LinkedIn & IFC study of migration

Luisa M Mimmi

November 24, 2017

1 CONCEPTUAL FRAMEWORK

A city's productivity can be simply defined as the net result of contrasting forces $Productivity = Agglomeration(positive) - Congestion(negative)$

Where $Agglomeration$ and $Congestion$ are a function of various dimensions $= f(Skills, Amenities, Form, Access)$

- Dimensions
  - **Skills**, a city's aggregate stock of human capital
  - **Amenities** - *attracting skills* - job opportunities, housing values, cultural attractions
  - **Form**, the size and spatial configuration of a city (*density* vs. *sprawl*, *wider metropolitan areas*)
  - **Access**, a city's connectedness (or barriers) to other cities, both at home and abroad, through the transportation network
- This comparative analysis will focus on the first dimension **Skills**, with hints to the other ones

```
library(ggplot2) # install.packages("ggplot2")
library(dplyr)
library(knitr)
library(datasets) # initialize
library(knitr)
library(kableExtra)
library(stats)
library(tidyr)
library(stringr)
library(stats)
```

```
getwd()
migr <- read.csv("migration.csv",fileEncoding="UTF-8-BOM")
demog <- read.csv("demographics.csv",fileEncoding="UTF-8-BOM")
```

2 DATA EXPLORATION BY COUNTRY

2.a Preliminary check on demog and migr

```
#   names(demog)      #see all header (column) names
demog[1:5,]          # Indexing (1st to 5 th rows only)
```

```
##   NEW_MEM_ID HIGHEST_DEGREE_OBTAINED SENIORITY
## 1          1                  doctor     Entry
## 2          2                  doctor   Partner
## 3          3                bachelor       CXO
## 4          4                  master     Entry
## 5          5                  master    Senior
##          EMPLOYER_INDUSTRY_SECTOR      POSITION_FUNCTION
## 1  Financial Services & Insurance Information Technology
## 2      Architecture & Engineering   Business Development
```

```
## 3      Retail & Consumer Products   Business Development
## 4           Technology - Hardware Information Technology
## 5 Government/Education/Non-profit              Education
```

```
str(demog)
```

```
## 'data.frame':    475316 obs. of  5 variables:
##  $ NEW_MEM_ID              : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ HIGHEST_DEGREE_OBTAINED : Factor w/ 4 levels "associate","bachelor",..: 3 3 2 4 4 2 2 2 4 3 ...
##  $ SENIORITY               : Factor w/ 11 levels "CXO","Director",..: 3 6 1 3 7 11 7 1 4 4 ...
##  $ EMPLOYER_INDUSTRY_SECTOR: Factor w/ 14 levels "Aero/Auto/Transport",..: 3 2 10 12 4 14 7 7 4 4 ..
##  $ POSITION_FUNCTION       : Factor w/ 26 levels "Accounting","Administrative",..: 13 4 4 13 7 18 16
```

```
summary(demog)       #see some summary statistics of each column
```

```
##    NEW_MEM_ID     HIGHEST_DEGREE_OBTAINED    SENIORITY
##  Min.   :     1   associate: 81072        Entry   :220187
##  1st Qu.:118830   bachelor :252952        Senior  :146986
##  Median :237658   doctor   : 40502        Manager : 46279
##  Mean   :237658   master   :100790        Training: 22047
##  3rd Qu.:356487                           Director: 21093
##  Max.   :475316                           VP      :  7884
##                                           (Other) : 10840
##                     EMPLOYER_INDUSTRY_SECTOR
##  Government/Education/Non-profit: 85999
##  Technology - Software          : 65725
##  Healthcare & Pharmaceutical    : 63281
##  Professional Services          : 60623
##  Financial Services & Insurance : 48297
##  Retail & Consumer Products     : 36494
##  (Other)                        :114897
##               POSITION_FUNCTION
##  Engineering           : 53792
##  Education             : 41195
##  Sales                 : 39617
##  Operations            : 36840
##  Research              : 31635
##  Information Technology: 31435
##  (Other)               :240802
```

```
#  sapply(demog, class)   # get class of all columns

#   names(migr)      #see all header (column) names
migr[1:5,]          # Indexing (1st to 5 th rows only)
```

```
##   NEW_MEM_ID WEEK_BEGINNING SOURCE_COUNTRY              SOURCE_REGION
## 1          1      1/31/2016  United States     San Francisco Bay Area
## 2          2      9/18/2016  United States Greater New York City Area
## 3          3      5/22/2016  United States     San Francisco Bay Area
## 4          4      7/24/2016  United States       Greater Detroit Area
## 5          5      1/24/2016  United States Greater New York City Area
##   DESTINATION_COUNTRY        DESTINATION_REGION
## 1       United States Greater Philadelphia Area
## 2      United Kingdom    London, United Kingdom
## 3       United States    Dallas/Fort Worth Area
## 4       United States       Greater Boston Area
```

```
## 5       United States    San Francisco Bay Area
```

```
str(migr)
```

```
## 'data.frame':    475316 obs. of  6 variables:
##  $ NEW_MEM_ID         : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ WEEK_BEGINNING     : Factor w/ 53 levels "1/10/2016","1/17/2016",..: 5 51 34 43 3 45 52 31 46 46
##  $ SOURCE_COUNTRY     : Factor w/ 3 levels "Australia","United Kingdom",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ SOURCE_REGION      : Factor w/ 347 levels "Aberdeen, United Kingdom",..: 272 126 272 118 126 333 1
##  $ DESTINATION_COUNTRY: Factor w/ 3 levels "Australia","United Kingdom",..: 3 2 3 3 3 3 3 3 3 3 ...
##  $ DESTINATION_REGION : Factor w/ 282 levels "Abilene, Texas Area",..: 106 157 61 93 222 168 104 11 9
```

```
summary(migr)       #see some summary statistics of each column
```

```
##    NEW_MEM_ID       WEEK_BEGINNING          SOURCE_COUNTRY
##  Min.   :     1   7/31/2016: 11460   Australia     :  8902
##  1st Qu.:118830   8/21/2016: 11433   United Kingdom: 36615
##  Median :237658   8/14/2016: 11384   United States :429799
##  Mean   :237658   9/11/2016: 11184
##  3rd Qu.:356487   8/28/2016: 11182
##  Max.   :475316   8/7/2016 : 10916
##                   (Other)  :407757
##                     SOURCE_REGION        DESTINATION_COUNTRY
##  Greater New York City Area: 37000   Australia     :  8647
##  Greater Los Angeles Area  : 23070   United Kingdom: 36199
##  San Francisco Bay Area    : 22643   United States :430470
##  Washington D.C. Metro Area: 19744
##  Greater Chicago Area      : 18521
##  Greater Boston Area       : 16898
##  (Other)                   :337440
##                    DESTINATION_REGION
##  Greater New York City Area: 38088
##  San Francisco Bay Area    : 36707
##  Washington D.C. Metro Area: 25549
##  Greater Los Angeles Area  : 23362
##  London, United Kingdom    : 20136
##  Greater Boston Area       : 17909
##  (Other)                   :313565
```

```
#  sapply(migr, class)   # get class of all columns
```

- **INSIGHTs**:
  - Data contains 347 origin regions and only 282 destinations
  - All Linkedin members (in data) have some tertiary education degree, ~ 50% are Entry level
  - Linkedin members (in data) are distributed in 14 sectors

Merge the 2 tables

```
both <- left_join(demog,migr, by="NEW_MEM_ID")
```

2.b Frequencies and Proportions

I'm interested in studying members distribution across categorical variables according to origin country

Percentages and Proportions of *HIGHEST DEGREE* across countries of origin

```
# Single variable
country <- table(both$SOURCE_COUNTRY)
# Proportions for a single variable table
prop.table(country)

# Cross table by two variables
xcountry <- xtabs(~ HIGHEST_DEGREE_OBTAINED +SOURCE_COUNTRY, both)
# xcountry
addmargins(xcountry)

# Proportions in Cross Table
prop.table(xcountry)                # proportion to total
prop.table(xcountry, margin = 1)    # proportion to row sum (DEGREE)
prop.table(xcountry, margin = 2)    # proportion to column sum (ORIGIN COUNTRY)

# Stratified Table
## 3rd variable as stratified variable
xcountry2 <- xtabs(~ HIGHEST_DEGREE_OBTAINED +SENIORITY +SOURCE_COUNTRY, both)
xcountry2
## flat table
ftable(xcountry2)
```

Proportions of HIGHEST DEGREE/Seniority/Sector/Position against Country of origin with *Dplyr*

```
#   Prop of members in each DEGREE to SUM of Country of origin
freq_OrigDegree <- both %>%
  group_by(both[,7],both[,2]) %>%
  summarise (n = n()) %>%
  mutate(freq = n / sum(n)) %>%
  mutate(rel.freq = paste0(round(100 * n/sum(n), 2), "%"))
freq_OrigDegree  #
```

```
## # A tibble: 12 x 5
## # Groups:   both[, 7] [3]
##    `both[, 7]`    `both[, 2]`      n   freq rel.freq
##    <fct>          <fct>        <int>  <dbl> <chr>
##  1 Australia      associate      255 0.0286 2.86%
##  2 Australia      bachelor      5632 0.633  63.27%
##  3 Australia      doctor         611 0.0686 6.86%
##  4 Australia      master        2404 0.270  27.01%
##  5 United Kingdom associate     1312 0.0358 3.58%
##  6 United Kingdom bachelor     22107 0.604  60.38%
##  7 United Kingdom doctor        3533 0.0965 9.65%
##  8 United Kingdom master        9663 0.264  26.39%
##  9 United States  associate    79505 0.185  18.5%
## 10 United States  bachelor    225213 0.524  52.4%
## 11 United States  doctor       36358 0.0846 8.46%
## 12 United States  master       88723 0.206  20.64%
```

```r
freq_OrigSniority <- both %>%
  group_by(both[,7],both[,3]) %>%
  summarise (n = n()) %>%
  mutate(freq = n / sum(n)) %>%
  mutate(rel.freq = paste0(round(100 * n/sum(n), 2), "%"))
freq_OrigSniority  # % remarkably similar in terms of Seniority  across countries of ORIGIN
```

```
## # A tibble: 31 x 5
## # Groups:   both[, 7] [3]
##    `both[, 7]` `both[, 3]`     n     freq rel.freq
##    <fct>       <fct>       <int>    <dbl> <chr>
##  1 Australia   CXO           119 0.0134   1.34%
##  2 Australia   Director      323 0.0363   3.63%
##  3 Australia   Entry        3542 0.398    39.79%
##  4 Australia   Manager      1150 0.129    12.92%
##  5 Australia   Owner          80 0.00899  0.9%
##  6 Australia   Partner        41 0.00461  0.46%
##  7 Australia   Senior       3331 0.374    37.42%
##  8 Australia   Training      154 0.0173   1.73%
##  9 Australia   Unpaid          1 0.000112 0.01%
## 10 Australia   VP            161 0.0181   1.81%
## # ... with 21 more rows
```

```r
freq_OrigSector <- both %>%
  group_by(both[,7],both[,4]) %>%
  summarise (n = n()) %>%
  mutate(freq = n / sum(n)) %>%
  mutate(rel.freq = paste0(round(100 * n/sum(n), 2), "%"))
freq_OrigSector  # % remarkably similar in terms of sector across countries
```

```
## # A tibble: 42 x 5
## # Groups:   both[, 7] [3]
##    `both[, 7]` `both[, 4]`                          n   freq rel.freq
##    <fct>       <fct>                            <int>  <dbl> <chr>
##  1 Australia   Aero/Auto/Transport                357 0.0401 4.01%
##  2 Australia   Architecture & Engineering         805 0.0904 9.04%
##  3 Australia   Financial Services & Insurance     926 0.104  10.4%
##  4 Australia   Government/Education/Non-profit   1716 0.193  19.28%
##  5 Australia   Healthcare & Pharmaceutical        723 0.0812 8.12%
##  6 Australia   Manufacturing/Industrial           372 0.0418 4.18%
##  7 Australia   Media & Entertainment              506 0.0568 5.68%
##  8 Australia   Oil & Energy                       251 0.0282 2.82%
##  9 Australia   Professional Services             1481 0.166  16.64%
## 10 Australia   Retail & Consumer Products         504 0.0566 5.66%
## # ... with 32 more rows
```

```r
freq_OrigPosition <- both %>%
  group_by(both[,7],both[,5]) %>%
  summarise (n = n()) %>%
  mutate(freq = n / sum(n)) %>%
  mutate(rel.freq = paste0(round(100 * n/sum(n), 2), "%"))
freq_OrigPosition  # % remarkably similar in terms of position across countries
```

```
## # A tibble: 78 x 5
## # Groups:   both[, 7] [3]
```

```
##    `both[, 7]` `both[, 5]`                      n   freq rel.freq
##    <fct>       <fct>                        <int>  <dbl> <chr>
##  1 Australia   Accounting                     208 0.0234 2.34%
##  2 Australia   Administrative                 162 0.0182 1.82%
##  3 Australia   Arts and Design                293 0.0329 3.29%
##  4 Australia   Business Development           663 0.0745 7.45%
##  5 Australia   Community and Social Services  439 0.0493 4.93%
##  6 Australia   Consulting                     247 0.0277 2.77%
##  7 Australia   Education                      521 0.0585 5.85%
##  8 Australia   Engineering                    830 0.0932 9.32%
##  9 Australia   Entrepreneurship               112 0.0126 1.26%
## 10 Australia   Finance                        508 0.0571 5.71%
## # ... with 68 more rows
```

- **INSIGHTs**:
  - US has a significantly higher # of Associates leaving (18% vs 3% and 4%)
  - Proportions seem remarkably similar in terms of Seniority / Sector / Position across countries of ORIGIN
  - At the country level there are no big differences… probably makes more sense looking at city or internally
  - Wonder if this is a subset of real Linkedin members or if there are particular similarities in the English-speaking countries

Percentages and Proportions - Seniority/Sector/Position against Country of DESTINATION- with *Dplyr*

I do the same analysis but looking at DESTINATION * Very similar results as per Origin

```
# prop DEGREE by country orf destin
freq_DestDegree <- both %>%
  group_by(both[,9],both[,2]) %>%
  summarise (n = n()) %>%
  mutate(freq = n / sum(n)) %>%
  mutate(rel.freq = paste0(round(100 * n/sum(n), 2), "%"))
freq_DestDegree  # US has a significantly higher # of Associates leaving (18% vs 3% and 4%)

# prop SENIORITY by country orf destin
freq_DestSniority <- both %>%
  group_by(both[,9],both[,3]) %>%
  summarise (n = n()) %>%
  mutate(freq = n / sum(n)) %>%
  mutate(rel.freq = paste0(round(100 * n/sum(n), 2), "%"))
freq_DestSniority  # % remarkably similar in terms of Seniority across countries

# prop SECTOR by country orf destin
freq_DestSector <- both %>%
  group_by(both[,9],both[,4]) %>%
  summarise (n = n()) %>%
  mutate(freq = n / sum(n)) %>%
  mutate(rel.freq = paste0(round(100 * n/sum(n), 2), "%"))
freq_DestSector  # % remarkably similar in terms of sector across countries

# prop POSITION by country orf destin
freq_DestPosition <- both %>%
```

```
  group_by(both[,9],both[,5]) %>%
  summarise (n = n()) %>%
  mutate(freq = n / sum(n)) %>%
  mutate(rel.freq = paste0(round(100 * n/sum(n), 2), "%"))
freq_DestPosition  # % remarkably similar in terms of position across countries

# 3-Way Frequency Table
# mytable <- table(both[,4],both[,2], both[,3])
# ftable(mytable)
```

3 DATA EXPLORATION BY CITY

interim manipulation

```
# create origin table
origin <- both %>% select(id=NEW_MEM_ID, degree=HIGHEST_DEGREE_OBTAINED, seniority=SENIORITY, sector=EM

# create destination table
destination <- both %>% select(id=NEW_MEM_ID, degree=HIGHEST_DEGREE_OBTAINED, seniority=SENIORITY, secto
```

3.1 Top 10 cities (US)

```
# Aggregate N flow (OUT) by City
library (knitr)
library(kableExtra)
aggreOrig <- origin  %>%
  group_by(cityO) %>%
  summarize(NumOutflow= n())  %>%
  mutate(freq = NumOutflow / sum(NumOutflow)) %>%
  mutate(rel.freq = as.numeric(paste0(round(100 * NumOutflow/sum(NumOutflow), 3))))  %>%
  arrange(desc(NumOutflow))

# Top 10 ORIGIN cities
aggreOrig_short <- aggreOrig[1:10,]
kable(aggreOrig_short, format = "html", caption = "Ranking of cities by Origin") %>%  kable_styling(boo
```

Ranking of cities by Origin

cityO

NumOutflow

freq

rel.freq

Greater New York City Area

37000

0.0778430

7.784

Greater Los Angeles Area

23070

0.0485361

4.854

San Francisco Bay Area

22643

0.0476378

4.764

Washington D.C. Metro Area

19744

0.0415387

4.154

Greater Chicago Area

18521

0.0389657

3.897

Greater Boston Area

16898

0.0355511

3.555

Dallas/Fort Worth Area

11592

0.0243880

2.439

Greater Philadelphia Area

11407

0.0239988

2.400

Greater Atlanta Area

11092

0.0233361

2.334

London, United Kingdom

10249

0.0215625

2.156

```r
# create destination table
destination <- both %>% select(id=NEW_MEM_ID, degree=HIGHEST_DEGREE_OBTAINED, seniority=SENIORITY, secto


# Aggregate N flow (IN) by City
 library (knitr)
library(kableExtra)
aggreDest <- destination  %>%
  group_by(cityD) %>%
  summarize(NumInflow= n())  %>%
  mutate(freq = NumInflow / sum(NumInflow)) %>%
  mutate(rel.freq = as.numeric(paste0(round(100 * NumInflow/sum(NumInflow), 3))))  %>%
  arrange(desc(NumInflow))

# Top 10 DESTINATION cities
aggreDest_short <- aggreDest[1:10,]
kable(aggreDest_short, format = "html", caption = "Ranking of cities by popular destination") %>% kable_
```

Ranking of cities by popular destination

cityD

NumInflow

freq

rel.freq

Greater New York City Area

38088

0.0801320

8.013

San Francisco Bay Area

36707

0.0772265

7.723

Washington D.C. Metro Area

25549

0.0537516

5.375

Greater Los Angeles Area

23362

0.0491505

4.915

London, United Kingdom

20136

0.0423634

4.236

Greater Boston Area

17909

0.0376781

3.768

Greater Chicago Area

16944

0.0356479

3.565

Dallas/Fort Worth Area

16157

0.0339921

3.399

Greater Seattle Area

15553

0.0327214

3.272

Greater Atlanta Area

14242

0.0299632

2.996

- **INSIGHTs**:
  - Intristingly, London is # 5 Origin but # 10 Destination
  - 9/10 top DESTINATION are the same as top ORIGIN which suggest there is mobility, but not necessarily the top destination are places where people stay
  - this can be explained by the American way of moving to and from the city of college
  - *those where all american !!!*

3.2 Top 10 cities (UK)

```
# Aggregate N flow (OUT) by City
library (knitr)
library(kableExtra)

# subset origin
originUK <- subset (origin , countryO == "United Kingdom")

aggreOrigUK <- originUK  %>%
  group_by(cityO) %>%
  summarize(NumOutflow= n())  %>%
  mutate(freq = NumOutflow / sum(NumOutflow)) %>%
  mutate(rel.freq = as.numeric(paste0(round(100 * NumOutflow/sum(NumOutflow), 3))))  %>%
```

```r
  arrange(desc(NumOutflow))  #%>% left_join(origin, by="cityO")

# Top 10 ORIGIN cities
aggreOrig_shortUK <- aggreOrigUK[1:10,]
kable(aggreOrig_shortUK, format = "html", caption = "Ranking of cities by Origin") %>%  kable_styling(bo
```

