DATA605_Project2

Harpreet Shoker

Contents

Environment setup

```
library(stringr)
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.2
```

```
##
## 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(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 3.4.3
```

```
library(knitr)
library(tibble)
```

```
## Warning: package 'tibble' was built under R version 3.4.3
```

```
library(ggplot2)
```

For 1st data set i have picked World population data from discussion by steven which has data set showing the population of the world's countries from 1980 to 2010

1 Data - World Population

Loading Data set for World population

```
countryData <- read.csv(file="https://raw.githubusercontent.com/Harpreet1984/DATA607/master/populationby
countryDataTidy <-  countryData %>% gather ("Year", "population", 2:32) #using gather()
```

```
## Warning: attributes are not identical across measure variables;
## they will be dropped
```

```
head(countryDataTidy)
```

```
##                             X  Year population
## 1             North America X1980  320.27638
## 2                   Bermuda X1980    0.05473
## 3                    Canada X1980    24.5933
## 4                 Greenland X1980    0.05021
## 5                    Mexico X1980   68.34748
## 6 Saint Pierre and Miquelon X1980    0.00599
```

Tidying Data

Removing all the rows that have NA and – as populations group data based on country. Calculate percentage change for the entire duration for each country.

```
countryDataFilteredGrouped <- countryDataTidy %>% filter (!population %in% c("--", NA) ) %>% group_by(X)
countryDataFilteredGrouped
```

```
## # A tibble: 229 x 2
##    X                   populationChange
##    <fct>                          <dbl>
##  1 Afghanistan                    0.936
##  2 Africa                         1.12
##  3 Albania                        0.118
##  4 Algeria                        0.839
##  5 American Samoa                 1.05
##  6 Angola                         0.938
##  7 Antigua and Barbuda            0.265
##  8 Argentina                      0.457
##  9 Armenia                       -0.122
## 10 Aruba                          0.749
## # ... with 219 more rows
```

Data Analaysis

Find the country that has the maximum population growth during this duration

```
countryWithMaxPopulationGrowth <-  countryDataFilteredGrouped %>% filter(populationChange ==  max(popula
countryWithMaxPopulationGrowth
```

```
## # A tibble: 1 x 2
##   X                   populationChange
##   <fct>                          <dbl>
## 1 United Arab Emirates            3.97
```

So here we found that the country that has maximum population growth during the entire duration is United Arab emirates

Find the country that has the minimum population growth during this duration

```
countryWithMinPopulationGrowth <-  countryDataFilteredGrouped %>% filter(populationChange ==  min(popula
countryWithMinPopulationGrowth
```

```
## # A tibble: 1 x 2
##   X          populationChange
##   <fct>                 <dbl>
## 1 Montserrat           -0.565
```

From the above result we can conclude that population percentage has decreased for montserrat by 5.64 %

2 Data - Time Spent (Male vs Female)

For 2nd data set Time spent as discussed by Nicholas this dataset on time use by gender and by country has variables include eating, sleeping, employment, travel, school, study, walking the dog, etc

Laoding Data

```
myurl <- "https://raw.githubusercontent.com/Harpreet1984/DATA607/master/TimeUse%20(1).csv"
time_info <- read.csv(myurl, header= TRUE,sep=",",stringsAsFactors=FALSE)
kable(time_info)
```

SEX	GEO.ACL00	Total	Personal.care	Sleep	Eating	Other.and.or.unspecified
Males	Belgium	24:00	10:45	8:15	1:49	0:42
Males	Bulgaria	24:00	11:54	9:08	2:07	0:39
Males	Germany (including former GDR from 1991)	24:00	10:40	8:08	1:43	0:49
Males	Estonia	24:00	10:35	8:24	1:19	0:52
Males	Spain	24:00	11:11	8:36	1:47	0:48
Males	France	24:00	11:44	8:45	2:18	0:41
Males	Italy	24:00	11:16	8:17	1:57	1:02
Males	Latvia	24:00	10:46	8:35	1:33	0:37
Males	Lithuania	24:00	10:53	8:28	1:32	0:53
Males	Poland	24:00	10:44	8:21	1:33	0:50
Males	Slovenia	24:00	10:31	8:18	1:33	0:40
Males	Finland	24:00	10:23	8:22	1:23	0:38
Males	United Kingdom	24:00	10:22	8:18	1:24	0:41
Males	Norway	24:00	10:06	7:56	1:25	0:45
Females	Belgium	24:00	11:11	8:34	1:50	0:47
Females	Bulgaria	24:00	11:38	9:07	1:55	0:36
Females	Germany (including former GDR from 1991)	24:00	10:58	8:15	1:46	0:56
Females	Estonia	24:00	10:30	8:26	1:12	0:53
Females	Spain	24:00	11:05	8:32	1:44	0:49
Females	France	24:00	11:53	8:55	2:11	0:46
Females	Italy	24:00	11:12	8:19	1:52	1:01
Females	Latvia	24:00	10:53	8:44	1:26	0:43
Females	Lithuania	24:00	10:56	8:35	1:26	0:56
Females	Poland	24:00	11:03	8:35	1:34	0:54
Females	Slovenia	24:00	10:32	8:25	1:26	0:41
Females	Finland	24:00	10:38	8:32	1:19	0:47
Females	United Kingdom	24:00	10:43	8:27	1:26	0:50
Females	Norway	24:00	10:27	8:10	1:20	0:56

