Librarian Census Data Code

Dan Anthony Dorado

2022-07-07

Table of Contents

# Abstract

The University of the Philippines School of Library and Information Studies (UPSLIS) conducted a study of Philippine Librarians from November 2018 to October 2019. This endeavor was the first-ever wide-scale study undertaken to collect the occupational profile of Librarians in the Philippines. The survey was designed to be an online questionnaire via Mentimeter and was presented for dissemination during the 2018 Philippine Librarians Association Inc. (PLAI) Congress at Novotel in Quezon City. To further disseminate, the link was posted on the UPSLIS Facebook account and was shared to library association group pages. A total of 685 responses were recorded which is only 7.5% of the almost 9000 registered librarians in the Philippines but can be used in making statistical inference with a 4% margin of error.

# About this document

This paper aims to discuss the methds and process used in cleaning, visuzlizing, and processing of the **Philippine Librarian Census Dataset** *(see* [*https://zenodo.org/record/6450452*](https://zenodo.org/record/6450452)*).* It details the:

* R packages used
* Data cleaning methods
* Data wrangling methods
* Data summarization methods
* Data computation methods
* Data visualization methods

The researchers hope that this would provide a framework for future LIS research to be more open by sharing theirs datasets and documentation. This would allow research verification and provide an opportunity for other researchers to use different theories and approaches to shared datasets in order to promote the culture of research in the LIS field.

# R packages and dataset

# Load needed package  
library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.2.1

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

## ✔ ggplot2 3.3.6 ✔ purrr 0.3.4  
## ✔ tibble 3.1.7 ✔ dplyr 1.0.9  
## ✔ tidyr 1.2.0 ✔ stringr 1.4.0  
## ✔ readr 2.1.2 ✔ forcats 0.5.1

## Warning: package 'ggplot2' was built under R version 4.2.1

## Warning: package 'tidyr' was built under R version 4.2.1

## Warning: package 'readr' was built under R version 4.2.1

## Warning: package 'dplyr' was built under R version 4.2.1

## Warning: package 'forcats' was built under R version 4.2.1

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(tm)

## Warning: package 'tm' was built under R version 4.2.1

## Loading required package: NLP

##   
## Attaching package: 'NLP'

## The following object is masked from 'package:ggplot2':  
##   
## annotate

library(infer)

## Warning: package 'infer' was built under R version 4.2.1

#Load the raw "librariancensus.csv" dataset  
dataset <- read\_csv("librariancensus.csv")

## Rows: 685 Columns: 67

## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (62): licensed, philippines, gender, working, industry, institution, ten...  
## dbl (5): voter, age, years\_service, gross\_salary, net\_salary  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

#Inspect the loaded dataset  
glimpse(dataset)

## Rows: 685  
## Columns: 67  
## $ voter <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,…  
## $ licensed <chr> "Yes, I have a valid and up-to-date PRC license fo…  
## $ philippines <chr> NA, "No, I am in the Philippines right now but onl…  
## $ gender <chr> NA, "Prefer not to say", "Woman", "Man", "Man", "M…  
## $ age <dbl> 59, 100, 40, 46, 32, 46, 48, 44, 25, 24, 34, 39, 3…  
## $ working <chr> NA, "No, I am a retiree.", "Yes, I am currently em…  
## $ industry <chr> NA, "information technology", "academic institutio…  
## $ years\_service <dbl> NA, 60, 14, 23, 10, 24, 23, 12, 5, 1, 12, 18, 14, …  
## $ institution <chr> NA, "private (BPO)", "government", "government", "…  
## $ tenure <chr> NA, "self-employed, freelancing or working with mu…  
## $ gross\_salary <dbl> NA, 100, 51, 53, 40, 56, 36, 38, 19, 11, 95, 58, 3…  
## $ net\_salary <dbl> NA, 100, 30, 38, 30, 42, 29, 26, NA, 10, 95, 36, 2…  
## $ benefits1 <chr> NA, "night differential pay", "paid vacation and/o…  
## $ benefits2 <chr> NA, NA, "monetary allowances", "monetary allowance…  
## $ benefits3 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, "h…  
## $ benefits4 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…  
## $ benefits5 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…  
## $ otherbenefits1 <chr> NA, "123", "Rice allowance\r\nClothing allowance\r…  
## $ otherbenefits2 <chr> NA, "123", NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,…  
## $ otherbenefits3 <chr> NA, "Hello Philippines and hello world", NA, NA, N…  
## $ otherbenefits4 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…  
## $ materials1 <chr> NA, "law materials and resources", NA, "archives (…  
## $ materials2 <chr> NA, NA, NA, "hardcopy theses and dissertations", N…  
## $ materials3 <chr> NA, NA, NA, "hardcopy books / pamphlets / etc", NA…  
## $ materials4 <chr> NA, NA, NA, "institutional files and records", NA,…  
## $ materials5 <chr> NA, NA, NA, "data sets", NA, "archives (physical)"…  
## $ materials6 <chr> NA, NA, NA, "electronic material (digital titles, …  
## $ materials7 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…  
## $ othermaterials1 <chr> NA, "Huehs", "Grading sheets\r\nExam papers", "Boo…  
## $ othermaterials2 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…  
## $ othermaterials3 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…  
## $ othermaterials4 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…  
## $ position <chr> NA, "supervisory / mid-management level", "supervi…  
## $ fiveyears <chr> NA, "No (planning to shift to other fields or indu…  
## $ pursue <chr> NA, "Hfeudnr", NA, NA, NA, NA, NA, NA, NA, NA, NA,…  
## $ education <chr> NA, "Doctor of Philosophy (Ph.D)", "Master\x92s de…  
## $ enrolled <chr> NA, "Yes", "Yes", "Yes", "No.", "Yes", "Yes", "No.…  
## $ completing <chr> NA, "Post-graduate Certificate or Diploma (non-LIS…  
## $ study\_again <chr> NA, NA, "No", "Yes", "Yes", "Yes", "Yes", "Yes", "…  
## $ `5yearsdegree` <chr> NA, NA, NA, "Short | Certificate | Diploma Program…  
## $ cpd1 <chr> NA, NA, "trainings and symposiums facilitated by p…  
## $ cpd2 <chr> NA, NA, "government facilitated trainings and symp…  
## $ cpd3 <chr> NA, NA, NA, NA, "within the work place (in-house t…  
## $ cpd4 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…  
## $ cpdsupport1 <chr> NA, NA, "from the government (scholarships, free t…  
## $ cpdsupport2 <chr> NA, NA, NA, "from the government (scholarships, fr…  
## $ cpdsupport3 <chr> NA, NA, NA, "from the work place (private sponsors…  
## $ cpdsatisfaction <chr> NA, NA, "Yes, the amount of trainings I received w…  
## $ cpdbarriers1 <chr> NA, NA, "CPD program sources that I really like", …  
## $ cpdbarriers2 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…  
## $ cpdbarriers3 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…  
## $ topics1 <chr> NA, NA, "Ethics\r\nCommunication skills\r\nCooking…  
## $ topics2 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…  
## $ topics3 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…  
## $ topics4 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…  
## $ topics5 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…  
## $ location <chr> NA, NA, "NCR", "NCR", "NCR", "NCR", "NCR", "NCR", …  
## $ worktravel <chr> NA, NA, "Yes", "Yes", "Yes", "Yes", "No", "No", "N…  
## $ filipino <chr> NA, NA, "No", "Yes", "Yes", "Yes", "No", "Yes", "Y…  
## $ dialect <chr> NA, NA, "Ilocano", NA, "Others", NA, "Kapampangan"…  
## $ foreignlang <chr> NA, NA, "None", NA, "Chinese", NA, NA, "English", …  
## $ type <chr> NA, NA, "Slis", "Library school", NA, "Library sch…  
## $ roles <chr> NA, NA, "Cleaner\r\nCook", NA, "Faculty", "Asst pr…  
## $ institutionposition <chr> NA, NA, "Head", "Asst prof 5", "Asst. Prof.", "Ass…  
## $ primaryjob <chr> NA, NA, NA, "Education", "Research, Extention, Tea…  
## $ affiliation <chr> NA, NA, "Plai\r\nCodhlis\r\nPatls", "Plai, patls, …  
## $ factor <chr> NA, NA, "Friends", "Similar thrusts", "colleagues"…