Ranking of cities by Origin

cityO

NumOutflow

freq

rel.freq

London, United Kingdom

10249

0.2799126

27.991

Manchester, United Kingdom

1170

0.0319541

3.195

Reading, United Kingdom

1002

0.0273658

2.737

Twickenham, United Kingdom

990

0.0270381

2.704

Oxford, United Kingdom

978

0.0267104

2.671

Birmingham, United Kingdom

836

0.0228322

2.283

Guildford, United Kingdom

828

0.0226137

2.261

Kingston upon Thames, United Kingdom

788

0.0215212

2.152

Coventry, United Kingdom

695

0.0189813

1.898

Cambridge, United Kingdom

668

0.0182439

1.824

```
# ======================================================#

# subset destination
destinationUK <- subset (destination , countryD == "United Kingdom")


# Aggregate N flow (IN) by City
 library (knitr)
library(kableExtra)
aggreDestUK <- destinationUK  %>%
  group_by(cityD) %>%
  summarize(NumInflow= n())  %>%
  mutate(freq = NumInflow / sum(NumInflow)) %>%
  mutate(rel.freq = as.numeric(paste0(round(100 * NumInflow/sum(NumInflow), 3))))  %>%
  arrange(desc(NumInflow))

# Top 10 DESTINATION cities
aggreDest_shortUK <- aggreDestUK[1:10,]
kable(aggreDest_shortUK, format = "html", caption = "Ranking of cities by popular destination") %>% kabl
```

Ranking of cities by popular destination

cityD

NumInflow

freq

rel.freq

London, United Kingdom

20136

0.5562585

55.626

Manchester, United Kingdom

1859

0.0513550

5.136

Birmingham, United Kingdom

949

0.0262162

2.622

Edinburgh, United Kingdom

803

0.0221829

2.218

Bristol, United Kingdom

754

0.0208293

2.083

Reading, United Kingdom

754

0.0208293

2.083

Cambridge, United Kingdom

719

0.0198624

1.986

Leeds, United Kingdom

695

0.0191994

1.920

Glasgow, United Kingdom

682

0.0188403

1.884

Oxford, United Kingdom

554

0.0153043

1.530

- **INSIGHTs**:

– Contrary to widespread mobility in US, London is origin of 28% of migrants and the Destination for 55% of them

3.3 Top 10 cities (Austr)

```
# Aggregate N flow (OUT) by City
library (knitr)
library(kableExtra)

# subset origin
originAustr <- subset (origin , countryO == "Australia")

aggreOrigAustr <- originAustr  %>%
  group_by(cityO) %>%
  summarize(NumOutflow= n())  %>%
  mutate(freq = NumOutflow / sum(NumOutflow)) %>%
  mutate(rel.freq = as.numeric(paste0(round(100 * NumOutflow/sum(NumOutflow), 3))))  %>%
  arrange(desc(NumOutflow))  #%>% left_join(origin, by="cityO")

# Top 10 ORIGIN cities
aggreOrig_shortAustr <- aggreOrigAustr[1:10,]
kable(aggreOrig_shortAustr, format = "html", caption = "Ranking of cities by Origin") %>%  kable_stylin
```

Ranking of cities by Origin

cityO

NumOutflow

freq

rel.freq

Sydney Area, Australia

3209

0.3604808

36.048

Brisbane Area, Australia

1995

0.2241069

22.411

Perth Area, Australia

1014

0.1139070

11.391

Adelaide Area, Australia

611

0.0686363

6.864

Canberra Area, Australia

556

0.0624579

6.246

Queensland, Australia

468

0.0525725

5.257

New South Wales, Australia

357

0.0401033

4.010

Newcastle Area, Australia

216

0.0242642

2.426

Western Australia, Australia

103

0.0115704

1.157

Toowoomba Area, Australia

74

0.0083127

0.831

```
# ======================================================#

# subset destination
destinationAustr <- subset (destination , countryD == "Australia")


# Aggregate N flow (IN) by City
 library (knitr)
library(kableExtra)
aggreDestAustr <- destinationAustr  %>%
  group_by(cityD) %>%
  summarize(NumInflow= n())  %>%
  mutate(freq = NumInflow / sum(NumInflow)) %>%
  mutate(rel.freq = as.numeric(paste0(round(100 * NumInflow/sum(NumInflow), 3))))  %>%
  arrange(desc(NumInflow))
```

```r
# Top 10 DESTINATION cities
aggreDest_shortAustr <- aggreDestAustr[1:10,]
kable(aggreDest_shortAustr, format = "html", caption = "Ranking of cities by popular destination") %>% k
```

Ranking of cities by popular destination

cityD

NumInflow

freq

rel.freq

Sydney Area, Australia

3885

0.4492888

44.929

Brisbane Area, Australia

2046

0.2366139

23.661

Canberra Area, Australia

674

0.0779461

7.795

Perth Area, Australia

532

0.0615242

6.152

Queensland, Australia

452

0.0522725

5.227

New South Wales, Australia

393

0.0454493

4.545

Adelaide Area, Australia

306

0.0353880

3.539

Newcastle Area, Australia

180

0.0208165

2.082

Western Australia, Australia

110

0.0127212

1.272

Toowoomba Area, Australia

69

0.0079796

0.798

- **INSIGHTs**:
  - Similar to UK, in Australiam Sydney is the origin of 36% of migrants and the Destination for 50% of them

3.4 Plots of top cities

JOIN the AGGREGATE by CITY to have Net flows in the same table (by City)

```
# Adding some variables for plots
aggreByCity <- full_join(aggreDest,aggreOrig, by = c("cityD" = "cityO"))
aggreByCity[c("NumInflow", "NumOutflow")][is.na(aggreByCity[c("NumInflow", "NumOutflow")])] <- 0

aggreByCity <-  aggreByCity %>%
  select(-freq.x,-rel.freq.x, -freq.y, -rel.freq.y) %>%  # get rid of meeaningless
  mutate (NetFlow= NumInflow -NumOutflow) %>% # Net
  mutate (NegOutFlow= -(NumOutflow)) %>% # neg sign
  mutate (Sign = ifelse(NetFlow > 0, "Positive", "Negative")) %>%
  mutate (colour= ifelse(NetFlow > 0, "positive", "negative")) %>%  mutate (city_copy = cityD)  %>%
  separate(city_copy, into = c("city_only", "metro area"), sep = ",")

summary(aggreByCity)
```

interim

```
# ====================== INTERIM ======================== #
names(aggreByCity)[1]<-"city"
names(migr)[4]<-"city"
# city_country <-full_join(aggreByCity, migr, by = c("city", "SOURCE_REGION")) %>%  # #select(city,coun

city_country <-full_join(aggreByCity, migr, by = "city") %>%
  select(city, country=DESTINATION_COUNTRY) %>%
  distinct () %>% # need the dup flag
  mutate(dupli = ifelse((city =="Greater New York City Area"|city =="San Francisco Bay Area" |city =="W
  city =="Greater Los Angeles Area"|city =="Greater Boston Area"|city =="Greater Chicago Area" |
  city =="Dallas/Fort Worth Area" |city =="Greater Seattle Area" |city =="San Francisco Bay Area") & cou
```

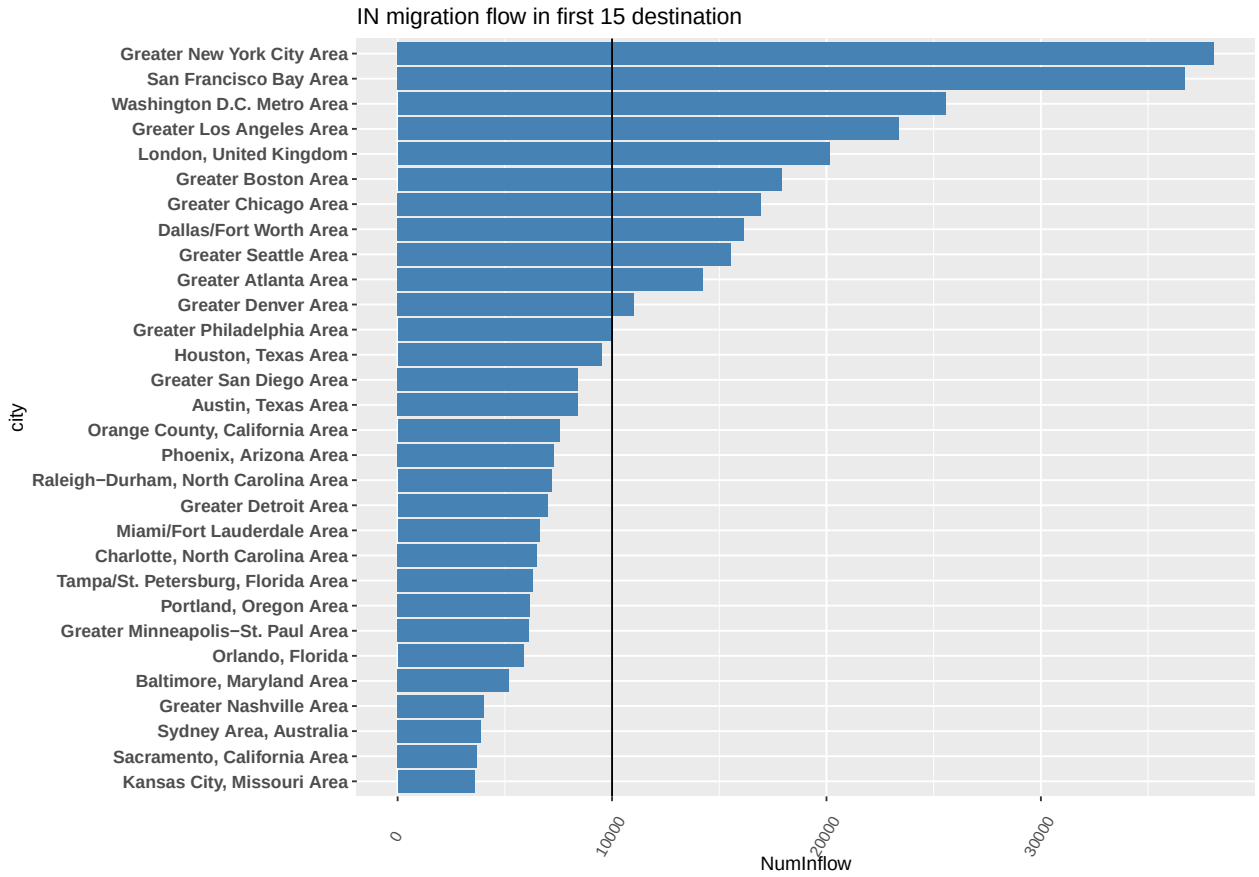
```
  mutate(dupli2 = ifelse((city =="London, United Kingdom") & country!="United kindom", "dup", "")) %>% 

  mutate(dupli3 = ifelse((city =="Sydney Area, Australia") & country!="Australia", "dup", "")) %>% # ne

  mutate(dupli4 = ifelse((city =="Perth Area, Australia") & country!="Australia", "dup", "")) %>% # nee

  mutate(dupli5 = ifelse((city =="Miami/Fort Lauderdale Area") & country!="United States", "dup", "")) 

  mutate(dupli6 = ifelse((city =="Brisbane Area, Australia") & country!="Australia", "dup", ""))
```

```
## Warning: Column `city` joining character vector and factor, coercing into
## character vector
```

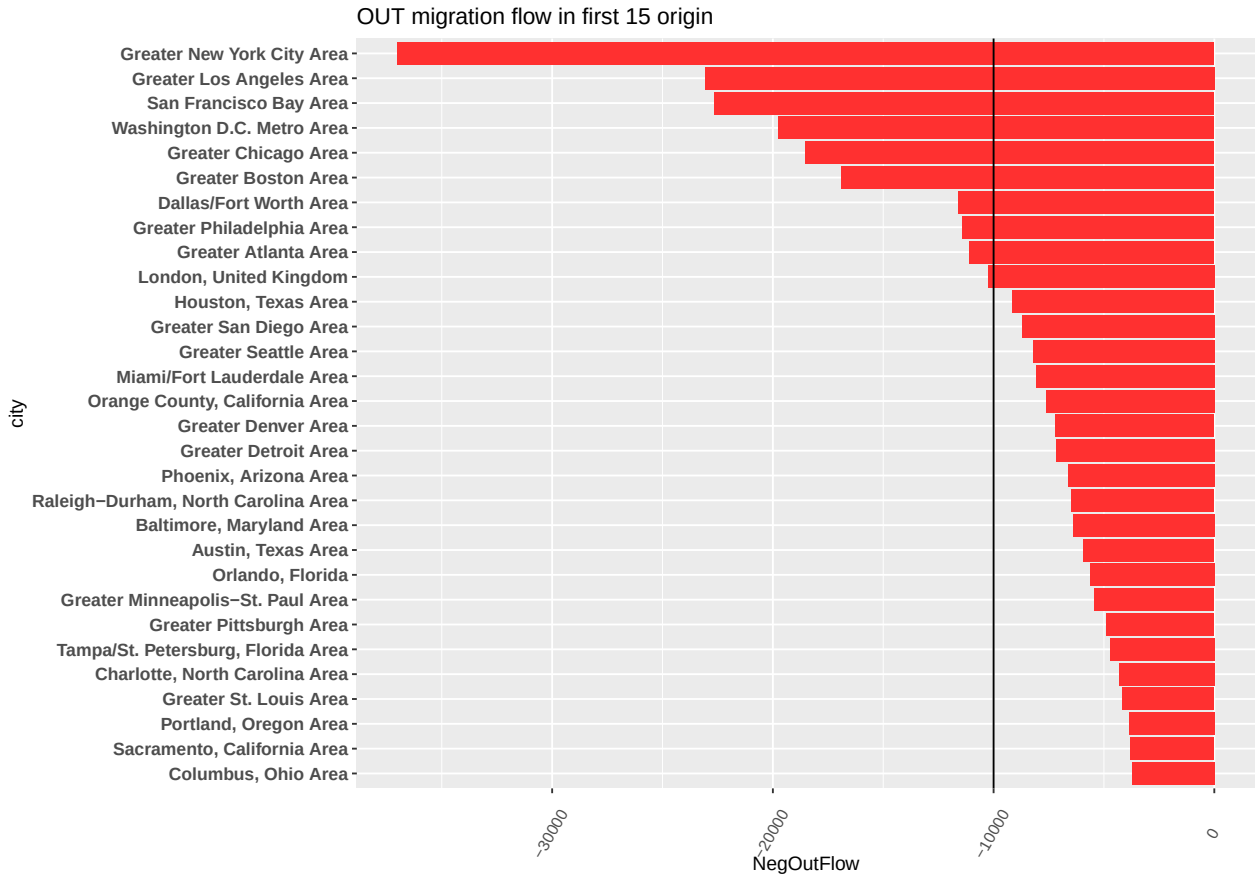
```
city_country <- subset(city_country, city_country$dupli!="dup" & city_country$dupli2!="dup" & city_coun


newRow <- data.frame(city ="London, United Kingdom", country ="United Kingdom" )

city_country <- rbind(city_country,newRow)
#city_country <- rbind(city_country, c("London, United Kingdom", "United kindom"))

names(aggreByCity)[1]<-"cityD"
# ======================================================== #
```

Major 30 (World - mostly US) cities TO people are migrating

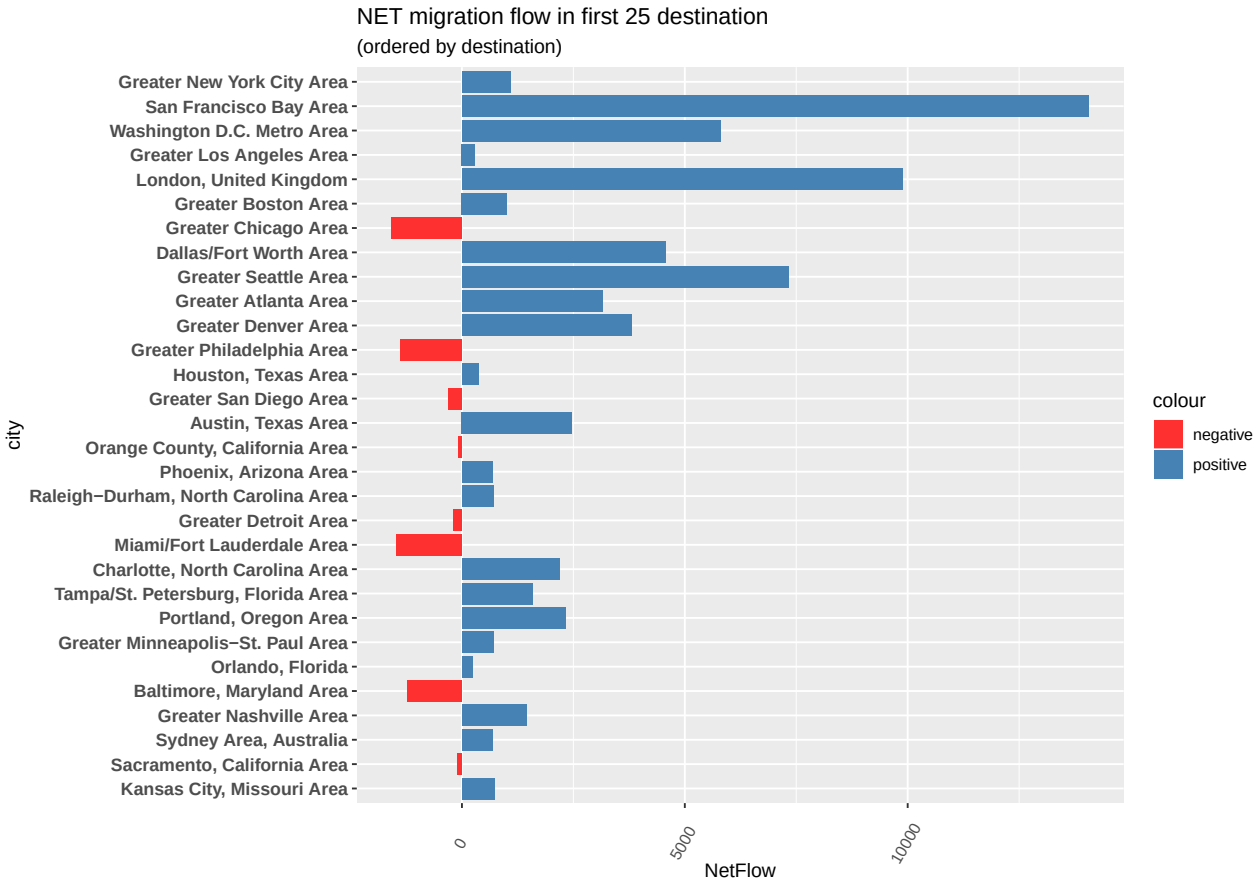
```
names(aggreByCity)[1]<-"city"

# 1) In  Flow in Top DESTINATION
Top_in <- aggreByCity %>% top_n(30, NumInflow)
  Top_in$city = with(Top_in, reorder(city,NumInflow)) # reorder Levels by Var

Top_in <- ggplot(data = Top_in,aes(city, NumInflow)) +
  geom_bar(stat = "identity", position="identity", fill = "steelblue") +
  geom_hline(yintercept=10000, color = "black", size=0.5) +
  theme(axis.text.x = element_text(angle=60, vjust=0.3), axis.text.y = element_text(size=10,face="bold"
  labs(title="IN migration flow in first 15 destination")


Top_in + coord_flip()
```



