Distribution of Public Service of Canada Employees by Designated Group and Salary Range

checking if our r is working

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

```
## [1] "Hello R world"
```

Analysis of Table 4

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)
tab04_eng <- read_excel("Documents/assignments/keira/cleaned/tab04-eng.xls", skip = 4, n_max = 22)
```

```
## New names:
## * ‘‘ -> ‘...3‘
## * ‘‘ -> ‘...5‘
## * ‘‘ -> ‘...6‘
## * ‘‘ -> ‘...8‘
## * ‘‘ -> ‘...9‘
## * ‘‘ -> ‘...11‘
## * ‘‘ -> ‘...12‘
## * ‘‘ -> ‘...14‘
## * ‘‘ -> ‘...15‘
```

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

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

```
tab04_eng <- clean_names(tab04_eng)
print(colnames(tab04_eng))
```

```
##  [1] "salary_range"                        "all_employees"
##  [3] "x3"                                  "women"
##  [5] "x5"                                  "x6"
##  [7] "aboriginal_peoples"                  "x8"
##  [9] "x9"                                  "persons_with_disabilities"
## [11] "x11"                                 "x12"
## [13] "members_of_a_visible_minority_group" "x14"
## [15] "x15"
```

```
selected_colnames <- c("salary_range", "all_employees", "women", "members_of_a_visible_minority_group",
subset_data <- tab04_eng[, selected_colnames]
subset_data <- subset_data[complete.cases(tab04_eng$salary_range), ]
head(subset_data)
```

```
## # A tibble: 6 x 6
##   salary_range all_employees women members_of_a_visible~1 persons_with_disabil~2
##   <chr>        <chr>         <chr> <chr>                  <chr>
## 1 Under 5,000  41            15    0                      0
## 2 5,000 to 9,~ 141           35    8                      7
## 3 10,000 to 1~ 67            35    0                      0
## 4 15,000 to 1~ 149           103   14                     8
## 5 20,000 to 2~ 260           191   28                     15
## 6 25,000 to 2~ 393           319   52                     22
## # i abbreviated names: 1: members_of_a_visible_minority_group,
## #   2: persons_with_disabilities
## # i 1 more variable: aboriginal_peoples <chr>
```

visualization of the data

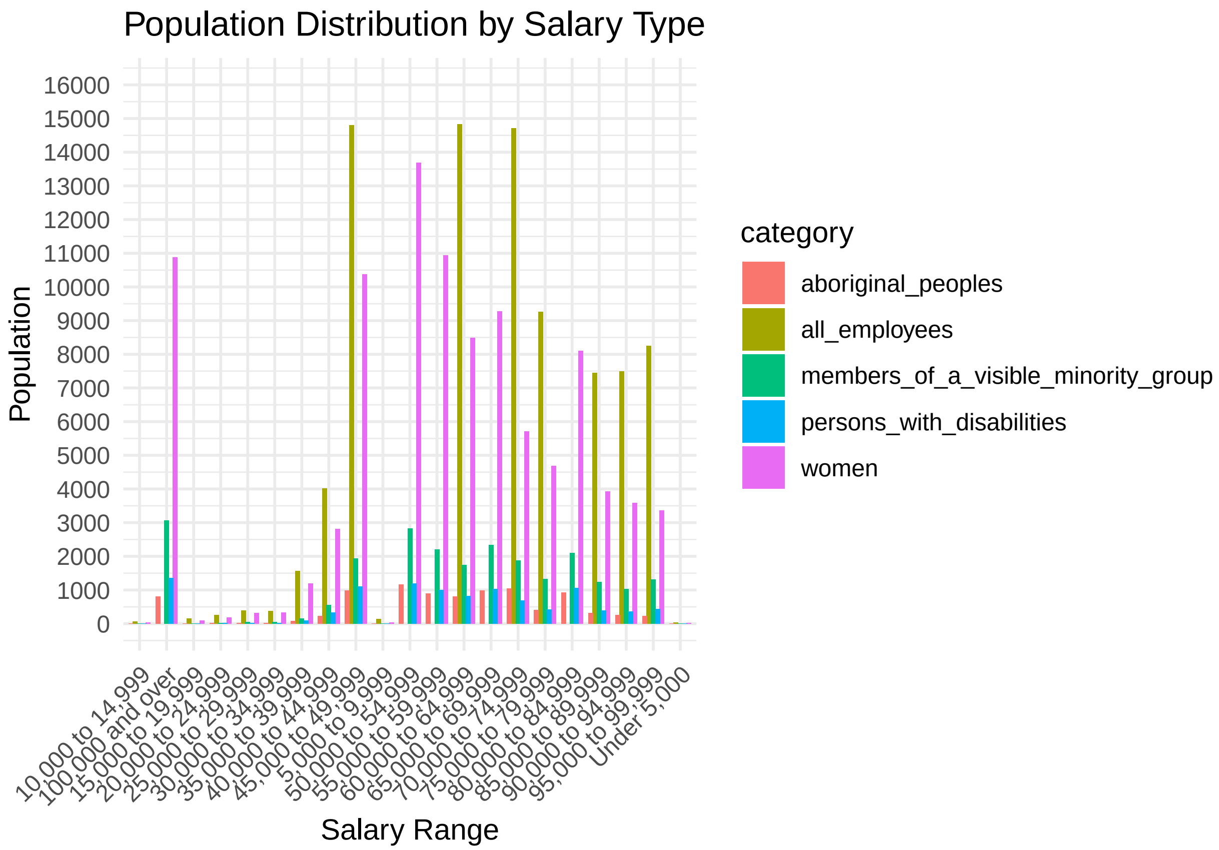
1. Drawing a bar graph showing the different salary ranges across different populations

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", -salary_range)

ggplot(subset_data_long, aes(x = salary_range, y = value, fill = category)) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.8)) +
  labs(title = "Population Distribution by Salary Type",
       y = "Population", x = "Salary Range") +
  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))
```

