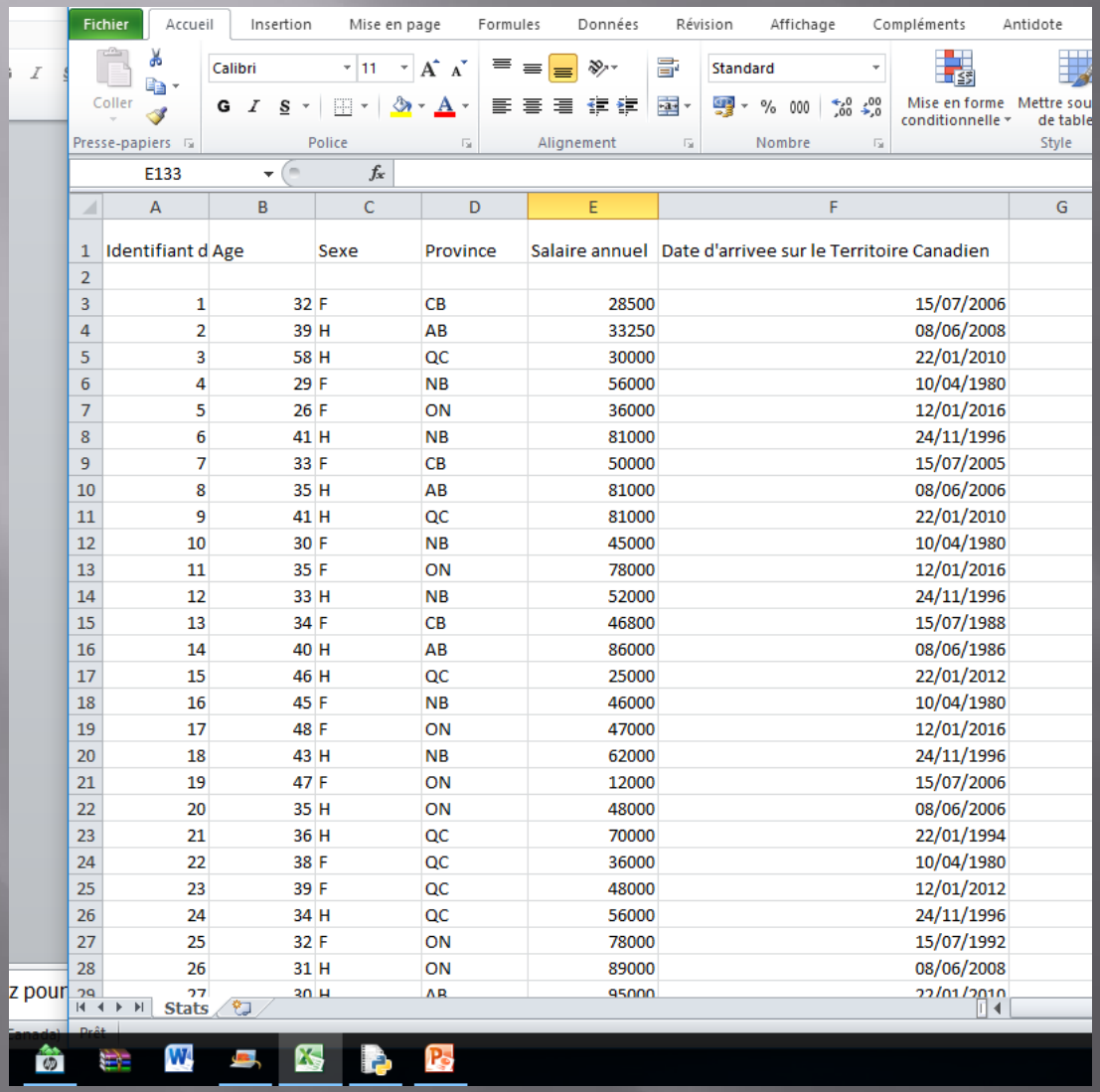
ANALYSE DE DONNÉES DES IMMIGRANTS REÇUS DEPUIS 1980

Analyses faites en langage Python par
Joel Sandé

Avertissement :

- Les chiffres qui suivent dans cete présentation sont completement FAUX.
- J'ai voulu juste mettre en évidence, à partir d'un cas pratique, la puissance d,analyse du Langage Python et sa capacité à nous générer des résultats sous forme visuelle.
- Je laisserai un lien pour télécharger le fichier csv d'òu les données sont tirées. Le code vous sera accessible sur demande à sandejoel@yahoo.ca (Gratuit si vous êtes un étudiant qui suit mon cours, ou un réprésentant du gouvernement).

