Final Project - Justice & Legal System Analysis by States

Patkhedkar Ninad

2020-11-19

Introduction

Effective Law and Order is one of the most important necessity of society. It has direct effect on living standard and crime rate in that area. The whole "Law and Order" aka Justice system consists of multiple functions which performs duties as below

- **Police Protection Services** - law enforcement, patrolling, traffic safety, parking meter read, animal warden etc.
- **Judicial and Legal Services** - civil and criminal functions of courts, state's attorneys, court reporters, register of wills etc.
- **Correction Services** - prisons, reformatories, rehabilitation centers, parole boards, pardon boards etc.

To effectively govern any area, all these functions need resources i.e. people to perform duties and funding.

In this project I will use public dataset about state wise employment and expenditures on various functions. I will investigate and find more information about how effectively Justice and Legal system is managed in different states. I will also try to find any correlation between population of state, expenses and number of employees etc.

Problem Statement Addressed

What are the factors that affect overall cost of Justice & Legal system?

Research Questions

1. How state population affects overall cost and per capita cost?
2. How number of employees in police protection services affect the cost of police protection?
3. How number of employees in judicial and legal services affect the cost of judicial and legal functions?
4. How number of employees in correction services affect the cost of correction functions?
5. Which states pays attractive salaries to police employees?

Data

Dataset source - Public dataset for fiscal 2016 from Bureau of Justice Statistics Justice Expenditure and Employment Extracts

Dataset description - Presents estimates of government expenditures and employment at the state level for the following justice categories: police protection (the function of enforcing the law), all judicial and legal functions (including prosecution, courts, and public defense), and corrections

Dataset files
1. jeee16t03.csv - Percent distribution of expenditure for the justice system by type of government
2. jeee16t08 - Per capita justice expenditure and full-time equivalent justice employment per 10,000 population
3. jeee16t11.csv - Justice system employment and payrolls of state governments

For Definitions, Methodology and other information, refer detailed guide - https://www.bjs.gov/content/pub/pdf/jeeeguide.pdf

How you addressed this problem statement

First I will clean each dataset file and create a consolidated dataset. Then I will identify and filter variables of interest. I will generate correlation,linear regression, and basic summary findings I will plot my findings to visualize the data results.

My initial hypothesis is population increases overall cost but should decrese per capita cost.

Data Prep

I will cleaup each data set and and create a consolidated dataset by "State" in this section. Here are few things I will perform as part of cleanup

- Simplify column names by coverting to lower case and spaces replaced by underscores
- Rename long columnn names e.g. population_2016_thousands to population_k
- Remove records of Total federal govt, Local city govt, Muncipalty govt data and keep only State data
- Remove records where population is not present
- Trim State abbreviations from records i.e. Virginia (VA) -> Virginia

```
library("dplyr")
library("janitor")
library("ggplot2")
library("knitr")
library("kableExtra")
library("tidyverse")
library("papeR")

options("width"=200)
options(scipen=999)
options(knitr.duplicate.label = "allow")
setwd("/cloud/project/completed/final_project")
# cleanup jeee16t03.csv
df_jee03 <- read.csv("jeee16t03.csv")
df_jee03 <- df_jee03 %>% clean_names() %>%
              rename(state = state_and_type_of_government) %>%
              rename(population_k = population_2016_thousands)
df_jee03 <- df_jee03 %>% filter(population_k != "-")
df_jee03$population_k <- suppressWarnings(as.numeric(as.character(df_jee03$population_k)))
df_jee03 <- df_jee03[-1,]
df_jee03$state <- sub("\\(.*", "",df_jee03$state )
trim <- function (x) gsub("^\\s+|\\s+$", "", x)
df_jee03$state <- trim(df_jee03$state)
# cleanup jeee16t08.csv
df_jee08 <- read.csv("jeee16t08.csv")
df_jee08 <-df_jee08 %>% clean_names() %>%
```

```
              rename(population = population_2016)
df_jee08 <-filter(df_jee08, state != "Total")
# cleanup jeee16t11.csv
df_jee11 <- read.csv("jeee16t11.csv")
df_jee11 <- df_jee11 %>% clean_names() %>% filter(state != "Total")
df_jee11$pp_average_earnings <- suppressWarnings(as.numeric(as.character(df_jee11$pp_average_earnings)))
df_jee11 <- na.omit(df_jee11)
```

