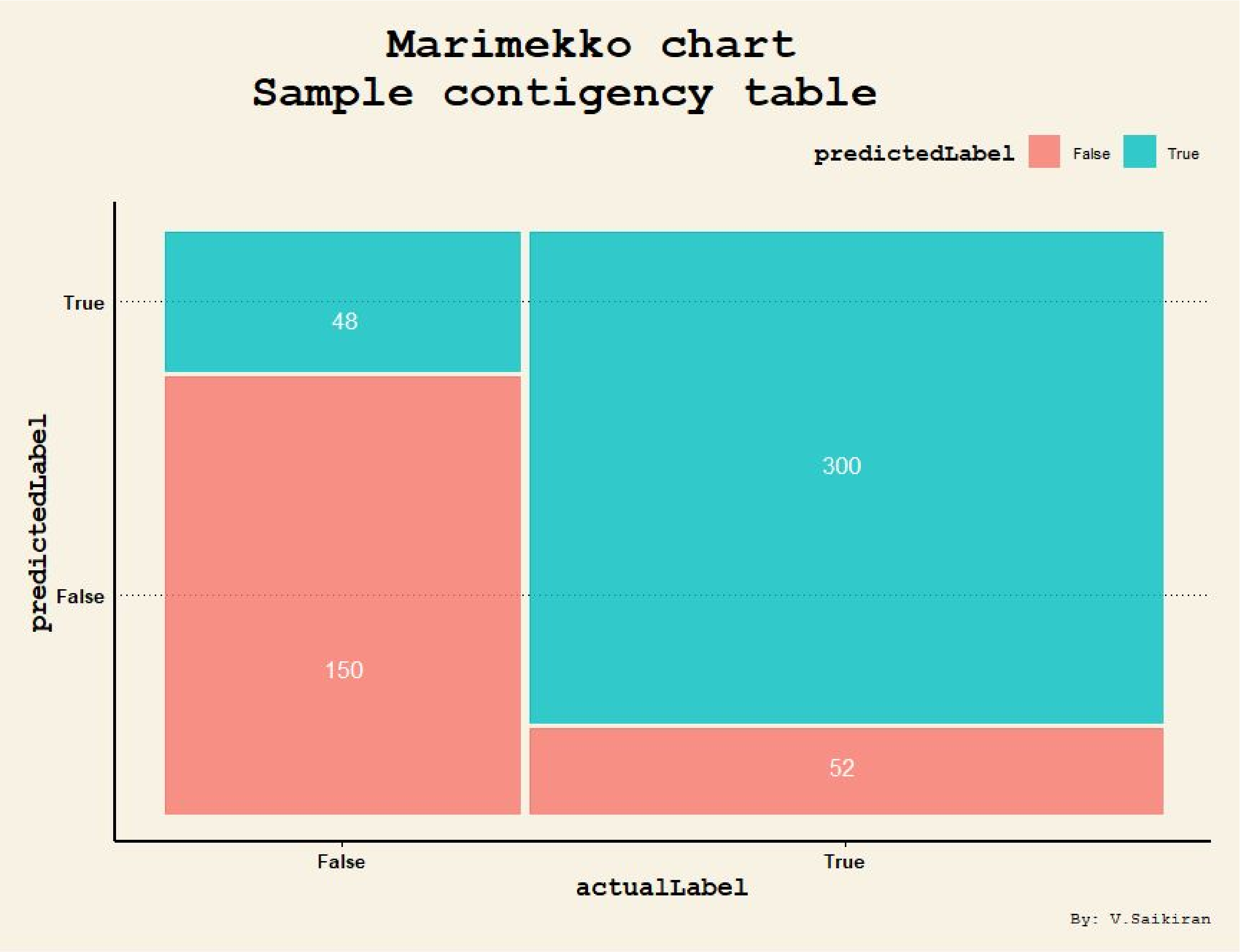
We have another chart which almost depicts the same kind of data. A sister version of Marimekko chart, the Bar Mekko chart.Below chart demonstrates this.



