MSDS Assignment One

Etana Disasa

10/28/2018

SUMMARY STATS

This series provides detailed industry statistics by geographic area for establishments of firms with paid employees. Data are shown on the 2012 North American Industry Classification System (NAICS) basis. These data was acquired from The United States Census Bereau. https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk#

In order to access the file in R and to analyze the file, the following libraries would need to be loaded.

```r
library(data.table)
library(e1071)
```

Our data was downloaded as a *.csv file. The line below imports the data as "ecosens" from the ECN_2012_US_55A1 which is in our working folder.

```r
ecosens <- read.csv("ECN_2012_US_55A1/ECN_2012_US_55A1_with_ann.csv")
```

This dataset has 356 Observations of 11 variable. GEO.id and GEO.id2 identify each state and teritory of the US along with STATES which denotes the names of each. SECTORID code and SECTORS provide unique ids and labels to identify the business sectors. This data was collected in the year 2012 which is in the YEAR. SECTORS lists the number of establishments; RCPTOT displays total revenue (in $1,000); PAYANN displays annual payment (in $1,000); PAYQTR1 records the payments made in the first financial quarter; and EMP displays the number of employees. Below is the summary.

```r
summary(ecosens)
```

```
##          GEO.id       GEO.id2             STATES       SECTORID
##  0400000US01:  7   Min.   : 1.00   Alabama   :  7   Min.   :    55
##  0400000US02:  7   1st Qu.:16.00   Alaska    :  7   1st Qu.:   551
##  0400000US04:  7   Median :29.00   Arizona   :  7   Median : 55111
##  0400000US05:  7   Mean   :29.01   Arkansas  :  7   Mean   :244078
##  0400000US06:  7   3rd Qu.:42.00   California:  7   3rd Qu.:551112
##  0400000US08:  7   Max.   :56.00   Colorado  :  7   Max.   :551114
##  (Other)    :314                   (Other)   :314
##                                                   SECTORS         YEAR
##  Corporate, subsidiary, and regional managing offices: 51   Min.   :2012
##  Management of companies and enterprises             :204   1st Qu.:2012
##  Offices of bank holding companies                   : 50   Median :2012
##  Offices of other holding companies                  : 51   Mean   :2012
##                                                             3rd Qu.:2012
##                                                             Max.   :2012
##
##      ESTAB             RCPTOT              PAYANN
##  Min.   :   1.00   Min.   :    3445   Min.   :       0
##  1st Qu.:  97.75   1st Qu.:  294426   1st Qu.:  141530
##  Median : 376.00   Median :  792982   Median : 1440146
##  Mean   : 755.13   Mean   : 1683527   Mean   : 4634811
##  3rd Qu.:1084.00   3rd Qu.: 1527616   3rd Qu.: 5817461
##  Max.   :5116.00   Max.   :20971148   Max.   :34921334
```

```
##
##     PAYQTR1              EMP
##  Min.   :       0   Min.   :     0
##  1st Qu.:   34621   1st Qu.:  1528
##  Median :  410577   Median : 18532
##  Mean   : 1351255   Mean   : 46321
##  3rd Qu.: 1714494   3rd Qu.: 62088
##  Max.   :10635080   Max.   :288253
##
```

Callilng the dataset at a data table helps us maneuver during the analysis process because datatables are both datatable and dataframe datatypes. Presently our dataset is a dataframe datatye.

```
class(ecosens)
```

```
## [1] "data.frame"
```

Therefore, converting the datatype into a datatable format looks like this. The new format now is named ecosnse.dt.

```
ecosens.dt <- as.data.table(ecosens)
class(ecosens.dt)
```

```
## [1] "data.table" "data.frame"
```

The number of establishements in each state provides an explanation as to why the employee numbers vary. For example, the average (mean) establishment per state is displayed below.

```
head(ecosens.dt[, mean(ESTAB), by=STATES])
```

```
##        STATES        V1
## 1:    Alabama  450.0000
## 2:     Alaska  114.2857
## 3:    Arizona  658.5714
## 4:   Arkansas  894.2857
## 5: California 3654.2857
## 6:   Colorado  717.1429
```

```
## Or the number of establishments with in each sectors are displayed below.
ecosens.dt[, mean(ESTAB), by=SECTORS]
```

```
##                                                 SECTORS        V1
## 1:              Management of companies and enterprises 1054.2157
## 2:                    Offices of bank holding companies   38.9200
## 3:                   Offices of other holding companies  121.8039
## 4: Corporate, subsidiary, and regional managing offices  894.2549
```