Now consolidate the dataset by joining on State field.

```
df_consolidated <- inner_join(df_jee08, df_jee03, by = "state")
df_consolidated <- inner_join(df_consolidated, df_jee11, by = "state")
```

We got 2 population fields in consolidated dataset.
population_k - represents population of state in thousands . . . basically round figure(k - stands for 1000)
population - represents actual count of population
We will keep field which represents population in thousands as its easy to follow for analysis. We will drop other population field.

```
df_consolidated <- df_consolidated %>% select(-c(population))
```

Now we have total 49 Observations and 41 variables in consolidated dataset. Though there are 50 states, data for state of "Hawai" had missing fields and hence omitted by na.omit() used above.I have identified below variables for my analysis work. These variables are from all 3 different files used.

- *state* - US state name
- *population_k* - Population in thousands
- *total_direct_expenditure* - Total expenditure in Thousands Dollars
- *total_justice_system_pc* - Per capita (10,000 population) cost in Dollars for justice system which include police, judicial and civil, correction functions
- *pp_total_employees* - Police protection services employees
- *jl_total_employees* - Judicial and Legal services employees
- *c_total_employees* - Correction services employees
- *police_protection_amount* - Police protection cost
- *judician_and_legal_amount* - Judicial and Legal cost
- *corrections_amount* - Correction services cost
- *pp_average_earnings* - Police protection average earnings in Dollars

```
df_consolidated <- df_consolidated %>% select(c(state,population_k,total_direct_expenditure,total_justi
str(df_consolidated)
```

```
## 'data.frame':    49 obs. of  11 variables:
##  $ state                     : chr  "Alabama" "Alaska" "Arizona" "Arkansas" ...
##  $ population_k              : num  4865 742 6945 2990 39209 ...
##  $ total_direct_expenditure : num  45277563 15808697 58975013 27299957 532948138 ...
##  $ total_justice_system_pc  : num  480 1298 710 504 1064 ...
##  $ pp_total_employees       : int  1303 683 1963 1223 11444 1274 2142 1100 4410 2625 ...
##  $ jl_total_employees       : int  3167 1406 2416 1667 6569 5191 6221 1835 19872 3571 ...
##  $ c_total_employees        : int  4664 2271 9700 5563 57809 7413 5834 2944 23740 16587 ...
##  $ police_protection_amount : int  1251270 370209 2261558 691059 17570133 1873320 1236997 348027 7848
##  $ judician_and_legal_amount: int  362060 254000 983419 220343 8675761 754162 826903 207990 2366274
##  $ corrections_amount       : int  722269 338005 1684710 595731 15468283 1313103 684159 308341 42485
##  $ pp_average_earnings      : num  3987 6649 5163 3877 8398 ...
```

So overall data selected for analysis looks like as below

Table 1: Justice and Legal System - 2016

state	population_k	total_direct_expenditure	total_justice_system_pc	pp_average_earnings
Alabama	4865	45277563	480.11	3987
Alaska	742	15808697	1297.65	6649
Arizona	6945	58975013	709.77	5163
Arkansas	2990	27299957	503.99	3877
California	39209	532948138	1063.89	8398
Colorado	5541	57293994	711.18	5981

```
kable(head(df_consolidated[,c("state","population_k","total_direct_expenditure","total_justice_system_p
```

Lets Summerize the data

```
summary(df_consolidated)
```

```
##     state            population_k   total_direct_expenditure total_justice_system_pc pp_total_employ
##  Length:49          Min.   :  584   Min.   :  7774956        Min.   : 450.1          Min.   :  200
##  Class :character   1st Qu.: 1906   1st Qu.: 23028337        1st Qu.: 555.8          1st Qu.:  834
##  Mode  :character   Median : 4678   Median : 47492664        Median : 660.3          Median : 1573
##                     Mean   : 6550   Mean   : 71056579        Mean   : 678.4          Mean   : 2108
##                     3rd Qu.: 7295   3rd Qu.: 82188266        3rd Qu.: 741.9          3rd Qu.: 2605
##                     Max.   :39209   Max.   :532948138        Max.   :1297.7          Max.   :11444
##  corrections_amount pp_average_earnings
##  Min.   :  137103   Min.   :3837
##  1st Qu.:  444268   1st Qu.:4656
##  Median :  903548   Median :5222
##  Mean   : 1584783   Mean   :5611
##  3rd Qu.: 1823340   3rd Qu.:6607
##  Max.   :15468283   Max.   :9058
```

```
kable(summarize(df_consolidated, type = "numeric"))
```