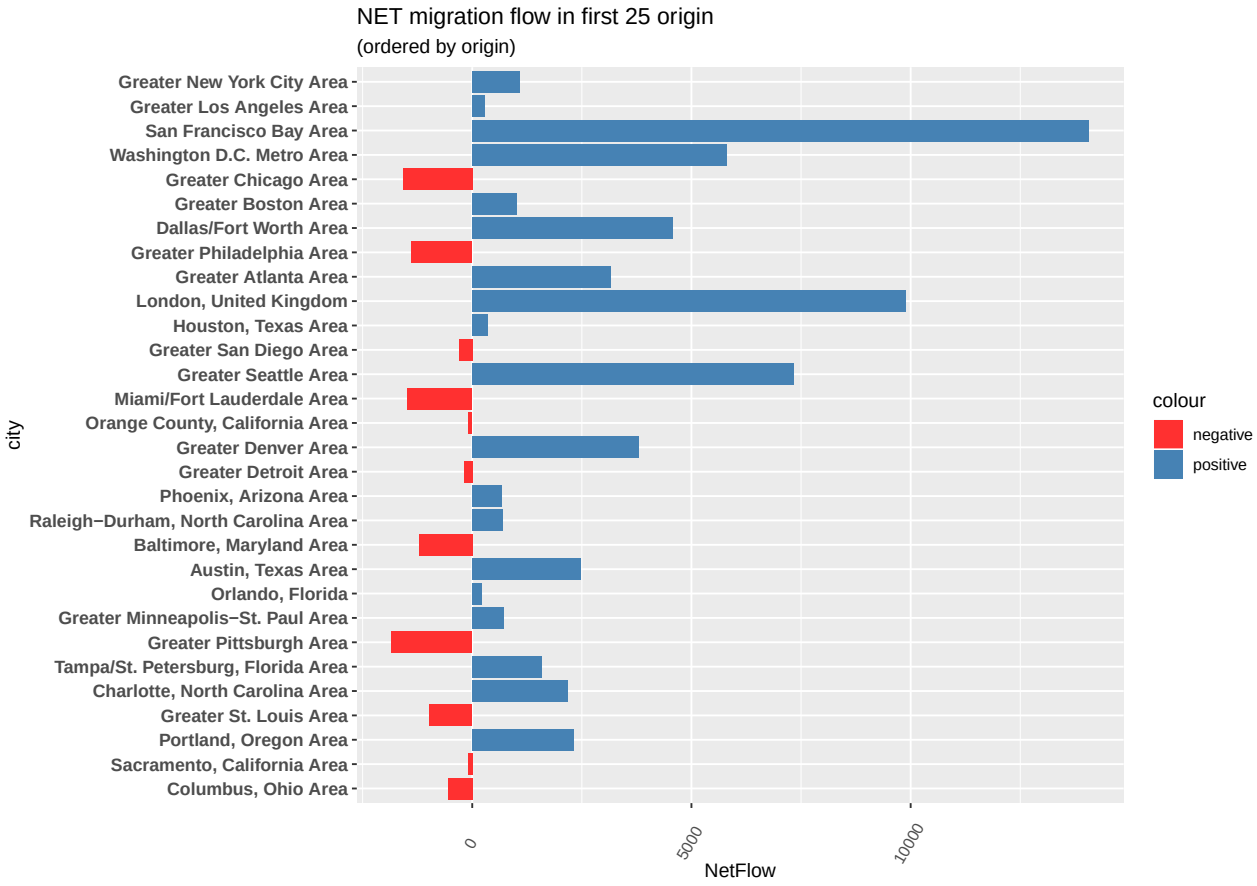

```
# 2) Out  Flow in Top ORIGIN
Top_out <- aggreByCity %>% top_n(30, NumOutflow)
  Top_out$city = with(Top_out, reorder(city,NumOutflow)) # reorder Levels by Var

Top_out <- ggplot(data = Top_out,aes(city, NegOutFlow)) +
  geom_bar(stat = "identity", position="identity", fill = "firebrick1") +
   geom_hline(yintercept=-10000, color = "black", size=0.5) +
  theme(axis.text.x = element_text(angle=60, vjust=0.3), axis.text.y = element_text(size=10,face="bold")
  labs(title="OUT migration flow in first 15 origin")

Top_out + coord_flip()
```




```
# plot side by side
# library(gridExtra)
# grid.arrange(in_flip, out_flip, ncol=2)

# 3.a)  NET flow in Top  DESTINATION
Top_net <- aggreByCity %>% top_n(30, NumInflow)
  Top_net$city = with(Top_net, reorder(city,NumInflow)) # reorder Levels by Var

Top_net <- ggplot(data = Top_net,aes(city, NetFlow)) +
  geom_bar(stat = "identity", position="identity",aes(fill = colour)) +
  theme(axis.text.x = element_text(angle=60, vjust=0.3), axis.text.y = element_text(size=10,face="bold")
  scale_fill_manual(values=c(positive="steelblue",negative="firebrick1")) +
  labs(title="NET migration flow in first 25 destination", subtitle="(ordered by destination)")

Top_net + coord_flip()
```




```r
# 3.b)  NET flow in Top  ORIGIN
Top_net <- aggreByCity %>% top_n(30, NumOutflow)
  Top_net$city = with(Top_net, reorder(city,NumOutflow)) # reorder Levels by Var

Top_net <- ggplot(data = Top_net,aes(city, NetFlow)) +
  geom_bar(stat = "identity", position="identity",aes(fill = colour)) +
  theme(axis.text.x = element_text(angle=60, vjust=0.3), axis.text.y = element_text(size=10,face="bold"
  scale_fill_manual(values=c(positive="steelblue",negative="firebrick1")) +
  labs(title="NET migration flow in first 25 origin", subtitle="(ordered by origin)")
Top_net + coord_flip()
```



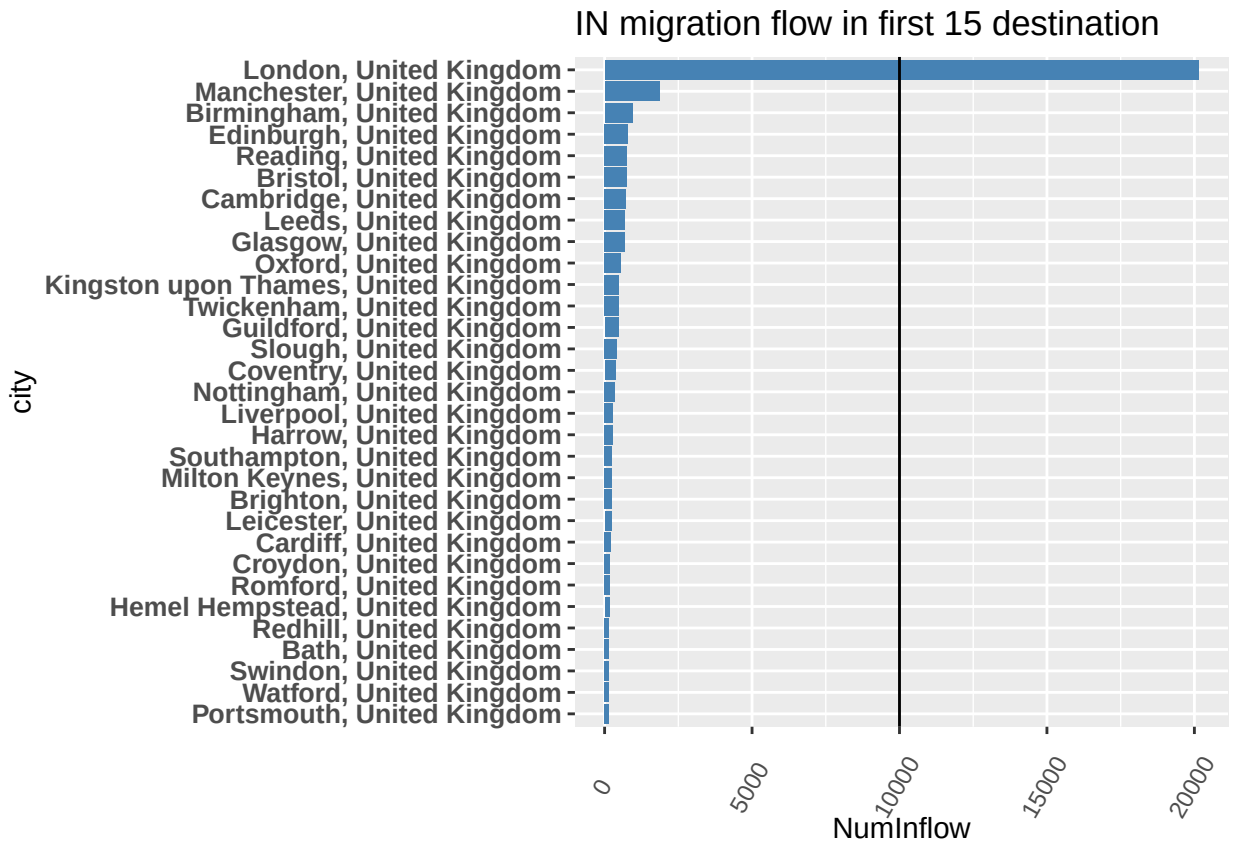

```
# all american
# lots of leaving in 2016 - especially in NY
```

```
# add country to aggreByCity
aggreByCity2 <- left_join(aggreByCity,city_country,by = "city")

# x[c("a", "b")][is.na(x[c("a", "b")])] <- 0
aggreByCity2[c("NumInflow", "NumOutflow")][is.na(aggreByCity2[c("NumInflow", "NumOutflow")])] <- 0
```

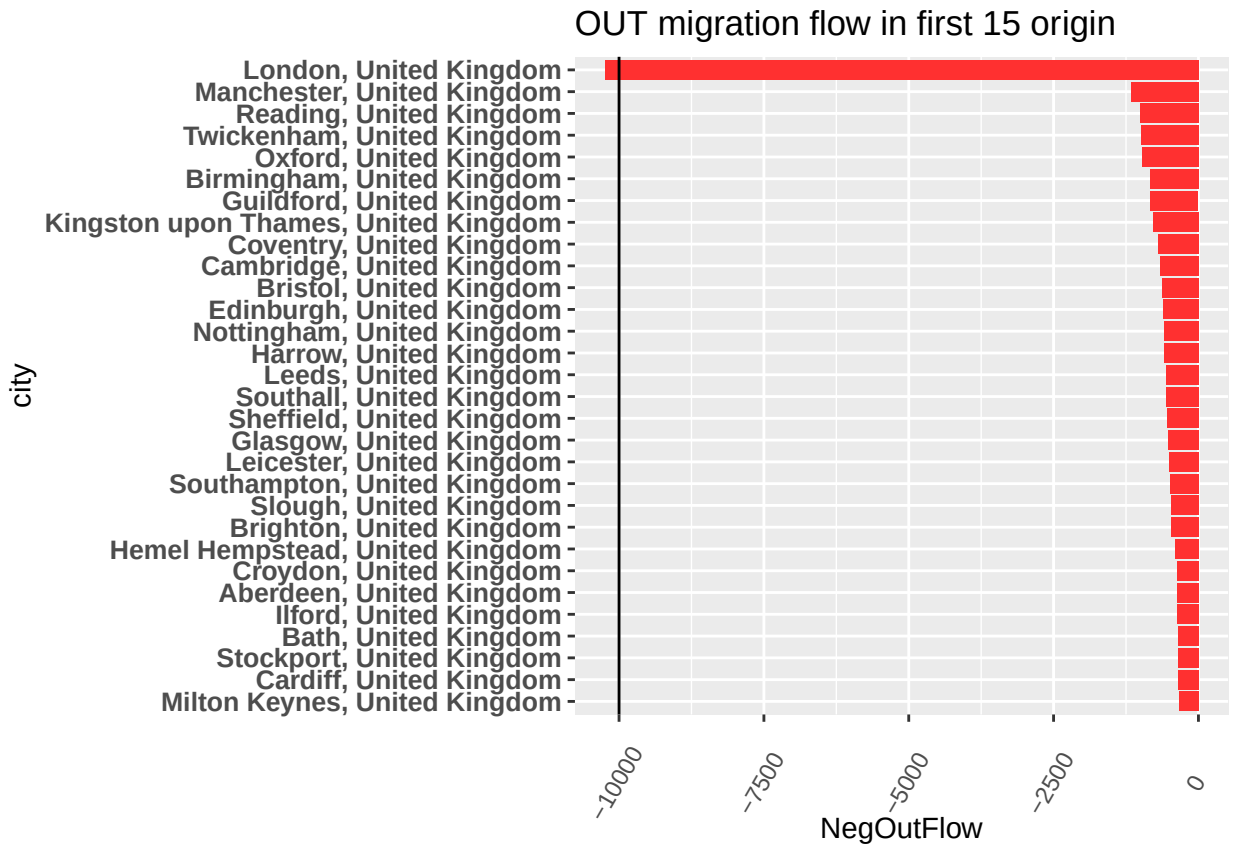
Major 20 (UK) cities TO / FROM / NET migration

```
# 1) In  Flow in Top DESTINATION

aggreByCityUK <- aggreByCity2 %>%
  filter (country == "United Kingdom")

Top_inUK <- aggreByCityUK %>%
  top_n(30, NumInflow)

Top_inUK$city = with(Top_inUK, reorder(city,NumInflow)) # reorder Levels by Var

Top_inUK <- ggplot(data = Top_inUK,aes(city, NumInflow)) +
  geom_bar(stat = "identity", position="identity", fill = "steelblue") +
  geom_hline(yintercept=10000, color = "black", size=0.5) +
  theme(axis.text.x = element_text(angle=60, vjust=0.3), axis.text.y = element_text(size=10,face="bold"
```

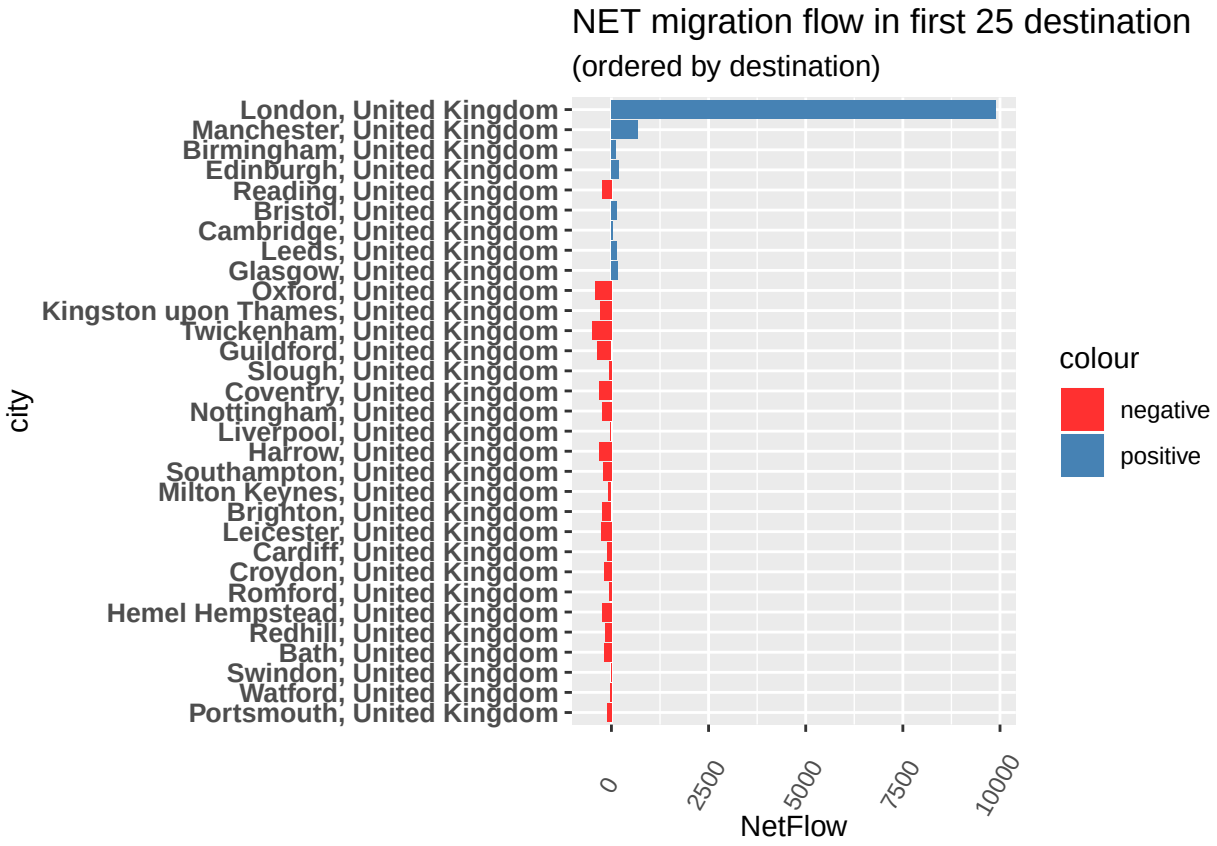
```
  labs(title="IN migration flow in first 15 destination")


Top_inUK + coord_flip()
```



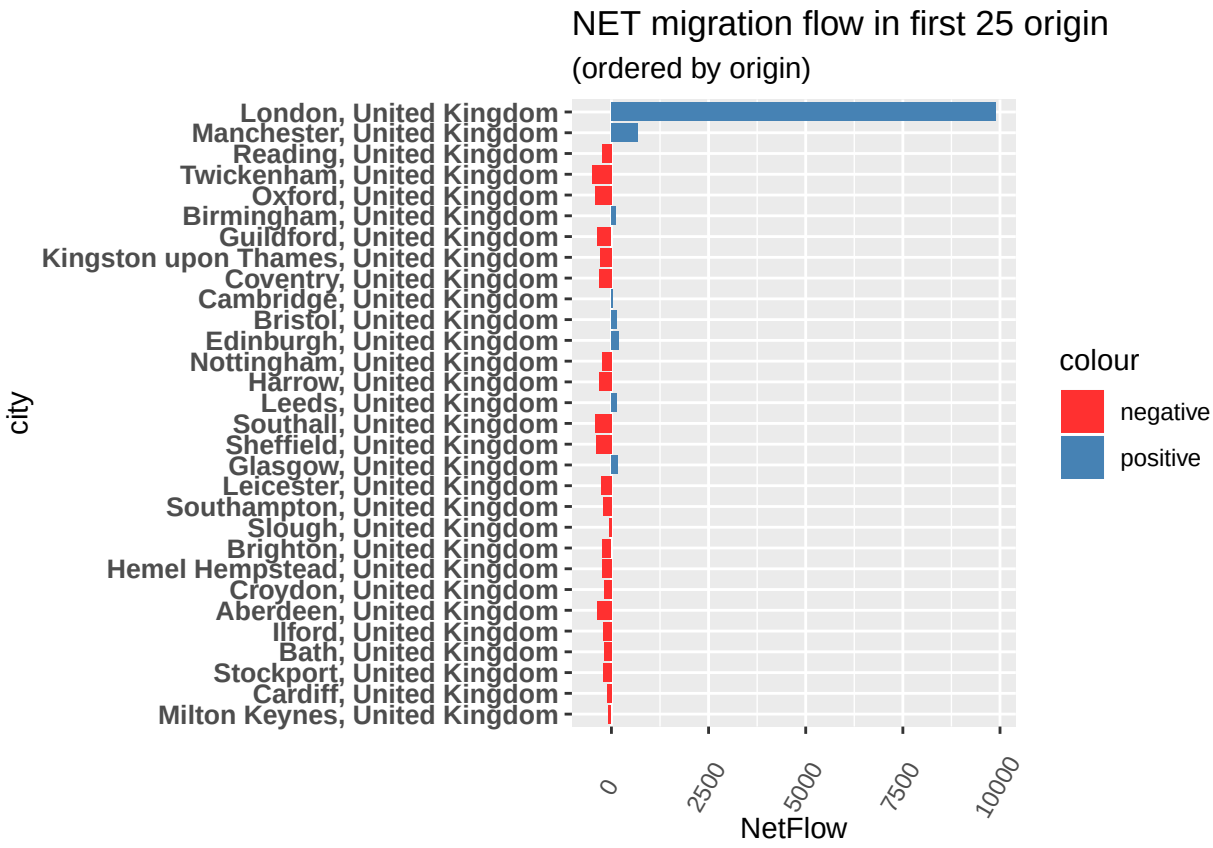

```
# 2) Out  Flow in Top ORIGIN
Top_outUK <- aggreByCityUK %>%
  filter (country == "United Kingdom") %>% # filter by country = UK
  top_n(30, NumOutflow)
  Top_outUK$city = with(Top_outUK, reorder(city,NumOutflow)) # reorder Levels by Var

Top_outUK <- ggplot(data = Top_outUK,aes(city, NegOutFlow)) +
  geom_bar(stat = "identity", position="identity", fill = "firebrick1") +
   geom_hline(yintercept=-10000, color = "black", size=0.5) +
  theme(axis.text.x = element_text(angle=60, vjust=0.3), axis.text.y = element_text(size=10,face="bold")
  labs(title="OUT migration flow in first 15 origin")

Top_outUK + coord_flip()
```




```
# 3.a)  NET flow in Top  DESTINATION
Top_netUK <- aggreByCityUK %>%
   filter (country == "United Kingdom") %>% # filter by country = UK
  top_n(30, NumInflow)
  Top_netUK$city = with(Top_netUK, reorder(city,NumInflow)) # reorder Levels by Var

Top_netUK <- ggplot(data = Top_netUK,aes(city, NetFlow)) +
  geom_bar(stat = "identity", position="identity",aes(fill = colour)) +
  theme(axis.text.x = element_text(angle=60, vjust=0.3), axis.text.y = element_text(size=10,face="bold"
  scale_fill_manual(values=c(positive="steelblue",negative="firebrick1")) +
  labs(title="NET migration flow in first 25 destination", subtitle="(ordered by destination)")
Top_netUK + coord_flip()
```