- Here we are comparing the growth in shipments of different mobile brands in the year (2014-2015) across china.
- Same as Marimekko charts, this chart also represents us with 2 axes having 100%. This chart additionally allows us to have a negative axis values on the y axis.
- A Bar Mekko chart can be used for showing market growth by segment.
- The growth rate is the **bar height** and **segment revenue** is the **bar width**. This chart would be useful for discussing the growth rate differences for large and small segments.

What are the famous applications where marimekko charts are used?

- One of the use cases involves showcasing of the **frequencies** of a **contingency table(**confusion matrix in Machine learning**)** where the area of each displayed cell is proportional to the whole



- These marimekko charts are widely used by **analysts** to make **business strategies**.
- These charts can also be used for visualizing **business processes**, e.g. to compare the **market share** of several **products** of several **brands.**

How can this chart be generated using R and ggplot? If additional libraries are required then discuss them. If R cannot generate this chart, then what open-source option one can use.

- Yes, one can use **ggplot** to build this chart. With the help of **geom_rect** which takes **xmin, xmax, ymin, ymax(start and end points of the rectangles on the plot).**
- There is a package named **ggmosaic** which provides us with **geom_mosaic** using which we can plot the **marimekko**(or **mosaic**) **chart**. I did not use it as it was not giving a proper chart for my data.
- One can also use the **mosaic** available **statsmodels.graphics.mosaicplot** a python library.

  Sample code for python package

  ```
  From statsmodels.graphics.mosaicplot import mosaic
  import matplotlib.pyplot as plt
  import seaborn as sns
  mosaic(dataset, [categoryVariabe1, categoryvariable2,....])
  plt.show()
  ```

I have used tableau public to get one of my marimekko charts(one with department wise acceptance/ rejection rates) as R packages are not giving proper charts for my data.

I am including the R codes with explanations to create Marimekko and barMekko charts.

Marimekko chart

Saikiran (194161016)

28/05/2020

```
#stacked bar plot
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(ggthemes)
library(ggplot2)


#creating the dataframe with required values like Gender(male, female)
application status(accepted or rejected), percentages showing the acceptance
rates for male and female
df12 <- data.frame(
  Gender = rep(c("Male", "Female"), each=2),
  status = rep(c("Accepted", "Rejected"),2),
  len = c(44.28,55.72,34.58,65.42))

#using geom_col to plot the stacked bar plot
p <- ggplot(df12, aes(x = Gender, y = len))+
  geom_col(aes(fill = status), width = 0.5)


#adding additional attributes like themes, setting axis ticks, positioning
the legend etc.
p+scale_colour_wsj("colors6", "") +
  theme_wsj()+
  labs(x = "Gender", y = "% of applicants",fill="status of application") +
  theme(text =
element_text(size=10),axis.title=element_text(size=14),axis.title.x =
element_text(size=14, face="bold", colour = "black"),    plot.caption =
element_text(size='8'),
    axis.title.y = element_text(size=14, face="bold", colour = "black"),
legend.justification = "top",
    axis.line.x = element_line(color="black", size = 1),
    axis.line.y = element_line(color="black", size = 1),
     panel.grid.major = element_line(colour = NA),legend.title =
element_text(color = "black", size = 15),
     panel.grid.minor = element_line(colour = NA))+scale_fill_brewer(palette
= "Set2")+scale_fill_manual(values =
c("#49517D","#CD4343"))+scale_y_continuous(label=function(x){return(paste(x,
"%"))})+labs(title = "University of California Berkeley 1973 \n
Graduate applications", caption = "created By : V.Saikiran \n
Data:https://en.wikipedia.org")
```

plot is removed as I have already demonstrated it in the earlier part of

report

creating a data frame which has segment names different segments width, and heights of the bars(partitioned w.r.t acceptance rates).

```
df <- data.frame(segment = c("Male", "Female"), segpct = c(66.14, 33.86),
Admitted = c(44.28,34.58), Rejected = c(55.72,65.42))
```

