Exercise-1

Himanshu Mayank

2023-03-07

Loading the connections dataset

```
library("tidyverse")
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.5.0
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library("igraph")
```

```
##
## Attaching package: 'igraph'
##
## The following objects are masked from 'package:dplyr':
##
##     as_data_frame, groups, 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:stats':
##
##     decompose, spectrum
##
## The following object is masked from 'package:base':
##
##     union
```

```
csv = read_csv('Connections.csv')
```

```
## Rows: 3030 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (6): First Name, Last Name, Email Address, Company, Position, Connected On
##
## 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.
```

```
csv
```

```
## # A tibble: 3,030 x 6
##    `First Name` `Last Name` `Email Address` Company             Posit~1 Conne~2
##    <chr>        <chr>       <chr>           <chr>               <chr>   <chr>
##  1 Immanuel     Tacky, MMAI <NA>            Canadian Tire Corp~ Data S~ 09 Mar~
##  2 Avneet (Avi) Kaur        <NA>            Environics Analyti~ Resear~ 09 Mar~
##  3 Pearl        Juneja      <NA>            United Nations Col~ Leader~ 09 Mar~
##  4 Marie-Hélène Gélinas     <NA>            CIMA+               Analys~ 09 Mar~
##  5 Omar         Lafif       <NA>            UAP Inc.            Specia~ 08 Mar~
##  6 Alex         Champavere  <NA>            Allianz Trade       Digita~ 08 Mar~
##  7 Amenah       Khan        <NA>            Best Buy Canada     Recrui~ 08 Mar~
##  8 Milena       Kumurdjieva <NA>            Canadian Tire Corp~ AVP, A~ 08 Mar~
##  9 Jiaxin (Myra) Xu         <NA>            Infosys             Data S~ 08 Mar~
## 10 Nikita       Tavkar      <NA>            CBI Health          Human ~ 07 Mar~
## # ... with 3,020 more rows, and abbreviated variable names 1: Position,
## #   2: `Connected On`
```

```
View(csv)
attach(csv)
```

```
## The following object is masked from package:ggplot2:
##
##     Position
```

```
df = csv %>% mutate(Company = str_trim(Company)) %>% drop_na(Company)
head(df,10)
```

```
## # A tibble: 10 x 6
##    `First Name` `Last Name` `Email Address` Company             Posit~1 Conne~2
##    <chr>        <chr>       <chr>           <chr>               <chr>   <chr>
##  1 Immanuel     Tacky, MMAI <NA>            Canadian Tire Corp~ Data S~ 09 Mar~
##  2 Avneet (Avi) Kaur        <NA>            Environics Analyti~ Resear~ 09 Mar~
##  3 Pearl        Juneja      <NA>            United Nations Col~ Leader~ 09 Mar~
##  4 Marie-Hélène Gélinas     <NA>            CIMA+               Analys~ 09 Mar~
##  5 Omar         Lafif       <NA>            UAP Inc.            Specia~ 08 Mar~
##  6 Alex         Champavere  <NA>            Allianz Trade       Digita~ 08 Mar~
##  7 Amenah       Khan        <NA>            Best Buy Canada     Recrui~ 08 Mar~
##  8 Milena       Kumurdjieva <NA>            Canadian Tire Corp~ AVP, A~ 08 Mar~
##  9 Jiaxin (Myra) Xu         <NA>            Infosys             Data S~ 08 Mar~
## 10 Nikita       Tavkar      <NA>            CBI Health          Human ~ 07 Mar~
## # ... with abbreviated variable names 1: Position, 2: `Connected On`
```

Filtering out the data and choosing top 5 companies my connections belong to

```
library(dplyr)

# group the data by Company and count the frequency
company_count <- df %>%
  group_by(Company) %>%
  summarise(count = n()) %>%
  arrange(desc(count))

# get the top 10 companies by frequency
top_5_companies <- company_count$Company[1:5]
head(top_5_companies,10)
```

```
## [1] "Tata Consultancy Services" "Amazon"
## [3] "Deloitte"                  "ZS"
## [5] "Infosys"
```

```
# filter the data for only the top 10 companies
csv <- df %>%
  filter(Company %in% top_5_companies)
csv$`Last_Name` <- substr(csv$`Last Name`, 1, 1)
csv <- csv %>%
  slice_sample(n = 50, replace = TRUE)
head(csv,10)
```

```
## # A tibble: 10 x 7
##    `First Name` `Last Name`  `Email Address`     Company Posit~1 Conne~2 Last_~3
##    <chr>        <chr>        <chr>               <chr>   <chr>   <chr>   <chr>
##  1 Ganesh Kumar Ramakrishnan <NA>                Infosys Senior~ 08 Apr~ R
##  2 Atri         Raha         <NA>                Deloit~ Analyst 22 Jul~ R
##  3 Parth        Girdhar      <NA>                ZS      Decisi~ 26 Sep~ G
##  4 Aravind      R            <NA>                Tata C~ Data S~ 25 Aug~ R
##  5 Sally        Bao          <NA>                Deloit~ Analys~ 05 Mar~ B
##  6 Navneeth     Sreenivasan  <NA>                Amazon  Softwa~ 08 Feb~ S
##  7 Aashna       Mahajan      mahajan.aashna@gma~ Amazon  Softwa~ 06 May~ M
##  8 Vibhas       Bogra        <NA>                Deloit~ SAP An~ 22 Feb~ B
##  9 Ganesh       sharma       <NA>                Tata C~ System~ 02 Jun~ s
## 10 Samkit       Shah         <NA>                Deloit~ DC Ana~ 19 Jan~ S
## # ... with abbreviated variable names 1: Position, 2: `Connected On`,
## #   3: Last_Name
```