Stats.csv



	A	B	C	D	E	F	G
1	Identifiant d	Age	Sexe	Province	Salaire annuel	Date d'arrivee sur le Territoire Canadien	
2							
3	1	32	F	CB	28500	15/07/2006	
4	2	39	H	AB	33250	08/06/2008	
5	3	58	H	QC	30000	22/01/2010	
6	4	29	F	NB	56000	10/04/1980	
7	5	26	F	ON	36000	12/01/2016	
8	6	41	H	NB	81000	24/11/1996	
9	7	33	F	CB	50000	15/07/2005	
10	8	35	H	AB	81000	08/06/2006	
11	9	41	H	QC	81000	22/01/2010	
12	10	30	F	NB	45000	10/04/1980	
13	11	35	F	ON	78000	12/01/2016	
14	12	33	H	NB	52000	24/11/1996	
15	13	34	F	CB	46800	15/07/1988	
16	14	40	H	AB	86000	08/06/1986	
17	15	46	H	QC	25000	22/01/2012	
18	16	45	F	NB	46000	10/04/1980	
19	17	48	F	ON	47000	12/01/2016	
20	18	43	H	NB	62000	24/11/1996	
21	19	47	F	ON	12000	15/07/2006	
22	20	35	H	ON	48000	08/06/2006	
23	21	36	H	QC	70000	22/01/1994	
24	22	38	F	QC	36000	10/04/1980	
25	23	39	F	QC	48000	12/01/2012	
26	24	34	H	QC	56000	24/11/1996	
27	25	32	F	ON	78000	15/07/1992	
28	26	31	H	ON	89000	08/06/2008	

Voici le contenu du fichier csv utilisé pour faire ces analyses

Plan de Présentation

- Introduction
- Méthode d'analyse
- Petites-Requêtes
  - Nombre d'enregistrements par province pour une année donnée
  - Nombre de Femmes qui ont immigrés depuis 2002 dans un province donnée, dont le salaire annuel actuel est supérieur ou égal à 45000 $
- Grosses-Requêtes
  - Statistiques des Nombres totaux d'enregistrement par province de 1980 à 2016
  - Statistiques pour toutes les provinces de la 2e petite requête.
  - Personnalisation
- Conclusion

Introduction

- Le Canada est un pays qui s'est bati sur sa diversité culturelle depuis les années 80.
- Pas mal d'immmigrants à la recherche d'un nouveau cadre géo-politiques immigrent au Canada.
- Je me suis donné le mandat de faire une analyse complète des données enregistrées de ces migrants.
- Je vous remercie de me suivre tout le long de mon analyse dont le code source est réalisé en language Python.

Méthodologie d'analyse

- Par habitude, J'aime bien me faire de petites requêtes (Fonctions) d'échauffement : C'est donc par celle-ci que nous allons commencer.
- En générale, lorsque ces requêtes sont établies, pour la suite, lorsqu'on a à faire avec de plus grosses requêtes, il suffit d'aller les rechercher une à une ou même les combiner pour se faciliter la tache lors d'une grosse requête.
- Ça devient un jeux d'enfant, et le code est plus facile à maintenir
- Et c'est parti …

Statistiques des Nombres totaux d'enregistrement par province de 1980 à 2016

Petites Requêtes

1) Nombre d'enregistrements pour une province donnée pour une année donnée

```
#=============================================================
#    Nombre_Enregistrements_Province_Annee(province, annee)
#=============================================================

def Nombre_Enregistrements_Province_Annee(province, annee):
    with open('Stats.csv','r') as csv_file:
        csv_reader = csv.reader(csv_file, delimiter=';')
        nombre = 0;
        for row in csv_reader:
            if row[3] == province and str(annee) in row[5]:
                nombre += 1

    csv_file.close()
    print nombre
    return nombre
```

Nombre_Enregistrements_Province_Annee('NB', 1996)

La réponse pour cette requête est 3

Petites Requêtes

En préparation d'une grosse requête qui s'étend sur toutes les années d'enregistrement, nous allons nous crérer une petite fonction qui stock dans un tableau tous les enregistrements pour une province donnée de l'année 1980 à 2016

```
#=============================================================================================
#   Enregistrements_Province(province, annee)
#=============================================================================================

def Enregistrements_Province(province, annee):
    Enregistrements_pro = []

    for ann in annee:
        Enregistrements_pro.append(Nombre_Enregistrements_Province_Annee(province, ann))
    return Enregistrements_pro
```

Petites Requêtes

```
#=============================================================================================================
#    Enregistrements_Province(province, annee)
#=============================================================================================================

def Enregistrements_Province_sexe_salaire(province, annee, sexe, salaire):
    Enregistrements_pro = []

    for ann in annee:
        Enregistrements_pro.append(Nombre_Enregistrements_sexe_Province_Annee(province, ann, sexe, salaire))
    return Enregistrements_pro
```

Petites requêtes

La 2eme partie de cette requette grosse requête revient à identifier les provinces.

