Distribution of Public Service of Canada Employees by Designated Group and Region of Work

checking if our r is working

```
print("Hello R world")
```

Analysis of Table 2

we are loading required libraries

Loading required libraries

```
#install.packages(c("readxl", "dplyr", "ggplot2", "tidyr"))
library(readxl)
library(janitor)
```

```
##
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
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(ggplot2)
library(tidyr)
```

loading the data and cleaning the names

we are going to load the data for table 1 and display the first few rows, just to ensure that our data is loaded successfully

we also cleaned the data to use numbers only, excluding the percentages

```
library(readxl)
tab02_eng <- read_excel("~/Documents/programming/R/velma/keira/tab02-eng.xls", skip = 4, n_max = 18)
```

```
## New names:
## * `` -> `...4`
## * `` -> `...6`
## * `` -> `...8`
## * `` -> `...10`
```

```
Sys.setlocale(category = "LC_CTYPE", locale = "en_US.UTF-8")
```

```
## [1] "en_US.UTF-8"
```

```
head(tab02_eng)
```

```
## # A tibble: 6 x 10
##   `Region of Work`        `All Employees` Women ...4  `Aboriginal Peoples` ...6
##   <chr>                   <chr>           <chr> <chr> <chr>                <chr>
## 1 <NA>                    Number          Numb~ %     Number               %
## 2 Newfoundland and Labra~ 2784            1177  42.2~ 173                  6.21~
## 3 Prince Edward Island    1554            980   63.0~ 44                   2.83~
## 4 Nova Scotia             8349            3567  42.7~ 412                  4.93~
## 5 New Brunswick           6387            3572  55.9~ 233                  3.64~
## 6 Quebec (without the NC~ 20427           10669 52.2~ 420                  2.05~
## # i 4 more variables: `Persons  with Disabilities` <chr>, ...8 <chr>,
## #   `Members of a Visible Minority Group` <chr>, ...10 <chr>
```

```
print(dim(tab02_eng))
```

```
## [1] 18 10
```

```
tab02_eng <- clean_names(tab02_eng)
selected_colnames <- c("region_of_work", "all_employees", "women", "members_of_a_visible_minority_group
head(tab02_eng)
```

```
## # A tibble: 6 x 10
##   region_of_work             all_employees women  x4    aboriginal_peoples x6
##   <chr>                      <chr>         <chr>  <chr> <chr>              <chr>
## 1 <NA>                       Number        Number %     Number             %
## 2 Newfoundland and Labrador  2784          1177   42.2~ 173                6.21~
## 3 Prince Edward Island       1554          980    63.0~ 44                 2.83~
## 4 Nova Scotia                8349          3567   42.7~ 412                4.93~
## 5 New Brunswick              6387          3572   55.9~ 233                3.64~
## 6 Quebec (without the NCR) † 20427         10669  52.2~ 420                2.05~
## # i 4 more variables: persons_with_disabilities <chr>, x8 <chr>,
## #   members_of_a_visible_minority_group <chr>, x10 <chr>
```

```r
subset_data <- tab02_eng[, selected_colnames]
#subset_data <- filter(subset_data, !is.na(region_of_work))
subset_data <- subset_data[complete.cases(tab02_eng$region_of_work), ]
head(subset_data)
```

```
## # A tibble: 6 x 6
##   region_of_work             all_employees women members_of_a_visible_minority~1
##   <chr>                      <chr>         <chr> <chr>
## 1 Newfoundland and Labrador  2784          1177  52
## 2 Prince Edward Island       1554          980   39
## 3 Nova Scotia                8349          3567  490
## 4 New Brunswick              6387          3572  154
## 5 Quebec (without the NCR) † 20427         10669 1999
## 6 NCR (Quebec) †             24333         13927 3162
## # i abbreviated name: 1: members_of_a_visible_minority_group
## # i 2 more variables: persons_with_disabilities <chr>, aboriginal_peoples <chr>
```