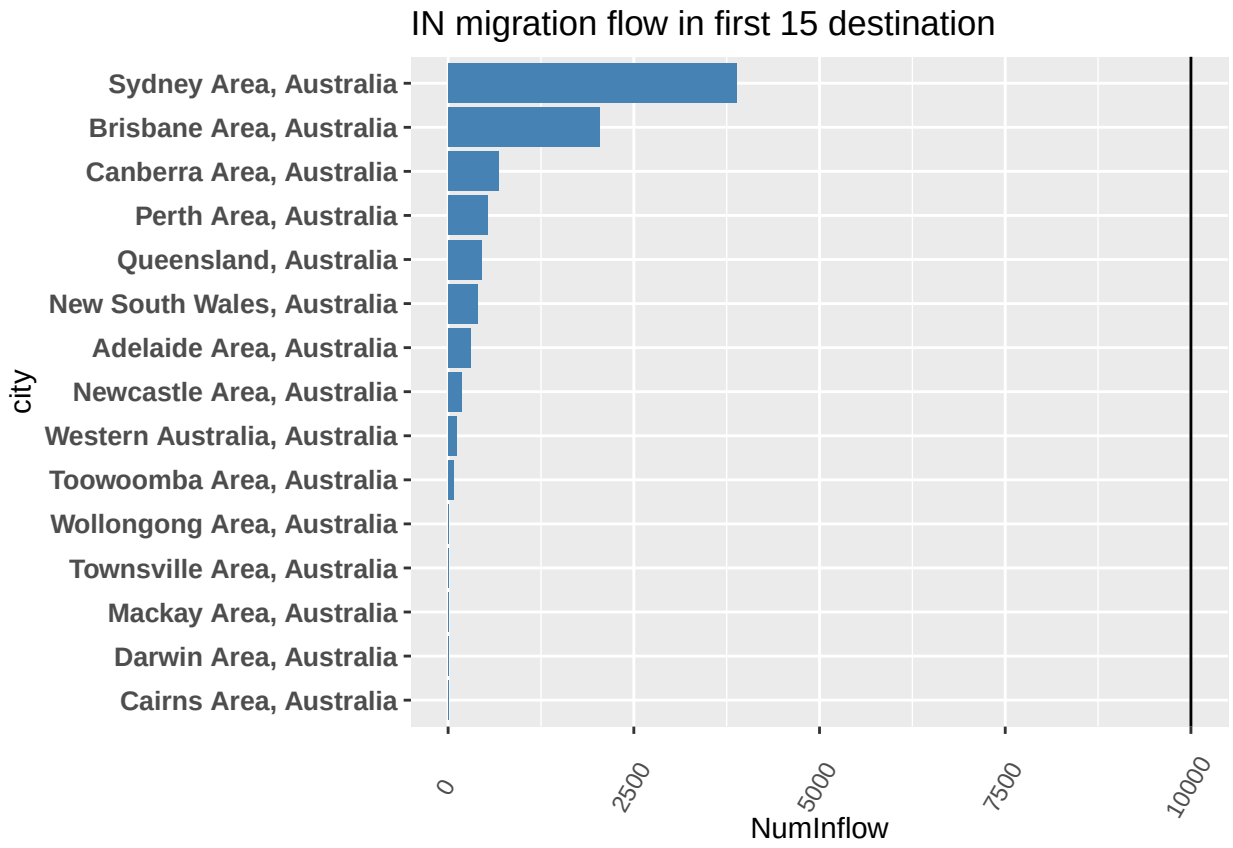

```
# 3.b)  NET flow in Top  ORIGIN
Top_netUK <- aggreByCityUK %>%
   filter (country == "United Kingdom") %>% # filter by country = UK
  top_n(30, NumOutflow)
  Top_netUK$city = with(Top_netUK, reorder(city,NumOutflow)) # reorder Levels by Var

Top_netUK <- ggplot(data = Top_netUK,aes(city, NetFlow)) +
  geom_bar(stat = "identity", position="identity",aes(fill = colour)) +
  theme(axis.text.x = element_text(angle=60, vjust=0.3), axis.text.y = element_text(size=10,face="bold")
  scale_fill_manual(values=c(positive="steelblue",negative="firebrick1")) +
  labs(title="NET migration flow in first 25 origin", subtitle="(ordered by origin)")
Top_netUK + coord_flip()
```



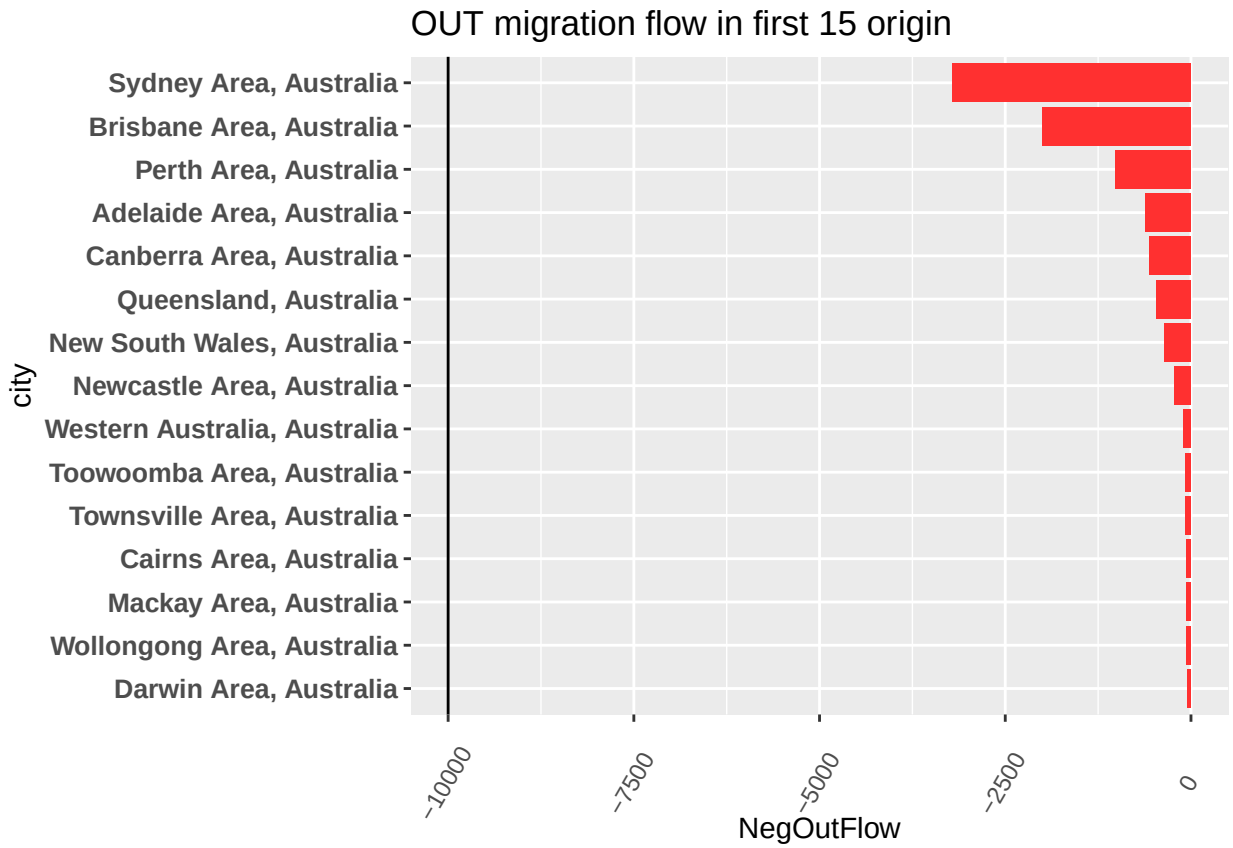
- **INSIGHTs**:
  - Contrary to US, London is a definitive outlier

Major 20 (AUSTRALIA) cities TO / FROM / NET migration

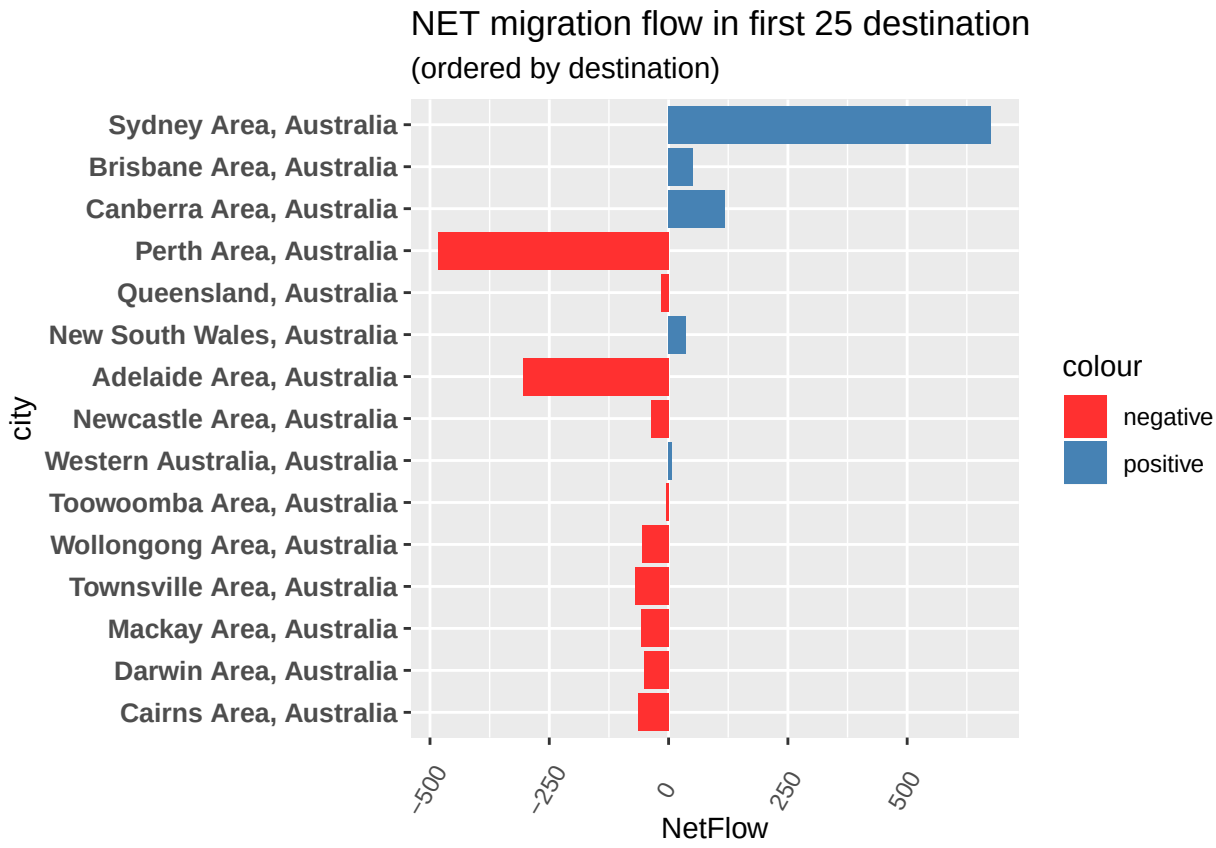
```
# 1) In  Flow in Top DESTINATION

aggreByCityAustr <- aggreByCity2 %>%
  filter (country == "Australia")

Top_inAustr <- aggreByCityAustr %>%
  top_n(30, NumInflow)

Top_inAustr$city = with(Top_inAustr, reorder(city,NumInflow)) # reorder Levels by Var

Top_inAustr <- ggplot(data = Top_inAustr,aes(city, NumInflow)) +
  geom_bar(stat = "identity", position="identity", fill = "steelblue") +
  geom_hline(yintercept=10000, color = "black", size=0.5) +
  theme(axis.text.x = element_text(angle=60, vjust=0.3), axis.text.y = element_text(size=10,face="bold")
  labs(title="IN migration flow in first 15 destination")


Top_inAustr + coord_flip()
```



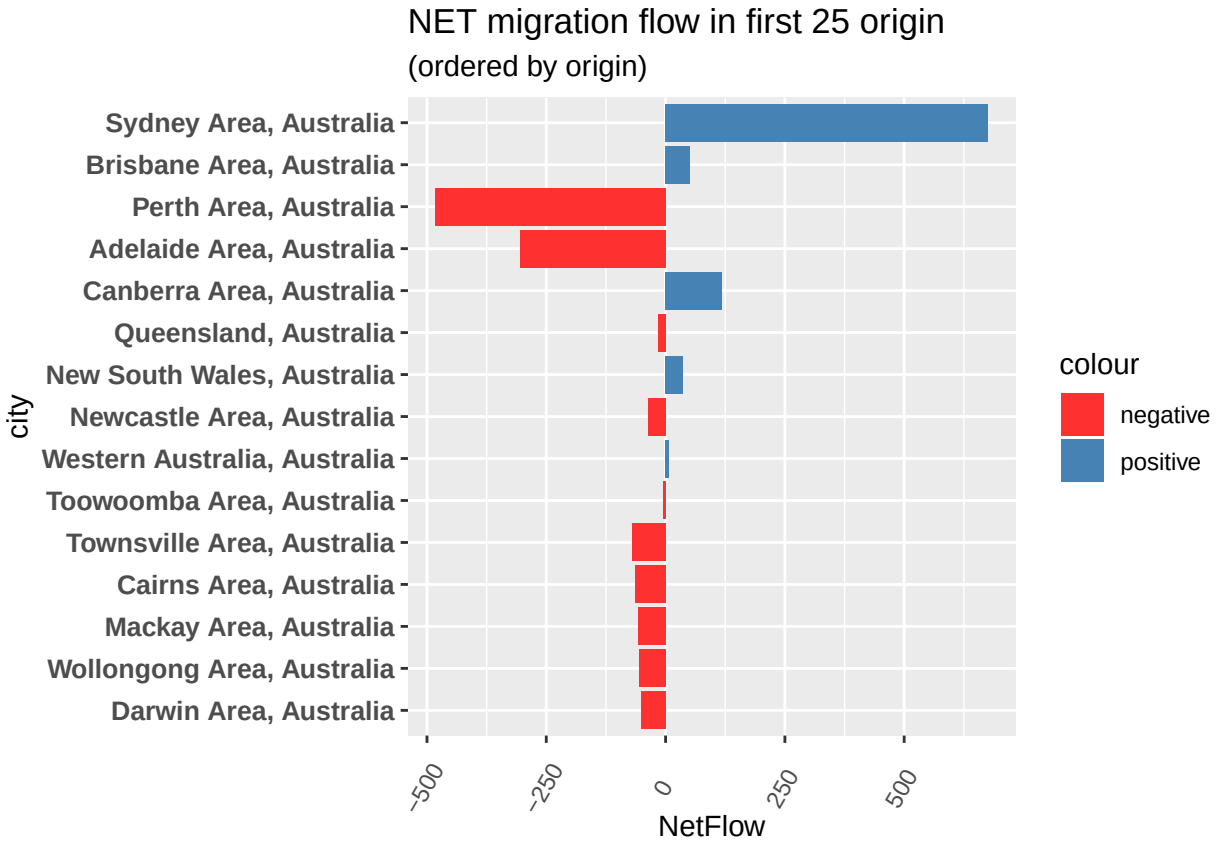

```
# 2) Out  Flow in Top ORIGIN
Top_outAustr <- aggreByCityAustr %>%
  filter (country == "Australia") %>% # filter by country = Austr
  top_n(30, NumOutflow)
  Top_outAustr$city = with(Top_outAustr, reorder(city,NumOutflow)) # reorder Levels by Var

Top_outAustr <- ggplot(data = Top_outAustr,aes(city, NegOutFlow)) +
  geom_bar(stat = "identity", position="identity", fill = "firebrick1") +
   geom_hline(yintercept=-10000, color = "black", size=0.5) +
  theme(axis.text.x = element_text(angle=60, vjust=0.3), axis.text.y = element_text(size=10,face="bold")
  labs(title="OUT migration flow in first 15 origin")

Top_outAustr + coord_flip()
```




```
# 3.a)  NET flow in Top  DESTINATION
Top_netAustr <- aggreByCityAustr %>%
   filter (country == "Australia") %>% # filter by country = Austr
  top_n(30, NumInflow)
  Top_netAustr$city = with(Top_netAustr, reorder(city,NumInflow)) # reorder Levels by Var

Top_netAustr <- ggplot(data = Top_netAustr,aes(city, NetFlow)) +
  geom_bar(stat = "identity", position="identity",aes(fill = colour)) +
  theme(axis.text.x = element_text(angle=60, vjust=0.3), axis.text.y = element_text(size=10,face="bold"
  scale_fill_manual(values=c(positive="steelblue",negative="firebrick1")) +
  labs(title="NET migration flow in first 25 destination", subtitle="(ordered by destination)")
Top_netAustr + coord_flip()
```




```
# 3.b)  NET flow in Top  ORIGIN
Top_netAustr <- aggreByCityAustr %>%
   filter (country == "Australia") %>% # filter by country = Austr
  top_n(30, NumOutflow)
  Top_netAustr$city = with(Top_netAustr, reorder(city,NumOutflow)) # reorder Levels by Var

Top_netAustr <- ggplot(data = Top_netAustr,aes(city, NetFlow)) +
  geom_bar(stat = "identity", position="identity",aes(fill = colour)) +
  theme(axis.text.x = element_text(angle=60, vjust=0.3), axis.text.y = element_text(size=10,face="bold")
  scale_fill_manual(values=c(positive="steelblue",negative="firebrick1")) +
  labs(title="NET migration flow in first 25 origin", subtitle="(ordered by origin)")
Top_netAustr + coord_flip()
```



4 BIVARIATE MEASURES OF ASSOCIATION

4.1 INflow by city vs highest degree

Is there any relation between where they choose to go and the highest degree they have? I use plots to check distributions type of DEGREE (Y) conditional on the city of destination (X)

- I check for :
  - *Existence*
  - *strnght*
  - *patterns* / direction

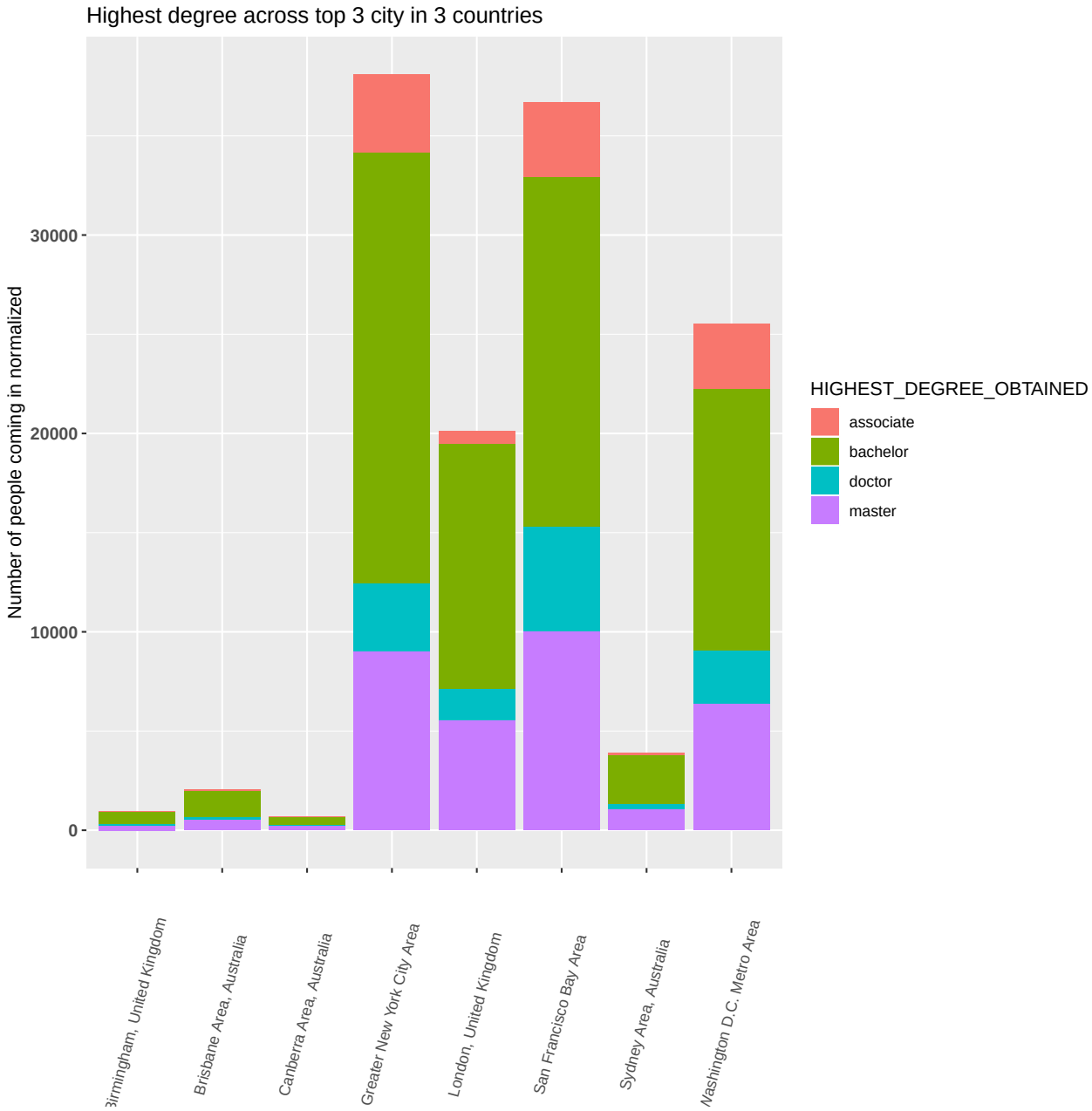
Greater New York City Area San Francisco Bay Area Washington D.C. Metro Area