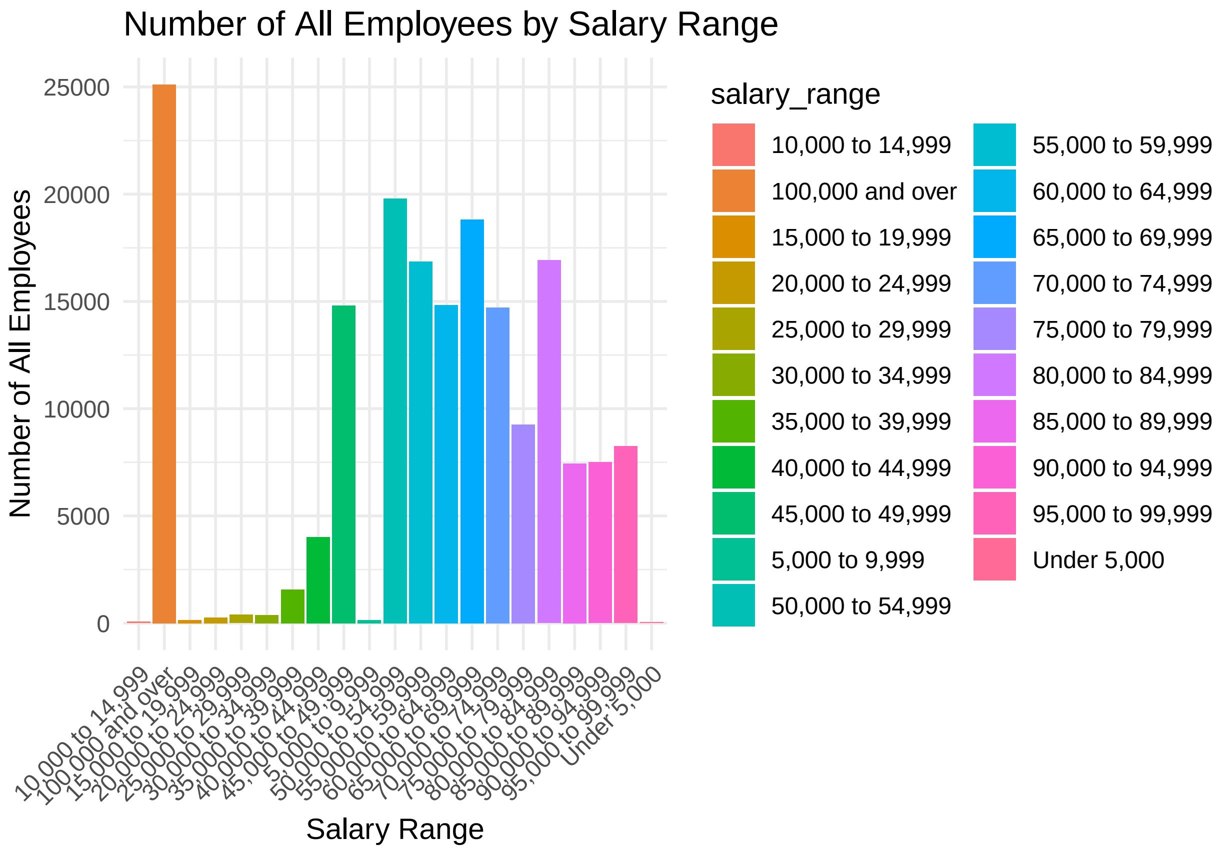


from the bar graph above you can see:

- that a few population is paid between 5000 - 9000

1. distribution of employment across different salary range

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

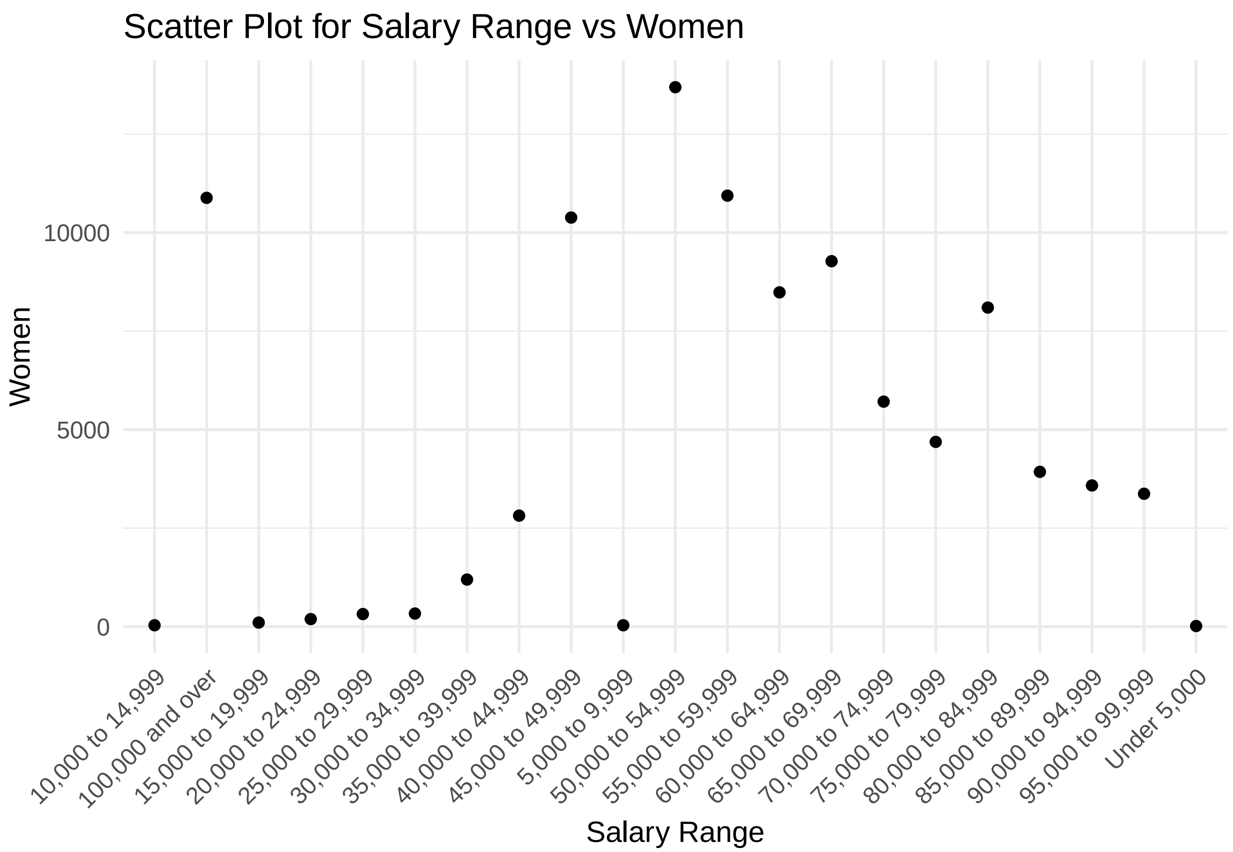


from the bar graph above you can deduce:

- a lot of employees are paid 100,000 and above
- there is a little population that is paid under 5,000

2. Scatter plot for distribution of women across different job sectors

```
ggplot(subset_data, aes(x = salary_range, y = women)) +