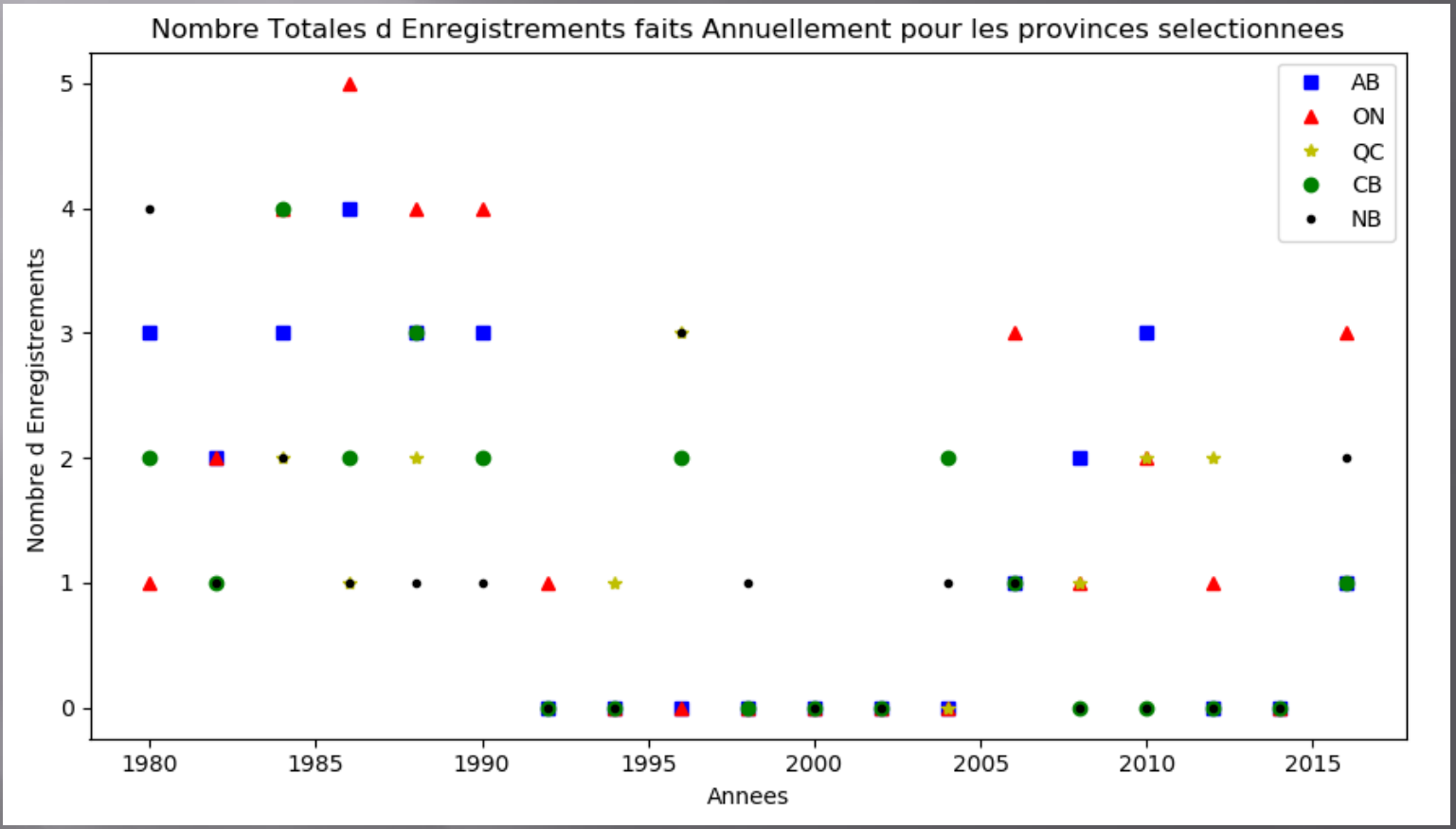
```
#=============================================================================================================================
#   Toutes_Les_Enregistrements()
#=============================================================================================================================

def Toutes_Les_Enregistrements():
    AB = []
    ON = []
    QC = []
    CB = []
    NB = []

    provinces = ['AB', 'ON', 'QC', 'CB', 'NB']
    annees = [1980, 1982, 1984, 1986, 1988, 1990, 1992, 1994, 1996, 1998, 2000, 2002, 2004, 2006, 2008, 2010, 2012, 2014, 2016]

    for prov in provinces:
        if prov == 'AB':
           AB = Enregistrements_Province('AB', annees)
        elif prov == 'ON':
           ON = Enregistrements_Province('ON', annees)
        elif prov == 'QC':
           QC = Enregistrements_Province('QC', annees)
        elif prov == 'CB':
           CB = Enregistrements_Province('CB', annees)
        elif prov == 'NB':
           NB = Enregistrements_Province('NB', annees)

    return AB, ON, QC, CB, NB
```

Nous y reviendrons dans Grosses requêtes

Grosse requête

```
#=============================================================================================================================================
#   Statistiques_des_Enregistrements()
#=============================================================================================================================================

def Statistiques_des_Enregistrements():
    AB, ON, QC, CB, NB = Toutes_Les_Enregistrements()
    annees = [1980, 1982, 1984, 1986, 1988, 1990, 1992, 1994, 1996, 1998, 2000, 2002, 2004, 2006, 2008, 2010, 2012, 2014, 2016]

    #print('annees = '+str(len(annees)))
    #print('AB = '+str(len(AB)))
    #print('ON = '+str(len(ON)))
    #print('QC = '+str(len(QC)))
    #print('CB = '+str(len(CB)))
    #print('NB = '+str(len(NB)))

    plt.plot(annees, AB, 'bs', label='AB')
    plt.plot(annees, ON, 'r^', label='ON')
    plt.plot(annees, QC, 'y*', label='QC')
    plt.plot(annees, CB, 'go', label='CB')
    plt.plot(annees, NB, 'k.', label='NB')

    #-----------------------------------
    plt.title('Nombre Totales d Enregistrements faits Annuellement pour les provinces selectionnees')
    plt.ylabel('Nombre d Enregistrements ')
    plt.xlabel('Annees')

    plt.legend()
    plt.show()
```