Manchester, United Kingdom London, United Kingdom Birmingham, United Kingdom

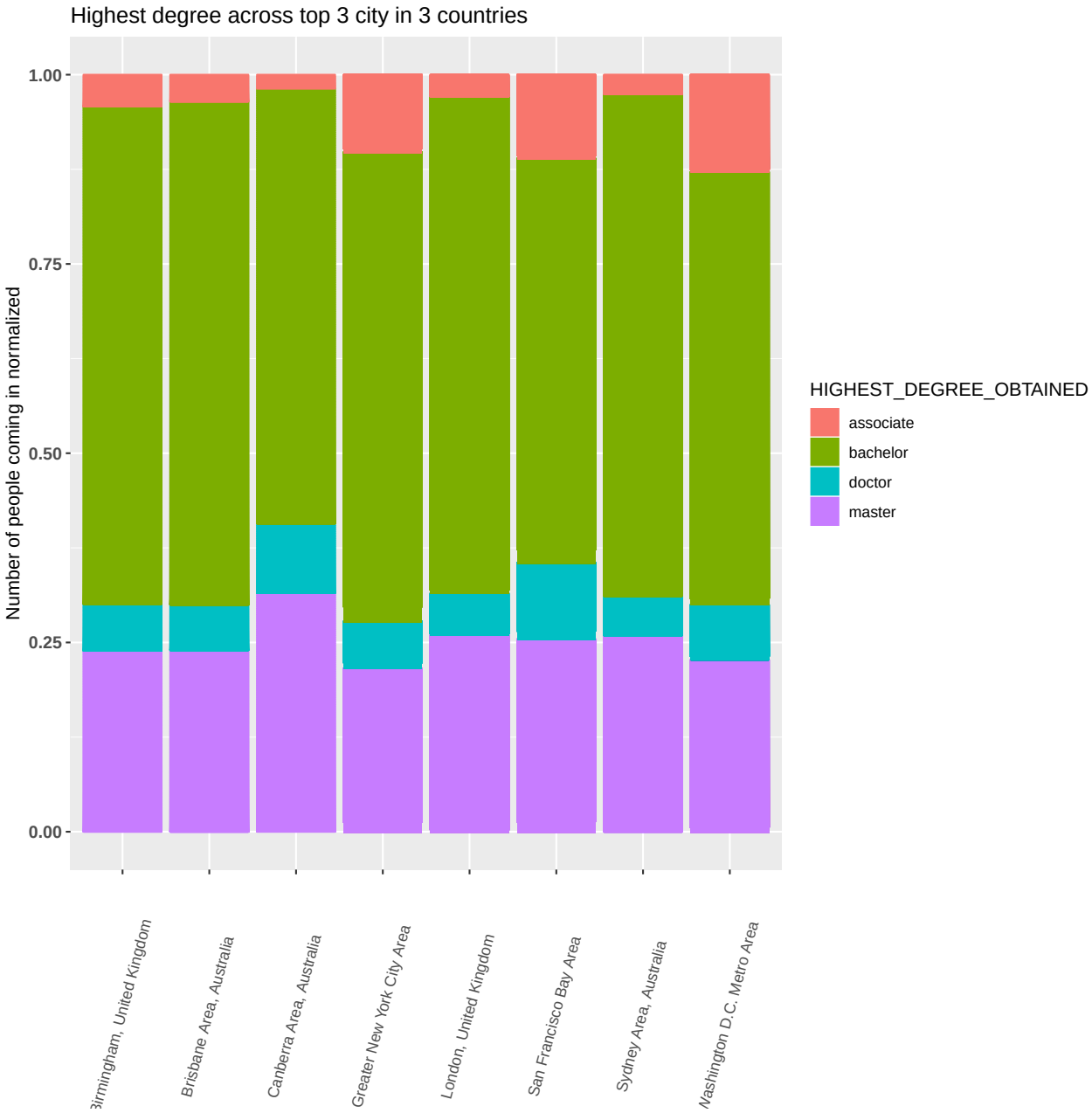
Sydney Area, Australia Brisbane Area, Australia Canberra Area, Australia

```
# 1) attempt for simplicity I select top 3 per country

#top3 <- both %>% filter(DESTINATION_REGION == "Greater New York City Area" |DESTINATION_REGION =="San

# 2) attempt

# top_subs <- subset(both, (DESTINATION_REGION == 'Greater New York City Area' | DESTINATION_REGION =='
```

```r
# 3) attempt
# When subsetting with [ names are always matched exactly
# z <- c(abc = 1, def = 2)
# z[c("a", "d")]

top_dest <- c("Greater New York City Area" , "San Francisco Bay Area",  "Washington D.C. Metro Area" ,

 # both_top <- both[as.character(  both$DESTINATION_REGION %in% top_dest), drop = T]

# data[data$Code %in% selected,]
# both_top <- both[both$DESTINATION_REGION %in% top_dest]
# both_top <- both[as.character(  both$DESTINATION_REGION %in% top_dest), drop = TRUE]

# 4) attemps
# data[data$Code == "A" | data$Code == "B", ]
top3cit <- both[both$DESTINATION_REGION == "Greater New York City Area" | both$DESTINATION_REGION == "Sa


# 5) attemps
# top_dest <- c("Greater New York City Area" , "San Francisco Bay Area",  "Washington D.C. Metro Area"


# top3cit_2 <-  both[both$DESTINATION_REGION %in%  top_dest, , drop =TRUE ]

 # mutate (Sign = ifelse(NetFlow > 0, "Positive", "Negative"))


# Explore

# mosaicplot(table(top3cit$DESTINATION_REGION, top3cit$HIGHEST_DEGREE_OBTAINED), ylab = "Political Part

# qplot
qplot(x = DESTINATION_REGION, data = top3cit, fill = HIGHEST_DEGREE_OBTAINED, geom = "bar") +
theme(axis.text.x = element_text(angle=75, vjust=0.3), axis.text.y = element_text(size=10,face="bold")
  labs(title="Highest degree across top 3 city in 3 countries" , x="", y="Number of people coming in no
```




```
# ggplot

#ggplot(data = Best_out,aes(cityD, NegOutFlow)) +
#  geom_bar(stat = "identity", position="identity", fill = "firebrick1") +
#   geom_hline(yintercept=-10000, color = "black", size=0.5) +
#  theme(axis.text.x = element_text(angle=60, vjust=0.3)) +
#  labs(title="OUT migration flow in first 15 destination")


ggplot(top3cit, aes(x=DESTINATION_REGION, y=NEW_MEM_ID, fill=HIGHEST_DEGREE_OBTAINED)) +
  geom_bar(aes(colour =HIGHEST_DEGREE_OBTAINED),stat="identity", position = "fill") +
  theme(axis.text.x = element_text(angle=75, vjust=0.3), axis.text.y = element_text(size=10,face="bold"
  labs(title="Highest degree across top 3 city in 3 countries" , x="", y="Number of people coming in no
```



- **INSIGHTs**:
  - Intuitively, it would seem San Francisco (follwed by Canberra) attracts the highest amount of doctors (Canberra also the highest group with master)
  - NY, San Francisco and Washington DC seem to receive many with "Associate" level: either young people go tehre to look for tehir first job

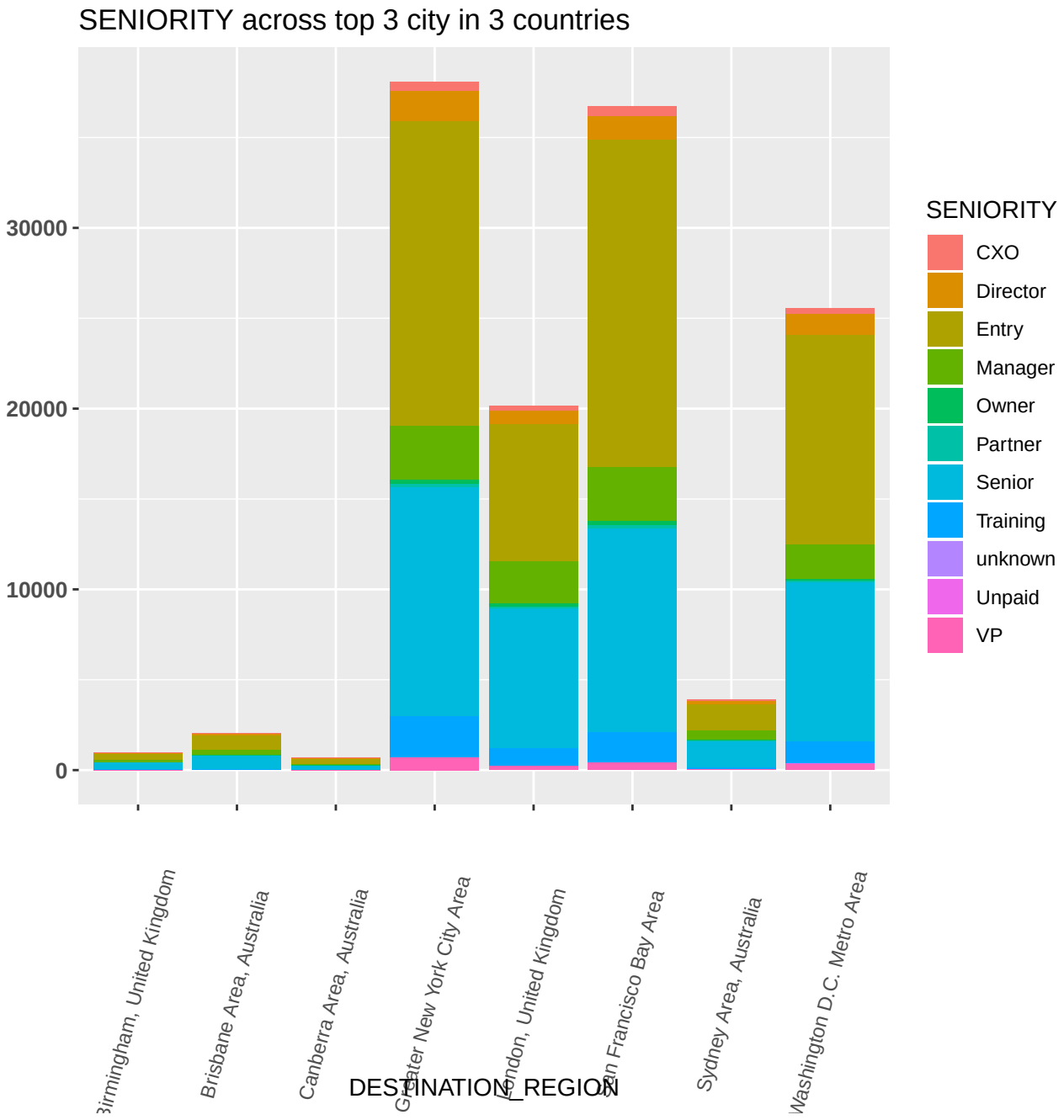
```
# cramer v   degree  X CITY OF DESTINATION (top 3)
x<- top3cit$DESTINATION_REGION
y<- top3cit$HIGHEST_DEGREE_OBTAINED

cv.test = function(x,y) {
  CV = sqrt(chisq.test(x, y, correct=FALSE)$statistic /
    (length(x) * (min(length(unique(x)),length(unique(y))) - 1)))
  print.noquote("Cramér V / Phi:")
```

```
  return(as.numeric(CV))
}

with(top3cit, cv.test(x, y)) # [1] Cramér V / Phi: 0.09052046
```

```
## [1] Cramér V / Phi:
```

```
## [1] 0.09052046
```

```
#
# mosaicplot(table(top3cit$DESTINATION_REGION, top3cit$SENIORITY), ylab = "Political Party", xlab = "Ta

# qplot
qplot(x = DESTINATION_REGION, data = top3cit, fill = SENIORITY, geom = "bar") +
theme(axis.text.x = element_text(angle=75, vjust=0.3), axis.text.y = element_text(size=10,face="bold") )
  labs(title="SENIORITY across top 3 city in 3 countries")
```



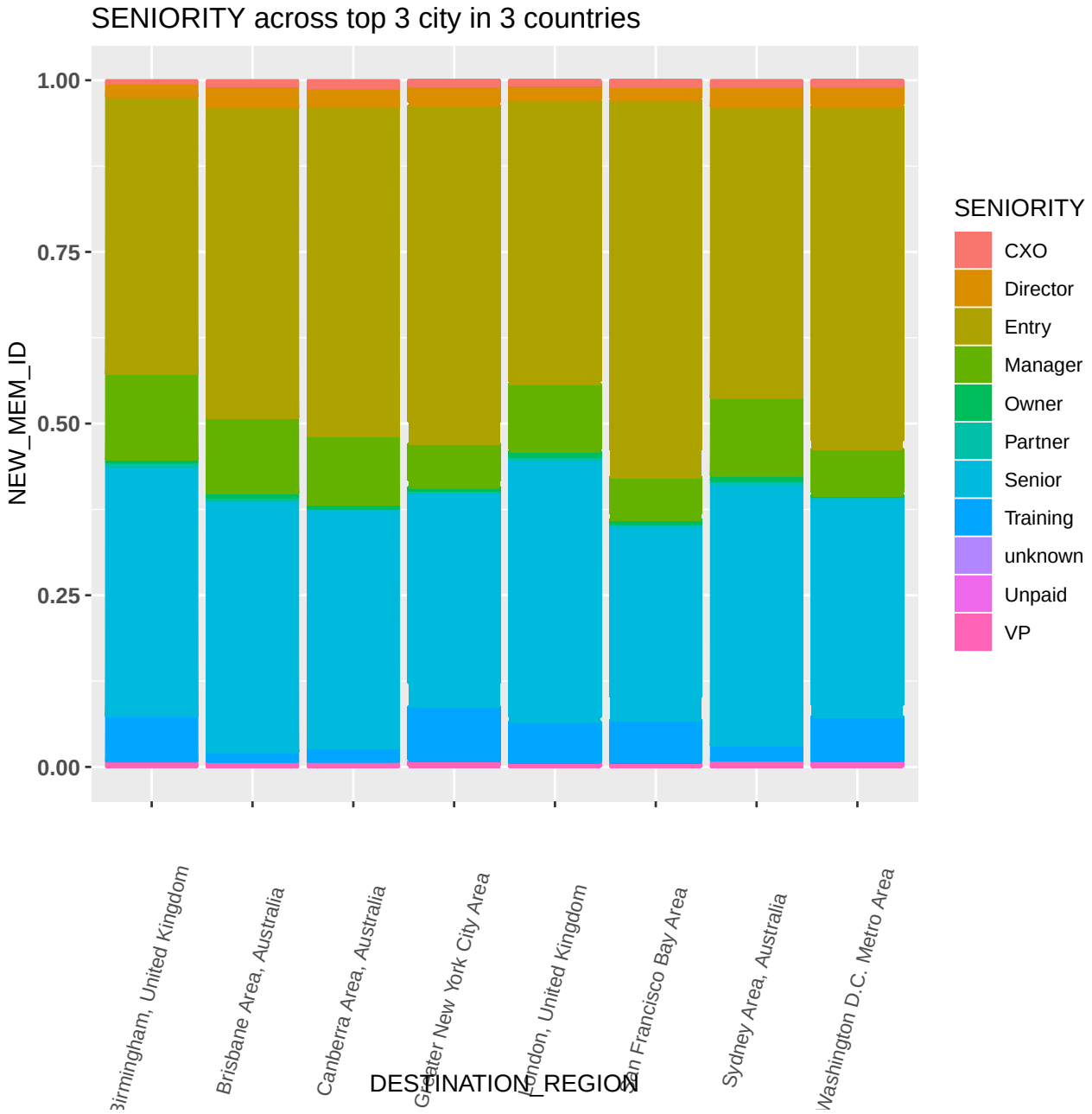

```
# Test statistic (assuming indipendence)
chisq <- chisq.test(x = table(top3cit$DESTINATION_REGION, top3cit$SENIORITY), correct = FALSE)
```

```
## Warning in chisq.test(x = table(top3cit$DESTINATION_REGION, top3cit
## $SENIORITY), : Chi-squared approximation may be incorrect
```

```
chisq
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(top3cit$DESTINATION_REGION, top3cit$SENIORITY)
## X-squared = NaN, df = 2810, p-value = NA
```

```
# ggplot
ggplot(top3cit, aes(x=DESTINATION_REGION, y=NEW_MEM_ID, fill=SENIORITY)) +  geom_bar(aes(colour =SENIOR
theme(axis.text.x = element_text(angle=75, vjust=0.3), axis.text.y = element_text(size=10,face="bold")
  labs(title="SENIORITY across top 3 city in 3 countries")
```



* **INSIGHTs**: + Intuitively, it would seem San Francisco (follwed by Canberra) attracts the highest amount of doctors (Canberra also the highest group with master) + NY, San Francisco and Washington DC seem to receive many with "Associate" level: either young people go tehre to look for tehir first job

```
# cramer v   SENIORITY   X CITY OF DESTINATION (top 3)
x<- top3cit$DESTINATION_REGION
y<- top3cit$SENIORITY

cv.test = function(x,y) {
```