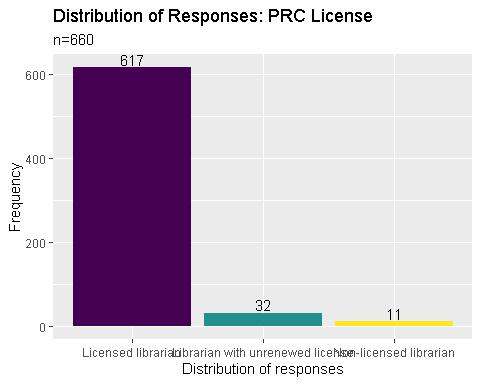
#controls how many digits will be printed  
options(digits=4)   
  
#Initialize a pseudorandom number generator to obtain the same output during simulation  
set.seed(12)

# Data Cleaning, Wrangling, and Visualization

## Question: Are you currently a licensed librarian (R.A. 9246) ?

* Yes, I have a valid and up-to-date PRC license for librarians
* No, I have not renewed my license
* No, I have never taken any licensure examination for librarians and/or never applied for a PRC license for librarians

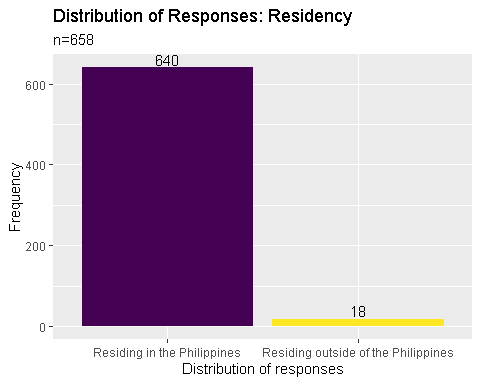
# Change the factor type  
dataset$licensed <- as.factor(dataset$licensed)  
# Recode responses  
dataset$licensed <- fct\_recode(dataset$licensed, "Licensed librarian" = "Yes, I have a valid and up-to-date PRC license for librarians", "Librarian with unrenewed license" = "No, I have not renewed my license", "Non-licensed librarian" = "No, I have never taken any licensure examination for librarians and/or never applied for a PRC license for librarians")  
# Rearrange the responses  
dataset$licensed <- factor(dataset$licensed, levels = c("Licensed librarian", "Librarian with unrenewed license", "Non-licensed librarian"))  
# Graph the data  
dataset %>%  
filter(!is.na(licensed)) %>%  
ggplot(aes(x = licensed, fill = licensed)) + geom\_bar() + geom\_text(aes(label = ..count..), stat = "count", vjust = -0.2, colour = "black") + xlab("Distribution of responses") + ylab("Frequency") + ggtitle("Distribution of Responses: PRC License", "n=660") + guides(fill = "none") + scale\_fill\_viridis\_d()



## Question: Are you currently residing in the Philippines?

* Yes, I am currently residing and staying in the Philippines
* No, I am in the Philippines right now but only for vacation
* No, I am currently residing outside the Philippines

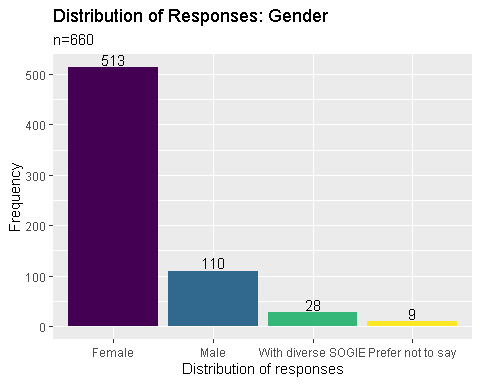
# Change the factor type  
dataset$philippines <- as.factor(dataset$philippines)  
# Recode responses  
dataset$philippines <- fct\_recode(dataset$philippines, "Residing in the Philippines" = "Yes, I am currently residing and staying in the Philippines", "Residing outside of the Philippines" = "No, I am in the Philippines right now but only for vacation", "Residing outside of the Philippines" = "No, I am currently residing outside the Philippines")  
# Rearrange the responses  
dataset$philippines <- factor(dataset$philippines, levels = c("Residing in the Philippines", "Residing outside of the Philippines"))  
# Graph the data  
dataset %>%  
filter(!is.na(philippines)) %>%  
ggplot(aes(x = philippines, fill = philippines)) + geom\_bar() + geom\_text(aes(label = ..count..), stat = "count", vjust = -0.2, colour = "black") + xlab("Distribution of responses") + ylab("Frequency") + ggtitle("Distribution of Responses: Residency", "n=658") + guides(fill = "none") + scale\_fill\_viridis\_d()



## Question: Kindly indicate which gender you identify the most

* Man
* Woman
* LGBTQ+
* Prefer not to say

# Change the factor type  
dataset$gender <- as.factor(dataset$gender)  
# Recode responses  
dataset$gender <- fct\_recode(dataset$gender, "With diverse SOGIE" = "LGBTQ+", "Male" = "Man", "Female" = "Woman")  
# Rearrange the responses  
dataset$gender <- factor(dataset$gender, levels = c("Female", "Male", "With diverse SOGIE", "Prefer not to say"))  
# Graph the data  
dataset %>%  
filter(!is.na(gender)) %>%  
ggplot(aes(x = gender, fill = gender)) + geom\_bar() + geom\_text(aes(label = ..count..), stat = "count", vjust = -0.2, colour = "black") + xlab("Distribution of responses") + ylab("Frequency") + ggtitle("Distribution of Responses: Gender", "n=660") + guides(fill = "none") + scale\_fill\_viridis\_d()

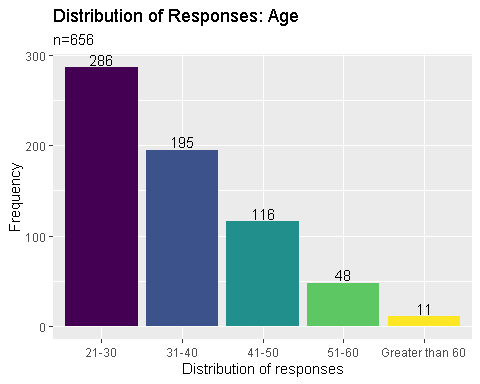


## Question: Kindly indicate your age

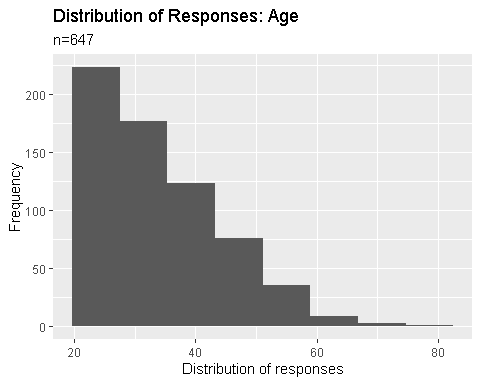
# Inspect the responses  
summary(dataset$age)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 18.0 25.0 31.0 33.8 40.0 100.0 24

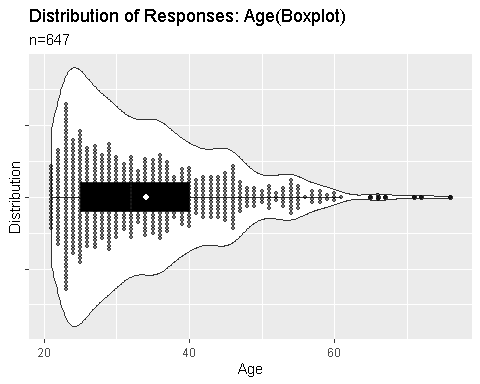
# Create a new column: age\_group  
age\_group <- cut(dataset$age, c(20,30,40,50,60,70,80), right = FALSE)  
# Attach age\_group to dataset  
dataset <- cbind(dataset, age\_group)  
# Recode responses  
dataset$age\_group <- fct\_recode(dataset$age\_group, "21-30" = '[20,30)', "31-40" = '[30,40)', "41-50" = '[40,50)', "51-60" = '[50,60)', "Greater than 60" = '[60,70)', "Greater than 60" = '[70,80)' )  
# Graph the data: bar graph  
dataset %>%  
filter(!is.na(age\_group)) %>%  
ggplot(aes(x = age\_group, fill = age\_group)) + geom\_bar() + geom\_text(aes(label = ..count..), stat = "count", vjust = -0.2, colour = "black") + xlab("Distribution of responses") + ylab("Frequency") + ggtitle("Distribution of Responses: Age", "n=656") + guides(fill = "none") + scale\_fill\_viridis\_d()