General Employee Analysis

Below is a few lines of the total number of employed individuals in each state and territory in the United States.

```
head(ecosens.dt[, sum(EMP), by=STATES])
```

```
##        STATES     V1
## 1:    Alabama  92660
## 2:     Alaska  31370
## 3:    Arizona 226715
## 4:   Arkansas 199965
## 5: California 1384785
## 6:   Colorado  275265
```

```r
# These makes up for a total of employed individuals Of 16,490,387 which is displayed  from below.
sum(ecosens.dt$EMP)
```

```
## [1] 16490387
```

Summary of Annual Payments made to employees in $1,000s.

```r
# The minimum amout of payment made by each sector is displayed below.
ecosens.dt[, min(PAYANN), by=SECTORS]
```

```
##                                                 SECTORS    V1
## 1:              Management of companies and enterprises   233
## 2:                    Offices of bank holding companies     0
## 3:                   Offices of other holding companies  3195
## 4: Corporate, subsidiary, and regional managing offices 72425
```

```r
# The median annual payment for employees in each sector is displayed below.
ecosens.dt[, median(PAYANN), by=SECTORS]
```

```
##                                                 SECTORS      V1
## 1:              Management of companies and enterprises 5635705
## 2:                    Offices of bank holding companies   25259
## 3:                   Offices of other holding companies   70391
## 4: Corporate, subsidiary, and regional managing offices 2999350
```

```r
# The average(mean) annual payment for employees is displayed below.
ecosens.dt[, mean(PAYANN), by=SECTORS]
```

```
##                                                 SECTORS         V1
## 1:              Management of companies and enterprises 6565523.25
## 2:                    Offices of bank holding companies   59881.06
## 3:                   Offices of other holding companies  142441.98
## 4: Corporate, subsidiary, and regional managing offices 5889559.41
```

```r
# The max annual payment for employees is displayed below.
ecosens.dt[, max(PAYANN), by=SECTORS]
```

```
##                                                 SECTORS       V1
## 1:              Management of companies and enterprises 34921334
## 2:                    Offices of bank holding companies   346976
## 3:                   Offices of other holding companies  1011694
## 4: Corporate, subsidiary, and regional managing offices 34137001
```

```r
# Standard Deviation by each sector is as follows.
ecosens.dt[, sd(PAYANN), by=SECTORS]
```

```
##                                                 SECTORS         V1
## 1:              Management of companies and enterprises 7460635.16
## 2:                    Offices of bank holding companies   69035.31
## 3:                   Offices of other holding companies  200916.57
## 4: Corporate, subsidiary, and regional managing offices 7488360.71
```

```r
# Across the United States, and across each sector, the summary of annual payment is as follows.
summary(ecosens.dt$PAYANN)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##        0   141530  1440146  4634811  5817461 34921334
```

```
# Standard Deviation
sd(ecosens.dt$PAYANN)
```