Tidying and analysis

```
timedata <- time_info
library(tibble)
timedata <- as_data_frame(timedata)
timedata <- timedata %>% rename(Country = GEO.ACL00)
```

Converting the time spent on personal care into minutes by removing the colon and then calculating mean of total time(personal care)

```
timedata_PC <- timedata %>%
               separate ("Personal.care" , c("PC_Min", "PC_sec"), sep=":")
timedata_PC <-  timedata_PC %>%
                mutate(PC_TotalSec= (as.numeric(PC_Min) * 60) + as.numeric(PC_sec))
timedata_PC <- timedata_PC %>%
               group_by(SEX) %>% summarise(mean= mean(PC_TotalSec))
kable(timedata_PC)
```

SEX	mean
Females	658.5000
Males	650.7143

Analysis:- Here from the above results we see that Females spent more time in Personal care compared to Males

Converting the time on eating in minutes by removing the colon and calculating mean of total time(eating)

```
timedata_Eat <- timedata %>%
              separate ("Eating" , c("E_Min", "E_sec"), sep=":")
timedata_Eat <-  timedata_Eat %>%
               mutate(E_TotalSec= (as.numeric(E_Min) * 60) + as.numeric(E_sec))
timedata_Eat <- timedata_Eat %>%
              group_by(Country,SEX) %>% summarise(mean= mean(E_TotalSec))
timedata_Eat
```

```
## # A tibble: 28 x 3
## # Groups:   Country [?]
##    Country  SEX      mean
##    <chr>    <chr>   <dbl>
##  1 Belgium  Females 110
##  2 Belgium  Males   109
##  3 Bulgaria Females 115
##  4 Bulgaria Males   127
##  5 Estonia  Females  72.0
##  6 Estonia  Males    79.0
##  7 Finland  Females  79.0
##  8 Finland  Males    83.0
##  9 France   Females 131
## 10 France   Males   138
## # ... with 18 more rows
```

Analysis :- From the above subset we can infer that Males spend more time in eating compared to females for most of the countries.

```
timedata_Sleep <- timedata %>%
              separate ("Sleep" , c("S_Min", "S_sec"), sep=":")
timedata_Sleep <-  timedata_Sleep %>%
               mutate(S_TotalSec= (as.numeric(S_Min) * 60) + as.numeric(S_sec))
timedata_Sleep <- timedata_Sleep %>%
              group_by(Country,SEX) %>% summarise(mean= mean(S_TotalSec))
timedata_Sleep
```

```
## # A tibble: 28 x 3
## # Groups:   Country [?]
##    Country  SEX      mean
##    <chr>    <chr>   <dbl>
##  1 Belgium  Females   514
##  2 Belgium  Males     495
##  3 Bulgaria Females   547
##  4 Bulgaria Males     548
##  5 Estonia  Females   506
##  6 Estonia  Males     504
```

```
##  7 Finland  Females   512
##  8 Finland  Males     502
##  9 France   Females   535
## 10 France   Males     525
## # ... with 18 more rows
```

Analysis :- From the above subset we can infer that Males spend less time in sleeping compared to females for most of the countries.