# Graph the data: histogram  
dataset %>%  
filter(!is.na(age)) %>%  
filter(age > 20) %>%  
filter(age < 90) %>%  
ggplot(aes(x = age, fill = age)) + geom\_histogram(bins = 8) + xlab("Distribution of responses") + ylab("Frequency") + ggtitle("Distribution of Responses: Age", "n=647") + guides(fill = "none") + scale\_fill\_viridis\_d()



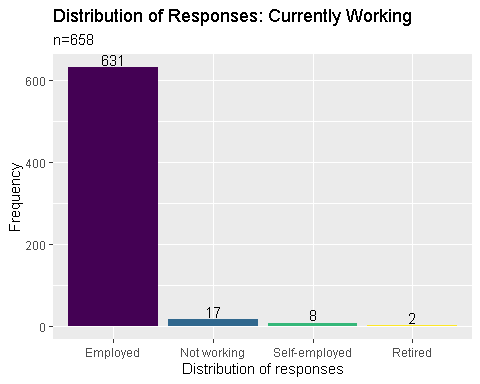
# Graph the data: boxplot + violin plot + dot plot  
dataset %>%  
filter(!is.na(age)) %>%  
filter(age > 20) %>%  
filter(age < 90) %>%  
ggplot(aes(x = 1, y = age)) + geom\_violin(adjust=.7) + geom\_boxplot(width = .1, fill = "black") + geom\_dotplot(aes(y = age), binaxis = "y", binwidth = .5, stackdir = "center", alpha = .5) + stat\_summary(fun = mean, geom = "point", fill = "white", shape = 21, size = 2.5) + coord\_flip() + theme(axis.text.y = element\_blank()) + xlab("Distribution") + ylab("Age") + ggtitle("Distribution of Responses: Age(Boxplot)", "n=647")



## Question: Are you currently employed and/or working?

* Yes, I am currently employed by an institution, agency, individual or similar.
* Yes, I am currently self-employed and/or doing freelance and consultancy work.
* No, I am currently not working for reasons other than retirement.
* No, I am a retiree.

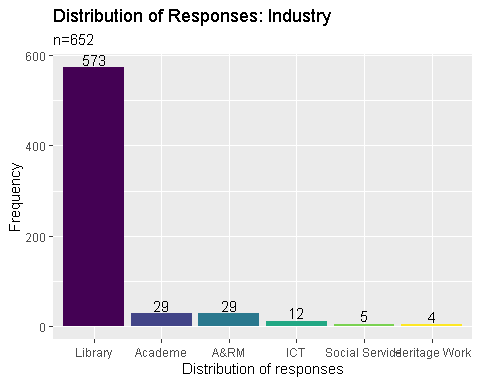
# Change the factor type  
dataset$working <- as.factor(dataset$working)  
# Recode responses  
dataset$working <- fct\_recode(dataset$working, "Retired" = 'No, I am a retiree.', "Not working" = 'No, I am currently not working for reasons other than retirement.', "Employed" = 'Yes, I am currently employed by an institution, agency, individual or similar.', "Self-employed" = 'Yes, I am currently self-employed and/or doing freelance and consultancy work.')  
# Rearrange the responses  
dataset$working <- factor(dataset$working, levels = c('Employed', 'Not working', 'Self-employed', 'Retired'))  
# Graph the data  
dataset %>%  
filter(!is.na(working)) %>%  
ggplot(aes(x = working, fill = working)) + geom\_bar() + geom\_text(aes(label = ..count..), stat = "count", vjust = -0.2, colour = "black") + xlab("Distribution of responses") + ylab("Frequency") + ggtitle("Distribution of Responses: Currently Working", "n=658") + guides(fill = "none") + scale\_fill\_viridis\_d()



## Question: Kindly indicate the setting, field or industry you are primarily working in

* library
* archives and/or records management
* academic institution (non-librarian position)
* information technology, computer science, software / database engineering and other similar line of work
* museum and heritage work
* social services

# Change the factor type  
dataset$industry <- as.factor(dataset$industry)  
# Recode responses  
dataset$industry <- fct\_recode(dataset$industry, "Academe" = 'academic institution (non-librarian position)',"A&RM" = 'archives and/or records management', "ICT" = 'information technology', "Library" = 'library', "Heritage Work" ='museum and heritage work', "Social Service" = 'social services')  
# Rearrange the responses  
dataset$industry <- factor(dataset$industry, levels = c('Library', 'Academe', 'A&RM', 'ICT', 'Social Service', 'Heritage Work'))  
# Graph the data  
dataset %>%  
filter(!is.na(industry)) %>%  
ggplot(aes(x = industry, fill = industry)) + geom\_bar() + geom\_text(aes(label = ..count..), stat = "count", vjust = -0.2, colour = "black") + xlab("Distribution of responses") + ylab("Frequency") + ggtitle("Distribution of Responses: Industry", "n=652") + guides(fill = "none") + scale\_fill\_viridis\_d()

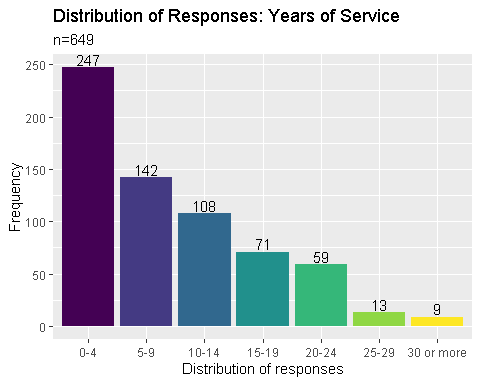


## Question: How long have you been working as a professional after passing the licensure exam for librarians (cumulative years)

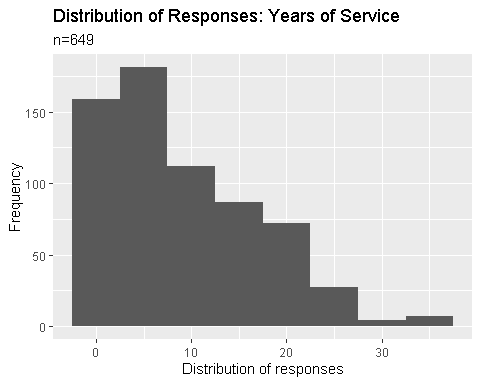
# Inspect the responses  
summary(dataset$years\_service)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 1.00 3.00 7.00 9.22 14.00 60.00 34

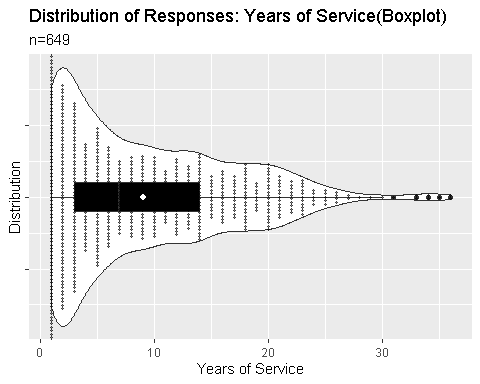
# Create a new column: years\_service\_group  
years\_service\_group <- cut(dataset$years\_service, c(0,5,10,15,20,25,30,35,40), right = FALSE)  
# Attach years\_service\_group to dataset  
dataset <- cbind(dataset, years\_service\_group)  
# Recode responses  
dataset$years\_service\_group <- fct\_recode(dataset$years\_service\_group, "0-4" = '[0,5)',"5-9" = '[5,10)', "10-14" = '[10,15)', "15-19" = '[15,20)', "20-24" = '[20,25)', "25-29" = '[25,30)',"30 or more" = '[30,35)', "30 or more" = '[35,40)')  
# Graph the data: bar graph  
dataset %>%  
filter(!is.na(years\_service\_group)) %>%  
ggplot(aes(x = years\_service\_group, fill = years\_service\_group)) + geom\_bar() + geom\_text(aes(label = ..count..), stat = "count", vjust = -0.2, colour = "black") + xlab("Distribution of responses") + ylab("Frequency") + ggtitle("Distribution of Responses: Years of Service", "n=649") + guides(fill = "none") + scale\_fill\_viridis\_d()



