week_5_project

Adejare Windokun

Monday, September 29, 2014

roject 2: Profiling a Data Set

Data obtained from: http://www.cms.gov/apps/ama/license-2011.asp?file=http://download.cms.gov/Research-Statistics-Data-and-Systems /Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Downloads/Medicare-Physician-and-Other-Supplier-PUF-a-CY2012.zip This data set contains all CMS payments to healthcare professionals for the year 2012, and includes such data as provider id, first name, last name, initials, speciality, amount billed,procedure code, amount paid, mean for amount paid etc The actual dataset consisits of 9 million records and is 1.7 GB in size so I am going to use a subset which consists of providers last whose names start with 'A'

This dataset is provider as an Excel file and I therefore opened and and saved it as a txt file. Direct import into R leads to data corruption I will be using the dplyr package in addtion to base R functions for the data profiling

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
require(plyr)
```

```
## Loading required package: plyr
```

```
require(dplyr)
```

```
## Loading required package: dplyr
##
## Attaching package: 'dplyr'
##
## The following objects are masked from 'package:plyr':
##
##     arrange, desc, failwith, id, mutate, summarise, summarize
##
## The following objects are masked from 'package:stats':
##
##     filter, lag
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
f= "C:/Users/jare/SkyDrive/WorkDocs/CUNY/607/Week 5/PUPD PUF A.txt"
data = read.table (f, header = TRUE,  sep = ",")
head(data)
```

```
##          npi nppes_provider_last_org_name nppes_provider_first_name
## 1 1003002494                     ANDERSON                     JOSEPH
## 2 1003002494                     ANDERSON                     JOSEPH
## 3 1003002494                     ANDERSON                     JOSEPH
## 4 1003002502                       ADKINS                      CAROL
## 5 1003002502                       ADKINS                      CAROL
## 6 1003002502                       ADKINS                      CAROL
##   nppes_provider_mi nppes_credentials nppes_provider_gender
## 1                 M             M.D.                     M
## 2                 M             M.D.                     M
## 3                 M             M.D.                     M
## 4                 J               PT                     F
## 5                 J               PT                     F
## 6                 J               PT                     F
##   nppes_entity_code nppes_provider_street1 nppes_provider_street2
## 1                 I               802 B ST
## 2                 I               802 B ST
## 3                 I               802 B ST
## 4                 I         1605 SCHERM RD
## 5                 I         1605 SCHERM RD
## 6                 I         1605 SCHERM RD
##   nppes_provider_city nppes_provider_zip nppes_provider_state
## 1          SAN RAFAEL          949013026                    CA
## 2          SAN RAFAEL          949013026                    CA
## 3          SAN RAFAEL          949013026                    CA
## 4           OWENSBORO          423015300                    KY
## 5           OWENSBORO          423015300                    KY
## 6           OWENSBORO          423015300                    KY
##   nppes_provider_country      provider_type
## 1                     US          Pathology
## 2                     US          Pathology
## 3                     US          Pathology
## 4                     US Physical Therapist
## 5                     US Physical Therapist
## 6                     US Physical Therapist
##   medicare_participation_indicator place_of_service hcpcs_code
## 1                                Y                O      88305
## 2                                Y                O      88342
## 3                                Y                O      G0416
## 4                                Y                O      97001
## 5                                Y                O      97035
## 6                                Y                O      97110
##              hcpcs_description line_srvc_cnt bene_unique_cnt
## 1   Tissue exam by pathologist          1797             165
## 2         Immunohistochemistry           568              72
## 3 Sat biopsy prostate 1-20 spc            25              25
```

```
## 4                Pt evaluation            31              31
## 5           Ultrasound therapy            47              22
## 6        Therapeutic exercises           983              58
##   bene_day_srvc_cnt average_Medicare_allowed_amt
## 1               168                      $133.43
## 2                73                      $131.74
## 3                25                      $866.27
## 4                31                       $68.93
## 5                46                       $10.72
## 6               331                       $26.69
##   stdev_Medicare_allowed_amt average_submitted_chrg_amt
## 1                     $20.95                   $263.23
## 2                     $16.22                   $255.64
## 3                      $0.00                  $1733.00
## 4                      $0.95                    $91.00
## 5                      $0.00                    $15.00
## 6                      $1.98                    $37.30
##   stdev_submitted_chrg_amt average_Medicare_payment_amt
## 1                   $41.17                      $105.53
## 2                   $31.27                      $102.43
## 3                    $0.00                      $693.02
## 4                    $0.00                       $47.75
## 5                    $0.00                        $8.03
## 6                    $9.38                       $20.75
##   stdev_Medicare_payment_amt
## 1                     $25.87
## 2                     $39.47
## 3                      $0.00
## 4                     $16.17
## 5                      $2.12
## 6                      $5.65
```