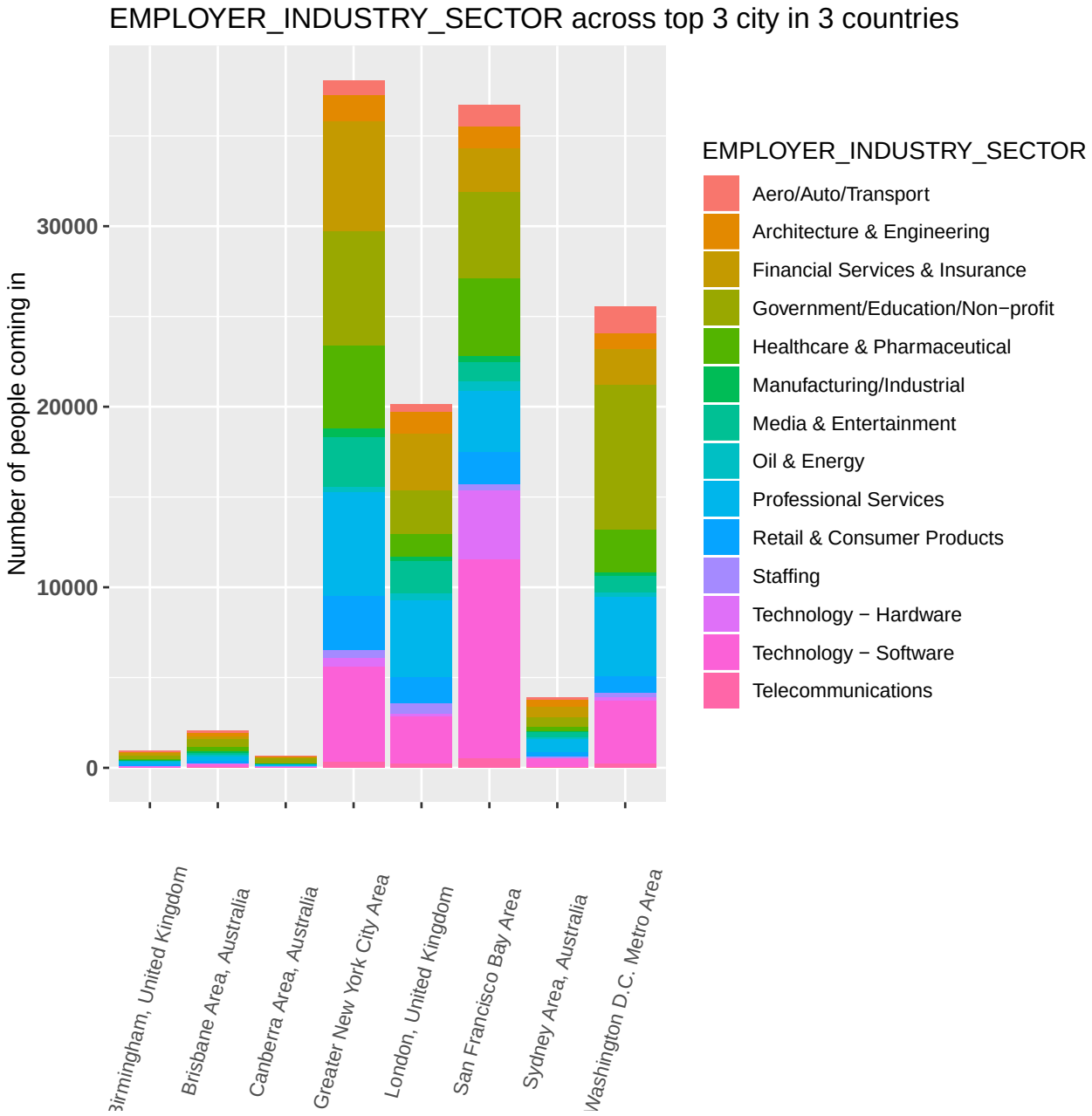
```
  CV = sqrt(chisq.test(x, y, correct=FALSE)$statistic /
    (length(x) * (min(length(unique(x)),length(unique(y))) - 1)))
  print.noquote("Cramér V / Phi:")
  return(as.numeric(CV))
}

with(top3cit, cv.test(x, y)) # [1] Cramér V / Phi: 0.04360073
```

```
## Warning in chisq.test(x, y, correct = FALSE): Chi-squared approximation may
## be incorrect
```

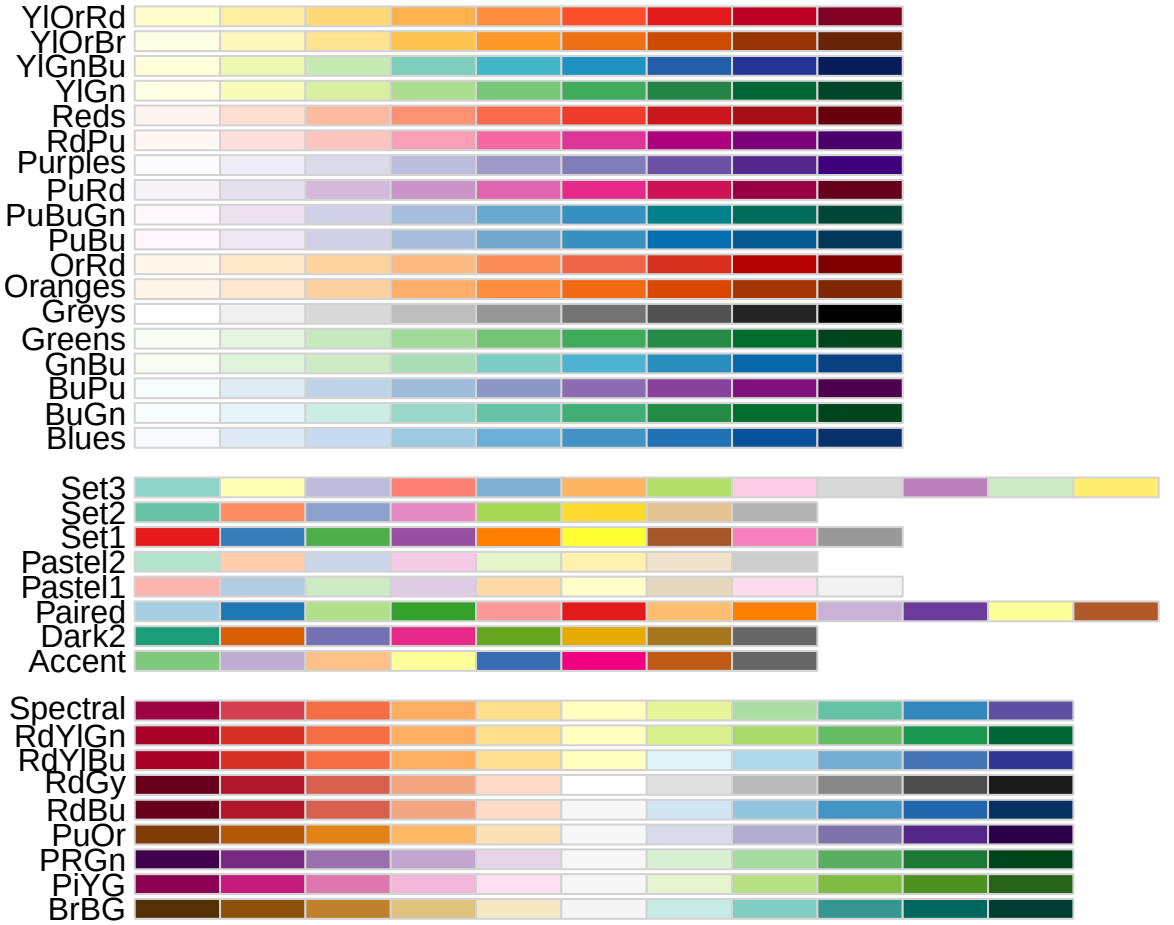
```
## [1] Cramér V / Phi:
```

```
## [1] 0.04360073
```

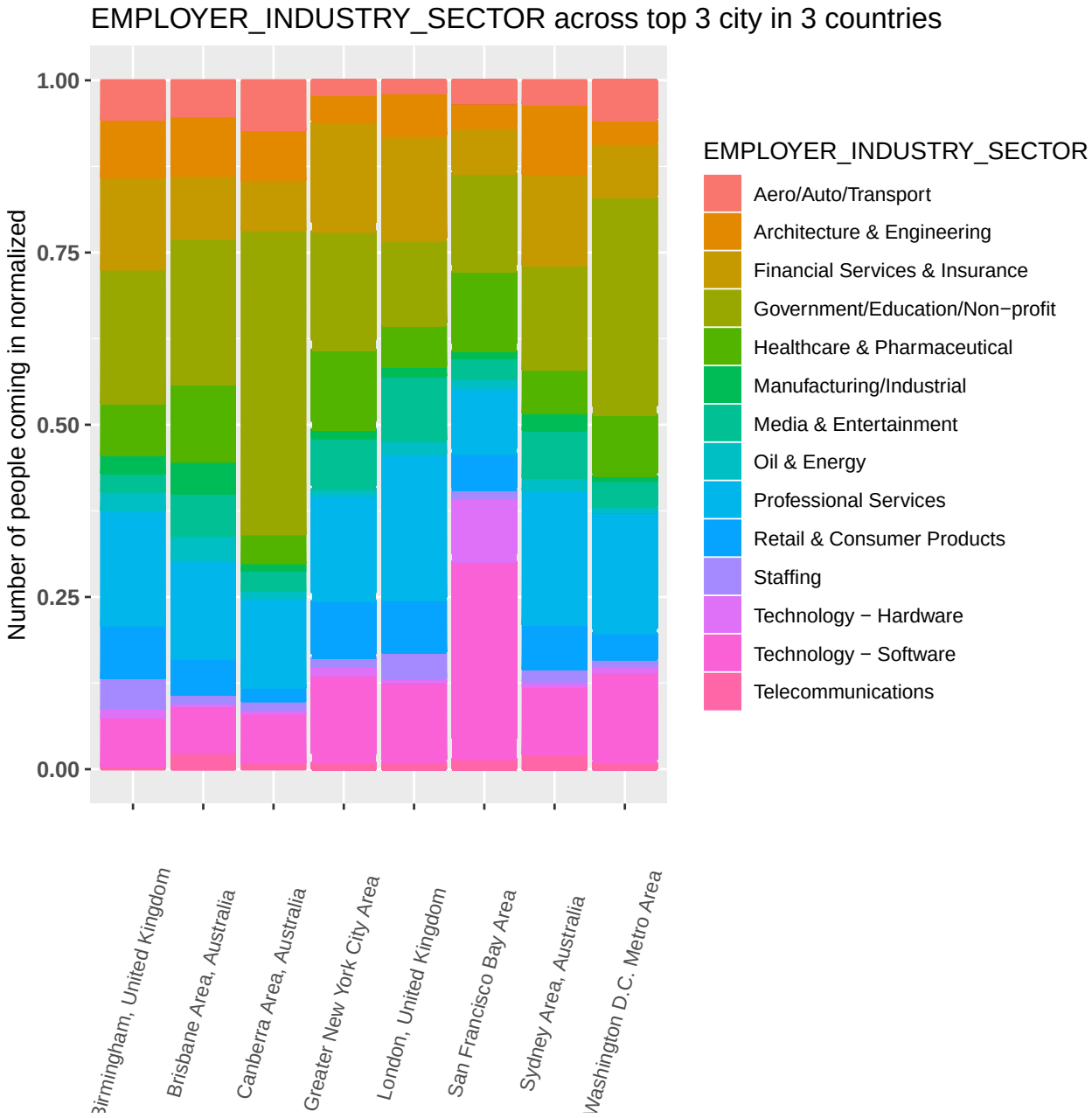
```
#
# mosaicplot(table(top3cit$DESTINATION_REGION, top3cit$EMPLOYER_INDUSTRY_SECTOR), ylab = "Political Par

# qplot
qplot(x = DESTINATION_REGION, data = top3cit, fill = EMPLOYER_INDUSTRY_SECTOR, geom = "bar") +
theme(axis.text.x = element_text(angle=75, vjust=0.3), axis.text.y = element_text(size=10,face="bold") )
 labs(title="EMPLOYER_INDUSTRY_SECTOR across top 3 city in 3 countries", x="", y="Number of people comi
```




```
# ggplot
library(RColorBrewer)
display.brewer.all()
```




```
ggplot(top3cit, aes(x=DESTINATION_REGION, y=NEW_MEM_ID, fill=EMPLOYER_INDUSTRY_SECTOR)) +
  geom_bar(aes(colour =EMPLOYER_INDUSTRY_SECTOR),stat="identity",  position = "fill") +
theme(legend.position = "right"    ,axis.text.x = element_text(angle=75, vjust=0.3), axis.text.y = eleme

   # scale_fill_brewer(palette = "Dark2") +
  # scale_fill_grey() +
  labs(title="EMPLOYER_INDUSTRY_SECTOR across top 3 city in 3 countries", x="", y="Number of people com
```



- **INSIGHTs**:
  - Intuitively, share of immigratns by sector seems to vary a lot in different cities
  - Extremely high number in Govv/ Educ/ Non Profit in Canberra & Washington
  - Extremely high number in Software + Hardware Technology in San Francisco
  - Extremely high number in Govv/ Educ/ Non Profit + Financial sErvices also in NY it would seem San Francisco (follwed by Canberra) attracts the highest amount of doctors (Canberra also the highest group with master)
  - NY, San Francisco and Washington DC seem to receive many with "Associate" level: either young people go tehre to look for tehir first job

CHECK A COUPLE OF CITIES FOR SECTOR X SEIORITY

```
freq_OrigDegree <- both %>%
  group_by(both[,7],both[,2]) %>%
  summarise (n = n()) %>%
  mutate(freq = n / sum(n)) %>%
  mutate(rel.freq = paste0(round(100 * n/sum(n), 2), "%"))
freq_OrigDegree  # US has a significantly higher # of Associates leaving (18% vs 3% and 4%)
```

```
## # A tibble: 12 x 5
## # Groups:   both[, 7] [3]
##    `both[, 7]`    `both[, 2]`      n   freq rel.freq
##    <fct>          <fct>        <int>  <dbl> <chr>
##  1 Australia      associate      255 0.0286 2.86%
##  2 Australia      bachelor      5632 0.633  63.27%
##  3 Australia      doctor         611 0.0686 6.86%
##  4 Australia      master        2404 0.270  27.01%
##  5 United Kingdom associate     1312 0.0358 3.58%
##  6 United Kingdom bachelor     22107 0.604  60.38%
##  7 United Kingdom doctor        3533 0.0965 9.65%
##  8 United Kingdom master        9663 0.264  26.39%
##  9 United States  associate    79505 0.185  18.5%
## 10 United States  bachelor    225213 0.524  52.4%
## 11 United States  doctor       36358 0.0846 8.46%
## 12 United States  master       88723 0.206  20.64%
```

Cramer's V

a measure of association for nominal variables. Effectively it is the Pearson chi-square statistic rescaled to have values between 0 and 1, as follows:

$$\phi_c = \sqrt{\frac{\chi^2}{N * (min(ncols, nrows) - 1)}}$$

where X^2 is the Pearson chi-square, nobs represents the number of observations included in the table, and where ncols and nrows are the number of rows and columns in the table, respectively.

For a 2 by 2 table, of course, this is just the square root of chi-square divided by the number of observations, which is also known as the ϕ coefficient.

Cramer's V varies from 0 (corresponding to no association between the variables) to 1 (complete association) and can reach 1 only when the two variables are equal to each other

```
# cramer v   Sector X CITY OF DESTINATION (top 3)
x<- top3cit$DESTINATION_REGION
y<- top3cit$EMPLOYER_INDUSTRY_SECTOR

cv.test = function(x,y) {
  CV = sqrt(chisq.test(x, y, correct=FALSE)$statistic /
    (length(x) * (min(length(unique(x)),length(unique(y))) - 1)))
  print.noquote("Cramér V / Phi:")
  return(as.numeric(CV))
}

with(top3cit, cv.test(x, y)) # [1] Cramér V / Phi: 0.1605205
```

```
## [1] Cramér V / Phi:
```

```
## [1] 0.1605205
```

```
# how about across all cities? (lower)
x<- both$DESTINATION_REGION
y<- both$EMPLOYER_INDUSTRY_SECTOR

cv.test = function(x,y) {
  CV = sqrt(chisq.test(x, y, correct=FALSE)$statistic /
    (length(x) * (min(length(unique(x)),length(unique(y))) - 1)))
  print.noquote("Cramér V / Phi:")
  return(as.numeric(CV))
}

with(both, cv.test(x, y)) # [1] Cramér V / Phi:  0.1264835
```

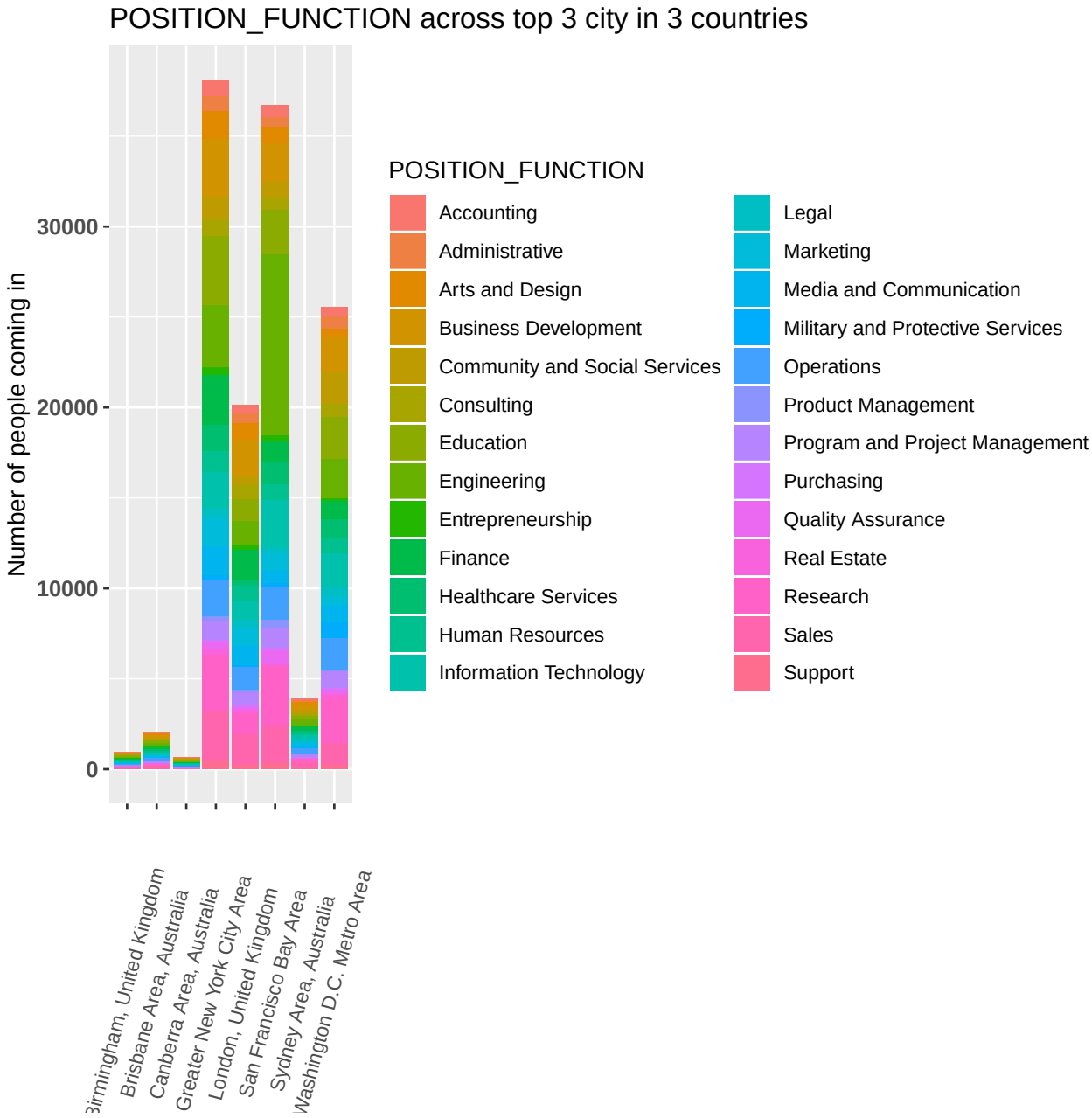
```
## Warning in chisq.test(x, y, correct = FALSE): Chi-squared approximation may
## be incorrect
```

```
## [1] Cramér V / Phi:
```

```
## [1] 0.1264835
```

position

```
#
# mosaicplot(table(top3cit$DESTINATION_REGION, top3cit$POSITION_FUNCTION), ylab = "Political Party", xl

# qplot
qplot(x = DESTINATION_REGION, data = top3cit, fill = POSITION_FUNCTION, geom = "bar") +
theme(axis.text.x = element_text(angle=75, vjust=0.3), axis.text.y = element_text(size=10,face="bold") )
  labs(title="POSITION_FUNCTION across top 3 city in 3 countries" , x="", y="Number of people coming in
```