Visualisation



Statistiques du Nombre de femmes qui ont immigré depuis 1980 et dont la

Petite requête

1) Nombre de femmes enregistrées depuis 1996 dans la province du Québec, dont le salaire annuel actuel est >= 45000

```
#=============================================================================
#   Nombre_Enregistrements_Sexe_Province_Annee(province, annee, sexe, salaire)
#=============================================================================

def Nombre_Enregistrements_sexe_Province_Annee(province, annee, sexe, salaire):
    with open('Stats.csv','r') as csv_file:
        csv_reader = csv.reader(csv_file, delimiter=';')
        nombre = 0;

        for row in csv_reader:
            date = row[5]
            ann = date[6:10]
            print ('ann = ' +str(ann)+ '\n')
            if row[3] == province and row[2] == sexe and row[4] >= salaire and annee >= int(ann):
               nombre += 1

    csv_file.close()
    print nombre
    return nombre
```

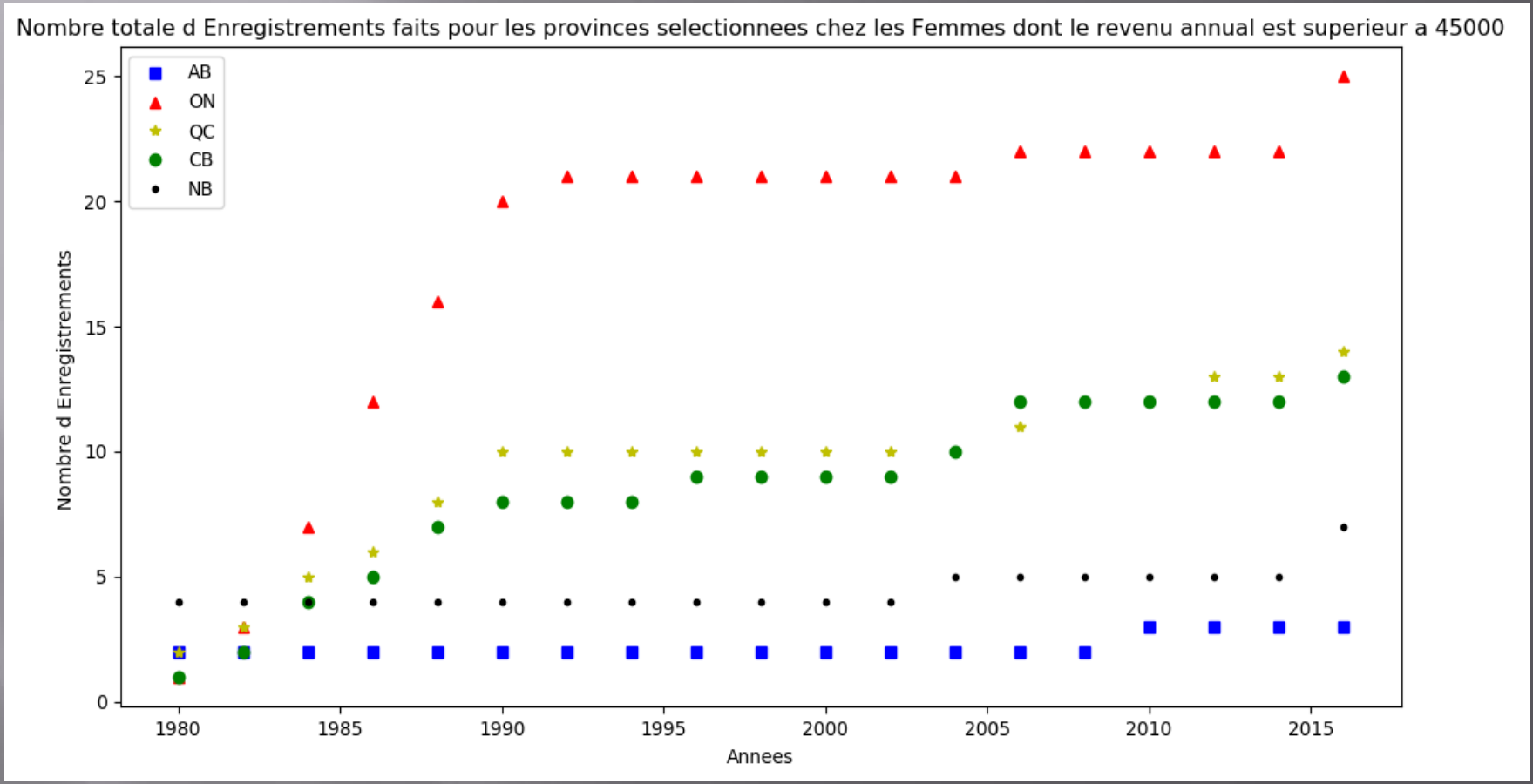
Nombre_Enregistrements_sexe_Province_Annee('QC', 1996, 'F', 45000)