# Graph the data: histogram  
dataset %>%  
filter(!is.na(years\_service)) %>%  
filter(years\_service < 40) %>%  
ggplot(aes(x = years\_service, , fill = years\_service)) + geom\_histogram(bins = 8) + xlab("Distribution of responses") + ylab("Frequency") + ggtitle("Distribution of Responses: Years of Service", "n=649") + guides(fill = "none") + scale\_fill\_viridis\_d()



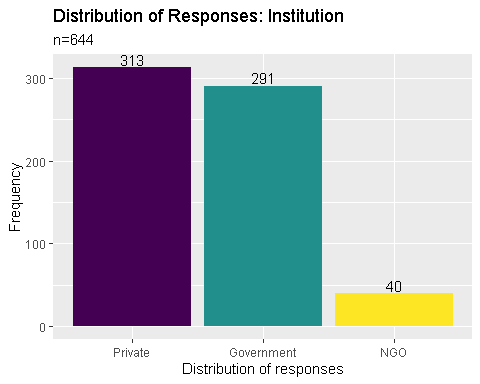
# Graph the data: boxplot + violin plot + dot plot  
dataset %>%  
filter(!is.na(years\_service)) %>%  
filter(years\_service < 40) %>%  
ggplot(aes(x = 1, y = years\_service)) + geom\_violin(adjust=.7) + geom\_boxplot(width = .1, fill = "black") + geom\_dotplot(aes(y = years\_service), binaxis = "y", binwidth = .3, stackdir = "center", alpha = .5) + stat\_summary(fun = mean, geom = "point", fill = "white", shape = 21, size = 2.5) + coord\_flip() + theme(axis.text.y = element\_blank()) + xlab("Distribution") + ylab("Years of Service") + ggtitle("Distribution of Responses: Years of Service(Boxplot)", "n=649")



## Question: Kindly indicate the type of institution (in terms of ownership) you are primarily working in

* government
* private (in-house)
* private (BPO)
* non-profit

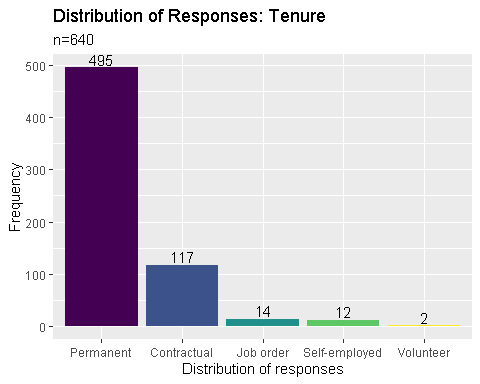
# Change the factor type  
dataset$institution <- as.factor(dataset$institution)  
# Recode responses  
dataset$institution <- fct\_recode(dataset$institution, "Government" = 'government', "NGO" = 'non-profit', "Private" = 'private (BPO)', "Private" = 'private (in-house)')  
# Rearrange the responses  
dataset$institution <- factor(dataset$institution, levels = c('Private', 'Government', 'NGO'))  
# Graph the data  
dataset %>%  
filter(!is.na(institution)) %>%  
ggplot(aes(x = institution, fill = institution)) + geom\_bar() + geom\_text(aes(label = ..count..), stat = "count", vjust = -0.2, colour = "black") + xlab("Distribution of responses") + ylab("Frequency") + ggtitle("Distribution of Responses: Institution", "n=644") + guides(fill = "none") + scale\_fill\_viridis\_d()



## Question: Kindly indicate your current tenure status

* permanent
* contractual
* voluntary work
* job order (J.O.)
* self-employed, freelancing or working with multiple institutions

# Change the factor type  
dataset$tenure <- as.factor(dataset$tenure)  
# Recode responses  
dataset$tenure <- fct\_recode(dataset$tenure, "Contractual" = 'contractual', "Job order" = 'job order (J.O.)', "Permanent" = 'permanent', "Self-employed" = 'self-employed, freelancing or working with multiple institutions', "Volunteer" = 'voluntary work')  
# Rearrange the responses  
dataset$tenure <- factor(dataset$tenure, levels = c('Permanent', 'Contractual', 'Job order', 'Self-employed', 'Volunteer'))  
# Graph the data  
dataset %>%  
filter(!is.na(tenure)) %>%  
ggplot(aes(x = tenure, fill = tenure)) + geom\_bar() + geom\_text(aes(label = ..count..), stat = "count", vjust = -0.2, colour = "black") + xlab("Distribution of responses") + ylab("Frequency") + ggtitle("Distribution of Responses: Tenure", "n=640") + guides(fill = "none") + scale\_fill\_viridis\_d()

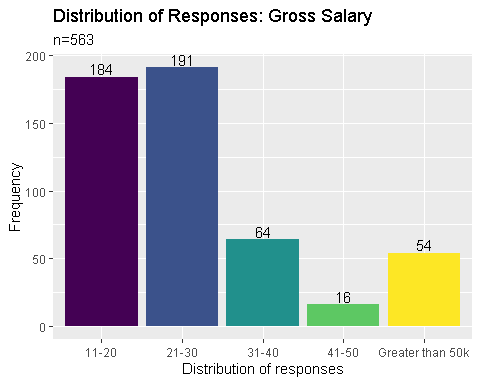


## Question: Kindly indicate your current monthly gross salary or income range (before deductions)

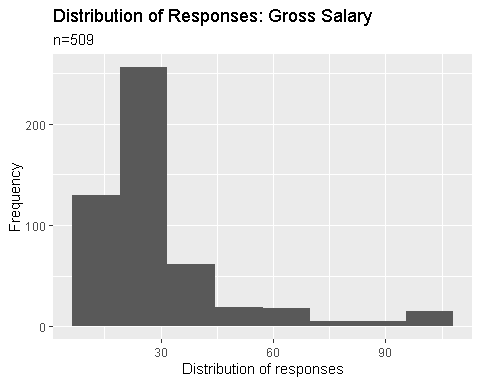
# Inspect the responses  
summary(dataset$gross\_salary)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 1.0 17.0 22.0 26.9 30.0 100.0 122

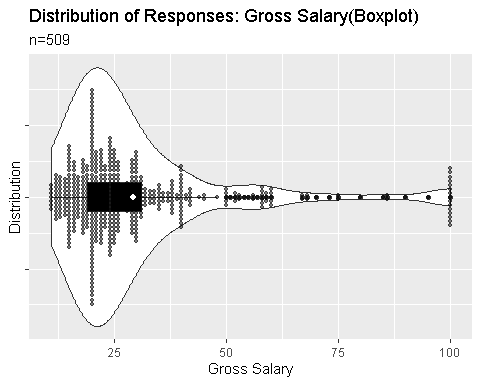
# Create a new column: gross\_salary\_group  
gross\_salary\_group <- cut(dataset$gross\_salary, c(10,20,30,40,50,60,70,80,90,100), right = TRUE)  
# Attach gross\_salary\_group to dataset  
dataset <- cbind(dataset, gross\_salary\_group)  
# Recode responses  
dataset$gross\_salary\_group <- fct\_recode(dataset$gross\_salary\_group, "11-20" = '(10,20]', "21-30" = '(20,30]', "31-40" = '(30,40]', "41-50" = '(40,50]', "Greater than 50k" = '(50,60]', "Greater than 50k" = '(60,70]', "Greater than 50k" = '(70,80]', "Greater than 50k" = '(80,90]', "Greater than 50k" = '(90,100]')  
# Graph the data: bar graph  
dataset %>%   
filter(!is.na(gross\_salary\_group)) %>%  
ggplot(aes(x = gross\_salary\_group, fill = gross\_salary\_group)) + geom\_bar() + geom\_text(aes(label = ..count..), stat = "count", vjust = -0.2, colour = "black") + xlab("Distribution of responses") + ylab("Frequency") + ggtitle("Distribution of Responses: Gross Salary", "n=563") + guides(fill = "none") + scale\_fill\_viridis\_d()



# Graph the data: histogram  
dataset %>%  
filter(!is.na(gross\_salary)) %>%  
filter(gross\_salary > 10) %>%  
ggplot(aes(x = gross\_salary, fill = gross\_salary)) + geom\_histogram(bins = 8) + xlab("Distribution of responses") + ylab("Frequency") + ggtitle("Distribution of Responses: Gross Salary", "n=509") + guides(fill = "none") + scale\_fill\_viridis\_d()