	N		Mean	SD		Min	Q1	Median	
population_k	49		6550.14	7308.96		584.00	1906.00	4678.00	72
total_direct_expenditure	49		71056579.41	91621466.80		7774956.00	23028337.00	47492664.00	821882
total_justice_system_pc	49		678.44	173.82		450.08	555.79	660.33	7
pp_total_employees	49		2107.80	2091.37		200.00	834.00	1573.00	26
jl_total_employees	49		3602.49	4108.31		470.00	1186.00	2287.00	40
c_total_employees	49		9028.63	10599.68		799.00	2927.00	5563.00	121
police_protection_amount	49		2207300.51	3039421.76		188210.00	458774.00	1251270.00	25019
judician_and_legal_amount	49		935305.20	1377625.10		80521.00	254000.00	571764.00	10355
corrections_amount	49		1584782.67	2433893.46		137103.00	444268.00	903548.00	18233
pp_average_earnings	49		5611.20	1370.49		3837.00	4656.00	5222.00	66

The dataset mostly contains continous variables. There are not categoriacal variables. Only State is categorical variable.

Analysis

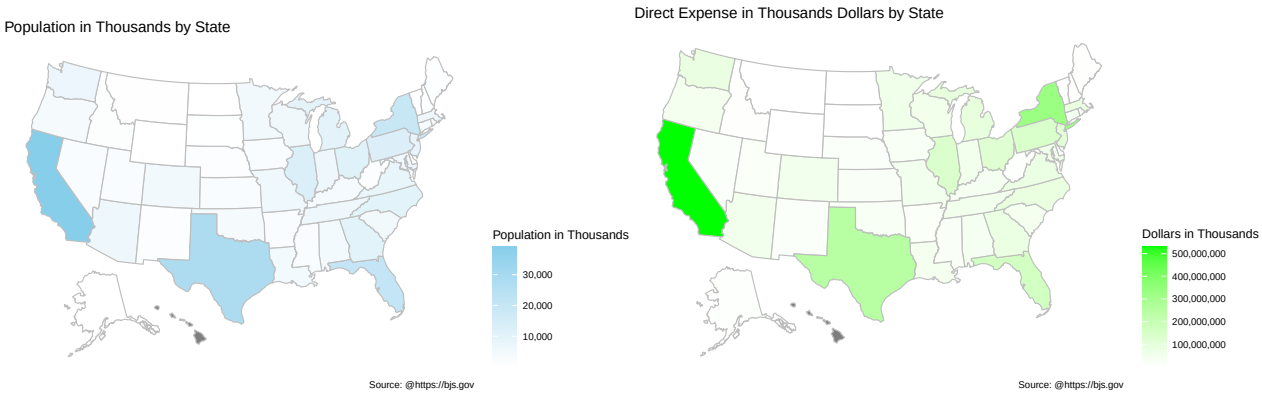
My initial hypothesis is overall cost of Legal and Justice system increases with Population. However at same time per Capta cost of system decreses with Population.