La réponse pour cette requête est 14

Petite requête

```
#=============================================================================================================================
#    Statistiques_des_Enregistrements()
#=============================================================================================================================

def Toutes_Les_Enregistrements_sexe_salaire():
    AB = []
    ON = []
    QC = []
    CB = []
    NB = []

    provinces = ['AB', 'ON', 'QC', 'CB', 'NB']
    annees = [1980, 1982, 1984, 1986, 1988, 1990, 1992, 1994, 1996, 1998, 2000, 2002, 2004, 2006, 2008, 2010, 2012, 2014, 2016]

    for prov in provinces:
        if prov == 'AB':
           AB = Enregistrements_Province_sexe_salaire('AB', annees, 'F', 45000)
        elif prov == 'ON':
           ON = Enregistrements_Province_sexe_salaire('ON', annees, 'F', 45000)
        elif prov == 'QC':
           QC = Enregistrements_Province_sexe_salaire('QC', annees, 'F', 45000)
        elif prov == 'CB':
           CB = Enregistrements_Province_sexe_salaire('CB', annees, 'F', 45000)
        elif prov == 'NB':
           NB = Enregistrements_Province_sexe_salaire('NB', annees, 'F', 45000)

    return AB, ON, QC, CB, NB
```

Visualisation



ON remarque qu'il y a plus d'immigrant venant en ONTARIO que dans les autres provinces

Personnalisation

ON peut même personnaliser la 2e requête :

- Varier les années de la requête
- Decider que c'est soit les statistiques des hommes ou celles des femmes ou même des deux
- Choisir le salaire désiré

Petite Requête

1) Varier les années de la requête

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

def Annees():
    annees = []
    annee_debut = input('a partir de quelle annee voulex -vous ces statistiques ? ')
    annee_fin = input('jusqu a quelle annee voulez-vous ces statistiques ? ')
    annee_fin = annee_fin+1;

    for x in range(annee_debut, annee_fin):
        annees.append(x)
    print annees
```

Petite Requête

2) Produire les statistiques des hommes et des Femmes, 3) mettre le salaire au choix

```
#=============================================================================================
#    Statistiques_des_Enregistrements()
#=============================================================================================

def Enregistrements_Personnalised_sexe_salaire():
    AB_F = []
    AB_H = []
    ON_F = []
    ON_H = []
    QC_F = []
    QC_H = []
    CB_F = []
    QC_H = []
    NB_F = []
    NB_H = []

    provinces = ['AB', 'ON', 'QC', 'CB', 'NB']
    annees = Annees() #[2004, 2006, 2008, 2010, 2012]
    salaire = input('Vous voulez des statistiques superieur a quel salaire ? ')

    for prov in provinces:
        if prov == 'AB':
           AB_F = Enregistrements_Province_sexe_salaire('AB', annees, 'F', salaire)
           AB_H = Enregistrements_Province_sexe_salaire('AB', annees, 'H', salaire)
        elif prov == 'ON':
           ON_F = Enregistrements_Province_sexe_salaire('ON', annees, 'F', salaire)
           ON_H = Enregistrements_Province_sexe_salaire('ON', annees, 'H', salaire)
        elif prov == 'QC':
           QC_F = Enregistrements_Province_sexe_salaire('QC', annees, 'F', salaire)
           QC_H = Enregistrements_Province_sexe_salaire('QC', annees, 'H', salaire)
        elif prov == 'CB':
           CB_F = Enregistrements_Province_sexe_salaire('CB', annees, 'F', salaire)
           CB_H = Enregistrements_Province_sexe_salaire('CB', annees, 'H', salaire)
        elif prov == 'NB':
           NB_F = Enregistrements_Province_sexe_salaire('NB', annees, 'F', salaire)
           NB_H = Enregistrements_Province_sexe_salaire('NB', annees, 'H', salaire)