# Graph the data: boxplot + violin plot + dot plot  
dataset %>%  
filter(!is.na(gross\_salary)) %>%  
filter(gross\_salary > 10) %>%  
ggplot(aes(x = 1, y = gross\_salary)) + geom\_violin(adjust=2) + geom\_boxplot(width = .1, fill = "black") + geom\_dotplot(aes(y = gross\_salary), binaxis = "y", binwidth = .9, stackdir = "center", alpha = .5) + stat\_summary(fun = mean, geom = "point", fill = "white", shape = 21, size = 2.5) + coord\_flip() + theme(axis.text.y = element\_blank()) + xlab("Distribution") + ylab("Gross Salary") + ggtitle("Distribution of Responses: Gross Salary(Boxplot)", "n=509")

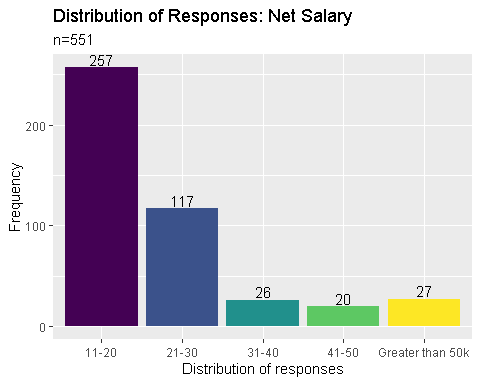


## Kindly indicate your current monthly net salary or income range (after deductions)

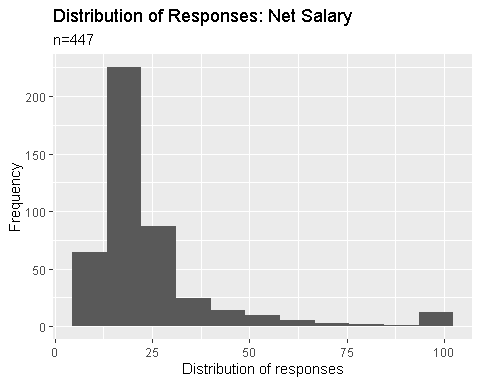
# Inspect the responses  
summary(dataset$net\_salary)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 1.0 12.0 18.0 21.1 24.0 100.0 134

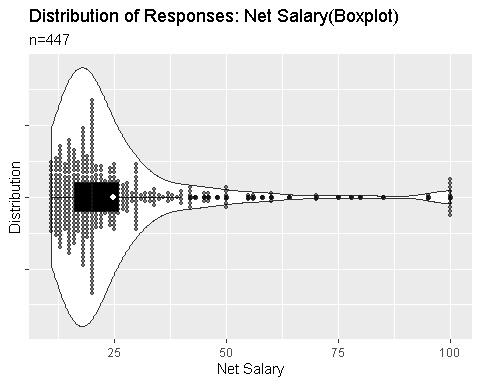
# Create a new column: gross\_salary\_group  
net\_salary\_group <- cut(dataset$net\_salary, c(10,20,30,40,50,60,70,80,90,100), right = TRUE)  
# Attach net\_salary\_group to dataset  
dataset <- cbind(dataset, net\_salary\_group)  
# Recode responses  
dataset$net\_salary\_group <- fct\_recode(dataset$net\_salary\_group, "11-20" = '(10,20]', "21-30" = '(20,30]', "31-40" = '(30,40]', "41-50" = '(40,50]', "Greater than 50k" = '(50,60]', "Greater than 50k" = '(60,70]', "Greater than 50k" = '(70,80]', "Greater than 50k" = '(80,90]', "Greater than 50k" = '(90,100]')  
# Graph the data: bar graph  
dataset %>%  
filter(!is.na(net\_salary\_group)) %>%  
ggplot(aes(x = net\_salary\_group, fill = net\_salary\_group)) + geom\_bar() + geom\_text(aes(label = ..count..), stat = "count", vjust = -0.2, colour = "black") + xlab("Distribution of responses") + ylab("Frequency") + ggtitle("Distribution of Responses: Net Salary", "n=551") + guides(fill = "none") + scale\_fill\_viridis\_d()



# Graph the data: bar graph  
dataset %>%  
filter(!is.na(net\_salary)) %>%  
filter(net\_salary > 10) %>%  
ggplot(aes(x = net\_salary, fill = net\_salary)) + geom\_histogram(bins = 11) + xlab("Distribution of responses") + ylab("Frequency") + ggtitle("Distribution of Responses: Net Salary", "n=447") + guides(fill = "none") + scale\_fill\_viridis\_d()



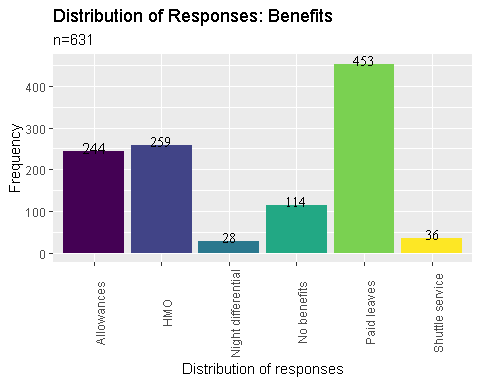
# Graph the data: boxplot + violin plot + dot plot  
dataset %>%  
filter(!is.na(net\_salary)) %>%  
filter(net\_salary > 10) %>%  
ggplot(aes(x = 1, y = net\_salary)) + geom\_violin(adjust=2) + geom\_boxplot(width = .1, fill = "black") + geom\_dotplot(aes(y = net\_salary), binaxis = "y", binwidth = .9, stackdir = "center", alpha = .5) + stat\_summary(fun = mean, geom = "point", fill = "white", shape = 21, size = 2.5) + coord\_flip() + theme(axis.text.y = element\_blank()) + xlab("Distribution") + ylab("Net Salary") + ggtitle("Distribution of Responses: Net Salary(Boxplot)", "n=447")



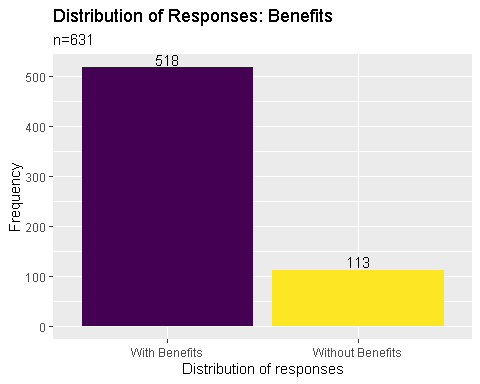
## Question: Kindly indicate additional benefits, if any (please check all that apply)

* free shuttle service
* night differential pay
* paid vacation and/or sick leaves
* health card and/or insurance other than PhilHealth
* monetary allowances
* NONE / Not Applicable

# Combine columns: benefits1, benefits2, benefits3, benefits4, benefits5  
dataset$benefits <- paste(dataset$benefits1, dataset$benefits2, dataset$benefits3, dataset$benefits4, dataset$benefits5)  
#Create new columns: shuttle, night\_differential, paid\_leaves, hmo, allowances, no\_benefits  
dataset <- transform(dataset, shuttle = grepl("shuttle", benefits), night\_differential = grepl("night", benefits), paid\_leaves = grepl("leaves", benefits), hmo = grepl("insurance", benefits), allowances = grepl("allowances", benefits), no\_benefits = grepl("NONE", benefits))  
# Summarize by counting the TRUE values per benefits  
dataset %>%  
summarise("Shuttle service" = sum(shuttle == TRUE), "Night differential" = sum(night\_differential == TRUE), "Paid leaves" = sum(paid\_leaves == TRUE), "HMO" = sum(hmo == TRUE), "Allowances" = sum(allowances == TRUE), "No benefits" = sum(no\_benefits == TRUE)) %>%  
gather() %>%  
ggplot(aes(x = key, y = value, fill = key)) + geom\_col(position = "dodge") + xlab("Distribution of responses") + ylab("Frequency") + ggtitle("Distribution of Responses: Benefits", "n=631") + guides(fill = "none") + scale\_fill\_viridis\_d() + theme(axis.text.x = element\_text(angle = 90)) +  
geom\_text(x = 1, y = 254, label = "244", colour = "black", size = 4) +   
geom\_text(x = 2, y = 269, label = "259", family = "serif", colour = "black", size = 4) +  
geom\_text(x = 3, y = 38, label = "28", family = "serif", colour = "black", size = 4) +  
geom\_text(x = 4, y = 124, label = "114", family = "serif", colour = "black", size = 4) +  
geom\_text(x = 5, y = 463, label = "453", family = "serif", colour = "black", size = 4) +  
geom\_text(x = 6, y = 46, label = "36", family = "serif", colour = "black", size = 4)