calculating the starting and ending points of the segments for x axis.

```
df$xmax <- cumsum(df$segpct)
df$xmin <- df$xmax - df$segpct
df$segpct <- NULL
df

##   segment Admitted Rejected   xmax  xmin
## 1    Male    44.28    55.72  66.14  0.00
## 2  Female    34.58    65.42 100.00 66.14
```

melting the data frame and storing start and end points of every combination of Gender and application status.

```
library(ggplot2)
library(reshape2)
library(plyr)


dfm <- melt(df, id=c("segment","xmin","xmax"))
dfm

##   segment  xmin   xmax variable value
## 1    Male  0.00  66.14 Admitted 44.28
## 2  Female 66.14 100.00 Admitted 34.58
## 3    Male  0.00  66.14 Rejected 55.72
## 4  Female 66.14 100.00 Rejected 65.42
```

calculating the start and end points for the acceptance rates on the y axis and storing them in the data frame.

```
dfm1 <- ddply(dfm, .(segment), transform, ymax = cumsum(value))
dfm1 <- ddply(dfm1, .(segment), transform,
     ymin = ymax - value)
dfm1

##   segment  xmin   xmax variable value   ymax  ymin
## 1  Female 66.14 100.00 Admitted 34.58  34.58  0.00
## 2  Female 66.14 100.00 Rejected 65.42 100.00 34.58
## 3    Male  0.00  66.14 Admitted 44.28  44.28  0.00
## 4    Male  0.00  66.14 Rejected 55.72 100.00 44.28
```

calculating the position to store the text showing the acceptance/ rejection percentages.

```
dfm1$xtext <- with(dfm1, xmin + (xmax - xmin)/2)
dfm1$ytext <- with(dfm1, ymin + (ymax - ymin)/2)
```

plotting the graph with necessary themes, capions, font sizes, axis labels, axis ticks, titles, legend positioning.
With the help of geom_rect which requires start and end points of rectangles on x as well as y axis. We have supplied the xmin, xmax, ymin,ymax that we have calculated earlier.

```
p <- ggplot(dfm1, aes(ymin = ymin, ymax = ymax,
     xmin = xmin, xmax = xmax, fill = variable))
p1 <- p + geom_rect(colour = I("grey"))
p2 <- p1 + geom_text(aes(x = xtext, y = ytext,label = paste(value,"%", sep =
""), size = 2),color="white")
p3 <- p2 + geom_text(aes(x = xtext, y = 103,
     label = paste(segment)), size = 4)+guides(size=FALSE)

p3 +scale_colour_wsj("colors6", "") +
  theme_wsj()+
  labs(x = "Gender", y = "% of applicants",fill="status of application") +
  theme(text =
element_text(size=10),axis.title=element_text(size=14),axis.title.x =
element_text(size=14, face="bold", colour = "black"),    plot.caption =
element_text(size='8'),
    axis.title.y = element_text(size=14, face="bold", colour = "black"),
legend.justification = "top",
    axis.line.x = element_line(color="black", size = 1),
    axis.line.y = element_line(color="black", size = 1),
     panel.grid.major = element_line(colour = NA),legend.title =
element_text(color = "black", size = 15),
     panel.grid.minor = element_line(colour = NA))+scale_fill_brewer(palette
= "Set2")+scale_fill_manual(values =
c("#49517D","#CD4343"))+scale_y_continuous(label=function(x){return(paste(x,
"%"))})+scale_x_continuous(label=function(x){return(paste(x,
"%"))})+labs(title = "University of California Berkeley 1973 \n
Graduate applications", caption = "created By : V.Saikiran \n
Data:https://en.wikipedia.org")
```