    return AB_F, AB_H, ON_F, ON_H, QC_F, QC_H, CB_F, CB_H, NB_F, NB_H, annees, salaire
```

Petite Requête

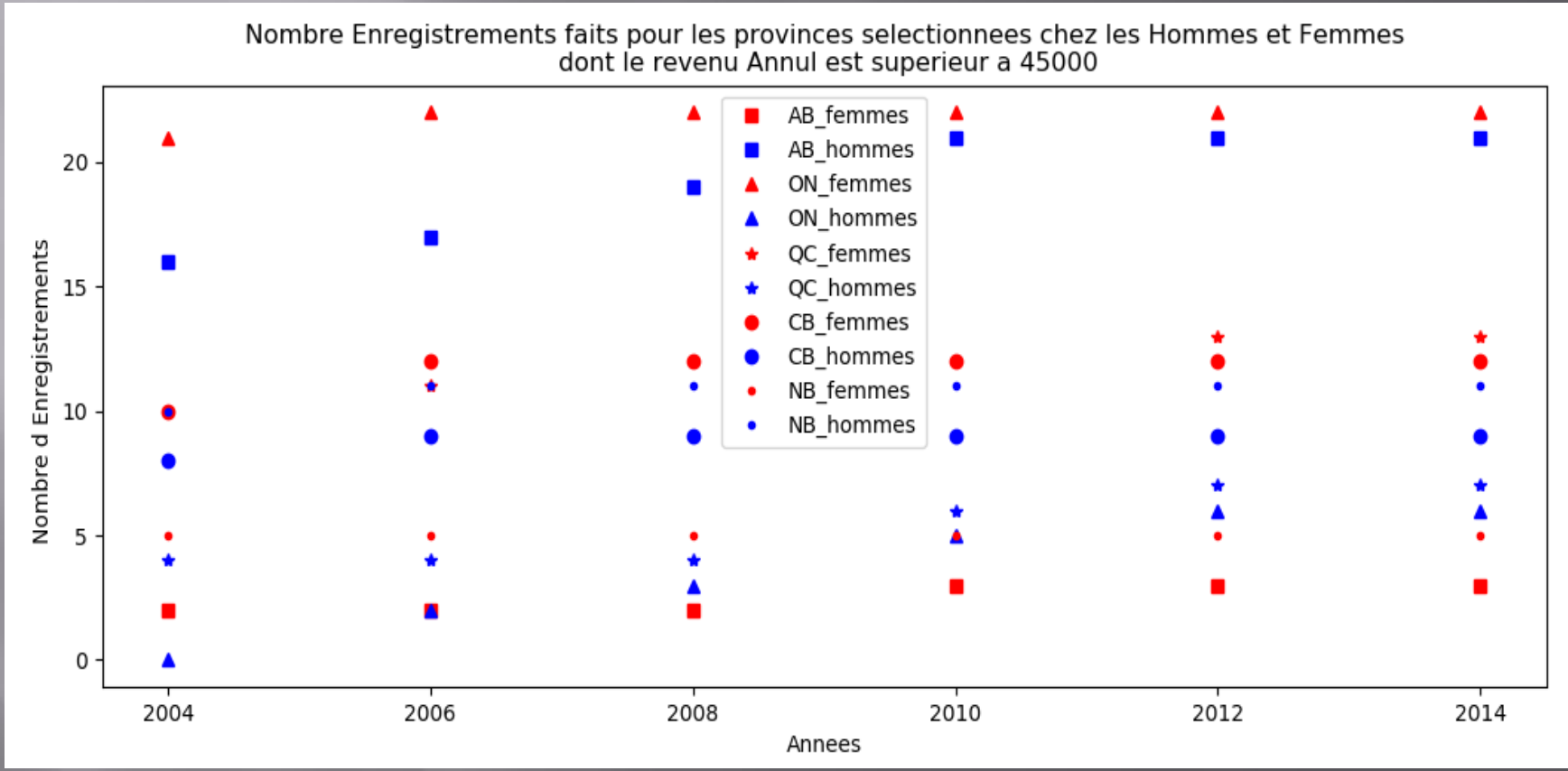
```
#=============================================================================================================================
#   Statistiques_des_Enregistrements_sexe_salaire()
#=============================================================================================================================

def Statistiques_des_Enregistrements_Personnalise_sexe_salaire():
    AB_F, AB_H, ON_F, ON_H, QC_F, QC_H, CB_F, CB_H, NB_F, NB_H, annees, salaire = Enregistrements_Personnalised_sexe_salaire()

    plt.plot(annees, AB_F, 'rs', label='AB_femmes')
    plt.plot(annees, AB_H, 'bs', label='AB_hommes')

    plt.plot(annees, ON_F, 'r^', label='ON_femmes')
    plt.plot(annees, ON_H, 'b^', label='ON_hommes')

    plt.plot(annees, QC_F, 'r*', label='QC_femmes')
    plt.plot(annees, QC_H, 'b*', label='QC_hommes')

    plt.plot(annees, CB_F, 'ro', label='CB_femmes')
    plt.plot(annees, CB_H, 'bo', label='CB_hommes')

    plt.plot(annees, NB_F, 'r.', label='NB_femmes')
    plt.plot(annees, NB_H, 'b.', label='NB_hommes')

    #-------------------------------------------------
    plt.title('Nombre Enregistrements faits pour les provinces selectionnees chez les Hommes et Femmes \n' +
              'dont le revenu Annul est superieur a ' +str(salaire))
```