visualization of the data

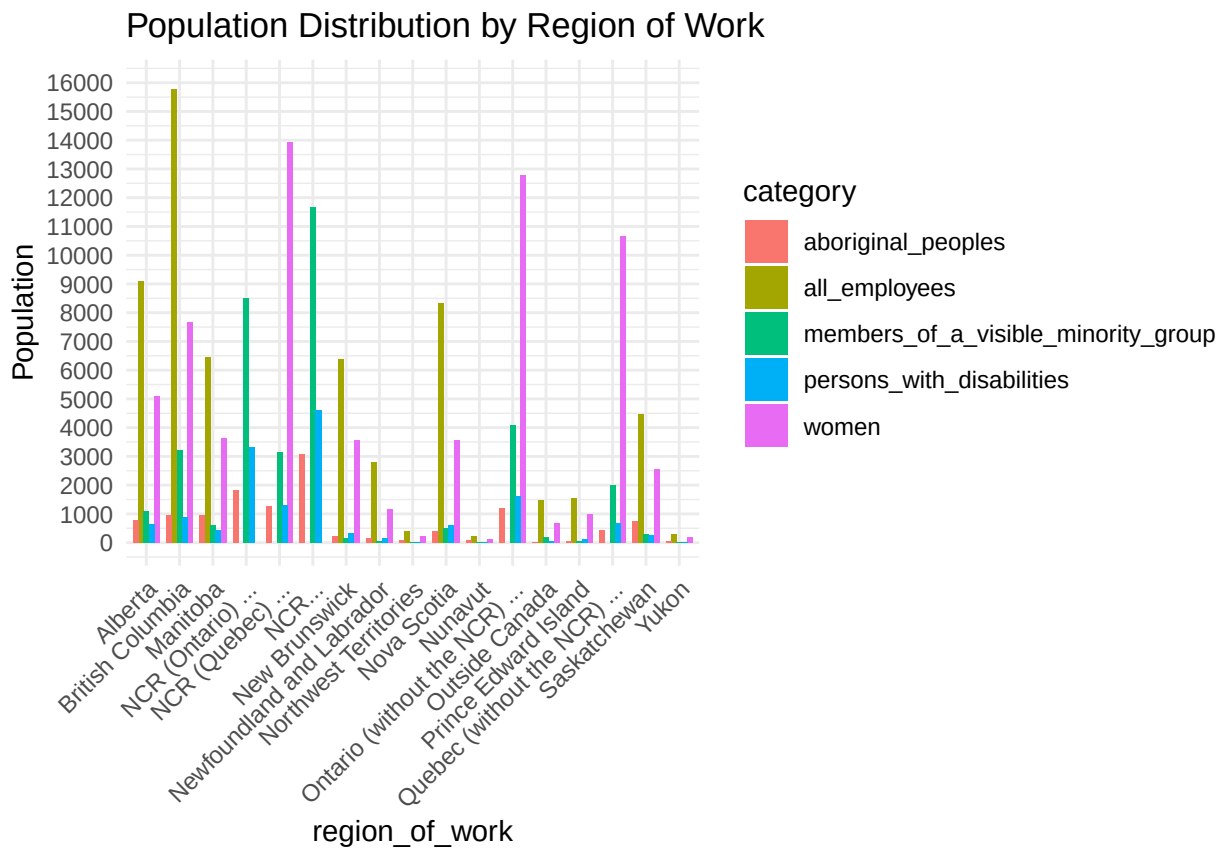
1. Drawing a bar graph showing the different distrbutions of employees categories across regions

i) Converted the data to numerical data
ii) Created a bar graph

```r
subset_data <- subset_data %>%
  mutate_at(vars(all_employees, women, members_of_a_visible_minority_group, persons_with_disabilities, a
            as.numeric)

subset_data_long <- subset_data %>%
  gather(key = "category", value = "value", -region_of_work)

ggplot(subset_data_long, aes(x = region_of_work, y = value, fill = category)) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.8)) +
  labs(title = "Population Distribution by Region of Work",
       y = "Population") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_y_continuous(limits = c(0, 16000), breaks = seq(0, 16000, by = 1000))
```