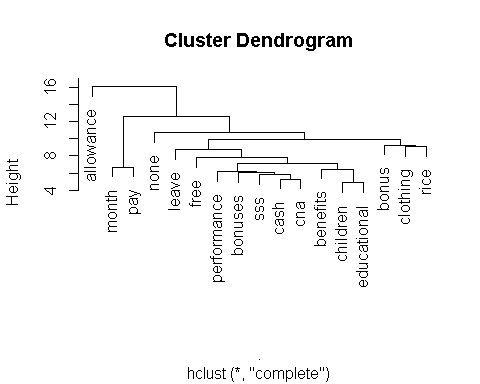


# Change the factor type  
dataset$benefits1 <- as.factor(dataset$benefits1)  
# Recode responses  
dataset$benefits6 <- fct\_recode(dataset$benefits1, "With Benefits" = 'free shuttle service', "With Benefits" = 'health card and/or insurance other than PhilHealth', "With Benefits" = 'monetary allowances', "With Benefits" = 'night differential pay', "Without Benefits" = 'NONE / Not Applicable', "With Benefits" = 'paid vacation and/or sick leaves')  
# Rearrange the responses  
dataset$benefits6 <- factor(dataset$benefits6, levels = c('With Benefits', 'Without Benefits'))  
# Graph the data  
dataset %>%  
filter(!is.na(benefits6)) %>%  
ggplot(aes(x = benefits6, fill = benefits6)) + geom\_bar() + geom\_text(aes(label = ..count..), stat = "count", vjust = -0.2, colour = "black") + xlab("Distribution of responses") + ylab("Frequency") + ggtitle("Distribution of Responses: Benefits", "n=631") + guides(fill = "none") + scale\_fill\_viridis\_d()



## Question: Other benefits not mentioned in the previous page

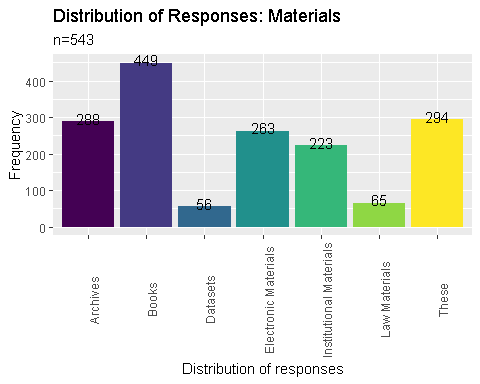
# Create a clean\_corpus function   
clean\_corpus <- function(corpus) {  
 corpus <- tm\_map(corpus, removePunctuation)  
 corpus <- tm\_map(corpus, removeNumbers)  
 corpus <- tm\_map(corpus, content\_transformer(tolower))  
 corpus <- tm\_map(corpus, removeWords, c(stopwords("en")))  
 corpus <- tm\_map(corpus, stripWhitespace)  
 return(corpus)  
}  
# Create a text dataset  
benefits.text <- paste(dataset$otherbenefits1, dataset$otherbenefits2, dataset$otherbenefits3, dataset$otherbenefits4)  
# Create a text corpus  
benefit.corpus <- benefits.text %>%  
VectorSource() %>%  
VCorpus()  
# Clean the created corpus  
benefit.corpus <- clean\_corpus(benefit.corpus)  
# Create a dendogram  
benefit.corpus %>%  
TermDocumentMatrix() %>%  
removeSparseTerms(sparse = 0.98) %>%  
as.matrix() %>%  
dist() %>%  
hclust() %>%  
plot()



## Question: What kind of materials are you currently working with? Kindly check all that applies.

* hardcopy theses and dissertations
* hardcopy books / pamphlets / etc
* electronic material (digital titles, digital archives, etc)
* law materials and resources
* archives (physical)
* data sets
* institutional files and records
* NA

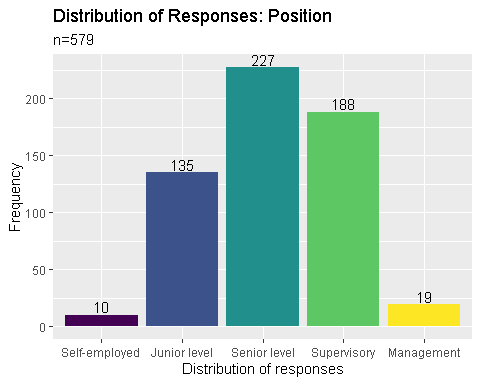
# Combine columns: materials1, materials2, materials3, materials4, materials5, materials6, materials7  
dataset$materials <- paste(dataset$materials1, dataset$materials2, dataset$materials3, dataset$materials4, dataset$materials5, dataset$materials6, dataset$materials7)  
# Create new columns: theses, books, electronic\_materials, law\_materials, datasets, institutional\_files  
dataset <- transform(dataset, theses = grepl("theses", materials), books = grepl("books", materials), electronic\_materials = grepl("electronic", materials), law\_materials = grepl("law", materials), datasets = grepl("data sets", materials), archives = grepl("archives", materials), institutional\_files = grepl("institutional files and records", materials))  
# Summarize by counting the TRUE values per materials  
dataset %>%  
summarise("These" = sum(theses == TRUE), "Books" = sum(books == TRUE), "Electronic Materials" = sum( electronic\_materials == TRUE), "Law Materials" = sum(law\_materials == TRUE), "Datasets" = sum(datasets == TRUE), "Archives" = sum(archives == TRUE), "Institutional Materials" = sum(institutional\_files == TRUE)) %>%  
gather() %>%  
ggplot(aes(x = key, y= value, fill = key)) + geom\_col(position = "dodge") + xlab("Distribution of responses") + ylab("Frequency") + ggtitle("Distribution of Responses: Materials", "n=543") + guides(fill = "none") + scale\_fill\_viridis\_d() + theme(axis.text.x = element\_text(angle = 90)) +  
geom\_text(x = 1, y = 298, label = "288", colour = "black", size = 4) +   
geom\_text(x = 2, y = 459, label = "449", colour = "black", size = 4) +   
geom\_text(x = 3, y = 66, label = "56", colour = "black", size = 4) +  
geom\_text(x = 4, y = 273, label = "263", colour = "black", size = 4) +  
geom\_text(x = 5, y = 233, label = "223", colour = "black", size = 4) +  
geom\_text(x = 6, y = 75, label = "65", colour = "black", size = 4) +  
geom\_text(x = 7, y = 304, label = "294", colour = "black", size = 4)



## Question: Kindly indicate your current work position

* entry level / junior staff
* regular / senior staff (non-supervisory)
* supervisory / mid-management level
* upper management level / top tier executive
* consultant / freelancer / self-employed

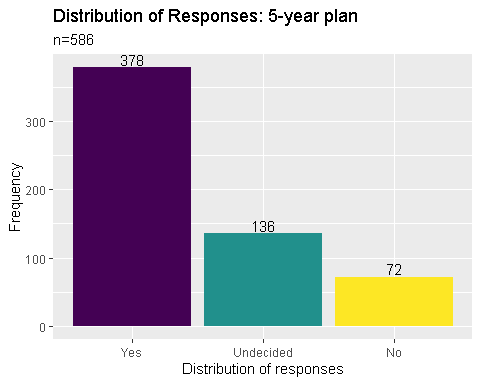
# Change the factor type  
dataset$position <- as.factor(dataset$position)  
# Recode responses  
dataset$position <- fct\_recode(dataset$position, "Self-employed" = 'consultant / freelancer / self-employed', "Junior level" = 'entry level / junior staff', "Senior level" = 'regular / senior staff (non-supervisory)', "Supervisory" = 'supervisory / mid-management level', "Management" = 'upper management level / top tier executive')  
# Graph the data  
dataset %>%  
filter(!is.na(position)) %>%  
ggplot(aes(x = position, fill = position)) + geom\_bar() + geom\_text(aes(label = ..count..), stat = "count", vjust = -0.2, colour = "black") + xlab("Distribution of responses") + ylab("Frequency") + ggtitle("Distribution of Responses: Position", "n=579") + guides(fill = "none") + scale\_fill\_viridis\_d()