Population and Direct Expenditure relationship

Lets see how population and direct cost of overall justice system looks like

```
library("usmap")
plot_usmap(data = df_consolidated, values = "population_k",  color = "grey", labels=FALSE) +
  scale_fill_continuous( low = "white", high = "skyblue",
                         name = "Population in Thousands", label = scales::comma
  ) +
  theme(legend.position = "right") +
  labs(title = "Population in Thousands by State", caption = "Source: @https://bjs.gov")

plot_usmap(data = df_consolidated, values = "total_direct_expenditure",  color = "grey", labels=FALSE) +
  scale_fill_continuous( low = "white", high = "green",
                         name = "Dollars in Thousands", label = scales::comma
  ) +
  theme(legend.position = "right") +
  labs(title = "Direct Expense in Thousands Dollars by State", caption = "Source: @https://bjs.gov")
```



As expected, I see similar patten on US map for population and cost. Lets check the co-relation between 2 variables.

```
cor(df_consolidated$population_k,df_consolidated$total_direct_expenditure)
```

```
## [1] 0.960584
```

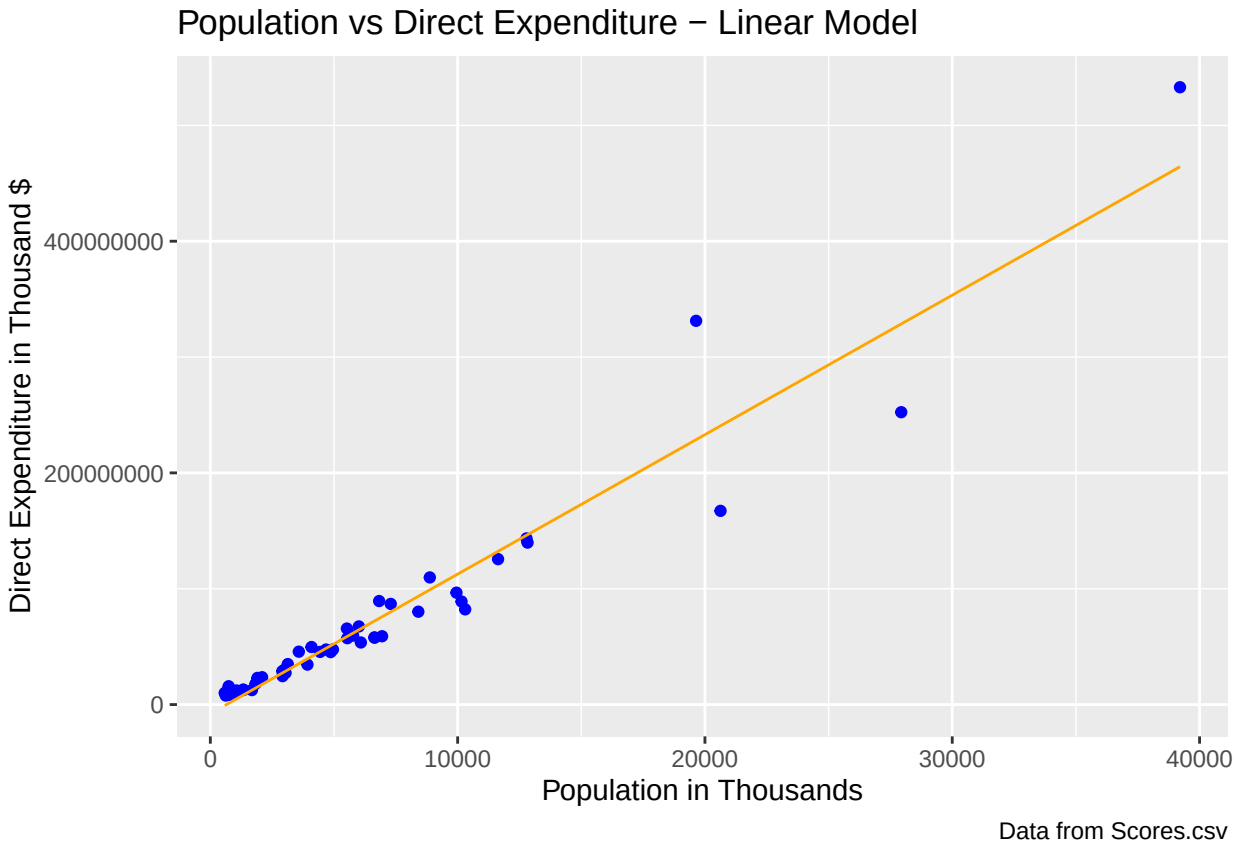
There is strong positive correlation. As population increases, it directly increases overall cost of justice and legal system. I will use linear regression model and check how it fits.

```
direct_expense_lm <-  lm(df_consolidated$total_direct_expenditure ~ df_consolidated$population_k)
summary(direct_expense_lm)
```

```
##
## Call:
## lm(formula = df_consolidated$total_direct_expenditure ~ df_consolidated$population_k)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -76183410  -5487534   1554209   7254478 102580965
##
## Coefficients:
##                               Estimate Std. Error t value            Pr(>|t|)
## (Intercept)                 -7816323.4  4960441.1  -1.576               0.122
## df_consolidated$population_k    12041.4      508.3  23.689 <0.0000000000000002 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25740000 on 47 degrees of freedom
## Multiple R-squared:  0.9227, Adjusted R-squared:  0.9211
## F-statistic: 561.2 on 1 and 47 DF,  p-value: < 0.00000000000000022
```

```
direct_expense_predict_df <- data.frame(total_direct_expenditure = predict(direct_expense_lm, df_consoli
## Plot the predictions against the original data
ggplot(data = df_consolidated, aes(y = total_direct_expenditure, x = population_k)) +
  geom_point(color='blue') +
  geom_line(color='orange',data = direct_expense_predict_df, aes(y=total_direct_expenditure, x=populati
  labs(
    title = "Population vs Direct Expenditure - Linear Model",
    caption = "Data from Scores.csv",
    x = "Population in Thousands",
    y = "Direct Expenditure in Thousand $"
  )
```