```
names(data)
```

```
##  [1] "npi"                              "nppes_provider_last_org_name"
##  [3] "nppes_provider_first_name"        "nppes_provider_mi"
##  [5] "nppes_credentials"                "nppes_provider_gender"
##  [7] "nppes_entity_code"                "nppes_provider_street1"
##  [9] "nppes_provider_street2"           "nppes_provider_city"
## [11] "nppes_provider_zip"               "nppes_provider_state"
## [13] "nppes_provider_country"           "provider_type"
## [15] "medicare_participation_indicator" "place_of_service"
## [17] "hcpcs_code"                       "hcpcs_description"
## [19] "line_srvc_cnt"                    "bene_unique_cnt"
## [21] "bene_day_srvc_cnt"                "average_Medicare_allowed_amt"
## [23] "stdev_Medicare_allowed_amt"       "average_submitted_chrg_amt"
## [25] "stdev_submitted_chrg_amt"         "average_Medicare_payment_amt"
## [27] "stdev_Medicare_payment_amt"
```

```
data = tbl_df(data)
```

create easier names for the columns, and select which of the colums we are going to use

```
subdata = select(data, npi, nppes_provider_gender, nppes_entity_code,
nppes_provider_state, provider_type, place_of_service,
average_Medicare_payment_amt)

n = c("provider_ID", "gender", "entity_code", "state", "provider_type",
"place_of_service", "amount_paid")
names(subdata) = n
subdata = tbl_df(subdata)
```

We have a total of 397,221 records with 7 columns

```
dim(subdata)

## [1] 397221      7
```

will remove the rows where the entity is not an individual ie I, this is because the dataset also contains payments to organisations

```
idata = filter(subdata, entity_code == 'I')
```

we are now left with 375,583 individual payments to providers

```
dim(idata)

## [1] 375583      7
```

The amount_paid which is the numeric amount that CMS paid for the serive got imported as a character, will convert it back to numerical

```
idata$amount_paid = as.numeric(idata$amount_paid)
str(idata)

## Classes 'tbl_df', 'tbl' and 'data.frame':    375583 obs. of  7 variables:
##  $ provider_ID     : int  1003002494 1003002494 1003002494 1003002502
1003002502 1003002502 1003002502 1003002502 1003002502 1003006107 ...
##  $ gender          : Factor w/ 3 levels "","F","M": 3 3 3 2 2 2 2 2 2 3
...
##  $ entity_code     : Factor w/ 2 levels "I","O": 1 1 1 1 1 1 1 1 1 1 ...
##  $ state           : Factor w/ 58 levels "AK","AL","AP",..: 7 7 7 21 21 21
21 21 21 43 ...
##  $ provider_type   : Factor w/ 89 levels "Addiction Medicine",..: 61 61 61
66 66 66 66 66 66 10 ...
##  $ place_of_service: Factor w/ 2 levels "F","O": 2 2 2 2 2 2 2 2 2 1 ...
##  $ amount_paid     : num  1000 607 32057 26524 33981 ...
```

Analysis Requirements You should include analysis of each variable. Summarize the values, identify any questionable values or outliers, and explain the (possible) significance of any missing values in the column. In addition, consider the possibilities of correlations among the variables. Look for any interesting patterns. (Do two columns correlate perfectly? Do missing values appear consistent across observations? These are just two such interesting

possibilities.) Consider whether there are any variables that should be recoded or binned. Do such transformations lead to further insights into the data set? Remember that your ultimate goal is to tell a story from the data. Include basic visualizations where appropriate.

provider_ID

**********Analysis The provider_ID field consists of a numeric code that CMS assigns to each provider and organization and it is a unique identifier it therefore allows the tracking of payments made to a single entity. By definition, this field cannot have a null or NA value The data below shows the total amount paid to each provider, the mean and standard deviation

```
summary.provider_ID = aggregate(amount_paid ~ provider_ID, data = idata, FUN
= function(x) c(sum = sum(x), mean = mean(x), sd = sd(x)))
head(summary.provider_ID)
```

```
##   provider_ID amount_paid.sum amount_paid.mean amount_paid.sd
## 1  1003002494           33664            11221          18045
## 2  1003002502          108408            18068           9893
## 3  1003006107          230095            15340          10930
## 4  1003007477           70650            23550           5569
## 5  1003010075           28123            14062           5239
## 6  1003010182            2673             2673             NA
```

The function the creates the sum, mean and sd returns a list in a column, so we need to separtate the list before using it Ideally should create a funtion that does this automatically

```
summary.provider_ID = cbind(summary.provider_ID[1],
(unlist(summary.provider_ID[,2])))
```

The top paid providers

```
head(arrange(summary.provider_ID, desc(sum)))
```

```
##   provider_ID     sum  mean    sd
## 1  1497708275 3866238 17654 10068
## 2  1053389957 3368383 24951  9454
## 3  1992754147 3088480 23942  8956
## 4  1770564411 3084336 23726  9879
## 5  1285682377 2936746 20977 10467
## 6  1720177207 2918935 24123 10341
```

The bottom paid providers The absent sd simply indicates that these providers where only paid for a single service and therefore only received a single payment