## Question: Within the next five (5) years, do you see your self working in the same industry?

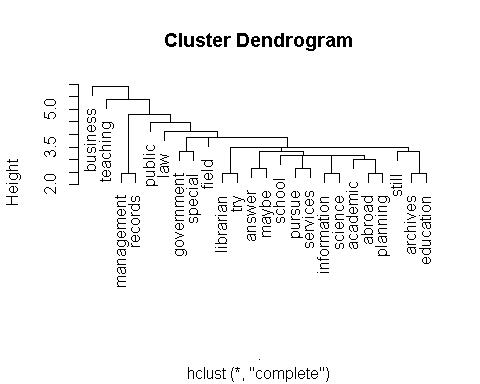
* Yes
* No (planning to shift to other fields or industry)
* Maybe / Undecided

# Change the factor type  
dataset$fiveyears <- as.factor(dataset$fiveyears)  
# Recode responses  
dataset$fiveyears <- fct\_recode(dataset$fiveyears, "Undecided" = 'Maybe / Undecided', "No" = 'No (planning to shift to other fields or industry)', "Yes" = 'Yes')  
# Rearrange the responses  
dataset$fiveyears <- factor(dataset$fiveyears, levels = c('Yes', 'Undecided', 'No'))  
# Graph the data  
dataset %>%  
filter(!is.na(fiveyears)) %>%  
ggplot(aes(x = fiveyears, fill = fiveyears)) + geom\_bar() + geom\_text(aes(label = ..count..), stat = "count", vjust = -0.2, colour = "black") + xlab("Distribution of responses") + ylab("Frequency") + ggtitle("Distribution of Responses: 5-year plan", "n=586") + guides(fill = "none") + scale\_fill\_viridis\_d()



## Question: If your answer was NO or MAYBE in the previous page, indicate the field you wish to pursue

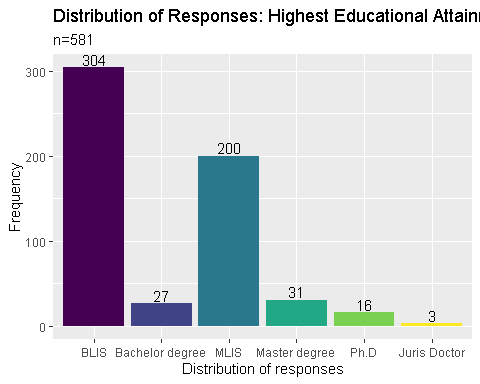
# Create a clean\_corpus function   
clean\_corpus <- function(corpus) {  
 corpus <- tm\_map(corpus, removePunctuation)  
 corpus <- tm\_map(corpus, removeNumbers)  
 corpus <- tm\_map(corpus, content\_transformer(tolower))  
 corpus <- tm\_map(corpus, removeWords, c(stopwords("en"), "none", "yes", "like", "library", "librarianship", "want", "work"))  
 corpus <- tm\_map(corpus, stripWhitespace)  
 return(corpus)  
}  
# Create a text corpus  
fiveyearplan.corpus <- dataset$pursue %>%  
VectorSource() %>%  
VCorpus()  
# Clean the created corpus  
fiveyearplan.corpus <- clean\_corpus(fiveyearplan.corpus)  
# Create a dendogram  
fiveyearplan.corpus %>%  
TermDocumentMatrix() %>%  
removeSparseTerms(sparse = 0.995) %>%  
as.matrix() %>%  
dist() %>%  
hclust() %>%  
plot()



## Question: Kindly indicate your highest educational attainment

* High School
* Bachelor’s degree (BLIS)
* Bachelor’s degree (non-BLIS)
* Master’s degree (MLIS)
* Master’s degree (non-MLIS)
* Juris Doctor / Bachelor of Laws
* Doctor of Philosophy (Ph.D)

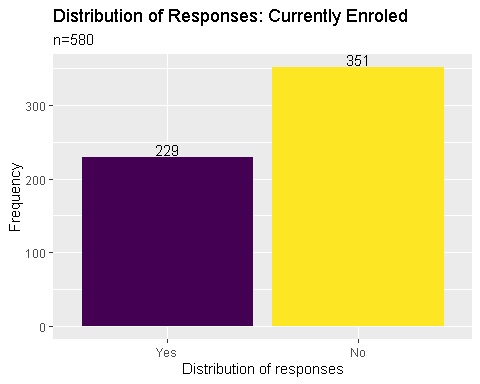
# Change the factor type  
dataset$education <- as.factor(dataset$education)  
# Recode responses  
levels(dataset$education)[1] <- "Bachelor degree"  
levels(dataset$education)[2] <- "BLIS"  
levels(dataset$education)[3] <- "MLIS"  
levels(dataset$education)[4] <- "Master degree"  
levels(dataset$education)[5] <- "Ph.D"  
levels(dataset$education)[6] <- "Juris Doctor"  
# Rearrange the responses  
dataset$education <- fct\_relevel(dataset$education, 'BLIS', 'Bachelor degree', 'MLIS', 'Master degree', 'Ph.D', 'Juris Doctor')  
# Graph the data  
dataset %>%  
filter(!is.na(education)) %>%  
ggplot(aes(x = education, fill = education)) + geom\_bar() + geom\_text(aes(label = ..count..), stat = "count", vjust = -0.2, colour = "black") + xlab("Distribution of responses") + ylab("Frequency") + ggtitle("Distribution of Responses: Highest Educational Attainment", "n=581") + guides(fill = "none") + scale\_fill\_viridis\_d()



## Question: Are you currently enrolled in formal studies? (Short Courses, Certificate Program, Diploma Program, Master Program, PhD)

* Yes
* No

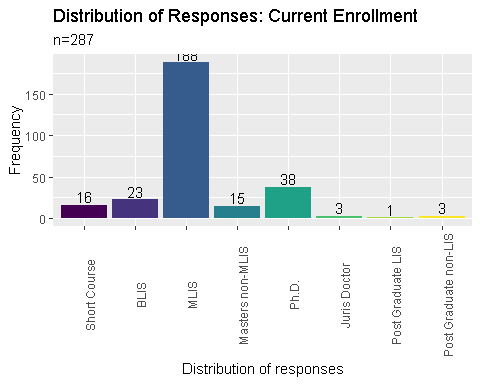
# Change the factor type  
dataset$enrolled <- as.factor(dataset$enrolled)  
# Recode responses  
levels(dataset$enrolled)[1] <- "No"  
# Rearrange the responses  
dataset$enrolled <- fct\_relevel(dataset$enrolled, 'Yes', 'No')  
# Graph the data  
dataset %>%  
filter(!is.na(enrolled)) %>%  
ggplot(aes(x = enrolled, fill = enrolled)) + geom\_bar() + geom\_text(aes(label = ..count..), stat = "count", vjust = -0.2, colour = "black") + xlab("Distribution of responses") + ylab("Frequency") + ggtitle("Distribution of Responses: Currently Enroled", "n=580") + guides(fill = "none") + scale\_fill\_viridis\_d()



## Question: If your answer was YES, indicate the degree you are currently completing

* Short | Certificate | Diploma Program
* Bachelor’s degree (BLIS)
* Master’s degree (MLIS)
* Master’s degree (non-MLIS)
* Juris Doctor / Bachelor of Laws
* Doctor of Philosophy (Ph.D)
* Post-graduate Certificate or Diploma (LIS)
* Post-graduate Certificate or Diploma (non-LIS)

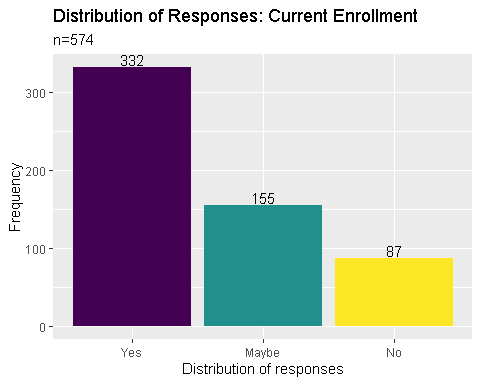
# Change the factor type  
dataset$completing <- as.factor(dataset$completing)  
# Recode responses  
levels(dataset$completing)[1] <- "BLIS"  
levels(dataset$completing)[2] <- "Masters non-MLIS"  
levels(dataset$completing)[3] <- "MLIS"  
levels(dataset$completing)[4] <- "Ph.D."  
levels(dataset$completing)[5] <- "Juris Doctor"  
levels(dataset$completing)[6] <- "Post Graduate LIS"  
levels(dataset$completing)[7] <- "Post Graduate non-LIS"  
levels(dataset$completing)[8] <- "Short Course"  
# Rearrange the responses  
dataset$completing <- factor(dataset$completing, levels = c("Short Course", "BLIS", "MLIS", "Masters non-MLIS", "Ph.D.", "Juris Doctor", "Post Graduate LIS", "Post Graduate non-LIS"))  
# Graph the data  
dataset %>%  
filter(!is.na(completing)) %>%  
ggplot(aes(x = completing, fill = completing)) + geom\_bar() + geom\_text(aes(label = ..count..), stat = "count", vjust = -0.2, colour = "black") + xlab("Distribution of responses") + ylab("Frequency") + ggtitle("Distribution of Responses: Current Enrollment", "n=287") + guides(fill = "none") + theme(axis.text.x = element\_text(angle = 90)) + scale\_fill\_viridis\_d()