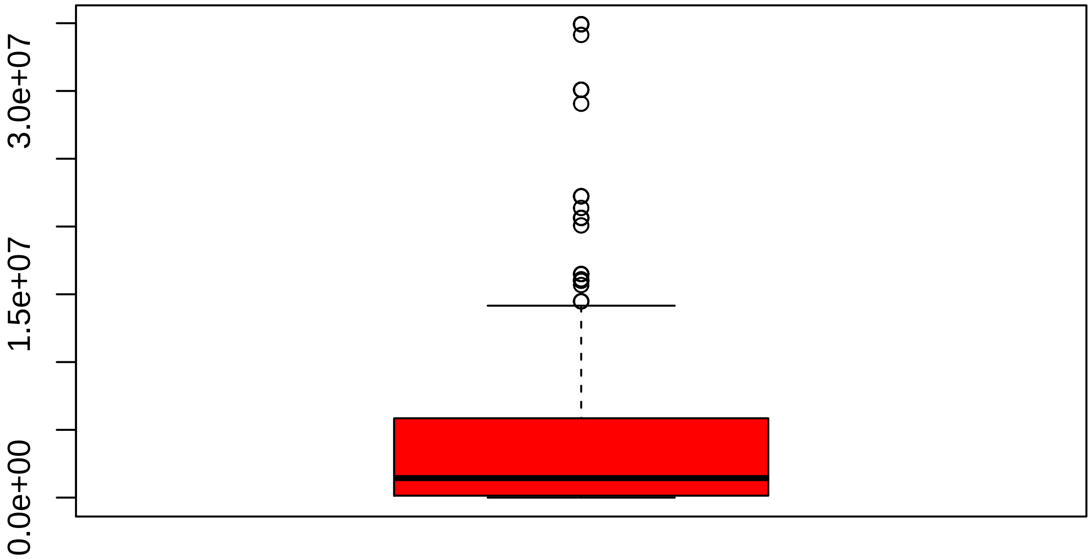
```
## [1] 6924484
```

PLOTTING

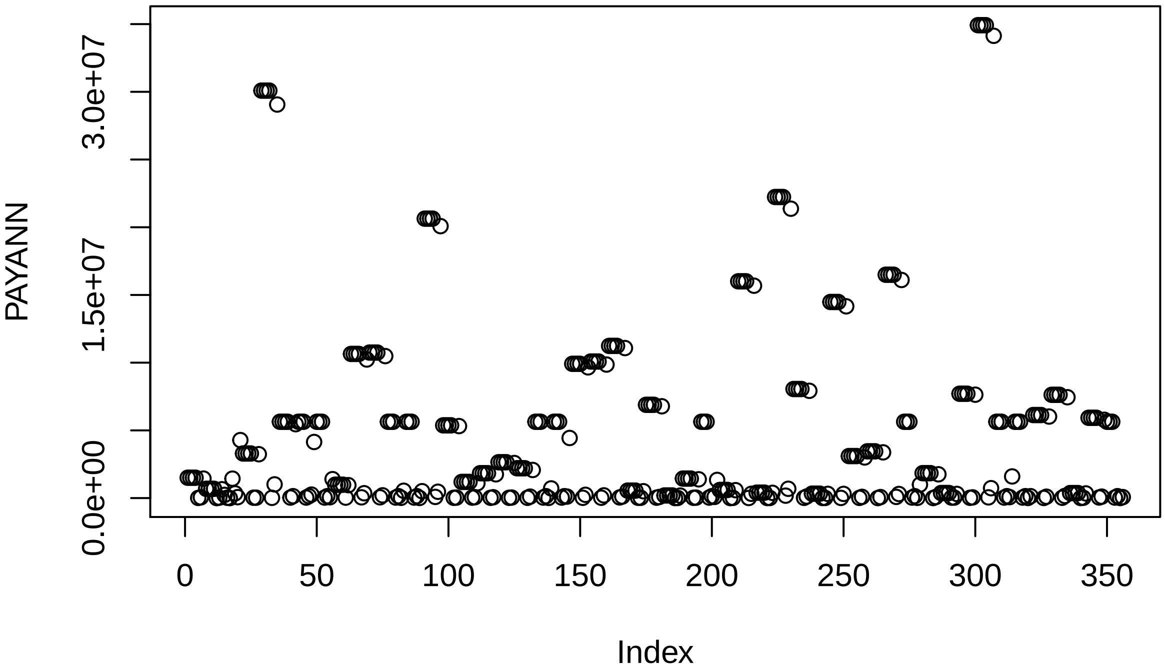
Below are boxplot, plot and histogram presentation of the annual payment (PAYANN) distribution made in the year 2012.

```
library(ggplot2)
# Boxplot
ecosens.dt [, boxplot(PAYANN, col="red")]
```