  geom_point() +
  labs(title = "Scatter Plot for Salary Range vs Women",
       x = "Salary Range",
       y = "Women") +
  theme_minimal() + theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

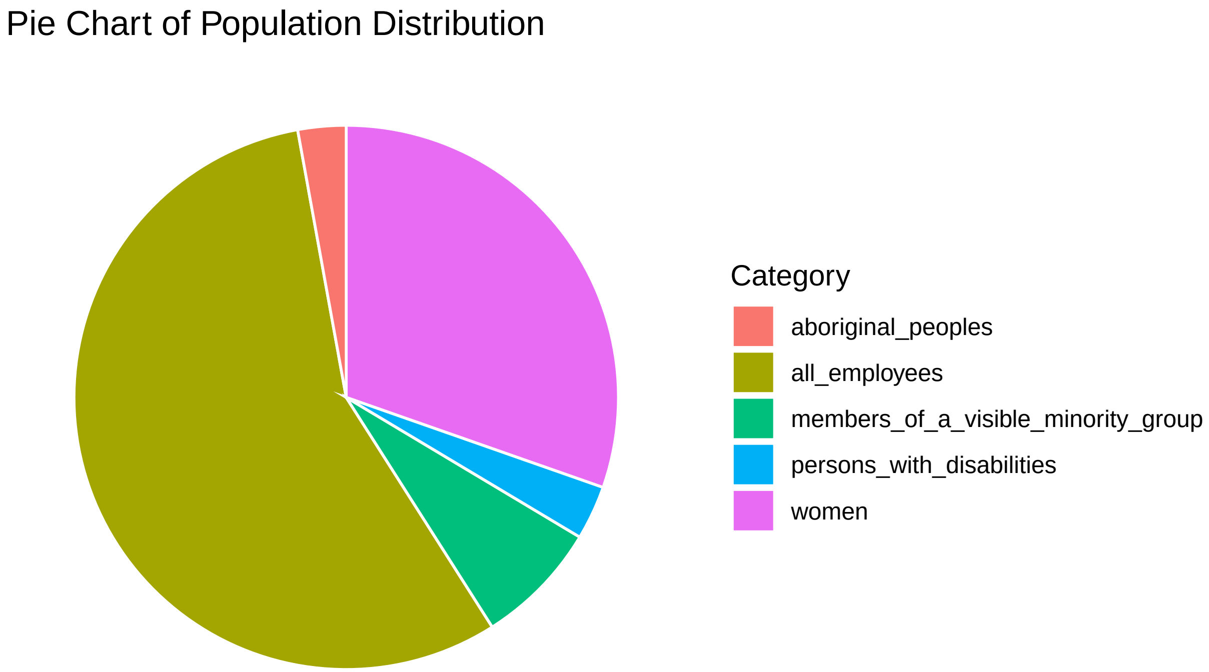


from the scatter plot above we can deduce:

- A lot of women are paid between 50,000 - 54,999
- there is low or zero women paid under 5,000

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