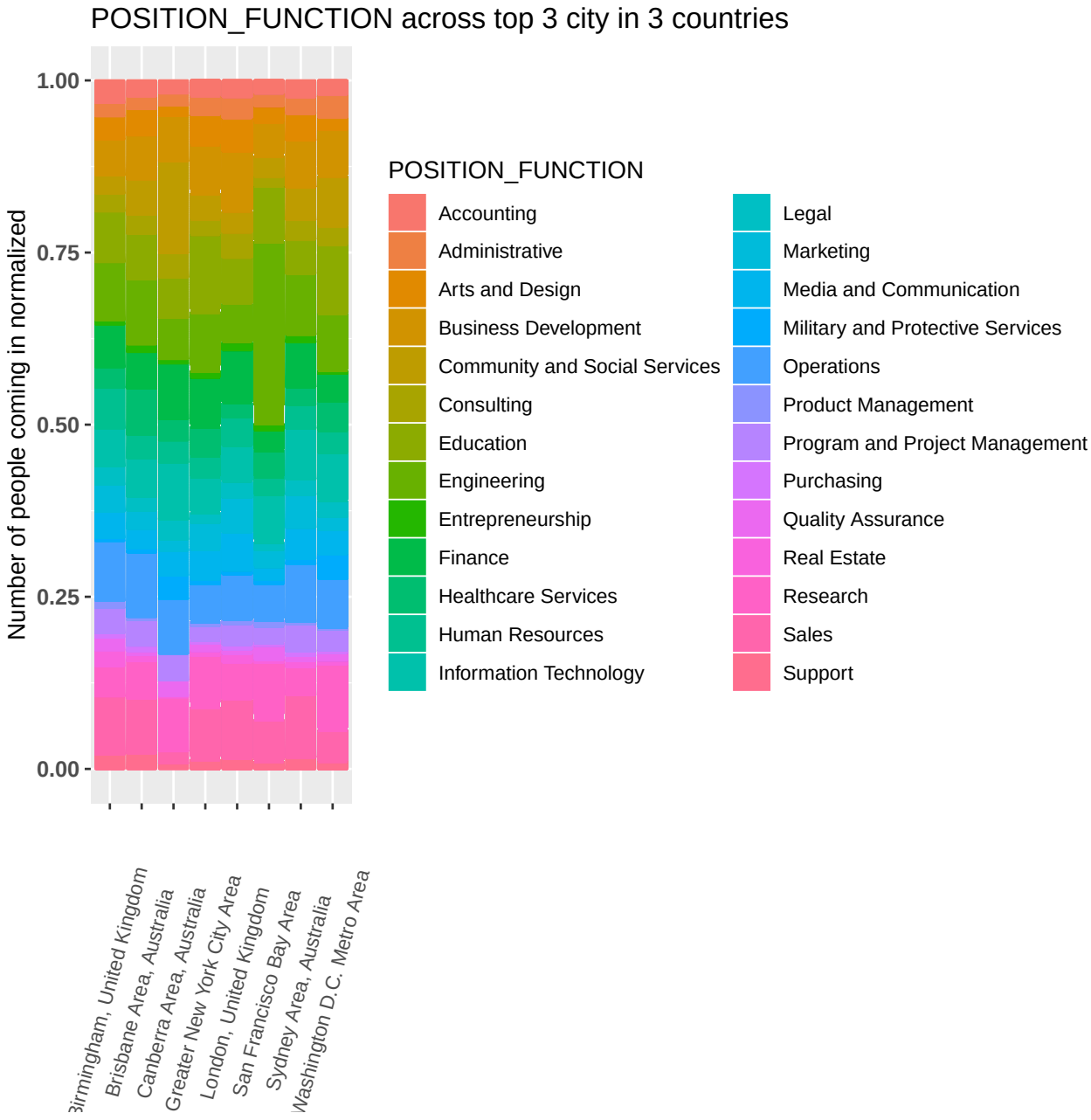

```
# Test statistic (assuming indipendence)
chisq <- chisq.test(x = table(top3cit$DESTINATION_REGION, top3cit$POSITION_FUNCTION), correct = FALSE)
```

```
## Warning in chisq.test(x = table(top3cit$DESTINATION_REGION, top3cit
## $POSITION_FUNCTION), : Chi-squared approximation may be incorrect
```

```
chisq
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(top3cit$DESTINATION_REGION, top3cit$POSITION_FUNCTION)
## X-squared = NaN, df = 7025, p-value = NA
```

```
# ggplot
ggplot(top3cit, aes(x=DESTINATION_REGION, y=NEW_MEM_ID, fill=POSITION_FUNCTION),  colour="black")    +
  geom_bar(aes(colour =POSITION_FUNCTION),stat="identity", position = "fill" )  +
theme(axis.text.x = element_text(angle=75, vjust=0.3), axis.text.y = element_text(size=10,face="bold") )
 labs(title="POSITION_FUNCTION across top 3 city in 3 countries" , x="", y="Number of people coming in n
```




```
# cramer v   POSITION_FUNCTION X CITY OF DESTINATION (top 3)
x<- top3cit$DESTINATION_REGION
y<- top3cit$POSITION_FUNCTION

cv.test = function(x,y) {
  CV = sqrt(chisq.test(x, y, correct=FALSE)$statistic /
    (length(x) * (min(length(unique(x)),length(unique(y))) - 1)))
```

```
  print.noquote("Cramér V / Phi:")
  return(as.numeric(CV))
}

with(top3cit, cv.test(x, y)) # [1] Cramér V / Phi: [1] 0.1318759
```

```
## Warning in chisq.test(x, y, correct = FALSE): Chi-squared approximation may
## be incorrect
```

```
## [1] Cramér V / Phi:
```

```
## [1] 0.1318759
```

```
# how about across all cities? (lower)
x<- both$DESTINATION_REGION
y<- both$POSITION_FUNCTION

cv.test = function(x,y) {
  CV = sqrt(chisq.test(x, y, correct=FALSE)$statistic /
    (length(x) * (min(length(unique(x)),length(unique(y))) - 1)))
  print.noquote("Cramér V / Phi:")
  return(as.numeric(CV))
}

with(both, cv.test(x, y)) # [1] Cramér V / Phi:  [1] 0.072097771
```

```
## Warning in chisq.test(x, y, correct = FALSE): Chi-squared approximation may
## be incorrect
```

```
## [1] Cramér V / Phi:
```

```
## [1] 0.07209771
```

5. DOMESTIC MIGRATION

comparative analysis of domestic / patterns of internal migration in each country

I will focus on the sector per city which seems the most significant bivariate association

construct different samples

```
# select only USA
both_USA <- both %>% filter(SOURCE_COUNTRY=="United States" & DESTINATION_COUNTRY=="United States")

# select only Australia
both_AUS <- both %>% filter(SOURCE_COUNTRY=="Australia" & DESTINATION_COUNTRY=="Australia")

# select only UK
both_UK <- both %>% filter(SOURCE_COUNTRY=="United Kingdom" & DESTINATION_COUNTRY=="United Kingdom")
```

construct USA internal sub-sample (for simplicity)

```
both_USA_N <- both_USA %>% group_by(DESTINATION_REGION) %>%  summarise(numIMM = n())
both_USA_N
```

```
## # A tibble: 212 x 2
##    DESTINATION_REGION             numIMM
##    <fct>                           <int>
##  1 Abilene, Texas Area                59
##  2 Albany, New York Area            1165
##  3 Albuquerque, New Mexico Area      335
##  4 Allentown, Pennsylvania Area      728
##  5 Anchorage, Alaska Area             79
##  6 Asheville, North Carolina Area    205
##  7 Athens, Georgia Area              311
##  8 Auburn, Alabama Area               60
##  9 Augusta, Georgia Area             201
## 10 Austin, Texas Area               8401
## # ... with 202 more rows
```

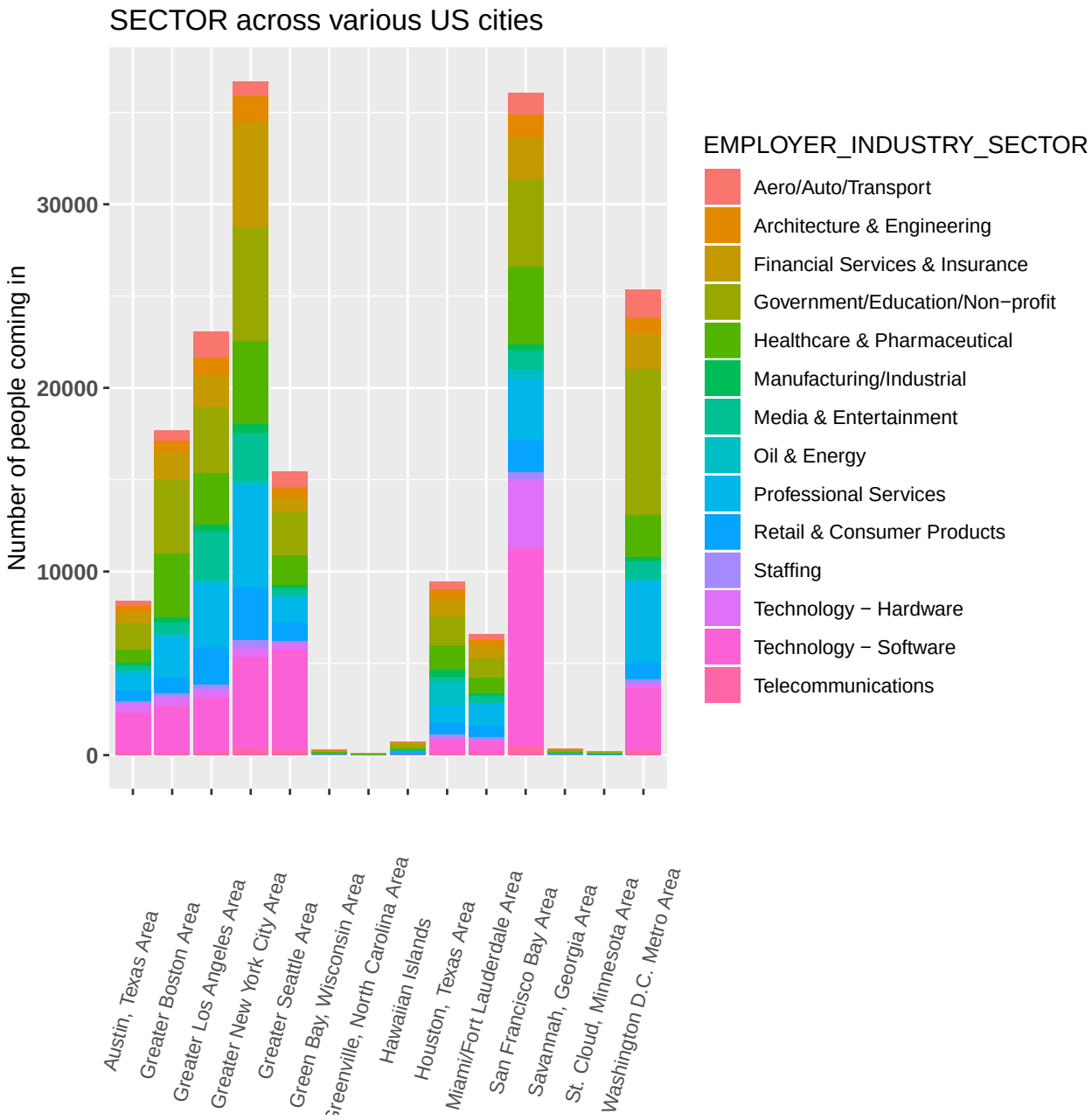
```
some_USA <- both_USA[both_USA$DESTINATION_REGION == "Greater New York City Area" | both_USA$DESTINATION_
                     | both_USA$DESTINATION_REGION == "Green Bay, Wisconsin Area" | both_USA$DESTINATION_
                     ,  ]

summary(some_USA)
```

```
##    NEW_MEM_ID     HIGHEST_DEGREE_OBTAINED    SENIORITY
##  Min.   :     4   associate:25165         Entry   :84953
##  1st Qu.:112192   bachelor :94375         Senior  :55988
##  Median :228366   doctor   :18926         Manager :15747
##  Mean   :231103   master   :41950         Training: 8991
##  3rd Qu.:348236                           Director: 7716
##  Max.   :475316                           VP      : 2776
##                                           (Other) : 4245
##                     EMPLOYER_INDUSTRY_SECTOR            POSITION_FUNCTION
##  Technology - Software          :33788       Engineering         :26427
##  Government/Education/Non-profit:33090       Education           :15830
##  Professional Services          :23385       Research            :15050
##  Healthcare & Pharmaceutical    :22185       Sales               :12712
##  Financial Services & Insurance :16524       Business Development:11655
##  Retail & Consumer Products     :11439       Operations          :11573
##  (Other)                        :40005       (Other)             :87169
##    WEEK_BEGINNING          SOURCE_COUNTRY
##  7/31/2016:  4510   Australia     :     0
##  8/14/2016:  4505   United Kingdom:     0
##  8/21/2016:  4493   United States :180416
##  8/28/2016:  4354
##  9/11/2016:  4327
##  6/5/2016 :  4289
##  (Other)  :153938
##                     SOURCE_REGION        DESTINATION_COUNTRY
##  Greater New York City Area: 15572   Australia     :     0
##  San Francisco Bay Area    : 10268   United Kingdom:     0
##  Greater Los Angeles Area  :  9648   United States :180416
##  Greater Boston Area       :  8611
##  Washington D.C. Metro Area:  7107
##  Greater Chicago Area      :  6984
##  (Other)                   :122226
##                   DESTINATION_REGION
##  Greater New York City Area:36695
```

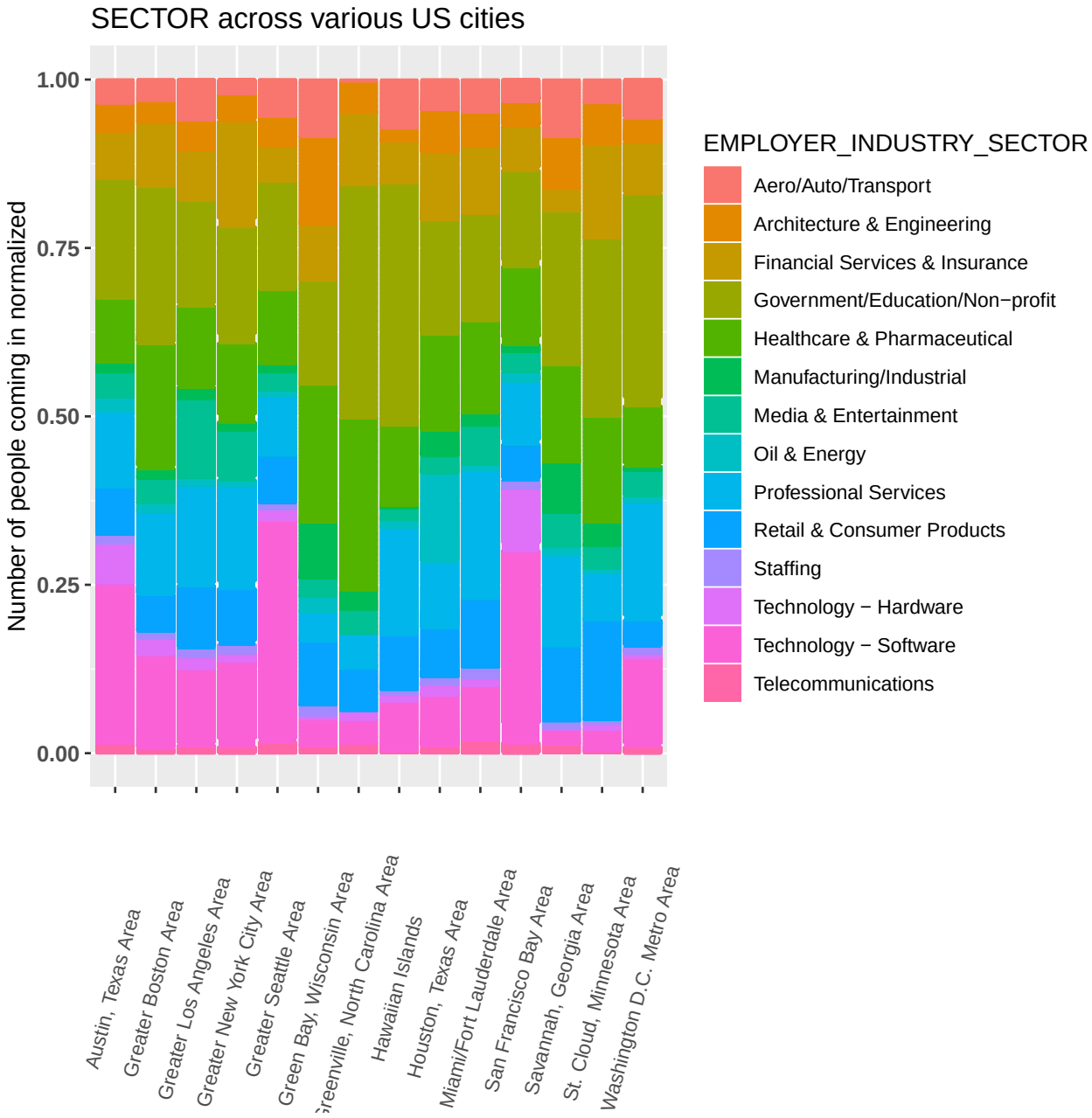
```
##  San Francisco Bay Area    :36040
##  Washington D.C. Metro Area:25335
##  Greater Los Angeles Area  :23073
##  Greater Boston Area       :17690
##  Greater Seattle Area      :15436
##  (Other)                   :26147
```

5.1 Bivariate measures of assiciation USA - sector

```
# qplot for visualization
qplot(x = DESTINATION_REGION, data = some_USA, fill = EMPLOYER_INDUSTRY_SECTOR, geom = "bar") +
theme(axis.text.x = element_text(angle=75, vjust=0.3), axis.text.y = element_text(size=10,face="bold") )
 labs(title="SECTOR across various US cities", x="", y="Number of people coming in")
```




```
# ggplot
ggplot(some_USA, aes(x=DESTINATION_REGION, y=NEW_MEM_ID, fill=EMPLOYER_INDUSTRY_SECTOR)) +
  geom_bar(aes(colour =EMPLOYER_INDUSTRY_SECTOR),stat="identity",  position = "fill") +
theme(legend.position = "right"    ,axis.text.x = element_text(angle=75, vjust=0.3), axis.text.y = eleme

   # scale_fill_brewer(palette = "Dark2") +
  # scale_fill_grey() +
  labs(title="SECTOR across various US cities", x="", y="Number of people coming in normalized")
```



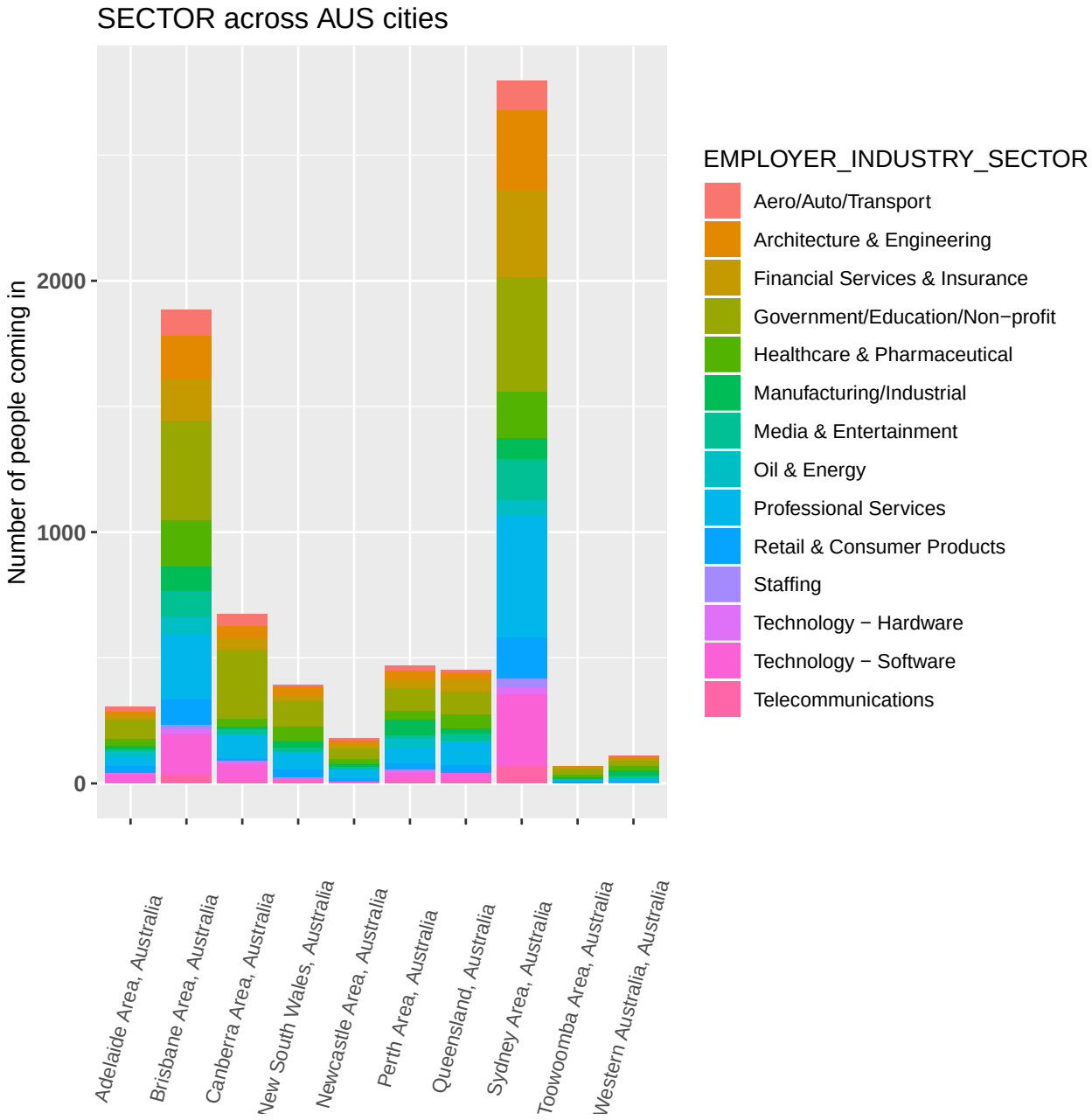

```
# cramer v   Sector X CITY OF DESTINATION (top 3)
x<- some_USA$DESTINATION_REGION
y<- some_USA$EMPLOYER_INDUSTRY_SECTOR

cv.test = function(x,y) {
  CV = sqrt(chisq.test(x, y, correct=FALSE)$statistic /
    (length(x) * (min(length(unique(x)),length(unique(y))) - 1)))
  print.noquote("Cramér V / Phi:")
  return(as.numeric(CV))
}

with(some_USA, cv.test(x, y)) # [1] Cramér V / Phi: 0.127 (less )
```

```
## Warning in chisq.test(x, y, correct = FALSE): Chi-squared approximation may
```