```
tail(arrange(summary.provider_ID, desc(sum)))
```

```
##       provider_ID sum mean sd
## 33886  1871704866 322  322 NA
## 33887  1407866890 317  317 NA
## 33888  1184658791 306  306 NA
## 33889  1255370243 267  267 NA
```

```
## 33890  1376762088 225  225 NA
## 33891  1184933293 100  100 NA
```

Ggender **********Analysis This is a character field that stores the entities gender, M = Male, F = Female, and null where the entity is an organization we have removed organizational entites therefore we only have M and F. The results of this are interesting. While Males where paid almost 4 times the total amount in aggregate than females the mean and sd payments are very similar, inidcating that this is more a function of the amount of males present in the dataset

```
summary1 = aggregate(amount_paid ~ gender, data = idata, FUN = function(x)
c(sum = sum(x), mean = mean(x), sd = sd(x)))
summary1
```

```
##   gender amount_paid.sum amount_paid.mean amount_paid.sd
## 1      F       1.750e+09        2.025e+04      1.131e+04
## 2      M       5.797e+09        2.005e+04      1.150e+04
```

This shows that we have twice as many Male providers as Female providers

```
ddply(idata,~gender,summarise,number_of_providers =
length(unique(provider_ID)))
```

```
##   gender number_of_providers
## 1      F               11634
## 2      M               22257
```

Entity_code **********Analysis not required - we are only looking at individual providers This is a character column that stores the values I for individual providers and O for organizations We have already filetered out and left only the individual proviers

State **********Analysis This is a character code that stores the state abbrevation of where the provider practices. A code of ZZ means the provider is not in the US. The other codes include the following:

'XX' = 'Unknown' 'AA' = 'Armed Forces Central/South America' 'AE' = 'Armed Forces Europe' 'AP' = 'Armed Forces Pacific' 'AS' = 'American Samoa' 'GU' = 'Guam' 'MP' = 'North Mariana Islands' 'PR' = 'Puerto Rico' 'VI' = 'Virgin Islands' 'ZZ' = 'Foreign Country'

The results are as expected. The highest paid states from CMS are Florida, Texas, California and New York, which incidentally tend to have larger populations and therefore larger elderly people whose insurance is therefore covered by Medicare

```
summary2 = aggregate(amount_paid ~ state, data = idata, FUN = function(x)
c(sum = sum(x), mean = mean(x), sd = sd(x)))
summary = cbind(summary2[1], (unlist(summary2[,2])))
summary
```

```
##    state       sum  mean    sd
## 1     AK   8603978 19963 11424
## 2     AL 137646943 19176 11559
## 3     AP    241853 18604 11209
```

```
## 4     AR  73490949 19671 11417
## 5     AS     10386 10386    NA
## 6     AZ 140219839 19969 11676
## 7     CA 589854082 19975 11549
## 8     CO  73187973 20302 11230
## 9     CT  93510427 20088 11327
## 10    DC  24416104 20296 11767
## 11    DE  29810403 20101 11709
## 12    FL 595139654 19695 11692
## 13    GA 200894202 19226 11601
## 14    GU    159485 19936 12178
## 15    HI  20685528 21216 10778
## 16    IA  85285646 20039 10960
## 17    ID  22525876 20572 10505
## 18    IL 378146374 20448 11450
## 19    IN 167489766 19980 11335
## 20    KS  63898898 20666 11097
## 21    KY 128138710 19569 11756
## 22    LA 113121724 20438 11164
## 23    MA 198450154 20719 11348
## 24    MD 173071925 20321 11591
## 25    ME  37823646 21297 10951
## 26    MI 329496831 19968 11562
## 27    MN 113040880 20211 11183
## 28    MO 173195889 20636 11260
## 29    MP    526484 15954 11215
## 30    MS  65950108 19817 11578
## 31    MT  15712081 19255 11338
## 32    NC 269310612 19608 11585
## 33    ND  21956813 20753 10658
## 34    NE  43637380 20613 10727
## 35    NH  25516802 21123 10937
## 36    NJ 254240852 19889 11623
## 37    NM  35569653 19783 11485
## 38    NV  58218081 19708 11744
## 39    NY 496833493 20375 11433
## 40    OH 309888027 20571 11320
## 41    OK  75118717 19778 11287
## 42    OR  57032257 20275 11222
## 43    PA 297575420 20603 11390
## 44    PR  29789426 20237 11075
## 45    RI  26833466 19965 11585
## 46    SC  99403550 19575 11546
## 47    SD  21647452 20519 10879
## 48    TN 182668767 19562 11442
## 49    TX 592334741 20019 11525
## 50    UT  47461162 20564 11261
## 51    VA 206677798 20215 11408
## 52    VI   1137793 24208 10680
## 53    VT  10128825 21551 10614
```

```
## 54    WA 111851633 20351 11165
## 55    WI 137833627 20621 11191
## 56    WV  71152424 20768 11250
## 57    WY   8648914 20942 10696
## 58    ZZ    932270 22197 11079
```