3. Data -sales-tax credits from the government of Canada

Loading data

```
craCreditBenefit <- read.csv(file="https://raw.githubusercontent.com/Harpreet1984/DATA607/master/CRA_Cr
kable(craCreditBenefit)
```

Province	GT_freq	GT_amount	X5K_freq	X5K_amount	X5K_10K_freq	X5K_10K_am
Newfoundland_and_Labrador	160300	$65,756	26410	$7,044	16150	$5,278
Prince_Edward_Island	45850	$18,648	8000	$2,115	4220	$1,363
Nova_Scotia	302210	$122,873	59800	$17,694	38340	$13,202
New_Brunswick	249780	$102,431	43160	$11,986	29500	$9,489
Quebec	2698620	$1,082,380	495090	$132,486	288830	$101,067
Ontario	3982840	$1,647,108	852320	$249,084	426320	$151,182
Manitoba	385370	$162,874	101660	$35,032	35830	$12,788
Saskatchewan	291210	$125,501	72980	$26,349	23750	$8,961
Alberta	921590	$384,471	208080	$65,556	80350	$28,722
British_Columbia	1390640	$561,578	304940	$89,447	124170	$42,769
Northwest_Territories	10740	$4,478	2890	$998	1420	$541
Yukon	9110	$3,788	1600	$442	730	$259
Nunavut	8530	$4,383	1680	$567	1740	$727
Outside_Canada	920	$436	350	$159	120	$44

tidying data set

```
craCreditBenefitAmountAnalysis <- craCreditBenefit %>% select(Province, X5K_amount, X5K_10K_amount, X10

craCreditBenefitAmountAnalysisTidy <- craCreditBenefitAmountAnalysis %>% gather("Category", "Amount",  1
```

```
## Warning: attributes are not identical across measure variables;
## they will be dropped
```

```
craCreditBenefitAmountAnalysisTidy$Amount = gsub(",", "",craCreditBenefitAmountAnalysisTidy$Amount)
craCreditBenefitAmountAnalysisTidy$Amount = gsub("\\$", "",craCreditBenefitAmountAnalysisTidy$Amount)
craCreditBenefitAmountAnalysisTidy$Amount = as.numeric (gsub("^$", "0",craCreditBenefitAmountAnalysisTi
```