## Question: Are you planning to study again in a formal university setting within the next five (5) years?

* Yes
* No
* Maybe

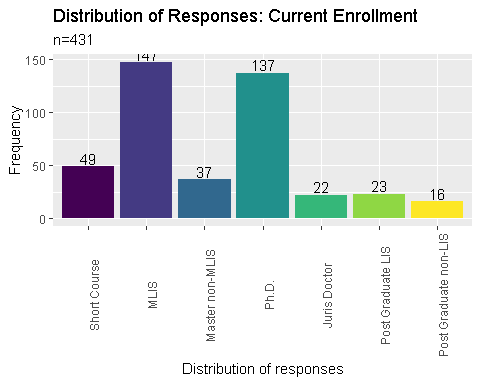
# Change the factor type  
dataset$study\_again <- as.factor(dataset$study\_again)  
# Rearrange the responses  
dataset$study\_again <- factor(dataset$study\_again, levels = c("Yes", "Maybe", "No"))  
# Graph the data  
dataset %>%  
filter(!is.na(study\_again)) %>%  
ggplot(aes(x = study\_again, fill = study\_again)) + geom\_bar() + geom\_text(aes(label = ..count..), stat = "count", vjust = -0.2, colour = "black") + xlab("Distribution of responses") + ylab("Frequency") + ggtitle("Distribution of Responses: Current Enrollment", "n=574") + guides(fill = "none") + scale\_fill\_viridis\_d()



## Question: If your answer was YES, kindly provide the degree you wish to study within the next five (5) years

* Short | Certificate | Diploma Program
* Bachelor’s degree (BLIS)
* Master’s degree (MLIS)
* Master’s degree (non-MLIS)
* Juris Doctor / Bachelor of Laws
* Doctor of Philosophy (Ph.D)
* Post-graduate Certificate or Diploma (LIS)
* Post-graduate Certificate or Diploma (non-LIS)

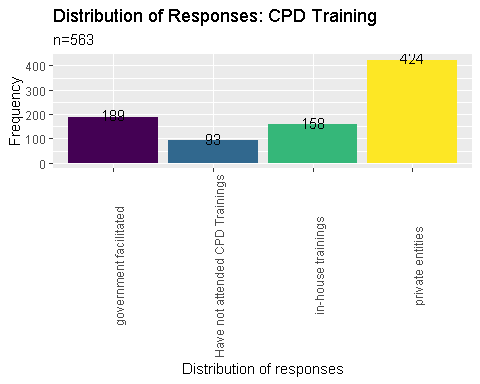
# Change the column name of '5yearsdegree'  
colnames(dataset)[40] <- "after5years"  
# Change the factor type  
dataset$after5years <- as.factor(dataset$after5years)  
# Recode responses  
levels(dataset$after5years)[1] <- "Master non-MLIS"  
levels(dataset$after5years)[2] <- "MLIS"  
levels(dataset$after5years)[3] <- "Ph.D."  
levels(dataset$after5years)[4] <- "Juris Doctor"  
levels(dataset$after5years)[5] <- "Post Graduate LIS"  
levels(dataset$after5years)[6] <- "Post Graduate non-LIS"  
levels(dataset$after5years)[7] <- "Short Course"  
# Rearrange the responses  
dataset$after5years <- factor(dataset$after5years, levels = c("Short Course", "MLIS", "Master non-MLIS", "Ph.D.", "Juris Doctor", "Post Graduate LIS", "Post Graduate non-LIS"))  
# Graph the data  
dataset %>%  
filter(!is.na(after5years)) %>%  
ggplot(aes(x = after5years, fill = after5years)) + geom\_bar() + geom\_text(aes(label = ..count..), stat = "count", vjust = -0.2, colour = "black") + xlab("Distribution of responses") + ylab("Frequency") + ggtitle("Distribution of Responses: Current Enrollment", "n=431") + guides(fill = "none") + theme(axis.text.x = element\_text(angle = 90)) + scale\_fill\_viridis\_d()



## Question: Where do you usually acquire your CPD trainings in LIS? Kindly check all that applies

* within the work place (in-house trainings)
* government facilitated trainings and symposiums
* trainings and symposiums facilitated by private entities, individuals and/or corporations
* NA (Have not attended CPD Trainings yet)

# Combine columns: cpd1, cpd2, cpd3, cpd4  
dataset$cpd <- paste(dataset$cpd1, dataset$cpd2, dataset$cpd3, dataset$cpd4)  
# Create new columns: inhouse, government, private, xattended  
dataset <- transform(dataset, inhouse = grepl("within the work place", cpd), government = grepl("government", cpd), private = grepl("private", cpd), xattended = grepl("attended", cpd))  
# Summarize by counting the TRUE values per materials  
dataset %>%  
summarise("in-house trainings" = sum(inhouse == TRUE), "government facilitated" = sum(government == TRUE), "private entities" = sum(private == TRUE), "Have not attended CPD Trainings" = sum(xattended == TRUE)) %>%  
gather() %>%  
ggplot(aes(x = key, y= value, fill = key)) + geom\_col(position = "dodge") + xlab("Distribution of responses") + ylab("Frequency") + ggtitle("Distribution of Responses: CPD Training", "n=563") + guides(fill = "none") + scale\_fill\_viridis\_d() + theme(axis.text.x = element\_text(angle = 90)) +  
geom\_text(x = 1, y = 199, label = "189", colour = "black", size = 4) +   
geom\_text(x = 2, y = 103, label = "93", colour = "black", size = 4) +   
geom\_text(x = 3, y = 168, label = "158", colour = "black", size = 4) +  
geom\_text(x = 4, y = 434, label = "424", colour = "black", size = 4)



## Where do you usually get funding or support for your trainings in LIS?

* from the work place (private sponsorship)
* from the government (scholarships, free trainings and seminars, etc)
* personal funding (personally attended trainings and symposiums and shouldered most of the expenses)
* NA (Have not attended CPD Trainings yet)

# Combine columns: cpdsupport1, cpdsupport2, cpdsupport3  
dataset$cpdsupport <- paste(dataset$cpdsupport1, dataset$cpdsupport2, dataset$cpdsupport3)  
# Create new columns: support\_inhouse, sgovernment, sprivate, sxattended  
dataset <- transform(dataset, support\_self = grepl("personal", cpdsupport), support\_government = grepl("government", cpdsupport), support\_private = grepl("private", cpdsupport), support\_xattended = grepl("attended", cpdsupport))  
# Summarize by counting the TRUE values per materials  
dataset %>%  
summarise("personal funding" = sum(support\_self == TRUE), "government facilitated" = sum(support\_government == TRUE), "private entities" = sum(support\_private == TRUE), "Have not attended CPD Trainings" = sum(support\_xattended == TRUE)) %>%  
gather() %>%  
ggplot(aes(x = key, y= value, fill = key)) + geom\_col(position = "dodge") + xlab("Distribution of responses") + ylab("Frequency") + ggtitle("Distribution of Responses: CPD Support", "n=563") + guides(fill = "none") + scale\_fill\_viridis\_d() + theme(axis.text.x = element\_text(angle = 90)) +  
geom\_text(x = 1, y = 162, label = "152", colour = "black", size = 4) +  
geom\_text(x = 2, y = 372, label = "362", colour = "black", size = 4) +  
geom\_text(x = 3, y = 300, label = "290", colour = "black", size = 4) +  
geom\_text(x = 4, y = 264, label = "254", colour = "black", size = 4)