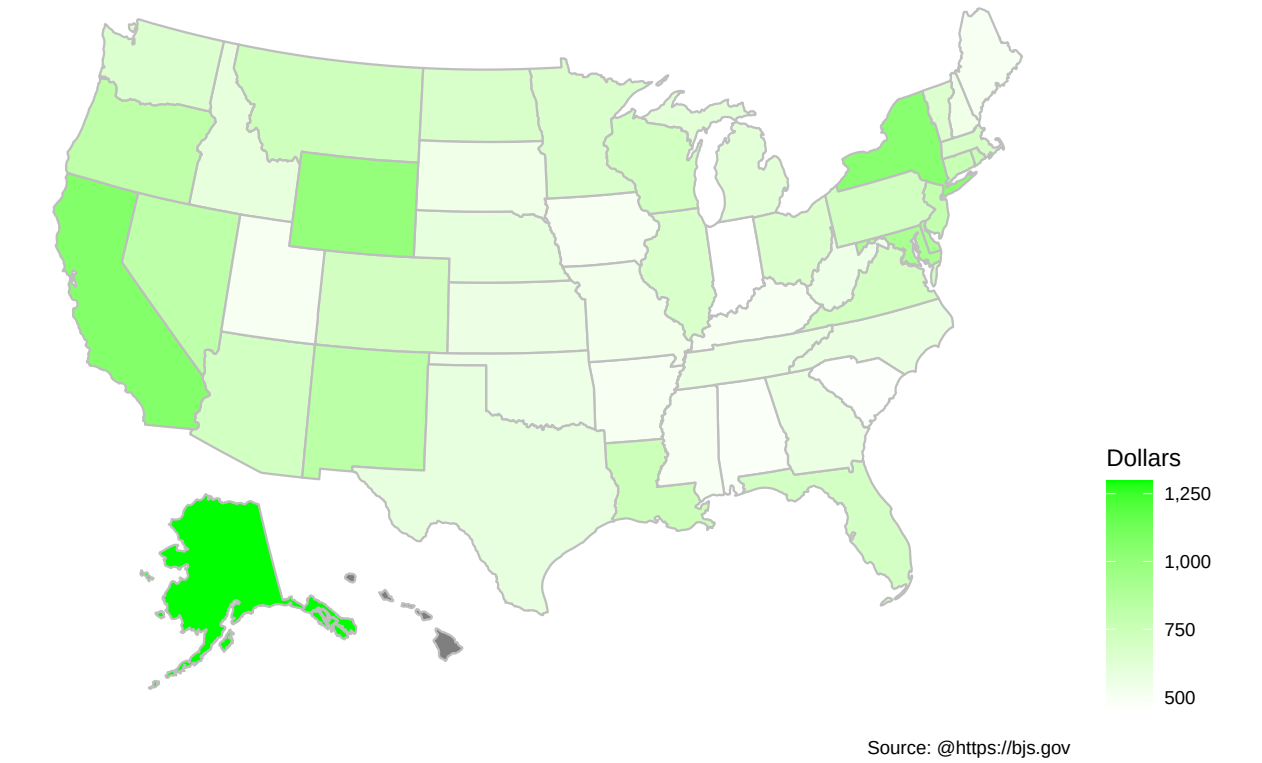
Looking at p-value and R-square model appears to correct.