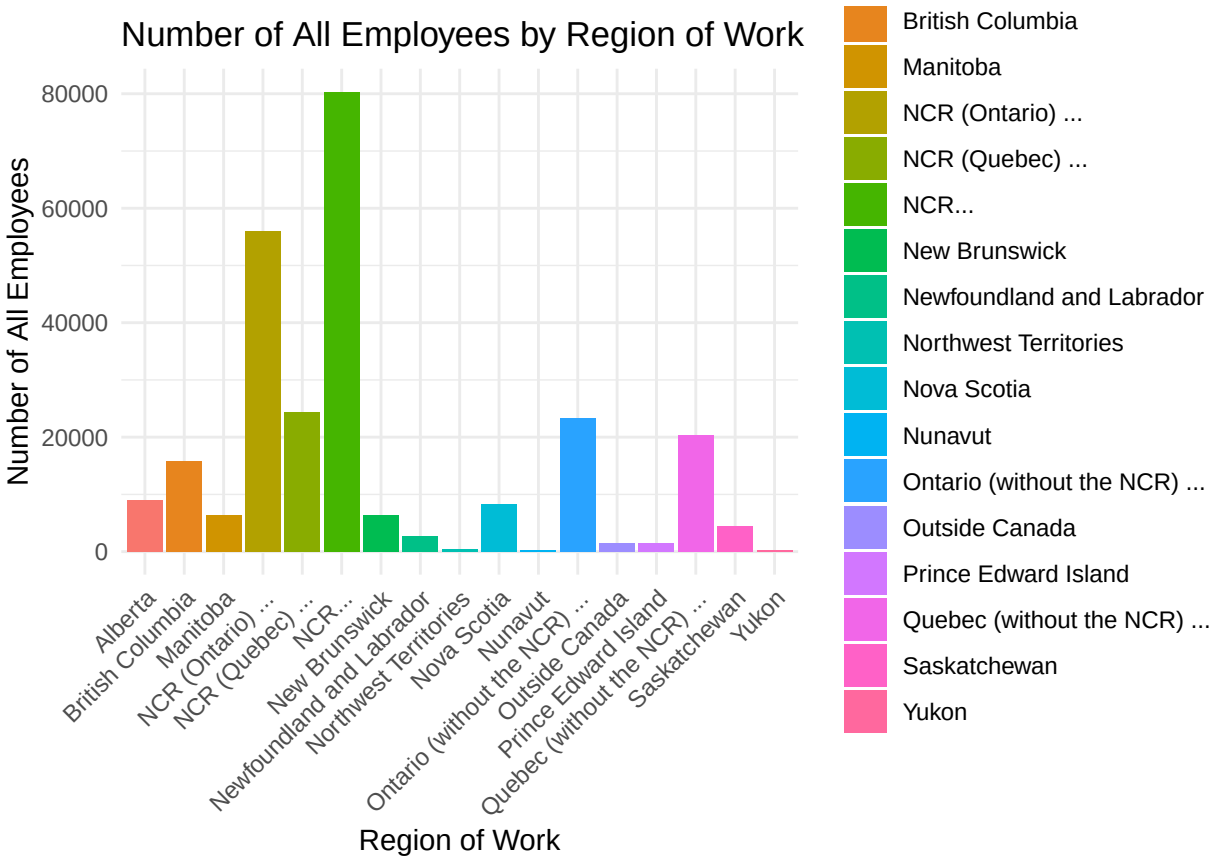


from the bar graph above you can see:

- that women are the second most employed across the different regions

1. distribution of employment across regions

```
ggplot(subset_data, aes(x = region_of_work, y = all_employees, fill = region_of_work)) +
  geom_bar(stat = "identity") +
  labs(title = "Number of All Employees by Region of Work",
       x = "Region of Work",
       y = "Number of All Employees") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

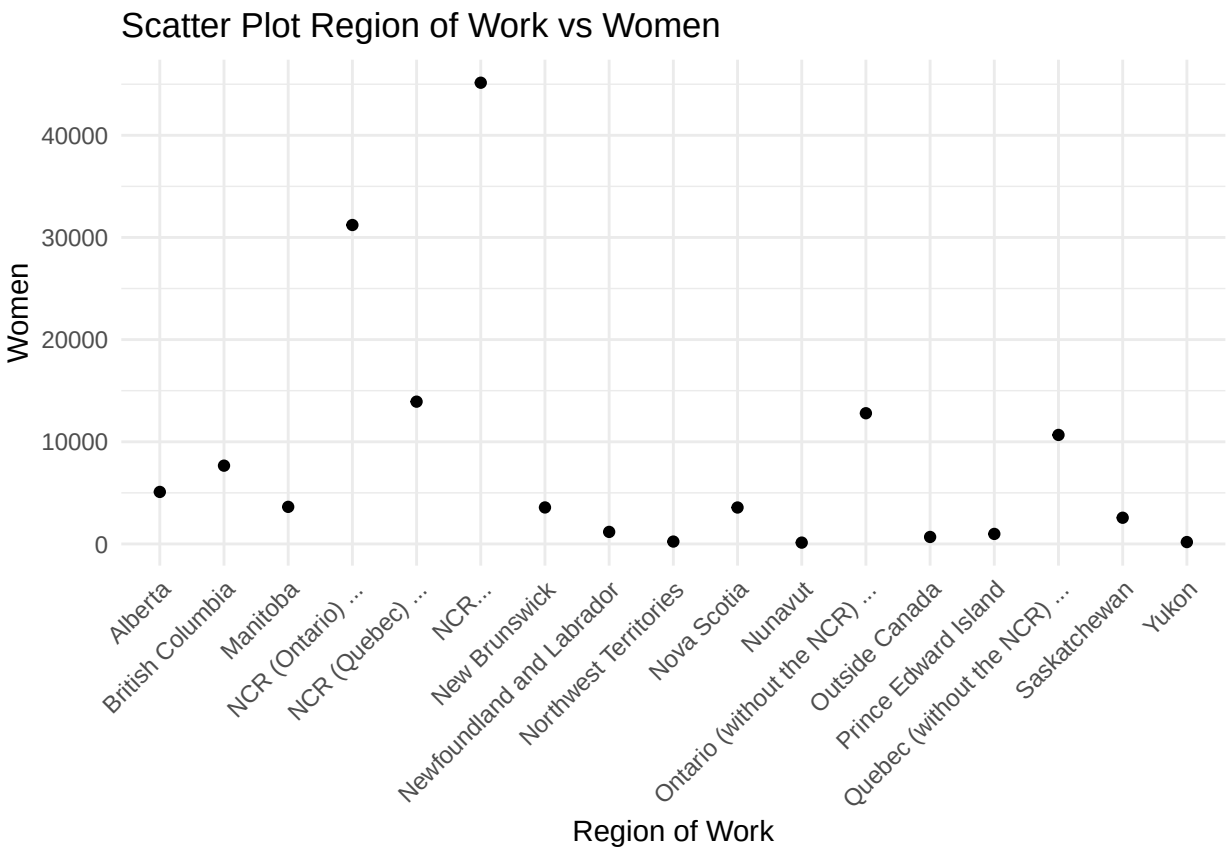


from the bar graph above you can deduce:

- there is a high rate of employment in NCR+ compared to other regions of work

2. Scatter plot for distribution of women across regions of work

```
ggplot(subset_data, aes(x = region_of_work, y = women)) +
  geom_point() +
  labs(title = "Scatter Plot Region of Work vs Women",
       x = "Region of Work",
       y = "Women") +
  theme_minimal() + theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