```
arrange(summary, desc(sum))
```

```
##    state       sum  mean    sd
## 1     FL 595139654 19695 11692
## 2     TX 592334741 20019 11525
## 3     CA 589854082 19975 11549
## 4     NY 496833493 20375 11433
## 5     IL 378146374 20448 11450
## 6     MI 329496831 19968 11562
## 7     OH 309888027 20571 11320
## 8     PA 297575420 20603 11390
## 9     NC 269310612 19608 11585
## 10    NJ 254240852 19889 11623
## 11    VA 206677798 20215 11408
## 12    GA 200894202 19226 11601
## 13    MA 198450154 20719 11348
## 14    TN 182668767 19562 11442
## 15    MO 173195889 20636 11260
## 16    MD 173071925 20321 11591
## 17    IN 167489766 19980 11335
## 18    AZ 140219839 19969 11676
## 19    WI 137833627 20621 11191
## 20    AL 137646943 19176 11559
## 21    KY 128138710 19569 11756
## 22    LA 113121724 20438 11164
## 23    MN 113040880 20211 11183
## 24    WA 111851633 20351 11165
## 25    SC  99403550 19575 11546
## 26    CT  93510427 20088 11327
## 27    IA  85285646 20039 10960
## 28    OK  75118717 19778 11287
## 29    AR  73490949 19671 11417
## 30    CO  73187973 20302 11230
## 31    WV  71152424 20768 11250
## 32    MS  65950108 19817 11578
## 33    KS  63898898 20666 11097
## 34    NV  58218081 19708 11744
## 35    OR  57032257 20275 11222
## 36    UT  47461162 20564 11261
## 37    NE  43637380 20613 10727
## 38    ME  37823646 21297 10951
## 39    NM  35569653 19783 11485
## 40    DE  29810403 20101 11709
## 41    PR  29789426 20237 11075
```

```
## 42    RI  26833466 19965 11585
## 43    NH  25516802 21123 10937
## 44    DC  24416104 20296 11767
## 45    ID  22525876 20572 10505
## 46    ND  21956813 20753 10658
## 47    SD  21647452 20519 10879
## 48    HI  20685528 21216 10778
## 49    MT  15712081 19255 11338
## 50    VT  10128825 21551 10614
## 51    WY   8648914 20942 10696
## 52    AK   8603978 19963 11424
## 53    VI   1137793 24208 10680
## 54    ZZ    932270 22197 11079
## 55    MP    526484 15954 11215
## 56    AP    241853 18604 11209
## 57    GU    159485 19936 12178
## 58    AS     10386 10386    NA
```

Provider_type **********Analysis This is a character column that stores the primary classification of the provider - ie the specialty. In terms of aggregate payments, as expected. Internal Medicine received the largest payments, followed by suprisingly radiology

```
summary3 = aggregate(amount_paid ~ provider_type, data = idata, FUN =
function(x) c(sum = sum(x), mean = mean(x), sd = sd(x)))
provider.summary = cbind(summary3[1], (unlist(summary3[,2])))
arrange(provider.summary, desc(sum))
```

```
##                               provider_type       sum  mean    sd
## 1                         Internal Medicine 1.310e+09 19511 11957
## 2                    Diagnostic Radiology 1.057e+09 22966 10177
## 3                         Family Practice 6.722e+08 18306 11461
## 4                              Cardiology 5.682e+08 19622 11719
## 5                       Orthopedic Surgery 1.987e+08 19573 10233
## 6                            Ophthalmology 1.974e+08 22565 10765
## 7                          Anesthesiology 1.847e+08 20040 12110
## 8                       Nurse Practitioner 1.837e+08 20267 10695
## 9                        Pulmonary Disease 1.778e+08 20645 12396
## 10                      Emergency Medicine 1.604e+08 20446 12102
## 11                                Podiatry 1.576e+08 21492  9671
## 12                     Gastroenterology 1.560e+08 17851 11867
## 13                             Neurology 1.487e+08 20963 11645
## 14                            Dermatology 1.479e+08 22133 11051
## 15                   Hematology/Oncology 1.459e+08 15549 12445
## 16                   Physician Assistant 1.381e+08 19671 10855
## 17                             Pathology 1.345e+08 22157  9932
## 18                        General Surgery 1.318e+08 20501 11689
## 19                             Nephrology 1.291e+08 19395 11789
## 20                    Physical Therapist 1.260e+08 18861  9265
## 21                              Urology 1.190e+08 18551 11838
## 22                             Optometry 1.077e+08 24774  9675
```