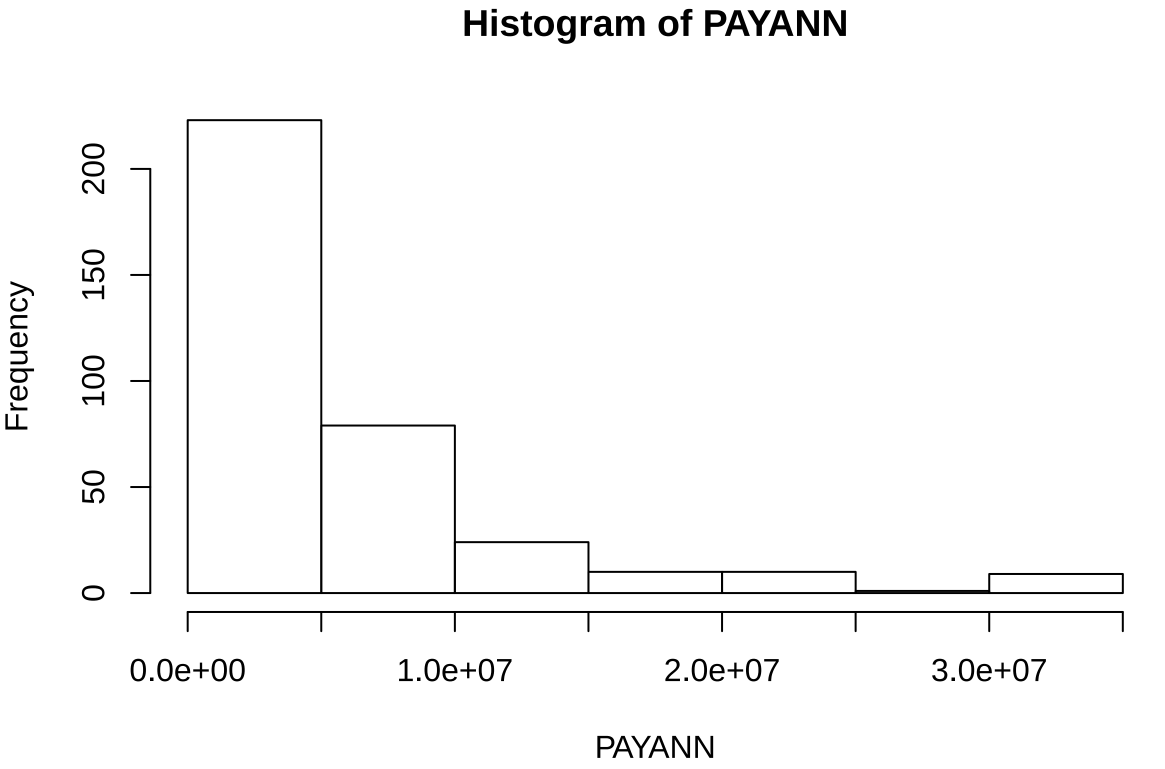
```
## Warning in data.table(stats = structure(c(0, 141104, 1440146, 5854485.5, :
## Item 1 is of size 5 but maximum size is 34 (recycled leaving remainder of 4
## items)
```



```
# Plot
ecosens.dt [, plot(PAYANN)]
```



```
# Histogram
ecosens.dt [, hist(PAYANN)]
```



Discussing the histogram

The histogram above appears to be positively skewed or skewed to the right. This is due to the fact that the mean (average) is greater than the median. The median is also very close the 1st quartile (25% quantile) which is the reason for the skewness.

BONUS STATISTICS

For the next section, we need "e1071" and "moments" libraries installed and loaded.

```
library(moments)
```

```
##
## Attaching package: 'moments'
```

```
## The following objects are masked from 'package:e1071':
##
##     kurtosis, moment, skewness
```

```
library(e1071)
```

Kurtosis

The kurtosis of payment is positive, which indicates that the payment distribution is mesokurtic. This is consistent with the fact that its histogram is not bell-shaped but skewed to the right.

```
ecosens.dt [, kurtosis(PAYANN)]
```

```
## [1] 8.92431
```

Skewness

The skewness of payment is 2.329521. It indicates that the payment distribution is skewed towards the right as it is displayed in the historam above.

```
ecosens.dt [, skewness(PAYANN)]
```

```
## [1] 2.329521
```

Moments

```

ecosens.dt [, all.moments(PAYANN)]
```

```
## [1] 1.000000e+00 4.634811e+06 6.929527e+13
```

Correlations

The correlation coefficient of PAYMENT (total payments made) and PAYQTR1 (payments made in the first quarter) is 0.90081. Since it is rather close to 1, we can conclude that the variables are positively linearly related.

```
ecosens.dt [, cor(PAYANN,PAYQTR1)]
```

```
## [1] 0.9890183
```

Quantiles

The quartiles are quantile values of each quarter. Here the 0% (min value) the 25% (1st quartile value), the 50% (2nd quartile and also the median), the 75% (3rd quartile value) and 100% (max value) are also displayed above under the summary of the PAYANN distribution.

```
ecosens.dt [, quantile(PAYANN)]
```

```
##         0%        25%        50%        75%       100%
##        0.0   141530.5  1440146.0  5817461.2 34921334.0
```

```
# Other than quartiles, quantiles could also be sepcified as shown below.
ecosens.dt [, quantile(PAYANN, c(.333, .666, 1.0))]
```