```
    plt.legend()
    plt.show()
```

Exécution

```
 RESTART: C:\Users\Admin\Desktop\Cegep Victoriaville\SITE_WEB_COURS\Informatique
\Data Science\Assurances_Emploi_DataScience\Stats_Demandes.py
a partir de quelle annee voulez -vous ces statistiques ? 2004
jusqu a quelle annee voulez-vous ces statistiques ? 2014
2004
2006
2008
2010
2012
2014
Vous voulez des statistiques superieur a quel salaire ? 45000
```

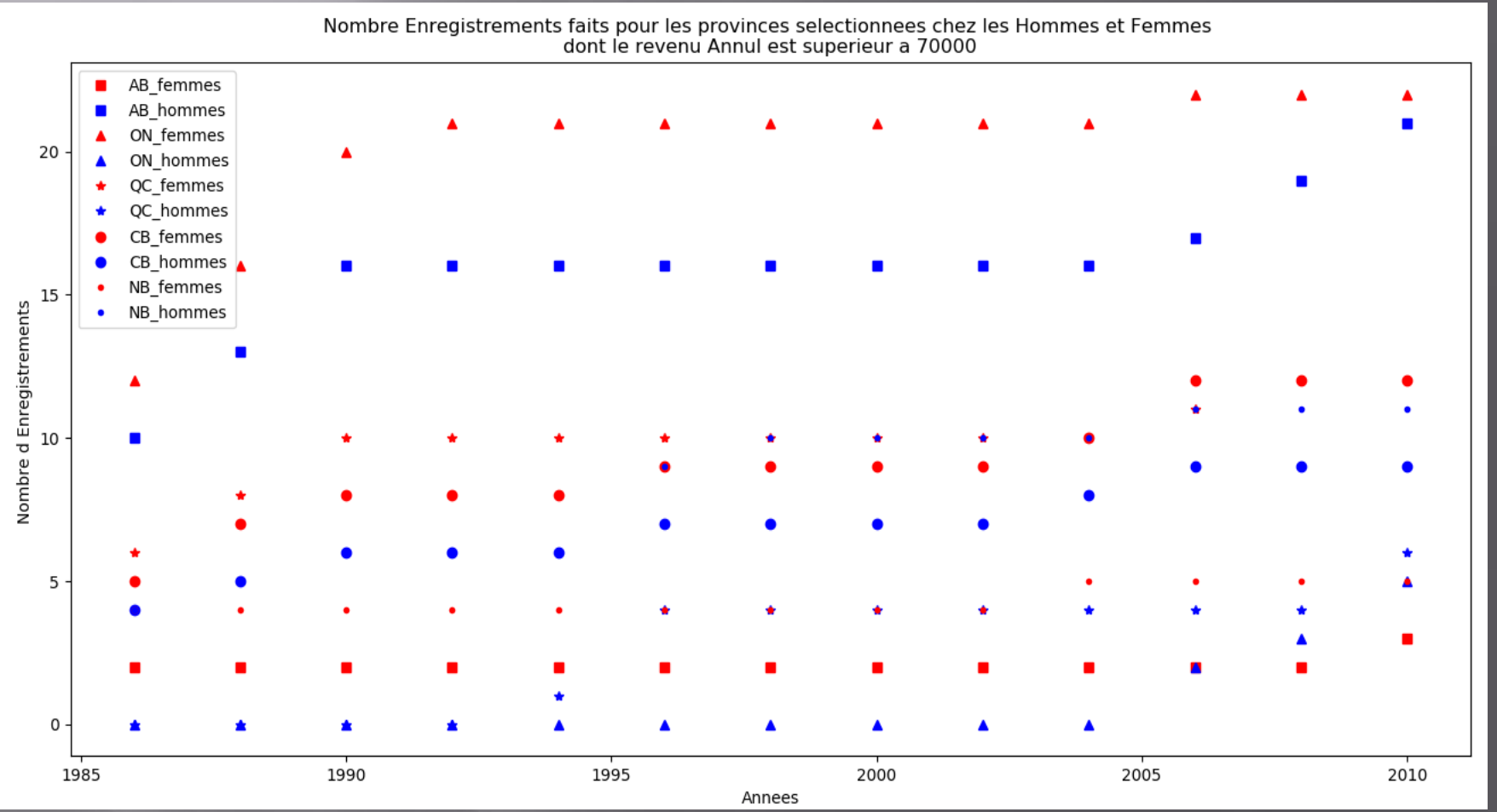
Visualisation



Exécution