```
## 23                        Obstetrics/Gynecology 1.004e+08 22808  9469
## 24                                   Psychiatry 9.371e+07 22244 10649
## 25       Physical Medicine and Rehabilitation 7.729e+07 21344 11519
## 26                              Otolaryngology 7.517e+07 20720 11553
## 27                                        CRNA 6.300e+07 19085 13168
## 28                                 Rheumatology 5.956e+07 18254 11431
## 29                            Radiation Oncology 5.378e+07 17364 11010
## 30                              Vascular Surgery 5.301e+07 18586 11870
## 31                              General Practice 5.200e+07 18833 11539
## 32                                 Endocrinology 4.757e+07 18289 12354
## 33                              Medical Oncology 4.457e+07 16039 12495
## 34                                  Neurosurgery 4.291e+07 21067 11146
## 35                            Infectious Disease 4.001e+07 20708 12523
## 36                    Cardiac Electrophysiology 3.456e+07 20365 10853
## 37                     Interventional Radiology 3.406e+07 21638 10968
## 38             Interventional Pain Management 3.269e+07 20043 11865
## 39                            Allergy/Immunology 2.359e+07 22045 11630
## 40               Critical Care (Intensivists) 2.354e+07 21097 12500
## 41                                  Chiropractic 2.324e+07 14325  4040
## 42                        Clinical Psychologist 2.050e+07 22408 10938
## 43          Plastic and Reconstructive Surgery 1.783e+07 21478 11045
## 44                            Geriatric Medicine 1.780e+07 19803 12149
## 45                                  Hand Surgery 1.656e+07 21092  8988
## 46                               Pain Management 1.622e+07 20556 11595
## 47                               Cardiac Surgery 1.512e+07 18084 12118
## 48             Licensed Clinical Social Worker 1.487e+07 22423  8037
## 49                               Thoracic Surgery 1.422e+07 17626 11965
## 50    Colorectal Surgery (formerly proctology) 1.121e+07 19769 11653
## 51                         Occupational therapist 1.107e+07 19810  8851
## 52         Audiologist (billing independently) 9.782e+06 15143  9981
## 53                            Pediatric Medicine 8.123e+06 19811 11427
## 54                        Gynecological/Oncology 6.166e+06 16532 12512
## 55                                    Hematology 5.521e+06 19237 12685
## 56                               Nuclear Medicine 5.242e+06 21138  9746
## 57         Certified Clinical Nurse Specialist 4.118e+06 22751  9645
## 58                             Surgical Oncology 3.976e+06 19981 12066
## 59           Osteopathic Manipulative Medicine 2.799e+06 19440 11133
## 60                 Anesthesiologist Assistants 2.763e+06 24451 10877
## 61                Oral Surgery (dentists only) 1.938e+06 20183 11255
## 62                           Preventive Medicine 1.937e+06 21517 11641
## 63                               Sports Medicine 1.586e+06 19825 11274
## 64                          Maxillofacial Surgery 1.533e+06 19910  9767
## 65                            Clinical Laboratory 1.422e+06 18472 12217
## 66 Registered Dietician/Nutrition Professional 1.407e+06 16553  5329
## 67                   Hospice and Palliative Care 1.122e+06 22445 12046
## 68                   Speech Language Pathologist 1.082e+06 22548 11600
## 69                       Certified Nurse Midwife 8.384e+05 24658  6018
## 70                   Peripheral Vascular Disease 7.779e+05 16910 12832
## 71                                Neuropsychiatry 7.107e+05 21536 12666
## 72             Unknown Physician Specialty Code 6.265e+05 20884 10665
```

```
## 73                          Geriatric Psychiatry 6.115e+05 21840 10322
## 74        Psychologist (billing independently) 4.992e+05 20799 11892
## 75                            Addiction Medicine 3.075e+05 21965 11887
## 76     Independent Diagnostic Testing Facility 1.273e+05 18182 12752
## 77                     Unknown Supplier/Provider 1.094e+05 15632 16221
## 78                                Sleep Medicine 2.566e+04 25656    NA
```

However, if we do arrange payments by mean payments, Sleep Medicine receives the highers followed by Psychologist, and Addiction Medicine. This may be a function of the dataset, as we are only looking a providers whose names start with 'A'.