Population and Per Capita Cost relationship

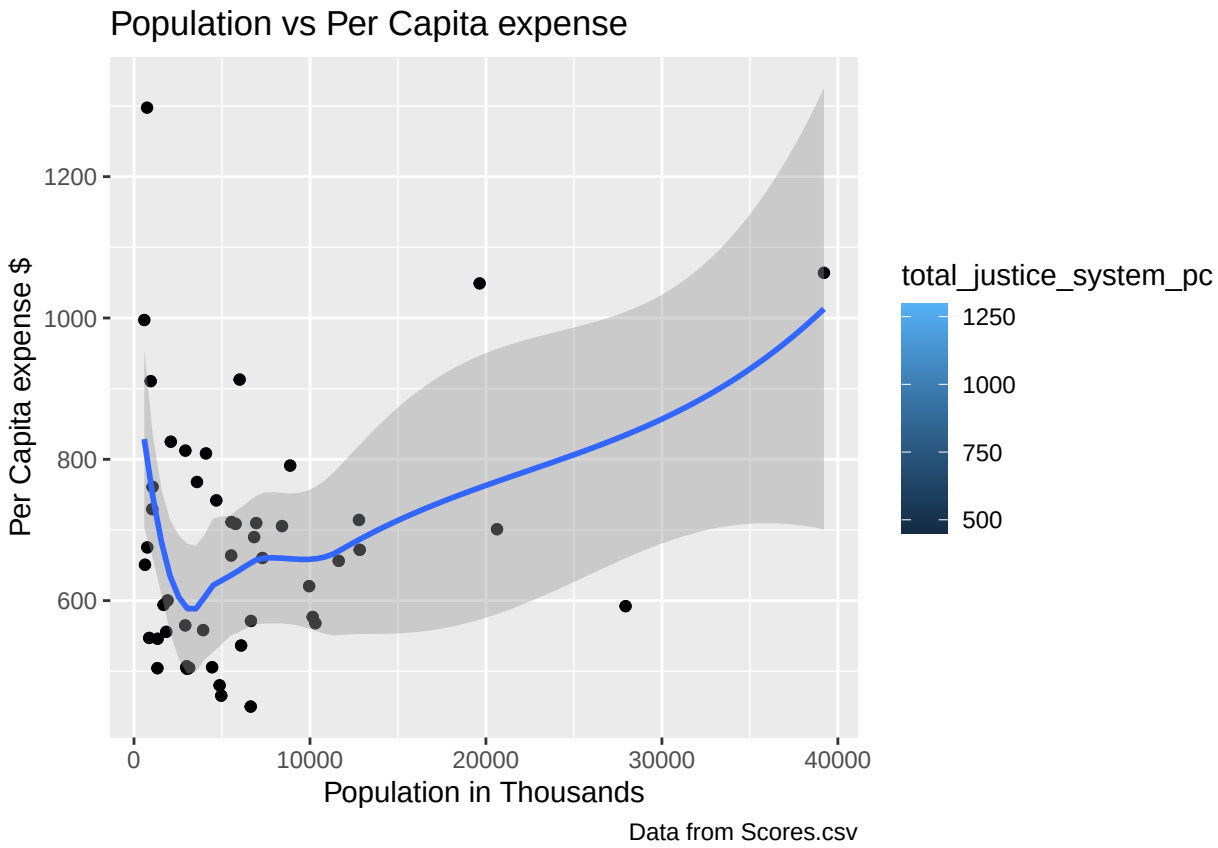
Lets check the correlation between population and per capita cost of justice system.

```
plot_usmap(data = df_consolidated, values = "total_justice_system_pc",  color = "grey", labels=FALSE) +
  scale_fill_continuous( low = "white", high = "green",
                         name = "Dollars", label = scales::comma
```

```
  ) +
  theme(legend.position = "right") +
  labs(title = "Justice system cost per capita by State", caption = "Source: @https://bjs.gov")
```




```
ggplot(df_consolidated, aes(x=population_k, y=total_justice_system_pc)) +
  geom_point(aes(color = total_justice_system_pc)) +
  geom_point() + geom_smooth() +
  labs(
    title = "Population vs Per Capita expense",
    caption = "Data from Scores.csv",
    x = "Population in Thousands",
    y = "Per Capita expense $"
  )
```




```
cor(df_consolidated$population_k,df_consolidated$total_justice_system_pc)
```

```
## [1] 0.2278301
```