Counting the number of connections from the sample for each company.For simplicity top 5 organisations is selected

```
count = csv %>% count(Company, sort=TRUE)
count
```

```
## # A tibble: 5 x 2
##   Company                      n
##   <chr>                    <int>
## 1 Tata Consultancy Services    20
```

```
## 2 Deloitte                     14
## 3 Amazon                        8
## 4 Infosys                       4
## 5 ZS                            4
```

Creating a new column with the first name with the last name initials

```
csv$last_initial <- substr(csv$`Last Name`, 1, 1)
csv$Full_Name <- paste(csv$`First Name`, csv$last_initial, sep = " ")
new_csv <- csv[, c("Full_Name", "Company")]
```

Create a new data frame called nodes by selecting only unique Full_Name values from the new_csv data frame, and then adding a new column called id with unique identifier values for each row.

```
nodes <- new_csv %>% distinct(Full_Name)
nodes <- nodes %>% rowid_to_column('id')
nodes
```

```
## # A tibble: 42 x 2
##       id Full_Name
##    <int> <chr>
##  1     1 Ganesh Kumar R
##  2     2 Atri R
##  3     3 Parth G
##  4     4 Aravind R
##  5     5 Sally B
##  6     6 Navneeth S
##  7     7 Aashna M
##  8     8 Vibhas B
##  9     9 Ganesh s
## 10    10 Samkit S
## # ... with 32 more rows
```

```
copy <- new_csv
colnames(copy) <- paste(colnames(copy), "2", sep="_")
```

Create a new data frame called cross by taking the cross-product of the new_csv and copy data frames, resulting in every possible combination of rows from both data frames. A new data frame called edges is created by filtering the cross data frame to include only rows where the Company and Company_2 columns are equal and the Full_Name and Full_Name_2 columns are not equal.

```
cross <- tidyr::crossing(new_csv, copy, .name_repair="minimal")
edges <- filter(cross, cross$Company == cross$Company_2 & cross$Full_Name != cross$Full_Name_2)
edges <- edges %>% select(Full_Name, Company, Full_Name_2, Company_2)
edges <- edges %>%
  left_join(nodes, by = c("Full_Name" = "Full_Name")) %>%
  rename(node_1 = id)
edges <- edges %>%
  left_join(nodes, by = c("Full_Name_2" = "Full_Name")) %>%
  rename(node_2 = id)
edges <- select(edges, node_1, node_2)
head(edges,10)
```

```
## # A tibble: 10 x 2
##    node_1 node_2
##     <int>  <int>
##  1      7     27
##  2      7     25
##  3      7     30
##  4      7      6
##  5      7     39
##  6     41      4
##  7     41     34
##  8     41      9
##  9     41     13
## 10     41     21
```

Creating a network from the given nodes and edges

```
library("tidygraph")
```

```
##
## Attaching package: 'tidygraph'
```