```
arrange(provider.summary, desc(mean))
```

```
##                                  provider_type       sum  mean    sd
## 1                                Sleep Medicine 2.566e+04 25656    NA
## 2                                    Optometry 1.077e+08 24774  9675
## 3                       Certified Nurse Midwife 8.384e+05 24658  6018
## 4                    Anesthesiologist Assistants 2.763e+06 24451 10877
## 5                           Diagnostic Radiology 1.057e+09 22966 10177
## 6                         Obstetrics/Gynecology 1.004e+08 22808  9469
## 7           Certified Clinical Nurse Specialist 4.118e+06 22751  9645
## 8                                  Ophthalmology 1.974e+08 22565 10765
## 9                    Speech Language Pathologist 1.082e+06 22548 11600
## 10                   Hospice and Palliative Care 1.122e+06 22445 12046
## 11              Licensed Clinical Social Worker 1.487e+07 22423  8037
## 12                        Clinical Psychologist 2.050e+07 22408 10938
## 13                                   Psychiatry 9.371e+07 22244 10649
## 14                                    Pathology 1.345e+08 22157  9932
## 15                                   Dermatology 1.479e+08 22133 11051
## 16                           Allergy/Immunology 2.359e+07 22045 11630
## 17                           Addiction Medicine 3.075e+05 21965 11887
## 18                          Geriatric Psychiatry 6.115e+05 21840 10322
## 19                     Interventional Radiology 3.406e+07 21638 10968
## 20                              Neuropsychiatry 7.107e+05 21536 12666
## 21                          Preventive Medicine 1.937e+06 21517 11641
## 22                                       Podiatry 1.576e+08 21492  9671
## 23            Plastic and Reconstructive Surgery 1.783e+07 21478 11045
## 24          Physical Medicine and Rehabilitation 7.729e+07 21344 11519
## 25                             Nuclear Medicine 5.242e+06 21138  9746
## 26                   Critical Care (Intensivists) 2.354e+07 21097 12500
## 27                                 Hand Surgery 1.656e+07 21092  8988
## 28                                 Neurosurgery 4.291e+07 21067 11146
## 29                                    Neurology 1.487e+08 20963 11645
## 30              Unknown Physician Specialty Code 6.265e+05 20884 10665
## 31          Psychologist (billing independently) 4.992e+05 20799 11892
## 32                                Otolaryngology 7.517e+07 20720 11553
## 33                            Infectious Disease 4.001e+07 20708 12523
## 34                             Pulmonary Disease 1.778e+08 20645 12396
## 35                               Pain Management 1.622e+07 20556 11595
## 36                               General Surgery 1.318e+08 20501 11689
```

```
## 37                           Emergency Medicine 1.604e+08 20446 12102
## 38                  Cardiac Electrophysiology 3.456e+07 20365 10853
## 39                          Nurse Practitioner 1.837e+08 20267 10695
## 40                Oral Surgery (dentists only) 1.938e+06 20183 11255
## 41              Interventional Pain Management 3.269e+07 20043 11865
## 42                                Anesthesiology 1.847e+08 20040 12110
## 43                             Surgical Oncology 3.976e+06 19981 12066
## 44                         Maxillofacial Surgery 1.533e+06 19910  9767
## 45                               Sports Medicine 1.586e+06 19825 11274
## 46                            Pediatric Medicine 8.123e+06 19811 11427
## 47                        Occupational therapist 1.107e+07 19810  8851
## 48                            Geriatric Medicine 1.780e+07 19803 12149
## 49    Colorectal Surgery (formerly proctology) 1.121e+07 19769 11653
## 50                           Physician Assistant 1.381e+08 19671 10855
## 51                                   Cardiology 5.682e+08 19622 11719
## 52                            Orthopedic Surgery 1.987e+08 19573 10233
## 53                             Internal Medicine 1.310e+09 19511 11957
## 54           Osteopathic Manipulative Medicine 2.799e+06 19440 11133
## 55                                   Nephrology 1.291e+08 19395 11789
## 56                                   Hematology 5.521e+06 19237 12685
## 57                                         CRNA 6.300e+07 19085 13168
## 58                            Physical Therapist 1.260e+08 18861  9265
## 59                              General Practice 5.200e+07 18833 11539
## 60                              Vascular Surgery 5.301e+07 18586 11870
## 61                                       Urology 1.190e+08 18551 11838
## 62                           Clinical Laboratory 1.422e+06 18472 12217
## 63                               Family Practice 6.722e+08 18306 11461
## 64                                 Endocrinology 4.757e+07 18289 12354
## 65                                  Rheumatology 5.956e+07 18254 11431
## 66     Independent Diagnostic Testing Facility 1.273e+05 18182 12752
## 67                                Cardiac Surgery 1.512e+07 18084 12118
## 68                              Gastroenterology 1.560e+08 17851 11867
## 69                              Thoracic Surgery 1.422e+07 17626 11965
## 70                            Radiation Oncology 5.378e+07 17364 11010
## 71                    Peripheral Vascular Disease 7.779e+05 16910 12832
## 72 Registered Dietician/Nutrition Professional 1.407e+06 16553  5329
## 73                          Gynecological/Oncology 6.166e+06 16532 12512
## 74                               Medical Oncology 4.457e+07 16039 12495
## 75                     Unknown Supplier/Provider 1.094e+05 15632 16221
## 76                           Hematology/Oncology 1.459e+08 15549 12445
## 77         Audiologist (billing independently) 9.782e+06 15143  9981
## 78                                  Chiropractic 2.324e+07 14325  4040
```

Place_of_service **********Analysis Place of service is a character filed that stores where the service took place, O = Office, F = Facility.

More payments were made to offices as compared to facilities (clinics, hospitals), as should be expected

```
summary4 = aggregate(amount_paid ~ place_of_service, data = idata, FUN =
function(x) c(sum = sum(x), mean = mean(x), sd = sd(x)))
place.summary = cbind(summary4[1], (unlist(summary4[,2])))
arrange(place.summary, desc(sum))
```