Map shows per capita cost hasn't decresed with state population. For states like California it is still high. My hypothesis was per capita cost would decrese with increase in population. Hence I was expecting strong negative correlation value. However correlation value shows weak positive correlation.

It means my hypothesis was wrong.

Police protection employee and cost relationship

```
cor(df_consolidated$pp_total_employees,df_consolidated$police_protection_amount)
```

```
## [1] 0.9177783
```

There is strong positive correlation between number of employees in Police protection services and expenditure It means as number of employees in department increases, the expenditure also increases.

Judician and Legal services employee and cost relationship

```
cor(df_consolidated$jl_total_employees,df_consolidated$judician_and_legal_amount)
```

```
## [1] 0.5162116
```

There is moderate positive correlation between number of employees in Police protection services and expenditure It means as number of employees in department increases, the expenditure increases moderatley.

This may be because Judicain and Legal system has other sources of revenues like court fees, penalties, motor vehicle registration, tax, license fees etc.

Correction services employee and cost relationship

```
cor(df_consolidated$c_total_employees,df_consolidated$corrections_amount)
```

```
## [1] 0.9433818
```

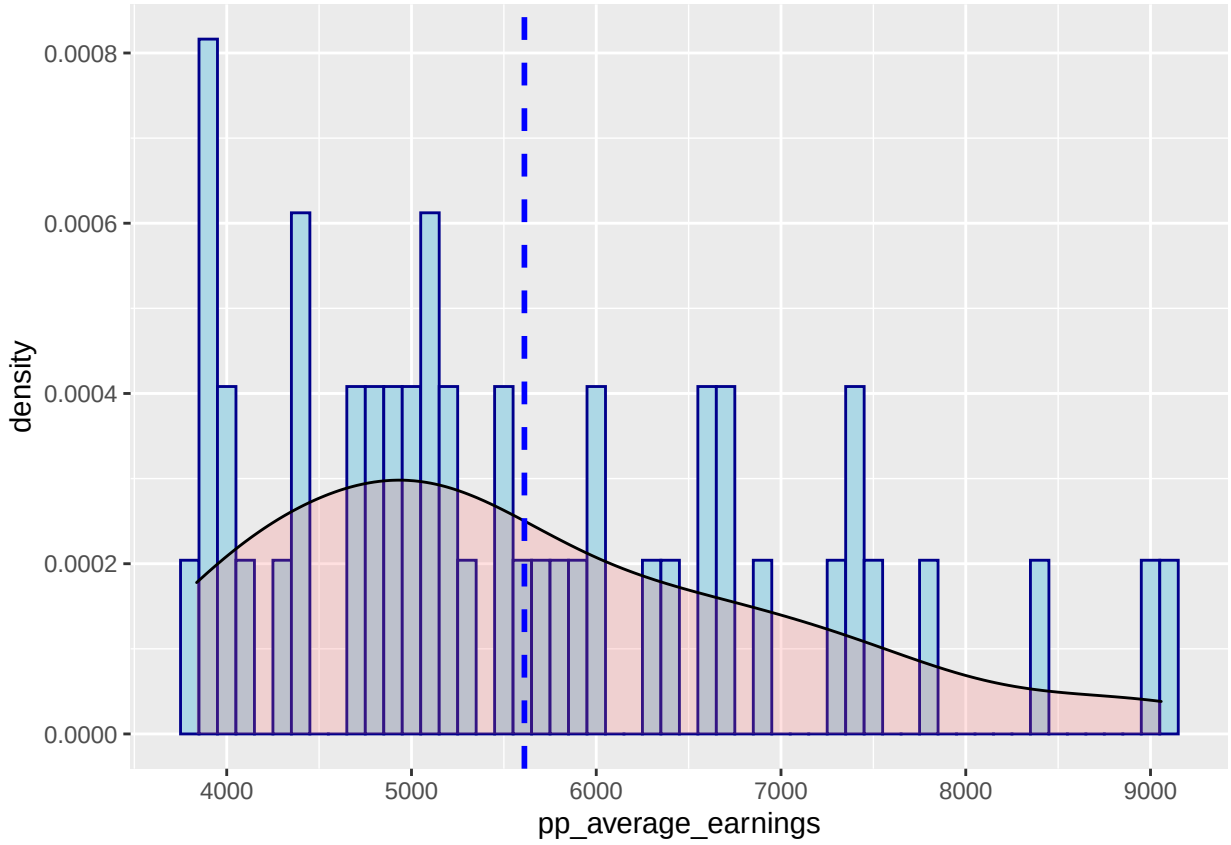
There is strong positive correlation between number of employees in correction services and expenditure It means as number of employees in department increases, the expenditure of department also increases.