```
 RESTART: C:\Users\Admin\Desktop\Cegep Victoriaville\SITE_WEB_COURS\Informatique
\Data Science\Assurances_Emploi_DataScience\Stats_Demandes.py
a partir de quelle annee voulez -vous ces statistiques ? 1986
jusqu a quelle annee voulez-vous ces statistiques ? 2010
1986
1988
1990
1992
1994
1996
1998
2000
2002
2004
2006
2008
2010
Vous voulez des statistiques superieur a quel salaire ? 70000
```

Visualisation



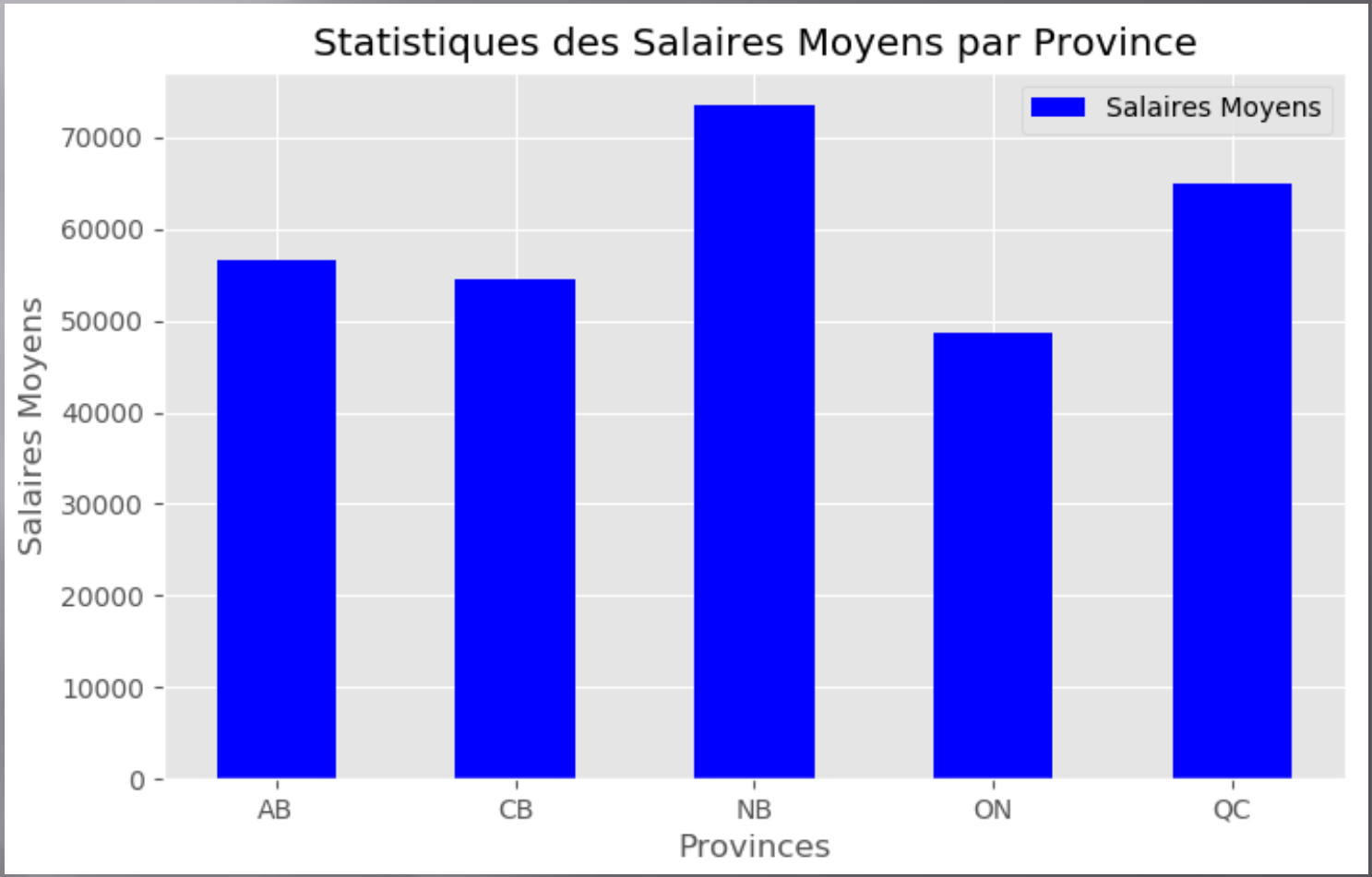
Petite requête supplémentaire

- Il n’est pas toujours exigé de faire des sous-requêtes;
- On peut faire des parcours avec 2 boucles imbriquées
- C’est le cas si on veut visualiser les salaires moyens totaux par province.

```
def Enregistrements_Province_salaire():
    provinces = ['AB','ON','QC','CB','NB']
    salaires = []
    salaire_AB = 0
    n_AB = 0
    salaire_ON = 0
    n_ON = 0
    salaire_QC = 0
    n_QC = 0
    salaire_CB = 0
    n_CB = 0
    salaire_NB = 0
    n_NB = 0

    with open('Stats.csv','r') as csv_file:
        csv_reader = csv.reader(csv_file, delimiter=';')
        nombre = 0;
        for x in range(0, 4):
            for row in csv_reader:
                if row[3] == 'AB' and row[4] != '':
                    salaire_AB = salaire_AB + int(row[4])
                    n_AB += 1
                elif row[3] == 'ON' and row[4] != '':
                    salaire_ON = salaire_ON + int(row[4])
                    n_ON += 1
                elif row[3] == 'QC' and row[4] != '':
                    salaire_QC = salaire_QC + int(row[4])
                    n_QC += 1
                elif row[3] == 'CB' and row[4] != '':
                    salaire_CB = salaire_CB + int(row[4])
                    n_CB += 1
                elif row[3] == 'NB' and row[4] != '':
                    salaire_NB = salaire_NB + int(row[4])
                    n_NB += 1

        salaire_AB = salaire_AB / n_AB
        salaire_ON = salaire_ON / n_ON
        salaire_QC = salaire_QC / n_QC
        salaire_CB = salaire_CB / n_CB
        salaire_NB = salaire_NB / n_NB

        salaires.append(salaire_AB)
        salaires.append(salaire_ON)
        salaires.append(salaire_QC)
        salaires.append(salaire_CB)
        salaires.append(salaire_NB)

    return salaires
```

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

def Statistiques_des_Salaires_Totaux_Par_Provinces():
    plt.style.use('ggplot')
    provinces = ['AB','ON','QC','CB','NB']
    salaires = Enregistrements_Province_salaire()

    plt.bar(provinces, salaires, label='Salaires Moyens', color='b', align='center', width=0.5)
    #-----------------------------------------------------------
    plt.title('Statistiques des Salaires Moyens par Province')
    plt.ylabel('Salaires Moyens')
    plt.xlabel('Provinces')

    plt.legend()
    plt.show()
```

Visualisation



Les résultats nous montrent que NB > QC > AB > CB > ON

Merci de m'avoir suivi dans cette aventure de la DataSciences

Me contacter au sandejoel@yahoo.ca
pour vos questions