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

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
library("ggraph")
network <- tbl_graph(nodes=nodes, edges=edges, directed=FALSE)
network
```

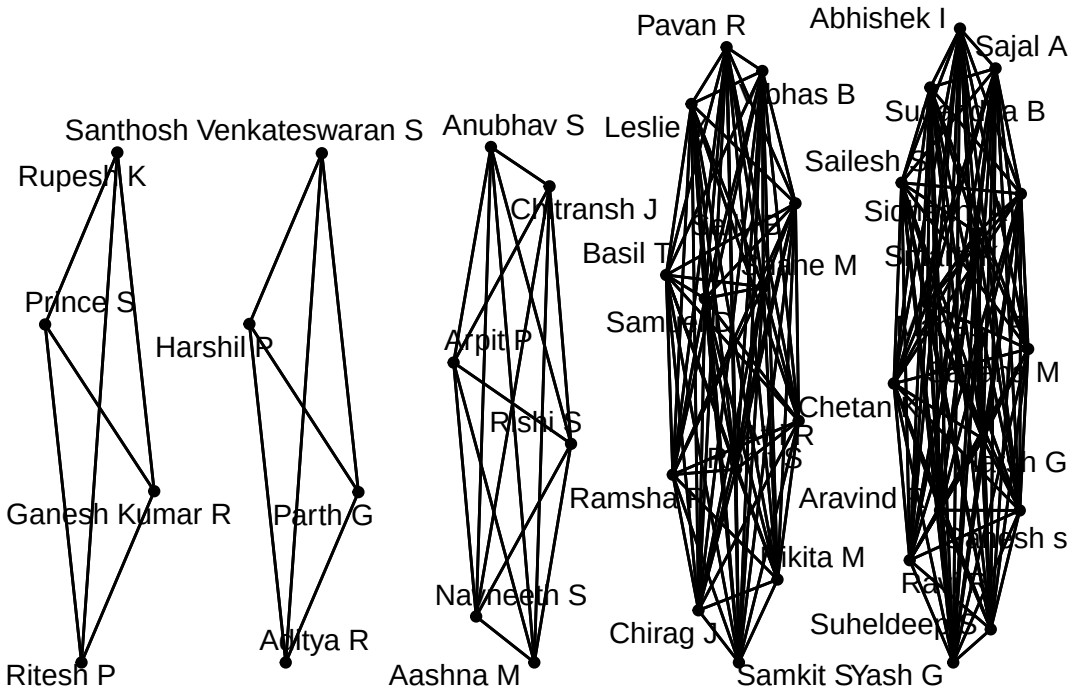
```
## # A tbl_graph: 42 nodes and 420 edges
## #
## # An undirected multigraph with 5 components
## #
## # Node Data: 42 x 2 (active)
##      id Full_Name
##   <int> <chr>
## 1     1 Ganesh Kumar R
## 2     2 Atri R
## 3     3 Parth G
## 4     4 Aravind R
## 5     5 Sally B
## 6     6 Navneeth S
## # ... with 36 more rows
## #
## # Edge Data: 420 x 2
##    from    to
##   <int> <int>
## 1     7    27
```

```
## 2     7    25
## 3     7    30
## # ... with 417 more rows
```

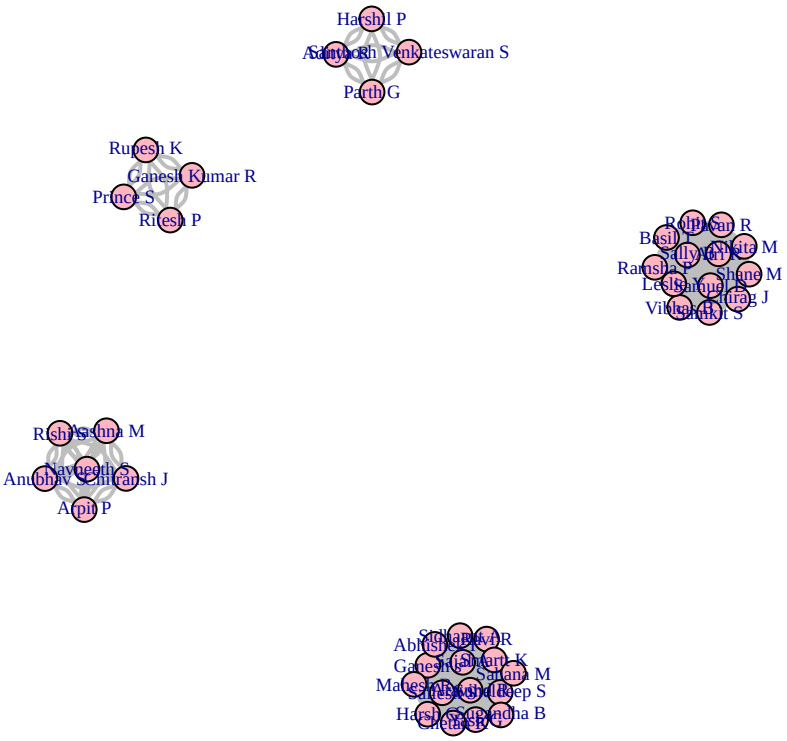
Plotting the graph. People from same organisation will share an edge

```
ggraph(network) +
geom_edge_link() +
geom_node_point() +
geom_node_text(aes(label=Full_Name), repel=TRUE) +
theme_graph()
```

```
## Using "stress" as default layout
```




```
graph <- graph_from_data_frame(edges, vertices = nodes, directed = FALSE)
V(graph)$name <- nodes$Full_Name
par(mar = rep(1, 4))
options(repr.plot.width = 100, repr.plot.height = 1000)
plot(graph, vertex.size = 7, vertex.color = "lightpink", vertex.label.cex = 0.6, edge.color = "gray", ed
```



Intuitively we can see that we get 5 different graphs ,one for each organisation since the data was filtered and sampled for top 5 ornagisation