```
##   place_of_service       sum  mean    sd
## 1                O 4.045e+09 18750 11408
## 2                F 3.502e+09 21908 11276
```

Amount_paid **********Analysis This is the actual dollar amount that CMS paid to each provider for service provided. I wil will subet the dataset to look at only anesthesiology as the provider_type for the rest of the excercise

```
str(idata)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    375583 obs. of  7 variables:
##  $ provider_ID     : int  1003002494 1003002494 1003002494 1003002502
1003002502 1003002502 1003002502 1003002502 1003002502 1003006107 ...
##  $ gender          : Factor w/ 3 levels "","F","M": 3 3 3 2 2 2 2 2 2 3
...
##  $ entity_code     : Factor w/ 2 levels "I","O": 1 1 1 1 1 1 1 1 1 1 ...
##  $ state           : Factor w/ 58 levels "AK","AL","AP",..: 7 7 7 21 21 21
21 21 21 43 ...
##  $ provider_type   : Factor w/ 89 levels "Addiction Medicine",..: 61 61 61
66 66 66 66 66 66 10 ...
##  $ place_of_service: Factor w/ 2 levels "F","O": 2 2 2 2 2 2 2 2 2 1 ...
##  $ amount_paid     : num  1000 607 32057 26524 33981 ...
```

```
anesthesia.data = subset(idata, provider_type == "Anesthesiology")
head(anesthesia.data)
```

```
## Source: local data frame [6 x 7]
##
##     provider_ID gender entity_code state  provider_type place_of_service
## 171  1003062175      M           I    IL Anesthesiology                F
## 172  1003062175      M           I    IL Anesthesiology                F
## 173  1003062175      M           I    IL Anesthesiology                F
## 174  1003062175      M           I    IL Anesthesiology                F
## 175  1003062175      M           I    IL Anesthesiology                F
## 176  1003062175      M           I    IL Anesthesiology                F
## Variables not shown: amount_paid (dbl)
```

```
summary.anesthesia = aggregate(amount_paid ~ gender, data = anesthesia.data,
FUN = function(x) c(sum = sum(x), mean = mean(x), sd = sd(x)))
```

Male anesthesiologists received almost 6 times the amount received by female anesthesiologists.

```
anesthesia.gender  = cbind(summary.anesthesia[1],
(unlist(summary.anesthesia[,2])))
anesthesia.gender
```

```
##   gender       sum  mean    sd
## 1      F  26489655 19521 12473
## 2      M 158235655 20129 12045
```

However, when we do look at the ratio of Male to Female anesthesiologists 5:1, this explains the descrepancy

```
ddply(anesthesia.data,~gender,summarise,number_of_providers =
length(unique(provider_ID)))
```