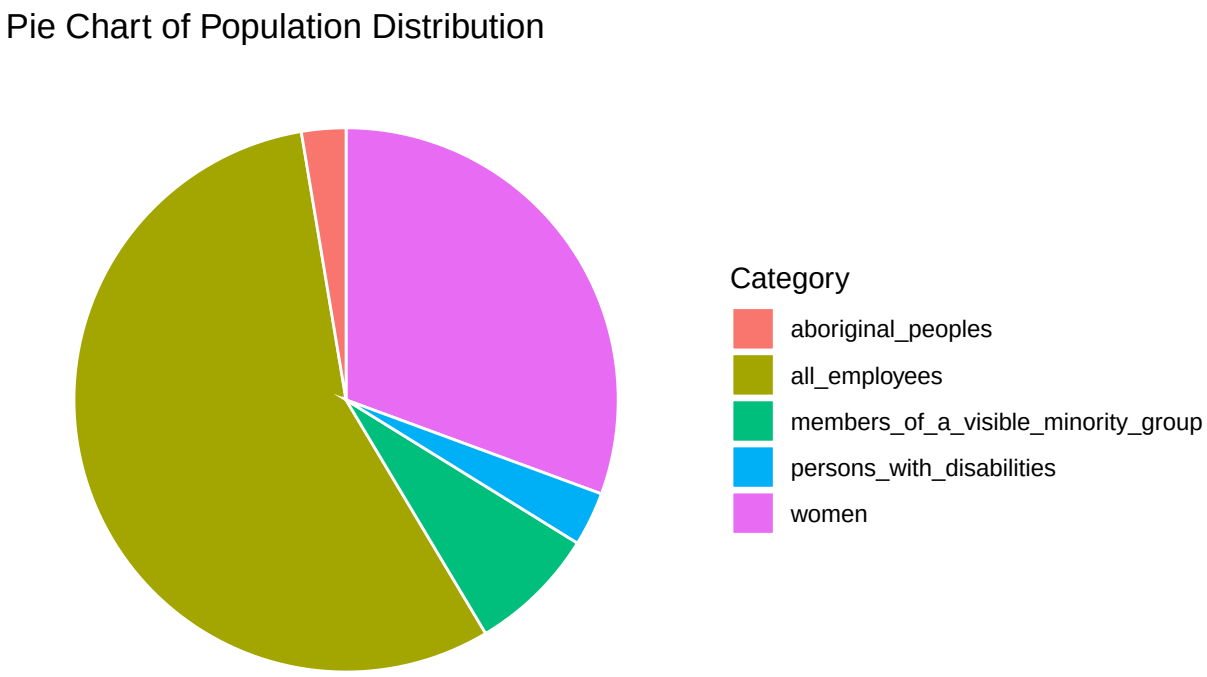


from the scatter plot above we can deduce:

- women employment is under 20,000 in various regions except for NCR‡

3. summary of how the employees are spread out

```
summary_data <- subset_data %>%
  summarise(
    all_employees = sum(all_employees),
    women = sum(women),
    members_of_a_visible_minority_group = sum(members_of_a_visible_minority_group),
    persons_with_disabilities = sum(persons_with_disabilities),
    aboriginal_peoples = sum(aboriginal_peoples)
  )
summary_data_long <- gather(summary_data, key = "category", value = "value")
ggplot(summary_data_long, aes(x = "", y = value, fill = category)) +
  geom_bar(stat = "identity", width = 1, color = "white") +
  coord_polar("y") +
  labs(title = "Pie Chart of Population Distribution",
       fill = "Category") +
  theme_minimal() +
  theme(axis.text = element_blank(),
        axis.title = element_blank(),
        panel.grid = element_blank())
```



from the pie chart above we can deduce:

- women are the second most employed category
- Aborginal people and person with disabilities have a few representation in the job industry