Police Protection staff earnings distribution

I was interested to check the distribution of earnings for police protection staff. They performed very important duty in our society. Which states would attractive destination for people looking forward for police protection jobs.

Lets check distribution with density plot.

```
p <- ggplot(df_consolidated, aes(x=pp_average_earnings)) +
  geom_histogram(aes(y=..density..),color="darkblue", fill="lightblue",binwidth = 100) +
  geom_density(alpha=.2, fill="#FF6666") +
  geom_vline(aes(xintercept=mean(pp_average_earnings)),
             color="blue", linetype="dashed", size=1)
p
```



Distributio is positively skewed. In majority of states earning looks to range between 4000 to 6000.

Table 2: Lucrative states for Police Protection jobs

	state	pp_average_earnings
2	Alaska	6649
5	California	8398
7	Connecticut	7381
8	Delaware	6665
12	Illinois	7387
17	Louisiana	7252
20	Massachusetts	9058
27	Nevada	6692
29	New Jersey	7526
31	New York	8999
38	Rhode Island	7823
42	Texas	6903

For better earnings, we will just filter out states where earnings are more than 3 quartile i.e. 75+ percentile. I will use quantile function.

```
pp_earning_df <- df_consolidated[ df_consolidated$pp_average_earnings > quantile(df_consolidated$pp_ave
kable(head(pp_earning_df[,c('state','pp_average_earnings')],50),caption="Lucrative states for Police Pr
```

Implications

This consolidated dataset provides lot of datapoints on how Justsice System in each state spends its funding.It allows policy makers to understand expenditure in various functions and provides insight into optimization as well as investments required in each area. These insights can be used for prediction of budgets for future years. The long term implication is on well being of society in the area.

Limitations

Its natural to compare different states on various variables in dataset. However it is important to note that, each State government is different and handles responsibilities differently. There is variation in scope of services offered by each state government. E.g. some state governments directly administer certain activities that elsewhere are undertaken by local governments, with or without fiscal aid. There is also variation in the division of responsibilities that exist for counties and cities.Governemntal structure, degree of,urbanization, and population density may affect the comparability of expenditure and employment data.

Concluding Remarks

Its natural to observe incrase in overall expenditure on justice system with increased population and staffing to perform duties. However per capita expenditure doesn't decrease with increased population. It means there are other non self-evident facotors which are adding a to overall cost. These factors could be buildings, communication devices, equipments, utility services,contracting services, forensic labs, retirement benefits to ex-employees etc.

Police protection and correction services expenditure increase with increase in employess with strong correlation. However expenditure in judician and legal services increase less moderately with increase in employyes. It

means Judicial and Legal services have sources of revenues from court fees, penalties, registration services, inspection services, license services etc.