```
##   gender number_of_providers
## 1      F                 264
## 2      M                1040
```

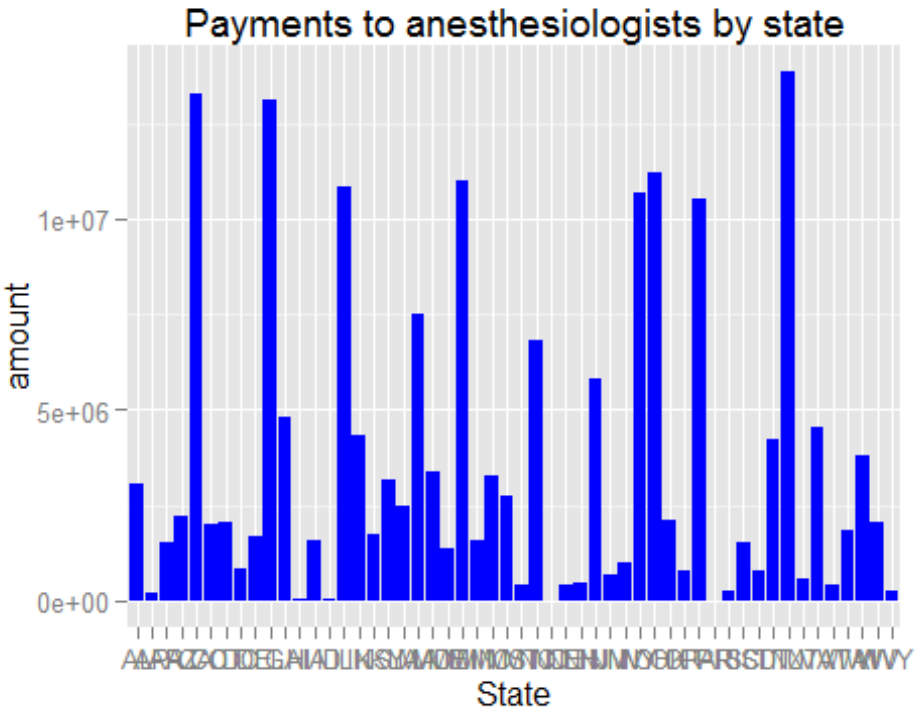
Looking a anesthesia payments per state, as expected the largest payments go to states with the largest Medicare Population Texas, California, Florida, and Ohio

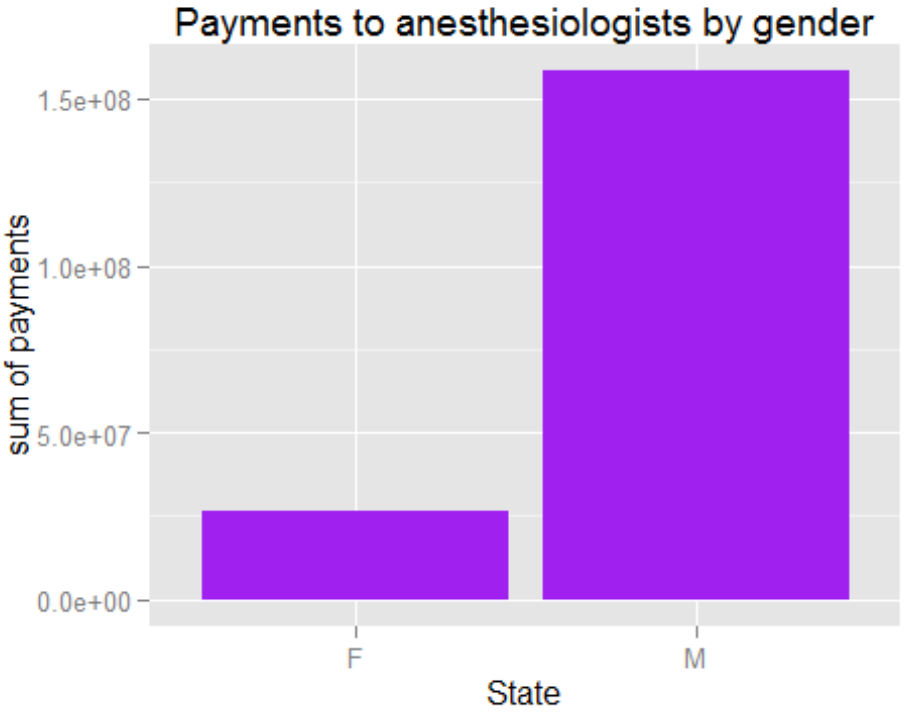
```
summary.state = aggregate(amount_paid ~ state, data = anesthesia.data, FUN =
function(x) c(sum = sum(x), mean = mean(x), sd = sd(x)))
anesthesia.state = cbind(summary.state[1], (unlist(summary.state[,2])))
#anesthesia.state
arrange(anesthesia.state, desc(sum))
```

```
##    state      sum  mean    sd
## 1     TX 13838808 19770 12543
## 2     CA 13253550 17671 11827
## 3     FL 13111931 19454 12534
## 4     OH 11188084 20604 11421
## 5     MI 10994089 21988 11637
## 6     IL 10842565 19607 12026
## 7     NY 10690152 17439 12375
## 8     PA 10529135 21444 12340
## 9     MA  7496389 21984 11960
## 10    NC  6835168 21630 11614
## 11    NJ  5822767 19474 13094
## 12    GA  4783471 20618 11739
## 13    VA  4547508 20484 11818
## 14    IN  4318303 17136 12323
## 15    TN  4202258 22838 11118
## 16    WI  3825204 21018 11739
## 17    MD  3356252 19513 12883
## 18    MO  3258915 22321 11450
## 19    KY  3171037 21719 11903
## 20    AL  3081129 22166 12447
## 21    MS  2751457 24349 10564
## 22    LA  2463425 22600 11705
## 23    AZ  2216780 16794 11624
## 24    OK  2112147 19926 11191
## 25    WV  2070081 25876 10321
## 26    CT  2037333 18190 12981
## 27    CO  1994803 20355 11828
```

```
## 28    WA  1823120 17700 11049
## 29    KS  1760693 22009 11107
## 30    DE  1673326 20917 12492
## 31    IA  1588160 22058 11188
## 32    MN  1551350 22814  9919
## 33    AR  1521868 19264 12074
## 34    SC  1511387 21591 12216
## 35    ME  1382256 19747 12311
## 36    NV   992865 13601 10524
## 37    DC   851243 19796 13408
## 38    SD   801943 23587 11388
## 39    OR   764590 17377 10808
## 40    NM   673963 19256 11900
## 41    UT   562952 18160 10330
## 42    NH   491119 19645 12260
## 43    MT   423017 21151 12619
## 44    VT   419509 17480 11876
## 45    NE   409529 22752 11311
## 46    WY   250883 16726  9716
## 47    RI   227972 28497  9256
## 48    AP   175054 17505 10291
## 49    HI    29429  7357  7941
## 50    ID    27786  9262  7069
## 51    ND    10708  5354   280
## 52    PR     7847  3924  1927
```

Graphics:





Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.