plot is removed as I have already demonstrated it in the earlier part of report

plotting the horizontal version by making x to y and y to axis and small modifications.

```
#code for variation in marimekko horizontal marimekko chart
p <- ggplot(dfm1, aes(ymin = xmin, ymax = xmax,
     xmin = ymin, xmax = ymax, fill = variable))
p1 <- p + geom_rect(colour = I("grey"))
dfm1$xtext
```

```
## [1] 83.07 83.07 33.07 33.07
```

```
p2 <- p1 + geom_text(aes(x = ytext, y = xtext,label = paste(value,"%", sep =
""), size = 2),color="white")
p3 <- p2 + geom_text(aes(x = 106, y = xtext,
     label = paste(segment)), size = 4)+guides(size=FALSE)

p3 +scale_colour_wsj("colors6", "") +
  theme_wsj()+
  labs(x = "% of applicants", y = "Gender",fill="status of application") +
  theme(text =
element_text(size=10),axis.title=element_text(size=14),axis.title.x =
element_text(size=14, face="bold", colour = "black"),    plot.caption =
element_text(size='8'),
    axis.title.y = element_text(size=14, face="bold", colour = "black"),
legend.justification = "top",
    axis.line.x = element_line(color="black", size = 1),
    axis.line.y = element_line(color="black", size = 1),
     panel.grid.major = element_line(colour = NA),legend.title =
element_text(color = "black", size = 15),
     panel.grid.minor = element_line(colour = NA))+scale_fill_brewer(palette
= "Set2")+scale_fill_manual(values =
c("#49517D","#CD4343"))+scale_y_continuous(label=function(x){return(paste(x,
"%"))})+scale_x_continuous(label=function(x){return(paste(x,
"%"))})+labs(title = "University of California Berkeley 1973 \n
Graduate applications", caption = "created By : V.Saikiran \n
Data:https://en.wikipedia.org")
```

plot is removed as I have already demonstrated it in the earlier part of report

plotting the bar mekko chart using the bar mekko chart available in mekko library.

```r
#bar mekko chart a sister of marimekko chart
library(mekko)
library(ggplot2)
df <- data.frame(
  Brand = c('Apple', 'Huawei', 'Oppo', 'Vivo','Xiaomi','All Others'),
  marketShare = c(200, 220, 150, 150,240,660),
  growth_Shipments = c(56,53,36,26, 23 ,-21)
  )

barMekko <- barmekko(df[order(-df$growth_Shipments),], Brand,
growth_Shipments, marketShare)
barMekko +
  labs(title = 'Growth shipments 2014-2015',caption = 'By: V.Saikiran \n
data:www.mekkographics.com')+
  scale_y_continuous(limits=c(-40, 60),label=function(x){return(paste(x,
"%"))})+
  scale_colour_wsj("colors6", "") +
  theme_wsj()+
  theme(text =
element_text(size=10),axis.title=element_text(size=14),axis.title.x =
element_text(size=14, face="bold", colour = "black"),plot.caption =
element_text(size='8'),
    axis.title.y = element_text(size=14, face="bold", colour = "black"),
legend.justification = "right",
    axis.line.x = element_line(color="black", size = 1),
    axis.line.y = element_line(color="black", size = 1))
```

plot is removed as I have already demonstrated it in the earlier part of report

plot demonstrating an application of Marimekko chart for contingency tables(confusion matrix)

```r
library(ggplot2)
library(ggmosaic)

df <-
data.frame(actualLabel=c('True','True','False','False'),predictedLabel=c('Tru
e','False','True','False'),value=c(300,52,48,150))
ggplot(data=df) +
  geom_mosaic(aes(weight=value, x=product(actualLabel),
fill=predictedLabel))+geom_text(aes(x = c(0.68,0.68,0.18,0.18), y =
c(0.6,0.08,0.85,0.25),label = paste(value), size =
2),color="white")+labs(title = '          Marimekko chart \n     Sample
contigency table',caption = 'By: V.Saikiran')+
  scale_colour_wsj("colors6", "") +
  theme_wsj()+
  theme(text = element_text(size=10),legend.title =
element_text(face='bold',size = 13),
  legend.text = element_text(size =
8),axis.title=element_text(size=14),axis.title.x = element_text(size=14,
face="bold", colour = "black"),plot.caption = element_text(size='8'),
    axis.title.y = element_text(size=14, face="bold", colour = "black"),
legend.justification = "right",
    axis.line.x = element_line(color="black", size = 1),
    axis.line.y = element_line(color="black", size = 1))+guides(size=FALSE)
```

plot is removed as I have already demonstrated it in the earlier part of report