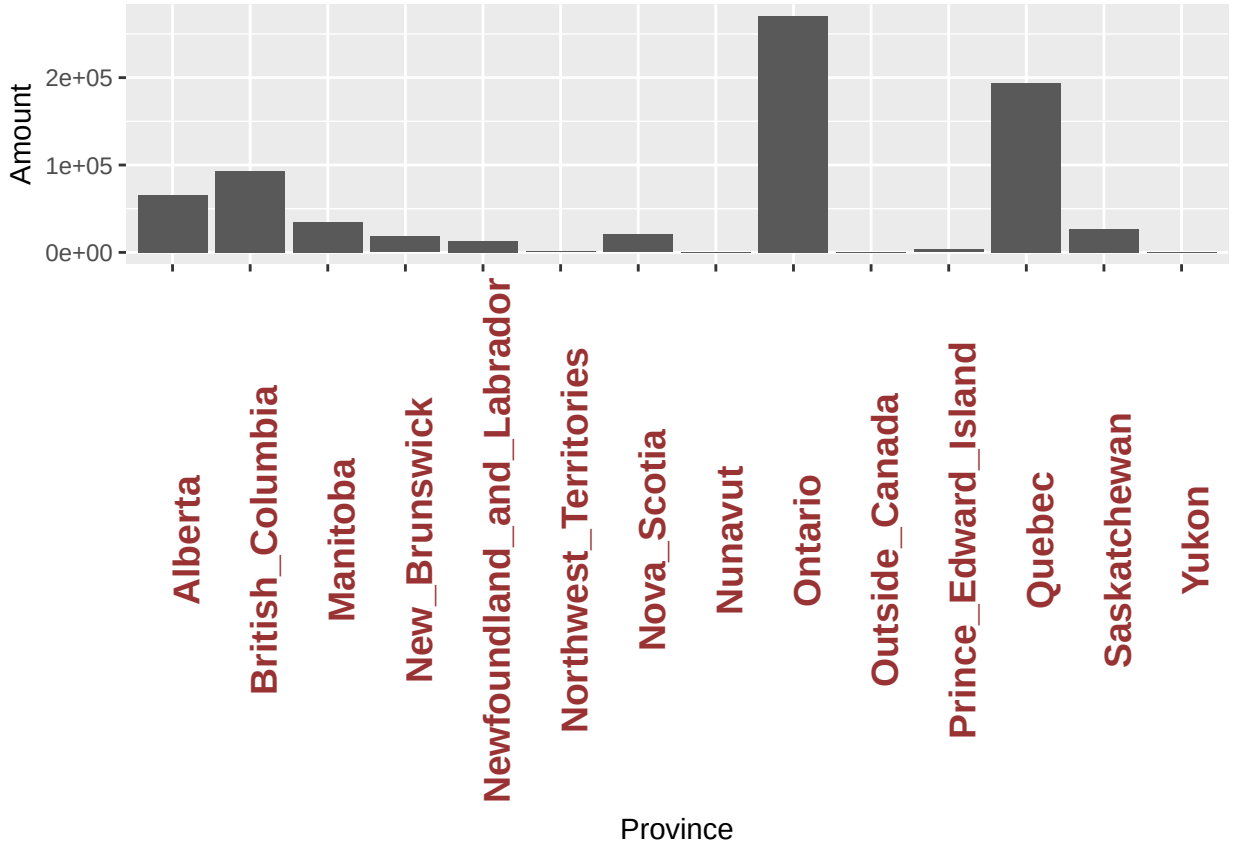
Analyzing data set

Here we are calculating Maximum Credit Benefit Per province

```
maxCreditBenefitByProvince <- craCreditBenefitAmountAnalysisTidy %>% group_by(Province) %>% summarise(A
kable(maxCreditBenefitByProvince)
```

Province	Amount
Alberta	65556
British_Columbia	93329
Manitoba	35032
New_Brunswick	18130
Newfoundland_and_Labrador	13023
Northwest_Territories	998
Nova_Scotia	20540
Nunavut	727
Ontario	270404
Outside_Canada	159
Prince_Edward_Island	3332
Quebec	194299
Saskatchewan	26349
Yukon	710

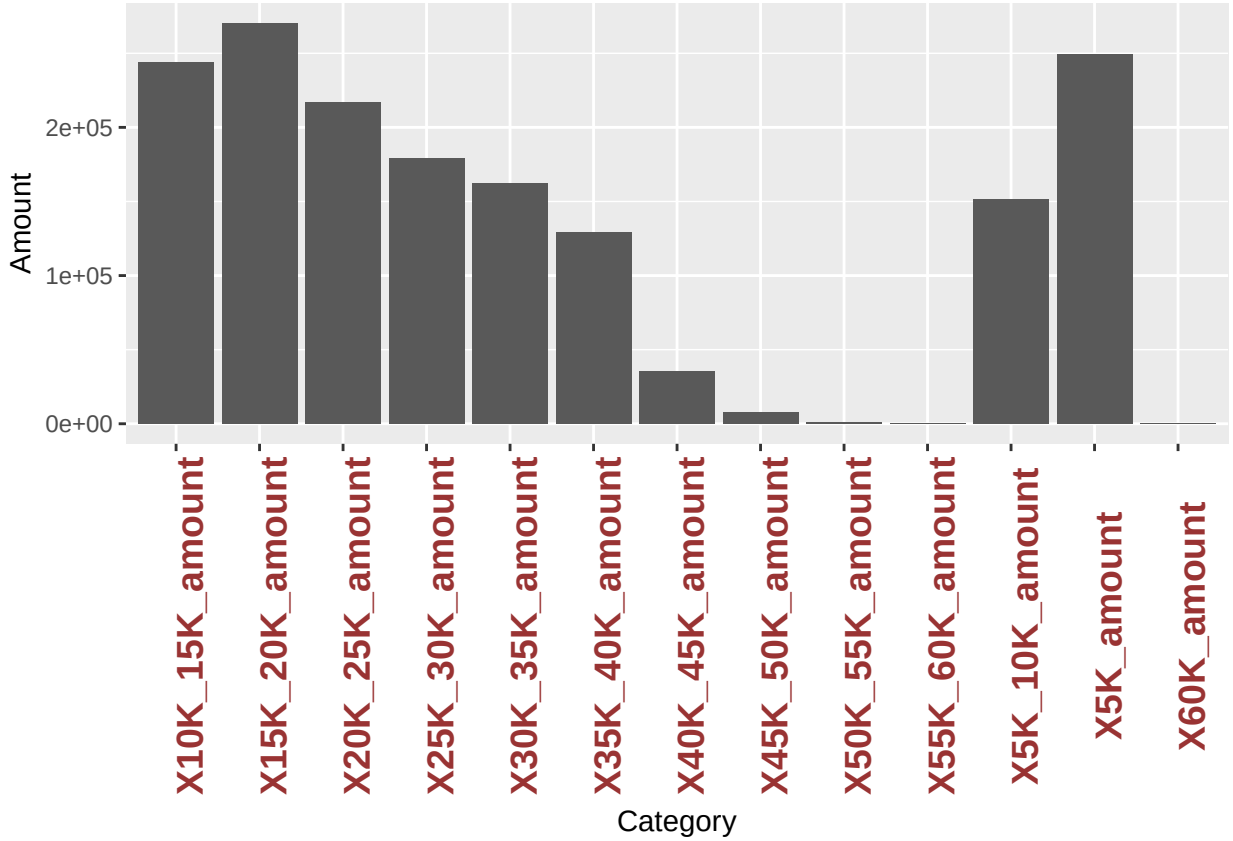
```
myggplot <- ggplot(maxCreditBenefitByProvince, aes(x = Province, y = Amount)) +
  geom_bar(stat = "identity")
myggplot <- myggplot + theme(axis.text.x = element_text(face="bold", color="#993333",
                           size=14, angle=90))
 myggplot
```




```
maxCreditBenefitByCategory <- craCreditBenefitAmountAnalysisTidy %>% group_by(Category) %>% summarise(Am
kable(maxCreditBenefitByCategory)
```