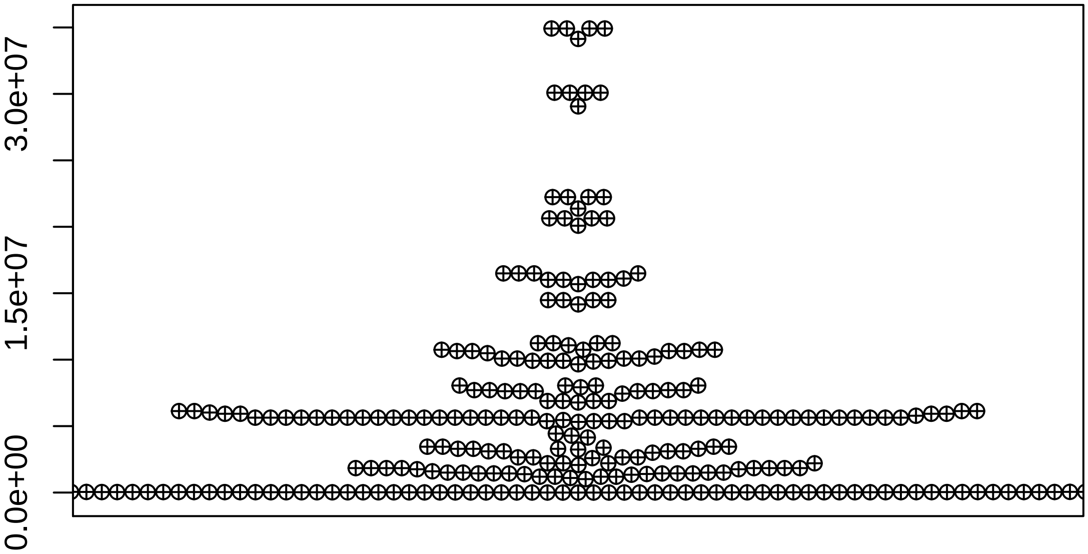
```
##    33.3%    66.6%     100%
##   327001  5635705 34921334
```

```
# This quantile values displayed are the thirdth of the PAYANN distribution.
```

BONUS PLOTS

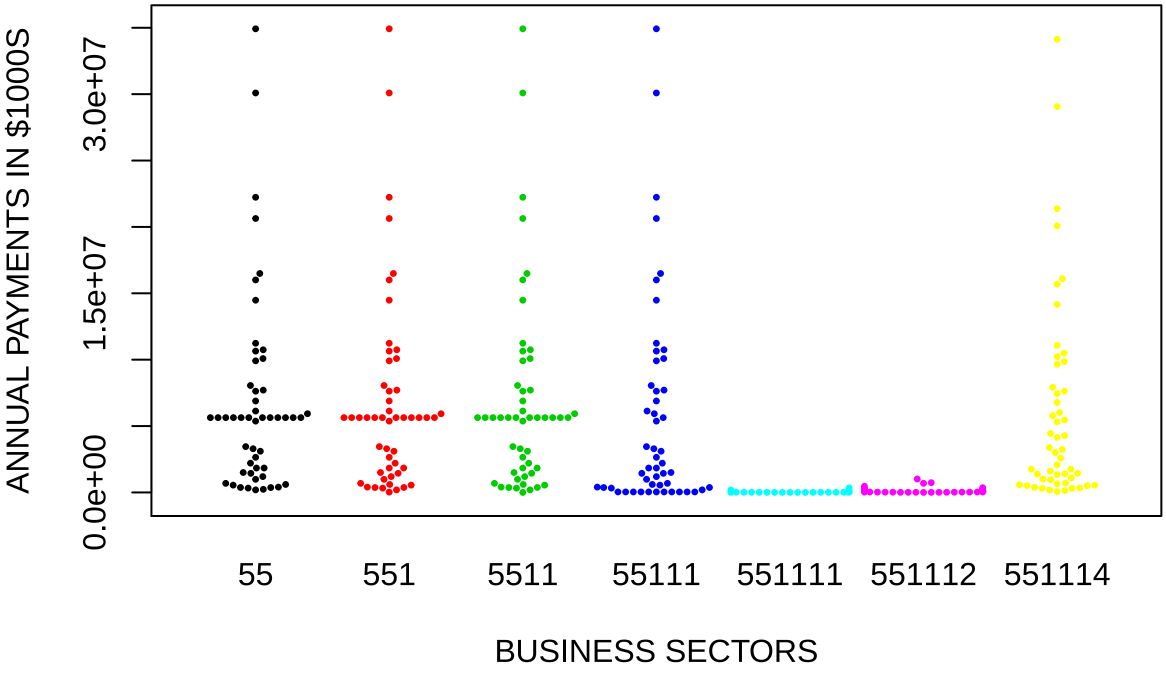
The following plots used here are Bee Swarm plots. The library "beeswarM" is downloaded, and loaded here from below. This plotting tool displays a distribution using one axis and plotting values left and right.

Beeswarm with one variable



Beeswarm with multiple variables

Below, the same tool is used to display PAYANN distribution against SECTORID. Furthermore, more attributes are included to color code the different SECTORID,lable the x and y axises and make sure the plots do not overlap.



RESOURCES

The United States Census Bereau. https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk#

R-TUTORIAL: An R Introduction to Statistics http://www.r-tutor.com/

R Documentation https://www.rdocumentation.org