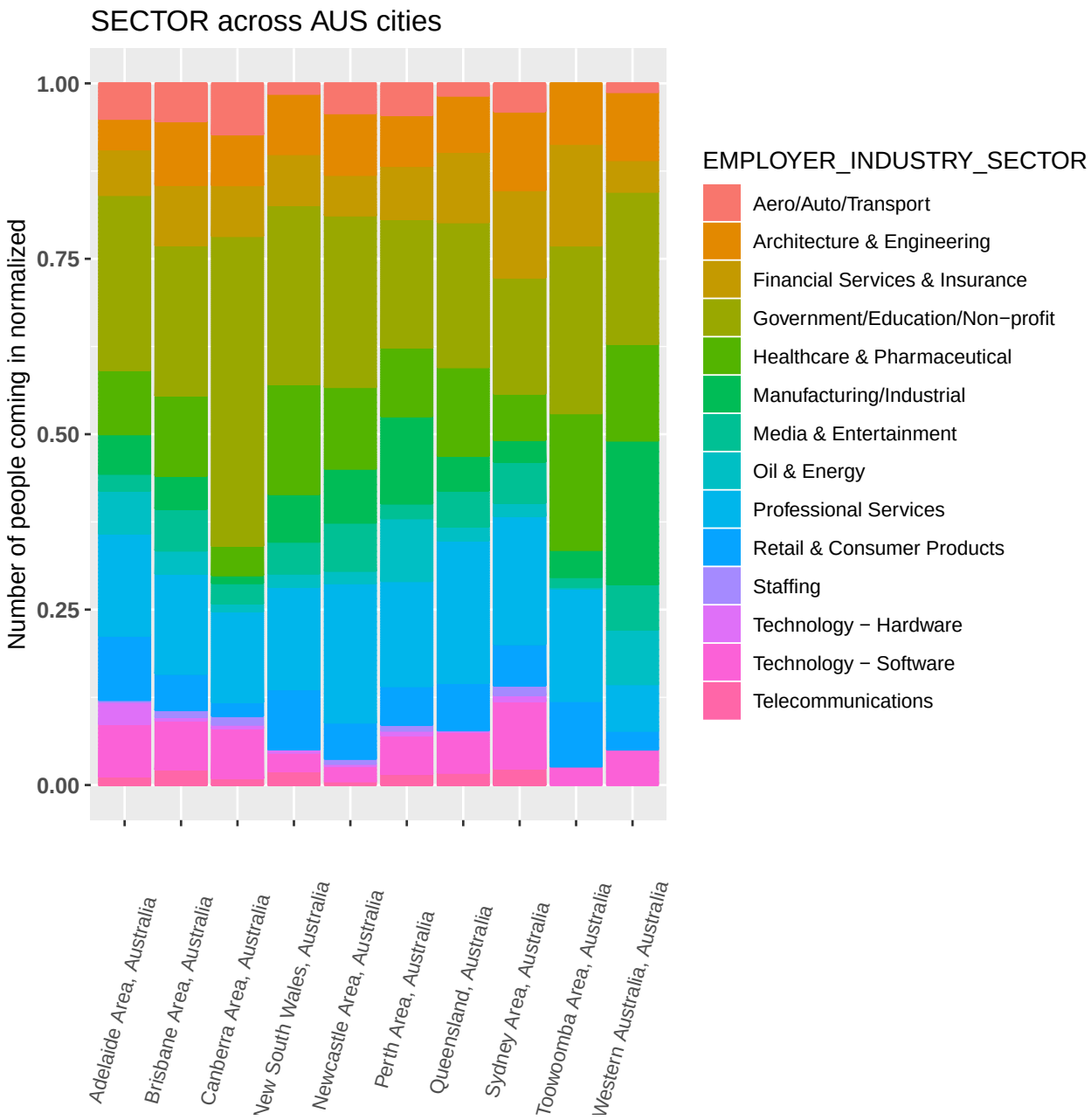
```
## be incorrect
## [1] Cramér V / Phi:
## [1] 0.1270301
```

5.2 Bivariate measures of association AUSTRALIA - sector

```
# qplot for visualization
qplot(x = DESTINATION_REGION, data = both_AUS, fill = EMPLOYER_INDUSTRY_SECTOR, geom = "bar") +
theme(axis.text.x = element_text(angle=75, vjust=0.3), axis.text.y = element_text(size=10,face="bold") )
 labs(title="SECTOR across AUS cities", x="", y="Number of people coming in")
```



```
# ggplot
ggplot(both_AUS, aes(x=DESTINATION_REGION, y=NEW_MEM_ID, fill=EMPLOYER_INDUSTRY_SECTOR)) +
  geom_bar(aes(colour =EMPLOYER_INDUSTRY_SECTOR),stat="identity",  position = "fill") +
theme(legend.position = "right"    ,axis.text.x = element_text(angle=75, vjust=0.3), axis.text.y = elem

   # scale_fill_brewer(palette = "Dark2") +
  # scale_fill_grey() +
  labs(title="SECTOR across AUS cities", x="", y="Number of people coming in normalized")
```




```
# cramer v   Sector X CITY OF DESTINATION (top 3)
x<- both_AUS$DESTINATION_REGION
y<- both_AUS$EMPLOYER_INDUSTRY_SECTOR
```

```
cv.test = function(x,y) {
  CV = sqrt(chisq.test(x, y, correct=FALSE)$statistic /
    (length(x) * (min(length(unique(x)),length(unique(y))) - 1)))
  print.noquote("Cramér V / Phi:")
  return(as.numeric(CV))
}

with(both_AUS, cv.test(x, y)) # [1] Cramér V / Phi: 0.127 (less )
```

```
## Warning in chisq.test(x, y, correct = FALSE): Chi-squared approximation may
## be incorrect
```

```
## [1] Cramér V / Phi:
```

```
## [1] 0.1074171
```

5.3 Bivariate measures of association UK - sector

construct UK internal sub-sample (for simplicity)

```
both_UK_N <- both_UK %>% group_by(DESTINATION_REGION) %>%  summarise(numIMM = n())
both_UK_N
```

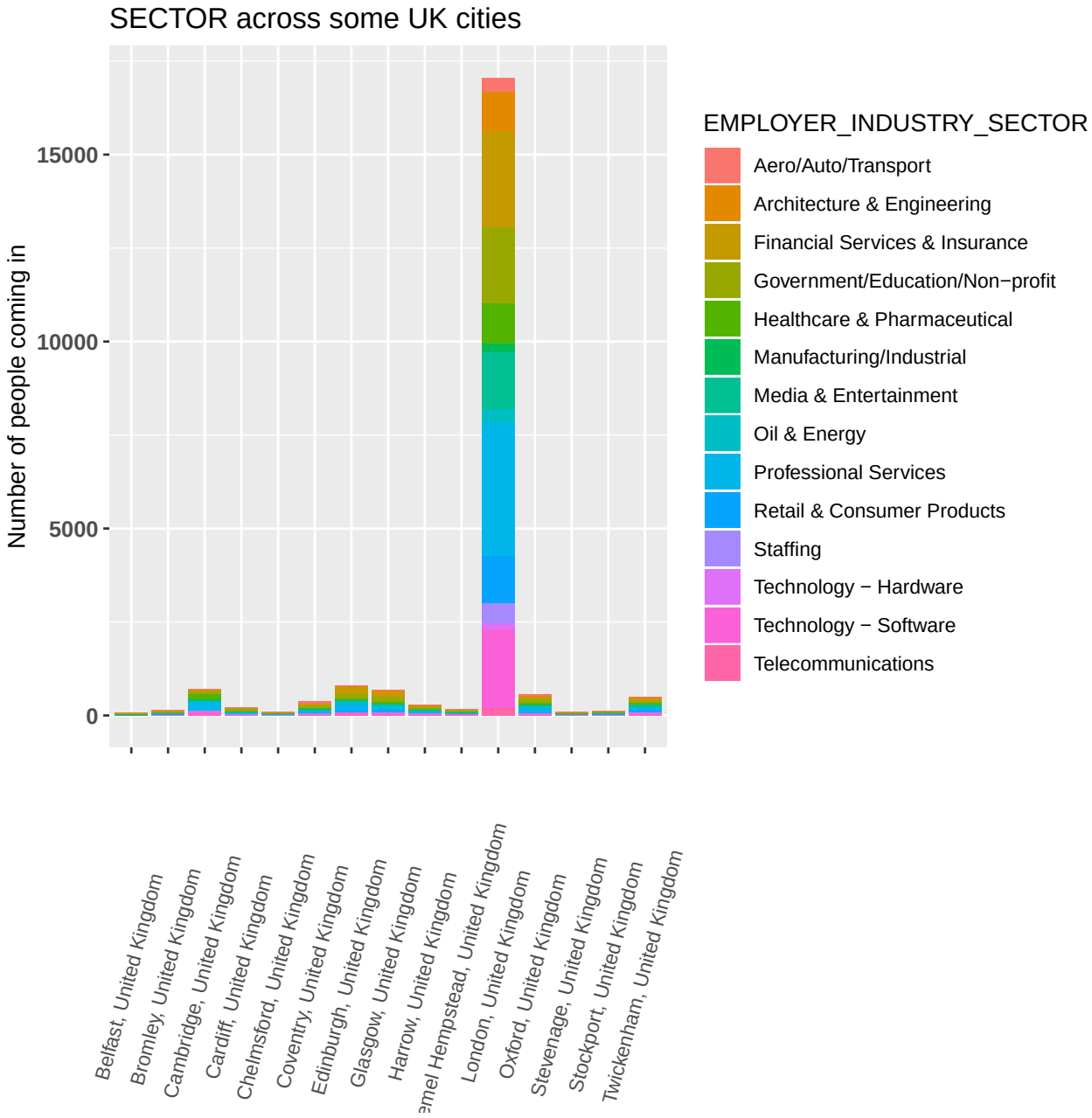
```
## # A tibble: 60 x 2
##    DESTINATION_REGION          numIMM
##    <fct>                        <int>
##  1 Bath, United Kingdom           160
##  2 Belfast, United Kingdom         66
##  3 Birmingham, United Kingdom     949
##  4 Bournemouth, United Kingdom     76
##  5 Brighton, United Kingdom       245
##  6 Bristol, United Kingdom        754
##  7 Bromley, United Kingdom        138
##  8 Cambridge, United Kingdom      719
##  9 Canterbury, United Kingdom      55
## 10 Cardiff, United Kingdom        226
## # ... with 50 more rows
```

```
some_UK <- both_UK[both_UK$DESTINATION_REGION == "London, United Kingdom" | both_UK$DESTINATION_REGION =
                   | both_UK$DESTINATION_REGION == "Bromley, United Kingdom" | both_UK$DESTINATION_REG
                   ,  ]

summary(some_UK)
```

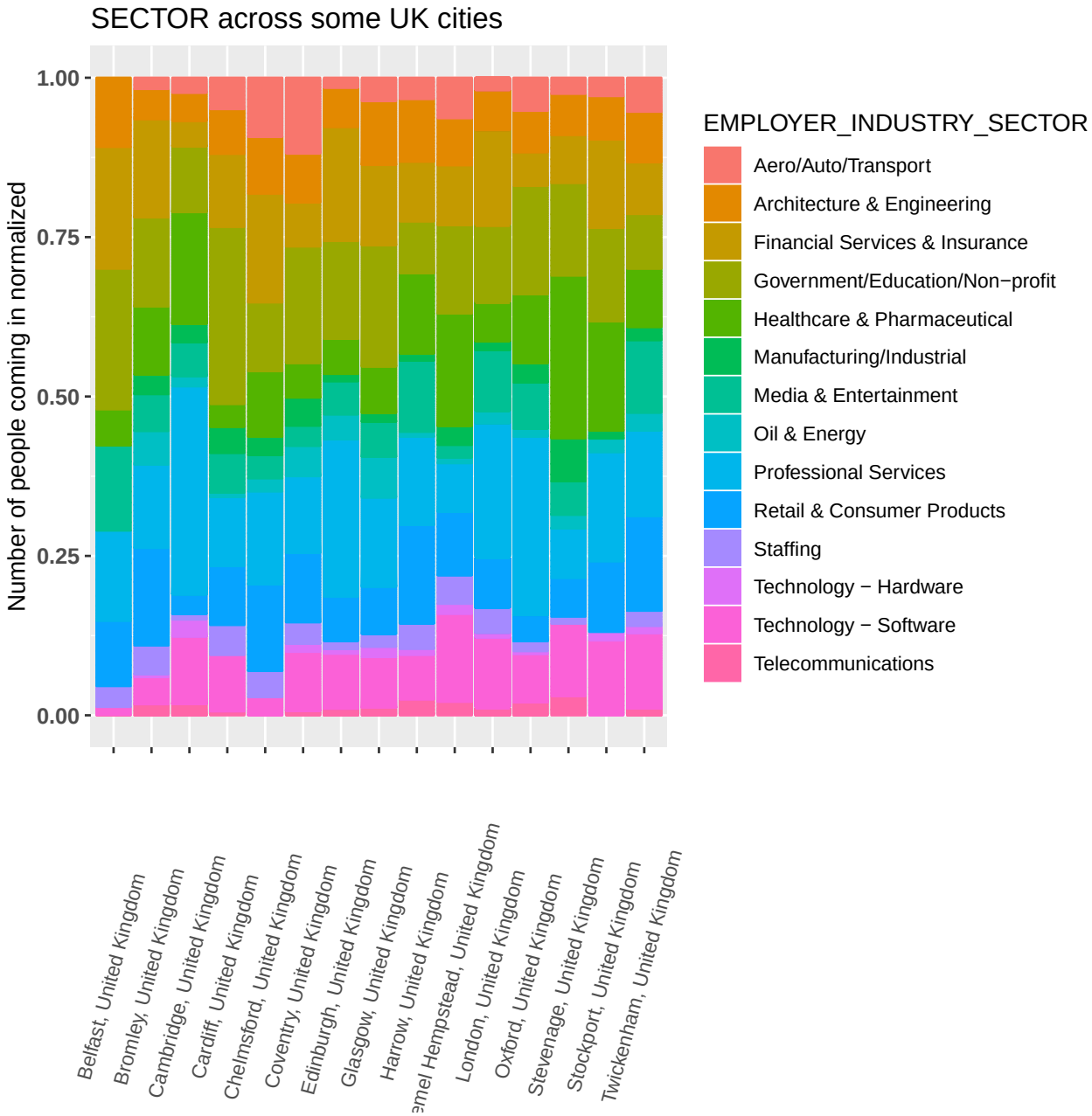
```
##    NEW_MEM_ID     HIGHEST_DEGREE_OBTAINED    SENIORITY
##  Min.   :    18   associate:  677         Entry   :8439
##  1st Qu.:163777   bachelor :13336         Senior  :8351
##  Median :286759   doctor   : 2093         Manager :2520
##  Mean   :271096   master   : 5773         Training:1057
##  3rd Qu.:385509                           Director: 754
##  Max.   :475315                           VP      : 254
##                                           (Other) : 504
##                      EMPLOYER_INDUSTRY_SECTOR            POSITION_FUNCTION
##  Professional Services          :4481        Business Development: 1927
##  Financial Services & Insurance :3093        Sales               : 1773
```

```
##  Government/Education/Non-profit:2757        Engineering         : 1572
##  Technology - Software          :2586        Finance             : 1539
##  Media & Entertainment          :1819        Operations          : 1464
##  Retail & Consumer Products     :1598        Research            : 1462
##  (Other)                        :5545        (Other)             :12142
##    WEEK_BEGINNING         SOURCE_COUNTRY
##  10/2/2016:  673   Australia     :    0
##  9/18/2016:  616   United Kingdom:21879
##  9/11/2016:  603   United States :    0
##  9/25/2016:  583
##  9/4/2016 :  575
##  10/9/2016:  565
##  (Other)  :18264
##                               SOURCE_REGION       DESTINATION_COUNTRY
##  London, United Kingdom              : 2192   Australia     :    0
##  Manchester, United Kingdom          :  816   United Kingdom:21879
##  Oxford, United Kingdom              :  767   United States :    0
##  Reading, United Kingdom             :  719
##  Twickenham, United Kingdom          :  712
##  Kingston upon Thames, United Kingdom:  699
##  (Other)                             :15974
##                   DESTINATION_REGION
##  London, United Kingdom    :17055
##  Edinburgh, United Kingdom :  803
##  Cambridge, United Kingdom :  719
##  Glasgow, United Kingdom   :  682
##  Oxford, United Kingdom    :  554
##  Twickenham, United Kingdom:  492
##  (Other)                   : 1574
```

```
# qplot for visualization
qplot(x = DESTINATION_REGION, data = some_UK, fill = EMPLOYER_INDUSTRY_SECTOR, geom = "bar") +
theme(axis.text.x = element_text(angle=75, vjust=0.3), axis.text.y = element_text(size=10,face="bold") )
 labs(title="SECTOR across some UK cities", x="", y="Number of people coming in")
```




```
# ggplot
ggplot(some_UK, aes(x=DESTINATION_REGION, y=NEW_MEM_ID, fill=EMPLOYER_INDUSTRY_SECTOR)) +
  geom_bar(aes(colour =EMPLOYER_INDUSTRY_SECTOR),stat="identity",  position = "fill") +
theme(legend.position = "right"    ,axis.text.x = element_text(angle=75, vjust=0.3), axis.text.y = eleme

   # scale_fill_brewer(palette = "Dark2") +
  # scale_fill_grey() +
  labs(title="SECTOR across some UK cities", x="", y="Number of people coming in normalized")
```




```
# cramer v   Sector X CITY OF DESTINATION (top 3)
x<- some_UK$DESTINATION_REGION
y<- some_UK$EMPLOYER_INDUSTRY_SECTOR

cv.test = function(x,y) {
  CV = sqrt(chisq.test(x, y, correct=FALSE)$statistic /
    (length(x) * (min(length(unique(x)),length(unique(y))) - 1)))
  print.noquote("Cramér V / Phi:")
  return(as.numeric(CV))
}

with(some_UK, cv.test(x, y)) # [1] Cramér V / Phi: 0.127 (less )
```

```
## Warning in chisq.test(x, y, correct = FALSE): Chi-squared approximation may
```

```
## be incorrect
## [1] Cramér V / Phi:
## [1] 0.07060599
```