Category	Amount
X10K_15K_amount	243898
X15K_20K_amount	270404
X20K_25K_amount	217190
X25K_30K_amount	179196
X30K_35K_amount	162303
X35K_40K_amount	129493
X40K_45K_amount	35557
X45K_50K_amount	7656
X50K_55K_amount	1028
X55K_60K_amount	97
X5K_10K_amount	151182
X5K_amount	249084
X60K_amount	18

```
library(ggplot2)
myggplot1 <- ggplot(maxCreditBenefitByCategory, aes(x = Category, y = Amount)) +
  geom_bar(stat = "identity")
myggplot1 <- myggplot1 + theme(axis.text.x = element_text(face="bold", color="#993333",
                          size=14, angle=90))
 myggplot1
```




```
craCreditBenefitFreqAnalysis <- craCreditBenefit %>% select(Province, X5K_freq, X5K_10K_freq, X10K_15K_

craCreditBenefitFreqAnalysisTidy <- craCreditBenefitFreqAnalysis %>% gather("Category", "Amount",  2:14
craCreditBenefitFreqAnalysisTidy$Amount = gsub(",", "",craCreditBenefitFreqAnalysisTidy$Amount)
```

```
craCreditBenefitFreqAnalysisTidy$Amount = as.numeric (gsub("^$", "0",craCreditBenefitFreqAnalysisTidy$Am
```

```
maxCreditFreqByProvince <- craCreditBenefitFreqAnalysisTidy %>% group_by(Province) %>% summarise(max(Am
kable(maxCreditFreqByProvince)
```

Province	max(Amount)
Alberta	208080
British_Columbia	304940
Manitoba	101660
New_Brunswick	NA
Newfoundland_and_Labrador	NA
Northwest_Territories	NA
Nova_Scotia	NA
Nunavut	NA
Ontario	852320
Outside_Canada	NA
Prince_Edward_Island	NA
Quebec	495090
Saskatchewan	72980
Yukon	NA

```
maxCreditFreqByCategory <- craCreditBenefitFreqAnalysisTidy %>% group_by(Category) %>% summarise(max(Am
kable(maxCreditFreqByCategory)
```

Category	max(Amount)
X10K_15K_freq	573760
X15K_20K_freq	563100
X20K_25K_freq	440530
X25K_30K_freq	362380
X30K_35K_freq	328470
X35K_40K_freq	308840
X40K_45K_freq	NA
X45K_50K_freq	NA
X50K_55K_freq	NA
X55K_60K_freq	NA
X5K_10K_freq	426320
X5K_freq	852320
X60K_freq	NA

Based on the above graphs, Ontario and montreal province got the maxmimum benefit credits and have highest frequencies. This makes sense as these two provinces are have highest number of working professionals.

Based on the general normal as the income bracket goes up the tax credit benefits decrease. This is confirmed with the graphs, there are substantial drops in the tax credit as the category goes above 40k.